JEP-TALN-RECITAL 2012

JEP : Journées d'Études sur la Parole
TALN : Traitement Automatique des Langues Naturelles
RECITAL : Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues

Actes de la conférence conjointe JEP-TALN-RECITAL 2012

Volume 3 : RECITAL

Éditeurs
Jorge Mauricio Molina Mejia
Didier Schwab
Gilles Sérasset

4 – 8 Juin 2012
Grenoble, France

Des versions imprimées de ces actes peuvent être achetées auprès de :

GETALP-LIG
Laurent Besacier
BP 53
38041 Grenoble Cedex 9
France
Laurent.Besacier@imag.fr

Préface

Pour la quatrième fois, après Nancy en 2002, Fès en 2004, et Avignon en 2008, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisent conjointement leurs principales conférences afin de réunir en un seul lieu les deux communautés du traitement de la parole et de la langue écrite pour favoriser les interactions entre nos deux communautés.

Plus précisément, la conférence JEP-TALN-RECITAL'2012 réunit cette année la vingt-neuvième édition des Journées d'Étude sur la Parole (JEP'2012), la dix-neuvième édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012) et la quinzième édition des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2012).

Nous avons souhaité organiser cet événement sur le campus universitaire de l'université de Grenoble, au plus proche des trois laboratoires co-organisateurs (LIG, LIDILEM, GIPSA-Lab). L'université Stendhal-Grenoble 3 (consacrée aux disciplines des humanités) nous accueille dans ses locaux à cette occasion.

Par ailleurs, JEP-TALN-RECITAL'2012 accueille quatre ateliers ; la septième édition du « DÉfi Fouille de Texte » (DEFT), la seconde édition du « DEfi Geste Langue et Signe » (DEGELS), ainsi que deux nouveaux auxquels nous souhaitons longue vie : « Interactions Langagières pour personnes Agées Dans les habitats Intelligents » (ILADI) et « Traitement Automatique des Langues Africaines – écrit et parole » (TALAf). Quatre conférenciers renommés ont accepté notre invitation pour des sessions plénières communes. Nous espérons que leur hauteur de vue et leur ouverture d'esprit permettront des discussions intéressantes et ouvriront des perspectives prometteuses.

Quelques informations sur les processus de sélection pour cette édition sont présentées ci-dessous. Nous remercions tous les relecteurs et membres des différents comités de programme pour leur travail ainsi que nos sociétés savantes : l'AFCP et l'ATALA (avec son comité permanent qui assure la continuité de la forme et du fond entre les diverses éditions).

Nous avons reçu 62 propositions d'articles longs pour TALN, parmi lesquels 24 ont été sélectionnés au moyen d'un processus de relecture consciencieux, soit un taux de sélection de 39 %. 61 articles courts ont été soumis parmi lesquels 29 ont été sélectionnés au moyen d'un processus de relecture identique à celui des articles longs, soit un taux de sélection de 48 %. Comme lors de l'édition précédente de TALN, les articles courts seront présentés sous forme de sessions orales brèves (2 minutes par publication) et de poster. 10 démonstrations seront également présentées au cours d'une session dédiée.

Concernant les JEP, 145 propositions ont été reçues. À l'issue de la réunion du comité de programme qui s'est tenue à Grenoble les 15 et 16 mars, 108 articles ont été sélectionnés (74 %). 28 articles seront présentés en session orale et 80 lors de sessions poster.

La désaffection grandissante des soumissions à RECITAL nous a conduit à proposer plusieurs innovations afin de remobiliser nos jeunes chercheurs. Tout d'abord, l'appel à communication a été étendu pour permettre la soumission de travaux préliminaires, projets de thèse, travaux des premiers mois de recherche (états de l'art, premières pistes...). Ensuite le processus de relecture a été modifié pour offrir à nos jeunes des relectures pédagogiques (encouragements, pistes) et permettre des échanges directs avec les relecteurs (relectures non-anonymes). Ces changements ont été accueillis très favorablement puisque nous avons reçu 42 propositions de communications parmi lesquelles 11 feront l'objet de présentations orales (27 %) et 17 de présentations sous forme de poster (40 %). Nous sommes également revenus à des sessions RECITAL spécifiques qui ne sont pas en parallèle avec des sessions TALN.

En ce qui concerne les actes, nous avons fourni de nouveaux styles optimisés pour une lecture à l'écran. Bien que les habitudes des auteurs aient été changées à cette occasion, nous espérons que les lecteurs nous feront des retours d'usage positifs. Un meilleur référencement des travaux présentés a aussi été l'une de nos préoccupations ; aussi avons-nous choisi de les faire référencer par l'ACL (*Association for Computational Linguistics*) dans l'*ACL Anthology* [1] pour une meilleure visibilité.

Nous vous souhaitons, chers lecteurs, un parcours passionnant et passionné au fil des nombreuses pages de ces actes et, pourquoi pas, des découvertes inattendues grâce au hasard et à votre sagacité ; découvertes qui seront les graines de nouvelles idées pour faire progresser nos champs de recherche.

Laurent Besacier, Président JEP

Hervé Blanchon & Georges Antoniadis, Présidents TALN

Didier Schwab & Jorge Mauricio Molina Mejia, Présidents RECITAL

1. http://www.aclweb.org/anthology/

Le mot de la présidente de l'Association pour le Traitement Automatique des Langues

L'Association pour le Traitement Automatique des Langues (ATALA [2]) soutient depuis 1959 les travaux de recherche fondamentale et appliquée en linguistique informatique.

En complément des travaux sur les modèles informatiques de la langue, il est primordial pour l'ATALA de renforcer ses liens avec des domaines connexes tels que le traitement de la parole ou la représentation des connaissances.

Ceci est d'autant plus important à un moment où, avec l'avènement des technologies de l'Internet et de l'information, les données écrites et parlées, qu'il était jusqu'alors très difficile de recueillir sont devenues, en un laps de temps très court, pléthores et très faciles d'accès. En quelques années seulement, nous sommes passé du rêve, avoir accès à plus de données, au cauchemar, avoir trop de données. L'Internet et l'utilisation généralisée des bases de données sont aujourd'hui la cause principale de la croissance exponentielle et continue des données en ligne.

De nos jours, grâce aux logiciels embarqués la plupart des types de dispositifs électroniques que nous utilisons quotidiennement sont en mesure de fournir des données pérennes. En effet, alors qu'auparavant la plupart des données disparaissaient après avoir été utilisées dans un but précis, les données sont maintenant stockées, fusionnées, distribuées et même revendues pour être analysées et interprétées dans le meilleur des cas, à des fins d'innovation ou d'avancée scientifique.

Dans un contexte en constante mutation, l'organisation conjointe entre l'AFCP et l'ATALA des journées TALN permet aux deux communautés d'échanger leurs méthodes d'analyse et de compréhension de ces données textuelles ou parlées afin de faire progresser la recherche en proposant de nouvelles méthodes et de nouveaux algorithmes sur lesquels s'appuyer pour développer de nouvelles technologies et services dans le domaine de l'analyse intelligente des données.

Frédérique Segond
Présidente de l'ATALA

2. http://www.atala.org/

Le mot de la présidente de l'Association Francophone de la Communication Parlée

Chers collègues,

Après les éditions de 1970 (1$^{\text{ères}}$ JEP), 1979 (10$^{\text{èmes}}$ JEP), et avec en 2000 un détour à Aussois (23$^{\text{èmes}}$ JEP), les Journées d'Etude sur la Parole sont de retour à Grenoble !

L'AFCP (Association Francophone de la Communication Parlée [3]) se réjouit de s'associer de nouveau à l'ATALA (Association pour le Traitement Automatique des Langues) pour l'organisation de cet événement commun que sont les JEP-TALN-RECITAL. Rappelons que depuis 2002, les communautés du traitement de la langue, orale comme écrite, se retrouvent périodiquement en un même lieu afin favoriser les échanges et stimuler l'émergence de projets de recherche commun. Les éditions passées, à Nancy en 2002, à Fès en 2004, à Avignon en 2008, ont été un réel succès et nous gageons que cette édition JEP-TALN-RECITAL'2012 sera de nouveau un moment fort de rencontres et d'échanges fructueux entre les différents acteurs de nos communautés.

Pour ce qui concerne cette 29$^{\text{ème}}$ édition des Journées d'Etude sur la Parole, 145 communications ont été soumises, ce qui est très satisfaisant (136 soumissions en 2010 à Mons, 130 en 2008 à Avignon). L'origine variée des soumissions (majoritairement de France, mais aussi de Belgique, de Suisse, du Canada, des Etats-Unis, de Tunisie, du Maroc,...) souligne une fois encore le caractère international de ces journées francophones, qui est une priorité de l'AFCP. Sur ces 145 soumissions, 108 ont été retenues, ce qui donne un taux d'acceptation de 74 % qui est similaire à celui de l'édition précédente. La couverture thématique des papiers retenus est vaste et reflète le dynamisme et la diversité des recherches sur la parole dans la communauté francophone.

Pour rappel, les communications aux JEP sont sélectionnées sur la base d'un article complet. Chaque soumission est évaluée par deux relecteurs. Le comité de programme, constitué des membres du CA de l'AFCP et de membres du comité d'organisation, se réunit pendant deux jours pour examiner les soumissions et leurs évaluations, certaines sont relues par un 3ème lecteur, et la sélection finale est effectuée. Les communications sélectionnées sont alors groupées par thèmes afin de définir les sessions thématiques de la conférence, et pour chaque session, des communications orales sont choisies. Les autres communications, qui seront présentées sous forme de posters, ne sont pas regroupées thématiquement de façon a avoir des sessions poster couvrant un large spectre d'intérêts. Il est donc à noter qu'aux JEP, la sélection entre communication orale et affichée s'effectue principalement sur la base d'un choix thématique pour les sessions orales et ne renvoie donc pas à un critère de qualité.

3. L'Association Francophone de la Communication Parlée (AFCP) est une structure d'animation et de réflexion de la communauté francophone travaillant sur la parole. `http://www.afcp-parole.org/`

Pour ces JEP, outres les traditionnelles bourses proposées aux étudiants et jeunes chercheurs, nous renouvelons notre action d'invitation de jeunes chercheurs appartenant à des laboratoires situés hors de France. Cinq jeunes chercheurs venant de Tunisie et d'Algérie ont été ainsi sélectionnés sur dossier et nous auront le plaisir de les accueillir à ces rencontres. Nous aurons également l'honneur de remettre lors de ces journées les prix de thèse édition 2010 et 2011, à Gwénolé Lecorvé et Juliette Kahn, respectivement.

Pour finir, l'AFCP est ravie de voir cette 29ème édition des Journées d'Etude sur la Parole se tenir à Grenoble. Grenoble est depuis longtemps un haut lieu de la recherche sur la parole et a toujours eu un rôle important dans la structuration et l'animation de notre communauté parole, tant au niveau national, qu'au niveau international. Après des restructurations difficiles du pôle parole grenoblois, nous ne pouvons que nous réjouir que l'ensemble des laboratoires grenoblois, sous l'impulsion du LIG, ait entrepris l'aventure commune qu'est l'organisation de cet événement important pour la communauté francophone. Au nom de l'AFCP, je tiens donc à remercier sincèrement tous les organisateurs de ces Journées, le LIG, le LIDILEM et le GIPSA-Lab et en particulier Laurent Besacier, pour son dynamisme et son investissement dans cette entreprise.

Au nom du comité de programme, je remercie aussi vivement les 114 relecteurs pour leur temps et leur travail fait dans un esprit constructif.

Enfin, je tiens à remercier tous les auteurs, conférenciers, et participants qui sont le moteur de notre communauté scientifique si sympathique.

Je vous souhaite à tous des journées et des rencontres enrichissantes et stimulantes.

Cécile Fougeron
Présidente de l'AFCP
Présidente du Comité de Programme des XXIXèmes JEP

Comité d'organisation de JEP-TALN-RECITAL'2012 :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3)
Véronique AUBERGÉ (Gipsa-Lab, CNRS)
Valérie BELLYNCK (LIG-GETALP, Grenoble INP)
Laurent BESACIER (LIG-GETALP, Université Grenoble 1)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2)
Francis BRUNET-MANQUAT (LIG-GETALP, Université Grenoble 2)
Emmanuelle ESPERANÇA-RODIER (LIG-GETALP, Université Grenoble 1)
Jérôme GOULIAN (LIG-GETALP, Université Grenoble 2)
Marie-Paule JACQUES (LIDILEM, Université Grenoble 3)
Olivier KRAIF (LIDILEM, Université Grenoble 3)
Alexandre LABADIÉ (LIG-GETALP, CNRS)
Thomas LEBARBÉ (LIDILEM, Université Grenoble 3)
Benjamin LECOUTEUX (LIG-GETALP, Université Grenoble 2)
Mathieu MANGEOT (LIG-GETALP, Université De Savoie)
Jorge Mauricio MOLINA MEJIA (LIDILEM, Université Grenoble 3)
Claude PONTON (LIDILEM, Université Grenoble 3)
François PORTET (LIG-GETALP, Grenoble INP)
Solange ROSSATO (LIG-GETALP, Université Grenoble 3)
Isabelle ROUSSET (LIDILEM, Université Grenoble 3)
Didier SCHWAB (LIG-GETALP, Université Grenoble 2)
Frédérique SEGOND (Pôle Innovation Viseo)
Gilles SÉRASSET (LIG-GETALP, Université Grenoble 1)
Agnès TUTIN (LIDILEM, Université Grenoble 3)
Michel VACHER (LIG-GETALP, CNRS)
Nathalie VALLÉE (Gipsa-Lab, CNRS)
Virginie ZAMPA (LIDILEM, Université Grenoble 3)

Comité de programme de JEP'2012 :

Présidents :

Laurent BESACIER (LIG-GETALP, Université Grenoble 1, France)
Cécile FOUGERON (LPP Paris)
Guillaume GRAVIER, IRISA et CNRS-INRIA Rennes)

Membres :

Gilles ADDA (LIMSI, Paris)
Melissa BARKAT-DEFRADAS (PRAXILING, Montpellier)
Loïc BARRAULT (LIUM, Le Mans)
Philippe BOULA DE MAREUIL (LIMSI, Paris)
Véronique BOULENGER (DDL Lyon)
Elisabeth DELAIS-ROUSSARIE (Lab. Linguistique Formelle, Paris)
Véronique DELVAUX (Univ. Mons, Belgique)

Didier DEMOLIN (Gipsa-Lab, Grenoble)
Laurence DEVILLERS (LIMSI, Paris)
Isabelle FERRANE (IRIT, Toulouse)
Emmanuel FERRAGNE (CLILLAC-ARP, Paris)
Corinne FREDOUILLE (LIA, Avignon)
Bernard HARMEGNIES (Univ. Mons, Belgique)
Fabrice HIRSCH (PRAXILING, Montpellier)
Thomas HUEBER (Gipsa-Lab, Grenoble)
Irina ILLINA (LORIA, Nancy)
David LANGLOIS (LORIA, Nancy)
Georges LINARES (LIA, Avignon)
Hélène LOEVENBRUCK (Gipsa-Lab, Grenoble)
Egidio MARSICO (DDL, Lyon)
Sylvain MEIGNIER (LIUM, Le Mans)
Christine MEUNIER (LPL, Aix en Provence)
Yohann MEYNADIER (LPL, Aix en Provence)
François PELLEGRINO (DDL, Lyon)
Pascal PERRIER (Gipsa-Lab, Grenoble)
François PORTET (LIG-GETALP, Grenoble)
Solange ROSSATO (LIG-GETALP, Grenoble)
Sophie ROSSET (LIMSI, Paris)
Marc SATO (Gipsa-Lab, Grenoble)
Christophe SAVARIAUX (Gipsa-Lab, Grenoble)
Christine SÉNAC (IRIT, Toulouse)
Rudolph SOCK (IPS, Strasbourg)
Annemie VAN HIRTUM (Gipsa-Lab, Grenoble)
Béatrice VAXELAIRE (IPS, Strasbourg)
Chakir ZEROUAL (LPP Paris et Univ. Sidi Mohamed Ben-abdellah, Fes, Maroc)

Relecteurs additionnels :

Martine ADDA-DECKER, LPP et LIMSI Paris)
Régine ANDRE-OBRECHT (IRIT, Toulouse)
Angélique AMELOT (LPP, Paris)
Corine ASTESANO (Univ. Toulouse 2 et LPL, Aix en Provence)
Véronique AUBERGÉ (LIG et GIPSA-Lab, Grenoble)
Nicolas AUDIBERT (LPP, Paris)
Gérard BAILLY (Gipsa-Lab, Grenoble)
Claude BARRAS (LIMSI, Paris)
Denis BEAUTEMPS (Gipsa-Lab, Grenoble)
Nathalie BEDOIN (DDL, Lyon)
Roxane BERTRAND (LPL, Aix en Provence)
Benjamin BIGOT (LIA, Avignon)
Frédéric BIMBOT (IRISA et CNRS-INRIA Rennes)
Anne BONNEAU (LORIA, Nancy)
Hélène BONNEAU-MAYNARD (LIMSI, Paris)
Hervé BREDIN (LIMSI, Paris)

Nathalie CAMELIN (LIUM, Le Mans)
Christian CAVE (LPL, Aix en Provence)
Claire PILLOT-LOISEAU (LPP, Paris)
Lise CREVIER-BUCHMAN (LPP, Paris)
Mariapaola D'IMPERIO (LPL, Aix en Provence)
Paul DELÉGLISE (LIUM, Le Mans)
Christian DICANIO (UC Berkeley, États-Unis)
Cong-Thanh DO (LIMSI, Paris)
Christelle DODANE (PRAXILING, Montpellier)
Driss MATROUF (LIA, Avignon)
Sophie DUFOUR (LPL, Aix en Provence)
Elie EL-KHOURY (LIUM, Le Mans)
Robert ESPESSER (LPL, Aix en Provence)
Yannick ESTÈVE (LIUM, Le Mans)
Martine FARACO (LPL, Aix en Provence)
Jérôme FARINAS (IRIT, Toulouse)
Dominique FOHR (LORIA, Nancy)
Teddy FURON (IRISA et CNRS-INRIA Rennes)
Maeva GARNIER (Gipsa-Lab, Grenoble)
Cedric GENDROT (LPP, Paris)
Alain GHIO (LPL, Aix en Provence)
Antoine GIOVANNI (CHU Marseille et LPL Aix en Provence)
Laurent GIRIN (Gipsa-Lab, Grenoble)
Pierre HALLE (LPP, Paris)
Sophie HERMENT (LPL, Aix en Provence)
Daniel HIRST (LPL, Aix en Provence)
Kathy HUET (Univ. Mons, Belgique)
Stephane HUET (LIA, Avignon)
Denis JOUVET (LORIA, Nancy)
Juliette KAHN (LNE Paris)
Sophie KERN (DDL, Lyon)
Hélène LACHAMBRE (IRIT, Toulouse)
Muriel LALAIN (LPL, Aix en Provence)
Antoine LAURENT (LIUM, Le Mans)
Gwénolé LECORVE (IDIAP Martigny (Suisse))
Thierry LEGOU (LPL, Aix en Provence)
Christophe LÉVY (LIA, Avignon)
Alain MARCHAL (LPL, Aix en Provence)
Odile MELLA (LORIA, Nancy)
Ilya OPARIN (LIMSI, Paris)
Caterina PETRONE (LPL, Aix en Provence)
Myriam PICCALUGA (LPL, Aix en Provence)
Julien PINQUIER (IRIT, Toulouse)
Serge PINTO (LPL, Aix en Provence)
Agnès PIQUARD-KIPFFER (LORIA, Nancy)

Michel PITERMANN (LPL, Aix en Provence)
Rachid RIDOUANE (LPP Paris)
Albert RILLIARD (LIMSI, Paris)
Mickael ROUVIER (LIUM, Le Mans)
Jérémi SAUVAGE (PRAXILING, Montpellier)
Jean-Luc SCHWARTZ (Gipsa-Lab, Grenoble)
Grégory SENAY (LIA, Avignon)
Willy SERNICLAES (ULB Bruxelles, Belgique)
Marion TELLIER (LPL, Aix en Provence)
Michel VACHER (LIG Grenoble)
Nathalie VALLÉE (Gipsa-Lab, Grenoble)
Anne VILAIN (Gipsa-Lab, Grenoble)
Coriandre VILAIN (Gipsa-Lab, Grenoble)
Emmanuel VINCENT (IRISA et CNRS-INRIA Rennes)
Pauline WELBY (LPL, Aix en Provence)

Comité de programme de TALN'2012 :

Présidents :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3, France)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2, France)

Membres :

Nicholas ASHER (IRIT, CNRS et Université Toulouse 3)
Frédéric BÉCHET (LIF, Aix Marseille Université)
Yves BESTGEN (Université Catholique de Louvain, Louvain-la-Neuve, Belgique)
Philippe BLACHE (LPL, CNRS et Université de Provence)
Christian BOITET (LIG-GETALP, Université Grenoble 1)
Malek BOUALEM (France Telecom Orange Labs, Lannion)
Narjès BOUFADEN (KeaText, Montréal, Canada)
Yllias CHALI (University of Lethbridge, Lethbridge, Canada)
Laurence DANLOS (ALPAGE, Université Paris 7)
Piet DESMET (ITEC, K.U.Leuven et K.U.Leuven KULAK, Belgique)
Mark DRAS (Macquarie University, Sydney, Australie)
Denys DUCHIER (LIFO, Université d'Orléans)
Marc DYMETMAN (XRCE, Grenoble)
Dominique ESTIVAL (University of Western Sydney, Sydney, Australie)
Cédrick FAIRON (Université Catholique de Louvain, Louvain-la-Neuve, Belgique)
Olivier FERRET (CEA LIST, Palaiseau)
Michel GAGNON (École Polytechnique de Montréal, Montréal, Canada)
Claire GARDENT (LORIA, Villers lès Nancy)
Nabil HATOUT (CLLE-ERSS, CNRS et Université Toulouse II)
Sylvain KAHANE (MODYCO-ALPAGE, Université Paris 10)
Laura KALLMEYER (Heinrich-Heine-Universität, Düsseldorf, Allemagne)
Mathieu LAFOURCADE (LIRMM, Université Montpellier 2)
Philippe LANGLAIS (DIRO, Université Montréal, Canada)
Guy LAPALME (RALI, Université Montréal, Canada)

Yves LEPAGE (IPS, Université Waseda, Japon)
Emmanuel MORIN (LINA, Université Nantes)
Adeline NAZARENKO (LIPN, Université Paris 13)
Luka NERIMA (LATL, Université Genève, Suisse)
Alain POLGUÈRE (Université de Lorraine et ATILF CNRS)
Laurent PRÉVOT (LPL, CNRS et Université de Provence)
Violaine PRINCE (LIRMM, Université Montpellier 2)
Jean-Philippe PROST (LIRMM, Université Montpellier 2)
Christian RETORÉ (LaBRI et INRIA, Université Bordeaux 1)
Sophie ROSSET (LIMSI, CNRS)
Didier SCHWAB (LIG-GETALP, Université Grenoble 2)
Holger SCHWENK (LIUM, Université du Maine, Le Mans)
Pascale SÉBILLOT (IRISA, INSA de Rennes)
Gilles SÉRASSET (LIG-GETALP, Université Grenoble 1)
Agnès TUTIN (LIDILEM, Université Grenoble 3)
Anne VILNAT (LIMSI, CNRS et Université Paris Sud)
François YVON (LIMSI, CNRS et Université Paris Sud)
Virginie ZAMPA (LIDILEM, Université Grenoble 3)
Pierre ZWEIGENBAUM (LIMSI, CNRS et INALCO)

Comité Scientifique de TALN'2012 :

Présidents :

Georges ANTONIADIS (LIDILEM, Université Grenoble 3, France)
Hervé BLANCHON (LIG-GETALP, Université Grenoble 2, France)

Membres :

Les membres du comité de programme aidés de . . .
Ramzi ABBES (Techlimed, Lyon)
Stergos AFANTENOS (IRIT, Université de Toulouse)
Salah AIT-MOKHTAR (XRCE, Grenoble)
Maxime AMBLARD (LORIA, Université de Lorraine)
Jean-Yves ANTOINE (LI, Université de Tours et Lab-STICC, CNRS)
Delphine BATTISTELLI (STIH, Université Paris 4)
Denis BECHET (LINA, Université de Nantes)
Patrice BELLOT (LSIS, Université Aix-Marseille)
Delphine BERNHARD (LiLPa, Université de Strasbourg)
Romaric BESANÇON (CEA-LIST, Saclay Nano-Innov)
Brigitte BIGI (LPL, Aix en Provence)
Julien BOURDAILLET (Xerox, États-Unis)
Caroline BRUN (XRCE, Grenoble)
Francis BRUNET-MANQUAT (LIG-GETALP, Université Grenoble 2)
Marie CANDITO (Alpage, Université Paris Diderot)
Thierry CHANIER (LRL, Clermont Université)
Vincent CLAVEAU (IRISA-CNRS, Rennes)
Nathalie COLINEAU (CSIRO ICT Centre, Marsfield, Australie)
Benoît CRABBÉ (Alpage, Paris 7)

Béatrice DAILLE (LINA, Université de Nantes)
Pascal DENIS (Alpage)
Iris ESHKOL-TARAVELLA (LLL, Université d'Orléans)
Cécile FABRE (CLLE-ERSS, Université Toulouse 2)
Benoit FAVRE (LIF, Université Aix-Marseille)
Dominic FOREST (Université de Montréal, Canada)
Karen FORT (INIST et LIPN, Paris 13)
George FOSTER (CNRC, Gatineau, Canada)
Nuria GALA (LIF, Université Aix-Marseille)
Bruno GAUME (CLLE-ERSS, Université Toulouse 2)
Éric GAUSSIER (LIG-GETALP, Université Grenoble 1)
Kim GERDES (LPP, Université Paris 3)
Jérôme GOULIAN (LIG-GETALP, Université Grenoble 2)
Benoît HABERT (ICAR, ENS Lyon)
Najeh HAJLAOUI (Institut de recherche Idiap, Martigny, Suisse)
Thierry HAMON (LimetBio, Université Paris 13)
Marie-Paule JACQUES (LIDILEM, Université Grenoble 1)
Guillaume JACQUET (XRCE, Grenoble)
Christine JACQUIN (LINA, Université de Nantes)
Adel JEBALI (Université Concordia, Montréal, Canada)
Leïla KOSSEIM (Université Concordia, Montréal, Canada)
Olivier KRAIF (LIDILEM, Université Grenoble 3)
Éric LAPORTE (LIGM, Université Paris-Est Marne-la-Valée)
Dominique LAURENT (Synapse, Toulouse)
Thomas LEBARBÉ (LIDILEM, Université Grenoble 3)
Anne-Laure LIGOZAT (LIMSI, ENSIIE)
Cédric LOPEZ (LIRMM, Université Montpellier 2)
Mathieu MANGEOT (LIG-GETALP, Université de Savoie)
Denis MAUREL (LI, Université de Tours)
Aurélien MAX (LIMSI, Université Paris-Sud)
Jasmina MILIĆEVIĆ (OLST, Dalhousie University, Canada)
Laura MONCEAUX (LINA, Université de Nantes)
Richard MOOT (LaBRI et SIGNES, Bordeaux)
Erwan MOREAU (Trinity College Dublin, Irlande)
Fabienne MOREAU (IRISA, Universtié Rennes 2)
Véronique MORICEAU (LIMSI, Université Paris-Sud)
Philippe MULLER (IRIT, Université de Toulouse)
Alexis NASR (LIF, Université Aix-Marseille)
Aurélie NÉVÉOL (NCBI, National Library of Medicine, États-Unis)
Jian-Yun NIE (RALI, Université de Montréal, Canada)
Cécile PARIS (CSIRO ICT Centre, Marsfield, Australie)
Yannick PARMENTIER (LIFO, Université d'Orléans)
Guy PERRIER (LORIA, Université de Lorraine)
Sylvain POGODALLA (LORIA, Vandoeuvre-lès-Nancy)
Thierry POIBEAU (LaTTiCe, Montrouge)
Claude PONTON (LIDILEM, Université Grenoble 3)

Andrei POPESCU-BELIS (Institut de recherche Idiap, Martigny, Suisse)
Carlos RAMISCH (LIG-GETALP, Grenoble)
Mathieu ROCHE (LIRMM, Université Montpellier 2)
Antoine ROZENKNOP (LIPN, Universtité Paris 13)
Benoît SAGOT (Alpage, INRIA Roquencourt)
Djamé SEDDAH (Alpage, Université Paris 4)
Kamel SMAÏLI (LORIA, Université de Lorraine)
Xavier TANNIER (LIMSI, Université Paris-Sud)
Isabelle TELLIER (LaTTiCe, Université Paris 3)
Juan-Manuel TORRES-MORENO (LIA, Université d'Avignon et des Pays de Vaucluse)
François TROUILLEUX (LRL, Université Clermont-Ferrand 2)
Lonneke VAN DER PLAS (IMS, Université de Stuttgart, Allemagne)
Fabienne VENANT (LORIA, Université Nancy 2)
Jacques VERGNE (GREYC, Université de Caen)
Éric VILLEMONTE DE LA CLERGERIE (Alpage, INRIA Roquencourt)
Eric WEHRLI (LATL, Université de Genève, Suisse)
Guillaume WISNIEWSKI (LIMSI, Université Paris-Sud)
Imed ZITOUNI (IBM T.J. Watson Research Center, Yorktown Heights, États-Unis)
Michael ZOCK (LIF, Marseille)
Amal ZOUAQ (Royal Military College of Canada et Athabasca University, Canada)
Mounir ZRIGUI (UTIC, Faculté des Sciences de Monastir, Tunisie)
Sandrine ZUFFEREY (ILC, Université Catholique de Louvain-la-Neuve, Belgique)

Comité de programme de RECITAL'2012 :

Présidents :

Jorge Mauricio MOLINA MEJIA (LIDILEM, Université Stendhal – Grenoble 3)
Didier SCHWAB (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)

Membres :

Vanessa ANDRÉANI (Société CFH et laboratoire ERSS, Université Toulouse 2 – Le Mirail)
Nicolas AUDIBERT (Laboratoire de Phonétique et Phonologie-CNRS, Université Sorbonne-Nouvelle)
Frédéric BÉCHET (Laboratoire d'Informatique Fondamentale de Marseille, Université d'Aix-Marseille)
Patrice BELLOT (LSIS, Université d'Aix-Marseille)
Valérie BELLYNCK (GETALP-LIG, Grenoble INP)
Farah BENAMARA (IRIT, Université Toulouse 3)
Christian BOITET (GETALP-LIG, Université Joseph Fourier – Grenoble 1)
Leila BOUTORA (LPL, Université d'Aix-Marseille, Marseille)
Francis BRUNET-MANQUAT (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)
François-Régis CHAUMARTIN (Société Proxem, Laboratoire Alpage, UMR INRIA, Université Paris 7)
Gaël DE CHALENDAR (CEA LIST, Palaiseau)
Achille FALAISE (GETALP-LIG, Société Floralis, Université Joseph Fourier-Grenoble 1)
Olivier FERRET (CEA LIST, Palaiseau)
Nuria GALA (LIF, Université d'Aix-Marseille)
Jérôme GOULIAN (GETALP-LIG, Université Pierre Mendes France – Grenoble 2)
Thierry HAMON (LIM&BIO, Université Paris 13)
Nicolas HERNANDEZ (LINA, CNRS 6241, Nantes)

Bernard JACQUEMIN (CREM, Université de Haute Alsace, Mulhouse)
Olivier KRAIF (LIDILEM, Université Stendhal – Grenoble 3)
Alexandre LABADIÉ (GETALP-LIG, Grenoble)
Mathieu LAFOURCADE (LIRMM, Université de Montpellier 2)
Guy LAPALME (RALI, Université de Montréal, Canada)
François LAREAU (CLT, Macquarie University, Australie)
Thomas LEBARBÉ (LIDILEM, Université Stendhal – Grenoble 3)
Benjamin LECOUTEUX (LIG-GETALP, Université Pierre Mendès France – Grenoble 2)
Yves LEPAGE (Université Waseda, Japon)
Mathieu LOISEAU (LIDILEM, Université Stendhal – Grenoble 3)
Cédric LOPEZ (LIRMM, Université Montpellier 2)
Denis MAUREL (Université François Rabelais Tours)
Aurélien MAX (LIMSI-CNRS & Université Paris-Sud)
Jean-Luc MINEL (MoDyCO, UMR 7114, Université Paris-Ouest Nanterre La Défense – CNRS)
Emmanuel MORIN (LINA, CNRS 6241, Nantes)
Yayoi NAKAMURA-DELLOYE (LCAO, Université Paris VII)
Claude PONTON (LIDILEM, Université Stendhal-Grenoble 3)
François PORTET (GETALP-LIG, Grenoble INP)
Laurent PREVOT (LPL, Université d'Aix-Marseille, Marseille)
Violaine PRINCE (LIRMM, Université Montpellier 2)
Jean-Philippe PROST (LIRMM, Université Montpellier 2)
Bali RANAIVO-MALANÇON (Universiti Sarawak Malaysia, Malaisie)
Christian RETORÉ (LaBRI, Université Bordeaux 1)
Mathieu ROCHE (LIRMM, Université Montpellier 2)
Solange ROSSATO (GETALP-LIG, Université Stendhal – Grenoble 3)
Azim ROUSSANALY (LORIA, Université de Lorraine)
Isabelle ROUSSET (LIDILEM, Université Stendhal – Grenoble 3)
Fatiha SADAT (Université du Québec à Montréal, Canada)
Tristan VANRULLEN (TVSI, Marseille)
Eric WEHRLI (LATL, Université de Genève, Suisse)
Virginie ZAMPA (LIDILEM, Université Stendhal – Grenoble 3)
Haifa ZARGAYOUNA (LIPN, Université Paris 13)
Michael ZOCK (CNRS-LIF, Marseille)
Mounir ZRIGUI (Faculté des Sciences, Université de Monastir, Tunisie)
Pierre ZWEIGENBAUM (LIMSI-CNRS, Orsay)

Conférenciers invités :

Ian Maddieson (Université de Californie, Berkeley, États-Unis)
Jacqueline Léon (Laboratoire d'histoire des théories linguistiques, CNRS, Paris)
Yoshinori Sagisaka (Université de Waseda, Japon)
Hans Uszkoreit (DFKI, Sarrebruck, Allemagne)

Sponsors :


Inria


upmf
Grenoble
Université Pierre-Mendès-France
Sciences sociales & humaines


UNIVERSITÉ
Stendhal
GRENOBLE 3


quaero


CELIFrance


GRENOBLE-ALPES
METROPOLE


Région
RhôneAlpes


cnrs


UNIVERSITE
JOSEPH FOURIER
SCIENCES TECHNOLOGIE MEDECINE


LNE
Le progrès, une passion à partager


Grenoble
INP


EUROPEAN
ELRA
LANGUAGE
RESOURCES
ASSOCIATION


SyLLabs


Pôle
GRENOBLE
cognition


Synapse
Développement


– LINGUA ET MACHINA –
L'ÉCRIT MULTILINGUE DANS L'ENTREPRISE


MGEN


EDF


VILLE DE
GRENOBLE


ami
ENTERPRISE INTELLIGENCE SOFTWARE


L I G


LIDILEM
UNIVERSITÉ
Stendhal
GRENOBLE 3


gipsa-lab


AFCP


ATALA

Table des matières

Segmentation non supervisée : le cas du mandarin

Pierre Magistry
Alpage, INRIA Paris–Rocquencourt & Université Paris Diderot

RÉSUMÉ

Dans cet article, nous présentons un système de segmentation non supervisée que nous évaluons sur des données en mandarin. Notre travail s'inspire de l'hypothèse de Harris (1955) et suit Kempe (1999) et Tanaka-Ishii (2005) en se basant sur la reformulation de l'hypothèse en termes de variation de l'entropie de branchement. Celle-ci se révèle être un bon indicateur des frontières des unités linguistiques. Nous améliorons le système de (Jin et Tanaka-Ishii, 2006) en ajoutant une étape de normalisation qui nous permet de reformuler la façon dont sont prises les décisions de segmentation en ayant recours à la programmation dynamique. Ceci nous permet de supprimer la plupart des seuils de leur modèle tout en obtenant de meilleurs résultats, qui se placent au niveau de l'état de l'art (Wang *et al.*, 2011) avec un système plus simple que ces derniers. Nous présentons une évaluation des résultats sur plusieurs corpus diffusés pour le *Chinese Word Segmentation bake-off II* (Emerson, 2005) et détaillons la borne supérieure que l'on peut espérer atteindre avec une méthode non-supervisée. Pour cela nous utilisons ZPAR en apprentissage croisé (Zhang et Clark, 2010) comme suggéré dans (Huang et Zhao, 2007; Zhao et Kit, 2008)

ABSTRACT

Unsupervized Word Segmentation

In this paper, we present an unsupervised segmentation system tested on Mandarine Chinese. Following Harris's Hypothesis in Kempe (1999) and Tanaka-Ishii (2005) reformulation, we base our work on the Variation of Branching Entropy. We improve on (Jin et Tanaka-Ishii, 2006) by adding normalization and Viterbi-decoding. This enables us to remove most of the thresholds and parameters from their model and to reach near state-of-the-art results (Wang *et al.*, 2011) with a simpler system. We provide evaluation on different corpora available from the Segmentation bake-off II (Emerson, 2005) and define a more precise topline for the task using cross-trained supervised system available off-the-shelf (Zhang et Clark, 2010; Zhao et Kit, 2008; Huang et Zhao, 2007)

MOTS-CLÉS : Apprentissage non-supervisé, segmentation, chinois, mandarin.

KEYWORDS: Unsupervized machine learning, segmentation, Mandarin Chinese.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 1–13,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Pour la plupart des langues utilisant l'alphabet latin, un découpage sur les espaces est une bonne approximation d'une segmentation en unités lexicales. L'écriture chinoise en revanche ne délimite pas ces unités par la typographie. Seules les marques de ponctuation indiquent une partie des frontières entre les unités lexicales qui peuvent être formées d'un ou plusieurs caractères chinois. L'étape de *tokenisation*, préalable à beaucoup de systèmes d'analyse automatique est de ce fait plus délicate. Pour les langues sans caractère d'espacement ou équivalent, on parle d'étape de segmentation en mot.

De nombreux systèmes de segmentation par apprentissage supervisé ont été proposés mais ils requièrent des corpus segmentés manuellement. Ceux-ci sont souvent spécifiques à un genre, un domaine ou une variété de mandarin et en l'absence d'un consensus sur la définition de ce qu'est un « mot », ils suivent des guides d'annotations qui divergent.

Les systèmes supervisés atteignent aujourd'hui des résultats satisfaisants lorsque le corpus approprié pour l'entraînement est disponible. Cependant, si l'on veut faire face à une plus grande diversité en genres et en domaines ou répondre à des questions plus théorique sur la caractérisation formelle des unités de langue, s'intéresser aux approches non-supervisées nous semble nécessaire.

De plus, il est important de souligner que le système que nous présentons ici n'est pas spécifique au mandarin, ni à la segmentation de séquence de caractères chinois en « mots ». Des expérimentations sur d'autres langues et à partir d'autres unités initiales (lettre, phonème, mot orthographique) sont en cours et donnent des résultats prometteurs. En l'état actuel de l'avancement de nos travaux, nous ne pouvons fournir d'évaluation complète que pour le mandarin. Nous limitons donc notre présentation à cette langue.

Cette article est organisé ainsi : après une courte présentation de l'état de l'art à la section 2, nous détaillons les méthodes et les difficultés d'évaluation des systèmes de segmentation non-supervisés à la section 3. Les sections 4 à 5 présentent le fonctionnement notre système et la section 6 les résultats obtenus.

2 État de l'art

Les systèmes de segmentation non supervisés reposent généralement sur un des trois types de mesures suivantes ou une combinaison des trois : le niveau de cohésion des unités obtenues (par exemple en utilisant l'information mutuelle, comme dans (Sproat et Shih, 1990)) ; le degré de séparation des unités obtenues (par exemple la diversité des contextes, (Feng *et al.*, 2004)) ou la probabilité d'une segmentation étant donnée une chaîne (Goldwater *et al.*, 2006; Mochihashi *et al.*, 2009).

Dans un article publié récemment, Wang *et al.* (2011) présentent une méthode baptisee « *Evaluation, Selection, Adjustment*. » (*ESA*). Cette méthode combine cohésion et séparation en une mesure à maximiser sur l'ensemble d'une séquence. Ils utilisent ensuite les résultats de leur système pour en modifier les paramètres (essentiellement les comptes de n-grammes) et réitérer le processus 10 à 30 fois. Ils obtiennent ainsi les meilleurs résultats actuels en segmentation non-supervisée du mandarin.

Les principaux inconvénients de l'approche ESA sont d'une part le fait qu'il faille itérer le processus sur le corpus environ 10 fois avant d'atteindre des niveaux de performance satisfaisants, et d'autre part la nécessité d'avoir à fixer le paramètre de couplage entre mesure de cohésion et mesure de séparation. Empiriquement, on constate une corrélation entre ce paramètre et la taille du corpus, mais cette corrélation dépend de la façon dont sont traités les caractères latins et les chiffres arabes au cours des prétraitements. De plus, calculer cette corrélation et choisir la valeur optimale du paramètre en question (ce que les auteurs appellent le *proper exponent*) nécessite un corpus segmenté à la main, ce qui contredit le caractère non-supervisé de l'approche. Toutefois, si parmi les différents types de prétraitements pour lesquels ESA a été évalué on se réfère aux configurations qui se rapprochent des nôtres, les résultats de Wang *et al.* (2011) avec leur approche ESA se situent tous aux alentours de 0,80 de f-mesure sur les mots.

L'approche plus ancienne de Jin et Tanaka-Ishii (2006) ne repose que sur une mesure de séparation, elle-même directement insipirée par une hypothèse linguistique formulée par Harris (1955). Reformulée au moyen de la notion d'entropie de branchement (*Branching Entropy*, *BE*) par Tanaka-Ishii (2005) en suivant les travaux de Kempe (1999), cette hypothèse peut s'énoncer comme suit : si les séquences de graphèmes, phonèmes, ou autres produites par l'homme étaient aléatoires, on s'attendrait à ce que l'entropie de branchement d'une séquence (estimée à partir de *n*-grammes en corpus) décroisse lorsque la longueur de la séquence croît. Ainsi, la variation de l'entropie de branchement (*Variation of the Branching Entropy*, *VBE*) devrait être systématiquement négative. Lorsque l'on observe au contraire une VBE positive, l'hypothèse de Harris conduit à conclure que l'on se situe à une frontière d'unités linguistiques. C'est sur la base de cette hypothèse que Jin et Tanaka-Ishii (2006) proposent un système qui segmente dès que la BE croît (c'est-à-dire que la VBE est positive) ou lorsqu'elle atteint un certain maximum. Les auteurs fixent la longueur maximale des séquences calculées à 6 et lisent le corpus de gauche à droite et de droite à gauche. À chaque intervalle entre deux caractères, ils peuvent donc observer jusqu'à 12 valeurs desquelles ils conservent le maximum.

Le principal inconvénient de l'approche de Jin et Tanaka-Ishii (2006) est que les décisions de segmentation sont prises très localement [1] et ne dépendent pas des segmentations voisines. De plus, ce système repose lui aussi sur des paramètres, et notamment le seuil sur la VBE au dessus duquel le système décide de segmenter (dans leur système, il y a segmentation dès lors que VBE≥ 0). En théorie, on pourrait décider de segmenter dès lors que la BE ne décroît pas suffisamment, ou à l'inverse ne segmenter que si la VBE est non seulement positive mais même au dessus d'un certain seuil non nul. À cet égard, placer le seuil à la valeur 0 peut être considéré comme une valeur par défaut, mais reste un paramètre adaptable. Enfin, Jin et Tanaka-Ishii ne prennent pas en compte le fait que la VBE pour un n-gramme n'est pas forcément comparable *a priori* avec la VBE pour un m-gramme dès lors que $n \neq m$: une normalisation est ici nécessaire, comme le suggèrent notamment Cohen *et al.* (2002).

Faute de place, nous ne décrirons pas ici d'autres systèmes que ceux de Wang *et al.* (2011) et de Jin et Tanaka-Ishii (2006). Un état de l'art plus exhaustif peut être trouvé dans les articles de (Zhao et Kit, 2008) et de (Wang *et al.*, 2011).

Dans cet article, nous montrons que l'on peut corriger les inconvénients du modèle de Jin et Tanaka-Ishii (2006) et atteindre des niveaux de performance comparables à ceux de l'état de l'art, c'est-à-dire de Wang *et al.* (2011), le tout avec un système plus simple.

1. Dans sa thèse, Jin utilise l'auto-apprentissage et le paradigme de la *minimum description length (MDL)* pour pallier à ce problème.

	mots		caractères	
Corpus	en tout	différents	en tout	différents
Academia Sinica (AS)	5 449 698	141 340	8 368 050	6 117
City University of Hong Kong (CITYU)	1 455 629	69 085	2 403 355	4 923
Peking University (PKU)	1 109 947	55 303	1 826 448	4 698
Microsoft Research (MSR)	2 368 391	88 119	4 050 469	5 167

TABLE 1 – Taille des corpus utilisés

3 Problèmes d'évaluation

Dans cet article, afin de pouvoir nous comparer au système de Wang *et al.* (2011), nous nous évaluons sur les corpus du second Bakeoff international de segmentation du chinois (*Second International Chinese Word Segmentation Bakeoff*, Emerson, 2005). Ces corpus couvrent 4 guides de segmentation différents, développés au sein de 4 institutions distinctes : l'Academia Sinica (AS), la *City University* de Hong-Kong (CITYU), l'université de Pékin (PKU) et Microsoft Research (MSR).

Des informations sur la taille des corpus sont données au tableau 1. Dans le cadre du *Bakeoff*, Les détails du contenu n'étaient pas connus des participants. Mais on peut noter que le corpus de PKU est constitué d'extraits du Quotidien du Peuple, journal de Pékin. Le corpus de CITYU est extrait du LIVAC (T'sou *et al.*, 1997), aussi constitué de textes de presse, mais d'origines plus variées. Le projet du LIVAC cherchant à rendre compte des variantes géographique du mandarin, il inclut des articles provenant de Pékin, Hong-Kong, Singapour ou Taïwan. Le corpus de l'AS est un corpus équilibré, qui rend essentiellement compte de la variante du mandarin utilisée à Taïwan. Aucune description du contenu du corpus de MSR n'est disponible à notre connaissance.

L'évaluation de systèmes non-supervisés est une problématique en soi. Un consensus sur une définition précise de la notion de *mot* restant difficile à atteindre, différents guides d'annotation pour la segmentation en mots ont été proposés et appliqués à divers corpus. L'évaluation de systèmes de segmentation supervisés peut être réalisée sur n'importe quel corpus, indépendamment du guide d'annotation sous-jacent, pour peu que les données d'entraînement et les données d'évaluation soient cohérentes. Cependant, pour les systèmes non-supervisés, il n'y a aucune raison d'obtenir des résultats plus proches de l'un des guides existants que d'un autre, plutôt que des résultats se situant quelque part entre les différents guides. Huang et Zhao (2007) propose d'utiliser l'entraînement et l'évaluation croisés de systèmes de segmentation supervisés pour avoir un ordre d'idée sur le taux de désaccord entre guides d'annotation. L'idée est donc d'entraîner puis d'évaluer un système supervisé sur deux corpus respectant deux guides d'annotation distincts, et d'en tirer une approximation de leur désaccord. C'est également un moyen d'estimer une borne supérieur de ce que l'on est en droit d'attendre de la part d'un système non-supervisé, qui n'a pas de raison d'être plus proche d'un guide d'annotation que ne le sont les autres guides existants (Zhao et Kit, 2008). Nous avons reproduit ce type de mesures sur nos 4 corpus au moyen du système supervisé ZPAR (Zhang et Clark, 2010), et nous avons trouvé une cohérence moyenne similaire à celle obtenue par Huang et Zhao (2007), de l'ordre de seulement 0,84 (f-mesure), qui sera donc notre *topline*. Par ailleurs, il est généralement admis que segmenter chaque caractère individuellement est une *baseline* raisonnable, puisque près de la moitié des mots-formes dans un corpus segmenté à la main sont des unigrammes. Une telle baseline obtient

un f-score d'environ 0,35.

Ces évaluations globales peuvent être raffinées en décomposant les résultats en fonction de la longueur des mots. Les mots de longueurs différentes ont en effet des distributions très dissemblables. Les évaluations par longueur donnent les résultats suivants : sur les unigrammes, les f-scores se situent entre 0,81 et 0,90, similaires aux résultats globaux. Les résultats pour les bigrammes sont légèrement meilleurs (0,85–0,92), mais bien plus bas sur les trigrammes, descendant entre 0,59 et 0,79. Or, dans un texte en mandarin, la majorité des occurrences sont des mots unigrammes ou bigrammes, mais le lexique est principalement composé de bigrammes et de trigrammes. Ceci vient du fait que les unigrammes sont souvent des mots grammaticaux à haute fréquence, alors que les trigrammes sont souvent le résultat d'affixations plus ou moins productives. Pour cette raison, les résultats uniquement calculés sur les occurrences ne pâtissent pas énormément de mauvaises performances sur les trigrammes, même si une proportion significative du lexique est ainsi mal traitée.

Une autre difficulté concernant l'évaluation et la comparaison entre systèmes non-supervisés est de prendre en compte de façon équitable les prétraitements et les connaissances *a priori* qui sont fournies aux systèmes. Par exemple, Wang *et al.* (2011) utilise différents niveaux de prétraitement (qu'ils appellent *settings* et que nous appelerons « configurations »). Dans les configurations 1 et 2, Wang *et al.* (2011) essayent de ne pas se reposer sur la ponctuation et l'encodage des caractères (notamment la distinction entre caractères chinois et latins). Cependant, ils optimisent indépendamment leur paramètre pour chaque configuration. Nous considérons donc que leur système prend en compte le niveau de prétraitement qui est effectué sur les caractères latins et les chiffres romains, et sait donc à quoi s'attendre en la matière. Dans leur configuration 3, les auteurs ajoutent la connaissance de la ponctuation en tant que frontières de mots, et leur configuration 4 ajoute à cela un prétraitement des caractères latins et des chiffres arabes, ce qui conduit à des résultats plus significatifs, moins questionnables et plus convaincants.

Nous sommes plus intéressés par une réduction du travail humain que par le déploiement à tout prix d'un système strictement non-supervisé. Nous ne pensons donc pas utile de nous empêcher de procéder à quelques prétraitements simples, tels que ceux discutés ci-dessus : détection des ponctuations, des caractères latins et des chiffres arabes[2]. C'est la raison pour laquelle nos expériences correspondent aux configurations 3 et 4 de Wang *et al.* (2011), et c'est à elles que nous nous comparons, en appliquant notre système aux mêmes corpus.

4 Variation de l'entropie de branchement

4.1 Formulation

Notre système repose sur l'hypothèse de Harris (1955) et sa reformulation par Kempe (1999) et Tanaka-Ishii (2005). Définissons à présent les notions sous-jacentes à notre système.

Soit un n-gramme $x_{0..n} = x_{0..1}\, x_{1..2} \ldots x_{n-1..n}$ dont le contexte droit $\chi_{\rightarrow}$ contient tous les caractères observés à sa droite dans le corpus. Nous définissons son *entropie de branchement droite* (*Right*

2. De simples expressions régulières peuvent aussi être envisagées pour traiter les cas non-ambigus de nombres et de dates utilisant les sinogrammes

Branching Entropy, *RBE*) comme suit :

$$\begin{aligned} h_{\rightarrow}(x_{0..n}) &= H(\chi_{\rightarrow} \mid x_{0..n}) \\ &= -\sum_{x \in \chi_{\rightarrow}} P(x \mid x_{0..n}) \log P(x \mid x_{0..n}). \end{aligned}$$

L'*entropie de branchement gauche* (*Left Branching Entropy*, *LBE*) est définie de façon symétrique : si l'on note $\chi_{\leftarrow}$ le contexte gauche de $x_{0..n}$, sa LBE est définie par :

$$h_{\leftarrow}(x_{0..n}) = H(\chi_{\leftarrow} \mid x_{0..n}).$$

La RBE $h_{\rightarrow}(x_{0..n})$ peut être considérée comme l'entropie de branchement (BE) de $x_{0..n}$ au cours d'un parcours de gauche à droite, alors que la LBE est la BE de $x_{0..n}$ au cours d'un parcours de droite à gauche.

À partir, d'une part, de $h_{\rightarrow}(x_{0..n})$ et $h_{\rightarrow}(x_{0..n-1})$, et d'autre part de $h_{\leftarrow}(x_{0..n})$ et $h_{\leftarrow}(x_{1..n})$, nous définissons la *variation de l'entropie de branchement* (*Variation of Branching Entropy*, *VBE*) dans les deux directions comme suit :

$$\begin{aligned} \delta h_{\rightarrow}(x_{0..n}) &= h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1}) \\ \delta h_{\leftarrow}(x_{0..n}) &= h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}). \end{aligned}$$

4.2 Observations intermédiaires

Après avoir reproduit les expériences de Jin et Tanaka-Ishii (2006), nous avons effectué une série d'observations des valeurs prises par l'entropie de branchement et ses variations. Certaines de ces observations ont motivé les modifications apportées au modèle que nous présenterons à la section suivante. Dans cette section, nous présentons les plus pertinentes de ces observations. Pour des questions de place disponible et pour éviter la redondance, les graphiques de cette section sont produits à partir du corpus de PKU uniquement.

4.2.1 Confirmation de l'hypothèse de Harris

Dans un premier temps, nous allons confirmer l'hypothèse de Harris sur nos données. Pour cela nous nous limitons à l'observation de la frontière droite des bigrammes de notre corpus (le choix des bigrammes étant motivé par leur représentativité tant en nombre d'occurences en corpus qu'en nombre d'entrées dans le lexique). Cette valeur est donc calculée pour chaque bigramme observé au moins deux fois dans le corpus. On affiche ensuite l'ensemble de ces valeurs sous forme d'une courbe de densité qui donne ainsi la répartition des valeurs prises par la variation d'entropie (c'est à dire les '$\delta h_{\rightarrow}(x_{0..2})$). On distingue ensuite de l'ensemble de tous les bigrammes ceux qui sont considérés comme des mots par l'annotation manuelle de ceux qui ne le sont pas. Le résultat est présenté figure 4.1.1. On observe que les mots valides (qui forment une très petite proportion de l'ensemble des bigrammes observés) se démarquent bien par une variation d'entropie plus grande à leur frontière droite. Cependant, on observe aussi une zone relativement importante de confusion, qui confirme la nécessité de chercher la segmentation optimale d'une phrase, et non simplement les frontières de façon indépendantes les unes des autres.

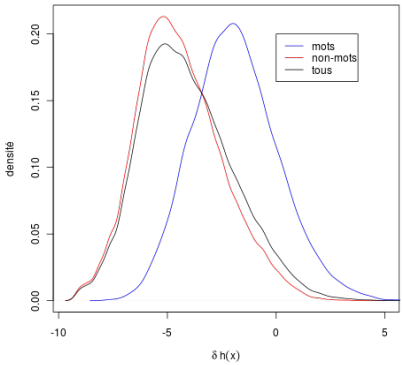


FIGURE 1 – Distribution des valeurs prises par la VBE à droite des bigrammes

On peut ensuite généraliser ce mode d'observation. En se demandant notamment si les valeurs de VBE à l'intérieur d'un mot sont aussi discriminantes que les valeurs qui en marquent les frontières. Pour différentes tailles de n-grammes mais en fixant n on va effectuer la même mesure et la même distinction entre mots et non-mots que précédemment mais cette fois-ci en prenant en compte les deux frontières gauche et droite ainsi que les valeurs observées à chaque inter-caractère à l'intérieur du n-gramme et dans chacun des deux sens de lecture possibles. Pour avoir une vue d'ensemble, on affiche ces résultats deux à deux sous la forme de courbes de niveaux. Les résultats pour les trigrammes sont présentés figure 2. On observe que des différentes valeurs observées les plus discriminantes sont sans conteste « gauche 1 » et « droite 3 », c'est à dire les entropies aux frontières. Il apparaît vraisemblable que la structure interne des unités morphologiquement complexes affectent la VBE, ce qui rend les valeurs internes plus difficiles à utiliser en pratique, contrairement à l'hypothèse suivie initialement dans (Magistry et Sagot, 2011).

4.2.2 Limites de la formulation par entropie

En présence de données aléatoires, on s'attend à ce que l'entropie de branchement diminue à mesure que la longueur de la chaîne considérée grandit. L'hypothèse de Harris nous fait dire que pour une chaîne donnée en langue naturelle, la variation de l'entropie de branchement lorsque l'on atteint une frontière sera anormalement élevée. Par ailleurs, on observe bien que pour des chaînes de même longueur, la variation de l'entropie de branchement aux frontières permet, au moins en partie, de distinguer les mots des non-mots. Toutefois, rien ne permet d'affirmer que cette distinction reste observable si l'on s'intéresse à des chaînes de longueurs différentes.

Nous avons donc cherché à observer les valeurs prises par la VBE aux frontières des n-grammes pour différentes valeurs de n. La figure 4.1.2 présente ces valeurs pour les uni- bi- et trigrammes. Elle montre qu'une normalisation ou qu'un recentrage de ces valeurs est nécessaire pour les rendre comparables.

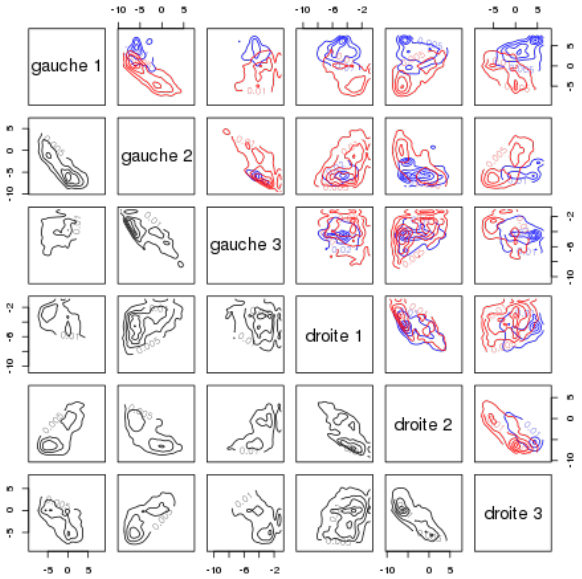


FIGURE 2 – Courbes de niveaux représentant la distribution des variations de l'entropie de branchement internes et aux frontières des trigrammes. La partie inférieur gauche du graphique (en noir) correspond à toutes les occurences de trigramme confondues, tandis que la partie supérieur droite distingue les mots (en bleu) des non-mots (en rouge). les dimensions indiquent qu'on s'intéresse à la variation d'entropie à gauche ou à droite d'un des trois caractères qui forment le trigramme, toujours en partant de l'extrémité opposée du trigramme.

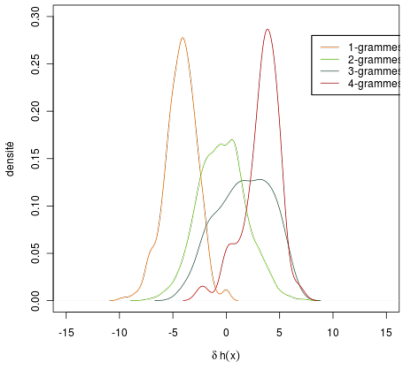


FIGURE 3 – VBE observée à droite des mots de différentes longueurs

4.2.3 Conséquences

Nous venons d'observer d'une part que l'information que l'on calcule aux frontières des mots est de loin la plus pertinente pour distinguer les unités lexicales recherchées, et d'autre part que si l'on cherche à comparer ces valeurs pour des unités de longueurs différentes, une normalisation est nécessaire. Ces observations motivent et permettent des modifications importantes du modèle de segmentation : les valeurs pertinentes sont moins nombreuses et indépendantes du contexte. Elles sont donc précalculables pour l'ensemble des n-grammes observés. De plus il nous faut les normaliser et chercher, pour une séquence de caractère donnée, à trouver la segmentation qui maximise ces mesures.

Ces modifications sont intégrées à notre algorithme de décodage présenté à la section suivante.

5 Algorithme de segmentation proposé

Les VBE ne sont pas directement comparables pour des chaînes de longeurs différentes, et doivent être normalisées. Nous recentrons les VBE, pour chaque longueur de chaîne, autour de 0. Pour cela, nous retranchons simplement à la VBE d'une chaîne de longeur k la moyenne des VBE de toutes chaînes de même longueur. Nous notons $\tilde{\delta h}_{\rightarrow}(x)$ et $\tilde{\delta h}_{\leftarrow}(x)$ les VBE normalisées. Pour simplifier, nous ne donnons que la définition de $\tilde{\delta h}_{\rightarrow}(x)$, celle de $\tilde{\delta h}_{\leftarrow}(x)$ en étant symétrique : pour chaque longueur k et chaque k-gramme x tel que $len(x) = k$, $\tilde{\delta h}_{\rightarrow}(x) = \delta h_{\rightarrow}(x) - \mu_{\rightarrow,k}$, où $\mu_{\rightarrow,k}$ est la moyenne des valeurs de $\delta h_{\rightarrow}(y)$ de tous les k-grammes y.

Il est important de noter que nous utilisons et normalisons la VBE et non l'entropie de branchement elle même. En effet, utiliser la BE contredirait l'hypothèse de Harris, puisque l'on ne

s'attendrait plus à ce que l'on ait $\tilde{h}(x_{0..n}) < \tilde{h}(x_{0..n-1})$ aux endroits qui ne sont pas des frontières de mots. De nombreux travaux utilisent pourtant la BE, normalisée ou non, et non la VBE, et obtiennent des résultats inférieurs à l'état de l'art (Cohen *et al.*, 2002).

Si nous ne basons nos décisions de segmentations que sur la VBE aux frontières de mots, chercher la meilleure segmentation d'une phrase revient à chercher en celle-ci les mots présentant les « meilleures frontières ». Cette qualité des frontières rejoint intuitivement (et empiriquement dans une certainement mesure, voir Magistry et Sagot (2011)) la notion d'autonomie syntaxique des unités qui composent la phrase. En termes de VBE, on peut définir la mesure d'autonomie d'un n-gramme comme $a(x) = \tilde{\delta}_{\leftarrow} h(x) + \tilde{\delta} h_{\rightarrow}(x)$.

On peut alors dire que plus l'autonomie $a(x)$ d'un n-gram x est grande, plus x est susceptible d'être un mot.

Avec cette mesure d'autonomie, on peut reformuler le problème de la segmentation d'une phrase comme la recherche du découpage qui maximise l'autonomie des mots qu'il délimite. Pour une séquence de caractères s, si on note $\mathrm{Seg}(s)$ l'ensemble de toutes les segmentations possibles, on cherche :

$$\arg\max_{W \in \mathrm{Seg}(s)} \sum_{w_i \in W} a(w_i) \times \mathrm{len}(w_i)$$

Où W est une segmentation délimitant les mots $w_0 w_1 \dots w_m$ et $\mathrm{len}(w_i)$ est la longueur d'un mot w_i, utilisée ici pour rendre comparables des segmentations aboutissant à des nombres de mots différents. Multiplier la mesure d'autonomie par la longueur du mot revient à attribuer un score aux caractères, qui contrairement aux mots sont en nombre constant entre les segmentations possibles d'une même chaîne.

Cette segmentation optimale en terme de VBEs est calculable simplement par programmation dynamique.

5.1 Décodage par programmation dynamique

Notre mesure d'autonomie d'un n-gramme donné est calculée à partir de tous ses contextes observés en corpus. Mais une fois calculée, elle ne dépend pas d'un contexte particulier. Elle ne dépend notamment pas du contexte observé spécifiquement au sein d'une chaîne en cours de segmentation.

Pour une chaîne donnée $u_{0..k}$ de longueur k, il y a 2^{k-1} segmentations possibles. Mais l'on peut remarquer que si l'on connaît la meilleure segmentation pour celle-ci et pour ses préfixes $u_{0..n}$, $n \leq k$, considérer un caractère supplémentaire et segmenter la chaîne $u_{0..k+1}$ ne nécessite que de considérer les appartenances possibles du caractère supplémentaire (le $k+1^{\text{ième}}$). Étant donné que nos mots sont contraints à être des séquences continues (insécables) de caractères, il nous suffit donc de considérer les cas suivants :

1. l'ajout du $k+1^{\text{ième}}$ caractère comme un mot de longueur 1 à fin de la meilleure segmentation de $u_{0..k}$
2. pour chaque prefixe $u_{0..n}$ de $u_{0..k}$ (avec $0 < n < k$), le cas où le $k+1^{\text{ième}}$ caractère est intégré à un mot unique de longueur $k-n$ qui vient s'ajouter à la fin de la meilleur segmentation de $u_{0..n}$

3. toute la chaîne $u_{0..k+1}$ est un mot unique de longueur $k+1$.

les cas 1 et 3 ci-dessus peuvent être vus comme les bornes du second cas, qui est le cas général si on prend $0 \leq n \leq k$. Ils sont explicités ici pour plus de clarté. Ce constat nous permet de reformuler la meilleure segmentation de $u_{0..k+1}$ à partir des meilleures segmentations de ses préfixes comme suit :

$$\underset{W \in \text{Seg}(u_{0..k+1})}{\arg\max} = \underset{V \in \bigcup_{n \leq k} \bigcup_{S \in \text{Seg}(u_{0..n})} S \cup \{u_{n..k+1}\}}{\arg\max} \sum_{w_i \in V} a(w_i) \times \text{len}(w_i)$$

Cette reformulation permet une programmation dynamique qui garde en mémoire les meilleures segmentations des $\text{Seg}(u_{0..n})$ et qui nous amène à ne considérer que $\sum_{n=2}^{k} n$ segmentations au lieu des 2^{k-1} théoriquement possibles. Cette méthode amène un surcoût négligeable pour $k < 5$ et devient de plus en plus intéressante à mesure que k grandit à partir de $k \geq 5$, ce qui est le plus souvent le cas.

6 Resultats et discussion

Nous avons évalué notre système sur les quatre corpus du *Bakeoff 2* et dans les configurations 2 et 3 telles que décrites à la Section 3. Nous comparons notre système (nVBE) aux résultats de Wang *et al.* (2011) ainsi qu'à notre propre implémentation de la stratégie « couper si une BE est croissante ». avec des variations de BE calculées dans les deux sens de lecture et pour toutes les longueurs de n-grammes, $1 \leq n \leq 6$. (à chaque position entre deux caractères, au plus 12 variations sont calculées, on segmente si au moins l'une d'entre elles est positive). Les résultats sont donnés Table 2. Les résultats filtrés par longueur de mot se trouvent Table 3.

Comme nous pouvons le voir, notre système est nettement meilleur que la stratégie de coupure sur accroissement de BE et obtient des scores comparables à ceux de ESA sans nécessiter de nombreuses itérations ni recourir à un paramètre.

Cela montre qu'on peut atteindre un bon niveau de segmentation en se basant uniquement sur une mesure de séparation. Lorsque celle-ci est maximisée pour une séquence donnée, il est raisonnable de penser qu'il existe une corrélation avec une éventuelle mesure de cohésion. Il n'est ainsi plus nécessaire d'avoir à trouver comment combiner les deux mesures.

On peut noter par ailleurs que l'évolution de nos résultats en fonction de la longueur des mots semble en accord avec la cohérence des guides d'annotation.

Nous ne pouvons fournir ici une analyse qualitative détaillée des résultats. Signalons tout de même que les erreurs observées nous semblent de même nature que nos observations antérieurs (Magistry et Sagot (2011)) et que celles présentes de la thèse de Jin. De nombreuses erreurs sont aussi liées aux dates et nombres écrits en chinois. Elles pourraient être écartées lors du prétraitement. D'autres erreurs concernent des morphèmes grammaticaux (« mots vides »)de haute fréquence et des affixes particulièrement productifs. Ces erreurs sont susceptibles de questionner les linguistes. Elle pourraient être corrigées en post-traitement par l'introduction de connaissances linguistiques.

Contrairement aux mots « pleins », ces mots vides ou morphèmes grammaticaux forment des classes fermées. De ce fait, introduire la connaissance linguistique nécessaire à leur bon traitement

System	AS	CITYU	PKU	MSR
Setting 3				
ESA bas	0.729	0.795	0.781	0.768
ESA haut	0.782	0.816	0.795	0.802
nVBE	0.758	0.775	0.781	0.798
Setting 4				
VBE> 0	0.63	0.640	0.703	0.713
ESA bas	0.732	0.809	0.784	0.784
ESA haut	0.786	0.829	0.800	0.818
nVBE	0.766	0.767	0.800	0.813

TABLE 2 – Évaluation sur les données du Bakeoff 2, suivant les configurations définies dans Wang *et al.* (2011). « Bas » et « haut » indiquent l'étendue des résultats obtenus par ESA pour différentes valeurs du paramètre du modèle. VBE> 0 segmente dès qu'une BE est croissante. nVBE correspond à la maximisation de la variation d'entropie de branchement normalisée aux frontières.

corpus	global	unigrammes	bigrammes	trigrammes
AS	0.766	0.741	0.828	0.494
CITYU	0.767	0.739	0.834	0.555
PKU	0.800	0.789	0.855	0.451
MSR	0.813	0.823	0.856	0.482

TABLE 3 – Resultats par longueur de mots (nVBE, configuration 4)

dans un système de segmentation ne nécessite qu'une quantité de travail limitée. Recourir à un système d'apprentissage supervisé ou symbolique pour traiter les classes de mots fermées et déléguer la gestion des classes ouvertes à un système non-supervisé nous semble être une voie prometteuse et linguistiquement pertinente.

Remarquons enfin que notre système obtient de bien meilleurs résultats sur les corpus de MSR et de PKU. Le corpus PKU étant le plus petit et AS le plus grand, la taille du corpus d'entraînement ne semble donc pas jouer à elle seule un rôle primordial pour expliquer les différences. En revanche, PKU est le corpus le plus homogène, il contient des articles qui sont tous issus du même journal. Le corpus AS au contraire est équilibré et présente une forte hétérogénéité des contenus. Le corpus CITYU est presque aussi petit que PKU mais contient des articles issus de journaux représentatifs de différentes variétés de mandarin, on peut donc s'attendre à ce que son contenu présente de grandes variations. Il semblerait donc que l'homogénéité des données d'entraînement soit aussi importante sinon plus que la quantité des données utilisées pour le bon fonctionnement du système présenté ici. Cette observation devra être vérifiée dans de prochains travaux. Si elle se confirmait, une étape de classification automatique des données d'entraînement pourrait être un prétraitement essentiel.

Références

COHEN, P., HEERINGA, B. et ADAMS, N. (2002). An unsupervised algorithm for segmenting categorical timeseries into episodes. *Pattern Detection and Discovery*, pages 117–133.

EMERSON, T. (2005). The second international chinese word segmentation bakeoff. *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133.

FENG, H., CHEN, K., DENG, X. et ZHENG, W. (2004). Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

GOLDWATER, S., GRIFFITHS, T. et JOHNSON, M. (2006). Contextual dependencies in unsupervised word segmentation. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 673–680.

HARRIS, Z. S. (1955). From phoneme to morpheme. *Language*, 31(2):190–222.

HUANG, C. et ZHAO, H. (2007). 中文分词十年回顾(chinese word segmentation : A decade review). *Journal of Chinese Information Processing*, 21(3):8–20.

JIN, Z. et TANAKA-ISHII, K. (2006). Unsupervised segmentation of chinese text by use of branching entropy. *In Proceedings of the COLING/ACL on Main conference poster sessions*, pages 428–435.

KEMPE, A. (1999). Experiments in unsupervised entropy-based corpus segmentation. *In Workshop of EACL in Computational Natural Language Learning*, page 713.

MAGISTRY, P. et SAGOT, B. (2011). Segmentation et induction de lexique non-supervisées du mandarin. *In Actes de TALN 2011, Montpellier*, pages 333–344.

MOCHIHASHI, D., YAMADA, T. et UEDA, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1-Volume 1*, pages 100–108.

SPROAT, R. W. et SHIH, C. (1990). A statistical method for finding word boundaries in chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

TANAKA-ISHII, K. (2005). Entropy as an indicator of context boundaries : An experiment using a web search engine. *In International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 93–105.

T'SOU, B., LIN, H., LIU, G., CHAN, T., HU, J., CHEW, C. et TSE, J. (1997). A synchronous chinese language corpus from different speech communities : Construction and applications. *Computational Linguistics and Chinese Language Processing*, 2(1):91–104.

WANG, H., ZHU, J., TANG, S. et FAN, X. (2011). A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3):421–454.

ZHANG, Y. et CLARK, S. (2010). A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 843–852.

ZHAO, H. et KIT, C. (2008). An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. *In The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), Hyderabad, India*.

Incrémentation lexicale dans les textes : une auto-organisation

Matthias Tauveron
Fonctionnements Discursifs et Traduction
UR LiLPa, Université de Strasbourg
matthias.tauveron@etu.unistra.fr

RESUME

Nous proposons une étude dynamique du lexique, en décrivant la manière dont il s'organise progressivement du début à la fin d'un texte. Pour ce faire, nous nous focalisons sur la co-occurrence généralisée, en formant un graphe qui représente tous les lemmes du texte et synthétise leurs relations mutuelles de co-occurrence. L'étude d'un corpus de 40 textes montre que ces relations évoluent d'une manière auto-organisée : la forme – et l'identité – du graphe de co-occurrence restent stables après une phase d'organisation terminée avant la 1ère moitié du texte. Ensuite, il n'évolue plus : les nouveaux mots et les nouvelles relations de co-occurrence s'inscrivent peu à peu dans le réseau, sans modifier la forme d'ensemble de la structure. La relation de co-occurrence généralisée dans un texte apparaît donc comme la construction rapide d'un système, qui est ensuite assez souple pour canaliser un flux d'information sans changer d'identité.

ABSTRACT

Lexical Incrementation within Texts: a Self-Organization

We propose here a dynamic study of lexicon: we describe how it is organized progressively from the beginning to the end of a given text. We focus on the "generalized co-occurrence", forming a graph that represents all the lemmas of the text and their mutual co-occurrence relations. The study of a corpus of 40 texts shows that these relations have a self-organized evolution: the shape and the identity of the graph of co-occurrence become stable after a period of organization finished before the first half of the text. Then they no longer change: new words and new co-occurrence relations gradually take place in the network without changing its overall shape. We show that the evolution of the "generalized co-occurrence" is the quick construction of a system, which is then flexible enough to channel the flow of information without changing its identity.

MOTS-CLES : Texte ; lexique ; co-occurrence généralisée ; auto-organisation

KEYWORDS: Text; lexicon; generalized co-occurrence; self-organization

1 Introduction

Le texte n'est pas qu'une suite linéaire de mots, propositions ou phrases. Il croît au fur et à mesure de son déroulement, à la manière d'une boule de neige et possède ainsi une dimension « incrémentielle » (Legallois, 2006). Nous proposons ici d'étudier cette incrémentation au niveau lexical. Nous montrons que le texte est un agencement complexe mais non anarchique de mots – comme (Adam, 2004, 35) l'a dit à propos des

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 15–28,
Grenoble, 4 au 8 juin 2012.

propositions – et proposons une description de la dynamique de cet agencement. Notre propos porte sur la « texture » ou la « textualité », c'est-à-dire la dimension formelle des textes et leur organisation, et ne concerne pas directement le sens ni sa construction en discours.

L'agencement des mots sera envisagé ici au travers de leurs relations de co-occurrence. Sous le nom de *co-occurrence généralisée*, l'étude de l'ensemble des relations de co-occurrence entre les lemmes d'un texte ou d'un corpus permet de révéler l'organisation en réseau de son lexique (Véronis, 2004, Viprey, 2006, Paranyushkin, 2010). Ces études en restent cependant à une image statique, montrant le réseau de co-occurrences tel qu'il se déploie une fois le texte lu dans son entier. Nous proposons ici une étude de la dynamique de cette co-occurrence généralisée en montrant son évolution du début à la fin d'un texte. Notre corpus de travail est formé en premier lieu de 20 textes courts (200 à 2000 mots, total de 11.015 mots), et en second lieu de 20 textes longs (5.000 à 22.500 mots, total de 192.477 mots).

Dans les cas typiques que nous observons, le réseau de co-occurrence généralisée croît très progressivement au début du texte. Cependant, de manière surprenante, à l'issue d'une première phase terminée avant la 1ère moitié du texte, ce réseau atteint un stade qui, sans être totalement figé, n'évolue plus guère par la suite. Ce réseau s'est donc formé une identité qui reste stable malgré l'apport ultérieur d'information. L'incrémentation lexicale dans la suite du texte ne fait que renforcer cette identité. Nous pensons qu'un tel comportement obéit à la définition que Moreno (2004) a donné des systèmes auto-organisés : ce sont des systèmes qui connaissent une *construction progressive de leur identité*. Nous expliciterons les arguments qui militent pour et contre la caractérisation du réseau de co-occurrence généralisée comme système auto-organisé, en montrant les enjeux linguistiques et cognitifs sous-jacents à cette question.

2 Méthode

2.1 Corpus

Notre corpus de textes courts est fait d'éditoriaux de revues scientifiques (revues *Développement durable et territoire*, *Vertigo*, *Langages*) et universitaires (revue *Savoir(s)*, « *magazine d'information de l'Université de Strasbourg* »). Une fois les premiers résultats obtenus, la démarche a été appliquée au corpus de textes longs : articles scientifiques (*Langages* 182 sur les théories du langage et les politiques des linguistes, *Intellectica*, 40 sur la notion de représentation) et encyclopédiques (*Wikipédia*, articles biographiques et analyses d'œuvres dans le domaine musical, grands articles portant sur la vie en société, tels « Religion », « Démocratie »). Nous disposons également d'un corpus de contrôle (textes totalement différents du corpus de travail sur lesquels on applique la même méthode, pour en tester le fonctionnement, le bien-fondé et la démarche) fait de deux textes poétiques d'Arthur Rimbaud issus des *Illuminations* (total : 1221 mots), choisis pour le caractère *a priori* désordonné de leur lexique. En tant

que tels, ils illustrent un cas extrême concernant l'organisation des mots dans un texte. On verra que leur étude permet de circonscrire certains comportements particuliers rencontrés dans le corpus de travail.

2.2Formation du graphe de co-occurrences

Le graphe de co-occurrences vise à donner une représentation visuelle synthétique de la co-occurrence généralisée dans un texte. Chaque lemme est représenté sous la forme d'un point (*nœud* du graphe) et un *lien* est tracé entre deux nœuds lorsque les lemmes correspondants sont co-occurrents. Tout lien est doté d'un *poids* qui indique le nombre de fois que la co-occurrence a été constatée dans le texte. On mesure ainsi directement l'importance de chaque lien. Nous considérons que deux unités sont co-occurrentes lorsqu'elles se trouvent à une distance de moins de trois mots[1], et ce, indépendamment des coupures de phrases.

La première étape du traitement consiste en une tokenisation et une lemmatisation faites en Perl grâce au dictionnaire fourni par l'ABU[2]. Le texte résultant est parcouru pour créer deux listes : celle des lemmes et celle des relations de co-occurrence. Ensuite, un programme Perl supplémentaire effectue, par l'intermédiaire du module *Graph* de Jarkko Hietaniemi[3], le calcul de la *betweenness centrality* des nœuds (définie *infra* en 2.3.1).

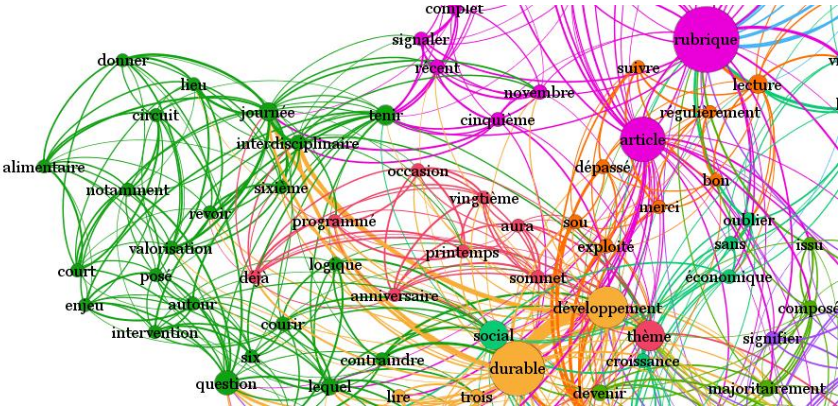


Figure 1 – Détail d'un graphe de co-occurrence pour un texte entier[4]

Notre choix de représenter les lemmes du texte sur le graphe (et non les formes rencontrées, ou encore des traits sémantiques) répond lui aussi à des raisons d'économie

[1] Le choix d'une distance aussi courte est motivée par la performance des outils de traitement, et par des questions de lisibilité du graphe. Comme nous l'a fait remarquer Tristan Vanrullen, accroître la taille de la fenêtre alourdit le graphe en multipliant les liens, ce qui génère par ailleurs un bruit important, notamment sur des textes longs portant sur plusieurs thématiques. La distance de 3 mots est un choix *a minima*.

[2] http://abu.cnam.fr/DICO/mots-communs.html

[3] http://search.cpan.org/~jhi/Graph-0.94/.

[4] Obtenu grâce à Gephi 0.7beta (www.gephi.org). Les couleurs signalent les tendances qu'ont les nœuds à être reliés : un nœud est plus lié aux nœuds de la même couleur que lui.

dans le traitement des données. Si la lemmatisation fait perdre des informations sur le sens des unités (Lemaire, 2008), nous pensons que cette perte n'est pas si significative dans notre problématique, formelle et non sémantique, d'organisation des mots dans les textes.

2.3Analyse des textes à l'aide du graphe de co-occurrences

2.3.1 L'organisation hiérarchique des graphes

L'intérêt du graphe ne réside pas dans la seule visualisation ergonomique qu'il propose. Il s'agit d'une structure mathématique pourvue de descripteurs permettant de caractériser numériquement et qualitativement les graphes, leurs nœuds et leurs liens. Nous nous intéressons ici spécifiquement à ceux qui révèlent son organisation hiérarchique[5].

Du fait des liens qu'ils nouent avec leurs voisins, et de leur position à l'échelle globale du graphe, certains lemmes ont une position centrale dans le réseau de co-occurrence. Parmi tous les indicateurs disponibles dans la littérature pour mesurer cette centralité, nous avons recours à la *centralité d'intermédiarité* (*betweenness centrality*, désormais BC). Notre choix se porte sur cette mesure pour trois raisons. En premier lieu, elle reflète l'intuition (Wasserman, Faust, 1994, 215) : les unités qui semblent les plus centrales à l'œil nu ont la BC la plus élevée. En second lieu, la BC est assez bien corrélée à la fréquence : un lemme fréquent a en général une BC élevée. La BC a cependant l'avantage de creuser les écarts entre les unités de fréquence similaire, et fait donc mieux apparaître les unités importantes. Enfin, elle renvoie à une forme pertinente d'organisation du lexique du texte : les unités ayant une BC élevée ont à la fois un rôle organisateur dans le graphe (du fait de leur position hiérarchique), et jouent un rôle d'intermédiaire entre les différentes notions du texte (Vergès, Bouriche, 2001, 69). Ce dernier aspect découle directement de la définition de ce paramètre.

La BC d'un nœud donné est en effet obtenue en additionnant la probabilité, pour tout couple de nœuds du graphe, que le nœud en question se trouve sur le plus court chemin reliant ces deux nœuds. Un nœud a donc une BC élevée si et seulement si beaucoup d'autres nœuds sont en relation directe avec lui ou sont obligés de passer par lui pour entrer en relation avec d'autres. Les nœuds dotés de la BC la plus élevée jouent donc un rôle constitutif dans le graphe de co-occurrence : c'est grâce à eux que se font la majorité des liens à échelle locale et à échelle globale. Un classement des nœuds par BC décroissante donne une image de l'organisation hiérarchique du lexique du texte. Une BC importante est le signe d'une position saillante dans le texte[6].

[5] Nous entendons par là que certains nœuds ont une position pré-éminente par rapport à d'autres. Sur la Figure 1, il s'agit notamment de *rubrique*, *article*, *développement* et *durable*. Soulignons que l'on pourrait chercher, indépendamment de cette organisation hiérarchique, d'autres formes d'organisation du graphe, notamment une organisation modulaire (Tauveron, 2012) que les paramètres que nous évoquons dans la section 5.2 permettent de caractériser.

[6] Contrairement à Boguraev, Neff (2000), notre définition de la saillance est purement interne au texte considéré, sans référence à une norme extérieure.

2.3.2 Une analyse longitudinale

L'originalité de notre étude réside dans le fait que nous nous focalisons sur l'évolution des graphes avec l'avancée du texte phrase après phrase[7]. Pour chaque texte, un premier graphe décrit la 1ère phrase, le 2ème graphe décrit les 2 premières, etc. Le texte entier n'est décrit que par le dernier graphe. Le choix d'une fenêtre s'élargissant à chaque étape (et non d'une « fenêtre glissante ») se justifie par le fait que nous travaillons sur un phénomène d'incrémentation. Une synthèse faite par un dernier programme Perl permet de comparer ces différentes étapes[8]. La comparaison est faite en relevant, dans l'évolution du texte au fil des étapes, les nœuds et les liens qui, au moins à un instant donné, ont une certaine prééminence. L'étude de l'ensemble de l'histoire de chacun de ces nœuds et de ces liens fournit une image d'ensemble des unités les plus saillantes, de ce point de vue, dans le texte. Ainsi, l'évolution du texte vue sous le seul angle de la co-occurrence généralisée permet d'étudier l'incrémentation lexicale qui a lieu dans le texte, non seulement sous son aspect quantitatif, mais d'étudier également sa construction.

3 Résultats

3.1 Trois types d'évolution

Dans notre corpus, nous avons mis en évidence trois grands types d'évolution d'ensemble de la co-occurrence généralisée.

Premier type, une croissance d'un bout à l'autre du texte : se créent sans arrêt de nouveaux nœuds et de nouveaux liens susceptibles de prendre une place prééminente assez rapidement.

Second type, une stabilisation progressive et unidirectionnelle du graphe : il commence par connaître une phase de croissance désordonnée, analogue au comportement précédent, mais cette croissance débouche rapidement sur un état stable, atteint en général avant la moitié du texte. Une fois ce palier atteint, le graphe n'évolue plus que dans le détail. C'est le comportement auto-organisé auquel nous avons fait référence en introduction ; il est d'ailleurs largement majoritaire dans notre corpus.

En troisième lieu, des cycles de stabilisation et de réorganisation successives. Le comportement précédent débouche subitement sur une nouvelle phase d'organisation au cours de laquelle est créé un nouvel état stable, différent du précédent.

Dans le corpus que nous avons utilisé, seul le second comportement est attesté de façon récurrente, comme le montre le tableau suivant :

[7] Nous ne pouvons revenir ici sur la question de la non-pertinence éventuelle de cette unité syntaxique dans l'étude de textes sur corpus, y compris dans le cas des corpus écrits.
[8] Ces synthèses sont représentées graphiquement à partir de la section 3.2.

Croissance permanente	3 textes	Textes brefs (Tous les textes de Rimbaud + un éditorial)
Stabilisation progressive et unidirectionnelle	33 textes	Textes brefs (éditoriaux) et tous les textes longs (scientifiques, encyclopédiques)
Cycles de stabilisation et de réorganisation	4 textes	Textes brefs (éditoriaux)

Tableau 1 – Répartition des textes selon les types d'évolution.

3.2 Les cas de stabilisation progressive et unidirectionnelle

3.2.1 Le corpus de textes brefs

Le début du texte[9] connaît éventuellement un sursaut de départ, phase brève (au plus un sixième du texte) au cours de laquelle les variations de centralité des nœuds et de poids des liens sont très rapides et imprévisibles, allant parfois d'un extrême à l'autre[10]. L'essentiel de l'évolution commence ensuite par une phase de stabilisation progressive. Les paramètres varient alors graduellement pour amener le texte à un certain stade organisé. Cette phase occupe en général entre un et deux cinquièmes du texte. Une fois atteint ce stade organisé, l'évolution est, qualitativement au moins, terminée. Les lemmes et les liens importants deviennent ensuite de plus en plus importants, à des vitesses différentes et imprévisibles. Ceux qui sont moins importants le restent, parfois en s'échangeant leurs places. Par contre, à quelques exceptions près, aucun nœud ou lien ne connaît d'important changement de classement passé ce stade. Il s'agit donc d'une phase d'aménagement de l'organisation. Le système que constitue la co-occurrence généralisée se donne au cours de la première phase une identité qui évolue ensuite sans se renier. Toute évolution ne fait que la raffermir, en accroissant le poids des unités importantes.

Dans notre corpus, un des cas les plus représentatifs est celui de l'éditorial du numéro de 1 de *Développement durable et territoire* (Figure 2, chaque courbe représente l'évolution d'un lemme au long du texte). Dans ce texte, la relation de co-occurrence généralisée a formé son identité pendant les 13 premières phrases du texte. La suite de l'évolution ne constitue qu'un réaménagement.

On peut mettre en évidence un phénomène analogue pour le comportement des liens. On voit sur les Figures 3 et 4 que le comportement des liens du graphe a la même histoire au cours de l'avancée du texte que celui des nœuds[11]. À ceci près que la construction de l'état

[9] Les différentes étapes sont qualitativement les mêmes dans tous les textes relevant de ce comportement. Seules changent la durée respective des phases et leur netteté.

[10] Cette forte variabilité s'explique par le fait que la mesure en question porte sur une portion du texte très réduite, et donc d'autant plus sujette à des variations aléatoires. On rencontre le même problème qu'avec un calcul de fréquence.

[11] Pour des raisons de lisibilité, nous répartissons les liens du graphe en 3 catégories. Les liens les plus importants (respectivement médians, moins importants) sont ceux qui, au moins une fois dans toute leur évolution, ont un poids qui dépasse 60% du poids maximal (respectivement 40%, 30%).

organisé du graphe a lieu sur une période plus étendue que pour les nœuds, à savoir entre la 11ème et la 15ème phrase.

Dans les détails, des différences parmi les Figures 3 et 4 révèlent un aspect de la dynamique du phénomène : moins les liens sont importants et plus ils évoluent vite au cours de la première phase. En effet, on constate pendant cette phase 4 montées abruptes sur le graphique des liens les plus importants et 10 sur le graphique suivants. C'est-à-dire que les liens les plus importants sont plus stables, et les liens les moins importants sont plus volatiles. Le poids des liens dénote donc la solidité et le caractère architectural des relations au cours d'une évolution. La relation de co-occurrence généralisée apparaît donc comme un véritable système, dont les parties les plus significatives sont pérennes, et les parties les plus légères sont adaptables au cours du temps[12].

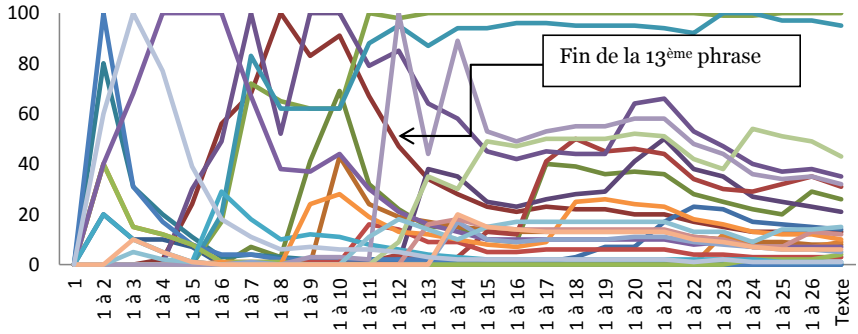


Figure 2 - Évolution de la BC des lemmes dans DDT-1

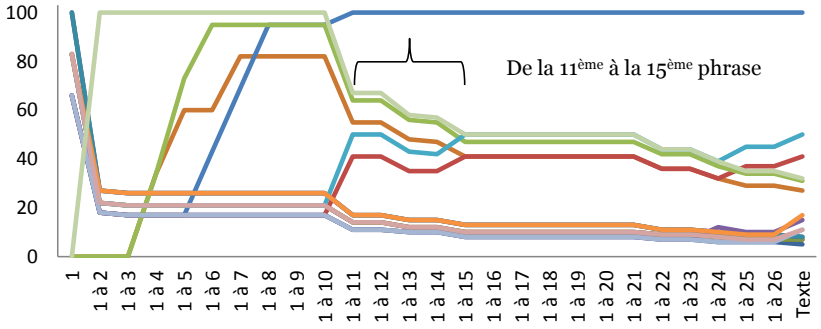


Figure 3 - Poids des liens les plus importants dans DDT-1-1

[12] On remarque sur Figure 4 que de nouveaux liens gagnent en importance dans la deuxième partie de l'évolution. Malgré le caractère unidirectionnel de ces évolutions, les parties les plus volatiles du graphe peuvent connaître des phases de croissance ponctuelles.

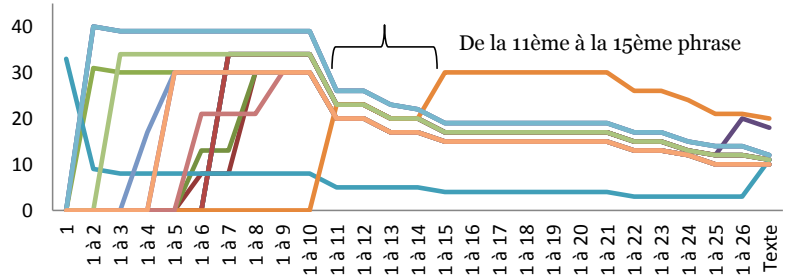


Figure 4 - Poids des liens les moins importants dans DDT-1-1

3.2.2 Le corpus de textes longs

Le comportement attesté sur la majorité des textes brefs du corpus de travail est également constatable sur l'ensemble du corpus de textes longs[13]. Il est même plus net sur ce second corpus. En effet, la phase initiale de stabilisation peut être – en proportion – largement plus courte, n'occupant qu'un dixième du texte dans certains cas. Il semble en particulier que plus le texte est long, plus la phase de stabilisation est courte en proportion. On le voit en particulier sur la Figure 5, représentant l'évolution d'un texte de 14.601 mots, divisé en 40 étapes de 15 phrases (article « Beethoven » de *Wikipédia*), et dans laquelle il apparaît bien que l'évolution de la BC des nœuds est finie dès la 4ème étape.

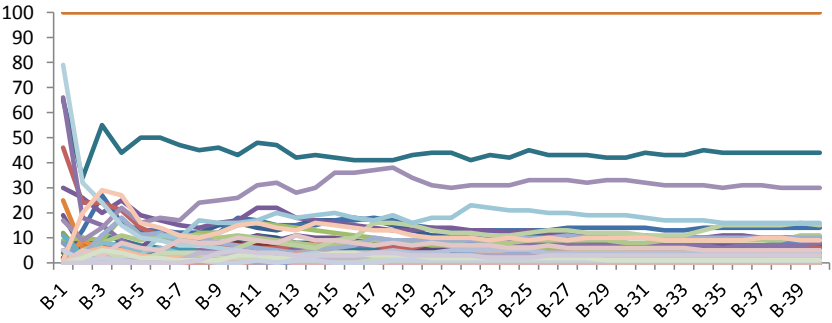


Figure 5 – BC des nœuds dans « Beethoven »

Le retard de l'évolution des liens par rapport à celle des nœuds est là encore constatable (Figure 6) : la phase d'organisation des liens du graphe s'étend jusqu'à l'étape 10, c'est-à-dire jusqu'à la fin du 1er quart du texte.

[13] Pour des raisons de lourdeur de traitement informatique, l'incrémentation est étudiée par paliers de 10 à 15 phrases, et non plus d'une seule. Le texte le plus long du corpus (22.500 mots) compte en effet 1.027 phrases.

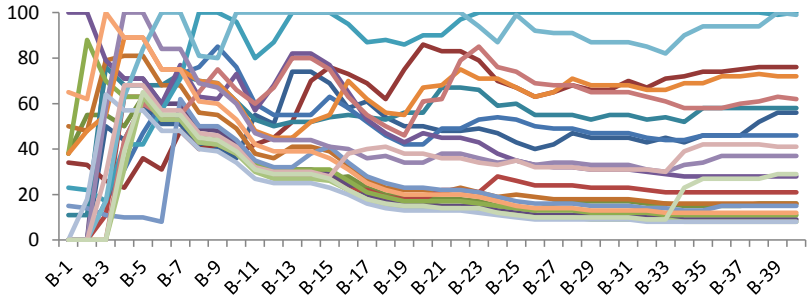


Figure 6 – Poids des liens les plus importants dans « Beethoven »

3.3 Les cas cycliques

Comme évoqué plus haut, nous avons pu mettre en évidence un fonctionnement cyclique sur un petit nombre de textes. Dans ces cas, l'évolution décrite en 3.2 se produit deux fois : une fois un état stable atteint, et maintenu pendant un laps de temps assez long, arrive une phase de réorganisation vers un autre état stable, maintenu lui aussi. On voit que le texte *DDT2-2* (Figure 7) alterne phases d'organisation (6 à 9, 21 à 30) et phases de stabilité (9 à 21, 30 à 41). La croissance de certaines unités qui a lieu lors de cette dernière phase (les deux courbes vertes correspondent à *politique* et à *culture*) ne bouleverse pas l'organisation d'ensemble. Par ailleurs, l'état du graphe lors des deux périodes de stabilité est bien distinct : c'est par exemple le lemme *notion* qui domine entre 9 et 21, alors que c'est *développement* qui domine après 30.

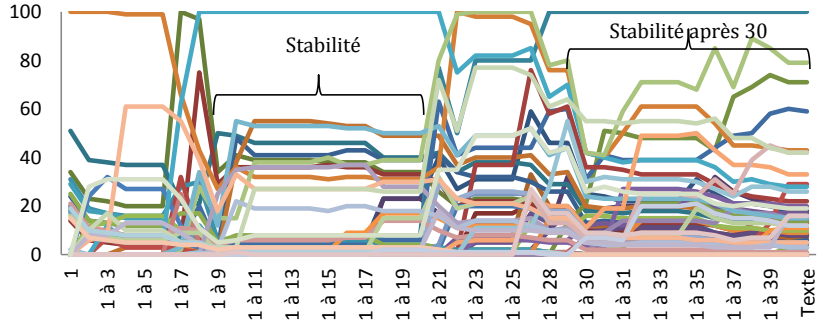


Figure 7 - Évolution de la BC des nœuds dans DDT2-2

Un retour à la lettre du texte permet de voir que ces deux phases de stabilité correspondent à des changements de thématique. La première partie du texte est consacrée à la résilience, qualifiée 5 fois de *notion*. Une phrase dénote ensuite un abandon de cette thématique (« *il semblerait qu'il n'y ait qu'un pas de la résilience à la*

résistance dans le cadre du développement durable ») et un passage à la thématique du développement durable (on trouve plus loin « *[é]videmment difficile de définir ici et en une phrase ce qu'est le développement durable* »)[14].

3.4Les cas de croissance permanente

Nous avons mené sur un corpus de contrôle fait de deux poèmes d'Arthur Rimbaud la même étude que sur le corpus de travail (Figure 8 : poème *Enfance*). Elle met en évidence un fonctionnement totalement différent des deux présentés plus haut, que l'on retrouve, de manière moins nette, sur un des textes du corpus (*Vertigo*3-1). Dans ces textes, la co-occurrence généralisée croît en permanence : au fur et à mesure qu'apparaissent de nouvelles phrases et de nouveaux mots, apparaissent également de nouveaux lemmes et de nouvelles relations de co-occurrence. On n'identifie pas de phase de stabilité à un quelconque moment. Comme sur le cas précédent, la plus haute valeur en BC est prise successivement par des lemmes différents, sans cette fois que cette alternance ne reflète des phases organisées.

Les trois textes marginaux qui connaissent ce comportement constituent en quelque sorte des cas limites. Dans les deux comportements organisés décrits plus haut, l'organisation en partie hiérarchique des graphes servait en effet à assurer la cohésion du réseau de co-occurrence généralisée, et permettait donc l'émergence d'une forme d'ensemble. Les nœuds les plus importants construisaient leur position de prééminence au cours d'une évolution mesurée, pour la garder par la suite. Au contraire, dans les cas de croissance permanente, aucun état stable n'est maintenu au cours du temps, ce qui n'empêche pas de considérer qu'on a affaire à une certaine forme d'organisation.

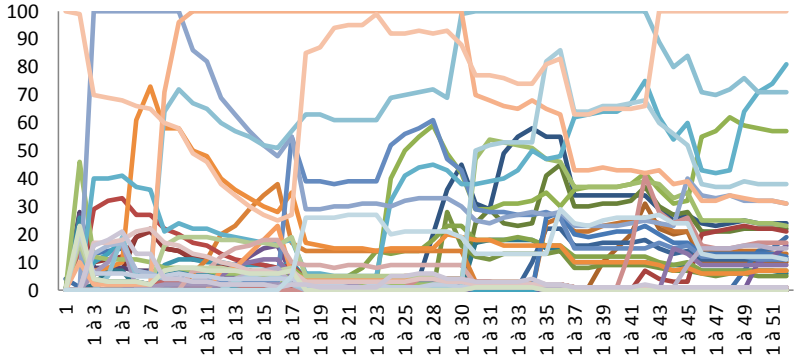


Figure 8 - Évolution de la BC des nœuds dans *Enfance* de Rimbaud

[14] Il faut noter cependant que le changement de thématique n'intervient pas à l'instant précis du changement de phase que nous identifions. Le grand nombre d'occurrences de *notion* au début du texte donne une certaine inertie à ce terme . Dans la mesure que nous faisons, sa position prééminente se maintient après le changement de thématique.

4 Discussion

4.1 Un système auto-organisé

Il est surprenant que le graphe ne connaisse pas systématiquement une croissance permanente d'un bout à l'autre du texte. Dans les faits, le cas normal est celui d'une stabilisation progressive et unidirectionnelle du graphe de co-occurrence[15]. Ce fonctionnement particulier peut s'expliquer en référence à un mécanisme général d'évolution de certains systèmes, appelé *auto-organisation* (van de Vijver, 2004).

Les observations faites montrent que l'évolution de la co-occurrence généralisée ne se fait pas au hasard, et ne prend pas la forme d'une croissance permanente. En effet, au fur et à mesure de l'avancée du texte, les mots et les relations de co-occurrence ne sont pas ajoutés de manière effrénée, ni placés anarchiquement mais sont disposés de manière systématique. Ce placement passe par une phase de stabilisation qui a lieu au début du texte, avant d'atteindre un état certes relativement stable, mais surtout susceptible de se réajuster pour permettre l'intégration des nouveaux contenus. C'est la force de cette structure : elle permet de canaliser le flux d'information continu dans la deuxième partie du texte. Elle repose sur la présence d'une certaine hiérarchie entre unités – les lemmes et les liens les plus importants, du fait de leur rôle architectural, étant moins volatiles. C'est cette organisation qui assure en premier lieu la stabilité du système et son existence (Ladrière, 2009), malgré la sollicitation (et le risque de désorganisation) que représente l'apport de nouveaux contenus. L'ensemble du processus montre donc une *construction progressive de l'identité du réseau de co-occurrence généralisée* (définitoire de l'auto-organisation chez Moreno, 2004). Ce réseau élabore donc son identité de système au cours d'une première phase. Par la suite, même si ce système ne reste pas figé et évolue, c'est sans modifier cette identité. Comme on l'a vu, comprendre l'évolution du système suppose de le considérer dans sa globalité. Cette évolution d'ensemble dépasse l'histoire de chacun des nœuds et des liens, et est imprévisible à partir de leurs évolutions ponctuelles (comme le montre le passage brusque de la phase chaotique initiale à la phase organisée). Par définition, l'organisation de la co-occurrence généralisée est donc *émergente*. Nous allons montrer de surcroît qu'elle semble *prévue* (par le scripteur du texte) *pour être identifiée* (par le lecteur) – il s'agit donc d'une *auto-organisation au sens faible* dans la typologie d'Atlan (2011, 194)[16].

Si on peut mettre le phénomène que nous avons constaté ici sur le compte d'une propriété générale de certains systèmes, nous allons montrer maintenant qu'il repose sur des contraintes et ressources de la cognition humaine.

[15] L'interprétation faite ici s'applique également, dans une moindre mesure, sur les textes présentant une évolution cyclique.

[16] Ces systèmes s'opposent chez l'auteur aux auto-organisations au sens fort, que sont par exemple les êtres vivants, dans lesquelles la structure a une finalité et une signification qui tirent leur origine de l'intérieur du système. De tels systèmes ont donc une évolution autonome. Un texte étant produit pour être lu, sa finalité et son fonctionnement sont nécessairement à mettre sur le compte du scripteur.

4.2Interprétation en termes cognitifs

4.2.1 Contraintes du système cognitif

Van Dijk, Kintsch (1983) ont supposé l'existence de mécanismes cognitifs permettant la structuration de l'information rencontrée dans les textes. Parmi eux, la suppression d'informations secondaires, et le regroupement de plusieurs informations en une seule. On peut supposer que ces mécanismes obéissent à la nécessité d'organiser le sens (et, indirectement, le lexique) du texte. Une des conséquences de ces phénomènes est la présence de thèmes (au sens de Hjørland, 2001), qui structurent le contenu sémantique du texte de manière notamment hiérarchique.

Le comportement auto-organisé que nous avons décelé révèle un des aspects de cette contrainte. En effet, le maintien de la prééminence d'unités importantes et la relative rigidité des liens les plus forts ont un rôle architectural dans la co-occurrence généralisée similaire à l'économie des thèmes dans un texte. Il apparaît donc que le caractère auto-organisé de la co-occurrence généralisée permet l'application des opérations de suppression et de structuration de l'information.

4.2.2 Ressources du système cognitif

Par ailleurs, le fonctionnement auto-organisé décrit ici n'est possible que grâce à certaines ressources du système cognitif. Deux des opérations cognitives fondamentales postulées par Langacker (1987, 116-140) – la sélection et la transformation – rendent ce type d'évolution possible.

Le fonctionnement des textes tel que nous l'avons montré repose en effet sur la capacité qu'a le système cognitif à sélectionner les unités et les relations les plus saillantes pour focaliser son attention sur elles et les placer au centre de sa construction du sens textuel. La modification permanente du système auto-organisé repose sur ses capacités de transformation : les représentations formées sont sans cesse déformées pour accepter de nouvelles informations arrivant certes par incrémentation successive, mais, et c'est notre conclusion, de manière structurée.

Le fonctionnement de ces ressources contribue à expliquer la difficulté à la lecture des deux textes poétiques de Rimbaud. L'impression de désordre qu'on peut avoir à leur sujet est démontrée ici, sur son seul versant lexical. Comme on l'a vu, ces textes font partie de ceux qui ne témoignent pas d'une stabilisation progressive : les opérations que sont la sélection et la transformation d'informations par le système cognitif sont dans ces cas inopérantes et ne permettent pas facilement l'élaboration d'un sens d'ensemble.

5 Conclusions

5.1 Perspectives d'application

Nous avons considéré ici (à la manière d'Utiyama, Isahara, 2001) que les thèmes des textes se définissaient par des relations de co-occurrence (fréquence des liens par

l'intermédiaire de leur poids, centralité). En suivant les idées de Hearst (1997) ou Ferret (2007), on pourrait appliquer ce travail en premier lieu à des fins de segmentation thématique. Nous avons en effet montré dans certains cas que notre méthode permettait de diviser les textes en différentes sections. Elle montre également dans d'autres cas (croissance permanente) que certains textes ne sont pas segmentables thématiquement.

Erkan, Radev (2004) et Xie (2005) ont par ailleurs montré que la centralité dans le lexique des textes permettait de mettre en évidence la structuration hiérarchique des informations, et donc de repérer les éléments qui doivent figurer dans leur résumé. La spécificité de notre méthode – envisager le texte dans sa dynamique – permettrait de repérer les passages qui apportent le plus d'informations. Si une telle étude confirme le fait que les informations essentielles se situent souvent au début des textes, nous avons pu montrer qu'il ne s'agit pas d'une généralité (dans les cas de croissance permanente).

En troisième lieu, la méthode présentée ici reposant sur le seul examen du lexique, c'est dans l'indexation des documents qu'elle connaîtrait ses meilleures applications. Elle pourrait permettre d'identifier les mots les plus essentiels du texte selon les passages et d'identifier leurs relations mutuelles mieux que ne le ferait une liste de mots-clés.

Enfin, l'examen des spécificités de l'organisation lexicale des textes permettrait de faire un pas dans la caractérisation du style et du genre d'un texte (voire d'un auteur)[17].

5.2 Nouvelles perspectives de recherche

La méthode présentée ici se doit d'être affinée et précisée en recourant à d'autres paramètres mathématiques donnant une meilleure description qualitative de la position des nœuds dans les graphes (*eigenvector centrality, clustering coefficient*, classes de modularité, etc.). L'effet de la taille de l'empan définissant la co-occurrence, d'un éventuel élagage du graphe ou de la lemmatisation restent également à caractériser en détail.

Références

ADAM, J.-M., (2004), *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

ATLAN, H., (2011), *Le Vivant post-génomique ou qu'est-ce que l'auto-organisation ?*, Paris : Odile Jacob.

BOGURAEV, B., NEFF, M., (2000), "Lexical Cohesion, Discourse Segmentation and Document Summarization," *Proceedings of RIAO'2000*, Paris (12–14 avril 2000).

ERKAN, G., RADEV, D., "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", *Journal of Articial Intelligence Research*, 22, 457-479.

FERRET, O. 2007. "Finding document topics for improving topic segmentation". *ACL 2007*. 480-487.

[17] Cette suggestion de Tristan Vanrullen portera tous ses fruits dans un prochain travail.

HEARST, M. 1997. “TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages”. *Computational Linguistics*, 23, 1, 33-64.

HJØRLAND, B., (2001), “Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content… and Relevance”, *Journal of the American Society for Information Science and Technology*, 52, 9, 774–778.

LADRIERE, J., (2009), « Système, épistémologie », *Encyclopaedia Universalis*.

LANGACKER, R., (1987), *Foundations of Cognitive Grammar. Theoretical Prerequisites*, Stanford : Standford UP.

LEGALLOIS, D., (2006), « Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison », *Langages*, 163, 3-9.

LEMAIRE, B., (2008), « Limites de la lemmatisation pour l'extraction de significations », *JADT 2008*, 725-732.

MORENO, A., (2004), « Auto-organisation, autonomie et identité », *Revue internationale de philosophie*, 228, 135-150.

PARANYUSHKIN, D., (2010), « Text network analysis », Conférence du *Performing Arts Forum*, http://noduslabs.com/research/pathways-meaning-circulation/, (14.09.2011).

TAUVERON, M., (2012), « Variation du sens lexical en discours : la co-occurrence généralisée valide une non-correspondance entre deux langues ». La cooccurrence : du fait statistique au fait textuel. Besançon, 9 février 2012.

UTIYAMA, M. ISAHARA, H.. 2001. “A statistical model for domain-independent text segmentation”. In *ACL’01*, 491–498.

VAN DE VIJVER, G., (2004), « Auto-organisation, autonomie, identité : Introduction », *Revue internationale de philosophie*, 228, 129-133.

VAN DIJK, T. KINTSCH, W., (1983), *Strategies of Discourse Comprehension*, Orlando : Academic Press.

VERGES, P., BOURICHE, B., (2001). “L’analyse des données par les graphes de similitude”. *Sciences humaines*, http://www.scienceshumaines.com/textesInedits/Bouriche.pdf.

VERONIS, J., (2004), “HyperLex : Lexical Cartography for Information Retrieval”, *Computer Speech & Language*, 18, 3, 223-252.

VIPREY, J.-M., (2006), “Structure non-séquentielle des textes”, *Langages*, 163, 71-85.

WASSERMAN, S., FAUST, K., (1994), *Social Network Analysis*, Cambridge UP.

XIE, Z., (2005), “Centrality Measures in Text Mining : Prediction of Noun Phrases that Appear in Abstracts”. *ACL’05, Proceedings of the Student Research Workshop*. Ann Arbor.

A Study of Heterogeneous Similarity Measures for Semantic Relation Extraction

Alexander Panchenko
Center for Natural Language Processing (CENTAL), Université catholique de Louvain
College Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgium)
alexander.panchenko@student.uclouvain.be

RÉSUMÉ

Etude des mesures de similarité hétérogènes pour l'extraction de relations sémantiques

L'article évalue un éventail de mesures de similarité qui ont pour but de prédire les scores de similarité sémantique et les relations sémantiques qui s'établissent entre deux termes, et étudie les moyens de combiner ces mesures. Nous présentons une analyse comparative à grande échelle de 34 mesures basées sur des réseaux sémantiques, le Web, des corpus, ainsi que des définitions. L'article met en évidence les forces et les faiblesses de chaque approche en contexte de l'extraction de relations. Enfin, deux techniques de combinaison de mesures sont décrites et testées. Les résultats montrent que les mesures combinées sont plus performantes que toutes les mesures simples et aboutissent à une corrélation de 0,887 et une Precision(20) de 0,979.

ABSTRACT

This paper evaluates a wide range of heterogeneous semantic similarity measures on the task of predicting semantic similarity scores and the task of predicting semantic relations that hold between two terms, and investigates ways to combine these measures. We present a large-scale benchmarking of 34 knowledge-, web-, corpus-, and definition-based similarity measures. The strengths and weaknesses of each approach regarding relation extraction are discussed. Finally, we describe and test two techniques for measure combination. These combined measures outperform all single measures, achieving a correlation of 0.887 and Precision(20) of 0.979.

MOTS-CLÉS : Similarité sémantique, Relations sémantiques, Similarité distributionnelle.

KEYWORDS: Semantic Similarity, Semantic Relations, Distributional Similarity.

1 Introduction

Semantic relations provide information about terms which have similar or related *meanings*. This kind of knowledge about language has proven to be valuable for various *NLP applications*, such as word sense disambiguation (Patwardhan *et al.*, 2003), query expansion (Hsu *et al.*, 2006), document categorization (Tikk *et al.*, 2003), or question answering (Sun *et al.*, 2005).

Let R be a set of synonymy, hypernymy, co-hypernymy, and associative relations between a set of terms C, established manually. A semantic relation extraction aims at discovering relations

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 29–42,
Grenoble, 4 au 8 juin 2012.

$\hat{R} \subseteq C \times C$ which would be as close to R as possible in terms of precision and recall :

$$\hat{R}^* = \arg\max_{\hat{R}} \frac{Precision(R,\hat{R}) \cdot Recall(R,\hat{R})}{Precision(R,\hat{R}) + Recall(R,\hat{R})}, Precision(R,\hat{R}) = \frac{|R \cap \hat{R}|}{|\hat{R}|}, Recall(R,\hat{R}) = \frac{|R \cap \hat{R}|}{|R|}.$$

The quality of the relations provided by existing extraction methods is still lower than the quality of manually constructed relations (see Section 5). This motivates the development of new relation extraction techniques.

One common approach to relation extraction is based on lexico-syntactic patterns such as those proposed by Hearst (1992). We use another extraction principle based on a *semantic similarity measure* between terms. The studied methods extract or recall pairs of semantically similar terms $\langle c_i, c_j \rangle$, but do not return the type of the relationship between them. Nonetheless, we suppose that the extractors must retrieve a mix of synonyms, hypernyms, co-hypernyms, and associations for practical use in NLP systems.

Existing similarity measures rely on one of these four sources of information – semantic networks (Resnik, 1995), Web corpus (Cilibrasi et Vitanyi, 2007), traditional corpora (Lin, 1998b), definitions of dictionaries (Lesk, 1986) or encyclopedia (Zesch *et al.*, 2008a). Prior research (Sahlgren, 2006; Heylen *et al.*, 2008; Panchenko, 2011) suggests that measures based on these sources of information are complementary. The goals of this work is to compare measures based on these four sources of information, and meta-measures combining information from different sources.

The main contributions of this paper are twofold. First, we present a comparative study of the heterogeneous baseline similarity measures. Several authors compared existing measures (see Section 5), but we do it on a large scale. We are the first to compare as many as 34 similarity measures based on the four sources of information listed above. Second, we present two combined metrics which use all the four information sources to calculate similarity (semantic networks, Web corpora, corpora, and definitions). Our experiments show that the measures based on complementary sources of information outperform all baseline measures by a wide margin achieving a correlation with human judgements up to 0.887 and Precision(20) up to 0.979 for the relation extraction task from a closed number of word pairs.

2 Similarity Measures

This section describes 34 knowledge-, web-, corpus-, and definition-based similarity measures, studied in this paper, as well as two combined measures.

Knowledge-based Measures We tested 6 knowledge-based measures based on WordNet (Miller, 1995) and SemCor corpus (Miller *et al.*, 1993) [1] : Inverted Edge Count (Jurafsky et Martin, 2009, p. 687), Leacock et Chodorow (1998), Resnik (1995), Jiang et Conrath (1997), Lin (1998a), and Wu et Palmer (1994). These measures use the following variables to compute the similarities : length of the shortest path in the network between terms c_i and c_j ; length of the shortest path from c_i to the lowest common subsumer (LCS) of c_i and c_j ; length of the shortest path from the root term to the LCS of c_i and c_j ; probability of c, estimated from a corpus ; probability of the LCS of c_i and c_j.

1. We used the implementation available in the package WordNet::Similarity (Pedersen *et al.*, 2004).

The complexity of the knowledge-based measures is mainly bounded by the computation time of the shortest paths between the nodes of the network. A limitation of these measures is that similarities can only be calculated between the 155.287 English terms encoded in the WordNet 3.0. For instance, since the named entity "TALN" is not present in WordNet, no relations between "TALN" and other words can be retrieved. Therefore, these measures are only able to *recall* provided beforehand lexico-semantic knowledge.

Web-based Measures Web-based metrics use the Web as a corpus in order to calculate similarities. They rely on the number of times terms co-occur in documents indexed by a Web search engine. In particular, web-based measures rely on the number of documents (hits) returned by the system by the query "c_i" and the number of hits returned by the query "c_i AND c_j".

We tested 9 measures relying either on Normalized Google Distance (NGD) (Cilibrasi et Vitanyi, 2007) or on Pointwise Mutual Information (PMI-IR) formula (Turney, 2001). We experimented with 5 NGD measures based respectively on BING, YAHOO, YAHOOBOSS, GOOGLE, and GOOGLE over the domain `wikipedia.org`, and with 4 PMI-IR measures based respectively on BING, YAHOOBOSS, GOOGLE, and GOOGLE over the domain `wikipedia.org`. [2]

The complexity of the web-based measures is mainly bounded by the maximum number of queries per second. For instance, BING allows not more than 7 queries per second for free ; GOOGLE allows 100 queries per day for free or 1000 queries for 5$; YAHOO asks 0.80$ for 1000 queries [3]. Web-based measures provide huge coverage of vocabulary in tens of languages. Therefore they are able to extract *new* lexico-semantic knowledge.

Corpus-based Measures We experimented with 13 measures which calculate the similarity between terms based on statistics derived from a corpus. Ten of them are based on the Distributional Analysis (Sahlgren, 2006; Curran, 2003). These distributional measures use 800M token corpus WACYPEDIA (Baroni *et al.*, 2009) tagged with TREETAGGER (Schmid, 1994) and dependency-parsed with MALTPARSER (Hall *et al.*, 2011). The distributional measures use context window or syntactic context techniques to calculate the similarities.

Our implementation of the distributional measures builds a feature matrix $\mathbf{F}$ from a corpus D, such that each term $c_i \in C$ is represented with a row-vector $\mathbf{f}_i$. The feature matrix is then normalized with Pointwise Mutual Information :

$$f_{ij} = log \frac{P(c_i, f_j)}{P(c_i)P(f_j)} = log \frac{f_{ij}}{n(c_i)\sum_i f_{ij}}. \quad (1)$$

Here, f_{ij} is an element of $\mathbf{F}$ is the number of times term c_i was represented with the feature f_j, $n(c_i)$ is the frequency of term c_i in the corpus. Finally, the similarity between the terms c_i and c_j is computed as the cosine between their respective feature vectors $\mathbf{f}_i, \mathbf{f}_j$:

$$s_{ij} = sim(c_i, c_j) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\left\|\mathbf{f}_i\right\| \left\|\mathbf{f}_j\right\|}. \quad (2)$$

Our choice of cosine among other metrics is in line with previous findings (Curran, 2003; Panchenko, 2011). The different distributional measures only vary in the way they build feature

2. Our own system is used in the experiments with measures based on BING (`http://www.bing.com/toolbox/bingdeveloper/`) and YAHOOBOSS (`http://developer.yahoo.com/search/boss/`), and Measures of Semantic Relatedness (MSR) web service (`http://cwl-projects.cogsci.rpi.edu/msr/`) is used for the measures based on GOOGLE and YAHOO !.

3. These rates were up-to date on April 2012. It is likely that the Bing API will be commercialized in future similarly to the YahooBoss.

vectors. The first seven measures perform a Bag-of-words Distributional Analysis (BDA). So, they construct the feature matrix **F** with the context window technique (Van de Cruys, 2010). We tested seven context window sizes – 1, 2, 3, 5, 8, 10 words, and a sentence. A term is represented with a bag of lemmas from a context window, passing a stop-word filter (around 900 words) and a stop part-of-speech filter (nouns, adjectives and verbs are kept).

The other three measures perform Syntactic Distributional Analysis (SDA). So, they construct the feature matrix **F** with the syntactic context technique (Lin, 1998b; Van de Cruys, 2010). Let the term $c_i =$ "cat" be linked with syntactic dependency $dt_j =$ OBJ with the word $w_k =$ "catch". Syntactic context of the term c_i is a bag of dependency-word pairs linked to it $\{\langle dt_j, w_k\rangle : w_k \notin Stoplist \wedge dt_k \in DT\}$, where DT is a set of dependency types used by a measure. [4]

In addition to these 10 distributional measures, we test 3 corpus-based measures available via the MSR web service. Two of them are based on the Factiva corpus (Veksler *et al.*, 2008), and use NGD and PMI-IR similarity functions (see above). The third measure rely on the Latent Semantic Analysis (Landauer et Dumais, 1997), trained on the TASA corpus (Veksler *et al.*, 2008). LSA calculates the similarity of terms with cosine (2) between term vectors in the "concept space".

The complexity of the corpus-based measures is mainly bounded by the time required to pre-process a corpus. In that respect, NGD and PMI-IR are the fastest methods, since they only require a corpus to be indexed in a standard way. BDA require more computational resources since pairwise similarities should be calculated between high-dimensional term vectors. Finally, LSA and SDA are the least scalable methods since the former performs a computationally heavy singular value decomposition of the term-document matrix, and the latter requires dependency parsing of the corpus. Similarly to web-based methods, corpus-based measures are able to extract relations between unknown terms. However, extraction capability of such measures is limited by the corpus – if "TALN" does not occur in the text then it would be impossible to obtain its relations.

Definition-based Measures We experimented with 6 measures which rely on explicit definitions of terms. The first four measures use definitions and relations of Wiktionary and abstracts of Wikipedia. [5] Our implementation of these four measures is similar to the techniques proposed by Zesch *et al.* (2008b). Our measures are different from the previously proposed in three aspects : (a) they represent each term c_i as a bag-of-words vector, while the measures of Zesch *et al.* (2008b) represent terms as concept vectors [6] ; (b) we use both texts from Wiktionary and Wikipedia in order to represent a term, which is not the case in the original work ; (c) we use semantic relations listed in Wiktionary to update similarity scores.

Algorithm 1 depicts pseudocode of these measures. First, it builds the definitions D for input terms C from the information available in Wiktionary and Wikipedia. The function $get\_wiktionary\_def$ returns for each term $c \in C$ a text composed of glosses, examples, quotations, related words, and categories found in Wiktionary (all meanings corresponding to a surface form of c are used). We remove syntax- and etymology-related categories such as "English nouns" or "Japanese proper names" with a stoplist of 94 words, such as "noun" or "esperanto". Next, the function $get\_wikipedia\_def$ returns for each term c a short abstract from the corresponding

4. We tested three models which use 6, 9, or 21 types of syntactic dependencies : $DT_6 =$ { NMOD, SBJ, OBJ, COORD, AMOD, IOBJ} ; $DT_9 =$ { NMOD, ADV, SBJ, OBJ, VMOD, COORD, AMOD, PRN, IOBJ } ; $DT_{21} =$ { NMOD, P, PMOD, ADV, SBJ, OBJ, VMOD, COORD, CC, VC, DEP, PRD, AMOD, PRN, PRT, LGS, IOBJ, EXP, CLF, GAP }.

5. We experimented with data downloaded on October 2011 from `www.wiktionary.org` and `www.dbpedia.org`.

6. An element f_{ij} of a *concept vector* equals to tf.idf score of term c_i in the definition d_j, while an element of *bag-of-words vector* f_{ij} equals to normalized frequency of word c_j in the definition d_i of term c_i.

Algorithm 1: Wiktionary-based sim.measure

Input: Terms C, $UseWikipedia$, Number of features β
Output: Similarity matrix, $\mathbf{S}$ $[C \times C]$

$D \leftarrow get\_wiktionary\_def(C)$;
if $UseWikipedia$ **then**
  $D \leftarrow D \cup get\_wikipedia\_def(C)$
$\mathbf{F} \leftarrow construct\_fmatrix(C, D, \beta)$;
$\mathbf{F} \leftarrow pmi(\mathbf{F})$;
$\mathbf{S} \leftarrow cos(\mathbf{F})$;
$\mathbf{S} \leftarrow update\_similarity(\mathbf{S})$;
return $\mathbf{S}$;

Algorithm 2: Relation fusion sim.measure

Input: Sim.matrices produced by N measures $\{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$, kNN threshold k
Output: Similarity matrix, $\mathbf{S}_{cmb}[C \times C]$

for i=1,N **do**
  $R_i \leftarrow threshold(\mathbf{S}_i, k)$;
  $\mathbf{R}_i \leftarrow relation\_matrix(R_i)$
$\mathbf{S}_{cmb} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{R}_i$;
return $\mathbf{S}_{cmb}$;

Wikipedia article (the name of the article must *exactly* match the term c). Next, the feature matrix $\mathbf{F}$ is constructed : each term $c_i \in C$ is represented as a bag-of-words vector $\mathbf{f}_i$, derived from its definition. These feature vectors are normalized with Pointwise Mutual Information (1). Pairwise similarities of terms are calculated with cosine (2). Finally, the pairwise similarities are corrected with the function $update\_similarity$. It assigns the highest similarity score to the pairs of terms which are directly related in Wiktionary :

$$s_{ij}^{updated} = \begin{cases} 1 & \text{if semantic relation } \langle c_i, c_j \rangle \text{ is listed in Wiktionary} \\ s_{ij} & \text{otherwise} \end{cases} \quad (3)$$

We tested four variations of this measure : two of them use only Wiktionary (1000 and 2500 features β), while the others use both Wiktionary and Wikipedia (1000 and 2500 features β). [7]

In addition to these four measures, we tested two measures based on WordNet glosses available in the package WordNet::Similarity : Extended Lesk (Banerjee et Pedersen, 2003) and Gloss Vectors (Patwardhan et Pedersen, 2006). The key difference between Wiktionary- and WordNet-based measures is that the latter uses definitions of related terms.

The complexity of the definition-based measures is mainly bounded by the time required to preprocess definitions and calculate pairwise similarities between them. In that respect, measures based on Wiktionary and WordNet are similar since they use the bag-of-word model to represent terms. The extraction capability of definition-based measures is limited by the number of available definitions. As of October 2011 WordNet contains 117.659 definitions (glosses) ; Wiktionary contains 536.594 definitions in English and 4.272.902 definitions on all languages ; Wikipedia has 3.866.773 English articles and 20.8 million of articles for all languages.

Combined Similarity Measures We tested two combination techniques – similarity and relation fusion. These methods take as input a set of similarity matrices $\{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$ produced by N combined measures. The output of a combination is a similarity matrix $\mathbf{S}_{cmb}$.

Similarity fusion combines N similarity measures with a simple mean over their respective pairwise similarity scores : $\mathbf{S}_{cmb} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{S}_i$.

Relation fusion keeps only the best relations provided by each measure ; then all these relations are merged. First, the algorithm retrieves the relations extracted by single measures with function

7. We used the JWKTL library (Zesch *et al.*, 2008a) as an API to Wiktionary, and `DBpedia.org` as a source of Wikipedia abstracts. In particular, we used this version of abstracts : `http://downloads.dbpedia.org/3.7/en/long_abstracts_en.nt.bz2`

$threshold$ (a kNN technique described in Section 3). Then each set of relations R_i is encoded in an adjacency matrix $\mathbf{R}_i$. An element of this matrix indicates if the terms c_i and c_j are related :

$$r_{ij} = \begin{cases} 1 & \text{if } \langle c_i, c_j \rangle \in R_k \\ 0 & \text{else} \end{cases} \quad (4)$$

The final similarity score is an average over adjacency matrices (line 4). In our experiments we empirically chose an internal kNN threshold k of 20%.

Expert approach was used to compose three groups of measures of the 34 measures. These groups of measures are combined with two techniques described above. The first group contains 4 measures (see Tables 1 and 2) : WN-Resnik, BDA-3-5000, SDA-21-100000, Def-WktWiki-1000. The second group contains 8 measures – the 4 previous ones plus WN-WuPalmer, LSA-Tasa, Def-GlossVec., and Def-Ext.Lesk. The third group contains 14 measures – the 8 previous ones plus WN-LeacockChodorow, WN-Lin, WN-JiangConrath, NGD-Factiva, NGD-Yahoo, and NGD-GoogleWiki. The running time required to calculate a similarity with a combined measure is close to the sum of times required by the measures used in a combination.

3 Evaluation

Our comparison of similarity measures is based on human judgments about semantic similarity and on semantic relations fixed manually by lexicographers [8].

Human Judgements This kind of evaluation is a standard and simple way to assess a semantic similarity measure. We used three classical human judgement datasets – MC (Miller et Charles, 1991), RG (Rubenstein et Goodenough, 1965) and WordSim353 (Finkelstein *et al.*, 2001) composed of 30, 65, and 353 pairs of terms respectively. Each of these datasets is composed of N tuples $\langle c_i, c_j, s_{ij} \rangle$, where c_i, c_j are terms, and s_{ij} is their similarity obtained by human judgement. Let $\mathbf{s} = (s_{i1}, s_{i2}, \ldots, s_{iN})$ be a vector of ground truth scores, and $\hat{\mathbf{s}} = (\hat{s}_{i1}, \hat{s}_{i2}, \ldots, \hat{s}_{iN})$ be a vector of similarity scores calculated by a measure. Then, the quality of the measure is assessed with Pearson and Spearman's correlation between $\mathbf{s}$ and $\hat{\mathbf{s}}$.

Semantic Relations This ground truth is composed of semantic relations $\langle c_i, type, c_j \rangle$, such as ⟨agitator, synonym, activist⟩, ⟨dishwasher, co-hyponym, freezer⟩, ⟨hawk , hypernym, predator⟩, and ⟨gun, synonym,weapon⟩. The dataset contains both meaningful and random relations. The evaluation is based on the number of correctly ranked relations. In order to extract relations R between a set of terms C, we follow a standard procedure. First, pairwise similarities between terms are calculated and saved in a $[C \times C]$ similarity matrix $\mathbf{S}$. The similarity scores are mapped to the interval $[0; 1]$. Second, each term c_i is linked with k% of its nearest neighbours : $\hat{R} = \bigcup_{i=1}^{|C|} \left\{ \left\langle c_i, c_j \right\rangle : (c_j \in \text{ top } k\% \text{ terms of } c_i) \wedge (s_{ij} \geq 0) \right\}, s_{ij} \in \mathbf{S}$.

Let $\hat{R}_k$ be a set containing top k % semantic relations for each target word c_i, and R be a set of all correct semantic relations. Then, Precision, Recall, F1-measure at k are calculated as follows : $P(k) = \frac{|R \cap \hat{R}_k|}{|\hat{R}_k|}, R(k) = \frac{|R \cap \hat{R}_k|}{|R|}, F(k) = \frac{P(k) \cdot R(k)}{P(k)+R(k)}$. Each "target" term c_i has roughly the same number of meaningful and random relations. That is why for a random measure $P(50) \approx 0.5$ and not $\frac{|\hat{R}|}{|C^2|} \approx 0$ as in the case of an open vocabulary relation extraction. We argue that this kind

8. Evaluation datasets and scripts are available at : `http://cental.fltr.ucl.ac.be/team/~panchenko/sre-eval/`

of evaluation should give a good idea about the relative performances of different measures. However, the performance scores in this evaluation should not be confused with the performance scores in an open-vocabulary relation extraction task. In this work, the quality of a similarity measure is assessed with the four statistics : $P(10)$, $P(20)$, $P(50)$, $F(50)$.

We used two semantic relation datasets : BLESS (Baroni et Lenci, 2011), and SN. The first one relates 200 target terms (100 animate and 100 inanimate nouns) to 8625 relatum terms with 26.554 semantic relations (14.440 are meaningful and 12.154 are random). Every relation has one of the following types : hypernymy, co-hypernymy, meronymy, attribute, event, or random. We built the SN (Semantic Neighbors) dataset in order to complement the BLESS, because it contains no synonyms. [9] SN relates 462 target terms (nouns) to 5910 relatum terms with 14.682 semantic relations (7341 are meaningful and 7341 are random). The SN contains synonyms coming from three sources : WordNet 3.0 (Miller, 1995), Roget's thesaurus (Kennedy et Szpakowicz, 2008), and a synonyms database [10].

4 Results

Human Judgements Table 1 presents correlations of the 34 single and the 3 combined measures with human judgements. We ranked the measures according to their Spearman's correlation. The best measures in each group (knowledge-, web-based etc.) are in bold. We observed that correlations of most web-based measures with human judgements are low and not significant in most of the cases. PMI-IR and NGD over Wikipedia are two exceptions. They provided the best results among the web measures. However, generally, knowledge-, corpus-, and definition-based measures perform far better than those relying on the Web as a corpus. Particularly high correlations with human judgements were observed for the following single similarity measures : *WN-Resnik, SDA-21-100000, Def-WktWiki-1000, BDA-3-5000*, and *WN-LeacockChodorow*. However, the similarity fusion of 14 measures *Cmb-Avg-14* outperformed all single measures on MC and RG datasets. In the same time, similarity fusion of 8 measures (*Cmb-Avg-8*) was better that any single measure on the WordSim353 pairs.

Semantic Relations Table 2 presents performance of the measures at relation extraction. We ranked the measures according to P(20) and P(50) statistics. We would like to recall that our evaluation procedure is different from an open vocabulary extraction and a random measure would achieve $P(50) \approx 0.5$ (see the first line of Table 2). The knowledge-,web-, corpus-,and definition-based measures are grouped and the best metrics in each group are in bold. Figure 1(c) depicts Precision-Recall graph of four variations of the definition-based measures. The following single measures provided the best scores in this evaluation : *WN-Resnik, SDA-21-100000, BDA-3-5000, Def-WktWiki-1000*, and *WN-WuPalmer*.

Our experiments showed that measures which use both Wiktionary and Wikipedia (denoted as *Def-WktWiki-**) are better on most of the datasets than measures relying only on Wiktionary (*Def-Wkt-**). In particular, *Def-WktWiki-1000* outperformed all definition-based measures, including those based on WordNet. On the BLESS dataset, the syntactic distributional analysis *SDA-21-10000* achieved the best precision among the single measures (0.953), while bag-of-words distributional analysis *BDA-3-5000* achieved the highest recall (0.835). On the SN dataset, the

9. SN dataset is available at `http://cental.fltr.ucl.ac.be/team/~panchenko/sre-eval/sn.csv`
10. `http://synonyms-database.downloadaces.com/`

WordNet-based measure WN-WuPalmer performed best achieving P(20) of 0.959 and P(50) of 0.764. However, the relation fusion of 8 measures (*Cmb-Rel-8*) outperformed all single measures on both datasets achieving P(20) of 0.975 and P(50) of 0.802 on the BLESS and P(20) of 0.971 and P(50) of 0.760 on the SN dataset.

Summary Results obtained on the human judgements and semantic relation datasets are overlapping but not identical. We used the following criterion in order to decide which measures are the best : a measure should be the best in its group (e. g. among corpus-based measures) in both types of evaluations. According to this criterion, the best single metrics are the WordNet measure *WN-Resnik*, the bag-of-words distributional measure *BDA-3-5000*, the syntactic distributional measure *SDA-21-100000*, and the measure *Def-WktWiki-1000* based on Wiktionary and Wikipedia. Figure 1 depicts distributions of similarity scores for these four most successful metrics. Our experiments showed that, for these measures there is a significant difference in distributions of scores of meaningful and random relations. This means that an appropriate kNN threshold level k clearly separates meaningful relations from the random ones. The best combined measure and the best measure overall is *Cmb-Rel-8*. It is based on the eight following measures : *WN-Resnik, BDA-3-5000, SDA-21-100000, Def-WktWiki-1000, WN-WuPalmer, LSA-Tasa, Def-GlossVec.*, and *Def-Ext.Lesk*. This result is interesting as combination of the four strongest measures (those listed in Figure 1 and denoted as *Cmb-*-4* in Tables 1 and 2) can benefit of redundancy provided by the additional weaker measures. Our results suggest that performance of the combinations based on 14 measures is very close to the performance of *Cmb-Rel-8* (see Figure 1(b) and Table 3). Thus, redundancy provided by the additional 6 measures does not improve the results with respect to the set of 8 measures.

Sim.Measure	MC Dataset		RG Dataset		WordSim353 Dataset	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Random	0.172 ***	0.056 ***	-0.060 ***	-0.047 ***	-0.158 ***	-0.122 ***
WN-Resnik	0.823	**0.784**	0.823	0.757	0.350	**0.330**
WN-Short.Path	0.755	0.724	0.782	0.788	0.366	0.290
WN-Leack.Chod.	0.779	0.724	0.841	**0.789**	0.313	0.295
WN-WuPalmer	0.768	0.742	0.800	0.775	0.270	0.330
WN-Lin	0.769	0.754	0.737	0.619	0.287	0.203
WN-JiangConrath	0.473 *	0.719	0.575	0.587	0.227	0.175
NGD-Bing	0.035 ***	0.063 ***	0.174 ***	0.181 ***	0.042 ***	0.058 ***
NGD-Yahoo	0.387 **	0.330 ***	0.448	0.445	0.290	0.254
NGD-Google	0.085 ***	0.019 ***	-0.013 ***	-0.012 ***	0.120 **	0.150 *
NGD-GoogleWiki	0.306 ***	0.334 ***	0.452	**0.501**	0.205	0.250
PMI-IR-Bing	0.079 ***	0.120 ***	0.116 ***	0.149 ***	0.000 ***	0.003 ***
PMI-IR-Google	0.046 ***	-0.107 ***	-0.061 ***	-0.039 ***	0.097 ***	0.113 **
PMI-IR-GoogleWiki	0.508 *	**0.498** *	0.401	0.411	0.254	**0.279**
BDA-sent-10000	0.642	0.638	0.694	0.703	0.383	0.362
BDA-1-5000	0.658	0.676	0.704	0.758	0.448	0.438
BDA-2-5000	0.667	0.638	0.698	0.734	0.441	0.439
BDA-3-5000	0.722	0.692	0.752	0.782	0.467	**0.465**
BDA-5-5000	0.710	0.683	0.755	**0.787**	0.467	0.455
BDA-8-5000	0.707	0.697	0.746	0.764	0.455	0.440
BDA-10-5000	0.710	0.718	0.746	0.764	0.443	0.425
SDA-6-100000	0.759	0.790	0.741	0.792	0.380	0.496
SDA-9-100000	0.756	0.790	0.732	0.787	0.384	0.491
SDA-21-100000	0.756	**0.790**	0.731	0.785	0.384	0.490
LSA-Tasa	0.737	0.694	0.645	0.604	0.527	0.565
NGD-Factiva	0.602	0.602	0.618	0.599	0.565	0.599
PMI-Factiva	0.312 ***	0.442 **	0.436	0.517	0.314	0.559
Def-WN-GlossVec.	0.566	0.653	0.647	0.738	0.383	0.322
Def-WN-Ext.Lesk	0.355 ***	0.792	0.340 *	0.717	0.209	0.409
Def-Wkt-1000	0.625	0.687	0.655	0.760	0.416	0.492
Def-Wkt-2500	0.625	0.687	0.655	0.760	0.382	0.527
Def-WktWiki-1000	0.704	**0.759**	0.701	**0.754**	0.453	**0.545**
Def-WktWiki-2500	0.704	0.759	0.701	0.754	0.416	0.520
Cmb-Avg-4	0.847	0.859	0.867	0.887	0.500	0.508
Cmb-Avg-8	0.858	0.858	0.867	0.883	0.537	**0.555**
Cmb-Avg-14	0.847	**0.859**	0.867	**0.887**	0.500	0.508

TABLE 1 – Evaluation on the human judgement datasets (MC, RG, and WordSim353). Here (*) means $p \leq 0.01$, (**) means $p \leq 0.05$, (***) means $p > 0.05$, otherwise $p \leq 0.001$. The best results for each group of measures are in bold. The very best results are in grey.

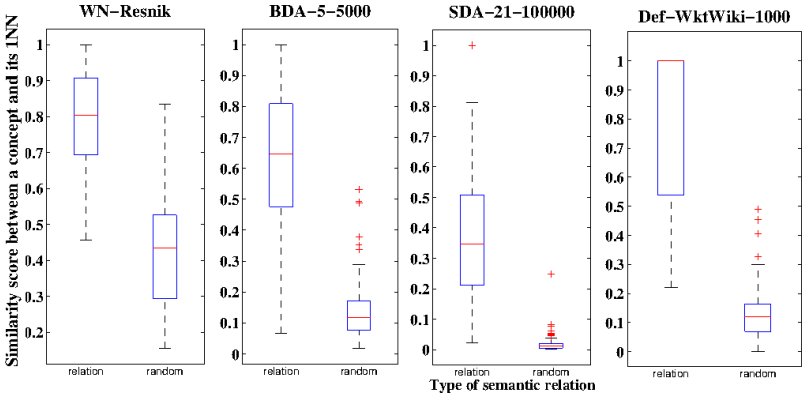


FIGURE 1 – Distribution of 1-NN similarity scores of the four best single measures on the BLESS dataset. Here "random" and "relation" are distributions of scores between random and meaningful relations. The distributions were calculated as suggested in (Baroni et Lenci, 2011).

Sim.Measure	BLESS Dataset				SN Dataset			
	P(10)	*P (20)*	*P(50)*	*F(50)*	*P(10)*	*P(20)*	*P(50)*	*F(50)*
Random	0.546	0.541	0.543	0.522	0.504	0.501	0.498	0.498
WN-Resnik	0.977	**0.958**	**0.718**	0.690	0.948	0.908	0.725	0.725
WN-Short.Path	0.967	0.925	0.722	0.693	0.981	0.947	0.752	0.752
WN-Leack.Chod.	0.967	0.925	0.722	0.693	0.982	0.951	0.756	0.756
WN-WuPalmer	0.978	0.938	0.706	0.678	0.979	**0.959**	**0.764**	0.764
WN-Lin	0.975	0.919	0.776	0.745	0.924	0.853	0.637	0.637
WN-JiangConrath	0.981	0.909	0.732	0.703	0.916	0.835	0.615	0.615
NGD-Bing	0.725	0.692	0.695	0.670	0.676	0.682	0.639	0.639
NGD-Yahoo	0.940	0.907	**0.782**	0.751	—	—	—	—
NGD-YahooBoss	0.847	0.843	0.747	0.718	—	—	—	—
NGD-Google	0.991	**0.934**	0.651	0.625	—	—	—	—
NGD-GoogleWiki	0.874	0.836	0.702	0.674	—	—	—	—
PMI-IR-Bing	0.675	0.650	0.692	0.667	0.610	0.608	0.647	0.647
PMI-IR-YahooBOSS	0.823	0.822	0.724	0.696	—	—	—	—
PMI-IR-Google	0.822	0.749	0.660	0.634	—	—	—	—
PMI-IR-GoogleWiki	0.791	0.761	0.676	0.649	—	—	—	—
BDA-sent-10000	0.962	0.920	0.799	0.767	0.941	0.898	0.724	0.724
BDA-1-5000	0.971	0.940	0.826	0.793	0.969	0.926	0.737	0.737
BDA-2-5000	0.966	0.939	0.829	0.796	0.970	0.929	0.738	0.738
BDA-3-5000	0.970	0.947	**0.835**	0.802	0.974	0.932	0.743	0.743
BDA-5-5000	0.975	0.946	0.833	0.800	0.971	0.929	0.744	0.744
BDA-8-5000	0.974	0.943	0.827	0.794	0.968	0.924	0.741	0.741
BDA-10-5000	0.972	0.941	0.821	0.789	0.962	0.922	0.737	0.737
SDA-6-100000	0.984	0.948	0.810	0.778	0.978	0.945	0.749	0.749
SDA-9-100000	0.984	0.951	0.809	0.777	0.977	0.945	0.753	0.753
SDA-21-100000	0.985	**0.953**	0.810	0.778	0.978	**0.946**	**0.753**	0.753
LSA-Tasa	0.967	0.936	0.801	0.769	0.901	0.839	0.637	0.637
NGD-Factiva	0.959	0.916	0.800	0.768	0.900	0.832	0.651	0.651
PMI-Factiva	0.903	0.860	0.816	0.784	0.826	0.768	0.606	0.606
Def-WN-GlossVec.	0.894	0.860	0.742	0.712	0.930	0.872	0.719	0.719
Def-WN-Ext.Lesk	0.940	0.870	0.716	0.687	0.950	**0.895**	0.653	0.653
Def-Wkt-1000	0.926	0.885	0.783	0.752	0.907	0.868	0.678	0.678
Def-Wkt-2500	0.915	0.882	0.754	0.754	0.928	0.898	0.704	0.704
Def-WktWiki-1000	0.942	**0.905**	**0.785**	0.725	0.917	0.878	0.696	0.696
Def-WktWiki-2500	0.931	0.891	0.765	0.734	0.937	0.912	**0.726**	0.726
Cmb-Avg-4	0.992	0.969	0.787	0.756	0.980	0.952	0.768	0.768
Cmb-Rel-4	0.989	0.970	0.737	0.708	0.975	0.943	0.696	0.696
Cmb-Avg-8	0.994	0.974	0.774	0.743	0.955	0.875	0.660	0.660
Cmb-Rel-8	0.994	0.975	**0.802**	0.770	0.989	**0.971**	**0.760**	0.760
Cmb-Avg-14	0.994	**0.979**	0.792	0.760	0.957	0.880	0.663	0.663
Cmb-Rel-14	0.994	0.973	0.811	0.779	0.987	0.966	0.759	0.759

TABLE 2 – Evaluation of the measures on the semantic relation datasets (BLESS and SN). Here *P(x)*, and *F(x)* are Precision, and F-measure as specified in Section 3. The best results for each group of measures are in bold. The very best results are in grey.

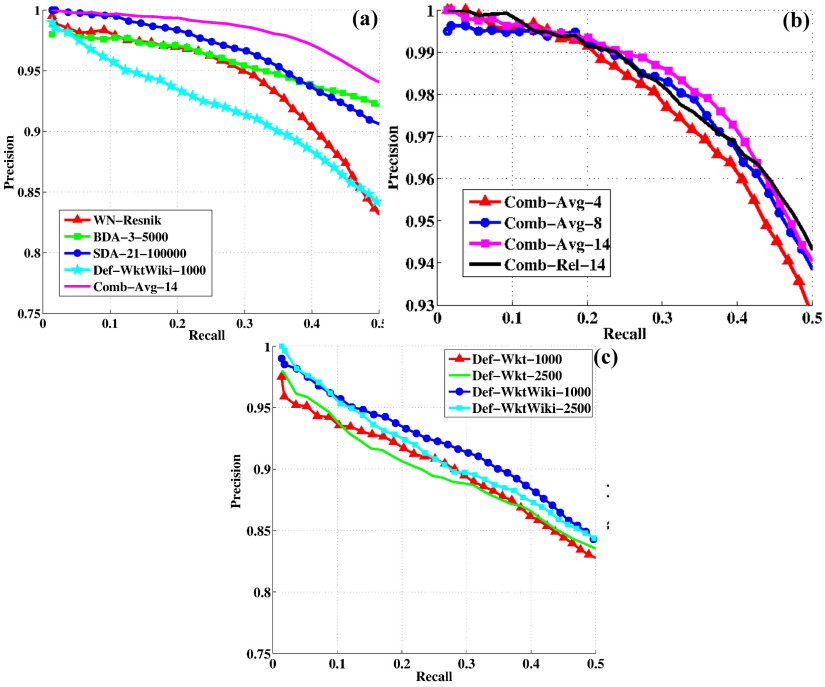


FIGURE 2 – Precision-Recall graphs of (a) the best single and combined measures ; (b) four combined measures ; (c) measures based on Wiktionary and Wikipedia.

Discussion There is a huge difference in performance between web-based and corpus-based measures. This is likely to be due to the noisy nature of the web documents (BDA/SDA use a more precise and linguistically motivated representation of a term) and the fact that the counts of a search engine API are rough approximations of the real counts. Similarly, the higher performance of the knowledge- and definition-based methods is likely due to the more linguistically precise representation of the terms. Some web measures yield significantly worst results than others. Following (Veksler *et al.*, 2008), we suggest that the variance in the results are due to differences in the corpora indexed by different search engines. For instance, Web measures over Wikipedia or Factiva provide better results since this corpora contain less noisy documents than the heterogeneous Web collection indexed by Bing.

Combined measures achieve higher precision and recall with respect to the single measures. First, this is due to the reuse of common lexico-semantic information (such as “car” being a synonym of “vechicle”) via knowledge- and definition-based measures. Measures based on WordNet and dictionary definitions achieve high precision as they rely on fine-grained manually constructed

resources. However, due to limited coverage of these resources they can only determine relations between a limited number of terms. On the other hand, measures based on web and corpora are nearly unlimited in their coverage, but provide less precise results. Combination of the measures let us keep high precision for frequent terms present in WordNet and dictionaries and at the same time calculate relations between rare terms unlisted in the handcrafted resources with web and corpus measures.

Second, combinations work well because, as it was found in previous research (Sahlgren, 2006; Heylen *et al.*, 2008; Panchenko, 2011), different measures provide complementary types of semantic relations. For instance, WordNet-based measures score high hypernyms, distributional analysis score high co-hypernymy and synonyms, etc. In that respect, a combination helps to recall more diverse relations. For example, a WordNet-based measure may return the hyponym ⟨salmon, seafood⟩, while a corpus-based measure would extract the co-hypernym ⟨salmon, mackerel⟩.

5 Related Work

There exists a significant body of literature about single measures discussed in this paper. However, just a few works compared different measures and their combinations. Furthermore, even less people evaluated the performance of these measures on the relation extraction task. One notable exception is the work of Curran et Moens (2002). The authors evaluated nine BDA measures and 14 weight functions and reported $Precision(5)$ of 0.52, and $Precision(10)$ of 0.45 for the best measure – Jaccard similarity with t-test weight function. Van de Cruys (2010) studied distributional measures and reported that : the optimal context window sizes for BDA is 2-5 words ; SDA is the best distributional measure. Budiu *et al.* (2007) compared LSA, PMI-IR, and GLSA. The authors found that GLSA performs better on the synonymy tests, while PMI-IR works better on the human judgement datasets. Agirre *et al.* (2009) compared 3 WordNet-based and 20 distributional measures (BDA and SDA) as well as their combinations. The authors found that a supervised combination of distributional and WordNet measures outperforms all measures on all datasets. Similarity measures which rely on Wikipedia, Wiktionary, WordNet and their combinations are described in the work of Zesch *et al.* (2007, 2008b). Navarro *et al.* (2009) described another method for extraction of synonyms from Wiktionary. Two promising measures which rely on Wikipedia were proposed by Strube et Ponzetto (2006) and Gabrilovich et Markovitch (2007).

Some studies compare the measures in context of NLP applications. For instance, Mihalcea *et al.* (2006) studied PMI-IR, LSA, and six WordNet-based measures on the text similarity task. The authors found that PMI-IR and Resnik are best corpus- and knowledge-based measures correspondingly ; and that an average over eight measures outperforms single measures. Budanitsky et Hirst (2006) found that the WN-JiangConrath is the best knowledge-based measure for the spelling correction application. Patwardhan et Pedersen (2006) report the same result for the task of word sense disambiguation. SDA was used by Grefenstette (1994) to induce a thesaurus.

In prior research, some attempts were made to combine baseline measures, including (Curran, 2002; Cederberg et Widdows, 2003; Mihalcea *et al.*, 2006; Agirre *et al.*, 2009). However, those studies did not take into account the whole range of existing information sources.

6 Conclusion

In this paper we compared 34 knowledge-, corpus-, web-, and definition-based measures on the task of predicting semantic similarity scores and semantic relations that hold between two terms. We also described and tested two techniques for measure combination. Our results show that the combined measures outperform all single measures achieving a correlation of 0.887 on RG dataset and $Precision(20)$ of 0.979 on the BLESS dataset. In the future research, we are going to estimate the precision of the relation extraction on the whole vocabulary C. The obtained relations will be applied in context of text classification and query expansion applications.

Références

AGIRRE, E., ALFONSECA, E., HALL, K., KRAVALOVA, J., PAŞCA, M. et SOROA, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. *In Proceedings of NAACL-HLT 2009*, pages 19–27.

BANERJEE, S. et PEDERSEN, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In IJCAI*, volume 18, pages 805–810.

BARONI, M., BERNARDINI, S., FERRARESI, A. et ZANCHETTA, E. (2009). The wacky wide web : A collection of very large linguistically processed web-crawled corpora. *LREC*, 43(3):209–226.

BARONI, M. et LENCI, A. (2011). How we blessed distributional semantic evaluation. *Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 1–11.

BUDANITSKY, A. et HIRST, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

BUDIU, R., ROYER, C. et PIROLLI, P. (2007). Modeling information scent : A comparison of lsa, pmi and glsa similarity measures on common tests and corpora. pages 314–332. In RIAO.

CEDERBERG, S. et WIDDOWS, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *In Proceedings HLT-NAACL*, pages 111–118.

CILIBRASI, R. L. et VITANYI, P. M. B. (2007). The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.

CURRAN, J. R. (2002). Ensemble methods for automatic thesaurus extraction. *In Proceedings of the EMNLP-02*, pages 222–229. ACL.

CURRAN, J. R. (2003). *From distributional to semantic similarity*. Thèse de doctorat, University of Edinburgh.

CURRAN, J. R. et MOENS, M. (2002). Improvements in automatic thesaurus extraction. *In Proceedings of the ACL-02 workshop on Unsupervised Lexical Acquisition*, pages 59–66.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G. et RUPPIN, E. (2001). Placing search in context : The concept revisited. *In WWW 2001*, pages 406–414. ACM.

GABRILOVICH, E. et MARKOVITCH, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In IJCAI*, volume 6, page 12.

GREFENSTETTE, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Springer.

HALL, J., NILSSON, J. et NIVRE, J. (2011). Single malt or blended ? a study in multilingual parser optimization. volume 43 de *Text, Speech and Language Technology*, pages 19–33. Springer.

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics*, pages 539–545. ACL.

HEYLEN, K., PEIRSMAN, Y., GEERAERTS, D. et SPEELMAN, D. (2008). Modelling word similarity : an evaluation of automatic synonymy extraction algorithms. *LREC'08*, pages 3243–3249.

HSU, M.-H., TSAI, M.-F. et CHEN, H.-H. (2006). Query expansion with conceptnet and wordnet : An intrinsic comparison. *Information Retrieval Technology*, pages 1–13.

JIANG, J. J. et CONRATH, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In ROCLING X*, pages 19–33.

JURAFSKY, D. et MARTIN, J. H. (2009). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

KENNEDY, A. et SZPAKOWICZ, S. (2008). Evaluating rogets thesauri. *ACL-08 HLT*, pages 416–424.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

LEACOCK, C. et CHODOROW, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database*, pages 265–283.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

LIN, D. (1998a). An Information-Theoretic Definition of Similarity. *In In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

LIN, D. (1998b). Automatic retrieval and clustering of similar words. *In Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. ACL.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI'06*, pages 775–780.

MILLER, G. A. (1995). Wordnet : a lexical database for english. *Communications of ACM*, 38(11):39–41.

MILLER, G. A. et CHARLES, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

MILLER, G. A., LEACOCK, C., TENGI, R. et BUNKER, R. T. (1993). A semantic concordance. *In Proceedings of the workshop on Human Language Technology*, pages 303–308. ACL.

NAVARRO, E., SAJOUS, F., BRUNO, G., PRÉVOT, L., SHUKAI, H., TZU-YI, K., MAGISTRY, P. et CHU-REN, H. (2009). Wiktionary and nlp : improving synonymy networks. *In Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, People's Web '09, pages 19–27. Association for Computational Linguistics.

PANCHENKO, A. (2011). Comparison of the baseline knowledge-, corpus-, and web-based similarity measures for semantic relations extraction. *GEMS Workshop (EMNLP)*, pages 11–21.

PATWARDHAN, S., BANERJEE, S. et PEDERSEN, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing*, volume 2588 de *LNCS*, pages 241–257. Springer Berlin.

PATWARDHAN, S. et PEDERSEN, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together*, page 1.

PEDERSEN, T., PATWARDHAN, S. et MICHELIZZI, J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. *In Demonstration Papers at HLT-NAACL 2004*, pages 38–41. ACL.

RESNIK, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In Proceedings of the 14th IJCAI conference.*, volume 1, pages 448–453.

RUBENSTEIN, H. et GOODENOUGH, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

SAHLGREN, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Thèse de doctorat, Stockholm University.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. pages 44–49.

STRUBE, M. et PONZETTO, S. P. (2006). Wikirelate ! computing semantic relatedness using wikipedia. *In Proceedings of the AAAI*, volume 21, pages 14–19.

SUN, R., JIANG, J., FAN, Y., HANG, T., TAT-SENG, C. et yen KAN, C. M. (2005). Using syntactic and semantic relation analysis in question answering. *In Proceedings of TREC*.

TIKK, D., YANG, J. D. et BANG, S. L. (2003). Hierarchical text categorization using fuzzy relational thesaurus. *KYBERNETIKA-PRAHA*, 39(5):583–600.

TURNEY, P. (2001). Mining the Web for Synonyms : PMI-IR versus LSA on TOEFL. *In Proceedings of the twelfth european conference on machine learning (ecml-2001)*.

Van de CRUYS, T. (2010). *Mining for Meaning : The Extraction of Lexicosemantic Knowledge from Text*. Thèse de doctorat, University of Groningen.

VEKSLER, V. D., GOVOSTES, R. Z. et GRAY, W. D. (2008). Defining the dimensions of the human semantic space. *In 30th Annual Meeting of the Cognitive Science Society*, pages 1282–1287.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd meeting on Association for Computational Linguistics*, pages 133–138.

ZESCH, T., GUREVYCH, I. et MÜHLHÄUSER, M. (2007). Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. *In HLT-NAACL 2007*, pages 205–208.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008a). Extracting lexical semantic knowledge from wikipedia and wiktionary. *In Proceedings of LREC'08*, pages 1646–1652.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008b). Using wiktionary for computing semantic relatedness. *In Proceedings of AAAI*, volume 2008, page 45.

Integrating Lexical, Syntactic and System-based Features to Improve Word Confidence Estimation in SMT

Luong Ngoc Quang
Laboratoire LIG, GETALP, Grenoble, France
Ngoc-Quang.Luong@imag.fr

RESUME

Intégration de paramètres lexicaux, syntaxiques et issus du système de traduction automatique pour améliorer l'estimation des mesures de confiance au niveau des mots

L'estimation des mesures de confiance (MC) au niveau des mots consiste à prédire leur exactitude dans la phrase cible générée par un système de traduction automatique. Ceci permet d'estimer la fiabilité d'une sortie de traduction et de filtrer les segments trop mal traduits pour une post-édition. Nous étudions l'impact sur le calcul des MC de différents paramètres : lexicaux, syntaxiques et issus du système de traduction. Nous présentons la méthode permettant de labelliser automatiquement nos corpus (mot correct ou incorrect), puis le classifieur à base de champs aléatoires conditionnels utilisé pour intégrer les différents paramètres et proposer une classification appropriée des mots. Nous avons effectué des expériences préliminaires, avec l'ensemble des paramètres, où nous mesurons la précision, le rappel et la F-mesure. Finalement nous comparons les résultats avec notre système de référence. Nous obtenons de bons résultats pour la classification des mots considérés comme corrects (F-mesure : 86.7%), et encourageants pour ceux estimés comme mal traduits (F-mesure : 36,8%).

ABSTRACT

Confidence Estimation at word level is the task of predicting the correct and incorrect words in the target sentence generated by a MT system. It helps to conclude the reliability of a given translation as well as to filter out sentences that are not good enough for post-editing. This paper investigates various types of features to circumvent this issue, including lexical, syntactic and system-based features. A method to set training label for each word in the hypothesis is also presented. A classifier based on conditional random fields (CRF) is employed to integrate features and determine the word's appropriate label. We conducted our preliminary experiment with all features, tracked precision, recall and F-score and we compared with our baseline system. Experimental results of the full combination of all features yield the very encouraging precision, recall and F-score for Good label (F-score: 86.7%), and acceptable scores for Bad label (F-score: 36.8%).

MOTS-CLES : Système de traduction automatique, mesure de confiance, estimation de la confiance, champs aléatoires conditionnels
KEYWORDS : Machine translation, confidence measure, confidence estimation, conditional random fields

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 43–56,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Statistical Machine Translation Systems in recent years have marked impressive breakthroughs with numerous fruitful achievements, as they produced more and more user-acceptable outputs. Nevertheless users have to face with some big questions that still remain open: are these translations ready to be published or some post-edit operations will be needed? Are they worth to be corrected or the re-translation from scratch is less time-consuming? It is undoubtedly that building a method which is capable of pointing out the correct parts as well as detecting the translation errors in each MT hypothesis is crucial to tackle these above issues. If we limit the concept "parts" to "words", the problem is called Word-level Confidence Estimation.

The objective of Word-level Confidence Estimation is to judge each word in the hypothesis as correct or incorrect by tagging it with an appropriate label. A classifier which has been trained beforehand by a feature set calculates the confidence score for MT output word, and then compares it with a threshold. All words with scores that exceed this threshold are categorized in the Good label set; the rest will belong to Bad label set.

Contributions of Confidence Estimation for the other aspects of Machine Translation are incontestable. Firstly, it assists the post-editors to quickly and intuitively identify the translation errors, and then they can determine whether to correct the sentence or re-translate it from scratch. This support gains lots of both post-editing time and efforts. Second, confidence score assigned to words is a potential clue to re-rank the MT hypothesis, thus improve its translation quality. Last but not least, it can be used by the translator in an interactive scenario (Gandrabur and Foster, 2003).

This article presents the integration of various types of features into CRF model to forecast the label for each word in the MT hypothesis. We organize the remaining parts as follow: in Section 2, we briefly review some previous researches related to confidence estimation at word level. The concept of CRF model, which we use to train our feature set will be introduced in Section 3. Section 4 details various system-based, lexical and syntactic features exploited for the classifier construction. Section 5 lists our settings to prepare for the preliminary experiments. The preliminary experiments together with their results are reported in Section 6. Lastly, section 7 concludes the paper and points out some perspectives.

2 Previous Work Review

To cope with Word-level Confidence Estimation problem, various approaches have been proposed, and most of them concentrate on the two major issues: which type of features and their combinations are efficient? And which classifier is the most suitable for training the feature sets?

In one of the earliest as well as most well-known work in this area, (Blatz et al., 2003) combine a considerable number of features by applying neural network and naïve Bayes learning algorithms. Among these features, the N-best lists based features, especially Word Posterior Probability (henceforth WPP) proposed in (Ueffing et al., 2003) have been shown to be one of the most effective system-based features by their

experimental results. The combination of WPP (with 3 different proposed variants) and the IBM-Model 1 based features are also confirmed to overwhelm all the other single ones, including heuristic and semantic features in terms of performance in (Blatz et al., 2004). Using solely N-best list, (Sanchis et al., 2007) suggest 9 different features and then adopt a smoothed naïve Bayes classification model for training them.

(Ueffing and Ney, 2005) introduce a novel approach which explicitly explores the phrased-based translation model for detecting word errors. The phrase is considered as a contiguous sequence of words and is extracted from word-aligned bilingual training corpus for both source and target sides. The confidence value of each target word is then computed by summing over all phrase pairs in which the target part contains this word. Experimental results indicate that their method yielded an impressive reduction of the classification error rate compared to the state-of-the-art ones on the same language pairs employed.

(Xiong et al., 2010) integrate the POS of target word with another lexical feature named null dependency link and train them by MaxEnt classifier. In their results, the linguistic features sharply outperform word posterior probability feature in terms of F-score and classification error rate.

Unlike most of previous work, (Soricut and Echihabi, 2010) applied solely the external features of MT system with the hope that their classifier can deal with various MT approaches, from statistical-based to rule-based one. Given an MT output, the BLEU score is forecast due to the regression model they developed.

(Bach et al., 2011) study a method to calculate the confidence score for both generated target words and sentences relied on a feature-rich classifier. The features employed include source side information, alignment context, and dependency structure. The integration between them and Word posterior probability and POS context of target language helps to augment marginally in F-score as well as the Pearson correlation with human judgment.

Our work differs from previous researches at these main points: firstly, we investigate and integrate various types of features: system-based features extracted from the MT outputs (N-best lists with the score of the log-linear model as well as source and target language model features), together with lexical and syntactic features to see if this combination helps to improve the classifier's performance. All results observed will be reported in Section 6. Secondly, instead of using Levenshtein alignment or TER-p for generating the training label, we propose to use TERp-A thanks to some advantages which will be pointed out in Section 5. Thirdly, we apply the CRF model for integrating our predictor features as well as to classify words in the test set, which has been proven to avoid limitations of Markov models and stochastic grammars (Lafferty et al., 2001).

3 Conditional Random Fields Model for Confidence Estimation

CRF (Lafferty et al., 2001) is a framework for building probabilistic models for segmenting and labeling sequence data. The core idea of CRF can be summarized as follow: let $X = (x_1, x_2, ..., x_N)$ be the random variable over data sequence to be labeled, $Y = (y_1, y_2, ..., y_N)$ be the output sequence obtained after the labeling task. In our case, X

ranges over words in the MT output, and Y represents the labels tagged for words. Each element $y_i\ (i=\overline{1,N})$ is assigned one value in the binary set $Y^N=\{Good, Bad\}$. The probability of sequence Y given X is written as:

$$p_\theta(Y \mid X)=\frac{1}{Z_\theta(X)}\exp\left\{\sum_{k=1}^{K}\theta_k F_k(X,Y)\right\} \tag{1}$$

where

$$F_k(X,Y)=\sum_{t=1}^{T} f_k(y_{t-1},y_t,x_t) \tag{2}$$

$\left\{f_k\right\}(k=\overline{1,K})$ is a set of feature functions, $\left\{\theta_k\right\}(k=\overline{1,K})$ are the associated parameter values, and $Z_\theta(x)$ is a normalization factor, in which, the value is calculated by:

$$Z_\theta(x)=\sum_{y\in Y^N}\exp\left\{\sum_{k=1}^{K}\theta_k F_k(X,Y)\right\} \tag{3}$$

In order to estimate the conditional maximum likelihood given T independent sequences $\{X^i, Y^i\}$ $(i=\overline{1,T})$ where X^i and Y^i contains N^i symbols, we have to minimize the negated conditional log-likelihood of the observations, with respect to θ:

$$\begin{aligned} l(\theta) &= -\sum_{i=1}^{T}\log p_\theta(Y_i \mid X_i) \\ &= \sum_{i=1}^{T}\left\{\log Z_\theta(X^i)-\sum_{k=1}^{K}\theta_k F_k(X^i,Y^i)\right\} \end{aligned} \tag{4}$$

The standard solution for this minimization is to apply an additional l^2 penalty term, determined by $\frac{\rho_2}{2}\left\|\theta\right\|_2^2$, where ρ_2 is a regularization parameter. The objective function is then a smooth convex function to be minimized over an unconstrained parameter space. Besides l^2, l^1 penalty calculated by $\rho_1\left\|\theta\right\|_1$ can also be exploited to perform the feature selection. It plays the role of controlling the amount of regularization as well as the number of extracted features. The combination of them helps to decrease the number of nonzero coefficients and to avoid the numerical problems which can appear in a huge dimensional parameter environment. The objective function corresponding to this combination will be $l(\theta)+\rho_1\left\|\theta\right\|_1+\frac{\rho_2}{2}\left\|\theta\right\|_2^2$.

Several optimization and regularization methods have been proposed to alleviate the parameter estimation issue. The most dominant algorithms among them are provided in WAPITI (Lavergne et al., 2010) – the CRF based labeling toolkit which we employed to combine our features, including: quasi-Newton (L-BFGS and OWL-QN), resilient propagation (R-PROP), stochastic gradient descent (SGD-L1), block-wise coordinate descent (BCD). We investigate the stochastic gradient descent to optimize our feature weights. In the labeling phase, we set the iterations for threshold from 0.3 to 1, with step of 0.025. In each loop, if the probability P("Good"|w) is greater than or equal this threshold, the corresponding word w will be tagged as "Good", and otherwise "Bad". This allows us to obtain a performance curve.

4 Exploitation of Various Kinds of Features

We explore three kinds of features, including:

4.1 System-based Features

They are the features extracted directly from our baseline SMT system based on Moses decoder options stated in Section 5.1, without the participation of any additional element. Based on the resource where features are found, they can be sub-categorized as following:

4.1.1 Target Side Features

We take into account the information in the MT output words, including:

- The word itself.
- The bi-gram sequences formed between current word and its precedence ($i-1 / i$) or successor ($i / i+1$).
- The trigram sequences formed between current word and its two precedent and two following words (including: $i-2 / i-1 / i$; $i-1 / i / i+1$; $i / i+1 / i+2$).

4.1.2 Source Side Features

Using the alignment information between each target and source sentence, we can easily track the source words which the target word is aligned to. Unlike IBM Model-1 (Brown et al., 1993a) which supposes that each target word can be aligned to at most one source word; we process also the situation in which a phrase in the source sentence translates as a single word in the target sentence. To facilitate the alignment representation, we applied the BIO[1] format. In case of multiple target words aligned with one source word, the first word's alignment information will be prefixed with symbols "B-" (means Begin); and "I-" (means Inside) will be added at the beginning of alignment information for each of the remaining ones. With the target words which are not aligned with any source word, alignment information will be represented as O.

Target words (MT output)	Source aligned words		Target words (MT output)	Source aligned words
The	B-le		to	B-de
public	B-public		look	B-tourner
will	B-aura		again	B-à|nouveau
soon	B-bientôt		at	I-à
have	I-aura		its	B-son

[1] See more at: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

the	B-l'		attention	B-attention
opportunity	B-occasion		.	B-.

TABLE 1 – Example of using BIO format to represent alignment information between source sentence and MT hypothesis.

Table 1 shows an example for this representation: since two target words "*will*" and "*have*" are aligned to "*aura*" in source sentence, the alignment information for them will be "*B-aura*" and "*I-aura*" respectively. In case a target word has multiple aligned source words (such as "*again*"), we separate these partners by symbol "|" after putting the prefix "B-" at the beginning.

4.1.3 Alignment Context Features

These features are proposed by (Bach et al., 2011) in regard with the intuition that collocation is a believable indicator for judging if a target word is generated by a particular source word. We also apply them in our experiments, containing:

- *Source alignment context features*: they are the patterns built from each target word and the surroundings of its source word. More precisely, we combine it with one word in the left (left source feature) or in the right (right source feature) of source word.
- *Target alignment context features*: similarly, we anchor the source word with all surroundings of the current target word. Since the window of size ± 2 is employed, it is obvious that 4 combinations will be generated.

4.1.4 Word Posterior Probability

As stated before, WPP has been proven to be one of the most prominent system-based features for confidence estimation. This is the likelihood of the word occurring in the target sentence, given the source sentence. Numerous knowledge sources have been proposed to calculate this value, such as word graphs, N-best lists, statistical word or phrase lexical. The key point here is to determine sentences in N-best lists that contain the word *e* under consideration in a fixed position *i*.

Let $p(f_1^J, e_1^I)$ be the joint probability of source sentence f_1^J and target sentence e_1^I. The word posterior probability of e occurring in position i is computed by aggregating probabilities of all sentences containing e in this position:

$$p_i(e \mid f_1^J) = \frac{p_i(e, f_1^J)}{\sum_{e'} p_i(e', f_1^J)} \tag{5}$$

where

$$p_i(e, f_1^J) = \sum_{I, e_1^I} \theta(e_i, e).p(f_1^J, e_1^I) \tag{6}$$

Here $\theta(.,.)$ is the Kronecker function. The normalization in equation (5) is

$$\sum_{e'} p_i(e', f_1^J) = \sum_{I, e_1^I} p(f_1^J, e_1^I) = p(f_1^J) \tag{7}$$

In this work, we investigate the word graph that represents MT hypotheses (Ueffing, Och, and Ney 2002; Zens and Ney 2005). Thanks to this graph, the posterior probability of word e in position i can be calculated by summing up the probabilities of all paths that contains an edge annotated with e in position i of the target sentence. We perform this summation by applying the forward-backward algorithm (Jurafsky and Martin, 2000). This algorithm also determines the total probability mass needed for normalization, as shown in equation (7).

4.1.5 Target and Source Language Model Based Features

Applying SRILM toolkit (Stolcke, 2002) with the bilingual corpus, we build the 4-gram language model for both target and source side. These language models permit to identify the n-gram (length of the longest sequence created by the current token and its previous ones in the language model) of each word in MT output as well as in the source sentence. For example, with the current token w_i: if the sequence $w_{i-2}w_{i-1}w_i$ appears in the language model but the sequence $w_{i-3}w_{i-2}w_{i-1}w_i$ does not, the n-gram value for w_i will be 3. The value set for each word hence ranges from 0 to 4. Similarly, we extract the n-gram value for the source word aligned to w_i as one more feature.

4.2 Lexical Features

One of the most prominent lexical features that have been widely explored in Confidence Estimation researches is Word's Part-Of-Speech (POS). This tag is assigned to each word in a sentence due to its syntactic and morphological behaviors to indicate its lexical category. In our work, we chose TreeTagger[1] tool for POS annotation task in both source and target sides.

We implement these following lexical characteristics:

- POS of current target word.
- Sequence of POS of all source words which this target word is aligned to, represented in BIO format like alignment representation mentioned in Section 4.1.2.
- Besides using POS of each word in the target side as one lexical feature, we also observed a window of size n (n = 2 and n = 3) over the neighboring target positions and build the n-gram sequence for POS. More specifically, with n = 2 we get the POS sequences $i-1, i$ and $i, i+1$; with n = 3 we have 3 sequences: $i-2, i-1, i$; $i-1, i, i+1$ and $i, i+1, i+2$.

4.3 Syntactic Features

Besides lexical features, the syntactic information of word in a sentence is also a potential clue for predicting its correctness. The intuition behind this is that if a word has grammatical relations with the others, it will be more likely to be correct than a word which has no relation. In order to obtain the links between words, we select the

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Link Grammar Parser[1] as our syntactic parser, allowing us to assign to each MT hypothesis a syntactic structure in which all pairs of words related together will be connected by a labeled link. In case of Link Grammar fails to find the full linkage for the whole sentence, it will skip each word one time until the sub-linkage for the remaining words has been successfully constructed. Based on this structure, we extract the following characteristics to build features:

- The Null Link property: does this word have link with the others or not?
- The total number of links this word has.

Another benefit yielded by Link Grammar Parser is the "constituent" tree (Penn tree-bank style phrase tree) to represent a sentence's grammatical structure (showing noun phrases, verb phrases, etc.). This constituent tree enables us to produce more syntactic features for word, including:

- Its constituent label.
- Its depth in the tree (or the distance between it and the tree root).

Figure 1 represents the syntactic structure as well as the constituent tree for a MT output: *"The government in Serbia has been trying to convince the West to defer the decision until by mid 2007."*.

```
linkparser> The government in Serbia has been trying to convince the West to defer the decision until by mid 2007 .
No complete linkages found.
Found 2368 linkages (1000 of 1000 random linkages had no P.P. violations) at null count 2
       Linkage 1, cost vector = (UNUSED=2 DIS=3 FAT=0 AND=0 LEN=28)

    +-------------------------------------------------------------------Xp------------------------------------------------------------------+
    |                                                                                                  +------------MVp-----------+            |
    +-------Wd------+----------Ss---------+                          +-----Os----+                   +------Os-----+              |            |
    |      +---Dmu---+---Mp---+--Js-+      +--PPf-+--Pg*b-+--TO--+---I---+      +-DG-+-TO-+--I--+     +---Ds--+              +----JT---+  +--RW--+
    |      |         |        |     |      |      |       |       |       |      |    |    |     |     |     |              |         |  |     |
LEFT-WALL the government.n-u in Serbia.l has.v been.v trying.v to.r convince.v the West to.r defer.v the decision.n [until] by [mid] 2007 . RIGHT-WALL

[S [NP [NP The government NP] [PP in [NP Serbia NP] PP] NP] [VP has [VP been [VP trying [S [VP to [VP convince [NP the West [S [VP to [VP defer [NP the decis
ion NP] until [PP by mid [NP 2007 NP] PP] VP] VP] S] NP] VP] VP] S] VP] VP] VP] . S]
```

FIGURE 1 - Example of parsing result generated by Link Grammar

It is intuitive to observe in the graphic representation that the words in brackets (including *"until"* and *"mid"*) have no link with the others, meanwhile the remaining ones have. For instance, the word *"trying"* is connected with *"to"* by the link "TO" and with *"been"* by the link "Pg*b". Hence the "Null Link" and "Total Number of Links" for the word *"mid"* are true, 0 and for the word *"trying"* are false, 2 respectively. The figure also brings us the constituent label and the distance to the root of each word. In case of *"government"*, these values are NP and 2, respectively.

5 Experimental Settings

5.1 French – English SMT System Construction

Our baseline French – English SMT system was constructed using the Moses toolkit (Koehn et al., 2007). This toolkit contains all of necessary components to train the

[1] http://www.link.cs.cmu.edu/link/

translation model. In our work, we kept the Moses's default setting: log-linear model with 14 weighted feature functions. To train the translation model, we used the Europarl and News parallel corpora that are used for WMT[1] evaluation campaign in 2010 (total 1,638,440 sentences). Our target language model is a standard n-gram language model trained using the SRI language modeling toolkit (Stocke, 2002) on the news monolingual corpus (48,653,884 sentences). More details on this baseline system can be found in (Potet et al., 2010).

Besides this, in the decoder phase, we also called some extended options of Moses for tracking both source and target sides information which is mandatory to build our system-based features later. The most pivotal options are listed in the Table 2.

Option name	Function
-print-alignment-info-in-n-best	Display source-to-target and target-to-source word-to-word alignments into the N-best list.
-n-best-list FILE SIZE [distinct]	Generate an n-best file of up to SIZE distinct sentences into file FILE.

TABLE 2 – Moses options employed for tracking alignment information and N-best lists.

5.2 Corpus Preparation

We use our above SMT System to generate the translation hypothesis for 10,881 source sentences taken from several news corpora of the WMT evaluation campaign (from 2006 to 2010). A post-edition task was implemented by using a crowd sourcing platform: Amazon's Mechanical Turk (MTurk), which allows a "requester" to propose a paid or unpaid work and a "worker" to perform the proposed tasks. To avoid the huge gaps between the hypothesis and its post-edition since the correctors can paraphrase or reorder words to form the smoother translation, we highly recommended them to keep the number of edit operations as low as possible, but still ensure the accuracy of this translation with the French sentence. A sub-set (containing 311 sentences) of these collected post-editions was evaluated by a former professional post-editor. Testing result showed that 87.1% of post-editions improve the hypothesis. Detailed description for the corpus construction can be found in (Potet et al., 2012). Finally we extracted randomly 10,000 sentences triples (including source sentence, translation hypothesis and post-edited hypothesis) to form the training set, and keep the remaining 881 sentence triples for the test set.

5.3 Word Label Setting Using TERp-A

In order to obtain the training labels for each word in the MT outputs, previous works have made several attempts. (Xiong et al., 2010) exploited the Levenshtein alignment between the hypothesis and its best reference translation for classifying a word as correct or incorrect. In another method, the Translation Error Rate (TER) alignment

[1] http://www.statmt.org/wmt10/

was performed by (Bach et al., 2011), yielding one of the following labels for each word: good, insertion, substitution and shift. Nevertheless these above studies expressed some drawbacks. The hypothesis and its reference may differ in word order even when they have close meaning. Levenshtein alignment may not be able to align shifted words; hence it leads the inaccurate classification results. TER can be considered as a better alignment tool as it overcomes the first approach by enabling the block movement of words in the MT hypothesis and treating it equally with the other edit operations in term of cost edit, however the exact matches quality still remains limited since it lacks some crucial linguistic edit operations, and its edit costs are not well correlated with various type of human judgments. In order to propose a better word label tagging, we utilize the TER-Plus[1] (or TERp) toolkit. TERp is an extension of TER, not only inheriting the success of this evaluation metric and alignment tool, but also eliminating its shortcomings by taking into account the linguistic edit operations, such as Stem matches, Synonyms matches and Phrase Substitutions besides the TER's conventional ones (Exact match, Insertion, Deletion, Substitution and Shift). These additions allow us to avoid categorizing the hypothesis word as Insertion or Substitution in case that it shares same stem, or belongs to the same synonym set represented by WordNet, or is the paraphrase of word in the reference. For our word label tagging task, we opted TERp-A, another version of TERp, in which each above-mentioned edit cost has been tuned to maximize the correlation with human judgment of Adequacy at the segment level (from the NIST Metrics MATR 2008 Challenge development data). Figure 3 illustrates the labels generated by TERp-A for one hypothesis and reference (post-edited sentence) pair.

Reference	The	consequence	of	the	fundamentalist	movement		also	has	its importance	.
		S			S	Y	I		D	P	
Hyp After Shift	The	result	of	the	hard-line	trend	is	also		important	.

FIGURE 3 – Example of training label task using TERp-A.

Each word or phrase in the hypothesis is aligned to a word or phrase in the reference with a type of edit: “I” for insertions, “S” for substitutions, “T” for stem matches, “Y” for synonym matches, “P” for phrasal substitutions, and “D” for deletions. We do not consider words marked with “D” since they appear only in the reference. The lack of a symbol indicates an exact match (we replace it with “E” thereafter). Since our objective in this work is to train a binary classifier, we re-categorize the obtained 6-label set into binary set: The E, T and Y are regrouped into Good category, whereas the S, P and I belong to the Bad category. Finally, we observed that out of total words (in both of training and test sets) are 85% labeled “G”, 15% labeled “B”.

5.4 Classifier Selection

Among the various CRF toolkits, we selected WAPITI to train our CRF model as well as to tag the binary label for each word in the test set. WAPITI – developed by LIMSI-

[1] http://www.umiacs.umd.edu/~snover/terp/doc v1.html

CNRS - is based on maxent, maximum entropy Markov and linear-chain CRF models. It is well suited for huge feature sets up to several billions and allows us to gain significantly in training time.

The training phase was conducted on our 10000 sentence set. In all experiments with different feature sets, we applied uniquely the Stochastic Gradient Descent (SGD) algorithm for L1-regularized model, which works by computing the gradient only on a single sequence at a time and making a small step in this direction, therefore it can quickly reach an acceptable solution for the model. In the train command, we set values for maximum number of iterations done by the algorithm (-maxiter), stop window size (--stopwin) and stop epsilon (--stopeps) to 200, 6, and 0.00005 respectively. We compared our binary classifier performance not only with the other ones, but also with two naive baselines that were previously created. In baseline 1, we labeled all words in the MT hypothesis as good translations. In baseline 2, we assigned them randomly into G or B with respect to the percentage between two labels like in the corpus (85% G, 15% B).

6 Experiments and Results

6.1 Evaluation Metrics

We evaluated the performance of our classifiers by using very common evaluation metrics: Precision, Recall and F-score. Suppose that we would like to calculate these values for label “B”. Let X be the number of words whose true label is B and have been tagged with this label by the classifier, Y is the total number of words classified as B, and Z is the total number of words which true label is B. Thanks to these concepts, Precision, Recall and F-score can be defined as follow:

$$\Pr = \frac{X}{Y} \quad Rc = \frac{X}{Z} \quad F = \frac{2 \times \Pr \times Rc}{\Pr + Rc} \tag{8}$$

These calculations can be applied in the same way for label “G”. It is straightforward to recognize that the higher precision is, the more precise our classification result will be. Meanwhile, the recall reflects our classifier's capability to retrieve the accurate label for words. F-score is the “harmonic balance” between the two.

6.2 Results and Analysis

We perform our preliminary experiment by training a unique classifier with the combination of all proposed features (21 features). The training algorithm and related parameters were discussed in Section 5.4. The values of precision and recall for “Good” and “Bad” label are tracked and their fluctuations corresponding to thresholds (from 0.3 to 1.0, step 0.025) are represented in Figure 3. Results indicate that in case of Bad label, recall increases nearly monotonously when threshold is enlarged incrementally (except the huge fluctuation from 0.58 to 1 when threshold reaches 1), whereas precision falls from 0.42 to 0.18. With Good label, the variation occurs in the opposite direction: recall drops almost regularly from 0.92 to 0.78, then falls down to 0 in the final iteration, meanwhile precision goes up marginally from 0.848 to 0.881.

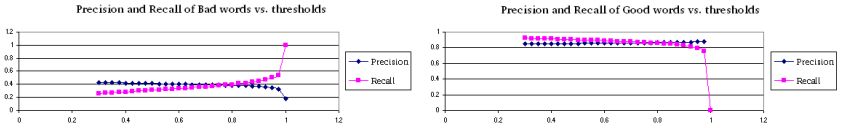


FIGURE 3 – Precision and Recall of labels vs. thresholds.

The curves representing the relationship between precision and recall of each class can be observed in Figure 4.

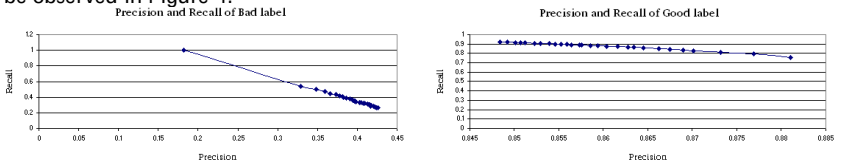


FIGURE 4 – Relationship between Precision and Recall of each label.

Table 3 reports the average values of Precision, Recall and F-score of these labels in the "all-features" system and the baseline systems. Results observed suggest that: (1) Good label is much better predicted than Bad label, (2) The combination of features helped to improve significantly the classifier's capacity to detect the translation errors (which the improvement of 28.55% in terms of F score for B label comparing with baseline 2).

System	Label	Pr(%)	Rc(%)	F(%)
All features	Good	86.01	87.47	86.66
	Bad	38.81	38.05	36.78
Baseline 1	Good	100.00	94.48	97.14
	Bad	0.00	-	-
Baseline 2	Good	85.23	94.47	89.61
	Bad	15.08	5.66	8.23

TABLE 3 – Average Precision, Recall and F-score for labels.

Compare to the result of (Bach et al., 2011), our F-score for G label is 11.16% better, however they outperform us in F-score for B label (27.02% higher). According to our analysis, this might be originated from the following reasons: (1) our training and testing corpus are much smaller than theirs (10.8K vs. 75K) and differ about language pairs, (2) in our corpus, the percentage of G words overwhelms B words (85% vs. 15%) and (3) the best combination of features has not been investigated yet in this paper. All of these issues will be further considered in our future work.

7 Conclusions and Perspectives

We presented an approach to confidence estimation at word level for machine translation which explores various kinds of features, including those from the MT system together with those related to lexical and syntactic function of word in a sentence. A CRF based model has been investigated to train these above features and form our binary classifier. Experimental results show that precision and recall obtained in Good label are very promising, and can be acceptable in Bad label. More meaningful scores are hopefully still ahead with a deeper investigation in each separated feature as well as their various combinations. The comparison with baselines system demonstrates enormous contributions of features towards the perfectibility of the classifier. We employed TERp-A toolkit to generate word labels which is better correlated to human judgment, then regrouped them to a binary set.

In future, this work can be extended in the following ways. Firstly, we plan to conduct the "feature selection" strategy to sort our feature set in the ascending order of their usefulness. From this result we will have better understanding about each feature and its combination with others, as well as eliminate those who are not interesting. Besides of this, we will investigate another type of feature named semantic feature based on some other knowledge resources like WordNet which hopefully can help to improve our state-of-the-art classifier's performance in terms of F-score, especially for Bad Label set. Another task will also be focused on is to find the most optimized methodology to conclude the confidence of whole sentence relied partially on the word-level confidence obtained from this current work.

References

ALBERTO SANCHIS, ALFONS JUAN, and ENRIQUE VIDAL (2007). Estimation of confidence measures for machine translation. In *Proceedings of the MT Summit XI*, Copenhagen, Denmark.

BONNIE DORR MATTHEW SNOVER, NITIN MADNANI and RICHARD SCHWARTZ (2008). TERp system description. In *MetricsMATR workshop at AMTA*.

DEYI XIONG, MIN ZHANG AND HAIZHOU LI (2010). Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th ACL*, pages 604–611, Uppsala, Sweden, July. Association for Computational Linguistics.

JESUS GIMENEZ and LLUIS MARQUEZ (2010b). Linguistic Features for Automatic MT Evaluation. *To Appear in Machine Translation.*

JOHN BLATZ, ERIN FITZGERALD, GEORGE FOSTER, SIMONA GANDRABUR, CYRIL GOUTTE, ALEX KULESZA, ALBERTO SANCHIS and NICOLA UEFFING (2004). Confidence estimation for machine translation. In *The JHU Workshop Final Report*, Baltimore, Maryland, USA, April.

J. LAFFERTY, A. MCCALLUM, and F. PEREIRA (2001). Conditional random fields: Probabilistic models for segmenting et labeling sequence data. In *Proc. ICML*.

LUCIA SPECIA, MARCO TURCHI, NICOLA CANCEDDA, MARC DYMETMAN, and NELLO CRISTIANINI (2009). Estimating the Sentence-Level Quality of Machine

Translation Systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona.

LUCIA SPECIA, ZHUORAN WANG, MARCO TURCHI, JOHN SHAWETAYLOR, and CRAIG SAUNDERS (2009). Improving the confidence of machine translation quality estimates. *In Proceedings of the MT Summit XII*, Ottawa, Canada.

NGUYEN BACH, FEI HUANG and YASER AL-ONAIZAN (2011). Goodness: A method for measuring Machine Translation Confidence. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 211–219, Portland, Oregon, June.

NICOLA UEFFING and HERMANN NEY (2007). Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

NICOLA UEFFING and HERMANN NEY (2005). Word-level confidence estimation for machine translation using Phrased-based translation models. *Proceedings HLT/EMNLP*, pages 763–770, Vancouver.

P. KOEHN, H. HOANG, A. BIRCH, C. CALLISON-BURCH, M. FEDERICO, N. BERTOLDI, B. COWAN, W. SHEN, C. MORAN, R. ZENS, et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL'07*, pages 177–180, Prague, Czech Republic, June.

PETER F. BROWN, STEPHEN A. DELLA PIETRA, VINCENT J. DELLA PIETRA and ROBERT L. MERCER (1993a). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

POTET MARION, EMMANUELLE ESPERANÇA-RODIER, LAURENT BESACIER and HERVE BLANCHON (2012). Collection of a Large Database of French-English SMT Output Corrections. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, May.

POTET MARION, LAURENT BESACIER and HERVÉ BLANCHON (2010). The LIG machine translation system for WMT 2010. In *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*, ACL Workshop. Uppsala, Sweden. 11-17 July.

RADU SORICUT and ABDESSAMAD ECHIHABI (2010). Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th ACL,* pages 612–621, Uppsala, Sweden, July. Association for Computational Linguistics.

S. GANDRABUR and G. FOSTER (2003). Confidence estimation for text prediction. *In Proceedings of CoNLL*, Edmonton, May.

STOLCKE, A. (2002), SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, Denver, USA, pp. 901-904.

SYLVAIN RAYBAUD, CAROLINE LAVECCHIA, DAVID LANGLOIS and KAMEL SMAILI (2009). Error detection for statistical machine translation using linguistic features. In *Proceedings of the 13th EAMT*, Barcelona, Spain, May.

THOMAS LAVERGNE, OLIVIER CAPPE and FRANÇOIS YVON (2010). Practical very large scale CRFs. In *Proceedings ACL*, pages 504–513.

Système de prédiction de néologismes formels : le cas des N suffixés par –IER dénotant des artefacts

Aurélie Merlo[1]
(1) STL UMR 8163, rue du Barreau BP 60149 59653 Villeneuve d'Ascq Cedex
aurelie.merlo@etu.univ-lille3.fr

RESUME

Nous présentons ici un système de prédiction de néologismes formels avec pour exemple la génération automatique de néologismes nominaux suffixés par –IER dénotant des artefacts (*saladier, brassière, thonier*). L'objectif de cet article est double. Il s'agira (i) de mettre en évidence les contraintes de la suffixation par –IER afin de les implémenter dans un système de génération morphologique puis (ii) de montrer qu'il est possible de prédire les néologismes formels. Ce système de prédiction permettrait ainsi de compléter automatiquement les lexiques pour le Traitement Automatique des Langues (TAL).

ABSTRACT

Prediction Device of Formal Neologisms : the Case of –IER Suffixed Nouns Denoting Artifacts

We'll introduce here a device that can predict neologisms using an example the automatical generation of nominal neologisms suffixed by –IER denoting artifacts (*saladier, brassière, thonier*). The aim of this article is double. We will first address the –IER suffixation constraints in order to take them into account on the implementation of our morphological generator. Second, we will describe our method to predict formal neologisms. Such a method will permit to automatically enrich NLP lexicons.

MOTS-CLES : morphologie constructionnelle, néologie, génération morphologique, incomplétude lexicale.
KEYWORDS : constructional morphology, neology, morphological generation, lexical incompleteness.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 57–70,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Certaines applications de Traitement Automatique des Langues (TAL) comme les traducteurs automatiques utilisent des lexiques (Sproat, 1992). Cependant, ces lexiques comme les dictionnaires (Sablayrolles, 2000 ; Sablayrolles, 2008) sont lacunaires dans la mesure où ils ne peuvent pas contenir l'ensemble des unités lexicales d'une langue. L'absence d'une unité lexicale dans un lexique de TAL peut alors poser problème. Ce problème, nommé dans la littérature scientifique l'incomplétude lexicale, a déjà été abordé dans de nombreux travaux. Certains de ces travaux proposent une typologie des mots inconnus (Dister & Fairon, 2004 ; Maurel, 2004 ; Cartoni, 2006 ; Blancafort & *al.*, 2010). D'autres travaux présentent des solutions pour palier cette incomplétude lexicale. Les travaux de (Dister & Fairon, 2004) soumettent la solution d'un repérage des mots inconnus dans un corpus québécois grâce au système GlossaNet. (Cartoni, 2006) propose l'implémentation de règles de construction afin d'analyser les mots construits [1]. (Blancafort & *al.,* 2010) proposent quant à eux un enrichissement dynamique d'un lexique TAL. Cela passe d'abord par une annotation en corpus des tokens inconnus pour une classification automatique. Puis, une validation manuelle est alors nécessaire pour décider de l'ajout temporaire ou définitif dans le lexique.

Dans le cadre de cet article, nous proposons une approche pour palier en partie l'incomplétude lexicale des lexiques de TAL. Dans la mesure où les néologismes formels ont une large place dans les mots inconnus (Maurel, 2004 ; Cartoni, 2006), nous proposons de les prédire afin de les intégrer dans les lexiques de TAL. Notre hypothèse est que les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel sont prédictibles. Mais seront-elles pour autant attestées dans l'usage ? Afin de vérifier notre hypothèse, nous avons élaboré un générateur automatique de néologismes formels à base de contraintes linguistiques. Nous avons choisi de générer automatiquement des néologismes formels nominaux suffixés par –IER[2] dénotant des artefacts (*saladier, brassière, thonier*). Le choix de cette suffixation s'explique par sa diversité référentielle (Corbin & Corbin, 1991).

Les objectifs de cet article sont alors (i) de mettre en évidence les contraintes linguistiques de la suffixation par –IER et (ii) de montrer qu'il est possible de prédire les néologismes formels.

Nous commencerons par un état de l'art des générateurs morphologiques existants afin de mettre en évidence l'originalité de notre système de génération automatique de néologismes formels. Puis, nous procèderons à l'analyse de la suffixation par –IER afin de mettre en évidence ses contraintes linguistiques qui seront par la suite implémentées au sein de notre système de génération morphologique. Nous commenterons les résultats obtenus afin de montrer les avantages et les limites de notre approche. Enfin, nous terminerons cet article sur les perspectives de recherche en morphologie constructionnelle que permet cette première approche de la prédiction des néologismes formels.

[1] Nous pensons ici à DériF (Namer, 2009).
[2] Nous avons adopté cette convention d'écriture afin de rassembler sous –IER les suffixes allomorphiques –*ier(e)* et –*er(e)*.

2 État de l'art des générateurs morphologiques

Il existe actuellement des systèmes de génération morphologique permettant de générer automatiquement des formes fléchies et/ou des formes dérivées.

Le premier de ces systèmes que nous présentons brièvement est le système PILAF (Procédures Interactives Linguistiques Appliquées au Français) (Courtin & *al.*, 1994). Le système PILAF a été élaboré dans le laboratoire CLIPS afin de réaliser des tâches d'analyse et de génération morphologique. Le système est composé d'une grammaire, d'un dictionnaire et de deux moteurs : l'un est consacré à l'analyse, l'autre à la génération. Ce système est capable de reconnaître et de générer 250 000 formes du français à l'aide d'un lexique comportant 35 000 entrées. Ce système, destiné à la morphologie flexionnelle, utilise un formalisme qui « permet le codage de règles de la morphologie dérivationnelle » (Courtin & *al.,* 1994 : 101).

Le « système dérivationnel » de (Tzoukermann & Jacquemin, 1997) est élaboré également pour fonctionner aussi bien en analyse qu'en génération. En génération, le système repose sur le principe de la concaténation et pour éviter la surgénération, (Tzoukermann & Jacquemin, 1997) proposent trois niveaux de filtrage : un filtrage lexical (les dérivés sont générés à partir de bases attestées dans un dictionnaire), un filtrage d'attestation en corpus et un filtrage sémantique collocatif (les dérivés sélectionnés doivent apparaître en contexte en collocation avec un même mot).

Le système Intex (Silverztein, 1993), dont la dernière version se nomme NooJ, est un environnement dans lequel il est possible d'élaborer des descriptions formalisées applicables sur corpus. Cet environnement comprend des dictionnaires électroniques (DELAF) et offre la possibilité de réaliser ses propres grammaires sous forme de graphes. Le système Intex est utilisé en traitement automatique de corpus mais également pour de la génération flexionnelle et/ou dérivationnelle.

Le système GéDériF est le pendant de DériF (Namer, 2009). C'est un système de génération automatique d'unités lexicales construites qui a une triple fonction : (i) « produire un lexique d'unités lexicales construites absentes des dictionnaires », (ii) « enrichir ce lexique d'informations constructionnelles et sémantiques, (iii) « constituer des micro-familles constructionnelles » (Dal & Namer, 2000). Les formes générées sont validées par une recherche en ligne. L'inconvénient est que « le générateur ne peut donc choisir la base des unités qu'il va générer automatiquement que parmi les mots déjà construits » (Dal & Namer, 2000) et analysés par DériF.

La présentation de ces systèmes de génération morphologique permet de mettre en évidence une limite commune. Ces systèmes offrent la possibilité de générer des mots dérivés et implémentent un certain nombre de règles dérivationnelles. Or, ces systèmes ne prennent pas en compte les contraintes sémantiques pouvant peser dans un processus dérivationnel.

À présent, dans le cadre de l'analyse de la suffixation par –IER, nous allons montrer en quoi notre système de génération morphologique est original.

3 Analyse de la suffixation par -IER

3.1 Méthodologie

Nous allons mettre en évidence les contraintes linguistiques pesant sur la suffixation par –IER. Nous avons choisi cette suffixation pour sa diversité référentielle (Corbin & Corbin, 1991). Dans le cadre de cet article, nous étudierons de plus près les dérivés nominaux suffixés par –IER dénotant des artefacts (*saladier, brassière, thonier*). Nous précisons que les résultats présentés sont extraits d'une étude plus large de la suffixation par –IER que nous avons réalisée dans le cadre d'un mémoire de recherche (Merlo, 2011).

Le corpus élaboré dans le cadre de ce mémoire provient de l'extraction du *TLF* d'une liste de candidats à la suffixation par –IER. Cette liste a été nettoyée afin de ne retenir que les lexèmes construits par la suffixation par –IER (élimination des emprunts par exemple du type *manager*). Nous avons par la suite sélectionné 530 lexèmes construits que nous avons analysés afin de faire apparaître les informations nécessaires à la prédiction de néologismes suffixés par –IER. Par conséquent, nous avons procédé à une analyse morpho-sémantique, morphophonologique et graphique de la suffixation par –IER.

Pour cet article ne retenant que les dérivés dénotant des artefacts, l'étude portera sur 119 lexèmes construits et 132 acceptions[3].

3.2 Analyse morpho-sémantique

3. 2. 1. Approche référentielle

L'analyse morpho-sémantique des dérivés suffixés par –IER a consisté en l'étude de la référence des dérivés et de leur base. Nous nous sommes inspiré de l'approche par classe référentielle proposée dans (Roché, 1998)[4]. Par convention, les classes référentielles seront entre crochets ([récipient] par exemple).

À partir des définitions du *TLF*, nous avons déterminé la classe référentielle des acceptions des dérivés de notre corpus d'étude. Notre analyse morpho-sémantique a pu confirmer l'existence de sept classes référentielles pour les dérivés suffixés par –IER dénotant des artefacts (*cf.* FIGURE 1 ci-dessous) que (Roché, 1998) avait mis en évidence.

[3] Nous avons choisi de prendre en compte les acceptions lorsque celles-ci n'étaient pas issues d'une dérivation sémantique. Ex : *médailler* (*s. v. médailler* dans le *TLF*) désigne à la fois un meuble contenant des médailles et un recueil de médailles.

[4] Une classe référentielle est définie comme la projection de propriétés sémantiques (Temple, 1996).

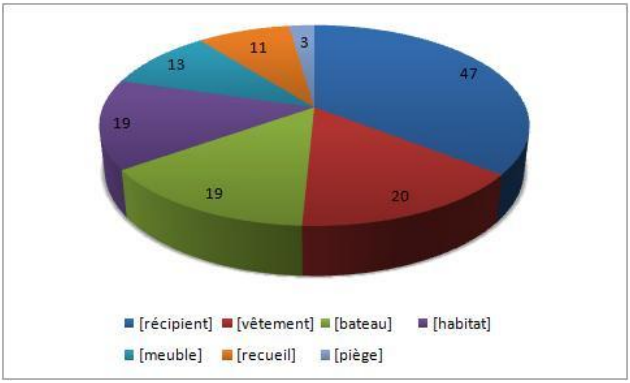


FIGURE 1 – Classe référentielle des dérivés suffixés par –IER dénotant des artefacts (contrainte n°1)

Du côté de la sortie, nous savons à présent que les dérivés suffixés par –IER dénotant des artefacts peuvent avoir pour référent un nom de récipient, un nom de vêtement, un nom de bateau, un nom d'habitat ou un nom de meuble. Néanmoins, il nous reste à déterminer si la suffixation par –IER sélectionne les bases en fonction d'une contrainte référentielle. Toujours à partir des définitions du *TLF*, nous avons déterminé la classe référentielle des bases de notre corpus (*cf.* TABLE 1 ci-dessous).

Classe référentielle du dérivé	Classe référentielle de la base	Nombre d'acceptions	Exemples
[bateau]	[poisson]	7	*baleinier*
	[partie du bateau]	5	*boulinier*
	[minéral]	3	*méthanier*
	[personne]	1	*négrier*
	[boisson]	1	*pinardier*
	[fruit]	1	*bananier*
	[rangement]	1	*vraquier*
[habitat]	[animal]	15	*pigeonnier*
	[personne]	2	*garçonnière*

	[minéral]	1	*terrier*
	[procès]	1	*volière*
[meuble]	[objet]	9	*bonnetière*
	[végétal]	3	*grainier*
	[durée]	1	*semainier*
[piège]	[animal]	3	*ratière*
[récipient]	[aliment]	22	*beurrier*
	[objet]	11	*bouquetier*
	[liquide]	8	*tisannière*
	[animal]	3	*turbotière*
	[minéral]	2	*sablier*
	[mode de cuisson]	1	*daubière*
[recueil]	[écrit]	7	*chansonnier*
	[objet]	2	*médaillier*
	[végétal]	2	*herbier*
[vêtement]	[partie du corps]	17	*jambière*
	[animal]	1	*grenouillère*
	[meuble]	1	*tablier*

TABLE 1 – Classe référentielle des bases (contrainte n°2)

Grâce à cette analyse de la classe référentielle des bases, à présent nous savons que pour générer par exemple un nom de recueil, la suffixation par –IER peut sélectionner des bases dont le référent désigne un écrit, un objet ou un végétal. Cette table permet également de mettre en évidence des préférences sémantiques telle que pour former un nom de meuble, la suffixation par –IER sélectionne de préférence une base dont le référent désigne un aliment (22 acceptions). Dans le cadre de la génération automatique de néologismes formels suffixés par –IER, nous ne tiendrons pas compte de ces préférences.

3.2.1 Variation en genre

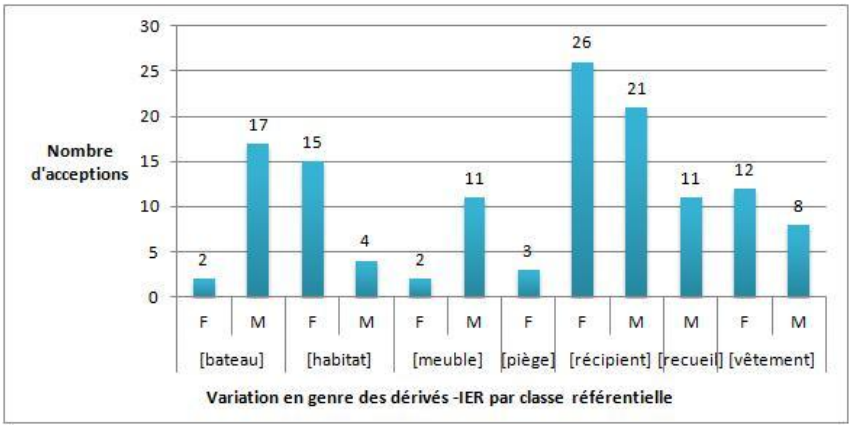


FIGURE 2 – Variation en genre des dérivés suffixés par –IER dénotant des artefacts (contrainte n° 3)

Prédire les néologismes suffixés par –IER dénotant des artefacts nécessite également de déterminer les cas de variation en genre. Selon (Roché, 1998 : 45), l'attribution du genre pour les artefacts s'applique par le biais d'« un déterminé implicite ou d'un terme générique » (*un (bateau) pétrolier*) ou par défaut au masculin lorsqu'il n'y a pas de déterminé implicite. Sur la FIGURE 2 ci-dessous, nous avons répertorié le nombre de formes féminines et masculines par classe référentielle des dérivés.

La FIGURE 2 nous permet de faire des choix quant à la flexion en genre des néologismes suffixés par –IER. En effet, nous voyons nettement se démarquer les formes masculines pour les classes référentielles [bateau], [meuble] et [recueil] et les formes féminines pour les classes référentielles [habitat], [piège], [récipient] et [vêtement]. Cependant, en l'absence de tests formels pour l'étiquetage en classe référentielle, nous admettons que l'attribution du genre ici est sujette à caution.

À présent, il nous reste à déterminer les contraintes formelles de la suffixation par –IER.

3.3 Analyse morphophonologique

Deux ensembles de contraintes pèsent plus particulièrement en morphophonologie : les contraintes de taille (Plénat, 1997) et les contraintes dissimilatives[5].

3.3.1 Contrainte de taille

Il s'agit de découvrir ici si la suffixation par –IER comporte des contraintes de taille sur l'entrée et/ou la sortie de ses règles de construction de lexème. Pour cela, nous avons calculé le nombre de syllabes de chaque base et de chaque dérivé de notre corpus (*cf.*

[5] Les contraintes de dissimilation rejettent la contiguïté de deux phonèmes identiques ou similaires.

TABLE 3 ci-dessous).

Taille de la base	Taille du dérivé	Exemple	Nombre de lexèmes
1	2	*beurrier*	41
1	3	*oeufrier*	16
2	2	*jarretière*	4
2	3	*grenouillère*	61
2	4	*vinaigrier*	1
3	4	*porcelainier*	9

TABLE 3 – Contrainte de taille pour les dérivés suffixés par –IER dénotant des artefacts (contrainte n° 4)

Ces résultats permettent de mettre en évidence que la suffixation par –IER ne sélectionne pas de base comportant plus de trois syllabes pour former des dérivés quadrisyllabiques. Par ailleurs, ces résultats montrent que la suffixation par –IER sélectionne de préférence des bases dissyllabiques pour former des dérivés trissyllabiques (61 lexèmes concernés). Enfin, à travers ces résultats, nous voyons que la suffixation par –IER utilise de préférence la concaténation avec un nombre important de lexèmes construits comportant une syllabe de plus par rapport à leur base.

Par conséquent, notre système de prédiction de néologismes suffixés par –IER dénotant des artefacts devra effectuer un tri formel (en plus d'un tri sur les classes référentielles) parmi les bases afin de ne sélectionner que les bases entre une et trois syllabes.

3.3.2 Contrainte sur l'attaque

Ce que nous appelons une attaque est le phonème placé au début de la syllabe contenant le suffixe –IER. Ce phonème provient en général de la finale consonantique de la base mais certaines attaques sont parfois des consonnes épenthétiques (comme dans *cafetière*), intercalées entre la base et le suffixe pour éviter un hiatus car « d'une façon générale, la dérivation par *–ier(e)* est gênée par une finale vocalique (sans consonne latente) » (Roché, 1998).

Nous avons relevé les attaques des 119 lexèmes construits de notre corpus (*cf.* FIGURE 3 ci-dessous).

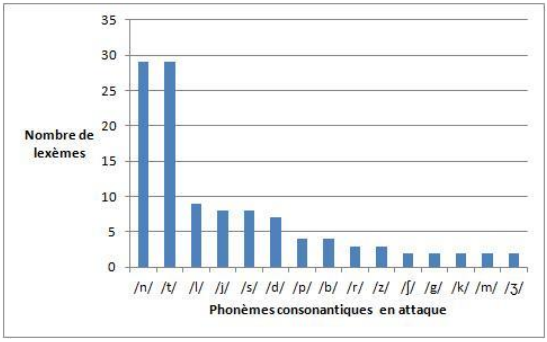


FIGURE 3 – Distribution des phonèmes consonantiques en attaque

Ce graphique montre que les dérivés suffixés par –IER comportant une attaque /n/ ou /t/ sont les plus nombreux. D'une manière générale, la suffixation par –IER tend à sélectionner des bases dont la finale consonantique est une alvéolaire (/t/ et /n/) et au contraire à éviter des bases dont la finale consonantique est un phonème labiodental (/v/ et /f/), bilabial (/p/, /b/ et /m/) ou vélaire (/k/ et /g/). Néanmoins, on ne peut pas parler de contraintes sur l'attaque mais plutôt de préférences ce qui n'est pas le cas lorsque la base se termine par une finale vocalique avec l'apparition systématique d'une consonne épenthétique.

3.4 Analyse graphique

3.4.1 Cas de changements graphiques

L'adjonction du suffixe –IER provoque des modifications graphiques. La première de ces modifications est l'apparition d'une consonne épenthétique entre la finale vocalique de la base et le suffixe –IER (contrainte n°5). Dans le cadre de notre mémoire, nous avons recensé 19 dérivés suffixés par -IER comportant une consonne épenthétique. Parmi eux des noms d'artefacts comme *cafetière* [récipient], *fourmilière* [habitat], *morutier* [bateau] ou encore *tabatière* [récipient]. Nous avons également découvert que le phonème /t/ est particulièrement utilisé comme consonne épenthétique.

Le second cas de modification graphique par adjonction du suffixe –IER est le changement d'accent qui traduit un changement d'aperture pour cause de contrainte dissimilative (contrainte n°6). Ainsi, parmi les noms d'artefacts, les dérivés *chéquier* [recueil] et *négrier* [bateau] ont subi un changement d'accent par rapport à leur base respective *chèque* et *nègre*.

Enfin, la gémination est le dernier cas de modification graphique relevé. La consonne finale de la base à tendance à se doubler avec la concaténation du suffixe –IER. Nous allons détailler ci-dessous les différents cas de figure de gémination.

3.4.2 Règles graphiques de concaténation du suffixe –IER

À partir de l'analyse de notre corpus, nous avons posé huit contraintes graphiques

régissant la concaténation du suffixe –IER :

- Contrainte n°1 : lorsque la base se termine par une consonne muette (*abricot*), il y a concaténation simple (*abricotier*).
- Contrainte n°2 : lorsque la base se termine par un phonème consonantique et par une voyelle muette (*agence*), la voyelle finale est supprimée et le suffixe –IER se concatène au radical de la base (*agencier*).
- Contrainte n°3 : lorsque la base se termine par un phonème vocalique et par une voyelle (*moru*), la consonne épenthétique *–t-* s'ajoute suivie de la concaténation du suffixe –IER (*morutier*).
- Contrainte n°4 : lorsque la base se termine par *–os, -as, -is* ou *–us*, (*matelas*) la consonne *–s-* s'ajoute suivie de la concaténation du suffixe –IER (*matelassier*).
- Contrainte n°5 : lorsque la base se termine par *–on*, la consonne *–n-* s'ajoute suivie de la concaténation du suffixe –IER (*boutonnière*).
- Contrainte n°6 : lorsque la base se termine par *–an* ou *–in*, il y a concaténation simple du suffixe –IER (*rubanier*, *jardinière*).
- Contrainte n°7 : lorsque la base se termine par *–il*, la consonne *–l-* s'ajoute suivie de la concaténation du suffixe –IER (*œillère*).
- Contrainte n°8 : lorsque la base contient dans sa dernière syllabe un accent grave (*trèfle*), cet accent devient un accent aigu dans le dérivé suffixé par –IER (*tréflier*).

En ce qui concerne les règles n°4 et n°5, elles ont été élaborées sur la base de la fréquence des phénomènes étudiés car il existe des contre-exemples à ces règles : *tamis>tamisier, thon>thonier*.

4 Implémentation

Nous avons dégagé les contraintes de la suffixation par –IER déterminant ainsi les connaissances linguistiques nécessaires à la génération automatique de néologismes suffixés par –IER. Cependant, la génération automatique de néologismes formels nécessite également un lexique de référence (ou lexique d'exclusion) et un lexique de bases annoté.

Bien que l'utilisation d'un dictionnaire présente des limites lorsque l'on étudie la néologie (Sablayrolles, 2000 ; Sablayrolles, 2008), nous avons choisi comme lexique de référence le lexique *Morphalou* élaboré à partir de la nomenclature du *TLF*. L'objectif de ce lexique de référence est de retenir uniquement les formes nouvelles générées. Si une forme générée est attestée dans le lexique de référence, celle-ci ne sera pas retenue.

L'élaboration du lexique de bases annoté a été plus complexe. Ce lexique devait comprendre les informations linguistiques nécessaires à l'application des contraintes de la suffixation par –IER. Ainsi, ce lexique devait contenir des indications sur la classe référentielle, le nombre de syllabe et une description graphique de la base afin de déterminer quelle règle graphique à appliquer. Au vu de ces informations, la ressource *Lexique* 3. 1. était une ressource utile en ce qui concerne le nombre de syllabes et la description orthographique (*cf.* TABLE 4 ci-dessous). La description orthographique permet de savoir si la finale de la base est vocalique (*café*) ou consonantique (*cafard*) ce qui a son importance pour l'application des contraintes graphiques n°1 et n°3. La

description phonétique permet de savoir si la base se termine par une finale consonantique et une voyelle (*agence*) ce qui a de l'importance pour l'application de la contrainte n°2. Enfin, la description orthographique des syllabes permet d'appliquer la contrainte n°8 du changement d'accent.

Lemme	Description orthographique	Description phonétique	Nombre de syllabes	Description orthographique des syllabes
café	CVCV	CVCV	2	ca-fé
cafard	CVCVCC	CVCVC	2	ca-fard
agence	VCVCCV	VCVC	2	a-gen-ce

TABLE 4 – Informations contenues dans *Lexique* 3. 1. et utiles au lexique des bases

Nous avons choisi aléatoirement 268 lexèmes de *Lexique* 3. 1. pour constituer notre lexique de bases avec pour seul objectif d'obtenir une homogénéité de classe référentielle. Puis, nous avons procédé à l'annotation des classes référentielles de ces lexèmes-bases. Aucune ressource actuellement en français ne possédant ce type d'information, nous nous sommes appuyés sur les définitions du *TLF* pour établir les classes référentielles.

5 Résultats

Notre hypothèse en introduction était que les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel sont prédictibles. Ici, il s'agissait de mettre en évidence les contraintes de la suffixation par –IER en s'attachant particulièrement à l'étude des noms d'artefacts.

Notre approche a permis de générer 544 nouvelles unités lexicales dont 222 dénotant des artefacts. Afin de vérifier notre hypothèse, nous avons interrogé le Web et déterminé quelles nouvelles unités lexicales étaient attestées. La FIGURE 4 ci-dessous dresse un bilan de cette vérification quantitative pour les néologismes dénotant des artefacts.

Ces résultats, dont notamment le nombre de néologismes formels générés attestés sur le Web (87 au total soit 39,19%), valident en partie notre hypothèse. Il apparaît possible de prédire les nouvelles unités lexicales qui satisfont les contraintes linguistiques liées à un procédé constructionnel. Il apparaît même possible de prédire la référence de ces nouvelles unités lexicales (18 au total parmi les attestés). Néanmoins, il apparaît également que notre système est en surgénération (135 non-attestés au total soit 60,81%).

Autant une attestation relevée sur le Web est sujette à caution (Kilgarriff & Grefenstette, 2003), autant l'absence d'attestation sur le Web est significative.

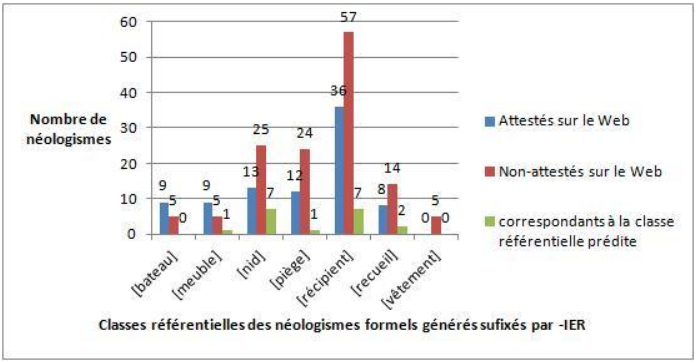


FIGURE 4 – Évaluation quantitative des résultats

Nous sommes face à trois types de résultats. Le premier concerne tous les cas de néologismes formels générés suffixés par –IER ne trouvant aucune attestation sur le Web. Cela provient-il d'une erreur de génération graphique ? Le néologisme généré *antilopière* par exemple n'est pas attesté sur le Web alors qu'il ne présente pas *a priori* d'erreur graphique. Cela provient-il alors d'une erreur de catégorisation de la classe référentielle de la base ? Cette hypothèse est plus probable dans la mesure où nous avons procédé sans tests formels mais à l'aide des définitions du *TLF*. Le second type de résultat conforte d'ailleurs cette hypothèse puisqu'il s'agit des néologismes formels générés suffixés par –IER attestés sur le Web mais pas dans la classe référentielle prédite. C'est ainsi le cas de *cochonnière* prédit dans la classe référentielle [piège] et attesté dans un contexte faisant référence à un véhicule[6]. Enfin, le dernier type de résultat concerne les attestations de néologismes générés suffixés par –IER dans les classes référentielles prédites tels que *choucroutière* prédit dans la classe [récipient]. D'autres hypothèses peuvent être avancées pour expliquer cette surgénération. Nous n'avons pas abordé la question de l'échangisme suffixal (Roché, 1997). La validation de notre hypothèse par le biais de la vérification des formes générées sur le Web est également discutable. Une autre alternative pourrait être une validation auprès d'un panel de locuteurs.

6 Conclusion : sur les pas des connaissances encyclopédiques

En conclusion, nous avons atteint nos deux objectifs qui étaient (i) de mettre en évidence les contraintes liées à la suffixation par –IER et (ii) de montrer qu'il est possible de prédire les néologismes formels par l'implémentation de ces contraintes au sein d'un système de génération morphologique.

Du point de vue applicatif, cette approche apporte des perspectives quant à l'incomplétude lexicale. Sur une thématique précise qui est l'étude des noms suffixés par –IER dénotant des artefacts, nous avons prédit 87 néologismes formels au total pouvant

[6] « René Thoré avec Kapi (mulassier) attelé à une cochonnière à 4 roues » (http://picasaweb.google.com/lh/photo/Pp3EyPQcQH1DaduXJK_9Mw)

directement être intégrés dans un lexique pour le TAL. Cette quantité est très insuffisante mais nous espérons améliorer le taux de prédiction prochainement. Du point de vue théorique, cette approche apporte de nouvelles perspectives de recherche en morphologie constructionnelle. Nous avons prédit la référence de 18 néologismes formels sur 87 attestés sur le Web grâce à une annotation en classe référentielle. Cela implique que notre approche de la prédiction par la notion de classe référentielle mériterait d'être approfondie. Tout d'abord, afin d'éviter l'utilisation du *TLF* pour la catégorisation référentielle des lexèmes, une première solution serait d'élaborer une série de tests formels en s'inspirant de travaux en sémantique-cognitive (notamment Fillmore, 1982 ; Langacker, 1987 ; Kleiber, 1990). Une seconde solution consisterait à utiliser des ontologies existantes telles que WOLF pour le français (Sagot & Fišer, 2008) ou WordNet (Fellbaum, 1998). Puis, à partir de là, il faudrait, par classe référentielle, faire émerger les contraintes extralinguistiques. Dans un travail de thèse que nous venons de débuter, nous tentons de démontrer l'hypothèse de la nécessité des connaissances encyclopédiques en morphologie constructionnelle (Aronoff, 1980 ; Clark & Clark, 1979). L'exemple du néologisme généré *cochonnière* met en lumière cette nouvelle hypothèse car si ce néologisme apparaît dans un contexte désignant un véhicule et non pas un piège c'est parce le cochon est un animal « domestique » que nous n'avons pas besoin de chasser. De la même manière, on relève *bananier* dont le référent désigne un nom de bateau formé sur une base dont le référent désigne un fruit mais pas *mirabellier* car la mirabelle ne se commercialise pas dans le monde entier et n'a donc pas besoin d'un transport en bateau. Les connaissances encyclopédiques permettraient ainsi de faire le tri entre ce qui est ce qui a une pertinence dénominative et ce qui n'en a pas.

Références

ARONOFF, M. (1980). Contextuals, *in Language*, 56, No. 4, pp. 744-758.

BLANCAFORT, S. J. H., RECOURCE, G., COUTO, J., SAGOT, B., STERN, R. et TEYSSOU, D. (2010). Traitement des inconnus : une approche systématique de l'incomplétude lexicale, *in Actes de TALN 2010*, Montréal : Canada.

CARTONI, B. (2006). Constance et variabilité de l'incomplétude lexicale. *Noûs* (3), pp. 10–13.

CLARK, E. V. et CLARK, H. (1979). When Nouns Surface as Verbs, *in Language*, 55, No. 4, pp. 767-811.

CORBIN, D. et CORBIN, P. (1991). Un traitement unifié du suffixe -ier(e), *in Lexique* 10, pp. 61-145.

COURTIN, J., DUJARDIN, D., GENTHIAL, D. et KOWARSKI, I. (1994). Analyse et génération morphologique avec le système PILAF, *in TAL* « Morphologie computationnelle », vol. 35, n° 2.

DAL, G. et NAMER, F. (2000). Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations, *in TAL* 41.

DISTER, A. et FAIRON, C. (2004). Extension des ressources lexicales grâce à un corpus dynamique, *in Lexicometrica*.

FELLBAUM, C. (1998). *WordNet: An Electronic Lexical Database*, Cambridge : MIT Press.

FILLMORE, C. J. (1982). Frame semantics in T. L. S. of Korea (Ed.), Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co.

KILGARRIFF, A et GREFENSTETTE, G. (2003). Introduction to the special issue on the Web as a corpus, *in Computational Linguistics, 29*(3), 333347.

KLEIBER, G. (1990). *La sémantique du prototype. Catégorie et sens lexical*, Paris : PUF.

LANGACKER, R.W. (1987). *Foundations of Cognitive Grammar*. Vol. I: Theoretical Prerequisites, Stanford: Stanford University Press.

MAUREL, D. (2004). Les mots inconnus sont-ils des noms propres ? *in Actes des JADT 2004.*

MERLO, A. (2011). *Élaboration d'un prototype de générateur automatique de néologismes formels : le cas des suffixés par –IER*. Mémoire de recherche en vue de l'obtention du Master professionnel « Lexicographie, Terminographie et Traitement Automatique de Corpus ». Université Charles de Gaulle, Lille 3.

NAMER, F. (2009). *Morphologie, lexique et Traitement Automatique des Langues – Le système DériF : TIC et Sciences cognitives*, London : Hermès Sciences Publishing.

PLENAT, M. (1997). Analyse morpho-phonologique d'un corpus d'adjectifs en –esque, *in Journal of French Language Studies*, 7 : 163–179.

ROCHE, M. (1997). Briard, bougeoir et camionneur : dérivés aberrants, dérivés possibles, *in* Corbin et al., éd. (1997), pp. 241–250.

ROCHE, M. (1998). *Deux études sur la dérivation en -ier(e)*, Toulouse, Carnets de grammaire (Rapports internes de l'ERSS, CNRS et Université de Toulouse-Le Mirail).

SAGOT, B. et FISER, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues, in TALN 2008, Avignon, France.

TZOUKERMANN, E. et JACQUEMIN, C. (1997). Analyse automatique de la morphologie dérivationnelle et filtrage de mots possibles, *in Silexicales* « Mots possibles et mots existants », n° 1, pp. 251 – 260.

SABLAYROLLES, J.-F. (2000). *La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes*, Paris : Honoré Champion.

SABLAYROLLES, J.-F. (2008). Néologie et dictionnaire(s) comme corpus d'exclusion, *in Néologie et terminologie dans les dictionnaires*, douzième Journée des dictionnaires, Université de Cergy-Pontoise, 17 mars 2004, sous la direction de Sablayrolles Jean-François, Paris : H. Champion.

SILBERZTEIN, M. (1993). Dictionnaires électroniques et analyse de texte : le système INTEX, Masson : Paris.

SPROAT, R. (1992). *Morphology and Computation*, Cambridge : MIT Press.

TEMPLE, M. (1996). *Pour une sémantique des mots construits,* Villeneuve d'Asq : Presses Universitaires du Septentrion.

Application d'un algorithme de traduction statistique à la normalisation de textos

Gabriel Bernier-Colborne[1]
(1) Observatoire de linguistique Sens-Texte
Université de Montréal
gabriel.bernier-colborne@umontreal.ca

RÉSUMÉ

Ce travail porte sur l'application d'une technique de traduction statistique au problème de la normalisation de textos. La méthode est basée sur l'algorithme de recherche vorace décrit dans (Langlais *et al.*, 2007). Une première normalisation est générée, puis nous appliquons itérativement une fonction qui génère des nouvelles hypothèses à partir de la normalisation courante, et maximisons une fonction de score. Cette méthode fournit une réduction du taux d'erreurs moyen par phrase de 33 % sur le corpus de test, et une augmentation du score BLEU de plus de 30 %. Nous mettons l'accent sur les fonctions qui génèrent la normalisation initiale et sur les opérations permettant de générer des nouvelles hypothèses.

ABSTRACT

Applying a Statistical Machine Translation Algorithm to SMS Text Message Normalization

We report on the application of a statistical machine translation algorithm to the problem of SMS text message normalization. The technique is based on a greedy search algorithm described in (Langlais *et al.*, 2007). A first normalization is generated, then a function that generates new hypotheses is applied iteratively to a current best guess, while maximizing a scoring function. This method leads to a drop in word error rate of 33% on a held-out test set, and a BLEU score gain of over 30%. We focus on the methods of generating the initial normalization and the operations that allow us to generate new hypotheses.

MOTS-CLÉS : Traduction statistique, normalisation de textos, algorithme de recherche vorace, modèle de langue.

KEYWORDS: Machine translation, SMS, text message, normalization, greedy search algorithm, language model.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 71–79,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les messages textes (SMS ou textos) contiennent fréquemment des formes qui ne sont pas conformes à l'orthographe ordinaire, ce qui rend leur traitement par des systèmes de traitement automatique de la langue problématique. La normalisation des textos consiste à « réécrire les textos au moyen d'une orthographe plus classique afin de les rendre plus facilement lisibles par un humain ou un ordinateur » (Yvon, 2008, p. 5) [1]. Par exemple, si on rencontre la forme « stai comment le ... », l'objectif est de produire une normalisation telle que « comment était le ... ».

Étant donné la popularité énorme des messages textes et des formes de communication apparentées, l'intérêt que pose la normalisation de ces messages a augmenté, ainsi le problème a-t-il inspiré de nombreux travaux depuis quelques années. Les différentes approches proposées font appel aux techniques de la correction orthographique, de la traduction statistique et de la reconnaissance automatique de la parole (Yvon, 2008). Par exemple, (Aw *et al.*, 2006) traitent le problème comme une tâche de traduction, où on vise à traduire l'anglais des textos en anglais standard. (Yvon, 2008) traite le problème comme une tâche de reconnaissance automatique de la parole (RAP), mais utilise également des techniques de correction orthographique ; une représentation phonétique des textos joue le rôle du modèle acoustique, et un modèle de langue est utilisé pour convertir les séquences de phones en séquences de mots. (Beaufort *et al.*, 2010) proposent pour leur part un système qui combine des techniques de correction automatique et de traduction statistique.

Ce travail porte sur l'application d'une technique de traduction statistique au problème de la normalisation de textos. Le problème consiste donc à « traduire » un texto en français standard. Ainsi, l'objectif de ce travail est de maximiser $p(f|e)$, où e désigne un texto et f, sa normalisation. On peut reformuler le problème ainsi en appliquant la loi de Bayes : $p(f|e) = p(f) \cdot p(e|f)$, ces deux termes étant déterminés par des modèles de langue et de traduction respectivement.

Une remarque concernant l'évaluation des techniques de normalisation de textos s'impose. Deux métriques sont souvent utilisées pour cette évaluation : certains auteurs utilisent le score BLEU (Papineni *et al.*, 2001), d'autres utilisent le taux d'erreur moyen par phrase (word error rate ou WER). Les deux métriques sont utilisées dans nos évaluations (ainsi que le taux de phrases erronées ou SER), et nous proposons qu'il est plus pertinent d'observer la réduction du WER, plutôt que le WER final, étant donné que les corpus de textos contiennent différentes quantités de formes à normaliser.

Les résultats présentés dans la littérature divergent beaucoup, et il est très délicat d'établir des comparaisons, notamment en raison des différences quant à la langue et la taille des corpus utilisés (en plus de l'utilisation de différentes métriques). (Aw *et al.*, 2006), qui travaillent sur la langue anglaise, obtiennent un score BLEU de 0,81. (Beaufort *et al.*, 2010) affirment que les systèmes à l'état de l'art obtiennent un WER de 11 %, et le système qu'ils proposent, qui exploite le corpus *SMS pour la science*, obtient un WER de 9,3 % et un score BLEU de 0,83. (Yvon, 2008) obtient un WER de 17,8 %, un résultat semblable à ce qu'on obtiendrait en utilisant un système générique de traduction statistique pour traiter ce problème. (Kobus *et al.*, 2008) obtiennent un WER de 16,5 % avec un système basé sur la métaphore de la RAP, de 12,3 % avec un système de traduction statistique, et d'environ 10,8% en combinant les deux systèmes.

1. Nous traduisons.

Le reste de cet article sera organisé de la façon suivante. Dans la section 2, nous décrirons les ressources utilisées dans le cadre de ce travail. La section 3 portera sur l'algorithme de recherche vorace que nous avons implémenté ; l'accent sera placé sur la fonction qui génère la normalisation initiale et la fonction qui génère de nouvelles hypothèses. Enfin, dans la section 4, nous analyserons les résultats obtenus.

2 Ressources

Trois ressources sont utilisées pour mettre en application l'algorithme vorace de recherche : un modèle de langue, un modèle de traduction et un corpus de textos annotés. Ce corpus est constitué de textos en français recueillis et annotés dans le cadre du projet Texto4Science (Langlais *et al.*, 2012). Chaque texto est accompagné d'une normalisation produite par un annotateur humain. Nous utilisons un corpus d'entraînement totalisant 11 000 textos alignés avec leur normalisation, un corpus de développement de 1135 paires et un corpus de test de 1000 textos non vus à l'entraînement, utilisé pour l'évaluation finale. Ce test est effectué seulement une fois, sur la meilleure version de notre système. Les autres résultats présentés proviennent tous d'évaluations sur le corpus de développement.

Le modèle de langue est un modèle trigramme avec lissage Kneser-Ney entraîné sur un corpus de français totalisant 673 000 phrases et 8, 6 millions de mots, qui comprend les textos normalisés du corpus d'entraînement.

Quant au modèle de traduction, nous utilisons un modèle probabiliste appris sur le corpus d'entraînement, de la forme $p(f|e)$ où e sont des mots de la langue des textos et f des mots du français normalisé. Le modèle est basé sur un alignement mot-à-mot entre f et e. Dans l'algorithme de recherche vorace décrit ci-dessous, la fonction qui génère de nouvelles hypothèses comprend une opération d'insertion de mots qui vise à combler les lacunes de ce modèle mot-à-mot. La simplicité de ce modèle, et de la fonction de score utilisée (voir section 3), est cohérente avec une approche par recherche vorace.

3 Algorithme

La technique mise en application ici est basée sur l'algorithme vorace de recherche décrit dans (Langlais *et al.*, 2007). Cet algorithme fait appel à trois fonctions : la première (*Seed*) génère une traduction initiale, la deuxième (*Score*) attribue aux traductions un score que l'on tente de maximiser, et la troisième (*Neighborhood*) génère, au moyen de différentes transformations, un ensemble d'hypothèses à tester à la prochaine itération, jusqu'à ce que le score plafonne. Dans (Langlais *et al.*, 2007), la fonction Seed choisit simplement la traduction la plus probable selon un modèle de traduction à segments ; la fonction Score est une combinaison log-linéaire de

modèles :

$$\begin{aligned} Score(e,f) = & \lambda_{lm} \log\, p_{lm}(f) & + \\ & \sum_i \lambda^i_{tm} \log\, p^i_{tm}(f|e) & - \\ & \lambda_w |f| & - \\ & \lambda_d p_d(e,f) \end{aligned}$$

où les λ sont des coefficients, p_{lm} est un modèle de langue, p^i_{tm} sont les différents modèles de traduction, $|f|$ est la longueur de la traduction et $p_d(e,f)$ est un modèle de distorsion.

L'algorithme vorace applique itérativement la fonction Neighborhood à une traduction courante et maximise le score jusqu'à ce qu'il plafonne.

Nous appliquons ici l'algorithme vorace au problème de la normalisation de textos. L'approche consiste globalement à :
– Générer une première normalisation plausible (Seed)
– Attribuer un score à cette normalisation (Score)
– Générer des nouvelles hypothèses au moyen de transformations (Neighborhood)
– Boucler les deux étapes précédentes jusqu'à ce que le score plafonne

3.1 Fonction Seed

Pour générer la normalisation initiale, deux méthodes sont comparées : recherche locale de la normalisation la plus probable pour chaque mot ; et identification de la meilleure normalisation par décodage de type Viterbi.

En ce qui concerne le décodage de type Viterbi, il est effectué à l'aide de la commande *Disambig* de SRILM (Stolcke, 2002), que nous utilisons pour produire la normalisation la plus probable étant donné une phrase source et un modèle de traduction. On peut également fournir à ce programme un modèle de langue afin qu'il maximise $p(e|f) \cdot p(f)$ plutôt que seulement $p(e|f)$.

3.2 Fonction Score

Nous simplifions la fonction de score de la façon suivante :

$$Score(e,f) = \lambda_{lm} \log\, p_{lm}(f) + \lambda_{tm} \log\, p_{tm}(e|f)$$

Le score utilisé maximise donc $p(e|f) \cdot p(f)$, ces deux probabilités étant déterminées au moyen des modèles de traduction et de langue. En ce qui concerne $p_{tm}(e|f)$, ce terme est calculé suivant la méthode IBM1 :

$$p(e_1^J|f_1^I) = \prod_{j=1}^{J} \left(\frac{1}{I} \sum_{i=0}^{I} p(e_j|f_i) \right)$$

Quant à $p_{lm}(f)$, nous calculons le produit des probabilités des trigrammes d'une phrase [2] (des tokens de début et de fin de phrase sont ajoutés). Ces probabilités sont tirées du modèle de langue.

3.3 Fonction Neighborhood

(Langlais *et al.*, 2007) décrivent six opérations mises en application dans la fonction Neighborhood, dont quelques-unes sont propres aux modèles à segments utilisés dans ce travail, alors que l'approche utilisée ici traduit (normalise) mot à mot. En revanche, les opérations *Swap*, qui intervertit deux mots adjacents, et *Replace*, qui remplace un segment dans la traduction par d'autres segments présents dans les modèles de traduction, s'appliquent très bien au modèle de traduction mot-à-mot. Nous appliquons aussi une opération que les auteurs ont suggérée, c'est-à-dire l'insertion de mots.

Celle-ci consiste à insérer des mots à n'importe quelle position dans une phrase, le vocabulaire des mots à insérer pouvant être déterminé de différentes façons. Nous mettons à l'épreuve deux variantes. L'opération *Insert_sp* insère seulement des mots que (Brown *et al.*, 1993) qualifient de *spurious*, c'est-à-dire des mots de la phrase cible qui ne sont alignés avec aucun mot dans la phrase source. Ceux-ci sont identifiés automatiquement à partir du modèle de traduction, en repérant tous les mots qui sont associés au mot vide. La deuxième opération, que nous appelons *Insert_tr*, insère d'autres traductions présentes dans le modèle de traduction pour les mots de la phrase source, l'objectif étant de combler les lacunes du modèle mot-à-mot, qui risque de proposer une traduction incorrecte dans les cas où un mot source doit être traduit par plus d'un mot cible.

En somme, la fonction Neighborhood fait appel à quatre opérations :
– Swap : intervertir deux mots adjacents
– Replace : remplacer un mot cible par d'autres équivalents potentiels
– Insert_tr : insérer d'autres équivalents potentiels d'un mot source
– Insert_sp : insérer des mots *spurious*
Les opérations Insert_sp et Swap seront utilisées dans toutes les versions évaluées ici sauf indication contraire, tandis que Replace et Insert_tr feront l'objet d'évaluation distinctes.

4 Analyse des résultats

4.1 Seed et Neighborhood

L'objectif principal de cette évaluation est de mettre à l'épreuve différentes façons d'obtenir la normalisation initiale (fonction Seed) et de générer des nouvelles hypothèses (Neighborhood). Avant de procéder à ces tests, nous avons d'abord enrichi manuellement la liste de mots *spurious* exploitée par l'opération Insert_sp. Une analyse rapide des mots extraits du modèle de traduction a montré que plus de la moitié étaient des mots de classes fermées. Nous avons complété les

2. Notre programme exploite un wrapper pour Python qui permet d'interroger SRILM (Madnani, 2009). Voir `http://www.desilinguist.org`.

listes d'articles, de déterminants démonstratifs et possessifs et de pronoms, ajoutant 32 mots à la liste. Une légère diminution du WER a été observée, à très faible coût.

Seed	IT	WER (%)	SER (%)	BLEU
Baseline		21,01	62,29	0,5683
Topword	Non	31,87	75,42	0,4202
	Oui	29,37	74,63	0,4382
Dis	Non	31,45	74,98	0,4237
	Oui	28,92	74,36	0,4456
Dis2	Non	14,05	53,92	0,7169
	Oui	12,22	48,63	0,7468
Dis3	Non	12,78	49,96	0,7394
	Oui	**11,05**	**43,88**	**0,7674**

TABLE 1 – Influence de Seed et de Insert_tr

Les scores qu'offrent différentes variantes de la fonction Seed sont présentées dans la table 1. Pour chacune des techniques, deux variantes de la fonction Neighborhood sont évaluées. Chacune comprend les opérations Swap et Insert_sp, mais nous activons et désactivons l'opération Insert_tr (indiqué dans la colonne *IT*). En ce qui concerne les variantes de Seed, *Topword* choisit simplement le mot cible le plus probable pour chaque mot source. *Dis* utilise le décodage Viterbi au moyen de Disambig, mais n'exploite aucun modèle de langue, seulement un modèle de traduction. *Dis2* exploite un modèle de langue bigramme et *Dis3*, un modèle trigramme. Enfin, pour déterminer le *baseline*, nous conservons simplement le texto de départ.

Les résultats montrent que les techniques naïves de génération de la normalisation initiale offrent des scores très pauvres, Topword et Dis obtenant des résultats à peu près équivalents. Or, lorsqu'on fournit un modèle de langue à Disambig, les scores deviennent nettement meilleurs. Cela suggère que cette implémentation de l'algorithme nécessite une normalisation initiale d'une certaine qualité.

Nous avons également évalué la fonction Replace, qui parcourt les mots de la source, extrait tous les équivalents du modèle de traduction, cherche la traduction du mot source dans la traduction courante, et la remplace par chacun des équivalents. Nous l'avons implémentée dans la version du programme qui obtient les meilleurs résultats, c'est-à-dire Dis3 avec Insert_tr, et le taux d'erreurs moyen par phrase ne diminue pas ; au contraire, il augmente d'environ 4 %, et le score BLEU diminue de 2 %. Il semble donc que l'opération Replace n'est pas bénéfique, du moins lorsque les normalisations initiales sont de bonne qualité. Nous montrerons dans la section suivante que le contraire est vrai lorsque celles-ci sont moins bonnes.

4.2 Amélioration des normalisations générées naïvement

Ayant identifié une combinaison de fonctions qui produit des résultats satisfaisants, nous cherchons à vérifier dans quelle mesure l'algorithme vorace de recherche améliore la qualité des normalisations fournies par la fonction Seed la plus naïve, c'est-à-dire Topword.

La table 2 présente les résultats de cette évaluation. *Dis3* indique les résultats qu'on obtient simplement en laissant à Disambig le soin de choisir la meilleure normalisation étant donné un modèle de traduction et un modèle de langue trigramme. *Greedy_search* désigne l'implémentation de l'algorithme qui obtient les meilleurs résultats : Dis3 est utilisé pour la traduction initiale, et la fonction Neighborhood comprend les opérations Swap, Insert_sp et Insert_tr. *TW* indique les résultats qu'on obtient par la méthode Topword, sans application de l'algorithme vorace. Par la suite, on montre comment la performance de l'algorithme vorace varie à mesure qu'on ajoute des opérations à la fonction Neighborhood : on désigne Swap par *SW*, Insert_sp par *IS*, Insert_tr par *IT* et Replace par *RE*.

Les résultats montrent que l'algorithme vorace n'améliore pas énormément la qualité des normalisations produites par Dis3, qui sont déjà beaucoup plus proches des normalisations de référence. Or, nous arrivons tout de même à réduire le taux d'erreurs moyen par phrase (WER) de presque moitié et à augmenter le score BLEU d'environ 35 % par rapport au baseline.

Si l'apport de l'algorithme vorace n'est pas énorme lorsque les normalisations initiales sont bonnes, il devient considérable lorsque celles-ci sont générées grossièrement. Les normalisations générées par Topword s'éloignent nettement des normalisations de référence, et Swap et Insert_sp ne les améliorent pas. Par contre, Replace (et dans une moindre mesure Insert_tr) est très bénéfique, offrant une réduction du taux d'erreurs moyen de l'ordre de 40 % et une augmentation du score BLEU d'environ 47 %. Ces gains sont attribuables, du moins en partie, au rôle que joue le modèle de langue, qui permet par ailleurs d'améliorer les normalisations générées par *Dis*, comme nous l'avons vu. Malgré ces gains, nous obtenons des meilleurs résultats lorsque les normalisations de départ sont déjà de bonne qualité, intégrant un modèle de langue. Rappelons aussi que, lorsque les normalisation initiales sont bonnes, Replace n'a pas un effet favorable. Il nous semble que ces observations correspondent aux intuitions qu'on peut avoir par rapport à cette approche de la traduction (ou normalisation).

4.3 Évaluation sur le corpus de test

Les résultats de l'évaluation finale, effectuée sur le corpus de test, sont présentés dans la table 3. Nous évaluons le système qui fournit les meilleurs résultats sur le corpus de développement : la normalisation de départ est générée par Dis3, et la fonction Neighborhood utilise les opérations Swap, Insert_sp et Insert_tr pour générer des nouvelles hypothèses. Tout d'abord, on observe que les textos contiennent une proportion nettement plus élevée de formes non standard que ceux

Méthode	WER (%)	SER (%)	BLEU
Baseline	21,01	62,29	0,5683
Dis3	13,01	51,98	0,7230
Greedy_search	11,05	43,88	0,7674
TW	30,42	75,51	0,4051
TW+SW+IS	31,87	75,42	0,4202
TW+SW+IS+IT	29,37	74,63	0,4382
TW+SW+IS+IT+RE	17,78	51,81	0,5947

TABLE 2 – Impact de l'algorithme vorace de recherche

	WER	SER	BLEU
Baseline	28,90	68,60	0,4677
Greedy_search	19,32	57,70	0,6189

TABLE 3 – Évaluation sur le corpus de test

du corpus de développement, le WER étant 37, 6 % plus élevé. Ainsi, le WER des normalisations produites passe de 11, 05 % (sur le corpus de développement) à 19, 32 %. De plus, la diminution du WER observée en test, de 33 %, est inférieure à la diminution observée pendant la phase de développement (47 %). Or, si toute différence de WER de 30 % est considérée significative (Yvon, 2008), il mérite d'être souligné que nos résultats dépassent ce seuil. En ce qui concerne le score BLEU, le score des normalisations produites est beaucoup plus faible lorsqu'on évalue sur le corpus de test, mais l'augmentation du score BLEU (32 %) est cohérente avec celle que nous avons observée pendant le développement (35 %).

5 Conclusion

Dans ce travail, nous avons mis en application un algorithme de recherche vorace utilisé en traduction statistique dans le but de normaliser des textos. L'accent a été placé sur les fonctions qui génèrent la normalisation initiale et aux opérations permettant de générer des nouvelles hypothèses.

L'approche qui obtient les meilleurs résultats consiste à générer la normalisation initiale par décodage de type Viterbi à partir des modèles de traduction et de langue ; à utiliser les opérations d'alternance et d'insertion de mots afin de générer des nouvelles hypothèses ; et à maximiser la fonction de score. Cette méthode engendre une diminution du taux d'erreurs moyen par phrase de 33 % lors de l'évaluation finale, et une augmentation du score BLEU de plus de 30 %.

L'opération Replace, qui consiste à remplacer des mots dans la normalisation courante par d'autres équivalents tirés du modèle de traduction, n'a pas un effet bénéfique lorsque les normalisations initiales sont de bonne qualité. Or, lorsque celles-ci sont générées par une simple recherche locale du mot cible le plus probable pour chaque mot source, l'opération Replace permet d'améliorer la qualité des normalisations, notamment grâce à l'apport du modèle de langue.

Ces techniques simples fournissent des résultats qui nous semblent intéressants. Il nous paraît donc profitable de traiter la normalisation des textos comme un problème de traduction intralinguistique.

Remerciements

Nous désirons remercier Philippe Langlais, ainsi que les relecteurs, pour leurs commentaires et leurs suggestions sur ce travail. Nous remercions M. Langlais ainsi que Fabrizio Gotti pour les ressources mises à notre disposition. Nous remercions également le Fonds de recherche du Québec – Société et culture pour son soutien financier.

Références

Aw, A., Zhang, M., Xiao, J. et Su, J. (2006). A Phrase-Based Statistical Model for SMS Text Normalization. *In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney (Australie). Association for Computational Linguistics.

Beaufort, R., Roekhaut, S., Cougnon, L.-A. et Fairon, C. (2010). A Hybrid Rule/Model-Based Finite-State Framework for Normalizing SMS Messages. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, Uppsala (Suède). Association for Computational Linguistics.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A. et Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Kobus, C., Yvon, F. et Damnati, G. (2008). Normalizing SMS : are Two Metaphors Better than One ? *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 441–448, Manchester (Angleterre). Coling 2008 Organizing Committee.

Langlais, P., Drouin, P., Paulus, A., Brodeur, E. R. et Cottin, F. (à paraître, 2012). Texto4science : a Quebec French Database of Annotated Short Text Messages. *In Proceedings of Language Resources and Evaluation Conference (LREC) 2012*, Istanbul (Turquie). ELRA.

Langlais, P., Patry, A. et Gotti, F. (2007). A Greedy Decoder for Phrase-Based Statistical Machine Translation. *In Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 104–113, Skövde (Suède).

Madnani, N. (2009). Querying and Serving N-gram Language Models with Python. *The Python Papers*, 4(2).

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2001). Bleu : A Method for Automatic Evaluation of Machine Translation. Rapport technique RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. *In Proceedings of ICSLP*, Denver (États-Unis).

Yvon, F. (2008). Reorthography of SMS Messages. Rapport technique 2008-18, LIMSI-CNRS.

Prémices d'une analyse syntaxique par transition pour des structures de dépendance non-projectives

Boris Karlov[1], Ophélie Lacroix[2]
(1) Université de Tver, 33, rue Zheliabov, 170000, Tver, Russie
(2) LINA, 2, rue de la Houssinière 44322 Nantes Cedex 3
bnkarlov@gmail.com
ophelie.lacroix@univ-nantes.fr

RÉSUMÉ
L'article présente une extension de l'analyseur traditionnel en dépendances par transitions adapté aux dépendances discontinues et les premiers résultats de son entraînement sur un corpus de structures de dépendances de phrases en français. Les résultats des premières expérimentations vont servir de base pour le choix des traits des configurations de calcul bien adaptés aux dépendances discontinues pour améliorer l'apprentissage des dépendances tête.

ABSTRACT
Beginnings of a Transition-Based Parsing for Non-Projectives Dependency Structures

This paper presents an extension of the traditional transition-based dependency parser adapted to discontinuous dependencies and the first results of its training on a dependency tree corpus of French. The first experimental results will be useful for the choice of parsing configuration features well adapted to discontinuous dependencies in order to ameliorate learning of head dependencies.

MOTS-CLÉS : analyse syntaxique par transitions, structure de dépendance non-projective, grammaire catégorielle de dépendance.

KEYWORDS: transition-based parsing, non-projective dependency structure, dependency categorial grammar.

1 Introduction

Il existe différentes méthodes d'analyses syntaxiques permettant de traiter les phrases du français. Ces analyses peuvent être syntagmatiques (par constituants) (Kow *et al.*, 2006; Vanrullen *et al.*, 2006) ou en dépendance (Nasr, 2004; Brunet-Manquat, 2005; Alfared *et al.*, 2011). Elles peuvent être guidées par les règles d'une grammaire (probabiliste ou non) ou être entainées sur un corpus. Depuis plusieurs années, les différentes méthodes d'analyses en dépendance (voir (Kübler *et al.*, 2009)) gagnent en interêt dans le domaine de l'analyse syntaxique. Dans cet article nous nous plaçons dans le cas d'une analyse syntaxique en dépendance, par transition, entraînée sur un corpus correct par rapport à une grammaire de dépendance (Alfared *et al.*, 2011).
Les méthodes d'analyses en dépendance permettent de produire des représentations d'une phrase plus expressives sémantiquement que celles par constituant. D'autres formalismes comme les grammaires d'arbres adjoints sont parfois utilisés pour prendre en compte l'aspect sémantique

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 81–94,
Grenoble, 4 au 8 juin 2012.

des phrases, mais ne permettent pas, par exemple, de révéler la relation de coréférence (Candito et Kahane, 1998). La représentation des phrases en structure de dépendance que nous avons choisi de mettre en avant ici, nous permet d'exprimer certaines relations qui ne sont pas toutes considérées dans les méthodes classiques par constituants. La figure 1 présente, entre autres, la relation distante existante entre les mots "*moins*" et "*que*" que l'on peut trouver lors d'une comparaison.

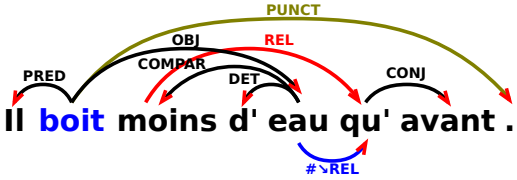


FIGURE 1 – Structure de dépendance de la phrase "Il boit moins d'eau qu'avant."

Il est donc possible de définir des dépendances croisées. Cependant une telle représentation (dite non-projective, voir la section 2.1) est peu utilisée dans le domaine de l'analyse en dépendance pour les phrases en français. Les analyses en dépendances courantes se basent principalement sur le même modèle que les analyses par constituants pour leur lien plus évident avec les grammaires formelles. Certains travaux consistent d'ailleurs à convertir des corpus annotés par constituants en corpus de dépendance (Candito *et al.*, 2009). Ce genre d'analyse produit donc des structures de dépendances projectives (sans dépendances croisées), proches des arbres syntagmatiques. On parle alors dans ce cas d'arbre de dépendance. Néanmoins, tout le potentiel des structures de dépendances n'est pas exploité lors d'une telle conversion. Comme le souligne (Rambow, 2010), une perte d'information serait inévitablement constatée si on tentait de convertir une structure de dépendance en structure par syntagme.
D'autre part, d'après (Nivre, 2011), les analyses du type analyse tabulaire (basée sur les premiers algorithmes tel que CKY ou Earley, voir (Kübler *et al.*, 2009)) ou analyse par satisfaction de contraintes (Maruyama, 1990) ne sont pas bien adaptées aux structures de dépendance non-projectives que nous souhaitons traiter. Le problème devient d'une complexité trop grande ou NP-complet. Les analyses par transition sur des structures de dépendance projectives pour des phrases en anglais donnent de bons résultats et l'adaptation aux structures de dépendance non-projectives se fait en une complexité au pire quadratique. Récemment, (Choi et Palmer, 2011) effectuent une analyse par transition sur des dépendances non-projectives pour des phrases en anglais. (Alfared *et al.*, 2011) travaillent sur un analyseur syntaxique semi-automatique permettant de produire des structures de dépendances projectives et non-projectives sur des phrases en français. Nous nous appuierons ici sur la grammaire catégorielle de dépendance qu'ils utilisent, (Dikovsky, 2011), pour définir les dépendances croisées existantes dans les structures de dépendance non-projectives. Nous allons alors présenter l'analyseur syntaxique par transition, que nous avons développé, permettant de définir des structures de dépendance non nécessairement projectives pour des phrases en français.

2 Structures de dépendance

Une structure de dépendance représente le résultat d'une analyse en dépendance pour une phrase donnée. Ces structures permettent de mettre en évidence les relations binaires entre les mots d'une phrase telles que les relations sujet-verbe, verbe-objet, etc. Les structures de dépendance sont à différencier des structures syntagmatiques qui sont le résultat d'une analyse par constituants. Mel'cuk met en évidence les différences, les avantages et les inconvénients de ces deux méthodes d'analyses dans (Mel'cuk, 1988) et propose la théorie "Sens-Texte" qui révèle l'aspect sémantique des dépendances entre les mots. Il met en avant, en outre, le fait que les structures de dépendances profondes sont invariantes selon l'ordre des mots dans la phrase. Ces idées sont aussi reprises plus récemment par (Kahane, 2001). Ici nous travaillons sur l'analyse des structures de dépendance de surface qui ont la particularité de garantir l'ordre des mots dans la phrase.

2.1 Projectivité, non-projectivité

De manière théorique, une structure de dépendance est un graphe orienté dont les noeuds sont les mots de la phrase à analyser et les arcs représentent les dépendances reliant ces mots. Dans une structure de dépendance, la relation de dominance est importante. Un mot domine de manière directe ses subordonnés, mais il domine aussi les subordonnés de ses subordonnés de manière transitive. Ainsi dans toute phrase la racine domine tous les mots. Un mot dominant est appelé un gouverneur.
Une structure de dépendance peut être projective ou non. Une structure est projective si pour chaque mot *w* tous les mots qui se trouvent entre *w* et ses subordonnés sont aussi dominés par *w*, comme illustré par la figure 2. Par ailleurs, les figures 1 et 3 présentent des structures de dépendances non-projectives. En effet, dans le cas de la figure 3, les mots "*attend*", "*son*" et "*mari*" sont compris entre le mot "*Elle*" et son subordonné "*pressée*" mais ne sont pas des subordonnés de "*Elle*", la structure n'est donc pas projective.

FIGURE 2 – Structure de dépendance de la phrase "La nature fait bien les choses."

2.2 Représentation utilisée

Les arcs représentant les dépendances sont de trois sortes différentes : projectifs, discontinus ou ancres. Les dépendances projectives sont des dépendances locales (courtes) qui ne se croisent pas les unes avec les autres. Les dépendances discontinues sont les dépendances non-projectives distantes qui peuvent croiser les dépendances projectives. Et les ancres sont des dépendances locales permettant de lier localement une dépendance discontinue. Effectivement, les subordonnés

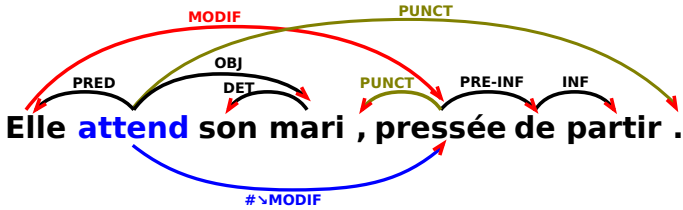


FIGURE 3 – Structure de dépendance de la phrase "Elle attend son mari, pressée de partir."

des dépendances ancres sont aussi subordonnés via des dépendances discontinues. Ces ancres sont importantes pour définir la grammaire présentée ensuite. La figure 3 permet d'illustrer chaque sorte d'arc énoncé ci-dessus. Dans cette figure, la dépendance discontinue est de type *modificateur* (le participe passé "*pressée*" est bien rattaché au pronom personnel "*Elle*"). Les dépendances discontinues, comme $MODIF$ [1] sont indiquées au dessus de la phrase tandis que les dépendances ancres liées aux dépendances discontinues, comme $\swarrow MODIF$, sont indiquées au dessous de la phrase. La *négation*, la *réfléxivité*, *l'apposition* et l'*aggrégation*, parmi d'autres types de dépendances, peuvent aussi engendrer des dépendances discontinues.

2.3 Grammaire catégorielle de dépendance (CDG)

(Dekhtyar et Dikovsky, 2008) exposent une méthode pour représenter les dépendances en catégories comme dans les grammaires catégorielles classiques (Bar-Hillel *et al.*, 1964). Cette grammaire a permis à (Alfared *et al.*, 2011) de produire le corpus que nous utiliserons ensuite pour notre analyse (voir section 4.1). L'idée est de représenter les dépendances projectives et les dépendances ancres par des catégories qui sont les types de ces dépendances et les dépendances non-projectives par des valences polarisées. Sur une phrase comme celle donnée en exemple dans la figure 4, on aura pour chaque mot les catégories suivantes : il $\mapsto [pred]$, y $\mapsto [\# \swarrow clit]^{\swarrow clit}$, est $\mapsto [\# \swarrow clit \backslash pred \backslash S / punct / aux]$, allé $\mapsto [aux]^{\nwarrow clit}$, . $\mapsto [punct]$.

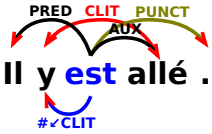


FIGURE 4 – Structure de dépendance de la phrase "Il y est allé."

Ces catégories sont ensuite utilisées dans une grammaire catégorielle enrichie de règles de dérivations, exposées dans la table 1, permettant de traiter les valences polarisées. La règle

1. $MODIF$ est le type attribué à une dépendance entre un mot et son modificateur, qui peut être un participe ou un adjectif. Les types de dépendances utilisés dans ces structures sont ceux employés par (Alfared *et al.*, 2011) pour construire leur corpus. Par ailleurs, les dépendances utilisées dans les figures de cet article sont étiquetées par les noms des groupes de dépendances auxquelles elles appartiennent (voir section 3.4) pour une illustration plus simple.

L permet d'éliminer les catégories comme dans les grammaires catégorielles classiques, et de concaténer les valences polarisées en une chaîne que l'on appelle *potentiel*. Pour la règle **I**, les valences sont traitées de la même manière tandis que la dérivation s'effectue sur les catégories itérables. De même, la règle Ω permet d'éliminer une catégorie itérable en conservant le potentiel tel qu'il était. Puis l'élimination des valences dans la dérivation (règle **D**) se fait sur le principe **FA** (First Available). Tout d'abord, des valences duales sont des valences de même catégorie dont les polarités sont $\swarrow$ et $\nwarrow$, ou $\nearrow$ et $\searrow$. Le principe FA indique que les valences duales les plus proches dans un potentiel sont des paires. Alors dans un potentiel $P_1(\swarrow C)P(\nwarrow C)P_2$, si $(\swarrow C)$ et $(\nwarrow C)$ (valences duales) n'apparaissent pas dans P, $(\swarrow C)$ et $(\nwarrow C)$ satisfont le principe **FA**.

$\mathbf{L}^1$	$C^{P_1}\left[C\backslash\beta\right]^{P_2} \vdash \left[\beta\right]^{P_1P_2}$
$\mathbf{I}^1$	$C^{P_1}\left[C^*\backslash\beta\right]^{P_2} \vdash \left[C^*\backslash\beta\right]^{P_1P_2}$
Ω^1	$\left[C^*\backslash\beta\right]^{P} \vdash \left[\beta\right]^{P}$
$\mathbf{D}^1$	$\alpha^{P_1(\swarrow C)P(\nwarrow C)P_2} \vdash \alpha^{P_1PP_2}$, si $(\swarrow C)(\nwarrow C)$ satisfait le principe **FA**

TABLE 1 – Règles de la grammaire catégorielle de dépendance généralisée

Un exemple d'arbre de dérivation, utilisant les valences polarisées pour les dépendances discontinues, dérivé de la phrase "*Il y est allé.*" est présenté par la figure 5.

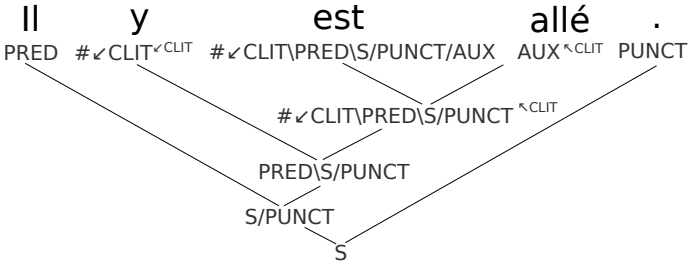


FIGURE 5 – Arbre de dérivation utilisant la CDG sur la phrase "Il y est allé."

3 Analyse par transition

Comme expliqué par (Nivre, 2008), une structure de dépendance peut être traduite en une suite de transitions. Le but d'un analyseur syntaxique par transition est de trouver une suite de transitions (opérations sur les configurations de calcul de l'analyseur) qui permette de construire une structure de dépendance correcte pour une phrase donnée. Pour faire ce travail, l'analyseur syntaxique doit s'appuyer sur un oracle qui lui indique quelle transition appliquer dans telle ou telle configuration. Les configurations et transitions sont des éléments essentiels de l'analyse par transition, qui peuvent varier selon l'utilisation que l'on veut en faire. Kübler, Donald et Nivre définissent clairement leurs emplois dans (Kübler *et al.*, 2009). Ici nous présentons un analyseur

par transition adapté aux dépendances discontinues, que nous avons développé. Le travail consiste tout d'abord à générer un oracle à partir d'un corpus de référence. Chaque structure de dépendance du corpus est traduite en un ensemble de couples configuration-transition qui sont ajoutées à l'oracle selon le nombre d'occurences d'une transition pour une configuration [2] donnée. L'oracle enregistre donc seulement la meilleure transition possible pour une configuration. Cet oracle sera alors utilisé pour déterminer une suite de transitions à appliquer, pour chaque phrase à analyser, permettant de construire des structures de dépendance. L'analyseur est donc déterministe car il s'appuie sur l'oracle pour choisir la transition à appliquer à chaque étape de l'analyse.

3.1 Configurations

Une configuration, pour une phrase, est un "état" dans lequel est l'analyseur syntaxique analysant cette phrase. On peut y appliquer une transition qui permettra de passer dans la configuration suivante.
Soit une phrase $s = w_1 w_2 ... w_n$. Une configuration pour une phrase est un quintuplet $(\sigma, \beta, \theta, i, E)$ où σ est une pile de mots w_i dans s, β est un tampon de mots w_i dans s, θ est un tampon de valences (un potentiel), i est la position du mot courant, et E est un ensemble d'arcs (w_i, r, t, w_j) où r est l'étiquette de l'arc (le type de la dépendance) et t est la sorte de l'arc (projectif, discontinu, ancre). La configuration initiale c_0 pour chaque phrase est :

$$([], [w_1, ..., w_n], [], 1, \emptyset).$$

La configuration finale, pour n'importe quel σ et E, est :

$$(\sigma, [], [], n+1, E).$$

3.2 Transitions

Les transitions sont des opérations s'appliquant aux configurations pour obtenir les configurations suivantes. Celles-ci sont présentées dans la table 2. Ces transitions se basent sur les transitions évoquées par (Kübler *et al.*, 2009). Nous y ajoutons des éléments permettant de traiter les structures de dépendances projectives aussi bien que les structures non-projectives. En effet, la transition *PutPotential* ajoute les valences polarisées dans le potentiel θ selon les règles de la table 1 de la grammaire catégorielle de dépendance vu à la section 2.3. Puis les transitions *DistLeft* et *DistRight* permettent de définir les dépendances discontinues de la structure de dépendance en éliminant les valences duales du potentiel θ, suivant la règle **D**. En outre, les transitions *LocalLeft* et *LocalRight* servent à définir les dépendances projectives ainsi que les dépendances ancres. Les ancres permettent de détecter les subordonnés discontinus.
Pour convertir une structure de dépendance en une suite de transitions, on utilise un algorithme qui applique les transitions dans un certain ordre. Pour cela, on considère toujours le mot en haut de la pile σ comme le mot courant de la configuration. L'algorithme essaie d'abord de trouver les dépendances discontinues de la structure. Donc dans un premier temps, il tente d'ajouter une valence dans le potentiel θ en appliquant la transition *PutPotential* si le mot courant est impliqué

2. Une configuration est représentée par un vecteur de traits : image de taille réduite de la configuration, voir section 3.4

dans une dépendance discontinue. Puis on applique les transitions *DistLeft* et *DistRight*, pour définir les dépendances distantes trouvées, jusqu'à ce qu'il n'y ait plus de valences duales. Ensuite l'algorithme traite les dépendances locales. Il commence donc par appliquer toutes les transitions *LocalLeft* possible pour le mot courant, de son subordonné le plus proche au plus à gauche. Puis si ce mot a des dépendances à droite, on le garde sur la pile et on applique la transition *Put* pour empiler le mot suivant sur σ. Néanmoins, si le mot courant n'a pas de subordonné droit et a un gouverneur sur la gauche alors il est possible d'appliquer la transition *LocalRight*. Le dernier cas possible est si le mot courant n'a plus aucun subordonné sur la droite et n'a pas non plus de gouverneur. Alors il s'agit de la racine et le processus est terminé.

Transition	Effet de son application sur une configuration
PutPotential $(p_1, ..., p_m)$	$(\sigma, w_k \mid \beta, \theta, k, E) \Rightarrow (\sigma, w_k \mid \beta, \theta p_1^k ... p_m^k, k+1, E)$ où les p_j sont des valences.
Put	$(\sigma, w_i \mid \beta, \theta, k, E) \Rightarrow (\sigma \mid w_i, \beta, \theta, next(k,i), E)$ où $next(k,l) = \begin{cases} k+1 & si\ k = l \\ k & sinon \end{cases}$
LocalLeft(d)	$(\sigma \mid w_i, w_j \mid \beta, \theta, k, E) \Rightarrow (\sigma, w_j \mid \beta, \theta, k, E \cup \{(w_i, d, TYPE(d), w_j)\})$
LocalRight(d)	$(\sigma \mid w_i, w_j \mid \beta, \theta, k, E) \Rightarrow (\sigma, w_i \mid \beta, \theta, next(k,j), E \cup \{(w_i, d, TYPE(d), w_j)\})$
DistLeft(v)	$(\sigma, \beta, \theta_1 \swarrow v^i \theta_2 \nwarrow v^j \theta_3, k, E) \Rightarrow (\sigma, \beta, \theta, k, E \cup \{(w_j, v, discontinuous, w_i)\})$ si $\swarrow v^i \nwarrow v^j$ est la paire la plus à gauche satisfaisant la condition FA.
DistRight(v)	$(\sigma, \beta, \theta_1 \nearrow v^i \theta_2 \searrow v^j \theta_3, k, E) \Rightarrow (\sigma, \beta, \theta, k, E \cup \{(w_i, v, discontinuous, w_j)\})$ si $\nearrow v^i \searrow v^j$ est la paire la plus à gauche satisfaisant la condition FA.

TABLE 2 – Les transitions utilisées par l'analyseur

3.3 Exemple de conversion

Pour illustrer la méthode de conversion d'une structure de dépendance en séquence de transitions voici un exemple pour la phrase "*Cette victoire, elle l'a méritée.*" dont la structure de dépendance est présentée par la figure 6. Cette phrase comprend des dépendances discontinues de type *clitique* et *coréférence*. La suite de transitions appliquées à cette phrase et les éléments de chaque configuration en découlants sont illustrés dans la table 3.

Transition	Pile σ	Tampon β	Potentiel θ	Ensemble d'arcs E
	[]	[Cette, victoire , elle l' a méritée .]	[]	∅
Put	[Cette]	[victoire , elle l' a méritée .]	[]	∅
PutPotential	[Cette]	[victoire , elle l' a méritée .]	[↙coref]	∅
LocalLeft	[]	[victoire , elle l' a méritée .]	[↙coref]	E = (victoire, projective, det, Cette)
Put	[victoire]	[, elle l' a méritée .]	[↙coref]	E
LocalRight	[]	[victoire elle l' a méritée .]	[↙coref]	E = E ∪ (victoire, projective, punct, ,)
Put	[victoire]	[elle l' a méritée .]	[↙coref]	E
Put	[victoire elle]	[l' a méritée .]	[↙coref]	E
PutPotential	[victoire elle]	[l' a méritée .]	[↙coref↖coref]	E
PutPotential	[victoire elle]	[l' a méritée .]	[↙coref↖coref↙clit]	E
DistLeft	[victoire elle]	[l' a méritée .]	[↙clit]	E = E ∪ (l', discontinuous, coref, victoire)
Put	[victoire elle l']	[a méritée .]	[↙clit]	E
LocalLeft	[victoire elle]	[a méritée .]	[↙clit]	E = E ∪ (a, anchor, clit, l')
LocalLeft	[victoire]	[a méritée .]	[↙clit]	E = E ∪ (a, projective, pred, elle)
LocalLeft	[]	[a méritée .]	[↙clit]	E = E ∪ (a, anchor, coref, victoire)
Put	[a]	[méritée .]	[↙clit]	E
PutPotential	[a]	[méritée .]	[↙clit↖clit]	E
DistLeft	[a]	[méritée .]	[]	E = E ∪ (méritée, discontinuous, clit, l')
LocalRight	[]	[a .]	[]	E = E ∪ (a, projective, aux, méritée)
Put	[a]	[.]	[]	E
LocalRight	[]	[a]	[]	E = E ∪ (a, projective, punct, .)

TABLE 3 – Conversion de la structure de dépendance de la phrase "Cette victoire, elle l'a méritée." en une suite de transitions

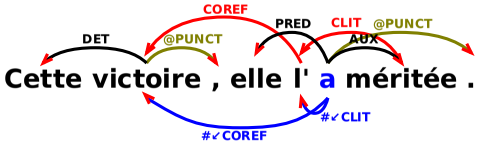


FIGURE 6 – Structure de dépendance de la phrase "Cette victoire, elle l'a méritée."

3.4 Oracles

L'idée de l'analyse par transition est de représenter partiellement les éléments des configurations (dont la taille n'est pas limitée) par les valeurs de vecteurs de traits. Le problème du choix des traits se pose alors. Différents traits peuvent être choisis pour entraîner un oracle à partir d'un corpus de structure de dépendance. Nous expérimentons ici avec huit vecteurs de traits différents. On peut donc choisir :
– l'utilisation ou non du nom complet de la classe grammaticale affectée au mot dans les structures de dépendance
– l'utilisation ou non des groupes de dépendances à la place des noms exacts des dépendances
– la taille du vecteur de traits.

Pour la classe grammaticale, on peut choisir d'utiliser le nom simple (court) de la classe ou le nom enrichi d'un paramètre (long). Par exemple pour le cas d'un mot de la classe adjectif quantificateur, le nom court correspondra à *Adj* et le nom long à *Adjquantifier*. La classe courte *Adj* rassemble les adjectifs modificateurs, quantifiants, restrictifs, comparatifs, etc... Un autre exemple est celui de la classe des noms (*N*). Elles peut regrouper les classes de noms longs *Nproper* (noms propres), *Ncommon* (noms communs), *Ntime* (noms à référence temporelle), etc...

Par ailleurs, les dépendances avec des fonctions syntaxiques proches peuvent être regroupées dans des groupes de dépendances. Il est alors possible de sélectionner une option traitant les dépendances selon le groupe de dépendances auxquelles elles appartiennent ou selon leurs types de dépendances précis. Par exemple pour le cas d'un mot dont la dépendance est du type objet accusatif, on utilisera par défaut le type de dépendance complet *a-obj*, mais on emploiera le type *OBJ* pour une représentation des types par groupe. Dans ce cas là, le type *OBJ* pourra aussi être assigné à une dépendance de type *g-obj* (objet génitif), *l-obj* (objet locatif) ou *o-obj* (objet oblique).

Le vecteur de trait peut être de taille 6 ou 8. Dans les deux cas, les vecteurs de traits, pour une configuration donnée à un moment de l'analyse, sont constitués de :
– la classe grammaticale du mot au sommet de la pile σ
– la classe grammaticale du premier mot du tampon β de mots
– le type de la dépendance la plus à gauche (c'est-à-dire, dont le subordonné est le plus à gauche) du mot en haut de la pile σ suivi de la sorte de l'arc (discontinu, projectif, ancre)
– le type de la dépendance la plus à droite du mot en haut de la pile σ suivi de la sorte de l'arc
– la dernière valence du potentiel et son orientation
– l'avant-dernière valence du potentiel et son orientation

Si le vecteur est de taille 8, il possède les deux traits suivants en plus :
– le type de la dépendance la plus à gauche du premier mot du tampon β suivi de la sorte de l'arc
– le type de la dépendance la plus à droite du premier mot du tampon β suivi de la sorte de l'arc
Ainsi nous générons huit oracles différents, exposés dans la table 4. L'oracle 1 est ainsi le plus général et l'oracle 8 est le plus précis.

	Taille du vecteur		*Classe grammaticale*		*Groupe*	
	6	8	*courte*	*longue*	*oui*	*non*
Oracle 1	√		√		√	
Oracle 2	√		√			√
Oracle 3	√			√	√	
Oracle 4	√			√		√
Oracle 5		√	√		√	
Oracle 6		√	√			√
Oracle 7		√		√	√	
Oracle 8		√		√		√

TABLE 4 – Les paramètres des différents oracles

4 Résultats des analyses et discussion

4.1 Corpus de dépendance

Le corpus utilisé pour cette analyse par transition est constitué d'un ensemble de corpus de phrases de style grammatical différent. Il a été annoté semi-automatiquement par des membres de l'équipe TALN du Lina et leurs associés, en utilisant le système *CDG Lab* (Alfared *et al.*, 2011) et une grammaire catégorielle du français (Dikovsky, 2011). Lors de chaque analyse syntaxique 90% du corpus est employé comme corpus d'entraînement pour produire un oracle et les 10% restants sont exploités par l'analyse syntaxique par transition et traduits en structures de dépendances. Le choix des phrases utilisées pour l'apprentissage et pour l'analyse se fait aléatoirement. Le corpus complet (apprentissage et analyse) comprend 2557 structures de dépendances dont 33490 mots. 41% des structures de dépendances de ce corpus contiennent au moins une dépendance discontinue mais les dépendances discontinues représentent seulement 4% du nombre de dépendances total du corpus.

4.2 Méthode d'expérimentation

Les fichiers de sortie de l'analyseur sont similaires aux fichiers d'entrée, de manière à ce que l'on puisse retrouver pour chaque phrase, les dépendances correctements assignées ou non entre les unités lexicales. Pour estimer les résultats, on calculera la f-mesure sur les têtes des dépendances. C'est à dire que l'on se contentera de vérifier si chaque mot de la phrase reçoit bien la dépendance du bon type. Toutefois, les dépendances discontinues ne sont pas comptabilisées, car les mots recevant une telle dépendance reçoivent aussi une dépendance ancre du même type. Le choix de calculer les résultats uniquement sur les têtes des dépendances plutôt que sur les dépendances elles-mêmes dépend de deux aspects importants :

– le volume du corpus n'est pas suffisant pour l'instant pour espérer avoir les bons assignements des dépendances et de leurs types en même temps
– ces types de résultats s'accordent avec des travaux futurs qui consisteront à déterminer les bonnes têtes des dépendances avant d'effectuer une analyse syntaxique en dépendance à partir de la CDG.

Le principe est donc de collecter, pour chaque type de dépendance apparaîssant dans le corpus d'entrée ou de sortie, le nombre de fois où celui-ci est assigné dans le corpus d'entrée, le nombre de fois où celui-ci est attribué à un mot du corpus de sortie et le nombre de fois où celui-ci est attribué au bon mot dans le corpus de sortie. La table 5 donne les résultats de ces assignations sur une analyse d'un corpus d'analyse possible (10% du corpus total, voir section 4.1) pour certains types de dépendances. Par exemple, le type *S* (la tête de la phrase) a été assigné sur 38 mots, dont 37 fois correctement, parmi les 255 mots typés *S* dans le corpus. Le type *DET* a toujours été assigné correctement sur les 117 fois où il a été trouvé pour un mot.

Type des dépendance	Nb d'occurences corpus d'entrée	Nb d'occurences corpus de sortie	Nb d'attributions correctes
S	38	255	37
PRED	169	314	168
DET	117	331	117
OBJ	69	218	64
COPUL	22	79	21
MODIF	26	113	24
CLIT	48	124	39
PUNCT	116	404	109

TABLE 5 – Exemple de résultats d'assignements des types (pour quelques types donnés) par comparaison entre un corpus d'entrée et de sortie

Pour chaque type de dépendance i de la table ainsi établie, on calcule la présicion et le rappel de la manière suivante :

$$\text{précision}_i = \frac{\text{Nb d'attributions correctes pour i}}{\text{Nb d'occurences corpus de sortie pour i}} \qquad \text{rappel}_i = \frac{\text{Nb d'attributions correctes pour i}}{\text{Nb d'occurences corpus d'entrée pour i}}$$

4.3 Résultats des expérimentations

Les valeurs de la f-mesure calculées selon la méthode décrite à la section 4.2 sont indiquées dans la table 6 en fonction de chaque oracle (voir la table 4 pour les caractéristiques des oracles). Les résultats sont une moyenne mesurée sur 20 itérations comprenant chacune la séparation du corpus en un corpus d'entraînement et un corpus de test, l'entraînement de l'oracle, l'analyse syntaxique par transition et le calcul de la f-mesure.

L'analyse des résultats montre que le choix du vecteur de traits pour l'oracle influe de manière significative sur l'attribution correcte ou non des types de dépendances sur les mots. Les meilleurs résultats se font avec les oracles dont les vecteurs de traits sont les plus informatifs. Le choix d'une classe grammaticale longue et d'une taille de vecteur de 8 est préférable pour obtenir de meilleurs résultats. En outre, le choix d'utiliser les groupes à la place des types exacts semble dépendre de la longueur de la classe grammaticale. Ces deux paramètres ont en effet un lien. Par exemple, si un mot de classe grammaticale N^3 à reçu le type de dépendance *a-obj*, en utilisant

3. voir section 3.4

	Oracle 1	Oracle 2	Oracle 3	Oracle 4
Précision	0.373	0.282	0.567	0.615
Rappel	0.187	0.151	0.258	0.294
F-mesure	0.250	0.197	0.355	0.398
	Oracle 5	Oracle 6	Oracle 7	Oracle 8
Précision	0.463	0.393	0.635	0.697
Rappel	0.192	0.169	0.313	0.362
F-mesure	0.272	0.237	0.420	0.477

TABLE 6 – Résultats des expérimentations sur les sorties de l'analyse selon les différents oracles

l'oracle 6, alors qu'il est de type *a-obj-d*, la dépendance sera considérée comme mal attribuée. Alors qu'avec l'oracle 5, la dépendance sera seulement *OBJ* dans les deux cas et sera donc correcte pour ce mot. Inversement, le problème se pose d'une autre manière entre les oracles 7 et 8. Lors de l'entraînement des oracles, la transition choisie dépend de la fréquence d'apparition d'une configuration dans le corpus. Un mot de classe grammaticale *Ntime* peut recevoir une dépendance de type objet ou préposition par exemple. Il se pourrait alors qu'en utilisant les groupes, l'ensemble des dépendances de type *PREPOS* soient plus nombreuses que l'ensemble des dépendances de types *OBJ* alors que sans utiliser les groupes, les dépendances de type *a-obj* soient plus nombreuses que chacune des dépendances de type *prepos-x* (où x=g|d|l|o|A|sel). Donc dans un cas, l'oracle assignerait une dépendance de type préposition alors que dans l'autre, il assignerait une dépendance de type objet.

En outre, la taille du vecteur de trait est significative dans le sens où les options ajoutées au vecteur de taille 8 apportent une information importante lors de l'analyse. En effet, de cette manière on connait la dépendance la plus à gauche et la dépendance la plus à droite de chacun des deux mots traités (le mot en haut de la pile σ et le premier mot du tampon β). On peut savoir par exemple si un verbe, étant la tête de la phrase, a déjà un sujet ou non. Alors qu'avec un vecteur de trait de taille 6, cette information n'est pas conservée et plusieurs sujets peuvent être attribués dans une même phrase lorsqu'il n'y a pas lieu de le faire.

4.4 Perspectives et améliorations à venir

Le choix d'une transition lors de l'analyse se fait de manière déterministe car, dans l'oracle, il n'y a qu'une seule transition possible pour une configuration donnée. En outre, plus les options choisies pour l'oracle sont généralistes, moins il y a de choix car des configurations différentes se retrouvent confondues dans un même vecteur de traits qui dirige l'analyse vers une seule transition pour des cas très différents. Il sera donc nécessaire par la suite de mémoriser dans l'oracle, en plus de l'ensemble configuration-transition le meilleur, quelques autres possibilitées de transition pour la configuration considérée.

De plus, l'analyse étant déterministe, celle-ci s'interrompt lorsqu'il n'y a pas de transition applicable. Beaucoup de structures de dépendances résultantes sont alors incomplètes ou parfois complétement vides. La proposition précedente, permettrait donc ensuite de pouvoir faire un

retour en arrière dans la séquence de transitions. Le nombre de retour en arrière possible pour une phrase devra être limité et le choix d'une nouvelle transition pourra se baser sur la seconde meilleure transition pour une certaine configuration. Ce problème est aussi une cause du faible score du rappel dans toutes les analyses. En effet, beaucoup de mots n'ont tout simplement pas d'attribution de type alors que les mots ayant une dépendance sont plus souvent correctement typés, d'où une précision meilleure que le rappel (voir exemple table 5).
Le choix d'une première transition puis d'une potentielle seconde meilleure transition devra se baser sur des données probabilistes. Ainsi, pour établir ces données il sera nécessaire d'expérimenter différents lissages avant de mémoriser les ensembles configuration-transitions dans l'oracle.

D'après les résultats des expérimentations, les oracles utilisant un vecteur de taille 8 sont plus efficaces car ils mémorisent les types des dépendances gauches et droites des deux mots traités lors de l'analyse. Cependant, ces informations ne sont pas pertinentes dans tous les cas. Notamment, le fait de savoir qu'un mot a une dépendance de type modificateur n'est pas déterminant pour la suite car rien n'empêche ce mot d'avoir d'autres modificateurs. Tous les types itérables, comme les modificateurs, les circonstantiels, les coordinations verbales, assignés à une dépendance ne seront donc pas mémorisés. Nous garderons uniquement des informations appropriées telles que les types prédicat, déterminant, objet, etc...

5 Conclusion

Les objectifs de cet article étaient de proposer des résultats d'expérimentations avec un analyseur syntaxique par transitions étendu aux dépendances discontinues et de préparer la base d'une comparaison de ces résultats avec un oracle guidant l'analyse de la CDG du français.
L'analyse par transition décrite dans cet article ne permet pas encore d'obtenir des résultats suffisants pour obtenir une bonne analyse en dépendance du français. Cependant, le système de transitions mis en place pour correspondre avec la grammaire CDG est tout à fait adaptée à l'ajout des dépendances discontinues par valences polarisées. De plus, nous voulions mettre en évidence les intérêts et les problèmes de ces premières expérimentations sur des phrases en français pour mettre en place un plan d'améliorations de l'analyse par transitions. Dans l'état actuel, l'algorithme établi pour analyser les phrases en structures de dépendances est déterministe et ne permet pas de faire des choix bien adaptés aux configurations traitées. L'aspect probabilistes mis en place pour calculer la fréquence d'apparition des configurations dans l'analyse des phrases du corpus devra être exploité pour sélectionner les meilleures transitions possibles pour une configuration et pas seulement la meilleure. En outre, les vecteurs de traits devront être mieux adaptés à l'information dont l'analyseur a besoin pour faire les bons choix.
Dès lors, les prochains résultats permettront de comparer cette analyse avec l'oracle qui guide l'analyse de la CDG du français et ainsi d'établir de nouvelles méthodes d'amorçages pour attribuer les têtes des dépendances avant l'analyse de la CDG.

Références

ALFARED, R., BÉCHET, D. et DIKOVSKY, A. (2011). "CDG Lab" : a toolbox for dependency grammars and dependency treebanks development. *In DEPLING 2011 (International Conference on Dependency Linguistics)*, Barcelona.

BAR-HILLEL, Y., GAIFMAN, C. et SHAMIR, E. (1964). *On categorial and phrase structure grammars*, pages 99–115. Addison-Wesley.

BRUNET-MANQUAT, F. (2005). Improving dependency analysis by syntactic parser combination. *In NLP-KE 2005 (Natural Language Processing and Knowledge Engineering)*, Wuhan.

CANDITO, M., CRABBÉ, B., DENIS, P. et GUÉRIN, F. (2009). Analyse syntaxique du français : des constituants aux dépendances. *In TALN 2009 (Traitement automatique des langues naturelles)*, Senlis.

CANDITO, M.-H. et KAHANE, S. (1998). Une grammaire tag vue comme une grammaire sens-texte précompilée. *In TALN 1998 (Traitement automatique des langues naturelles)*, Paris.

CHOI, J. D. et PALMER, M. (2011). Getting the most out of transition-based dependency parsing. *In ACL 2011 (Association for Computational Linguistics)*, Portland.

DEKHTYAR, M. et DIKOVSKY, A. (2008). *Generalized Categorial Dependency Grammars*, pages 230–255. LNCS 4800. Springer.

DIKOVSKY, A. (2011). Categorial dependency grammars : from theory to large scale grammars. *In DEPLING 2011 (International Conference on Dependency Linguistics)*, Barcelona.

KAHANE, S. (2001). Grammaires de dépendance formelles et théorie sens-texte. *In TALN 2001 (Traitement automatique des langues naturelles)*, Tours.

KOW, E., PARMENTIER, Y. et GARDENT, C. (2006). Semtag, the loria toolbox for tag-based parsing and generation. *In TAG+8 (The Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms)*, Sydney.

KÜBLER, S., MCDONALD, R. et NIVRE, J. (2009). *Dependency parsing*. Morgan et Claypool.

MARUYAMA, H. (1990). Structural disambiguation with constraint propagation. *In ACL 1990 (Association for Computational Linguistics)*, Pittsburgh.

MEL'CUK, I. (1988). *Dependency syntax : Theory and Practice*. State University of New York Press.

NASR, A. (2004). *Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement*. Habilitation à diriger des recherches, Université Paris 7, UFR d'informatique.

NIVRE, J. (2008). *Algorithms for Deterministic Incremental Dependency Parsing*, pages 513–553. Volume 34 (4). Massachusetts Institute of Technology.

NIVRE, J. (2011). Bare-bones dependency parsing. *In NODALIDA 2011 (Nordic Conference of Computational Linguistics)*, Riga.

RAMBOW, O. (2010). The simple truth about dependency and phrase structure representations. *In NAACL HTL 2010 (North American Chapter of the Association for Computational Linguistics - Human Language Technologies)*, Los Angeles.

VANRULLEN, T., BLACHE, P. et BALFOURIER, J.-M. (2006). Constraint-based parsing as an efficient solution : Results from the parsing evaluation campaign easy. *In LREC 2006 (Conference on Language Resources and Evaluation)*, Genoa.

Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu

Marion Baranes[1,2]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris
(2) viavoo, 69 rue Danjou, 92100 Boulogne Billancourt
marion.baranes@viavoo.fr

RÉSUMÉ

Dans ce papier, nous introduisons le problème que pose la correction orthographique sur des corpus de qualité très dégradée tels que les messages publiés sur les forums, les sites d'avis ou les réseaux sociaux. Nous proposons une première architecture de correction qui a pour objectif d'éviter au maximum la sur-correction. Nous présentons, par ailleurs l'implémentation et les résultats d'un des modules de ce système qui a pour but de détecter si un mot inconnu, dans une phrase de langue connue, est un mot qui appartient à cette langue ou non.

ABSTRACT

Towards Automatic Spell-Checking of Noisy Texts : General Architecture and Language Identification for Unknown Words

This paper deals with the problem of spell checking on degraded-quality corpora such as blogs, review sites and social networks. We propose a first architecture of correction which aims at reducing overcorrection, and we describe its implementation. We also report and discuss the results obtained thanks to the module that detects whether an unknown word from a sentence in a known language belongs to this language or not.

MOTS-CLÉS : Correction automatique, détection de langue, données produite par l'utilisateur.

KEYWORDS: Spelling correction, language identification, User-Generated Content.

1 Introduction

Les outils de traduction, d'extraction de sentiments ou encore de fouille de textes sont de plus en plus utilisés. La majorité de ces outils s'appuient sur des corpus relativement propres. Si une personne choisit de travailler sur des données plus altérées, rédigées sur le web par exemple, sa tâche se complexifie. Il est donc important de pouvoir nettoyer ces textes afin d'appliquer par la suite les traitements voulus. Être capable de normaliser et de corriger automatiquement devient alors un réel besoin. Un besoin dans la mesure où ces derniers n'altèrent pas la qualité du texte. Un correcteur qui ferait de la sur-correction (qui corrigerait par exemple tous les mots inconnus) engendrerait la perte de nombreuses informations. Cette correction n'a pas besoin d'être parfaite. Dans le cadre de la fouille de texte par exemple, les informations détectées sont très souvent faites à l'aide de mots-clefs et de grammaires locales qui ne prennent pas forcément la flexion des mots en compte. Si une faute d'accord n'est pas corrigée, l'information sera tout de même détectée. Il vaut donc mieux un outil qui sous-corrige un texte plutôt qu'un qui le sur-corrige.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 95–108,
Grenoble, 4 au 8 juin 2012.

La correction orthographique n'est pas un sujet nouveau. Les travaux qui y font référence sont nombreux (voir section 2). Toutefois, ils sont rarement adaptés aux types de textes de qualité parfois très dégradée que l'on analyse lorsque l'on traite des données réelles dites « produites par l'utilisateur » (*User-Generated Content*) comme les blogs, les forums ou encore, les réseaux sociaux. Notre objectif est de combler ce manque, en développant une architecture et des technologies dédiées à la normalisation orthographique et typographique automatisée de corpus réels très bruités. Pour ce faire, nous nous appuierons sur un corpus provenant du web uniquement composé de messages clients (voir section 3) dont tous nos exemples sont tirés. Ce projet étant en cours, nous avons déjà une idée des résultats que nous souhaitons obtenir. L'architecture de correction est déjà en partie définie et est en cours d'implémentation. Elle est composée de nombreux modules qui, au fur et à mesure, amélioreront la qualité du texte et détermineront s'il convient ou non de corriger un mot inconnu (un mot inconnu pouvant correspondre à une faute d'orthographe, une entité nommée, un mot étranger, un emprunt ou encore un néologisme). Ces modules permettront, entre autres, de limiter les cas de sur-correction. Dans cet article, nous nous concentrerons sur l'un de ces modules dont le but est de déterminer si, dans un texte dont on connaît la langue, un mot inconnu correspond à un mot étranger ou non.
Cet article est structuré comme suit. Nous commencerons par dresser un état de l'art du domaine (section 2). Puis, nous décrirons les objectifs et principes généraux de correction que nous comptons implémenter (section 3). Nous préciserons dans cette même section les premières étapes préalables au module décrit en section 4 et pour lequel nous disposons de résultats évalués. Enfin, nous ferons un point sur le travail réalisé et sur nos perspectives (section 5).

2 État de l'art de la correction orthographique automatique

La mise en place d'un système de correction dépend beaucoup du type de corpus que l'on veut corriger. C'est pourquoi nous ferons état de la grande diversité de ces derniers dans le paragraphe qui suit avant de dresser un panorama des travaux réalisés dans le domaine.
Les correcteurs automatiques ne tendent pas à corriger les mêmes types de fautes, notamment en fonction de la provenance de ces corpus. Ainsi, un texte retranscrit par un processus de reconnaissance optique des caractères (OCR) ne contiendra pas les mêmes erreurs qu'un texte journalistique, qu'un texto (ou SMS), un mail ou encore un message posté sur un médias social du web (réseaux sociaux, forums, etc.). À tous ces canaux, correspondent des documents qui divergent en fonction de leurs tailles, de leurs contenus, de leurs objectifs, du type de vocabulaire utilisé (spécialisé ou non, familier ou soutenu), ou encore de l'aisance qu'a le locuteur avec la langue utilisée (écrit-il dans sa langue maternelle ou non ?). De ce fait, on peut supposer que les fautes produites dépendront aussi de ces critères et ne seront pas systématiquement de même nature. Par exemple, un texte OCR a plus de chance de contenir des fautes liées à la similarité typographique « graphique » (« *l* » vs « *1* ») tandis qu'un message de forum contiendra plus probablement des erreurs de proximité phonétique et/ou typographique (proximité des lettres sur un clavier). Le type de fautes à corriger variera ainsi en fonction des corpus sélectionnés.
On distingue généralement les deux types d'erreurs suivants dans les textes : les fautes lexicales et les fautes grammaticales (Kukich, 1992). Sont placés dans la catégorie des fautes lexicales tous les mots qui ne figurent pas dans le dictionnaire (voir exemple 1a), contrairement aux fautes grammaticales qui ne peuvent être détectées que si le contexte est pris en compte (voir exemple 1b). Ces deux types d'erreurs sont souvent traités séparément dans la littérature.

(1) a. j'**aimrai resrever** 2 billets

b. **ces** nul tu captes pas en montagne je le sais car je **lait**

À ses débuts, la correction lexicale se faisait indépendamment du contexte et s'appuyait notamment sur des règles de correction typographique (suppression, ajout d'un caractère, substitution d'une lettre avec une autre et inversion de deux lettres) à effectuer afin d'obtenir un mot correctement orthographié (Damerau, 1964; Kernighan *et al.*, 1990). Néanmoins on a rapidement réalisé que cette technique ne suffisait pas. Par exemple, si on observe la phrase : « *J'espère que vous* **réallisez** *que vos produits sont* **tjs hor** *de* **pri**... », on constate que seul « *réallisez* » a de réelles chances d'être bien corrigé. Le mot « *pri* » pourrait l'être aussi mais ses corrections possibles sont nombreuses, donnant lieu à une ambiguïté : doit-il être corrigé par « *prie* », « *pris* », « *prit* » ou « *prix* » ? Il en est de même pour « *hor* ». D'autres solutions furent donc proposées par la suite.

Dès les années 1990, Kukich (1992) publie un panorama des diverses techniques existantes de l'époque. Beaucoup de travaux, y compris de très récents, ont choisi d'utiliser des modèles de langage *n*-grammes afin de prendre en compte le contexte du mot à corriger. Ces *n*-grammes sont généralement composés de tokens (Brill et Moore, 2000; Carlson et Fette, 2007; Park et Levy, 2011) mais cette solution n'est pas parfaitement satisfaisante. Si le contexte du mot à corriger est mal orthographié, ce qui est le cas dans l'exemple proposé ci-dessus, une solution intermédiaire serait alors d'utiliser des *n*-grammes phonétiques (Toutanova et Moore, 2002) ou de prendre en compte ces deux types de *n*-grammes (Boyd, 2009). De cette manière, les fautes d'orthographe ne modifiant pas la phonétique d'un mot ne pourraient pas altérer les résultats du correcteur : le mot « *hor* » de notre exemple ne gênerait pas la correction de « *pri* » et inversement. Ces modèles peuvent être associés ou non à un modèle d'erreur et s'appuient généralement sur de nombreux paramètres supplémentaires tels que la position d'une erreur dans un mot, la catégorie du mot à corriger, sa longueur ou encore sa phonétique. Une autre approche consiste à prendre en compte la mesure de similarité distributionnelle qui existe entre une phrase contenant une erreur et ses candidats de correction possibles (Li *et al.*, 2006).
Par ailleurs, avec l'essor des nouvelles formes de communication, de nouvelles approches ont été proposées pour traiter le langage texto (ou SMS) : en passant par la phonétisation du texte à corriger (Kobus *et al.*, 2008), en ajoutant des ressources lexicales à un correcteur lexical dit classique (Guimier De Neef et Fessard, 2007) (ce qui permet de systématiser certaines corrections comme « *tjs/toujours* ») ou encore en s'appuyant sur des modèles entraînés sur des textes bruités alignés avec leur contrepartie nettoyée (Beaufort *et al.*, 2010).
Enfin, l'émergence d'internet a eu des répercussions sur plusieurs techniques de correction qui l'ont considéré comme une source d'informations pertinente. C'est, par exemple, le cas de Chen *et al.* (2007) qui proposent une méthode enrichie par des résultats de requêtes, produites sur des moteurs de recherche.
Corriger tous les mots qui ne figurent pas dans le dictionnaire nécessite par ailleurs de détecter les entités nommées, les néologismes ou encore les emprunts. Les quelques approches qui ont choisi de traiter ce problème reposent généralement soit sur des lexiques spécifiques (Beaufort *et al.*, 2010; Kobus *et al.*, 2008) soit sur le contexte (Li *et al.*, 2006). Par exemple, Han et Baldwin (2011) proposent d'utiliser un classifieur qui s'appuie sur les paramètres de dépendances syntaxiques reliant le mot susceptible d'être mal orthographié aux mots présents dans son contexte afin de déterminer s'il doit être corrigé ou non.

Comme nous l'avons dit précédemment, les fautes grammaticales donnent lieu à des mots existants dans le dictionnaire (par exemple : *conseil/conseille* ou *ai/est/et/hait/...*). Pour les

corriger, s'appuyer sur le contexte est donc inévitable. Si on prend la phrase « *Enfin un objet qu'***ont** *peut emporter partout avec* **sois** », on ne pourra corriger le mot « *ont* » en « *on* » qu'à condition de prendre en compte ses mots voisins. Pour cela plusieurs méthodes ont été proposées comme utiliser des systèmes de règles (Mangu et Brill, 1997), faire de la classification (Rozovskaya et Roth, 2010), mettre en place des modèles *n*-grammes contenant des catégories grammaticales associées à des paramètres contextuels (Golding et Schabes, 1996) ou encore des modèles *n*-grammes de tokens. C'est cette dernière solution, souvent combinée à de gros corpus, à des modèles d'erreurs, à différents types de paramètres, et/ou à des mesures de similarité, qui est la plus utilisée (Carlson et Fette, 2007; Islam et Inkpen, 2009; Stehouwer et van Zaanen, 2009; Gao *et al.*, 2010). Certaines études proposent aussi la combinaison de ces différentes approches. C'est par exemple le cas de Xu *et al.* (2011) qui utilisent un modèle trigramme et un classifieur lors des différentes étapes de leur correcteur. La majorité de ces techniques, bien que différentes, fonctionne avec la même logique. Elles tentent dans un premier temps de détecter une erreur puis, créent ensuite une liste de candidats de correction possibles pour enfin choisir la correction la plus probable.

3 Architecture du correcteur orthographique envisagé

3.1 Objectifs généraux

Cette thèse vise à mettre en place un correcteur capable de corriger des textes provenant du web. Ce sont des textes dont la taille, le contenu, la langue et la qualité rédactionnelle sont variables. En fonction du canal choisi par l'internaute et de l'internaute lui-même, le message sera plus ou moins bruité. De plus, si on observe plus attentivement le contenu de certains d'entre eux (voir l'exemple 2), on constate qu'on ne peut se restreindre à un module de correction uniquement lexical ou grammatical.

(2) regardée vraimen se don vous avez besoin et ne vous fait pas avoir par leurre pub com quoi il se souci des leur clientèle allée voir vite ailleur

Peu d'approches procèdent à une correction à la fois lexicale et grammaticale (cf. cependant Carlson et Fette (2007)). Dans cette optique, nous envisageons un correcteur contextuel. Les différentes études faites sur la correction se cantonnent souvent à une seule méthode, proposant ensuite de faire varier la valeur de certains paramètres afin d'optimiser les résultats de leur correcteur. Ne sachant quelle est la meilleure approche, nous songeons à aborder le problème en comparant et combinant ces dernières. Nous nous attarderons donc aussi bien sur des systèmes utilisant des *n*-grammes (phonétiques ou non) que sur des systèmes par règles ou sur des systèmes d'alignement automatique. Nous n'écarterons pas non plus la possibilité de nous appuyer sur des ressources lexicales plus adaptées à nos corpus ou d'inclure des informations propres à la sémantique distributionnelle. De cette manière nous pourrons ensuite choisir la combinaison la plus performante.

Bien que les travaux décrits dans cet article ne concernent que le français, nous voudrions, à terme, un correcteur qui puisse corriger plusieurs langues. Par conséquent, il ne s'agit pas de ré-implémenter des travaux adaptés à l'anglais, mais de trouver la bonne combinaison qui nous permettra de traiter de manière efficace et indépendante plusieurs langues différentes (telles que l'anglais, le français, l'allemand, l'espagnol ou encore l'italien). Le caractère multilingue que peut avoir un correcteur n'apparaît que très peu dans la littérature (cf. cependant Reynaert (2004)).

Le corpus que nous utilisons comme objet d'étude est uniquement constitué de messages client très divers les uns des autres. Cette hétérogénéité s'explique par le fait qu'ils proviennent de canaux différents (réseaux sociaux, forums, sites d'avis, mails, blogs, enquêtes de satisfaction). Après une rapide étude sur corpus, nous avons pu constater que le nombre et le type de faute d'un message varie en fonction du canal utilisé. Nous voudrions donc pouvoir adapter de manière automatique et dynamique notre correction au texte à corriger et ainsi être capable, par exemple, d'en proposer une plus légère pour un mail que pour un message provenant d'un réseau social.

3.2 Prétraitements nécessaires

Actuellement, nous n'avons pas encore défini tous les détails de l'architecture de notre système de correction. Nous ne donnerons donc pas d'indications supplémentaires à ce sujet. Néanmoins, le correcteur que nous voulons mettre en place étant contextuel, la qualité du texte l'entourant aura automatiquement des conséquences sur ses performances. Et ce, même si cette dernière est prise en compte dans l'implémentation du système. Nous proposons donc de procéder à une série de prétraitements qui ont pour but de normaliser en grande partie le texte, d'y détecter les mots ou groupes de mots que l'on pourrait ignorer pendant la suite du traitement ou encore de pré-corriger certains mots. Comme le montre la figure 1 ces prétraitements se partitionnent en plusieurs modules.

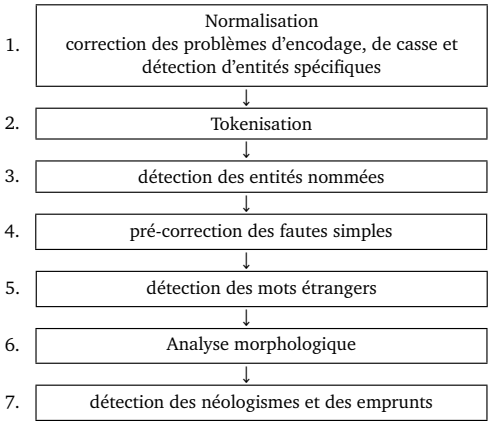


FIGURE 1 – Architecture des prétraitements nécessaires

Les modules 1, 2 et 3 visent à nettoyer le texte de toutes ses erreurs d'encodage, des fautes de casse [1] et tend à détecter les entités nommées [2] contenant de la ponctuation ou d'autres caractères spéciaux (URL, smiley, adresse mail, date,...). La reconnaissance de ces entités a lieu à

1. La correction des fautes de casse consistera entre autres à ajouter une majuscule manquante en début de phrase ou encore à renormaliser les phrases entièrement écrites en majuscules.
2. Nous reprenons ici la notion d'entité nommées telle qu'elle est utilisée dans Sagot et Boullier (2008)

cet endroit afin d'éviter certains cas qui pourraient induire le tokeniseur en erreur. Le module de détection des entités nommées (module 3), quant à lui, fait référence aux autres types d'entités nommées tels que les noms propres ou encore les acronymes. Puisque nous ne traitons pour l'instant que des textes en français, ces premiers modules peuvent être en grande partie gérés par SxPipe (Sagot et Boullier, 2008). [3]
Nous souhaitons ensuite (dans le module 4) proposer une pré-correction aux fautes aisément détectables dont la correction ne laisserait place à aucune ambiguïté (ex : *nooooon* → *non*).
Enfin, il est important de pouvoir définir, lorsque l'on rencontre un mot inconnu non intégré à une entité nommée, si ce mot appartient à la langue du texte que l'on souhaite corriger ou non (module 5). Si c'est le cas, il sera analysé morphologiquement (module 6) afin de déterminer s'il correspond, ou non, à une faute, un emprunt, un néologisme ou encore à une entité nommée qui n'aurait pas été détectée dans les modules précédents (module 7). Dans le cas contraire, il y a fortes chances pour que ce mot soit un mot dit « étranger ». Il sera donc soit laissé tel quel soit corrigé avec un faible coût de correction (ex : faute d'accent). C'est uniquement une fois que tous ces prétraitements auront été mis en place que nous pourrons nous concentrer sur notre système de correction.

4 Détecter automatiquement l'appartenance d'un mot à une langue

4.1 Présentation du module

L'un de nos modules de prétraitement (module 5) doit donc déterminer si un mot, inconnu d'un lexique de référence du français et non traité par les modules précédents, correspond à un mot « étranger ». Nous définissons cette notion de mot « étranger » de façon opérationnelle comme suit. Un mot inconnu est « étranger » si, notamment parce qu'il est un emprunt à une autre langue n'ayant pas été adapté morphologiquement, il ne doit pas faire l'objet d'une correction orthographique par un outil traitant du français. À l'inverse, les inconnus qui sont catégorisés comme « français » seront notamment des fautes lexicales dont on sera amené à essayer de corriger l'orthographe, ou des néologismes (y compris des emprunts adaptés). Par ailleurs, après avoir observé les mots inconnus qui apparaissaient dans nos corpus, nous avons constaté que la quasi-totalité des mots que nous voudrions annoter comme « étrangers » sont en réalité des mots anglais (cf. exemples 3c et 3d et table 2). Nous allons appuyer notre module sur l'approximation suivante : l'identification des inconnus « étrangers » pourra se faire à l'aide d'approches cherchant à distinguer d'une part des inconnus *anglais* et d'autre part des inconnus français [4]. Considérons les exemples suivants :

(3) a. Il y a **marqer ke** la carte n'est pas disponible

b. J'ai tellement débranché pour **rebooter**, **reseter** que je pourrais le faire les yeux fermés

c. y a-t'il la possibilité de les utiliser **online** ?

d. **OMFG** si **no fake go** sinon joli montage

3. Cette chaîne traite déjà d'autres langues que le français, mais elle devra être améliorée avant que nous ne l'utilisions pour traiter d'autres langues.

4. Parmi les 36 000 mots parcourus pour annoter nos mots étrangers, nous n'avons trouvé que des mots anglais.

Les mots en gras de l'exemple 3a correspondent à des mots mal orthographiés. Quant à ceux de l'exemple 3b, ils pourraient plutôt être considérés comme des emprunts adaptés. Dans ces deux cas, nous considérons que ces mots appartiennent à la langue française. Nous voudrions, par la suite, pouvoir les analyser morphologiquement afin de déterminer s'il s'agit d'emprunts adaptés, de fautes d'orthographe ou encore de néologismes. Les mots en gras des exemples 3c et 3d sont des mots anglais, donc « étranger ». Ils détiennent un sens particulier dans ces phrases et nous ne voudrions surtout pas tenter de les corriger en les considérant comme des erreurs produites par l'internaute.
Cette tâche peut sembler similaire aux travaux réalisés dans le domaine de la détection de langue. La détection de langue s'appuie souvent, soit sur des connaissances linguistiques, soit sur des méthodes statistiques (Grefenstette, 1995; Giguet, 1998). Plusieurs approches permettent de s'appuyer sur des connaissances linguistiques. On peut choisir de n'exploiter que les mots grammaticaux ou les mots les plus courts d'une langue (Johnson, 1993; Ingle, 1976), de s'appuyer sur un lexique de chaque langue, ou encore de ne prendre en compte que les suites de lettres qui n'apparaissent que dans une langue (Dunning, 1994). En parallèle, on peut aussi choisir de s'appuyer sur des méthodes plus statistiques telles que les *n*-grammes simples (Cavnar et Trenkle, 1994; Martins et Silva, 2005), en utilisant l'entropie relative (Sibun et Reynar, 1996) ou associés à des modèles de Markov (Dunning, 1994). Par exemple, sur des données issues du Web, qui sont plus proches de nos données que des textes plus littéraires, Martins et Silva (2005) rapportent une précision variant de 80% à 100% pour la classification de pages parmi 12 langues en utilisant les *n*-grammes ainsi que quelques heuristiques complémentaires. Bien que notre module a une tâche similaire aux travaux décrits ci-dessus, il se distingue de ces derniers de part le fait qu'il ne veut non pas connaître la langue d'un texte, mais uniquement celle d'un mot isolé au sein d'un texte dont la langue est connue. Nous ne pouvons donc prendre en compte le contexte de ce mot.

4.2 Mise en place du système de classification

Les systèmes de classification permettent de prédire la valeur/classe d'un objet à partir d'un ensemble de données. Dans notre cas, nous voulons prédire la classe (mot français ou mot étranger) d'un mot inconnu. Pour mettre en place notre système il nous faut :
– définir un corpus duquel on extraira diverses informations (par exemple, des fréquences, divers types de traits, etc.) ; ces informations serviront de données d'entraînement ; ces données dépendent à la fois du corpus choisi et de la façon dont elles en sont extraites ;
– sélectionner un système de classification permettant d'apprendre un modèle probabiliste à partir des données d'entraînement ;
– mettre en place un corpus d'évaluation représentatif de nos données réelles qui nous permettra d'évaluer notre module.

4.2.1 Construction des données d'entraînement

Pour mettre en place notre module, nous avons besoin de choisir un corpus d'entraînement pour chaque classe de notre classifieur. Il nous faut donc un corpus de mots français et un corpus de mots anglais (qui correspond à la classe des mots étrangers). L'extraction des données d'entraînement peut se faire à partir de différents types de corpus.

– Des ressources lexicales utilisées comme corpus, qui prennent en compte les formes fléchies de chaque lemme du français et de l'anglais. Nous avons ainsi réalisé des expériences préliminaires en utilisant le lexique Le*fff* du français (Sagot, 2010) et sa contrepartie anglaise EnLex.
– Des corpus correspondant à des textes « propres », comme des corpus journalistiques, des extraits de livres ou encore des articles Wikipédia.
– Des corpus plus « bruités », dits *produits par l'utilisateur* (*User-Generated Content*). C'est par exemple le cas des corpus WaCKy Baroni *et al.* (2009), construits par l'aspiration d'un grand nombre de pages Internet, qui vont d'articles journalistiques à des messages extraits de forums. Ce type de corpus contient donc en quantité importante des textes de qualité dégradée et contenant de nombreux néologismes, plus proches de ceux que nous avons à traiter que les corpus « propres ». Des corpus WaCKy ont été constitués pour plusieurs langues. Le corpus WaCKy de référence pour le français est frWaC, celui pour l'anglais est ukWaC.

Nous avons entraîné nos systèmes de classification sur des corpus relevant de ces trois cas. Le corpus provenant de ressources lexicales nous semblait pertinent de part sa richesse en mots distincts. Néanmoins, des évaluations préliminaires sur le corpus de référence nous ont permis de constater que la présence de trop nombreux mots rares dans les données d'entraînement impactait la qualité de nos modèles. Un second corpus contenant des textes assez propres, constitué de Wikipedia (français et anglais), du corpus Brown (anglais) et du corpus de l'Est Républicain (français) s'est avéré légèrement meilleur mais insuffisant. Cela s'explique par la différence de qualité rédactionnelle présente entre les données d'entraînement et les données de référence. Nous avons obtenu de meilleurs résultats en nous entraînant sur les corpus WaCKy, plus proches des nôtres. Les systèmes de classification introduits par la suite auront, par conséquent, tous été entraînés sur ces corpus. Les mots contenus dans frWaC ne sont pas tous des mots en français : il contient également des mots inconnus, dont de nombreux mots français mal orthographiés (comme dans nos corpus) ainsi que des mots étrangers. Toutefois, nous faisons l'approximation consistant à ignorer ces derniers et à considérer frWaC comme un corpus approprié pour apprendre ce qu'est un mot « français ». Autrement dit, un mot à annoter « anglais » sera plus caractéristique de ukWaC que de frWaC, même s'il apparaît dans ce dernier.

4.2.2 Systèmes mis en place

Baseline Nous avons, dans un premier temps, implémenté un système de classification naïf qui repose simplement sur les fréquences des tokens dans frWaC et dans ukWaC. Nous faisons l'hypothèse que les mots présents dans les données que nous avons à traiter ont de fortes chances d'apparaître dans ces corpus. Notamment, un mot « français » mal orthographié a de bonnes chances d'apparaître avec la même orthographe fautive dans frWaC, mais pas dans ukWaC, ou du moins à une fréquence moindre.
Notre baseline fonctionne de la manière suivante. Lorsqu'un mot est inconnu du lexique de référence utilisé, ici le Le*fff*, on compare son nombre d'occurrences dans les corpus ukWaC et frWaK. Si ce mot est plus fréquent dans le premier, on considère qu'il est étranger, dans le cas contraire, il est annoté comme français. Si le mot inconnu n'apparaît dans aucun des deux corpus, ce système naïf lui attribue aléatoirement une des deux langues.

Système proposé Pour aller au-delà de ce classifieur naïf, nous avons défini plusieurs jeux de traits permettant de modéliser les mots présents dans nos corpus d'entraînement. Les expériences présentées ci-dessous reposent sur trois jeux de traits (illustrés dans la table 1).

– Comme indiqué plus haut, les travaux sur la reconnaissance de langue s'appuient beaucoup sur les systèmes *n*-grammes. Nous avons donc extrait des mots du frWaC et ukWaC les *n*-grammes qui les composent [5] et nous avons construit un trait booléen pour chaque *n*-gramme obtenu. Nous avons fait diverses expériences en utilisant soit une seule classe de *n*-grammes (par exemple, seulement les trigrammes), soit deux (par exemple, les bigrammes et les trigrammes).
– Nous avons également rajouté des traits booléens issus de la discrétisation du rapport entre la fréquence d'un mot donné dans le frWaC et celle du même mot dans le ukWaC [6]. L'utilisation de ces traits est indiquée par l'abréviation « freq-ratio » dans les tableaux ci-dessous.
– Un inconvénient des traits de type freq-ratio est qu'ils ne prennent pas en compte la significativité statistique du rapport de fréquences : un mot attesté une fois dans l'un des corpus et deux fois dans l'autre sera dans la même classe qu'un mot attesté 1 000 fois dans le premier et 2 000 fois dans le second. C'est pourquoi nous avons également réalisé des expériences en utilisant comme traits des classes de t-test permettant de mesurer la significativité de l'écart entre la fréquence d'un mot dans frWaC et celle de ce même mot dans ukWaC [7]. Ces traits sont indiqués par la mention « t-test » dans les tableaux ci-dessous.

Inconnu	fréq. ukWaC/frWaC	2-grammes	freq-ratio [8]	t-test[8]
access	797 734/8 898	_a, ac, cc, ce, es, ss, s_	F-R6 $(33 \leq 89 < 100)$	TT5 $(-6476 \leq -728 < -11)$
vanquish	641/51	_v, va, an, nq, qu, ui,...	F-R7 $(3 \leq 12 < 33)$	TT5 $(-6476 \leq -18 < -11)$
activié	0/27	_a, ac, ct, ti, iv, vi,...	F-R1 $(0 \leq 0 < 0,01)$	TT3 $(3,6 \leq 6,2 < 16)$
regler	12/970	_r, re, eg, gl, le, er, r_	F-R2 $(0,01 \leq 0,01 < 0,05)$	TT1 $(16 \leq 37 < 7425)$

TABLE 1 – Illustration des traits de notre module : *n*-grammes (ici bigrammes), freq-ratio et t-test

Nous avons alors construit nos données d'apprentissage en assignant à tous les mots de frWaC (resp. ukWaC) la classe « français » (resp. « anglais ») et diverses combinaisons des traits ci-dessus. Nous avons entraîné sur ces différents jeux de données d'entraînement le système de régression binomiale implémenté dans MegaM [9] (Daumé III, 2004). Chaque combinaison de traits (par exemple, bigrammes + t-test) conduit à un modèle différent. Face à un inconnu à classifier, il suffit alors d'en extraire les traits correspondant à l'un des modèles puis de calculer la prédiction de ce modèle au vu de ces traits.

4.3 Évaluation

4.3.1 Données de référence

Nous avons constitué manuellement, à partir de notre corpus, des données d'évaluation contenant des mots annotés comme « français » et comme « étranger ». Afin que ces mots soient représentatifs

5. *n* varie de 1 à 4 dans les résultats rapportés ici.
6. Cette discrétisation a été réalisée comme suit : les données conduisant à un rapport de fréquences supérieur ou égal à 1 ont été réparties en 4 classes de taille identique ; il en est de même pour les données conduisant à un rapport de fréquences inférieur à 1. On obtient donc 8 classes au total.
7. La discrétisation a été réalisée également en deux fois 4 classes, avec un t-test de 0 comme pivot.
8. Traits représentés ainsi : « Classe (x<val<y) » : Classe contenant un mot dont la valeur est comprise entre x et y.
9. http ://www.cs.utah.edu/ hal/megam/

des cas d'inconnus à traiter, nous avons conservé, pour chaque classe (« français » et « étranger »), les 564 premiers inconnus rencontrés dans notre corpus [10]. Nos données de référence, constituées par conséquent de 1 128 mots inconnus, correspondent à l'ensemble de ces inconnus annotés. Elles contiennent donc autant de mots « français » que de mots « étrangers ». Un échantillon des 15 premiers mots inconnus annotés de chaque type est représenté dans la table2.

Inconnus annotés comme « français »	*abitacle, abonment, achet, actionet, activié, additionels, adébloquée, adhére, adhére, adoore, afi, agreabl, aimmerais, aixenProvence, ala*
Inconnus annotés comme « étranger »	*access, add, advanced, advantage, adventure, adventures, after, again, agency, agreement, airline, airport, all, allOffTheLights, american*

TABLE 2 – Échantillon de mots inconnus présents dans les données de référence

Les mots inconnus annotés français correspondent en grande partie à des fautes d'orthographe et, à quelques néologismes et emprunts. Ceux annotés anglais sont plutôt des mots utilisés dans le monde du web, des noms de jeux ou films non traduits et des mots mal orthographiés. Ils correspondent aussi à des mots composant une phrase isolée en anglais au sein d'un message en français. Ces derniers apparaissent peu dans nos textes. La quantité de mots inconnus annotés comme français et étranger n'est donc pas représentative de leurs fréquences d'apparition. Néanmoins, les considérer directement comme « français » conduirait à faire de la sur-correction. Bien que ces mots n'apparaissent pas dans les dictionnaires du français, beaucoup sont présents dans les corpus WaCKy. La répartition des mots inconnus présents ou non dans ces corpus en fonction de leur annotation est représentée à la table 3.

	Annotés « français »	Annotés « étranger »	Total
Mots présents dans les corpus WaCKy	402	556	958
Mots absents des corpus WaCKy	162	8	170
Total	564	564	1 128

TABLE 3 – Informations quantitatives sur les données de référence

En constituant manuellement ce corpus, nous avons choisi d'évaluer les performances de notre système de manière isolée. Cela suppose le bon fonctionnement des étapes préalables à ce module. Il est donc évident que, si l'un de ces prétraitements génèrent des erreurs, les résultats de ce module seront impactés et seront très probablement moins bons.

4.3.2 Résultats et discussion

Cette évaluation s'appuie sur les taux d'erreurs de chacun de nos modèles avec notre corpus de référence. La présence d'un mot dans frWaC et ukWaC peut avoir de réelles conséquences sur nos résultats. Pour cette raison, nous présenterons tout d'abord les résultats obtenus avec uniquement les mots de notre corpus de référence existants dans les corpus frWak et ukWak puis, les résultats obtenus avec ceux absents [11]. Considérons la table 4 qui contient les taux d'erreurs de nos modèles en fonction des traits choisis et combinés. On constate qu'utiliser la fréquence

10. Environ 26 000 mots ont été parcourues pour trouver les inconnus « français » et près de 34 000 pour les « étrangers »
11. Les données concernant les mots de notre corpus de référence présents ou non dans les corpus WaCKy sont indiquées table 3.

d'un mot améliore considérablement nos résultats si ce mot est présent dans les corpus WaCKy. Avec les unigramme, ce taux d'erreur passe ainsi de 0,305 à 0,066. On constate, par ailleurs, que notre baseline, s'appuyant sur la fréquence des mots connus du corpus WaCKy, obtient un moins bon taux d'erreur (0,073) que ceux obtenu par notre système. Ce constat est satisfaisant puisqu'il illustre le fait que l'entraînement de nos traits *n*-grammes a un effet positif sur nos résultats.
Nous avons ensuite évalué les mots absents des corpus WaCKy (cf. la table 5). Dans ce contexte, les résultats obtenus avec nos traits de fréquence et notre baseline (50%) ne sont pas significatifs puisqu'ils s'appuient sur les mots présents de ces corpus. Les résultats des *n*-grammes seuls ne sont, quant à eux, pas surprenants puisqu'ils sont similaires à ceux des mots connus (table 4).

	n-grammes uniquement	*n*-grammes + freq-ratio	*n*-grammes + t-test
1-gramme	0,305	0,073	0,066
2-gramme	0,216	0,073	0,070
3-gramme	0,167	0,073	0,080
4-gramme	0,119	0,082	0,082
1 à 2-gramme	0,212	0,073	0,070
1 à 3-gramme	0,157	0,074	0,082
2 à 3-gramme	0,169	0,074	0,080

TABLE 4 – Taux d'erreur de nos modèles sur les mots présents dans le WaCKy

	n-grammes uniquement	*n*-grammes + freq-ratio	*n*-grammes + t-test
1-gramme	0,322	0,953	0,333
2-gramme	0,222	0,883	0,257
3-gramme	0,152	0,784	0,199
4-gramme	0,134	0,678	0,211
1 à 2-gramme	0,240	0,901	0,263
1 à 3-gramme	0,193	0,871	0,205
2 à 3-gramme	0,170	0,743	0,181

TABLE 5 – Taux d'erreur de nos modèles sur les mots absents du WaCKy

La dernière table présente les taux d'erreur obtenus lorsqu'on évalue la totalité de notre corpus de référence. Il montre que si nous utilisons uniquement des modèles *n*-grammes, nos modèles sont peu satisfaisants. Notre baseline, dont le taux d'erreur général est de 0,136, se révèle même meilleur. Cela s'explique par le fait que les corpus frWaC et ukWaCk contiennent beaucoup de mots que nous souhaitons annoter (958/1128). De plus, on constate que nos modèles purement *n*-gramme ont un taux d'erreur qui décroit au fur et à mesure que la valeur de *n* augmente. Cela est dû au fait que plus la taille d'un *n*-gramme grandit plus on a de fortes chances d'y stocker des mots entiers. Ce sont donc les approches qui prennent en compte uniquement les modèles *n*-gramme qui produisent le plus d'erreurs. Les modèles combinant les *n*-grammes et les traits de fréquence freq-ratio et t-test sont pour ces mêmes raisons plus performants. Enfin, les résultats obtenus par la combinaison des *n*-grammes et du t-test valide bien l'idée qu'avoir des valeurs statistiques plus significatives du rapport de fréquence permet d'optimiser nos résultats.
Notre modèle atteint 90% de bonnes classifications. Ces résultats sont satisfaisants dans la mesure où, si on se réfère par exemple aux résultats de Grefenstette (1995), on constate que sur un texte français restreint à un ou deux mots il obtient 69,2% de bonnes détections avec un modèle trigramme et de 30,8% de bonnes détections avec un modèle linguistique qui ne prend en compte que les mots courts de la langue.

	n-grammes uniquement	*n*-grammes + freq-ratio	*n*-grammes + t-test
1-gram	0,307	0,206	0,107
2-gram	0,217	0,195	**0,098**
3-gram	0,165	0,180	**0,098**
4-gram	0,129	0,172	0,102
1 à 2-gram	0,216	0,198	**0,099**
1 à 3-gram	0,162	0,194	0,101
2 à 3-gram	0,169	0,175	**0,095**

TABLE 6 – Taux d'erreur de nos modèles sur la totalité de nos données de référence

5 Conclusion et perspectives

À l'heure où le traitement automatique des langues s'intéresse de plus en plus aux données réelles dites « produites par l'utilisateur » (*User-Generated Content*), nous avons expliqué la nécessité d'avoir un outil de normalisation et de correction qui permettrait un meilleur fonctionnement des outils plus adaptés à des corpus propres de type journalistique. Cette tâche délicate est nécessaire lorsqu'on travaille sur des textes de qualité dégradée puisqu'elle doit manipuler le texte précautionneusement sans l'altérer en le sur-corrigeant. L'architecture modulaire de correction décrite ici vise à réduire ces risques. Bien qu'encore en cours d'implémentation, certains modules de notre correcteur sont d'ores et déjà fonctionnels. C'est le cas du module présenté à la section 4. Ce module, qui permet de détecter si un mot qui ne figure pas dans un dictionnaire de référence du français correspond à un mot dit « étranger » , obtient des résultats satisfaisants (plus de 90% de classification correcte). Et ce, d'autant plus si on prend en compte la complexité de cette tâche de par le fait que l'on ne peut s'appuyer sur le contexte des mots qui nous intéressent. La suite de nos travaux seront dans la continuité du schéma présenté (section 3). Notre prochain objectif sera donc de mettre en place un analyseur morphologique qui nous guidera dans la classification des mots inconnus annotés « français » par le module de détection des mots étrangers. Ces mots inconnus, ayant déjà été filtrés par les modules de détection d'entités nommées et de mots étrangers, il nous ne restera plus qu'à essayer de prédire s'il s'agit d'emprunts adaptés, de néologismes ou encore de fautes d'orthographe afin d'achever cette phase de prétraitements pour les mots inconnus.

6 Remerciements

Je remercie Benoît Sagot et Geoffrey Doucy (directeur R&D de viavoo) pour tous leurs conseils.

Références

BARONI, M., BERNARDINI, S., FERRARESI, A. et ZANCHETTA, E. (2009). The Wacky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

BEAUFORT, R., ROEKHAUT, S., COUGNON, L.-A. et FAIRON, C. (2010). A Hybrid Rule/Model-Based Finite-State Framework for Normalizing SMS Messages. *In Proceedings of the 48th Annual*

Meeting of the Association for Computational Linguistics (ACL'10), pages 770–779, Uppsala, Suède.

BOYD, A. (2009). Pronunciation modeling in spelling correction for writers of English as a foreign language. *In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Student Research Workshop and Doctoral Consortium*, pages 31–36, Boulder, Colorado.

BRILL, E. et MOORE, R. C. (2000). An Improved Error Model for Noisy Channel Spelling Correction. *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, Hong Kong.

CARLSON, A. et FETTE, I. (2007). Memory-based context-sensitive spelling correction at web scale. *In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA'07)*, pages 166–171.

CAVNAR, W. B. et TRENKLE, J. M. (1994). N-gram based text categorization. *In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.

CHEN, Q., LI, M. et ZHOU, M. (2007). Improving Query Spelling Correction Using Web Search Results. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 181–189, Prague, Czech Republic.

DAMERAU, F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

DAUMÉ III, H. (2004). Notes on cg and lm-bfgs optimization of logistic regression.

DUNNING, T. (1994). Statistical Identification of Language. *In Technical report CRL MCCS-94-273*, Computing Research Lab,New Mexico State University.

GAO, J., LI, X., MICOL, D., QUIRK, C. et SUN, X. (2010). A large scale ranker-based system for search query spelling correction. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 358–366, Beijing, Chine.

GIGUET, E. (1998). Méthode pour l'analyse automatique de structures formelles sur documents multilingues. *Thèse de doctorat, spécialité Informatique*.

GOLDING, A. R. et SCHABES, Y. (1996). Combining Trigram-based and Feature-based Methods for Context-sensitive Spelling Correction. *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 71–78, Santa Cruz, États-Unis.

GREFENSTETTE, G. (1995). Comparing two language identification schemes. *In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT 1995)*, Rome, Italie.

GUIMIER DE NEEF, É. et FESSARD, S. (2007). Évaluation d'un système de transcription de SMS. *In Proceedings of the 26th International Conference on Lexis and Grammar*, Bonifacio, France.

HAN, B. et BALDWIN, T. (2011). Lexical normalisation of short text messages : makn sens a #twitter. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 368–378, Portland, États-Unis.

INGLE, N. C. (1976). A language identification table. *The Incorporated Linguist*, 15(4):98–101.

ISLAM, A. et INKPEN, D. (2009). Real-word spelling correction using Google Web IT 3-grams. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1241–1249, Singapour. Association for Computational Linguistics.

JOHNSON, S. (1993). Solving the problem of language recognition. *In Technical report*, School of Computer Studies, University of Leeds.

KERNIGHAN, M. D., CHURCH, K. W. et GALE, W. A. (1990). A Spelling Correction Program Based on a Noisy Channel Model. *In Proceedings of the 13th conference on Computational linguistics (CoLing'90)*, pages 205–210, Helsinki, Finland.

KOBUS, C., YVON, F. et DAMNATI, G. (2008). Transcrire les SMS comme on reconnaît la parole. *In Actes de TALN 2008*, pages 128–138, Avignon, France.

KUKICH, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, 24(4):377–439.

LI, M., ZHANG, Y., ZHU, M. et ZHOU, M. (2006). Exploring distributional similarity based models for query spelling correction. *In Proceedings the 44th Annual Meeting of the Association for Computational Linguistics and of the 21th conference on Computational linguistics (ACL-CoLing 2006)*, pages 1025–1032, Syndey, Australie.

MANGU, L. et BRILL, E. (1997). Automatic Rule Acquisition for Spelling Correction. *In Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, pages 187–194, Nashville, États-Unis.

MARTINS, B. et SILVA, M. J. (2005). Language identification in web pages. *In Proceedings of the 2005 ACM symposium on Applied computing*, SAC '05, pages 764–768, New York, NY, USA.

PARK, Y. A. et LEVY, R. (2011). Automated whole sentence grammar correction using a noisy channel model. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 934–944.

REYNAERT, M. (2004). Multilingual Text Induced Spelling Correction. *In Proceedings of the 20th International Conference on Computational Linguistics*, Genève, Suisse.

ROZOVSKAYA, A. et ROTH, D. (2010). Generating Confusion Sets for Context-Sensitive Error Correction. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, MIT Stata Center, États-Unis.

SAGOT, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In Proceedings of LREC 2010*, La Valette, Malte.

SAGOT, B. et BOULLIER, P. (2008). SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, pages 155–188.

SIBUN, P. et REYNAR, J. C. (1996). Language Identification : Examining the Issues. *In Proceedings of SDAIR-96, the 5th Symposium on Document Analysis an Information Retrieval*, pages 125–135.

STEHOUWER, H. et van ZAANEN, M. (2009). Language models for contextual error detection and correction. *In Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference (CLAGI'09)*, pages 41–48, Athènes, Grèce.

TOUTANOVA, K. et MOORE, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 144–151, Philadelphie, États-Unis.

XU, W., TETREAULT, J., CHODOROW, M., GRISHMAN, R. et ZHAO, L. (2011). Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1291–1300, Edinburgh, Royaume-Uni.

Création d'un multi-arbre à partir d'un texte balisé : l'exemple de l'annotation d'un corpus d'oral spontané

Julie Belião
LPP - Université Paris Sorbonne Nouvelle (ILPGA) - CNRS - UMR 7018
MoDyCo - Université Paris Ouest Nanterre La Défense - CNRS - UMR 7114
julie@beliao.fr

RÉSUMÉ

Dans cette étude, nous nous intéressons au problème de l'analyse d'un corpus annoté de l'oral. Le système d'annotation considéré est celui introduit par l'équipe des syntacticiens du projet Rhapsodie. La principale problématique qui sous-tend un tel projet est que la base écrite sur laquelle on travaille est en réalité une transcription de l'oral, balisée par les annotateurs de manière à délimiter un ensemble de structures arborescentes. Un tel système introduit plusieurs structures, en particulier macro et micro-syntaxiques. Du fait de leur étroite imbrication, il s'est avéré difficile de les analyser de façon indépendante et donc de travailler sur l'aspect macro-syntaxique indépendamment de l'aspect micro-syntaxique. Cependant, peu d'études jusqu'à présent considèrent ces problèmes conjointement et de manière automatisée. Dans ce travail, nous présentons nos efforts en vue de produire un outil de parsing capable de rendre compte à la fois de l'information micro et macro-syntaxique du texte annoté. Pour ce faire, nous proposons une représentation partant de la notion de multi-arbre et nous montrons comment une telle structure peut être générée à partir de l'annotation et utilisée à des fins d'analyse.

ABSTRACT

Creating a Multi-Tree from a Tagged Text : Annotating Spoken French

This study focuses on automatic analysis of annotated transcribed speech. The annotation system considered has been recently introduced to address the several limitations of classical syntactic annotations when faced to natural speech transcriptions. It introduces many different components such as embedding, piles, kernels, pre-kernels, discursive markers etc.. All those components are tightly coupled in a complex tree structure and can hardly be considered separately because of their close intrication. Hence, a joint analysis is required but no analysis tool to handle them all together was available yet. In this study, we introduce such an automatic parser of annotated transcriptions of speech and present the corresponding framework based on multi-trees. This framework permits to jointly handle separate aspects of speech such as macro and micro syntactic levels, which are traditionnaly considered separately. Several applications are proposed, including analysis of the transcribed speech by classical parsers designed for written language.

MOTS-CLÉS : Arbres syntaxiques, unité illocutoire, unités rectionnelles, micro-syntaxe, macro-syntaxe, entassement.

KEYWORDS: Syntactic trees, illocutionary unit, microsyntax, macrosyntax, piles.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 109–122,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le projet ANR Rhapsodie (Rhapsodie, 2012) a pour but de créer un corpus de trois heures de parole transcrite, annoté en prosodie et en syntaxe, qui serve de référence pour le français parlé. Une analyse et une annotation prosodique et syntaxique sont réalisées indépendamment l'une de l'autre sur l'ensemble du corpus, de manière à permettre une étude intono-syntaxique(Benzitoun *et al.*, 2009)(Benzitoun *et al.*, 2010)(Lacheret-Dujour *et al.*, 2011). Nous nous concentrerons ici sur la tâche d'exploitation informatique du corpus annoté syntaxiquement.

La problématique qui sous-tend le système d'annotation syntaxique de Rhapsodie est que le français parlé transcrit ne présente que peu de similitudes par rapport à la syntaxe de l'écrit pour pouvoir être traité directement par des parsers syntaxiques tels que FRMG (de la Clergerie *et al.*, 2009). Les transcriptions de l'oral sur lesquelles ont travaillé les syntacticiens ne sont ni ponctuées, ni segmentées et comportent un grand nombre de phénomènes propres à l'oral tels que les disfluences, les greffes (décrites en section 3.3), etc. Ce sont ces particularités inhérentes à la transcription du français parlé qui posent problème aux parsers classiques. Les syntacticiens de Rhapsodie ont donc développé un système d'annotation syntaxique centré sur le cas de l'oral (Benzitoun *et al.*, 2009)(Benzitoun *et al.*, 2010). Leurs travaux se basent sur ceux de l'école d'Aix (Blanche-Benveniste *et al.*, 1990) et sur la syntaxe de dépendance introduite par Tesnière (Tesnière, 1959). Ce système de balisage manuel dispose de suffisamment de souplesse pour rendre compte d'un grand nombre de phénomènes relatifs à la micro et à la macro-syntaxe. Cependant, aucune représentation informatique n'était jusqu'à présent disponible pour exploiter ce formalisme de manière automatisée. Dans cette étude, nous introduisons une telle représentation et montrons comment elle peut être mise à profit pour l'analyse de la parole transcrite annotée.

Dans un premier temps, nous présenterons brièvement les différents niveaux micro et macro-syntaxiques considérés en section 2. Ensuite, nous préciserons le système d'annotation utilisé en section 3. Enfin, la notion de multi-arbre sera discutée pour sa représentation informatique en section 4 puis exploitée dans le but de procéder à une analyse automatique de la parole annotée.

2 Phénomènes micro et macro syntaxiques

2.1 Unités rectionnelles

L'approche adoptée ici est une approche *"de bas en haut"* (Benzitoun *et al.*, 2010; Lacheret-Dujour *et al.*, 2011). Une Unité Rectionnelle (UR) est une unité construite autour d'une tête, qui n'est à priori syntaxiquement dépendante d'aucun élément de rang supérieur dans le texte. La rection est caractérisée par les contraintes imposées sur une position donnée en termes de parties du discours, de marques morphologiques et de possibilités de restructuration (commutation avec un pronom, effacement, passivation, clivage, etc.). Il est important de souligner le fait que les UR ne sont pas définies dans l'absolu. C'est toujours relativement à un texte donné que l'on peut affirmer raisonnablement que certaines constructions ne dépendent d'aucune catégorie du contexte. Les UR, unités micro-syntaxiques sont souvent considérées comme les unités significatives maximales et sont définies à la fois par leur connexité rectionnelle interne et par leur autonomie externe (Berrendonner, 2002) : *"La micro-syntaxe vise à décrire des constructions syntaxiques conçues comme des ensembles rectionnels complets" (Benzitoun* et al.*, 2010)*.

2.2 Unités Illocutoires

Parallèlement à l'UR, il y a l'Unité Illocutoire (UI) dont la délimitation est liée à la reconnaissance de la force illocutoire qui peut affecter un segment dans un texte. UR et UI sont des unités relativement autonomes qui ont leurs propres règles de formation et leurs propres combinatoires. L'UI fait partie de la macro-syntaxe et *"on appelle unité illocutoire une portion de discours comportant un unique acte illocutoire, soit une assertion, soit une interrogation, soit une injonction*". (Benzitoun *et al.*, 2010)

Les syntacticiens de Rhapsodie ont considéré que ces deux modules de l'analyse syntaxique sont complémentaires mais que la sortie de l'un ne constitue pas l'entrée de l'autre. Ainsi les UI sont constituées d'UR variées, allant de l'interjection à des constructions plus complexes à plusieurs enchâssements. Les UI peuvent donc combiner plusieurs UR, mais leurs frontières ne coïncident pas forcément entre elles.

Le principe d'annotation consiste à segmenter par une balise adéquate dès que l'on ne peut plus effectuer de rattachement micro-syntaxique à l'intérieur du texte. Dans le cadre de notre étude, ce travail est effectué manuellement.

Chaque UI se décompose en un certain nombre d'unités, prosodiquement marquées — c'est du moins l'hypothèse qui est faite : (Blanche-Benveniste, 1997)(Cresti, 2000) — que l'on appelle composantes illocutoires (CI). Ces unités sont nommées suivant leur position par rapport au noyau. Le noyau (kernel) est l'UR qui comporte la force illocutoire de l'UI. Les autres UR, dépourvues de force illocutoire, s'associent au noyau et sont appelées : prénoyaux (prekernel) à gauche du noyau, in-noyaux (inkernel) dans le noyau ou post-noyaux (postkernel) à droite du noyau.

3 Balisage de la transcription

Dans cette section, nous présentons le système de balisage manuel introduit dans (Benzitoun *et al.*, 2009)(Benzitoun *et al.*, 2010) et permettant d'annoter la transcription selon les niveaux de la micro et macro-syntaxes.

3.1 Balisage des UI

– Les UI sont délimitées par le symbole **//** qui est une marque de fin d'UI [1] :
a. ***on peut après passer le concours de l'agreg pour enseigner à l'université*** **//** *(échantillon M103-Corpus Rhapsodie)*
– Par défaut, le symbole **//**, qui marque la fin d'une UI, marque aussi la fin d'une UR. Cependant, UR et UI ne se correspondent pas toujours. Le symbole **//+** (le + indique de manière générale une relation de rection) indique que l'UR se poursuit après la fin de l'UI.
b. ***c'est un Chinois //+ très riche //*** *(échantillon D210-Corpus Rhapsodie)*
– La notion de parallélisme permet d'identifier la mise en paralèlle de deux UIs. Les parallèlismes sont annotés par le signe **//=**.
c. "oh" tout est relatif //= tout est relatif // (échantillon D009-Corpus Rhapsodie)

1. Pour des raisons pédagogiques, tous les exemples donnés dans cet article sont simples et ne comportent chacun qu'une partie des phénomènes étudiés afin de les mettre en évidence. Cela-dit, il est entendu que le formalisme présenté est opérationnel et a été testé pour plus de trois heures de français parlé spontané. La plupart du temps l'ensemble des phénomènes sont réalisés simultanément.

3.2 Balisage des marqueurs d'UR et de Composante illocutoire

3.2.1 Pré-noyau, post-noyaux, in-noyaux

– Le pré-noyau, annoté **<** (ou **<+** si relation de rection).
a. ***bien évidemment <*** *c'est vrai pour la peinture religieuse en Occident // (échantillon M202-Corpus Rhapsodie)*
b. ***au début <+*** *il n'y avait pratiquement pas d' informatique // (échantillon D005-Corpus Rhapsodie)*
– Les symboles **>** et **>+** signalent les post-noyaux :
a. ***^*** *et "euh" Charlot s'est accusé* ***> plutôt que de laisser la jeune fille s'accuser*** *// (échantillon M024-Corpus Rhapsodie)*
b. ***^*** *mais vous étiez auprès des femmes* ***>+ là-bas //*** *(échantillon D204-Corpus Rhapsodie)*
– L'in-noyau est annoté par les symboles **( )** et **(+)** :
a. une rallonge à venir ***(également)*** *dans le secteur automobile // (échantillon M206-Corpus Rhapsodie)*
b. le cri de Job ***(+ que nous avons entendu dans la première lecture)*** *retentit à nos oreilles // (échantillon M203-Corpus Rhapsodie)*

3.2.2 Introducteurs

– Une UI peut commencer par un ou plusieurs introducteurs. Ces éléments ont la fonction de préciser la nature de la relation entre l'UI qu'ils introduisent et d'autres UI dans le discours (notamment l'UI qui précède). On les annote par le symbole **^**.
a. ***^donc*** *c'est pas normal qu'ils arrivent en CP ne parlant pas français // (échantillon D002-Corpus Rhapsodie)*
b. ***^et*** *tu arrives à la fontaine place Notre Dame // (échantillon M001-Corpus Rhapsodie)*
– Sont annotés avec le même symbole les marqueurs d'entassement ou joncteurs comme *et, ou, mais,* etc :
c. { les uns | ***^et*** *les autres } (échantillon M203-Corpus Rhapsodie)*

3.3 Balisage des enchâssements et parenthèses

Une UI peut se trouver à l'intérieur d'une autre UI. On distingue deux cas, les enchâssements et les insertions.
– Le discours rapporté dans cet exemple, "*casse-toi pauvre con*" forme une UI. Par contre, "*il a dit*" n'est ni une UI complète, ni une UR complète. On considère donc que "*casse-toi pauvre con*" dans "*il a dit casse-toi pauvre con*" est régi par le verbe dire.
a. il a dit ***[casse-toi > pauvre con //]*** *//*
– La greffe est la réalisation d'une UI au sein d'une UI. *"Il s'agit du procédé qui consiste à remplir une position syntaxique à l'aide d'une autre catégorie que celle attendue"* (Deulofeu, 1999). Ces deux types d'enchâssement sont annotés par des crochets et un marqueur de fin d'UI **[//]** :
b. vous avez dit que ***[disons ma carrière pour simplifier //]*** *témoigne de ma bonne conduite // (échantillon D201-Corpus Rhapsodie)*
– L'enchâssement ne contient pas toujours une UI, en effet il peut aussi contenir des sous-composantes d'une composante illocutoire (CI). Ici il s'agit d'un enchâssement d'une proposition avec un pré-noyau mais qui n'est pas une UI :
c. ce qui fait que ***[au moment de la guerre < nous étions toujours en Bretagne]*** *// (échantillon D003-Corpus Rhapsodie)*

– On parle d'insertion d'UI chaque fois qu'une UI vient interrompre momentanément une autre UI. On utilise les parenthèses simples **( )** pour délimiter l'UI insérée.
*d. "euh" et sinon < les spécialités { les m~ | un { peu moins **(je sais pas si c'est ça qui vous intéress*
***//)** | petit peu moins } } prises < "bah" { { c'est les | c'est les } spécialités à risques //+ | { la gynéc*
obstétrique (par exemple) | la cancérologie } } // (échantillon D006-Corpus Rhapsodie)

3.4 Balisage des Entassements

Les entassement font normalement partie de la micro-syntaxe, l'entassement, aussi appelé pile (voir (Gerdes et Kahane, 2009)(Kahane et Pietrandrea, 2012)(Kahane, 2012)), est un dispositif de connexion syntaxique qui relie tous les éléments qui occupent la même position syntaxique à l'intérieur de l'UR. On utilise les symboles **{** et **}** pour marquer le début et la fin de la liste et **|** pour signaler le ou les points de jonction dans le prolongement des listes paradigmatiques et de l'analyse en grille proposées dans (Blanche-Benveniste, 1990). Les conjoints ne sont pas nécessairement des constituants micro-syntaxiques mais peuvent être des disfluences par exemple :

a. ^ et **{ la | la }** Loire est en bas // *(échantillon D003-Corpus Rhapsodie)*

4 L'arbre complet et son exploitation

La réalisation d'un balisage manuel permettant d'encoder de l'information macro et micro-syntaxique implique de réaliser un parsing de ce balisage afin de pouvoir l'exploiter. Il y a trois raisons importante pour cela :

1. La création à partir du balisage d'une multi-arborescence macro et micro-syntaxique dans la transcription afin d'en extraire des arbres topologiques et d'entassement.
2. Pouvoir à partir des différents parcours de cet arbre, fournir une version dépliée de la transcription afin de faciliter la tâche de l'analyseur syntaxique automatique.
3. À partir de l'arbre initial et des résultats du parser automatique, restituer l'ordre original des mots de la transcription et procéder à leur intégration dans la structure arborescente initialement annotée.

Dans le but de réaliser ces différentes tâches, il a dont fallu implémenter un algorithme capable de réaliser ces différentes tâches.

La grammaire de balisage développée ne rentrant pas dans les cadres classiques des grammaires non-contextuelles ou même des grammaires-contextuelles d'ordre *k*, il a été nécessaire de mettre au point un parser *ad-hoc* pouvant permettre l'analyse de l'intégralité des symboles du balisage. L'objectif de cette partie est de mettre en évidence l'utilité de réunir en tant qu'objets, l'information micro, macro-syntaxique et d'entassement dans un même arbre appelé multi-arbre. Pour des raisons de place, nous ne détaillons pas ici l'agorithme de parsage, décrit dans (Beliao et Liutkus, 2012), mais nous concentrons plutôt sur l'utilité de la représentation arborée qu'il produit.

On appellera arbre une structure de données qui peut se représenter sous la forme d'une hiérarchie dont chaque élément est appelé nœud. Dans notre cas, nous avons choisi d'implémenter un parser qui construit un arbre multiple ou englobant qui représente à la fois l'information micro et macro-syntaxique. Ainsi chaque nœud correspond à un type d'unité et est typé en tant qu'unité illocutoire, entassement, enchâssement, marqueur discursif et de manière générale tout typage donné par le balisage. Ainsi, cet arbre intègre toutes les informations contenues dans le balisage.

L'implémentation d'un tel arbre présente l'avantage de pouvoir être parcouru selon un point de vue macro et micro-syntaxique ou les deux en même temps, permettant différents traitements impliquant ou non toute l'information.

Nous allons présenter en sous-section 4.1 une série d'exemples qui nous permettront de constater que le balisage peut se représenter efficacement sous la forme d'un arbre. Nous présenterons ensuite en sous-section 4.2 l'opération de dépliage, qui consiste, à partir de l'arbre, à générer un ensemble de phrases susceptibles d'être traitées par un analyseur syntaxique automatique. Nous présenterons ensuite en sous-section 4.3 l'opération que nous avons évoquée plus haut et qui consiste à extraire du multi-arbre les noeuds désirés pour en obtenir un arbre particularisé. Nous verrons que l'obtention de l'arbre topologique et de l'arbre des entassements sont des cas particuliers de cette opération de projection. Nous évoquerons ensuite la phase de repliage en section 4.4, qui consiste à réintégrer au multi-arbre l'information donnée par le parser automatique. Enfin, nous montrerons en sous-section 4.5 comment le multi-arbre arbre peut être utilisé de manière naturelle pour convertir l'information d'annotation dans un format structuré tel que XML.

4.1 Exemples d'arbres

Considérons l'exemple suivant :

a. vous avez dit que [disons ma carrière pour simplifier //] témoigne de ma bonne conduite // (échantillon D201-Corpus Rhapsodie)

Cet exemple [2] peut être représenté comme indiqué dans la figure 1. En effet, on voit que cette phrase est composée de deux UI, la deuxième étant enchâssée dans le noyau de la première par une greffe. Cette deuxième UI contient un noyau.

```
|=> root:
    |=> iu:
        |=> ru:
            |=> kernel: vous avez dit que [...] témoigne de ma bonne conduite
                |=> graft:
                    |=> ui:
                        |=> ru:
                            |=> kernel: disons ma carrière pour simplifier
```

FIGURE 1 – Représentation macro de a.

Dans cet exemple, on n'a pas encore considéré le phénomène d'entassement. Considérons donc l'exemple suivant :

b. les fêtes y sont { plus | plus } nombreuses // (échantillon D101-Corpus Rhapsodie)

On voit que cette UI contient une UR de type noyau.

Cependant, certains de ses éléments : "plus, plus" sont entassés (disfluence) selon deux couches. On observe donc par le balisage que ce segment est un entassement inclut dans un noyau. Le typage noyau correspond à une information macro-syntaxique indépendante du typage des entassements qui relève de la micro-syntaxe. On peut assimiler ces deux informations à deux niveaux ou encore deux dimensions différentes du discours et plusieurs approches sont envisageables ici.

2. Pour des raisons didactiques les numéros d'identifiant des lexèmes ont été remplacés par les token-mots dans les exemples donnés.

La première approche consisterait à représenter de manière indépendante les dépendances macro-syntaxiques (noyaux, pré-noyaux, enchâssements, etc) et les informations d'entassements sous la forme de deux arbres "projetés". Elle présenterait l'avantage d'offrir directement au spécialiste une représentation pertinente selon le point de vue désiré. Ainsi, on obtiendrait deux arbres donnés en figure 2, un premier arbre contenant l'information macro et un deuxième arbre contenant l'information des entassements.

Information macro :

```
|=> root:
    |=> iu:
        |=> ru:
            |=> kernel: les fêtes y sont plus plus nombreuses
```

Information des entassements :

```
|=> root:
    |=> ui: les fêtes y sont [...] nombreuses
        |=> pile:
            |=> layer: plus
            |=> layer: plus
```

FIGURE 2 – Abres donnant l'information macro (haut) et l'information des entassements pour l'exemple b. (bas)

Cependant, comme on le verra en sous-section 4.2, certaines taches ne sont plus réalisables si une telle disjonction est faite car l'information portée par l'un des deux typage sera perdue.

Par conséquent, la deuxième approche consiste à intégrer l'ensemble de ces informations dans la même structure, c'est à dire de considérer dans le même arbre l'information macro et micro-syntaxique. Par exemple, on peut représenter l'UI considérée par l'arbre donné en figure 3 (l'idéal étant un graphe) :

```
|=> root:
    |=> iu:
        |=> ru
            |=> kernel: les fêtes y sont [...] nombreuses
            |=> pile:
                |=> layer:
                    |=> kernel: plus
                |=> layer:
                    |=> kernel: plus
```

FIGURE 3 – Représentation macro+entassement de b.

Cet arbre, dit "multiple" contient ainsi l'ensemble des informations contenues dans le balisage, ce qui est nécessaire pour certains traitements comme on le verra en sous-section 4.2. Cependant, il a l'inconvénient pour le syntacticien de ne pas représenter de manière conventionnelle l'information syntaxique. Cela dit, il est possible de ne garder de cet arbre que l'information micro ou macro-syntaxique de manière à obtenir des représentations plus conventionnelles comme on le verra en section 4.3.

Considérons un autre exemple un peu plus complexe :

c. il faut avoir un don spécial parce que [la psychiatrie < { c' est | c' est } quelque chose] // (échantillon D006-Corpus Rhapsodie)

L'ensemble peut être représenté sous la forme de l'arbre donné en figure 4 :

```
|=> root:
    |=> iu:
        |=> ru:
            |=> kernel: il faut avoir un don spécial parce que
            |=> embedded:
                    |=> prekernel: la psychiatrie
                    |=> pile:
                        |=> layer:
                            |=> kernel: c' est
                        |=> layer:
                            |=> kernel: c' est
                    |=> kernel: quelque chose
```

FIGURE 4 – Représentation macro+entassement+micro de *c.*

4.2 Dépliage

L'une des premières applications possibles de l'arbre est de procéder au dépliage du texte. On entend par dépliage du texte un réarrangement des entassements permettant ensuite une analyse syntaxique automatique par un programme informatique. En effet les structures d'entassement, d'enchâssement et de parenthèsage, particulièrement courantes à l'oral, ne sont pas analysables en l'état par les analyseurs syntaxiques qui sont calibrés pour l'écrit. Les parsers ne savent pas traiter les disfluences et font encore beaucoup d'erreurs sur les coordinations (difficulté avec le rattachement du deuxième conjoint). Le dépliage va donc explorer chaque chemin de l'entassement et donne une UR bien formée sans entassement (Gerdes et Kahane, 2009). Il est nécessaire de fournir des segments syntaxiques débarrassés de tout phénomènes de l'oral à ce type de programme.

Considérons le premier exemple suivant :

a. les fêtes y sont { plus | plus } nombreuses // (échantillon D101-Corpus Rhapsodie)

Le dépliage correspondant est donné figure 5.

```
Unrolled sentences:
les fêtes y sont plus nombreuses
les fêtes y sont plus nombreuses
```

FIGURE 5 – Dépliage correspondant à l'exemple a.

Pour obtenir ce dépliage, il nous a fallu prendre en compte les éléments micro-syntaxiques (ici l'entassement) et macro-syntaxiques (l'UI composée d'un noyau), pour ce faire le multi-arbre donné figure 3 est nécessaire.

Soit à présent l'exemple suivant :

b. ^ alors le petit fauteuil { que j'ai { là | à côté } | que je veux rhabiller } a toujours été appelé par mes parents fauteuil-crapaud // (échantillon D009-Corpus Rhapsodie)

Le multi-arbre obtenu est représenté sur la figure 6 et le dépliage qui en résulte sur la figure 7.

```
|=> root:
    |=> iu:
        |=> ru:
            |=> intro: alors
        |=> ru:
            |=> kernel: le petit fauteuil [...] a toujours
                                été appelé par mes parents fauteuil-crapaud
            |=> pile:
                |=> layer:
                    |=> kernel: que j'ai
                    |=> pile:
                        |=> layer:
                            |=> kernel: là
                        |=> layer:
                            |=> kernel: à côté
                |=> layer:
                    |=> kernel: que je veux rhabiller
```

FIGURE 6 – multi-arbre de *b.*

```
alors
le petit fauteuil que j'ai là a toujours été appelé par mes parents fauteuil crapaud
le petit fauteuil que j'ai à côté a toujours été appelé par mes parents fauteuil crapaud
le petit fauteuil que je veux rhabiller a toujours été appelé par mes parents fauteuil crapaud
```

FIGURE 7 – UR dépliées extraites du multi-arbre de b. et envoyée au parser automatique

Ici on constate que chaque alternative d'entassement (chaque couche qui compose la pile) est explorée, chaque couche de chaque pile va venir occuper le rôle syntaxique qu'elle doit occuper. On obtient donc autant d'alternatives qu'il y a de piles et de couches dans une pile. On notera également que les éléments de type introducteurs (ici : *alors*), marqueurs discursifs etc [...] sont "séparés" des autres UR car eux aussi peuvent perturber l'analyse syntaxique automatique. En aucun cas ils ne seront ignorés, ils seront réintégrés (cf section 4.4) aux autres UR après l'analyse en ligne du parser FRMG par un algorithme de "repliage" (cf (Beliao et Liutkus, 2012)).

De plus, il est important de voir que l'information d'entassement n'est pas suffisante à elle seule pour fournir le dépliage de l'arbre, mais que le multi-arbre complet est bien nécessaire à cette tâche. En effet l'information d'entassement est d'ordre micro-syntaxique et n'est pas suffisante pour la tâche de dépliage car on l'a vu plus haut, les marqueurs discursifs aussi peuvent poser problème et que l'on a donc besoin parallèlement de l'information macro. Considérons l'exemple suivant :

c. il faut avoir un don spécial parce que [la psychiatrie < { c' est | c' est } quelque chose] //
(échantillon D006-Corpus Rhapsodie)

L'UI contient un enchâssement dans lequel on remarque une UR de type pré-noyau et un noyau contenant un entassement. Si l'on ignore l'une de ces informations le dépliage ne serait que partiel. Si l'on considère seulement l'information en UR, le pré-noyau "la psychiatrie" sera effectivement extrait mais on obtiendra une UI contenant la disfluence "*c'est c'est*", ce qui ne manquerait pas de provoquer un problème au moment de passage dans le parser automatique. Une fois de plus le multi-arbre s'avère indispensable.

```
la psychiatrie
il faut avoir un don spécial parce que c' est quelque chose
il faut avoir un don spécial parce que c' est quelque chose
```

FIGURE 8 – Dépliage résultant de l'analyse de *c.*

Le dépliage obtenu figure 8 nous permet de constater que le pré-noyau “la psychiatrie” a bien été sorti des phrases générées, ce qui n’aurait pas été possible si l’information topologique avait été éliminée par la considération d’un arbre simple.

4.3 Projections

On a vu précédemment que l’arbre complet était nécessaire à l’opération de dépliage. En effet si l’on veut obtenir un tri des phénomènes à extraire provisoirement des UI et les multiples possibilités qu’offrent les piles, la concomitance de ces informations au sein du même arbre est indispensable.

Pour faire un parallèle géométrique, on peut difficilement conceptualiser un hypercube. L’opération de projection ou d’extraction consiste ainsi à diminuer le nombre de dimensions présentes dans le multi-arbre, de manière à se focaliser sur un point de vue particulier, cela revient à extraire l’arbre voulu, par la sélection des noeuds désirés.

Prenons l’exemple suivant :

a. ^alors ce que je souhaiterais faire de ma vie < c’est { devenir professeur d’italien à savoir certifié | donc "euh" enseigner { au collège | ^ainsi ^qu’au lycée } } // (échantillon M103-Corpus Rhapsodie)

Le multi-arbre obtenu est donné en figure 9 et l’arbre macro-syntaxique correspondant est donné figure 10. En revanche si l’on souhaite étudier uniquement les phénomènes d’entassement, on peut obtenir un arbre d’entassement par l’extraction des nœuds d’entassement uniquement, donnée figure 11. Les arbres projetés macro-syntaxiques et d’entassement sont obtenus comme leur nom l’indique par des projections des nœuds voulus sur le multi-arbre. Admettons que l’on veuille l’arbre d’entassement, il suffit de n’afficher que les nœuds relatifs à l’information d’entassement etc...

```
|=> root:
    |=> iu:
        |=> ru:
            |=> intro: alors, ainsi qu'
        |=> ru:
            |=> prekernel: ce que je souhaiterais faire de ma vie
        |=> ru:
            |=> kernel: c'est [...]
            |=> pile:
                |=> layer:
                    |=> kernel: devenir professeur d'italien à savoir certifié
                |=> layer:
                    |=> kernel: donc enseigner
                    |=> pile:
                        |=> layer:
                            |=> kernel: au collège
                        |=> layer:
                            |=> kernel: au lycée
        |=> ru:
            |=> discursivemarker: euh
```

FIGURE 9 – Multi-arbre résultant de l’analyse de *a.*

```
|=> root:
    |=> iu:
        |=> ru:
            |=> intro: alors, ainsi qu'
        |=> ru:
            |=> prekernel: ce que je souhaiterais faire de ma vie
        |=> ru:
            |=> kernel: c'est devenir professeur d'italien à savoir
                                certifié, donc enseigner, au collège, au lycée
        |=> ru:
            |=> discursivemarker: euh
```

FIGURE 10 – Arbre projeté de macro-syntaxe résultant de l'analyse de *a*.

```
|=> root:
    |=> pile:
        |=> layer: devenir professeur d'italien à savoir certifié
        |=> layer: euh donc enseigner
            |=> pile:
                |=> layer: au collège
                |=> layer: ainsi qu'au lycée
```

FIGURE 11 – Arbre projeté des entassements résultant de l'analyse de *a*.

Ces arbres projetés ont été obtenus par l'application d'un algorithme de regroupement des nœuds sur l'arbre complet. Pour des raisons de place, on ne rentrera pas dans les détails de cet algorithme ici [3].

4.4 Repliage des résultats du parser automatique

Nous avons vu que la phase de dépliage visait à simplifier le parsage par un analyseur automatique. Pour chaque dépliage possible d'une UI, une analyse automatique produit un ensemble de traits syntaxiques et un système de dépendance. Le *repliage* consiste à repercuter tous ces traits et liens de dépendance sur la transcription originale. Cette phase complexe est rendue possible par l'identification des éléments communs aux différents dépliages et par le fait que les données sont représentés comme des objets, pouvant avoir plusieurs attributs et liens entre eux.

Pour chaque lexème, on obtient ainsi autant de rôles syntaxiques et de liens de dépendance que le nombre d'UI dépliées dans lesquelles ce lexème apparait. L'avantage de cette phase est qu'elle permet de désambiguiser l'analyse syntaxique de certains lexèmes, par exemple en choisissant pour trait syntaxique (genre, nb etc...) de chaque mot celui qui apparait le plus de fois ou, en cas d'égalité, de choisir le dernier (critère de proximité). Ainsi, dans :

a. *"j'ai acheté { un | une } baguette //"*

le lexème baguette reçoit le trait *féminin,* malgré les segments dépliés contradictoires envoyés au parser automatique :

```
j'ai acheté un baguette
j'ai acheté une baguette
```

Une propriété intéressante de l'approche par dépliage/repliage est qu'elle permet — avec seulement des modifications mineures — de considérer plusieurs annotations différentes du même texte. En effet, chaque annotation différente produirait son propre lot de dépliages à analyser. Le repliage permettrait alors de rendre compte des ces différentes annotation et rendrait possible une plus grande robustesse en cas d'ambiguités dans les choix des annotateurs.

3. l'algorithme est consultable dans le rapport technique (Beliao et Liutkus, 2012).

Pour finir, la phase de repliage permet de visualiser l'ensemble des traits et dépendances ainsi construits directement sur la transcription originale.

4.5 Conversion en formats structurés

L'un des objectifs du traitement du balisage est l'obtention de données structurées. On cherche à générer à partir du corpus annoté l'ensemble des arbres topologiques et des arbres d'entassement possibles. Pour ce faire, l'équipe de recherche a opté pour une structure XML, ce format sert de format d'import-export pour le corpus annoté et la base SQL du projet. À terme, tous les résultats des différentes phases d'annotations syntaxiques du corpus sont donc appelés à être chargés dans une base de données SQL. Le tables relationnelles de la base sont enrichies à partir de ces fichiers XML. Le format XML des différentes phase d'annotaion sert également d'input au logiciel Vakyartha-Arborator (Gerdes, 2012) qui permet à l'équipe des syntacticiens de procéder à une phase de vérification et correction manuelle après les phases de projection, dépliage et repliage des données annotées.

La problématique qui se pose est donc de convertir des structures de données obtenues vers des fichiers structurés XML. La représentation du balisage sous la forme d'arbre permet d'effectuer cette tâche de manière triviale à partir du multi-arbre. Pour l'opération de projection de la micro ou de la macro-syntaxe, des algorithmes récursifs très simples permettent de convertir un arbre en format structuré de type XML.

Pour l'exemple a. on aura la représentation de la figure 12 pour la représentation topologique et la figure 13 pour la représentation de l'entassement.

a. ^ qui (donc) reste { toute seule | fort étonnée } // (échantillon M002-Corpus Rhapsodie)

```
<constree const_type="topology" id="a">
    <const type="iu">
        <const type="intro">
            <const const_type="lexeme" id="qui"/>
        </const>
        <const type="inkernel">
            <const const_type="lexeme" id="donc"/>
        </const>
        <const type="kernel">
            <const const_type="lexeme" id="reste"/>
            <const const_type="lexeme" id="toute"/>
            <const const_type="lexeme" id="seule"/>
            <const const_type="lexeme" id="fort"/>
            <const const_type="lexeme" id="étonnée"/>
        </const>
    </const>
</constree>
```

FIGURE 12 – Arbre topologique résultant de l'analyse de *a.*

```
<constree const_type="pile" id="a">
    <const type="pile">
        <const type="layer">
            <const const_type="lexeme" id="toute"/>
            <const const_type="lexeme" id="seule'"/>
        </const>
        <const type="layer">
            <const const_type="lexeme" id="fort"/>
            <const const_type="lexeme" id="étonnée"/>
        </const>
    </const>
</constree>
```

FIGURE 13 – Arbre d'entassement résultant de l'analyse de *a.*

Après le passage des UI dépliées dans l'analyseur automatique on procède au repliage des UIs dépliées et on obtient —après modification de certains traits et ajout de certains liens— un arbre de dépendance au format XML. Pour l'UI de exemple a. nous obtiendrons alors l'arbre de dépendance XML de la figure 14.

a. les fêtes y sont { plus | plus } nombreuses //

```
<dependency id="dep33" markupIU="les␣fêtes␣y␣sont␣{␣plus␣|␣plus␣}␣nombreuses␣//">
    <link depid="plus" func="dep" govid="nombreuses" id="func402"/>
    <link depid="plus" func="dep" govid="nombreuses" id="func403"/>
    <link depid="les" func="dep" govid="fêtes" id="func404"/>
    <link depid="fêtes" func="sub" govid="sont" id="func405"/>
    <link depid="y" func="ad" govid="sont" id="func406"/>
    <link depid="sont" func="root" id="func407"/>
    <link depid="nombreuses" func="pred" govid="sont" id="func408"/>
</dependency>
```

FIGURE 14 – Arbre de dépendance obtenu après repliage des UI dépliées résultant de a.

Le multi-arbre n'est pas généré en format XML dans le cadre du projet, il n'est utilisé que comme structure relais permettant l'ensemble des traitements nécessaires à la réalisation des tâches de dépliage et de projection.

5 Conclusion

Dans cette présentation on a proposé une systématisation informatique du système d'annotation du corpus Rhapsodie pour son exploitation par un parser FRMG. Cette proposition allie une représentation sous forme d'arbre, adaptée au formalisme souhaité, et les différents algorithmes permettant de mettre en œuvre cette proposition. Dans cette étude, nous nous sommes concentrés sur la présentation de cette représentation et sur les différents traitements qu'elle permet. La présentation des traitements informatiques correspondants fait l'objet d'un rapport technique indépendant.

Il a été vu que l'implémentation dans un multi-arbre de la totalité de l'information encodée dans le balisage est nécessaire pour certains traitements, tels que le dépliage, montrant qu'une exploitation conjointe des niveaux micro et macro-syntaxiques est parfois nécessaire. Ce multi-arbre peut aisément être projeté — ou particularisé — pour ne plus inclure qu'un sous ensemble des informations qu'il contient.

Un grand nombre de points évoqués dans cette étude peuvent faire l'objet de travaux ultérieurs. Tout d'abord, il est possible d'étendre la présente étude au cas où plusieurs annotations sont

disponibles pour la même transcription. Ensuite, il ne semble pas que la notion d'arbre, limitée au cas où chaque noeud n'a qu'un seul père, permette de rendre compte de tous les liens de dépendance envisageables. Un graphe, plus général, pourrait être plus adéquat dans ce but.

Références

Beliao, J. et Liutkus, A. (2012). Rapport technique provisoire des algorithmes utilisés pour le parsing d'un corpus de francais oral annoté. Rapport technique, HAL : halshs-00682283 version 1.

Benzitoun, C., Dister, A., Gerdes, K., Kahane, S. et Marlet, R. (2009). annoter du des textes tu te demandes si c'est syntaxique tu vois. *The 28th Conference on Lexis and Grammar*, Arena Romanistica 4, Presses de l'Université de Bergen:16–27.

Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P. et Sabio, F. (2010). Tu veux couper là faut dire pourquoi. propositions pour une segmentation syntaxique du francais parlé. *Actes du Congrès Mondial de Linguistique Francaise*, La Nouvelle Orléans.

Berrendonner, A. (2002). Morpho-syntaxe, pragma-syntaxe et ambivalences sémantiques. *Andersen, N. Nolke (éds). Macro-syntaxe et macro-sémantique. Actes du colloque d'Aarhus*, pages 23–41.

Blanche-Benveniste, C. (1990). Un modèle d'analyse syntaxique 'en grilles' pour les productions orales. *Anuario de Psicologia Liliane Tolchinsky (coord.) Barcelona*, vol. 47:11–28.

Blanche-Benveniste, C. (1997). Approches de la langue parlée en francais. *Paris : Ophrys*.

Blanche-Benveniste, C., Bilger, M., Rouget, C. et van den Eynd, K. (1990). Le français parlé. études grammaticales. *Paris, CNRS Éditions.*

Cresti, E. (2000). Corpus di italiano parlato. *Florence, Accademia della Crusca.*

de la Clergerie, E., Sagot, B., Nicolas, L. et Guénot, M.-L. (2009). Frmg : évolutions d'un analyseur syntaxique tag du francais. *11th International Conference on Parsing Technologies (IWPT'09)*.

Deulofeu, J. (1999). *Recherches sur les formes de la prédication dans les énoncés assertifs en français contemporain (le cas des énoncés introduits par le morphème que)*. Thèse de doctorat, Université Paris 3.

Gerdes, K. (2012). Arborator : A tool for collaborative dependency annotation. *http ://arborator.ilpga.fr/vakyartha/*.

Gerdes, K. et Kahane, S. (2009). Speaking in piles : Paradigmatic annotation of french spoken corpus. *Proceedings of the Fifth Corpus Linguistics Conference, Liverpool.*

Kahane, S. (2012). De l'analyse en grille à la modélisation des entassements. *(à paraitre) Hommage à Claire Blanche-Benveniste, Presses de l'université de Provence.*, in S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio éds.

Kahane, S. et Pietrandrea, P. (2012). Typologie des entassements en francais. *In Actes de la conférence Linx*.

Lacheret-Dujour, A., Kahane, S., Pietrandrea, P., Avanzi, M. et Victorri, B. (2011). Oui mais elle est où la coupure, là ? Quand syntaxe et prosodie s'entraident ou se complètent. *Langue française, Paris-Larousse*, 170:61–80.

Rhapsodie (2012). Site du projet rhapsodie, corpus prosodique de référence en francais parlé. *http ://rhapsodie.risc.cnrs.fr*.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Librairie C. Klincksieck.

Construction automatique d'un lexique de modifieurs de polarité

Noémi Boubel
UCLouvain, Cental, Place Blaise Pascal, 1, B-1348 Louvain-la-Neuve, Belgique
noemi.boubel@uclouvain.be

RÉSUMÉ

La recherche présentée [1] s'inscrit dans le domaine de la fouille d'opinion, domaine qui consiste principalement à déterminer la polarité d'un texte ou d'une phrase. Dans cette optique, le contexte autour d'un mot polarisé joue un rôle essentiel, car il peut modifier la polarité initiale de ce terme. Nous avons choisi d'approfondir cette question et de détecter précisément ces modifieurs de polarité. Une étude exploratoire, décrite dans des travaux antérieurs, nous a permis d'extraire automatiquement des adverbes qui jouent un rôle sur la polarité des adjectifs auxquels ils sont associés et de préciser leur impact. Nous avons ensuite amélioré le système d'extraction afin de construire automatiquement un lexique de structures lexico-syntaxiques modifiantes associées au type d'impact qu'elles ont sur un terme polarisé. Nous présentons ici le fonctionnement du système actuel ainsi que l'évaluation du lexique obtenu.

ABSTRACT

Automatic Construction of a Contextual Valence Shifters Lexicon

The research presented in this paper takes place in the field of Opinion Mining, which is mainly devoted to assigning a positive or negative label to a text or a sentence. The context of a highly polarized word plays an essential role, as it can modify its original polarity. The present work addresses this issue and focuses on the detection of polarity shifters. In a previous study, we have automatically extracted adverbs impacting the polarity of the adjectives they are associated to and qualified their influence. The extraction system has then been improved to automatically build a lexicon of contextual valence shifters. This lexicon contains lexico-syntactic patterns combined with the type of influence they have on the valence of the polarized item. The purpose of this paper is to show how the current system works and to present the evaluation of the created lexicon.

MOTS-CLÉS : fouille d'opinion, modifieurs de valence affective, modifieurs de polarité.

KEYWORDS: opinion mining, contextual valence shifters.

1. avec le soutien de Wallonie-Bruxelles International

1 Introduction et état de l'art

Le champ de recherche de la fouille d'opinion regroupe des tâches diverses, notamment celle de distinguer le positif du négatif et de définir de cette façon la polarité (ou *valence*) d'un texte. D'un point de vue terminologique, nous préférerons le terme *polarité* au terme anglais *valence* afin d'éviter l'ambiguité avec le concept français de *valence* en syntaxe. Ces dernières années, les recherches dans ce domaine se sont fortement développées, comme on peut le voir dans la vue d'ensemble donnée par (Pang et Lee, 2008). Les tâches à accomplir se sont diversifiées et spécialisées, selon les contraintes industrielles ou le niveau de précision voulu.
Il est progressivement apparu qu'une des entraves importantes à l'efficacité des systèmes de fouille d'opinion était la prise en compte du contexte. En effet, la présence d'un terme négatif dans une phrase, par exemple, ne signifie pas forcément que la phrase est négative. Ce terme peut effectivement être nié, tempéré, intégré dans un contexte hypothétique, etc. Ainsi, de nombreux phénomènes contextuels ont un impact sur un terme polarisé dans un texte. Notre objectif ici est d'identifier et de décrire précisément ces phénomènes.

1.1 Etat de l'art

Dans le domaine de la fouille d'opinion, rares sont les travaux dont le sujet central d'étude traite des phénomènes qui ont un impact sur la polarité d'un terme. Zaenen et Polanyi (2004) tentent, dans cette optique, de décrire tous les cas où le contexte modifie un terme polarisé. Leur étude, en anglais, postule l'existence d'éléments contextuels appelés *contextual valence shifters* qui modifient la valeur initiale d'un terme. Leur hypothèse de travail est que la valence de termes polarisés peut être renforcée ou affaiblie par la présence d'autres items lexicaux, par la structure du discours et le type de texte, ou enfin par des facteurs culturels.

Sur leur impulsion, des travaux de plus en plus nombreux, introduisent dans des systèmes de classification d'opinion, une certaine prise en compte du contexte, plus ou moins complète et riche (Kennedy et Inkpen, 2006; Musat et Trausan-Matu, 2010; Taboada *et al.*, 2011). Les ressources utilisées (comme des listes d'adverbes) sont généralement définies intuitivement. La terminologie anglophone qui traite de ces concepts relativement récents est assez diverse et peu stabilisée. Plusieurs notions sont utilisées, certaines se complètent ou se recouvrent. Ainsi, Zaenen et Polanyi (2004) traitent des concepts de *contextual valence shifter*, ou *modifier* et plus précisément de la *negation* (qui inverse la polarité) et des *intensifiers* (qui l'intensifient ou l'atténuent). Kennedy et Inkpen (2006) précisent ensuite cette terminologie et divisent les modifieurs en trois types : la *negation*, les *intensifiers* (qui ont la seule faculté d'intensification) et les *diminishers* (qui atténuent la force d'un terme polarisé). Signalons également le concept d'*intensifiers* défini par (Quirk *et al.*, 1985) comme des éléments qui ont un impact sur l'intensité de la polarité d'un terme. Ils se classent en deux grandes catégories : les *amplifiers* qui amplifient l'intensité sémantique du voisinnage lexical (*very*), et les *downtoners*, qui atténuent cette intensité (*slightly*). De plus, au-delà de l'idée d'intensité, sont développés également les concepts de *polarity influencers* (Wilson *et al.*, 2009), *non veridical context* (Zwarts, 1995; Giannakidou, 1998) ou *irrealis markers* (Taboada *et al.*, 2011).

Cette terminologie n'est pas, à notre connaissance, développée dans les recherches francophones. Vernier *et al.* (2009) et Petrakis *et al.* (2009) prennent en compte, dans certains cas, le contexte autour de termes polarisés ou la combinaison de plusieurs termes polarisés, mais ne reprennent

pas le concept de *contextual valence shifter* tel qu'il est défini plus haut. Nous parlerons ici de *modifieur de polarité* (aussi appelés *modifieur de valence affective*), et des notions d'*intensifieurs*, *atténuateurs*, et *inverseurs*, à la suite des travaux de (Zaenen et Polanyi, 2004) et (Kennedy et Inkpen, 2006).

1.2 Objectifs

Sur la base des constatations ci-dessus, notre objectif est d'extraire automatiquement, à partir d'un corpus, des phénomènes contextuels qui ont un impact sur des termes polarisés, autrement dit construire de façon automatique une liste de modifieurs à partir d'un corpus. Nous nous limitons ici à l'étude de toutes structures (ou patrons) lexico-syntaxiques dans lesquelles sont intégrés des termes polarisés et qui ont un impact sur la polarité d'un terme. Il s'agira par exemple du syntagme prépositionnel qui associe la préposition *sans* à un nom, patron lexico-syntaxique qui inverse la polarité du nom. Nos travaux antérieurs ont conduit à la création d'une méthodologie, qui repère des structures de ce type susceptibles d'être des modifieurs de polarité.
Cette méthodologie, appliquée à l'extraction d'adverbes (plus précisément de syntagmes adjectivaux modifiés par des adverbes), est décrite dans (Boubel et Bestgen, 2011). Les résultats de cette étude exploratoire permettent de supposer que certaines caractéristiques statistiques peuvent prédire des caractéristiques sémantiques (et prédire donc en particulier l'impact sémantique du modifieur sur un terme polarisé). Une analyse linguistique des adverbes extraits, présentée dans (Boubel, 2011) a ensuite été menée afin de vérifier cette hypothèse. Cette analyse a mis en évidence trois types d'adverbes partageant des caractéristiques statistiques communes. Il est apparu que les adverbes de chaque catégorie remplissent également un rôle sémantique similaire. Trois cas ont ainsi été dégagés [2] :

– le modifieur *intensifie* le terme polarisé auquel il est associé (« (. . .) le film est *absolument* jubilatoire. ») ;
– le modifieur *inverse* ou *atténue* la polarité du terme (« C'est absurde, *peu* crédible, inintéressant (...). ») ;
– le modifieur apparait dans une *structure évaluative plus large*, comme une *comparaison* ou une *concession*, et met souvent en relation plusieurs termes polarisés (« On l'eût aimé *moins* glacé, *plus* fiévreux, *plus* emporté. ») ; ce dernier cas se distingue des précédents car le modifieur n'a pas un impact direct sur un terme précis.

Notre objectif ici est d'améliorer et de perfectionner la méthodologie d'extraction sur deux points :

1. Dépasser le cadre des adverbes : notre première étude a démontré la pertinence des adverbes comme modifieurs potentiels ; nous cherchons maintenant à déterminer dans quelle mesure d'autres catégories syntaxiques peuvent également être modifiantes. Pour cela, nous avons adapté le système à la détection de toutes relations de dépendance syntaxique éventuellement modifiantes.
2. Automatiser le classement des modifieurs : la méthodologie d'extraction ne définit pas, à l'origine, la nature du modifieur (son impact sur le terme polarisé). Nous avons automatisé, dans notre système actuel, le classement des modifieurs dans une des trois catégories définies plus haut.

2. Cette catégorisation pourra être améliorée ou modifiée selon nos différentes investigations dans le domaine et notamment selon les résultats obtenus dans cette étude.

L'outil fournit donc maintenant en sortie une liste de *structures lexico-syntaxiques modifiantes* classées en *trois groupes*. L'objectif principal de l'article est d'évaluer la pertinence des résultats obtenus, et la performance du système. L'évaluation manuelle et systématique effectuée nous permet également de juger de la pertinence de notre catégorisation de modifieurs et de mettre en lumière d'autres phénomènes contextuels intéressants.

Dans la suite de cet article, nous décrivons l'approche adoptée et la méthodologie proposée, avant d'évaluer les résultats et de conclure.

2 Méthodologie proposée

La méthodologie d'extraction a été développée en collaboration avec Yves Bestgen et est décrite en détail dans (Boubel et Bestgen, 2011). Nous rappelons ici les grandes lignes de l'approche. Nous expliquons ensuite plus en détail l'automatisation du classement des modifieurs grâce à l'ajout de règles basées sur les résultats statistiques.

2.1 Approche

Nous nous basons sur deux ressources : un corpus contenant des énoncés dont on connaît la polarité, et un lexique de termes positifs ou négatifs. L'idée de départ de notre approche est de s'intéresser au contexte linguistique des termes issus du lexique. En effet, on peut supposer que l'impact du contexte sera différent selon qu'il porte sur un terme positif ou négatif : (1) dans un texte négatif, (2) dans un texte positif ou (3) dans un texte présentant une opinion mitigée. Nous nous limitons à l'étude de structures lexico-syntaxiques, et étudions en conséquence les *relations de dépendance syntaxique* mettant en jeu un terme polarisé. L'objectif est de rendre compte, grâce à des techniques statistiques, de l'influence du contexte sur la polarité d'un terme et de dégager les contextes lexico-syntaxiques qui induisent toujours le même impact.

2.2 Ressources et outils utilisés

Le corpus utilisé est constitué d'extraits de critiques de films issus du site Allociné [3]. Ce site rassemble, pour un même film, de brefs extraits (pouvant aller d'un syntagme à quelques phrases) d'articles provenant de différents journaux donnant une opinion sur le film. Ces extraits sont classés selon la teneure de l'opinion sur une échelle de 1 à 5 : les avis très négatifs ont la note de 1, et les avis très positifs la note de 5. Le corpus contient 77561 critiques et environ 2 millions de mots.

Le lexique que nous avons utilisé (lexique classant des termes selon leur polarité) a été constitué automatiquement grâce à la méthode de (Vincze et Bestgen, 2011). Cette méthode permet de construire un lexique de polarité adapté à un domaine précis en se basant sur un corpus de textes traitant du domaine voulu et sur une liste de mots dont la polarité a été déterminée en demandant à des personnes de les juger sur un échelle allant de 1 à 7. Le lexique à notre disposition a donc été créé sur la base de notre corpus de critiques de films et est donc adapté

3. http ://www.allocine.fr/

au vocabulaire du cinéma ("*nanar*" est classé dans les termes négatifs, par exemple). Pour un meilleur résultat, nous n'avons retenu, du lexique de départ, que les termes positifs et négatifs les plus extrêmes et avons ensuite fait un bref nettoyage des listes afin d'en supprimer les termes incohérents. Nous obtenons donc une liste de 846 termes négatifs et 857 termes positifs.

Enfin, nous utilisons l'analyseur syntaxique XIP de Xerox (Aït-Mokhtar *et al.*, 2002) pour extraire les relations de dépendance syntaxique mettant en jeu un terme polarisé.

2.3 Traitement

2.3.1 Analyses statistiques

Grâce aux différentes ressources et outils décrits ci-dessus, nous obtenons un corpus annoté en dépendances. Nous identifions automatiquement les syntagmes contenant un terme polarisé et obtenons en sortie une liste de syntagmes, la polarité du terme polarisé contenu dans le syntagme, et la note de la critique dans laquelle chaque syntagme a été extrait (cf. table 1).

Relation extraite :	DETERM_INT(<NOUN :fluidité_fluidité :>,<DET :quelle_quel :>)
Terme polarisé :	*fluidité*
Polarité du terme :	positif
Phrase brute :	« *Quel rythme, quelle fluidité.* »
Note de la critique :	5/5
Structure lexico-syntaxique :	DETERM_INT(<NOUN : :>,<DET :quel :>)

TABLE 1 – Les différentes caractéristiques d'un syntagme extrait

Afin de juger de l'impact des syntagmes extraits sur la polarité, nous retirons le terme polarisé et recherchons au sein du corpus la structure lexico-syntaxique obtenue. Cette méthode nous a permis de déterminer la fréquence d'apparition de chacune de ces structures dans le corpus selon deux critères : (1) la note de la critique, (2) la polarité du terme polarisé. Nous analysons ces données d'un point de vue statistique en construisant une table de contingence pour chaque structure selon ces deux critères (nous ne retenons que les structures ayant une fréquence de plus de 20 quand elles sont associées à au moins une des deux polarités). Nous analysons chaque table au moyen du test du chi-carré, et évaluons de cette manière s'il y a indépendance entre la note de la critique et la fréquence de la relation. Nous retenons ensuite les relations pour lesquelles le résultat du test est significatif (seuil de 0,05) et en calculons les résidus ajustés. De cette façon, nous mettons en évidence les structures lexico-syntaxiques dont la note de la critique et la polarité du terme associé ont un effet sur leur distribution dans le corpus. Nous nous appuyons enfin sur la valeur des résidus ajustés significatifs (résidus dont la probabilité est inférieure à 0,05) pour parler d'une relation *surreprésentée* (résidu ajusté positif) ou *sous-représentée* (résidu ajusté négatif) dans une note particulière avec un terme d'une certaine polarité.
La table 2 montre les résultats statistiques obtenus pour le syntagme nominal modifié par l'adjectif *total*. Lorsqu'elle est associée à un nom positif, cette structure obtient un chi-carré de 16,99, correspondant à une p-value de 0,002, valeur inférieure au seuil de 0,05. Nous considérons donc le test significatif, et calculons les résidus ajustés. Les valeurs des résidus ajustés significatifs indiquent que la structure, lorsqu'elle est associée à un nom positif, est sous-représentée dans les notes 2 et 3, et surreprésentée dans les notes 4 et 5. La même analyse est effectuée lorsque *total* est associé à un nom négatif. Ainsi, le syntagme s'avère notamment surreprésenté dans les

critiques positives avec un nom positif (« (...) un film d'*une finesse totale* ») et surreprésenté dans les critiques négatives avec un nom négatif (« (...) d'une bêtise abyssale, d'*une abjection totale*. »)

[total - **nom positif**] : χ^2=16.99 (pvalue : 0,002)		[total - **nom négatif**] : χ^2=38,57 (pvalue : 0,000)	
Note	Résidus Ajustés	Note	Résidus Ajustés
1	-1,04 (proba : 0,30) non-significatif	**1**	**5,14** (proba : 0,0000) **surreprésentation**
2	-2,24 (proba : 0,0253)sous-représentation	**2**	**2,58** (proba : 0,0099) **surreprésentation**
3	-2,23 (proba : 0,0259)sous-représentation	3	-2,05 (proba : 0,0401) sous-représentation
4	**2,18** (proba : 0,0292) **surreprésentation**	4	-2,73 (proba : 0,0063) sous-représentation
5	**2,48** (proba : 0,0132) **surreprésentation**	5	-0,83 (proba : 0,40) non-significatif

TABLE 2 – Caractéristiques statistiques de l'adjectif *total*

2.3.2 Règles d'attribution automatique d'étiquettes

Afin de finaliser la procédure d'extraction, nous cherchons à définir automatiquement l'impact des structures lexico-syntaxiques significatives extraites sur un terme polarisé.
Dans (Boubel, 2011), notre analyse linguistique nous avait amenée à analyser plus en détail les résultats statistiques. Nous avions ainsi mis en relation les surreprésentations et sous-représentations caractéristiques de ces adverbes avec leur rôle sémantique. Globalement, trois tendances se sont dégagées (la surreprésentation s'est avérée plus informative que la sous-représentation) :

1. Les adverbes surreprésentés avec des adjectifs dont la polarité coïncide avec celle de la critique, comme pour l'adverbe *profondément* (table 3), *intensifient* souvent la valeur des adjectifs auxquels ils sont associés : « (...) Un film gonflé et *profondément attachant* (...) ».
2. Les adverbes surreprésentés avec un adjectif dont la polarité ne coïncide pas avec celle de la critique, comme on peut le voir dans la table 4, remplissent souvent un rôle d'atténuateur ou d'inverseur : « Ici, c'est épuisant et *jamais crédible*. ».
3. Les adverbes surreprésentés dans les notes mitigées, comme c'est le cas pour l'adverbe *parfois* (table 5) sont souvent inclus dans des structures rhétoriques plus larges comme des mécanismes d'opposition, de concession ou de comparaison : « La poésie trash du réalisateur ne faiblit pas, contrairement au rythme de son récit *parfois répétitif*. ».

Sur ces constatations, nous avons donc dégagé trois catégories d'adverbes ayant un rôle sémantique différent. Pour plus de facilité pour la suite de notre étude, nous abrégeons ces différentes catégories par *intensifieurs* (point 1 ci-dessus), *inverseurs* (point 2) et *concessifs* (point 3), bien que ces termes soient réducteurs.

ADJMOD(<ADJ : :>,<ADV :profondément :>)					
Avec mots **positifs**	1/5	2/5	3/5	**4/5**	**5/5**
surreprésenté				•	•
non-significatif	•				
sous-représenté		•	•		

TABLE 3 – Surreprésentations et sous-représentations de l'adverbe *profondément*

ADJMOD(<ADJ : :>,<ADV :jamais :>)					
Avec mots **négatifs**	1/5	2/5	3/5	**4/5**	**5/5**
surreprésenté				•	•
non-significatif					
sous-représenté	•	•	•		

TABLE 4 – Surreprésentations et sous-représentations de l'adverbe *jamais*

ADJMOD(<ADJ : :>,<ADV :parfois :>)					
Avec mots positifs	1/5	2/5	**3/5**	4/5	5/5
surreprésenté			•		
non-significatif	•	•			•
sous-représenté				•	
Avec mots négatifs	1/5	2/5	**3/5**	4/5	5/5
surreprésenté			•		
non-significatif				•	•
sous-représenté	•	•			

TABLE 5 – Surreprésentations et sous-représentations de l'adverbe *parfois*

Il est raisonnable de penser que ces conclusions peuvent être valables pour d'autres structures que les adverbes. C'est donc sur la base de l'étude empirique présentée ci-dessus que nous avons défini un ensemble de règles qui attribuent un score à une structure pour chaque type de modifieurs. Nous conférons ainsi un score de 1 à 10 à une structure pour chacune des trois classes de modifieurs, grâce à une dizaine de règles par classe. Ces règles, que nous ne détaillons pas ici pour une question de clarté et de place, se basent sur les propriétés statistiques du modifieur (surreprésentations et sous-représentations) en fonction de deux critères : la polarité du terme et la note de la critique. Ainsi, une structure obtient un score élevé : (1) dans la classe des intensifieurs lorsqu'elle est surreprésentée dans une critique dont la polarité coïncide avec celle du terme polarisé ; (2) dans la classe des inverseurs lorsqu'elle est surreprésentée dans une critique dont la polarité n'est pas celle du terme polarisé ; (3) dans la classe des concessifs lorsqu'elle est surreprésentée dans les critiques mitigées (note 3/5). Nous classons ensuite la structure dans la catégorie de modifieurs qui obtient le score le plus élevé.
De cette façon, le syntagme nominal modifié par l'adjectif *total* (cf table 2) obtient un score de 8 comme intensifieur, 0 comme inverseur et 2 commme concessif. Cette structure est en effet surreprésentée dans les critiques **positives** lorsqu'elle est associée à un nom **positif**, et surreprésentée dans les critiques **négatives** lorsqu'elle est associée à un nom **négatif**, ce qui lui confère un score d'intensification élevé. Le syntagme nominal modifié par l'adjectif *total* est donc répertorié comme intensifieur avec un score de 8.

3 Évaluation

3.1 Résultats de l'extraction

Au terme de notre expérimentation, nous obtenons une liste de 243 structures lexico-syntaxiques supposées modifiantes dont 108 intensifieurs, 74 inverseurs, 52 concessifs. De plus, 9 structures

obtiennent un score ex-aequo maximal dans deux catégories : 1 structure classée autant comme un inverseur que comme un concessif, et 8 structures en tant qu'intensifieurs et concessifs. Les scores obtenus pour chacune des structures sont relativement peu élevés : seulement 16 structures ont un score de 7 ou plus, et 127 structures, soit un peu plus de 50%, obtiennent un score entre 3 et 4.
Les catégories syntaxiques des éléments modifieurs au sein de la structure lexico-syntaxique sont diverses, on trouve notamment 53 adverbes, 63 adjectifs, 66 noms, 20 prépositions. La table 6 présente les 10 structures remportant les scores les plus élevés.

Structure	Score	Modifieur	Exemple
NEGAT(<VERB : :>)	10	Inverseur	Ce palace *ne mérite* vraiment pas d'étoile...
DETERM(<NOUN : :>,<DET :un :>)	9	Intensifieur	*Un ratage* assez complet
VMOD_POSIT1(<VERB : :>,<ADV :pas :>)	9	Inverseur	Pas drôle, *pas rythmé* (...)
ADJMOD(<ADJ : :>,<ADV :trop :>)	8,5	Concessif	Certains trouveront l'esquisse *trop caricaturale*, mais (...)
NMOD_POSIT1(<NOUN : :>,<ADJ :total :>)	8	Intensifieur	un film d'une *finesse totale*
PREPOBJ(<NOUN : :>,<PREP :sans :>)	8	Inverseur	C'est un ouvrage *sans grâce*
ADJMOD(<ADJ : :>,<ADV :moins :>)	8	Concessif	C'est à peine *moins racoleur* que (...)
ADJMOD(<ADJ : :>,<ADV :parfois :>)	7,5	Concessif	Jamais malsain, mais *parfois vulgaire*, (...)
ADJMOD(<ADJ : :>,<ADV :assez :>)	7,5	Concessif	Tout cela n'est pas méchant, mais *assez vain*.
DETERM(<NOUN : :>,<DET :quelque :>)	7,5	Concessif	À peine *quelques lueurs* dans un océan de niaiseries

TABLE 6 – Liste des 10 premières structures extraites obtenant le score le plus élevé.

3.2 Évaluation des résultats

Pour évaluer la pertinence des résultats, nous avons parcouru manuellement la liste des 243 structures extraites afin de déterminer dans quelle mesure elles modifient effectivement la polarité d'un terme et de quelle façon. Il s'agit donc d'une évaluation qualitative des résultats obtenus. Bien entendu, il sera indispensable d'effectuer une analyse quantitative et d'évaluer l'apport de cette liste de modifieurs au sein d'une application de fouille d'opinion. Ces deux évaluations n'ont pas le même objectif et se complètent, c'est pourquoi il est intéressant de les mener à bien toutes les deux.
De nombreux retours sur corpus ont été nécessaires pour comprendre le rôle des structures extraites dans une critique. Certaines structures font référence à des constructions lexico-syntaxiques plus larges, que l'analyseur syntaxique ne peut pas restituer dans son intégralité. C'est le cas, par exemple, de la relation *PREPOBJ* d'un verbe modifié par la préposition *par*, qui fait référence à la construction plus large : [*il finit par - VB*]). Dans ce cas, nous avons jugé et classé les structures ou expressions dans leur intégralité. Nous les classons donc comme correctes lorsqu'elles sont effectivement modifiantes.

3.2.1 Cas pertinents

Notre évaluation manuelle nous a conduit à juger 87 structures pertinentes sur 243, dont 37 intensifieurs, 22 inverseurs et 27 concessifs. Le dernier modifieur est la structure adjectivale modifiée par l'adverbe *peu* classée autant comme un inverseur que comme un concessif (avec un score de 4 pour les deux catégories). Cette double fonction inverseur-concessif n'est pas incohérente dans l'absolu. Pour l'adverbe *peu*, en particulier, il est parfaitement concevable de lui attribuer un rôle d'atténuateur (« Pour addicts *peu exigeants*. »), mais aussi de le voir intégrer une structure comparative ou concessive plus large (« Aussi *peu* réaliste que morale »). En revanche, le double classement intensifieur-inverseur est plus problématique en soi. Aucune des 8 structures dans cette situation (comme la préposition *à travers* ou le complément du nom *homme*) n'a été retenue. Les structures mises en évidence ici sont variées. Elles sont souvent composées d'adverbes et d'adjectifs, catégories syntaxiques largement utilisées dans d'autres travaux du domaine. Toutefois d'autres structures rarement traitées apparaissent également. Nous détaillons ici les résultats pour chaque catégorie de modifieurs.

La liste des 37 intensifieurs, dont des exemples sont reportés dans la table 7, est principalement constituée, à plus de 90%, de structures lexico-syntaxiques contenant des adjectifs ou des adverbes. On remarquera toutefois la présence de certaines formes plus complexes, comme le complément du nom "*de l'année*", qui amplifie clairement la polarité du nom auquel il est associé.

Structure	Score	Exemple
NMOD_POSIT1(<NOUN : :>,<ADJ :total :>)	8	Un *total* enchantement.
NMOD_POSIT1(<NOUN : :>,<ADJ :véritable :>)	5	(...) un *véritable* nanar sans intérêt (...)
NMOD_POSIT1(<NOUN : :>,<ADJ :tel :>)	5	Dommage que les personnages, les gags et le scénario (...) dégagent un *tel ennui*.
ADJMOD(<ADJ : :>,<ADV :profondément :>)	4	Un film *profondément* généreux.
NMOD_POSIT1(<NOUN : :>,<NOUN :année :>)	2	(...) le ratage le plus spectaculaire et inattendu *de l'année*.

TABLE 7 – Exemple d'intensifieurs

Ensuite, les 22 inverseurs ont des formes plus hétérogènes, les catégories syntaxiques qui les composent sont plus diverses. Il s'agit également en majorité d'adverbes, mais peu d'adjectifs sont présents. Les stratégies utilisées pour exprimer l'inversion ou l'atténuation semblent en effet plus variées. Les structures qui ont obtenu les scores les plus hauts sont clairement des inverseurs, comme on peut le voir dans la table 8. Notons que la structure *VMOD(<VERB : :>,<CONJQUE :que :>)* correspond à la construction syntaxique restrictive "*ne..que*".

Enfin, les 27 concessifs sont en majorité des structures contenant des adverbes, mais d'autres structures plus diverses apparaissent, de la même façon que pour les inverseurs. Là encore, les adjectifs sont peu nombreux (au nombre de 3 : *certain, inégal, même*). Au vu de la liste à juger, nous sommes amenée à reconsidérer quelque peu la définition de cette catégorie. On y trouve en effet des éléments très divers, mais qui expriment tous, et grâce à des stratégies plus ou moins directes, un avis mitigé, une opinion nuancée, ou une hésitation, comme on peut l'entrevoir dans les exemples de la table 9. Ainsi, les constructions plus complexes (*vb - sans déplaisir*, *finir par - vb*) font référence à des stratégies qui expriment un avis mitigé ou peu enthousiaste.

Structure	Score	Exemple
NEGAT(<VERB : :>)	10	(...) son énergie, (...) *ne* desservent jamais l'ensemble
VMOD_POSIT1(<VERB : :>,<ADV :pas :>)	9	La poésie et le ton (...) ne fonctionnent *pas*.
PREPOBJ(<NOUN : :>,<PREP :sans :>)	8	C'est un ouvrage *sans grâce*.
ADJMOD(<ADJ : :>,<ADV :jamais :>)	5	Ici, c'est épuisant et *jamais crédible*.
VMOD(<VERB : :>,<CONJQUE :que :>)	5	un exercice de style où la vie ne palpite *que* trop rarement.
PREPOBJ(<NOUN : :>,<PREP :en dépit de :>)	3	Hélas, *en dépit de la générosité* du propos, (...)

TABLE 8 – Exemple d'inverseurs

Structure	Score	Exemple
DETERM(<NOUN : :>,<DET :quelque :>)	7,5	(..) arrache cependant *quelques sourires* désabusés.
PREPOBJ(<VERB : :>,<PREP :par :>)	7,5	(...) finit *par séduire* ; (...) finit *par lasser*
CONNECT(<VERB : :>,<CONJ :si :>)	7,5	Mais, avouons-le, *si* le mélo *fonctionne* c'est surtout grâce à (...)
NMOD_POSIT1(<NOUN : :>,<ADJ :certain :>)	7	(...) une *certaine froideur* habite son exercice de style virtuose.
NMOD_POSIT1(<NOUN : :>,<ADJ :inégal :>)	4	(...) avec un *bonheur inégal*.
PREPOBJ(<NOUN : :>,<PREP :malgré :>)	4	Bref, *malgré la sincérité* du réalisateur, Bella Ciao est un film raté.
VMOD_POSIT1(<VERB : :>,<NOUN :déplaisir :>)	3,5	(...) cette comédie (...) *se déguste* sans *déplaisir*.

TABLE 9 – Exemple de concessifs

Après ce tour d'horizon, il est intéressant de se pencher sur la nature des éléments pertinents. Précédemment, nous avions limité notre étude aux adverbes et montré leur pertinence en tant que modifieurs de polarité. Au vu des résultats obtenus ici pour les autres catégories syntaxiques, ils semblent occuper un rôle central. Signalons tout de même que, dans la mesure où la catégorie des adverbes est relativement fermée, ils vont être plus fréquents et donc plus facilement déclarés significatifs par le test. Dans cet article, nous avons cherché à évaluer la pertinence des autres catégories syntaxiques. Les résultats sont moins nombreux, mais ils viennent compléter notre lexique et apportent donc une information non-négligeable.

1. Verbes : Très peu de verbes sont extraits et aucun n'est pertinent. Certaines structures verbales sont extraites cependant par l'intermédiaire d'autres relations de dépendance (des relations qui mettent en jeu des prépositions, par exemple). C'est le cas de l'expression [*il finit par - VB*]. On pourrait envisager d'autres expressions pertinentes comme [*être loin de - VB*] ou [*passer à côté de - NOM*]. Il serait intéressant d'étudier à quel point les verbes peuvent effectivement avoir un impact sur un terme polarisé.
2. Noms : Peu de noms se révèlent être des modifieurs pertinents. De nombreux noms ont pourtant été extraits. C'est la catégorie la plus fréquemment source d'erreur. Ils semblent avoir effectivement un impact important, mais cet impact est difficile à justifier et à juger.

3. Adjectifs : Les adjectifs sont souvent utilisés comme les indices principaux d'une polarité. L'extraction de nombreux adjectifs ici montre qu'ils peuvent aussi être des modifieurs, en particulier pour l'intensification.

Pour résumer, plus de 50% des modifieurs jugés pertinents correspondent à des structures contenant un adverbe (associé à un adjectif ou à un verbe polarisé). Environ 22% des modifieurs ne sont pas des adjectifs ou des adverbes (noms, prepositions, déterminants...). Ceux-ci sont principalement des inverseurs et des concessifs. Il semble donc que des stratégies plus diverses soient utilisées pour exprimer la concession et l'inversion que pour l'intensification. L'intensification agit en effet souvent de façon plus directe sur un terme polarisé grâce à une relation syntaxique locale. Au contraire, l'apport d'une nuance, quelle qu'elle soit, ou d'une inversion, s'exprimera plutôt au niveau de l'organisation du discours, ou grâce à des formulations plus complexes. Il serait intéressant de se pencher plus particulièrement sur ces phénomènes.

3.2.2 Erreurs ou incohérences de l'extraction

Lors de notre évaluation manuelle, un certain nombre de résultats se sont révélés être réellement inappropriés et ne pas avoir leur place dans le contexte d'une évaluation ou d'une opinion.
Tout d'abord, 17 éléments proviennent d'erreurs d'extraction ou d'erreurs induites par la méthodologie (erreurs de l'analyse syntaxique ou dans le lexique). Les scores de ces éléments sont généralement peu élevés. D'autre part, une cinquantaine de structures se sont avérées plus problématiques, dans la mesure où il est difficile de déterminer en quoi elles ont un impact sur la polarité d'un terme. Certaines font référence à des constructions très courantes, que l'analyse statistique a jugées significatives. C'est le cas du syntagme nominal introduit par le déterminant indéfini *un*, classé comme intensifieur avec un score de 9. Ces cas obtiennnent parfois des scores élevés. D'autres structures apparaissent dans des contextes divers. Il est alors difficile de leur définir un rôle sémantique clair. Enfin, nous avons considéré que 3 structures extraites effectivement modifiantes, reprises dans la table 10, ont été mal classées. Deux d'entre elles ont un score très faible.

Structure	Score	Type du modifieur
OBJ(<VERB : :>,<PRON :rien :>)	4	Concessif
NMOD_POSIT1(<NOUN : :>,<ADJ :bon :>)	1	Inverseur
NMOD_POSIT1(<NOUN : :>,<ADJ :efficace :>)	1	Inverseur

TABLE 10 – 3 structures mal classées

3.2.3 Autres phénomènes contexuels

A côté des cas corrects et des réelles incohérences, notre système met également en évidence des éléments qui jouent un certain rôle dans l'expression d'une polarité ou d'une évaluation. Ce ne sont pas des modifieurs, mais ils apportent un éclairage intéressant sur les phénomènes qui apparaissent dans le contexte de termes polarisés.

Parmi ces phénomènes contextuels particuliers, 19 éléments font en fait partie du vocabulaire du cinéma et désignent des caractéristiques du film soumises à jugement, et donc, dans ce cadre souvent associées à des termes positifs ou négatifs (table 11).

Structure	Exemple
NMOD_POSIT1(<NOUN : :>,<NOUN :effet :>) NMOD_POSIT1(<NOUN : :>,<NOUN :dialogue :>)	« surenchère *d'effets* » « pauvreté *des dialogues* », « musicalité *des dialogues* »

TABLE 11 – Constructions contenant des termes issus du vocabulaire du cinéma

Nous avons ensuite classé 30 éléments comme étant eux-mêmes polarisés. Comme notre système se base sur des structures lexico-syntaxiques, certains éléments polarisés s'avèrent être des structures complexes. C'est le cas par exemple du complément du nom "*de maître*", clairement positif, comme on peut le voir dans les expressions "*main de maître*" ou "*travail de maître*". Les lexiques de polarité classiques prennent souvent peu en compte ce type d'expressions à mots multiples, se concentrant sur les mots simples. Il serait avantageux d'utiliser également ces expressions polarisées pour une tâche automatique de fouille d'opinion.

Enfin, nous extrayons également un certain nombre de structures diverses (36) qui jouent un rôle dans l'expresson d'une évaluation. Il n'est pas possible de les considérer comme des modifieurs, dans le sens où elles n'ont pas un réel impact sur un terme polarisé, mais elles occupent une place importante dans le langage évaluatif. Il ne s'agit pas de structures figées plus larges, que l'analyseur syntaxique n'a pas pu restituer dans son ensemble, mais plutôt d'expressions ou formulations diverses souvent utilisées qui aceptent de nombreuses variantes. Ici, elles sont souvent le reflet de formulations utilisées pour juger un film. Des exemples en sont donnés dans la table 12. Ces structures peuvent avoir différents rôles. Certaines sont plutôt positives et négatives dans le strict contexte du cinéma. D'autres semblent plus servir à introduire un jugement, sans être en elles-mêmes polarisées.

Structure	Exemple
PREPOBJ(<VERB : :>,<PREP :à :>)	« c'est à voir » ; « c'est à découvrir »
OBJ(<VERB : :>,<PRON :ça :>)	« (...) n'a pas mérité ça » ; « on a vu ça 100 fois » ; « on n'a jamais vu ça »
NMOD_POSIT1(<NOUN : :>,<NOUN :série :>)	« comédie de série b » ; « série z » ; « série télé »
OBJ(<VERB : :>,<NOUN :intérêt :>)	« n'offre aucun intérêt » ; « éveille l'intérêt » ; « capte l'intérêt »

TABLE 12 – Constructions évaluatives diverses

3.2.4 Performance de l'extraction

Au terme de cette étude, une première remarque doit être faite sur le nombre d'extractions. Il s'avère en effet moins important que l'on aurait pu le supposer au départ, et peu de structures obtiennent un score élevé (30 éléments avec un score de 5 ou plus, et 16 avec 7 ou plus). Cela s'explique en partie par le fait que les relations de dépendance que l'on extrait, relativement précises, ont chacunes des fréquences peu élevées dans le corpus, ce qui complique l'analyse statistique.
En ce qui concerne la performance proprement dite du système, 87 structures ont donc été considérées comme bien classées sur les 243 structures proposées, soit environ 35%. Il nous

faut mesurer ce résultat par le fait qu'aucun seuil minimal de score n'a été appliqué sur la liste complète. Les structures qui obtiennent un score élevé sont cependant relativement pertinentes. Ainsi sur les 30 structures ayant un score de 5 ou plus, 23 ont été jugées pertinentes (et 14 structures pertinentes sur 16 pour un score égal à 7 ou plus). D'autre part, la méthodologie n'extrait finalement qu'environ 30% de réelles incohérences. Certaines extractions sont en effet pertinentes dans le langage de l'évaluation, comme nous l'avons montré précédemment. Enfin, le système a tendance à déterminer de façon correcte le type du modifieur lorsque celui-ci est pertinent. Trois éléments seulement se révèlent mal classés. Ces résultats sont récapitulés dans la table 13.

		Intens.	Invers.	Concess.	Scores ex æquo	Total	
cas pertinents		**37**	**22**	**27**	**1**	87	**35,80%**
Autres	vocabulaire du cinéma	12	2	5	0	19	
phénomènes	expressions d'évaluation	23	8	4	1	36	**34,98%**
contextuels	polarisés eux-mêmes	13	11	6	0	30	
Erreurs	mal classé	0	2	1	0	3	
	non-pertinent	15	13	7	3	38	**29,22%**
	erreurs dues à l'extraction ou à la méthodologie	5	11	1	0	17	
	cas difficile à trancher	3	5	1	4	13	
		108	**74**	**52**	**9**	**243**	

TABLE 13 – Répartition des modifieurs évalués manuellement

En conclusion, notre méthodologie nous a permis d'identifier des modifieurs pertinents en dépassant le cadre de l'adverbe et d'extraire des éléments variés, typiques du langage de l'évaluation. Elle met en avant des phénomènes peu traités dans d'autres travaux, et plus détaillés, dans la mesure où l'on traite de structures lexico-syntaxiques et non de termes simples. Les résultats ont permis de mettre en avant deux fonctionnements un peu différents : l'intensification, portée par des adjectifs et des adverbes, souvent avec un impact direct sur un terme polarisé, et l'expression de l'inversion, de l'atténuation ou d'une nuance, portée par des contructions plus diverses et complexes.

4 Conclusion et perspectives

Le travail présenté ici identifie des modifieurs de polarité grâce à l'étude de structures lexico-syntaxiques qui mettent en jeu un terme polarisé. Cette étude se révèle être un bon point de départ pour se rendre compte concrètement de divers phénomènes de modification. Elle met en avant en particulier, de par la méthodologie utilisée, les éléments qui ont un impact direct et local sur un terme d'une certaine polarité. Il s'avère que ces éléments sont relativement limités. D'autres stratégies, plus diverses et complexes, apparaissent. Ces stratégies expriment souvent une atténuation ou une inversion, se situent plutôt au niveau de l'organisation du discours et associent fréquemment plusieurs termes polarisés.

D'une part, il sera nécessaire de compléter cette analyse qualitative par une analyse quantitative en intégrant les modifieurs extraits ici dans un système de fouille d'opinion. L'objectif est de savoir si la prise en compte de ces structures modifiantes améliore la détection de la polarité d'un syntagme ou d'une phrase.
D'autre part, cette étude a montré l'intérêt d'approfondir la recherche sur les stratégies de

modification sans se limiter aux relations de dépendance syntaxique. Deux pistes, en particulier, seraient intéressantes à explorer : (1) étude approfondie des phénomènes d'inversion de la polarité, (2) définition de règles de composition (afin de pouvoir déterminer la polarité d'une phrase qui contient plusieurs termes polarisés).

Références

Aït-Mokhtar, S., Chanod, J. et Roux, C. (2002). Robustness beyond shallowness : Incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144.

Boubel, N. (2011). Extraction automatique de modifieurs de valence affective dans un texte. etude exploratoire appliquée au cas de l'adverbe. *In Travaux du Cercle belge de Linguistique*, volume 6.

Boubel, N. et Bestgen, Y. (2011). Une procédure pour identifier les modifieurs de la valence affective d'un mot dans des textes. *In Actes de TALN11*, volume 2, pages 137–142, Montpellier.

Giannakidou, A. (1998). *Polarity sensitivity as (non) veridical dependency*, volume 23. J. Benjamins.

Kennedy, A. et Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Musat, C. et Trausan-Matu, S. (2010). The impact of valence shifters on mining implicit economic opinions. *Artificial Intelligence : Methodology, Systems, and Applications*, pages 131–140.

Pang, B. et Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).

Petrakis, S., Klenner, M., Ailloud, E. et Fahrni, A. (2009). Composition multilingue de sentiments. *In Actes de TALN2009*, Senlis.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. et Crystal, D. (1985). *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. et Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Vernier, M., Monceaux, L., Daille, B. et Dubreil, E. (2009). Catéorisation des évaluations dans un corpus de blogs multi-domaine. http ://hal.archives-ouvertes.fr/hal-00405407/fr/.

Vincze, N. et Bestgen, Y. (2011). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. *In Actes de TALN11*, volume 1, pages 223–234, Montpellier.

Wilson, T., Wiebe, J. et Hoffmann, P. (2009). Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Zaenen, A. et Polanyi, L. (2004). Contextual valence shifters. *In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 106–111.

Zwarts, F. (1995). Nonveridical contexts. *Linguistic Analysis*, 25:286–312.

Une plate-forme générique et ouverte pour l'acquisition des expressions polylexicales

Carlos Ramisch
LIG-GETALP, Grenoble, France
INF-UFRGS, Porto Alegre, Brésil
Carlos.Ramisch@imag.fr

RÉSUMÉ

Cet article présente et évalue une plate-forme ouverte et flexible pour l'acquisition automatique d'expressions polylexicales (EPL) à partir des corpus monolingues. Nous commençons par une motivation pratique suivie d'une discussion théorique sur le comportement et les défis posés par les EPL dans les applications de TAL. Ensuite, nous décrivons les modules de notre plate-forme, leur enchaînement et les choix d'implémentation. L'évaluation de la plate-forme a été effectuée à travers une applications : la lexicographie assistée par ordinateur. Cette dernière peut bénéficier de l'acquisition d'EPL puisque les expressions acquises automatiquement à partir des corpus peuvent à la fois accélérer la création et améliorer la qualité et la couverture des ressources lexicales. Les résultats prometteurs encouragent une recherche plus approfondie sur la manière optimale d'intégrer le traitement des EPL dans de nombreuses applications de TAL, notamment dans les systèmes traduction automatique.

ABSTRACT

An Open and Generic Framework for the Acquisition of Multiword Expressions

In this paper, we present and evaluate an open and flexible methodological framework for the automatic acquisition of multiword expressions (MWEs) from monolingual textual corpora. We start with a pratical motivation followed by a theoretical discussion of the behaviour and of the challenges that MWEs pose for NLP applications. Afterwards, we describe the modules of our framework, the overall pipeline and the design choices of the tool implementing the framework. The evaluation of the framework was performed extrinsically based on an application : computer-assisted lexicography. This application can benefit from MWE acquisition because the expressions acquired automatically from corpora can both speed up the creation and improve the quality and the coverage of the lexical resources. The promising results of previous and ongoing experiments encourage further investigation about the optimal way to integrate MWE treatment into NLP applications, and particularly into machine translation systems.

MOTS-CLÉS : Expressions polylexicales, extraction lexicale, lexique, mesures d'association, corpus, lexicographie.

KEYWORDS: Multiword expression, lexical extraction, lexicon, association measures, corpus, lexicography.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 137–149,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le terme *expression polylexicale* (EPL, en anglais *multiword expression*) comprend un grand nombre de phénomènes linguistiques qui engendrent des constructions variées telles que les expressions idiomatiques (*payer les yeux de la tête*), les expressions figées (*a priori*), les noms composés (*appareil photo*), les constructions à verbe support (*rendre visite*), etc. Il n'existe pas une définition unique et communément acceptée pour le terme *expression polylexicale*, car il peut être défini comme une « combinaison arbitraire et récurrente de mots » (Smadja, 1993) ou « une unité syntaxique et sémantique dont le sens exact ou la connotation ne peuvent pas être dérivés directement et sans ambiguïté du sens ou de la connotation de ses composantes » (Choueka, 1988) ou simplement comme une « interprétation idiosyncrasique qui dépasse la limite du mot (ou les espaces) » (Sag *et al.*, 2002), avec la propriété qu'elle « doit être répertoriée dans un lexique » (Evert, 2004, p 17.).

La tendance que les mots ont à s'attirer mutuellement, c'est-à-dire le phénomène clé derrière le concept d'EPL, a lieu dans la zone floue entre le lexique et la grammaire. Cela constitue un véritable défi pour les systèmes de TAL classiques. De plus, les EPL sont omniprésentes dans une langue, figurant fréquemment dans le langage oral et écrit de tous les jours, ainsi que dans les communications spécialisées techniques et scientifiques. Parmi les caractéristiques notables des EPL décrites dans la littérature, les plus importantes sont :

- **Caractère arbitraire** : Parfois, des constructions syntaxiquement et sémantiquement valides ne sont pas du tout naturelles simplement parce que les locuteurs natifs de la langue ne les utilisent pas. Même si ces constructions sont tout à fait compréhensibles, elles paraissent étranges et constituent des marqueurs d'un usage non natif. Cela rend l'apprentissage des EPL difficile pour les apprenants d'une langue qui, malgré leur connaissance du lexique et des règles grammaticales générales, n'ont pas assez d'expérience sur l'usage de cette langue. Smadja (1993, p. 143 à 144) illustre cela en présentant huit façons différentes de se référer à l'indice Dow Jones de la bourse de New York, dont seulement quatre sont acceptables.
- **Institutionnalisation** : Les EPL sont récurrentes, car elles correspondent à des façons habituelles de s'exprimer. Jackendoff (1997) estime qu'elles correspondent à la moitié des entrées du lexique d'un locuteur natif. Sag *et al.* (2002) remarquent que cela pourrait être une sous-estimation si l'on prend en compte les EPLs dans les domaines spécialisées, où elles constituent le noyau des connaissances exprimées et représentées.
- **Non-compositionnalité** : Le sens de l'expression entière ne peut pas toujours être déduit directement des sens de ses parties. Par conséquent, la compositionnalité des EPL varie dans un continuum allant des expressions complètement compositionnelles (*appareil photo*) à celles qui sont complètement opaques/idiomatiques (*casser sa pipe*).
- **Hétérogénéité** : Les EPLs sont difficiles à définir car elles englobent une grande quantité de phénomènes. Cela les rend difficiles à traiter par les applications de TAL, qui ne peuvent pas utiliser une approche unifiée. Souvent, les applications de TAL utilisent une des multiples typologies ou schémas de classification existants [1].
- **Non-substituabilité** : Il n'est pas possible de remplacer une partie d'une EPL par un mot proche ou équivalent (synonyme, hyperonyme, etc). Cette propriété motive la notion d'*anti-collocation* (Pearce, 2001), qui correspond à une combinaison de mots maladroite ou inhabituelle (*café corsé* vs *?café consistant*).

1. Par exemple, Smadja (1993) classe les EPL selon leur fonction syntaxique dans la phrase alors que Sag *et al.* (2002) les classent en fonction de leur degré de flexibilité (syntaxique).

$\texttt{en}_{\text{SRC}}$	*I* **paid** *my poor parents* **a visit**
$\texttt{pt}_{\text{TA}}$	*Eu* **pago** *os meus pais pobres* **uma visita**
$\texttt{pt}_{\text{REF}}$	*Eu* **fiz uma visita** *aos meus pobres pais*
$\texttt{fr}_{\text{TA}}$	*J'ai* **payé** *mes pauvres parents* **une visite**
$\texttt{fr}_{\text{REF}}$	*J'ai* **rendu visite** *à mes pauvres parents*
$\texttt{en}_{\text{SRC}}$	*Students pay* **an arm and a leg** *to park on campus*
$\texttt{pt}_{\text{TA}}$	*Estudantes pagam* **braço e uma perna** *para estacionar no campus*
$\texttt{pt}_{\text{REF}}$	*Estudantes pagam* **os olhos da cara** *para estacionar no campus*
$\texttt{fr}_{\text{TA}}$	*Les étudiants paient* **un bras et une jambe** *pour se garer sur le campus*
$\texttt{fr}_{\text{REF}}$	*Les étudiants paient* **les yeux de la tête** *pour se garer sur le campus*
$\texttt{en}_{\text{SRC}}$	*It shares the* **translation-invariance and homogeneity properties** *with the central moment*
$\texttt{pt}_{\text{TA}}$	*Ele compartilha a* **tradução invariância e propriedades de homogeneidade** *com o momento central*
$\texttt{pt}_{\text{REF}}$	*Ele compartilha as* **propriedades de invariância por translação e de homogeneidade** *com o momento central*
$\texttt{fr}_{\text{TA}}$	*Il partage la* **traduction-invariance et propriétés d'homogénéité** *avec le moment central*
$\texttt{fr}_{\text{REF}}$	*Il partage les* **propriétés d'invariance par translation et d'homogénéité** *avec le moment central*

TABLE 1 – Phrases contenant des EPL qui posent problème pour un système de TA empirique.

– **Lexicalisation** : Quelque part dans les applications de TAL, l'information, indiquant qu'un ensemble ou séquence de mots est « indissociable », doit être disponible. Les concepteurs du système doivent donc choisir l'endroit où chaque EPL sera représentée : on ne peut pas les énumérer une à une dans le lexique (sous-génération), ni toutes les inclure dans les règles de la grammaire comme des combinaisons libres (sur-génération). Identifier le degré de lexicalisation de chaque (classe d') EPL est important pour toutes les tâches d'analyse et de génération en TAL.

Dans ce travail, nous adoptons la définition proposée par Calzolari *et al.* (2002), qui définissent les EPL comme :

> ... différents phénomènes liés qui peuvent être décrits comme une séquence [ou groupe] de mots [2] à voir comme une unité à un certain niveau d'analyse linguistique.

Cette définition générique et volontairement vague peut être instanciée selon les besoins des applications. Par exemple, le tableau 1 [3] montre des erreurs générées par un système de traduction automatique (TA) empirique. Pour ce système, une EPL est une séquence de mots qui doit être traduite comme une unité. Sinon, le système génèrera des erreurs, c'est-à-dire des constructions agrammaticales ou artificielles (phrase 1), des traductions littérales maladroites d'expressions idiomatiques (phrase 2) et des mauvais choix lexicaux et syntaxiques dans les textes spécialisés (phrase 3). Dans un système de TA experte, ces EPL seraient typiquement représentées comme

2. Cette définition est réductrice car elle ne considère que les séquences de mots. Nous étendons la notion de séquence vers des groupes de mots, c'est-à-dire des séquences contigües mais aussi des groupes de mots non adjacents liés syntaxiquement et/ou sémantiquement par leur contexte d'occurrence.

3. Source en anglais ($\texttt{en}_{\text{SRC}}$) à partir du web. Traductions automatiques (TA) en portugais (`pt`) et en français (`fr`) fournies par Google Translate (`http://translate.google.com/`) le 18 février 2012. Traductions de référence (REF) fournies par des locuteur natifs.

une entrée à part entière dans le lexique (ou plus précisément, dans le dictionnaire de tournures), sans quoi les mêmes erreurs pourraient survenir.

Ces exemples illustrent l'importance de traiter les EPL dans les systèmes de TA. Plus généralement, le traitement d'EPL peut accélérer et aider à éliminer des ambiguïtés dans de nombreuses applications de TAL, par exemple :

- **Lexicographie** : Church et Hanks (1990) utilisent un environnement lexicographique comme cadre d'évaluation, en comparant la recherche manuelle et intuitive avec la méthode automatique proposée.
- **Reconnaissance optique de caractères** : Supposons qu'un système de reconnaissance optique de caractères a autant de chances de reconnaître les mots *poule* et *poêle* dans *poule/poêle élevée en plein air*. La connaissance d'une EPL utilisant la première option l'aide à choisir [4].
- **Désambiguïsation lexicale** : Les EPL ont tendance à être moins polysémiques que leurs composantes isolées. Finlayson et Kulkarni (2011) illustrent que le mot *world* possède neuf significations possibles dans WordNet 1.6, *record* en possède quatorze, mais *world record* en a seulement une.
- **Étiquetage morpho-syntaxique et analyse syntaxique** : des publications en analyse et étiquetage morpho-syntaxique indiquent que les EPL peuvent aider à éliminer les ambiguïtés syntaxiques (Seretan, 2008; Constant et Sigogne, 2011).
- **Recherche d'information** : Lorsqu'une EPL telle que *pop star* est indexée comme une unité dans un système de recherche d'information, la précision du système s'améliore (Acosta *et al.*, 2011).
- **Apprentissage de langues étrangères** : Puisque les EPL sont très difficiles à apprendre pour des locuteurs non natifs, les dictionnaires et les ressources pédagogiques contenant des entrées polylexicales peuvent être très utiles dans l'enseignement des langues. Un exemple d'une telle ressource est le dictionnaire italien de collocations décrit par Spina (2010).
- **Traduction automatique** : les expériences montrent que l'inclusion d'EPL dans les systèmes de TA améliore la qualité de la traduction (Carpuat et Diab, 2010; Stymne, 2011).

Cet article porte sur le traitement des EPL, allant de l'acquisition automatique à leur intégration dans des applications. Dans un premier temps, nous présentons un bref tour d'horizon de la bibliographie en acquisition automatique d'EPL (§ 2). Dans un deuxième temps, nous décrivons le modèle conceptuel et la plate-forme logicielle développée pour l'acquisition des EPL (§ 3). Par la suite, nous présentons la validation de cette plate-forme dans le cadre de la lexicographie assistée par ordinateur (§ 4). Les expériences en cours montrent des résultats prometteurs mais des enquêtes supplémentaires seront nécessaires pour mieux comprendre les apports des EPL aux systèmes de TAL et en particulier aux systèmes de TA (§ 5).

2 État de l'art

Plusieurs projets ont eu pour objectif de compiler manuellement des ressources lexicales comprenant des EPL. Par exemple, pour le français, le LADL constitue depuis de nombreuses années des listes de noms composés compatibles avec le formalisme/outil Unitex (Gross, 1986). Ce formalisme, cependant, présente des avantages et des inconvénients en ce qui concerne le

4. En fait, cela se fait souvent à l'aide des modèles de langage à *n*-grammes, mais les *n*-grammes ne peuvent pas modéliser adéquatement tous les types d'EPL. Prenons l'exemple d'une expression très flexible telle que **take** *patient risk factors and comfort* **into account**.

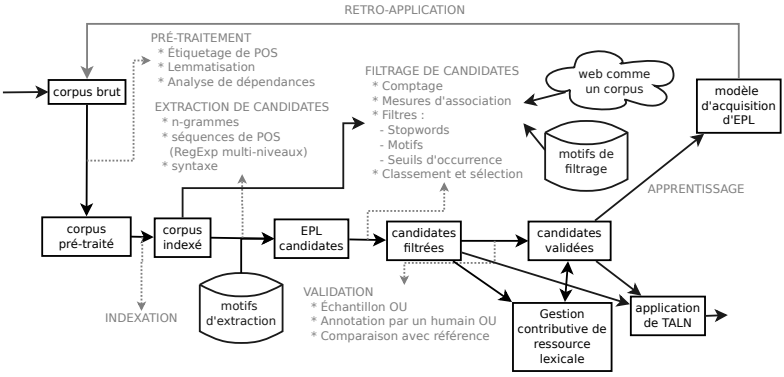


FIGURE 1 – Schéma du mwetoolkit — plate-forme d'acquisition d'EPL à partir de corpus.

compromis entre pouvoir d'expression et temps d'apprentissage (Graliński *et al.*, 2010). La compilation manuelle de ressources lexicales étant très onéreuse, la recherche des dernières années s'est concentrée sur l'acquisition automatique des EPL, dans le but d'accélérer le travail des lexicographes.

Parmi les premiers travaux développant des méthodes automatiques d'identification d'EPL, il y a celui de Smadja (1993). Il a proposé et développé l'outil Xtract pour l'extraction de collocations à partir de textes à travers une combinaison de *n*-grammes et d'une mesure d'information mutuelle. Sur des textes communs, Xtract a une précision de l'ordre de 80%. Depuis lors, de nombreux progrès ont été réalisés, soit sur l'extraction d'EPLs en général (Dias, 2003), soit en se concentrant sur un type d'EPL spécifique, tel que les collocations (Pearce, 2002), les verbes à particule (Ramisch *et al.*, 2008) et les noms composés (Keller et Lapata, 2003).

Une approche indépendante du type devenue très populaire consiste à utiliser des mesures d'association (Evert et Krenn, 2005), qui ont été appliquées avec des degrés variables de succès. Un des avantages de cette approche est qu'elle est indépendante de la langue. Ceci est particulièrement important car les travaux sur l'anglais prédominent (Pearce, 2002; Ramisch *et al.*, 2008), même si des travaux sur les EPL dans plusieurs langues ont été publiés, par exemple Dias (2003) pour le portugais, Evert et Krenn (2005) pour l'allemand et de Cruys et na Villada Moirón (2007) pour le néerlandais.

3 Acquisition d'EPL

Nous proposons une nouvelle plate-forme, décrite dans la figure 1. Elle intègre de multiples techniques parmi celles citées ci-dessus et couvre l'ensemble du pipeline de l'acquisition d'EPL. Cette plate-forme a été mise en œuvre dans un outil libre appelé mwetoolkit [5]. Ici, nous ne résumons que les aspects principaux de la plate-forme, qui a été décrite précédemment dans

5. `http://mwetoolkit.sourceforge.net/`

d'autres publications (Ramisch *et al.*, 2010a,b; Araujo *et al.*, 2011).

Le point de départ de l'extraction est un corpus monolingue de texte brut. Avant l'application du mwetoolkit, il est possible de prétraiter le corpus, à condition que les outils de prétraitement soient disponibles pour la langue cible, en l'enrichissant avec des étiquettes morpho-syntaxiques, des lemmes et de la syntaxe de dépendance. Pour simplifier la représentation des données, nous avons choisi d'utiliser seulement les informations d'analyse syntaxique qui peuvent être représentées comme un attribut du mot, excluant ainsi toute forme d'arborescence. Néanmoins, il est possible de représenter les relations syntaxiques de dépendance à travers une paire ⟨ type de relation, mot père ⟩ qui attribue à chaque mot le mot dont il dépend et l'étiquette qui décrit le type de relation (par exemple, objet direct, modificateur, sujet).

Ensuite, à l'aide des connaissances linguistiques d'experts, de l'intuition, de l'observation empirique et/ou des exemples, il faut décrire les EPL cibles à travers des motifs multiniveaux dans un formalisme similaire aux expressions régulières. Il est possible d'utiliser plusieurs niveaux d'analyse simultanément, par exemple : on veut extraire tous les noms qui sont l'objet direct du verbe *cuisiner*. L'application de ces motifs sur un corpus indexé génère une liste d'EPL candidates.

Pour le filtrage, un grand nombre de méthodes est disponible, allant de simples seuils de fréquence à des listes de mots interdits (*stopwords*), en passant par des mesures d'association plus sophistiquées. Les mesures d'association disponibles (score t, coefficient de Dice, information mutuelle et rapport de vraisemblance) sont, à l'exception de la dernière, applicables à des candidates de longueur arbitraire, ce qui n'est pas toujours le cas dans les outils de linguistique de corpus traditionnels.

Finalement, les candidates résultantes filtrées sont soit directement injectées dans une application de TAL, soit validées manuellement avant l'application. Il est possible de réutiliser les candidates pour l'apprentissage d'un modèle. Ce modèle sera appliqué sur de nouveaux corpus afin d'identifier et d'extraire automatiquement de nouvelles EPL selon les caractéristiques de celles acquises auparavant.

À ce jour, il n'y a pas de consensus sur une méthode optimale d'acquisition d'EPL. Il n'est donc pas possible de déterminer s'il existe une méthode unique pour toutes les EPL, ou alors s'il faudrait chercher une combinaison de méthodes ou un sous-ensemble de méthodes qui fonctionne mieux pour un type d'EPL en particulier. La plupart des travaux récents se concentrent sur l'extraction d'EPL à partir de corpus prétraités (Seretan, 2008) et sur le filtrage automatique et le tri grâce à des mesures d'association (Evert, 2004; Pecina, 2010), mais peu d'auteurs fournissent une vue d'ensemble de la chaîne de traitement d'EPL.

Un des avantages de la plate-forme et de l'outil proposés dans cet article est qu'ils modélisent le processus d'acquisition par des tâches modulaires qui peuvent être enchaînés de plusieurs façons. Chaque tâche a de multiples techniques disponibles pour leur accomplissement. Par conséquent, il est hautement personnalisable et permet un paramétrage détaillé selon les types d'EPL cibles, contrairement à des outils similaires tels que NSP [6] et UCS. [7]

De plus, les techniques développées ne dépendent pas d'une longueur fixe d'expressions candidates (par exemple, les paires de mots) ni sur l'hypothèse de contiguïté. Grâce à cette souplesse, cette méthodologie peut être facilement appliquée à un grand nombre de langues, de types

6. http://search.cpan.org/dist/Text-NSP
7. http://www.collocations.de/software.html

d'EPL et de domaines, ne dépendant pas d'un formalisme donné ou d'un outil. [8] Pour une langue donnée, si certains outils de prétraitement tels que les étiqueteurs morpho-syntaxiques et/ou analyseurs sont disponibles, il n'y a pas de raison pour ne pas s'en servir. Dans ce cas, intuitivement, les résultats seront bien meilleurs que sur du texte brut non analysé. Mais comme toutes les langues ne disposent pas de ces outils, la méthodologie a été conçue pour être appliquée même dans le contexte où aucun outil de prétraitement n'est disponible. Dans l'avenir, nous voudrions valider cette hypothèse en réalisant l'extraction d'EPL sur un corpus dans une langue pauvrement dotée.

4 Résultats

Ici, nous présentons les résultats de l'évaluation de notre plate-forme dans le cadre de la lexicographie assistée par ordinateur. Tout d'abord, nous introduisons une nouvelle classification pour les différents axes d'évaluation de l'acquisition automatique d'EPL (§ 4.1). Ensuite, nous résumons l'évaluation quantitative et qualitative extrinsèque de la plate-forme d'acquisition d'EPL proposée précédemment (§ 4.2).

4.1 Évaluation de l'acquisition d'EPL

Comme l'a souligné Pecina (2005), « l'évaluation des méthodes d'extraction de collocations est une tâche complexe. D'une part, les différentes applications exigent différents [...] seuils. D'autre part, les méthodes donnent des résultats distincts selon les intervalles de leurs scores d'association ». Nous structurons l'évaluation de l'acquisition d'EPL selon les critères suivants :

- **Selon la nature des mesures :**
  - **Quantitative** : consiste à évaluer l'acquisition à travers des mesures objectives telles que la précision, le rappel, la F-mesure et la précision moyenne. Alors que de nombreux articles calculent uniquement la précision sur les premières *n* EPL retournées, il faut aussi évaluer le rappel, ce qui est rarement fait. Néanmoins, cela est d'une importance capitale dans l'attribution de l'utilité d'une méthode. Une méthode très précise qui n'extrait qu'une douzaine d'expressions quand il y a en réalité des milliers d'expressions à récupérer n'est pas plus efficace que la force brute ou la recherche manuelle. La quantité d'EPL découvertes est un facteur aussi important que leur qualité, et il est difficile d'évaluer combien d'EPL sont « suffisantes » pour que l'acquisition automatique soit utile (Villavicencio *et al.*, 2005; Church, 2011).
  - **Qualitative** : le but de l'évaluation qualitative est d'obtenir une compréhension approfondie des EPL obtenues et des erreurs commises par la méthode d'acquisition. Cela consiste à observer les motifs récurrents en analysant les listes résultantes en termes de leur adéquation à l'application cible. Cette évaluation est souvent itérative, c'est-à-dire que les améliorations possibles sont retro-appliquées sur la méthode d'acquisition, avec une nouvelle évaluation, et ainsi de suite.
- **Selon l'objectif de l'acquisition :**

8. Toutefois, la méthodologie est conçue pour traiter les langues qui utilisent des espaces pour séparer les mots. Ainsi, lorsque l'on travaille avec du chinois, du japonais, ou même avec des noms composés en allemand, un prétraitement supplémentaire est nécessaire.

Langue	Type	Corpus	Mots	Cand.	EPL	Publication
anglais	VàP	Europarl-frg	13M	5,3K	875	(Ramisch *et al.*, 2012)
grec	NC	Europarl	26M	5K	815	(Linardaki *et al.*, 2010)
portugais	EV	PLN-BR-FULL	29M	407K	773	(Duran *et al.*, 2011)

TABLE 2 – Acquisition d'EPL appliquée à la lexicographie assistée par ordinateur.

  - **Intrinsèque** : la plupart des résultats d'évaluation publiés dans les références bibliographiques citées dans cet article sont intrinsèques, c'est-à-dire, ils considèrent les EPL en elles-mêmes, directement, en tant que produit final d'un processus. Même si elle présente de nombreuses limitations, l'évaluation intrinsèque donne toujours une estimation de qualité qui permet une comparaison inter-méthodes fiable.
  - **Extrinsèque** : il est souvent plus facile d'estimer la qualité du résultat pour une tâche de TAL concrète que pour une liste d'EPL dont on ne connait pas l'application. Ainsi, l'évaluation extrinsèque, c'est-à-dire l'utilisation des EPL dans une application de TAL extérieure, peut être très concluante pour démontrer si les EPL acquises sont utiles.
- **Selon le type d'EPL :**
  - **Fondée sur les types** : certaines expressions ne sont pas ambiguës et peuvent être annotées hors contexte, comme des entrées dans un lexique. C'est souvent le cas quand il s'agit de noms composés, de termes techniques et de constructions à verbe support. La décision de savoir si une séquence de mots est une EPL, dans ce type d'annotation, est indépendante du contexte dans lequel elle apparaît.
  - **Fondée sur les occurrences** : cette annotation doit être effectuée quand les EPL cibles sont ambiguës, comme les verbes à particule et les expressions idiomatiques. Hors contexte, il est impossible de dire si les mots doivent être traités comme une unité ou indépendamment.

D'une part, les résultats de ces *évaluations intrinsèques* sont souvent vagues ou peu concluants. Bien qu'ils fassent la lumière sur les paramètres optimaux pour un scénario donné, ils sont difficiles à généraliser et ne peuvent pas être directement appliquées à d'autres configurations. La qualité des EPL acquises mesurée par des critères objectifs dépend de la langue, du domaine et du type de la construction cible, ainsi que de la taille et du genre du corpus, des ressources déjà disponibles [9], des filtres appliqués, des étapes de prétraitement, ...

D'autre part, l'*évaluation extrinsèque* consiste à insérer les EPL acquises dans des applications de TAL réelles et à évaluer l'impact de ces nouvelles données sur la performance globale du système. Ainsi, une contribution originale du présent travail est l'application de l'évaluation extrinsèque de l'acquisition d'EPL sur une étude de cas : la lexicographie assistée par ordinateur. Notre objectif à long terme est d'étudier (1) quel est l'impact de ces EPL sur les applications de TAL en général et (2) la (ou les) meilleure(s) méthodes(s) pour les intégrer dans le pipeline complexe de l'application cible.

9. Il est inutile d'acquérir des EPL déjà présentes dans le dictionnaire.

4.2 Lexicographie assistée par ordinateur

Nous avons travaillé pour cette évaluation en collaboration avec des collègues linguistes et lexicographes expérimentés, dans le but de créer de nouvelles ressources lexicales contenant des EPL. Les langues des ressources sont l'anglais, le grec et le portugais. Le tableau 2 résume les résultats de chaque évaluation.

Nous avons extrait les verbes à particule (VàP) à partir d'un fragment de la partie anglaise du corpus Europarl. [10] Nous avons considéré un VàP comme étant formé par un verbe (à l'exception de *be* et *have*) suivi d'une particule prépositionnelle [11] éloignée d'au plus 5 mots après le verbe. [12] Nous avons obtenu 5 302 candidates à VàP qui apparaissent plus d'une fois dans le corpus. L'évaluation de ces candidates a été effectuée de façon automatique, en les comparant avec un dictionnaire de référence. Parmi les candidates, 875 ont été classifiées comme de véritables VàP. Cette évaluation quantitative et fondée sur les types est une première étape d'une expérience en cours sur l'intégration de ces constructions dans un système de traduction automatique.

Pour le grec, il existe une vaste littérature portant sur les propriétés linguistiques des EPL, mais les approches informatiques sont encore limitées (Fotopoulou *et al.*, 2008). Dans nos expériences, nous avons extrait de la partie grecque du corpus Europarl, étiquetée morpho-syntaxiquement, des noms composés (NC) correspondant aux motifs suivants : adjectif-nom, nom-nom, nom-déterminant-nom, nom-préposition-nom, préposition-nom-nom, nom-adjectif-nom et nom-conjonction-nom. Les candidates ont été comptées dans deux corpus et classées par quatre mesures d'association. Les premières 150 candidates selon chaque mesure d'association ont été évaluées intrinsèquement par trois locuteurs natifs. Ainsi, chaque annotateur a jugé environ 1 200 candidates. Finalement, les annotations ont été combinées, entraînant la création d'un lexique avec 815 EPL nominales en grec.

L'objectif du travail avec les expressions verbales (EV) en portugais était de réaliser une analyse qualitative de ces constructions bien comme de la méthode d'acquisition. Nous avons étiqueté morpho-syntaxiquement le corpus PLN-BR-Full [13] et ensuite nous avons effectué quelques itérations d'une phase d'évaluation qualitative par un lexicographe expérimenté suivie d'une nouvelle phase d'extraction. Finalement, nous avons extrait des séquences de mots correspondant aux motifs verbe-[déterminant]-nom-préposition, verbe-préposition-nom, verbe-[préposition/déterminant]-adverbe et verbe-adjectif. Le processus d'extraction a ainsi conduit à une liste de 407 014 candidates qui ont ensuite été filtrées avec des mesures d'association.

Durant l'évaluation quantitative fondée sur les types, un annotateur humain expert a validé manuellement 12 545 candidates, parmi lesquelles 699 ont été annotées comme des EV compositionnelles et 74 comme des EV idiomatiques. Ensuite, une analyse fine de chaque motif d'extraction a été réalisée dans le but de trouver des corrélations entre la flexibilité syntaxique et les propriétés sémantiques telles que la compositionnalité. Les retours fournis par le lexicographe ont montré que l'outil est plus flexible et plus efficace que les concordanciers traditionnels, car il permet d'identifier des EPL candidates sans fixer une liste pré-définie de verbes support. Ainsi, des constructions non prototypiques ont pu être identifiées grâce à l'utilisation de notre plate-forme.

10. `http://statmt.org/europarl`
11. *up, off, down, back, away, in, on.*
12. Même si la particule pourrait apparaître plus loin que 5 mots après le verbe, de tels cas sont suffisamment rares pour être ignorées dans cette expérience.
13. `www.nilc.icmc.usp.br/plnbr`

5 Conclusions et perspectives

Dans cet article, nous avons décrit une plate-forme pour l'acquisition des EPL à partir de corpus monolingues. La contribution principale de ce travail réside dans le fait qu'il représente une étape vers l'intégration des EPL automatiquement extraites dans des applications réelles. Premièrement, nous avons proposé un *cadre méthodologique* unifié, ouvert et flexible pour l'acquisition automatique des EPL. Deuxièmement, nous avons effectué une vaste *évaluation de l'acquisition d'EPL*, afin de disséquer l'influence des différents types de ressources utilisées dans l'acquisition sur la qualité des EPL résultantes. De plus, nous avons proposé une nouvelle taxonomie qui classifie les résultats de l'évaluation d'acquisition d'EPL selon trois axes principaux.

Nous sommes actuellement en train de développer des méthodes pour l'intégration des EPL verbales acquises automatiquement dans un système de TA empirique. Les EPL verbales sont des constructions syntaxiquement flexibles. Par conséquent, elles constituent un défi pour les systèmes de TA empirique fondés sur les séquences de mots. Après quelques expériences préliminaires, nous avons constaté le besoin d'appliquer une méthode identification fondée sur la syntaxe. En même temps, nous analysons les alignements lexicaux et les entrées de la table de traduction du système sur un ensemble d'EPL prototypiques, afin de mieux comprendre l'impact des EPL sur les résultats du système. Enfin, nous voulons réaliser des expériences sur la simplification d'EPL, par exemple, le remplacement d'un verbe polylexical comme *come back* par sa forme simple *regress*. Ainsi, l'intuition qui veut que l'on fasse ressembler la langue source à la langue cible facilite la tâche d'apprentissage d'alignements inter-langue (Stymne, 2011). Comme ces améliorations dépendent du paradigme de TA choisi, nous voulons également évaluer les stratégies pour l'intégration des EPL verbales dans les systèmes de TA experte tels que ITS2 (Wehrli et Ramluckun, 1993) et Etap-3 (Apresian *et al.*, 2003).

En dépit d'un important effort de recherche dans ce domaine, le traitement d'EPL dans les applications de TAL est encore un problème ouvert. Bien sûr, ceci n'est pas vraiment une surprise dans la mesure où les linguistes ont démontré la complexité de ce problème depuis des décennies (Sag *et al.*, 2002). Au début des années 2000, Schone et Jurafsky (2001) ont posé la question si l'identification automatique d'EPL était un problème résolu, et la réponse que cet article apporta à l'époque fut négative. De même, les préfaces du numéro spécial sur les EPL de la revue *Language Resources and Evaluation* (Rayson *et al.*, 2010) et de l'atelier MWE 2011 (Kordoni *et al.*, 2011) indiquent que, d'un point de vue pratique, plusieurs défis sont à relever dans le but d'obtenir des résultats moins artificiels pour les EPL dans les systèmes de TAL. Nous croyons que, à long terme, le présent travail de recherche contribuera à la conception d'applications de TAL qui intègrent pleinement le traitement des EPL comme une étape très importante dans la constitution du lexique et de la grammaire. Néanmoins, étant donné la complexité du problème, ce traitement doit être continuellement amélioré, car ils nous semble peu probable que, dans un avenir proche, on puisse proposer une solution définitive et unifiée pour le traitement des EPL dans les applications de TAL.

Remerciements

Je remercie mes tuteurs Christian Boitet et Aline Villavicencio, ainsi que les collègues qui ont contribué activement à ce travail : Evita Linardaki, Magali Sanchez Duran et Vitor De Araujo. Merci aux réviseurs pour leurs suggestions et à Antoine Gay pour la relecture. Ce travail est financé par une allocation du Ministère de l'Enseignement Supérieur et de la Recherche et par le projet CAMELEON (CAPES-COFECUB 707-11).

Références

Acosta, O., Villavicencio, A. et Moreira, V. (2011). Identification and treatment of multiword expressions applied to information retrieval. *In* (Kordoni *et al.*, 2011), pages 101–109.

Apresian, J., Boguslavsky, I., Iomdin, L. et Tsinman, L. (2003). Lexical functions as a tool of ETAP-3. *In Proc. of the First MTT Conference (MTT 2003)*.

Araujo, V. D., Ramisch, C. et Villavicencio, A. (2011). Fast and flexible MWE candidate generation with the mwetoolkit. *In* (Kordoni *et al.*, 2011), pages 134–136.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C. et Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *In Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain. ELRA.

Carpuat, M. et Diab, M. (2010). Task-based evaluation of multiword expressions : a pilot study in statistical machine translation. *In Proc. of HLT : The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California. ACL.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *In RIAO'88*, pages 609–624.

Church, K. (2011). How many multiword expressions do people know ? *In* (Kordoni *et al.*, 2011), pages 137–144.

Church, K. W. et Hanks, P. (1990). Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

Constant, M. et Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. *In* (Kordoni *et al.*, 2011), pages 49–56.

de Cruys, T. V. et na Villada Moirón, B. (2007). Semantics-based multiword expression extraction. *In* Gregoire, N., Evert, S. et Kim, S. N., éditeurs : *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 25–32, Prague, Czech Republic. ACL.

Dias, G. (2003). Multiword unit hybrid extraction. *In* Bond, F., Korhonen, A., McCarthy, D. et Villavicencio, A., éditeurs : *Proc. of the ACL Workshop on MWEs : Analysis, Acquisition and Treatment (MWE 2003)*, pages 41–48, Sapporo, Japan. ACL.

Duran, M. S., Ramisch, C., Aluísio, S. M. et Villavicencio, A. (2011). Identifying and analyzing brazilian portuguese complex predicates. *In* (Kordoni *et al.*, 2011), pages 74–82.

Evert, S. (2004). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.

EVERT, S. et KRENN, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

FINLAYSON, M. et KULKARNI, N. (2011). Detecting multi-word expressions improves word sense disambiguation. *In* (Kordoni *et al.*, 2011), pages 20–24.

FOTOPOULOU, A., GIANNOPOULOS, G., ZOURARI, M. et MINI, M. (2008). Automatic recognition and extraction of multiword nominal expressions from corpora (in greek). *In Proceedings of the 29th Annual Meeting, Department of Linguistics*, Aristotle University of Thessaloniki, Greece.

GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M. et MAKOWIECKI, F. (2010). Computational lexicography of multi-word units : How efficient can it be ? *In* (Laporte *et al.*, 2010), pages 1–9.

GROSS, M. (1986). Lexicon - grammar the representation of compound words. *In Proc. of the 11th COLING (COLING 1986)*.

JACKENDOFF, R. (1997). Twistin' the night away. *Language*, 73:534–559.

KELLER, F. et LAPATA, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Comp. Ling. Special Issue on the Web as Corpus*, 29(3):459–484.

KORDONI, V., RAMISCH, C. et VILLAVICENCIO, A., éditeurs (2011). *Proc.of the ACL Workshop on MWEs : from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA. ACL.

LAPORTE, E., NAKOV, P., RAMISCH, C. et VILLAVICENCIO, A., éditeurs (2010). *Proc.of the COLING Workshop on MWEs : from Theory to Applications (MWE 2010)*, Beijing, China. ACL.

LINARDAKI, E., RAMISCH, C., VILLAVICENCIO, A. et FOTOPOULOU, A. (2010). Towards the construction of language resources for greek multiword expressions : Extraction and evaluation. *In* PIPERIDIS, S., SLAVCHEVA, M. et VERTAN, C., éditeurs : *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

PEARCE, D. (2001). Synonymy in collocation extraction. *In WordNet and Other Lexical Resources : Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46.

PEARCE, D. (2002). A comparative evaluation of collocation extraction techniques. *In Proc. of the Third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain. ELRA.

PECINA, P. (2005). An extensive empirical study of collocation extraction methods. *In Proc. of the ACL 2005 SRW*, pages 13–18, Ann Arbor, MI, USA. ACL.

PECINA, P. (2010). Lexical association measures and collocation extraction. *Lang. Res. & Eval. Special Issue on Multiword expression : hard going or plain sailing*, 44(1-2):137–158.

RAMISCH, C., ARAUJO, V. D. et VILLAVICENCIO, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. *In Proc. of the ACL 2012 SRW*, Jeju, Republic of Korea. ACL.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010a). Multiword expressions in the wild ? the mwetoolkit comes in handy. *In* LIU, Y. et LIU, T., éditeurs : *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010b). mwetoolkit : a framework for multiword expression identification. *In Proc. of the Seventh LREC (LREC 2010)*, Malta. ELRA.

RAMISCH, C., VILLAVICENCIO, A., MOURA, L. et IDIART, M. (2008). Picking them up and figuring them out : Verb-particle constructions, noise and idiomaticity. *In* CLARK, A. et TOUTANOVA, K., éditeurs : *Proc. of the Twelfth CoNLL (CoNLL 2008)*, pages 49–56, Manchester, UK. The Coling 2008 Organizing Committee.

RAYSON, P., PIAO, S., SHAROFF, S., EVERT, S. et MOIRÓN, B. V. (2010). Multiword expressions : hard going or plain sailing ? *Lang. Res. & Eval. Special Issue on Multiword expression : hard going or plain sailing*, 44(1-2):1–5.

SAG, I., BALDWIN, T., BOND, F., COPESTAKE, A. et FLICKINGER, D. (2002). Multiword expressions : A pain in the neck for NLP. *In Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 de *LNCS*, pages 1–15, Mexico City, Mexico. Springer.

SCHONE, P. et JURAFSKY, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem ? *In* LEE, L. et HARMAN, D., éditeurs : *Proc. of the 2001 EMNLP (EMNLP 2001)*, pages 100–108, Pittsburgh, PA USA. ACL.

SERETAN, V. (2008). *Collocation extraction based on syntactic parsing*. Thèse de doctorat, University of Geneva, Geneva, Switzerland.

SMADJA, F. A. (1993). Retrieving collocations from text : Xtract. *Comp. Ling.*, 19(1):143–177.

SPINA, S. (2010). The dictionary of italian collocations : Design and integration in an online learning environment. *In Proc. of the Seventh LREC (LREC 2010)*, Malta. ELRA.

STYMNE, S. (2011). Pre- and postprocessing for statistical machine translation into germanic languages. *In Proc. of the ACL 2011 SRW*, pages 12–17, Portland, OR, USA. ACL.

VILLAVICENCIO, A., BOND, F., KORHONEN, A. et MCCARTHY, D. (2005). Introduction to the special issue on multiword expressions : Having a crack at a hard nut. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):365–377.

WEHRLI, E. et RAMLUCKUN, M. (1993). ITS-2 : an interactive personal translation system. *In Proc. of the 6th Conf. of the EACL (EACL 1993)*, page 476, Utrecht, The Netherlands. ACL.

Adaptation d'un système de reconnaissance d'entités nommées pour le français à l'anglais à moindre coût

Mohamed Hatmi
LINA, UMR 6241,Université de Nantes
mohamed.hatmi@univ-nantes.fr

RÉSUMÉ

La portabilité entre les langues des systèmes de reconnaissance d'entités nommées est coûteuse en termes de temps et de connaissances linguistiques requises. L'adaptation des systèmes symboliques souffrent du coût de développement de nouveaux lexiques et de la mise à jour des règles contextuelles. D'un autre côté, l'adaptation des systèmes statistiques se heurtent au problème du coût de préparation d'un nouveau corpus d'apprentissage. Cet article étudie l'intérêt et le coût associé pour porter un système existant de reconnaissance d'entités nommées pour du texte bien formé vers une autre langue. Nous présentons une méthode peu coûteuse pour porter un système symbolique dédié au français vers l'anglais. Pour ce faire, nous avons d'une part traduit automatiquement l'ensemble des lexiques de mots déclencheurs au moyen d'un dictionnaire bilingue. D'autre part, nous avons manuellement modifié quelques règles de manière à respecter la syntaxe de la langue anglaise. Les résultats expérimentaux sont comparés à ceux obtenus avec un système de référence développé pour l'anglais.

ABSTRACT

Adapting a French Named Entity Recognition System to English with Minimal Costs

Cross-language portability of Named Entity Recognition systems requires linguistic expertise and needs human effort. Adapting symbolic systems suffers from the cost of developing new lexicons and updating grammar rules. Porting statistical systems on the other hand faces the problem of the high cost of annotation of new training corpus. This paper examines the cost of adapting a rule-based Named Entity Recognition system designed for well-formed text to another language. We present a low-cost method to adapt a French rule-based Named Entity Recognition system to English. We first solve the problem of lexicon adaptation to English by simply translating the French lexical resources. We then get to the task of grammar adaptation by slightly modifying the grammar rules. Experimental results are compared to a state-of-the-art English system.

MOTS-CLÉS : Reconnaissance d'entités nommées, approche symbolique, portabilité entre les langues.

KEYWORDS: Named entity recognition, symbolic approache, cross-language portability.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 151–161,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La reconnaissance des entités nommées (REN) est une sous-tâche de l'extraction d'information consistant à délimiter et à catégoriser certaines expressions linguistiques autonomes et mono-référentielles (Ehrmann, 2008). Ces dernières correspondent traditionnellement à l'ensemble des noms propres (noms de personnes, de lieu et d'organisation) ainsi que certaines expressions numériques et temporelles (expressions de dates, de temps, de pourcentages, etc.). La délimitation et la catégorisation des entités nommées sont connues sous le nom d'annotation qui consiste généralement à encadrer une entité nommée par le biais d'une balise de début et de fin mentionnant sa typologie.

Certaines langues ont suscité beaucoup d'intérêt, notamment via les campagnes d'évaluations telles que MUC (Grishman et Sundheim, 1996) pour l'anglais et le japonais, CONLL (Tjong Kim Sang, 2002) pour l'espagnol et l'allemand et ESTER (Galliano *et al.*, 2009) pour le français. Plusieurs systèmes ont été développés pour ces différentes langues. La plupart d'entre eux utilisent soit des méthodes symboliques soit des méthodes statistiques. Les systèmes symboliques sont basés sur des lexiques (listes de prénom, de pays, etc.) et sur un ensemble de règles de réécriture (Maurel *et al.*, 2011; Brun et Ehrmann, 2010; Stern et Sagot, 2010). D'un autre côté, les systèmes statistiques sont basés sur un modèle appris à partir d'un corpus préalablement annoté (Bikel *et al.*, 1997; Raymond et Fayolle, 2010; Béchet et Charton, 2010).

La portabilité entre les langues des systèmes de REN est coûteuse en termes de temps et de connaissances linguistiques requises. Par exemple, les systèmes symboliques nécessitent un traitement lourd incluant, notamment, la modification des règles contextuelles et le développement de nouvelles listes de mots déclencheurs et de noms propres. La rareté et le coût de construction de ces ressources représentent un problème majeur pour certaines langues (Poibeau, 2003; Gamon *et al.*, 1997). D'un autre côté, les systèmes probabilistes se heurtent au problème de disponibilité des corpus d'apprentissages. Ces derniers ne sont pas faciles à constituer et ne sont pas disponibles pour toutes les langues. Plusieurs travaux visent à automatiser le processus d'annotation en exploitant l'encyclopédie multilingue Wikipédia (Nothman *et al.*, 2009) et les corpus multilingues comparables (Klementiev et Roth, 2006).

Dans cet article, nous examinons l'intérêt et le coût associé pour porter un système existant de REN pour du texte bien formé vers une autre langue. Nous nous sommes appuyés pour cela sur le système symbolique français Nemesis (Fourour, 2002). Nous présentons une méthode permettant de porter Nemesis vers l'anglais à moindre coût. Nous décrivons le système Nemesis dans la section 2 et les corpus d'évaluation dans la section 3. Nous présentons ensuite le détail du processus d'adaptation dans la section 4. La section 5 présente les résultats sur les corpus ayant servi aux évaluations et les compare aux résultats obtenus par Stanford Named Entity Recognizer [1] (Finkel *et al.*, 2005), un système de REN développé pour l'anglais. Pour terminer, dans la section 6, nous discutons les apports et les limites de cette approche.

1. Ce système est disponible gratuitement à : http ://nlp.stanford.edu/software/CRF-NER.shtml

2 Nemesis : un système symbolique de REN pour le français

Nemesis (Fourour, 2002) est un système qui permet la délimitation et la catégorisation des entités nommées développé pour le français et pour du texte bien formé (c'est-à-dire qui respecte les règles du français écrit). Il se base essentiellement sur les indices internes et externes définis par McDonald (1996). L'architecture de Nemesis se compose principalement de trois modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles.

Prétraitement lexical : segmentation du texte en occurrences de formes et de phrases, puis association des sigles à leur forme étendue.
Projection des lexiques : les lexiques ont été construits soit manuellement, soit automatiquement à partir du Web. Les éléments composant ces lexiques (79 476 éléments) sont répartis en 45 listes selon les catégories dans lesquelles ils sont utilisés : prénom connu, mot déclencheur d'un nom d'organisation (l'élément fait partie de l'entité nommée : « Fédération française de handball »), contexte d'un nom de personne (l'élément appartient au contexte gauche immédiat de l'entité nommée, mais ne fait pas partie de celle-ci : « philosophe Emmanuel Kant »), fin d'un nom d'organisation (l'élément est la dernière forme composant l'entité nommée : « Conseil régional », « Coupe du monde de football »), etc.
La projection des lexiques consiste à associer les étiquettes liées aux lexiques aux différentes formes du texte. Une forme peut avoir plusieurs étiquettes (Washington|prénom-connu|lieu-connu).
Application des règles : les règles de réécriture permettent l'annotation du texte par des balises identifiant les entités nommées (délimitation et catégorisation). Elles sont basées sur des étiquettes sémantiques référant à une forme capitalisée ou à une forme appartenant à un lexique. En tout, Nemesis utilise 93 règles qui s'exécutent dans un ordre prédéfini. Lorsque plusieurs règles s'appliquent, Nemesis opte pour la règle ayant la priorité la plus élevée. Voici un exemple de règles de réécriture :

```
$Clé-oronyme $Article-min [ $Forme-capitalisée+ ]→ ORONYME
```

et le résultat de son application :

"montagne du <ORONYME> Mont-Blanc </ORONYME>"

L'évaluation de Nemesis a été réalisée sur un corpus composé de textes issus du journal Le Monde et du Web (31 000 mots). Les performances sur l'ensemble des entités nommées montrent un rappel de 79 % et une précision de 91 % (Fourour, 2003).

3 Description des corpus et mesure des performances

Trois corpus dont deux en langue française et un autre en langue anglaise ont été utilisés dans nos expérimentations.

Le corpus de référence anglais, *BBN Pronoun Coreference and Entity Type Corpus* [2], est composé d'articles provenant de *Wall Street Journal* manuellement annoté en entités nommées. Le guide d'annotation comporte 12 catégories principales (*Person*, *Facility*, *Organization*, *GPE*, *Location*, *Nationality*, *Product*, *Event*, *Work of Art*, *Law*, *Language*, *Contact-Info*) et plusieurs sous-catégories. Seules les entités sont prises en compte dans l'étiquette. Les formes qui ne font pas partie de l'entité elle-même sont exclues de l'annotation (Mr. <PERSON> Spoon </PERSON>). Ce corpus a été divisé en deux parties, développement (3/4 du corpus) et test (1/4 du corpus).

Le corpus de référence français (Stern et Sagot, 2010) [3] est constitué de dépêches provenant de l'Agence France-Presse (AFP) et contient des annotations manuelles des entités de type Personne, Lieu et Organisation (comprenant les noms d'entreprises). Les formes non constitutives du nom de l'entité lui-même sont exclues de l'annotation (M. <PERSON> Spoon </PERSON>). La taille de ce corpus est bien plus faible que celle du corpus anglais. Pour des raisons de comparaison, nous avons également eu recours à un corpus non annoté constitué des articles du journal Le Monde (2007).

Dans ce travail, seules les entités communes entre les deux corpus annotés ont été retenues pour les expériences (Personne, Organisation, Lieu et Entreprise). La table 1 décrit les différents corpus utilisés dans ce travail. Les performances sont mesurées en termes de rappel et

Corpus	**Langue**	**Nb de mots**	**Nb d'entités**
Corpus BBN (développement)	anglais	938 330	46 478
Corpus BBN (test)	anglais	235 274	11 930
Corpus AFP	français	38 831	1 497
Corpus Le Monde	français	1 010 000	-

TABLE 1: Description des corpus

de précision. Le rappel est défini par le nombre d'entités correctement étiquetées au regard du nombre d'entités étiquetées dans la référence. La précision est le nombre d'entités correctement étiquetées au regard du nombre d'entités correctement et incorrectement étiquetées. La F-mesure combine ces deux mesures.

$$Rappel = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ dans\ la\ référence} \quad (1)$$

$$Précision = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ fournis} \quad (2)$$

$$F - mesure = \frac{2 * rappel * précision}{rappel + précision} \quad (3)$$

2. Catalogue LDC n° LDC2005T33
3. Ce corpus est disponible librement dans le cadre de la distribution de SXPipe

4 Adaptation de Nemesis à l'anglais

La mesure des performances des systèmes de REN dépend directement de la cohérence entre les annotations manuelles et les annotations automatiques. Nous avons donc commencé par ajuster les règles de délimitation et de catégorisation de Nemesis aux normes d'annotation des corpus d'évaluation.

Notre objectif est de porter Nemesis vers l'anglais d'une façon simple et peu coûteuse. La méthode proposée consiste à adapter séparément et séquentiellement les deux principaux éléments constitutifs de Nemesis : les lexiques et les règles de réécriture.

4.1 Adaptation des lexiques

Nous avons construit l'ensemble des lexiques pour l'anglais en traduisant tout simplement les lexiques existants pour le français. La traduction est faite automatiquement en utilisant un dictionnaire bilingue [4] sans aucune information contextuelle. Cela concerne principalement l'ensemble des lexiques de mots déclencheurs. Les lexiques des noms de personne et d'entreprise sont conservés. Cette tâche n'est pas coûteuse en termes de temps et ne demande pas une expertise linguistique (des outils en ligne comme *Google Translate* peuvent être utilisés pour les langues pour lesquelles les dictionnaires électroniques ne sont pas disponibles). Lorsque le dictionnaire comporte plusieurs traductions pour un mot, l'ensemble des traductions sont conservées (nous ne traitons pas à ce niveau les problèmes de polysémie). Une fois la phase de traduction terminée, nous avons utilisé une liste des mots outils en anglais pour écarter certaines entrées pouvant produire de bruit, par exemple le mot *even* qui se présente comme un nom de personne dans le lexique français. En définitive, l'ensemble des lexiques pour l'anglais compte 83 305 entrées.

4.2 Adaptation des règles

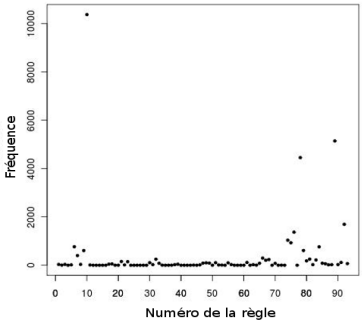
4.2.1 Classification des règles

Avant d'adapter les règles, nous avons commencé par mesurer la fréquence et la précision relatives de chacune des règles utilisées par Nemesis (93 règles). La fréquence représente le nombre de fois où chaque règle est déclenchée pour reconnaître une entité nommée. La figure 1 présente les résultats obtenus sur le corpus Le Monde pour le français (1a) et sur le corpus de développement BBN pour l'anglais (après adaptation des lexiques) (1b). Les observations montrent que la loi de Zipf est respectée pour les deux langues : un nombre limité de règles est à l'origine de la plupart des entités extraites, les autres règles sont rarement déclenchées. Par exemple pour l'anglais, 11 règles couvrent 86% des entités extraites. On remarque aussi que de nombreuses règles ne sont pas déclenchées (38 pour l'anglais contre 17 pour le français).

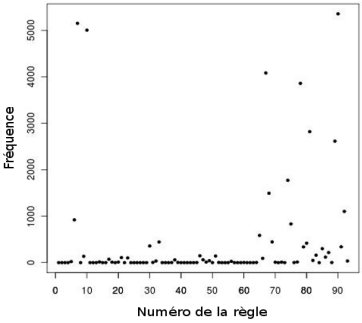
La précision mesure la quantité d'entités correctement reconnues parmi les réponses retournées. En effet, ce n'est pas parce qu'une règle est fréquente qu'elle est pour autant précise.

4. Catalogue ELRA-M0033 (http ://catalog.elra.info/product_info.phpproducts_id=666&language=fr)

Nous avons calculé la précision de chacune des règles déclenchées pour le français sur le corpus AFP et pour l'anglais sur le corpus de développement BBN. La figure 2 présente les résultats obtenus pour l'anglais. Plusieurs règles déclenchées obtiennent une précision relativement faible (32 règles ont une précision inférieure à 50%, ce qui représente environ 60% des règles déclenchées). En se basant sur les critères de fréquence et de précision, nous avons ensuite classé



(a) Français (corpus Le Monde)



(b) Anglais (corpus de développement BBN)

FIGURE 1: Nombre de déclenchements des règles pour le français et pour l'anglais

les règles en différentes catégories, par exemple : règles fréquentes [5] ayant une bonne précision [6] dans les deux langues (7 règles), règles fréquentes pour l'anglais avec une faible précision [7] (3 règles), règles déclenchées seulement pour le français (19 règles), règles non déclenchées dans les deux langues (15 règles), etc. Ces différentes catégories vont nous permettre de déterminer quelles sont les règles à modifier pour la langue anglaise.

5. Fréquence > 1 000
6. Précision > 70 %
7. Précision < 50 %

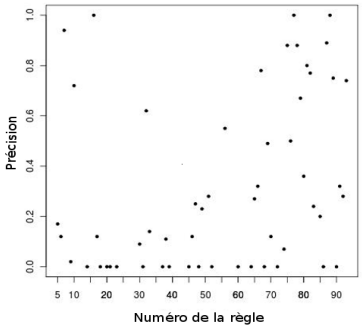


FIGURE 2: Précision des règles déclenchées pour l'anglais (corpus BBN)

4.2.2 Adaptation des règles

Pour adapter les règles, nous avons sélectionné celles n'appartenant pas aux catégories de bonne précision pour l'anglais. Nous avons également éliminé certaines règles déclenchées seulement pour le français car elles sont très spécifiques à cette langue, par exemple :

`[ $Clé-organisation $Article-min $Item $Article-min $Forme-capitalisée+ ] → ORGANISATION`

et un exemple d'application :

"<ORGANISATION> Centre de recherche de Solaize </ORGANISATION>"

Nous avons ensuite modifié les règles sélectionnées de manière à respecter la syntaxe de la langue anglaise (29 règles), comme la règle suivante pour le français :

`$Fonction $Adjectif-de-nationalité [ $Forme-capitalisée+ ] → PERSON`

et le résultat de son application :

"Le président français <PERSON> Jacques Chirac </PERSON>"

et la règle pour l'anglais après adaptation :

`$Adjectif-de-nationalité $Fonction [ $Forme-capitalisée+ ] → PERSON`

et le résultat de son application :

"The French President <PERSON> Jacques Chirac </PERSON>"

5 Résultats

Pour des raisons de comparaison, nous avons dans un premier temps appliqué Nemesis au corpus de test BBN et au corpus AFP. Aucune adaptation n'a été réalisée. La table 2 présente les performances réalisées par catégorie. Nous pouvons remarquer que la reconnaissance des noms de personne demeure satisfaisante (perte d'environ 5 points de F-mesure). Cela s'explique par le fait que les lexiques de Nemesis contiennent des prénoms anglais et qu'il y a des règles de réécriture communes. La reconnaissance se voit fortement dégradée pour les autres catégories. Nous avons ensuite mesuré l'apport de l'adaptation des lexiques et des règles. La table 3 affiche

	Nemesis Français	**Nemesis Anglais (sans aucune adaptation)**
	Corpus AFP (français)	**Corpus de test BBN (anglais)**
	F1 (P/R)	**F1 (P/R)**
Personne	77,33 (85,83/71,03)	71,82 (66,65/77,85)
Lieu	88,82 (90,24/87,44)	37,46 (45,61/31,79)
Organisation	54,24 (65,04/46,51)	1,5 (1,1/2)
Entreprise	37,73 (54,05/30)	10 (36,83/5,84)

TABLE 2: Performances de Nemesis pour le français et l'anglais sans adaptation

les gains obtenus pour chaque adaptation. Ces résultats sont comparés à ceux obtenus avec un système natif (Stanford NER).

	Nemesis Anglais (adaptation lexiques)	**Nemesis Anglais (adaptation lexiques et règles)**	**Stanford NER**
	Corpus BBN (anglais)	**Corpus BBN (anglais)**	**Corpus BBN (anglais)**
	F1 (P/R)	**F1 (P/R)**	**F1 (P/R)**
Personne	77,3 (74,48/80,34)	79,51 (81,08/78,01)	90,9 (88,07/93,9)
Lieu	74,38 (72,5/76,35)	79,17 (78,59/79,76)	90,8 (87,6/94,2)
Organisation	21,02 (24,2/18,58)	38,15 (41,9/35,02)	84,6 (89,2/80,04)
Entreprise	27,52 (50,72/18,88)	30,15 (68,34/19,34)	

TABLE 3: Performances de Nemesis pour le français et l'anglais

L'adaptation des lexiques a permis un apport significatif concernant la reconnaissance des noms de lieu (environ 37 points de F-mesure). La reconnaissance des noms d'organisation et d'entreprise se voit améliorée mais elle reste toutefois faible. En effet, ce problème semble lié à une couverture insuffisante des lexiques de Nemesis et à une ambiguité liée à la catégorisation des organisations et des entreprises (entreprise catégorisée en tant qu'organisation et vice-versa).

L'adaptation des règles montre un gain relativement bon pour la catégorie organisation

(environ 17 points de F-mesure) et une légère amélioration concernant les autres catégories. Les résultats globaux sont bien en dessous de ceux obtenus avec Stanford NER, surtout pour la catégorie organisation. Stanford NER est un système à base d'apprentissage développé pour l'anglais. Pour ce dernier, les noms d'entreprises sont catégorisés comme étant organisation.

6 Discussion

Cette approche peu coûteuse pour porter Nemesis vers l'anglais montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. Elle montre ses limites pour la reconnaissance des noms d'organisation et d'entreprise. L'analyse des résultats obtenus nous a permis de souligner deux problèmes récurrents de la reconnaissance des entités nommées : la délimitation des entités nommées et la délimitation des catégories.

Les règles d'annotation sont loin de faire consensus. Elles suscitent toujours des vives discussions et plusieurs remises en cause du guide d'annotation, par exemple dans le cadre de la campagne ETAPE (Évaluations en Traitement Automatique de la Parole). Le premier problème rencontré concerne la délimitation des entités nommées. En effet, les règles de délimitation de Nemesis ne sont pas les mêmes que celles utilisées pour annoter le corpus AFP et le corpus BBN. Par exemple, Nemesis inclut les titres dans les noms de personne (<PERSON> M. Dorgan </PERSON>) alors que ces derniers ne sont pas inclus dans le corpus AFP (M. <PERSON> Dorgan </PERSON>). Nous avons dû adapter les règles de délimitation de Nemesis aux spécificités du corpus traité. Cependant, plusieurs erreurs de délimitation ont été relevées. Le deuxième problème concerne la délimitation des catégories. Nemesis adopte une catégorisation fine (5 catégories et 30 sous-catégories). Cette typologie a l'avantage de pouvoir s'adapter aux typologies moins fines en regroupant des sous-catégories. Nous avons donc adapté la typologie de Nemesis en fonction des catégories du corpus AFP et du corpus BBN. Toutefois, beaucoup d'erreurs sont dues à une incohérence de catégorisation. Par exemple, la notion d'organisation et d'entreprise n'est pas identique entre Nemesis et le corpus BBN (l'entité « Federal Reserve Board » est annotée comme étant une entreprise par le système Nemesis alors qu'elle est considérée comme étant une organisation dans le corpus BBN).

7 Conclusion et perspectives

Cet article présente une méthode peu coûteuse pour porter un système symbolique de REN dédié pour le français vers l'anglais. L'adaptation est basée principalement sur les ressources développées pour le français. Elle consiste à traduire les lexiques français et à adapter légèrement quelques règles de la grammaire. L'évaluation du système adapté montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. En revanche, les résultats restent insuffisants pour les noms d'organisation et d'entreprise. Un traitement plus approfondi est nécessaire pour ces deux catégories. Pour cela, nous comptons mesurer l'impact d'un enrichissement des lexiques de Nemesis (notamment les listes d'organisation et d'entreprise). D'un autre côté, nous envisageons d'adapter les règles de réécriture automatiquement et de tester cette méthode sur d'autres langues qui sont moins proches du français que l'anglais.

Références

Béchet, F. et Charton, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, pages 5338–5341, Dallas, Texas, USA.

Bikel, D. M., Miller, S., Schwartz, R. et Weischedel, R. (1997). Nymble : a high-performance learning name-finder. *In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington, DC, USA.

Brun, C. et Ehrmann, M. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ester 2. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada.

Ehrmann, M. (2008). *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7, France.

Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, USA.

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *In Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 265–274, Nancy, France.

Fourour, N. (2003). Apport du web dans la reconnaissance des entités nommées. *In Revue Québécoise de Linguistique (RQL)*, pages 41–60.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. *In Proceedings of the 10th conference Interspeech*, pages 2583–2586, Brighton, UK.

Gamon, M., Lozano, C., Pinkham, J. et Reutter, T. (1997). Practical experience with grammar sharing in multilingual nlp. *In Workshop From research to commercial applications : making NLP work in practice*, pages 49–56, Madrid, Spain.

Grishman, R. et Sundheim, B. (1996). Message Understanding Conference-6 : a brief history. *In Proceedings of the 16th conference on Computational linguistics (COLING'06)*, pages 466–471, Copenhagen, Denmark.

Klementiev, A. et Roth, D. (2006). Named entity transliteration and discovery from multilingual comparable corpora. *In Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06)*, pages 82–88, New York, USA.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I. et Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *In Traitement Automatique des Langues (TAL)*, pages 69–96.

McDonald, D. D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In Corpus processing for lexical acquisition*, pages 21–39. MIT Press, Cambridge, MA, USA.

Nothman, J., Murphy, T. et Curran, J. R. (2009). Analysing wikipedia and gold-standard corpora for ner training. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 612–620, Athens, Greece.

POIBEAU, T. (2003). The multilingual named entity recognition framework. *In Proceedings of the 10th Conference on European chapter of the Association for Computational Linguistics (EACL'03)*, pages 155–158, Budapest, Hungary.

RAYMOND, C. et FAYOLLE, J. (2010). Reconnaissance robuste d'entités nommés sur de la parole transcrite automatiquement. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, pages 19–23, Montréal, Canada.

STERN, R. et SAGOT, B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada.

TJONG KIM SANG, E. F. (2002). Introduction to the conll-2002 shared task : language-independent named entity recognition. *In Proceedings of the 6th Workshop on Computational Language Learning (CoNLL'02)*, pages 155–158, Taipei, Taiwan.

État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie

Mounira Manser
LIM&BIO (EA3969)
Université Paris 13
93017 Bobigny Cedex
France
manser.mounira@gmail.com

RÉSUMÉ

L'accès au contenu des textes de spécialité est une tâche difficile à réaliser. Cela nécessite la définition de méthodes automatiques ou semi-automatiques pour identifier des relations sémantiques entre les termes que contiennent ces textes. Nous distinguons les approches de TAL permettant d'acquérir ces relations suivant deux types d'information : la structure interne des termes ou le contexte de ces termes en corpus. Afin d'améliorer la qualité des relations acquises et faciliter leur réutilisation en corpus, nous nous intéressons à la prise en compte du contexte dans une méthode d'acquisition de relations de synonymie basée sur l'utilisation de la structure interne des termes. Nous présentons les résultats d'une expérience préliminaire tenant compte de l'usage des termes dans un corpus biomédical en anglais. Nous donnons quelques pistes de travail pour définir des contraintes sémantiques sur les relations de synonymie acquises.

ABSTRACT

State of the Art on the Acquisition of Semantic Relations between Terms : Contextualisation of the Synonymy Relations

Accessing to the context of specialised texts is a crucial but difficult task. It requires automatic or semi-automatic methods dedicated to the identification of semantic relations between terms appearing in the texts. NLP approaches for acquiring semantic relations between terms can be distinguished according to the type of information : the internal structure of the terms and the term context. In order to improve the quality of the acquired synonymy relations and their reusability in other corpora, we aim at taking into account the context into an approach based on the internal structure of the terms. We present the results of a preliminary experiment taking into account the use of the terms in a English biomedical corpora. This experiment will be helpful to add semantic constraints to the already acquired synonymy relations.

MOTS-CLÉS : Acquisition de relations, Synonymie, Relations sémantiques, Terminologie, Domaine Biomédical, Corpus de spécialité.

KEYWORDS: Relation Acquisition, Synonymy, Semantic Relations, Terminology, Biomedical Domain, Specialised corpora.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 163–175,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Devant la masse considérable de donnés disponibles, la difficulté d'extraire et de rechercher l'information à partir des textes devient de plus en plus importante et demande un effort de structuration et d'ordonnancement des données. Les ressources terminologiques répondent partiellement à ce besoin en proposant les termes du domaine et différents types de relations (synonymie, hypéronymie ou des relations plus spécifiques au domaine comme *a-pour-symptôme*). La projection des termes issus d'une terminologie ne suffit pas (Cabré, 1999; Hamon, 2005; Spasic *et al.*, 2005; McIntosh et Curran, 2009). Il est nécessaire de disposer de terminologies structurées afin d'être capable de s'adapter aux usages dans les documents manipulés. Pour cela, des approches automatiques ou semi-automatiques doivent être mises en œuvre. Des approches de TAL ont ainsi été proposées pour aider à l'identification des termes et des relations permettant de structurer ces listes de termes. Cependant, la qualité des résultats peut varier en termes de rappel (les méthodes sont trop restrictives) ou de précision (les ambiguïtés ou la polysémie sont peu ou mal prises en compte).

Nous nous intéressons ici à l'acquisition de relations de synonymie entre termes. Notre travail consiste à proposer et à tester des méthodes dans le but d'améliorer l'acquisition de relations de synonymie produites par SynoTerm (Hamon et Nazarenko, 2001). Il s'agit de filtrer des relations de synonymie en exploitant les informations contextuelles et sémantiques liées aux termes ou à leur composants. Le travail est réalisé sur des données issues du domaine biomédical en langue anglaise. Nous travaillons à la fois avec les ressources terminologiques Gene Ontology et UMLS, ainsi qu'avec un ensemble des résumés Medline (voir section 3).

Dans cet article nous présentons d'abord les relations sémantiques pouvant être fournies par une ressource terminologique (section 2.1). Puis nous présentons un état de l'art des approches d'acquisition de ces relations (section 2). La section 3 est consacrée au matériel que nous avons utilisé pour nos expériences. Nous décrivons à la section 4 les pistes de travail pour l'amélioration de l'acquisition de relations de synonymie et les résultats d'une expérience préliminaire (section 5).

2 Approches pour l'identification des relations sémantiques entre les termes

Différents types d'approches de TAL permettent d'extraire des relations sémantiques entre les termes. L'acquisition de telles relations est réalisée soit en exploitant la structure interne des termes issus de corpus ou de terminologie (section 2.2), soit en s'appuyant sur le contexte de ces termes en corpus (section 2.3). Dans cette section, nous présentons tout d'abord les types de relations pouvant apparaître dans une terminologie puis un panorama des approches et des types de relations qu'elles peuvent permettre d'acquérir.

2.1 Relations sémantiques dans une terminologie

Les terminologies visent à recenser les termes d'un domaine de spécialité, c'est-à-dire les unités linguistiques désignant un concept, un objet ou un processus (Bourigault et Jacquemin, 2000),

mais aussi les relations sémantiques qu'entretiennent ces termes entre eux. Plusieurs types de relations sémantiques sont proposées par les ressources terminologiques (Sager, 1990), le choix d'inclure un type de relation étant surtout dépendant de l'usage de la ressource :

– les **relations taxinomiques** : Ce type de relation structure des termes dans une arborescence.

  Les relations d'**hypéronymie** (`is-a/est-un`) relient un terme général à un terme spécifique. Par exemple, nous avons les relations *oxidase* `is-a` *enzyme* ou *contractil fiber* `is-a` *fiber*.

  Les **relations partitives** (méronymie ou partie-tout) sont utilisées pour définir une relation entre deux termes ou l'un est une partie de l'autre. On a par exemple la relation de méronymie *nucleus* `partie-de` *cell*.

– Les **relations sémantiques lexicales** regroupent deux types de relations :

  les relations de `synonymie` ou d'équivalence qui relient les termes possédant le même sens, par exemple *red blood* et *erythrocyte* ;

  les relations d'`antonymie` ou d'opposition qui relient les termes ayant des sens contraires, par exemple *anabolosm* et *catabolism*

– Les **relations inter-hiérarchiques** (transversales) relient les termes appartenants à des branches distinctes d'une ou plusieurs hiérarchies. Ces relations sont très variables suivant les domaines. Par exemple, la relation `localisée-dans` permet de lier les termes *division cellulaire* et *cellule*.

2.2 Exploitation de la structure interne des termes

L'acquisition des relations sémantiques basée sur la structure interne des termes a donné lieu à de nombreux travaux. Ceux-ci utilisent différents types d'informations issus de l'analyse linguistique des termes (informations morphologiques, syntaxiques et plus rarement sémantiques) et/ou des indices statistiques comme la fréquence ou la productivité. La combinaison de ces travaux s'avèrent très utiles pour la structuration d'une liste de termes (Daille, 2003).

Le partage de bases morphologiques communes peut aider à l'acquisition de relations sémantiques. Ainsi, dans le domaine biomédical, (Zweigenbaum et Grabar, 2000) ont exploité ces types d'informations morphologiques entre des termes et leur productivité pour identifier les relations de synonymie, d'hypéronymie ou inter-hiérarchiques (*acide, acido, acidité, acidurie, acidocitose, acidophile*). La dérivation est également très utile pour identifier des variantes terminologiques en corpus (Jacquemin, 1997; Grabar et Hamon, 2006) ou structurer sémantiquement un lexique dans un contexte multilingue (Namer et Baud, 2007). Ainsi, le premier travail permet d'identifier une relation entre *stenosis of the aorta* et *aortic stenosis*. Tandis que dans le second, il est possible de lier les termes *proctorrhagia* et *colorrhagia*, ou les adjectifs *bactérioïde* et *bactériforme*.

D'autres travaux utilisent des techniques d'apprentissage par analogie sur l'emploi des fonctions lexicales pour identifier des relations inter-hiérarchiques entre les termes (Claveau et L'Homme, 2005). Par exemple, les auteurs exploite la fonction lexicale connue entre *connecteur* et *connecter* (caractérisant la relation inter-hiérarchique `instrument_pour`), pour identifier la même relation ou la même fonction lexicale entre *éditeur* et *éditer*.

L'analyse syntaxique des termes permet généralement d'identifier des relations d'hypéronymie. Ainsi, de nombreux travaux exploitent l'inclusion lexicale, c'est-à-dire l'hypothèse que lorsqu'un

terme est lexicalement inclus dans un autre, une relation d'hypéronymie peut être établie (par exemple, *fatty acids **essential** / fatty acids*). Pour cela, ils s'appuient sur une décomposition en tête/expansion des termes (Bourigault, 1994; Bodenreider *et al.*, 2001; Grabar et Zweigenbaum, 2002). Une analyse fine des relations induites par inclusion lexicale montre cependant que celle-ci permet également de produire des relations inter-hiérarchiques (Ibekwe-SanJuan, 2005). Dans Faster (Jacquemin, 1997), des variantes terminologiques sont identifiées à l'aide des mécanismes syntaxiques d'insertion, de juxtaposition et de coordination. Une approche similaire mais basée uniquement sur le remplacement de chaînes de caractères est utilisée par (Verspoor *et al.*, 2003) pour acquérir des relations d'hypéronymie dans les termes de Gene Ontology. La variation verbo-nominale combinée à un apprentissage inductif peut également être pris en considération pour identifier des relations sémantiques (Bouillon *et al.*, 2000).

Bien qu'en général les informations sémantiques sont plutôt utilisées dans des méthodes basées sur le contexte des termes, quelques travaux prenant en compte la structure interne des termes exploitent ces types d'informations. Il est alors nécessaire de faire l'hypothèse de la compositionnalité des termes et de s'appuyer sur la présence d'un invariant syntaxique. Les indices sémantiques sont alors des relations entre les composants des termes. Ainsi, le processus d'acquisition de variation morpho-syntaxique proposé par (Jacquemin, 1997) peut être étendu en exploitant des relations de synonymie (Jacquemin, 1999). D'autres travaux visent à propager les relations sémantiques sur les termes complexes (Hamon et Nazarenko, 2001) en combinant différentes ressources lexicales. La qualité des relations inférées dépend de la spécialisation des relations sémantiques initiales par rapport au domaine. Un moyen d'obtenir ces relations initiales spécifiques au domaine consiste à utiliser les relations sémantiques issues d'une terminologie (Verspoor *et al.*, 2003; Hamon et Grabar, 2008). Les relations initiales induites doivent cependant être filtrées ou contextualisées, comme nous le présentons à la section 4.

2.3 Prise en compte du contexte des termes

Les approches exploitant le contexte des termes s'appuient sur l'hypothèse que la sémantique des termes peut être identifiée avec les contextes dans lesquels ils apparaissent. Ainsi, outre la désambiguïsation sémantique des termes, l'étude du contexte des termes fournit des indices importants et parfois indispensables pour acquérir et caractériser les relations sémantiques qu'ils entretiennent entre eux.

Nous nous intéressons ici principalement aux travaux réalisés sur les langues de spécialité et qui s'intéressent essentiellement à identifier les relations hiérarchiques et les relations inter-hiérarchiques. Les relations de synonymie sont plus rarement identifiées de cette manière.

2.3.1 Définition de patrons lexico-syntaxiques

Une des principales stratégies pour acquérir des relations sémantiques à l'aide du contexte des termes consiste à définir des patrons lexico-syntaxiques caractéristiques de la relation visée (Hearst, 1992; Auger et Barrière, 2008). Les patrons sont définis à partir d'observations en corpus et permettent d'extraire des relations d'hypéronymie. Par exemple, le patron *NP , NP * , or other NP* appliqué à la phrase "*Bruises, wounds, broken bones or other injuries...*" identifie des relations d'hypéronymie *bruise* `is_a` *Injury*, *wound* `is_a` *injury* et *broken bone* `is_a` *injury*. L'utilisation

d'une méthode d'identification automatique des patrons lexico-syntaxiques en corpus permet d'affiner les observations réalisées et d'obtenir de meilleurs résultats (Morin, 1999). Des relations transversales peuvent également être acquises grâce à cette stratégie (Hamon *et al.*, 2010; Røst *et al.*, 2010).

Si les approches basées sur les patrons exploitent en général des informations lexicales et morpho-syntaxiques dans un contexte relativement restreint, certains travaux cherchent à utiliser des relations de dépendance syntaxique (Snow *et al.*, 2005). Les patrons syntaxiques sont alors construits par apprentissage supervisé sur des chemins de dépendance syntaxique entre les termes. Un filtrage sur la longueur de ces chemins est ensuite appliqué. C'est également le cas lorsqu'il s'agit d'identifier des relations inter-hiérarchiques. Les patrons lexico-syntaxiques servent de base à la définition de modèle d'apprentissage exploitant les CRF (Yang et de Roeck, 2010) ou les SVM (Grouin *et al.*, 2010). Dans ces deux travaux, les entités mises en relations ainsi que leur types sémantiques sont connues, ce qui est généralement pas le cas dans les approches classiques visant à acquérir des relations sémantiques entre termes. Les indices trouvés en corpus dans le contexte peuvent être exploités pour inférer des patrons par programmation logique inductive (Martienne et Morin, 1999; Claveau et L'Homme, 2004). Des relations d'hypéronymie, dans le premier travail, ou des relations entre des verbes et des noms, dans le second, peuvent ainsi être identifiées.

On peut aussi remarquer que les patrons lexico-syntaxiques sont plus rarement utilisés pour l'acquisition de relations de synonymie. Cependant, il semble que certains domaines de spécialité, comme la biologie, s'y prêtent mieux (Weissenbacher, 2004; Mccrae et Collier, 2008). Ce phénomène particulier est probablement dû aux pratiques de renommage des noms de gènes lors de la découverte de leur fonction, l'explicitation de termes, mais aussi aux efforts dans ce domaine pour améliorer la recherche d'information et l'interopérabilité sémantique entre les terminologies existantes et les textes. Ainsi, par exemple le patron *also known* permet d'extraire la relation de synonymie dans l'extrait suivant : *reculatory factor (IRF) also know IRF-8*.

Des contextes plus larges peuvent être exploités pour acquérir des relations entre termes. Ainsi, des contextes riches en connaissance (*Knowledge-Rich Context*) sont définis selon (Meyer, 2001) comme étant un contexte qui contient des termes d'un domaine spécialisé et des models (patterns) de connaissaces. Des relations hiérarchiques (hypéronymie et méronymie) et des relations transversales (Schumann, 2011) entre les termes et les models sont alors obtenues. Par exemple, la phrase *tNF kappa B is a potent mediator of specific gene expression in human monocytes and has been shown to play a role in transcription of the HIV-1 genome in promonocytic leukemias* est défini comme un contexte riche en connaissance pour la relation d'hyperonymie *tNF kappa B* / *potent mediator*.

2.3.2 Exploitation de la distribution contextuelle des termes

Une autre utilisation du contexte des termes consiste à réaliser une analyse distributionnelle (Harris, 1990) pour regrouper des termes partageant des contextes (à l'origine syntaxique). Par exemple, les termes *insuffisance rénale* et *détresse respiratoire* sont sémantiquement proches car ils partagent les mêmes contextes *prise en charge d'une insuffisance rénale* et *prise en charge d'une détresse respiratoire*, *apparition d'une insuffisance rénale* et *apparition d'une détresse respiratoire*. Il est ainsi possible d'identifier une relation de proximité sémantique entre des termes (Bourigault *et al.*, 2004), voire des relations de synonymie (Ferret, 2011). Par exemple, dans (Grefenstette,

1994), l'analyse distributionnelle appliquée à un corpus médical permet de repérer plusieurs types de relations : synonymie (*large* / *important* / *great*), méronymie (*patient* / *group*) ainsi que des relations d'hypéronymie *patient* / *woman*.

Cette méthode a été également utilisée par (Resnik, 1993) afin de mettre en évidence les relations sémantiques associées aux termes. Il s'agit alors de remplacer les termes présents dans les contextes par leurs classes sémantiques, issues de WordNet. Par exemple les termes *infirmier* et *docteur* sont remplacés par la classe *profession de santé* de WordNet. Il est également possible d'adopter une approche mixte qui combine l'analyse distributionnelle et les patrons lexico-syntaxiques (Caraballo, 1999). Les termes sont d'abord regroupés en fonction des contextes partagés, et des patrons sont alors appliqués pour identifier des relations d'hypéronymie.

Enfin, nous pouvons également mentionner l'approche proposée par (Cimiano *et al.*, 2000) pour acquérir des relations taxonomiques de manière semi-automatique. L'analyse syntaxique des phrases est utilisée pour construire des contextes formels (chaque nom est caractérisé par un ensemble d'attributs composé par des verbes et pour lesquels le nom apparaît comme un argument). Les noms partageant les mêmes contextes seront utilisés pour former un treillis. L'Analyse Formelle de Concepts (FCA) est alors appliquée pour repérer les appariements entre les concepts (Ganter et Wille, 1999). Cette approche vise à structurer les connaissances sous forme d'une hiérarchie à partir d'un ensemble d'entité. Les entités sont représentées sous forme d'un treillis de Galois. Le treillis est composé d'un ensemble d'individus (les objets formels, par exemple des termes), d'un ensemble de caractéristique (les attributs formels, par exemple des contextes) et d'une relation binaire entre les objets et les attributs. Il est également possible d'extraire des relations autre que des relations d'hypéronymie, notamment des relations inter-hiérarchique à l'aide de l'analyse relationnelle de concepts (Bendaoud *et al.*, 2010).

2.3.3 Discussion et analyse

Les travaux présentés ci-dessus reflètent l'importance de ce champ scientifique et montrent une certaine diversité dans les approches permettant d'acquérir des relations en termes issus de domaine de spécialité. Dans cet article, nous proposons de combiner des approches basées sur des informations externes et internes aux termes. Notre objectif se rapproche des travaux de (Resnik, 1993). Mais ici, nous nous situons dans un domaine de spécialité et nous visons à contextualiser sémantiquement les relations de synonymie acquises par une approche exploitant la structure interne des termes.

3 Matériel

Dans cette section, nous décrivons le matériel que nous avons à notre disposition pour nos expériences. Nous avons sélectionné des ressources terminologiques (UMLS et Gene Ontology) et des corpus issus du domaine biomédical (corpus Genia et BioNLP). Nous envisageons également d'utiliser des ressources générales (notamment WordNet) pour étiqueter sémantiquement nos corpus de travail.

3.1 Ressources terminologiques et lexicales

– **Gene ontology** : Gene Ontology (GO) (The Gene Ontology Consortium, 2000) est une ressource terminologique dont l'objectif est de décrire le rôle des gènes dans les organismes (prokaryotes et eukaryotes) ainsi que leurs produits géniques. Elle propose 54 453 concepts et 94 161 termes. Les termes de GO sont structurés en trois arbres hiérarchiques : processus biologiques, fonctions moléculaires et composants cellulaires. Le vocabulaire de GO est structuré à l'aide de trois types de relations : l'hypéronymie (119 430 relations), méronymie (29 573 relations) et la synonymie (101 254 relations). Actuellement, nous n'utilisons que les relations de synonymie.
– **UMLS** (Lindberg *et al.*, 1993) Unified Medical Language System (UMLS) est une ressource terminologique biomédicale. Développé par la National Library of Medicine (NLM), elle regroupe plus d'une centaine de thésaurus de différentes langues dans un méta-thesaurus. Celui-ci organise 700 000 concepts au sein d'un réseau sémantique composé de 134 types sémantiques et structuré par 54 relations sémantiques hiérarchisées par le lien is-a. Les types sémantiques associés aux termes de l'UMLS seront utilisés pour définir les contraintes sémantiques sur les mots ou les termes pour lesquels nous avons acquis des relations de synonymie.
– **WordNet** : WordNet (Fellbaum, 1998) est une base de données lexicale, développée à l'Université de Princeton en 1985. Son objectif est de structurer le contenu sémantique et lexicale de la langue anglaise. WordNet est organisé en ensembles de synonymes appelés synsets. À chaque synset, correspond un concept. Les relations lexicales présentes dans WordNet ne lient que les mots de la même catégorie morpho-syntaxique. Il existe donc quatre hiérarchies (pour les noms, les verbes, les adjectifs et les adverbes). Dans la version 3.0, WordNet contient 155 287 mots avec 117 798 noms, 11 529 verbes, 21 479 adjectifs et 4 481 adverbes organisés en 117 659 synset. Nous envisageons d'exploiter, dans un deuxième temps, les relations issues de WordNet pour étendre le processus d'acquisition de relations de synonymie, et d'exploiter les synsets dans la définition des contraintes sémantiques.

3.2 Corpus de travail

Notre approche vise à contextualiser les relations de synonymie acquises automatiquement. Nous exploiterons deux corpus de spécialité, constitués d'un ensemble d'articles issus de la base de données Medline [1].

Le corpus GENIA [2] contient, dans sa version 3.0, 2 000 résumés Medline au format XML, recueillis à l'aide des termes MESH : "Human", "Blood Cells", et "Transcription Factors" (Kim *et al.*, 2003). Il est composé de 400 000 mots. Il est annoté avec différents niveaux d'informations linguistiques et sémantiques.

Dans un premier temps, nous utilisons le corpus d'entraînement de la campagne BioNLP2011 [3] (Pyysalo *et al.*, 2011). Celui-ci est un ensemble de 800 articles de Medline issus du corpus GENIA. Il est composé de 176 146 mots. Il est également annoté avec des informations linguistiques et sémantiques et notamment des relations inter-hiérarchiques *Protein/component* et *subunit/complex*.

1. http ://www.ncbi.nlm.nih.gov/Entrez/
2. http ://www.nactem.ac.uk/genia/genia-corpus
3. http ://2011.bionlp-st.org/

4 Prise en compte du contexte des termes dans l'acquisition de relations de synonymie

Notre travail s'appuie sur la méthode implémentée dans SynoTerm pour acquérir des relations de synonymie entre termes complexes (Hamon et Nazarenko, 2001). Cette méthode se base sur l'hypothèse que des relations de synonymie peuvent être propagées à travers le principe de compositionnalité. Trois règles sont proposées pour inférer des relations de synonymie entre termes complexes à partir de relations élémentaires de synonymie entre mots [4] (étape d'inférence sur la figure 1). Ainsi, deux termes complexes sont considérés comme synonymes si au moins un de leur composant dans la même position syntaxique sont synonymes. Par exemple, étant donné la relation de synonymie entre les mots *infection* et *sepsis*, les termes *wound infection* et *wound sepsis* sont identifiés comme synonymes. Les auteurs ont également montré que la qualité des relations inférées dépend de l'origine des relations entre les termes simples : des relations de synonymie issues d'un dictionnaire de langue générale contribuent à augmenter le rappel, tandis que des relations spécialisées acquises sur un corpus du domaine permettent d'améliorer la précision.

La difficulté étant de disposer de relations entre termes simples, spécifiques au domaine, l'approche inverse a été définie pour induire des relations élémentaires à partir des relations fournies par une ressource terminologique (étape d'induction sur la figure 1) (Hamon et Grabar, 2008). L'application de la méthode sur Gene Ontology permet d'acquérir 3 707 relations de synonymie avec une précision de 0,72. Nous avons pu parfois observer des décalages sémantiques entre le type de la relation issue de la ressource terminologique et la relation induite : alors que la relation initiale est une relation de synonymie, la relation induite exprime un autre type de relation. Par exemple, à la figure 1, *cell receptor complex* et *lymphocyte receptor complex* sont des synonymes dans GO tandis que la relation induite est considérée comme une relation d'hypéronymie. On peut supposer que cette modification de type sémantique entre la relation induite et la relation entre les termes complexes est dû à un usage particulier qui peut être capturé à travers le contexte terminologique ou l'usage en corpus. De plus, la synonymie étant une relation contextuelle (Cruse, 1986), il semble important de prendre en compte le contexte lors de l'acquisition des relations élémentaires.

L'objectif de notre travail est ainsi de définir une méthode permettant de contextualiser les relations élémentaires induites. Nous souhaitons pouvoir associer des catégories sémantiques issues du contexte des termes pour contraindre l'utilisation des relations induites.

Pour cela, nous avons identifié deux pistes de travail :

1. Filtrage des relations induites par leur usage dans un corpus. Les relations élémentaires induites à partir d'une ressource terminologique (étape d'induction) sont d'abord exploitées sur un corpus pour inférer des relations entre termes complexes (étape d'inférence). L'objectif est d'identifier les relations élémentaires utiles et leur associer un poids ou une confiance plus importante. Ici nous faisons l'hypothèse que les relations incorrectes ne devraient pas être utilisées dans un corpus. Ce travail permettra également d'identifier des contextes lexicaux puis sémantiques utiles pour contextualiser les relations élémentaires.
2. Exploitation d'informations sémantiques associées aux termes. Il s'agit de contraindre le champ d'application des relations entre les termes simples en exploitant les informations

4. ou des termes moins complexes, c'est-à-dire de longueur plus petite.

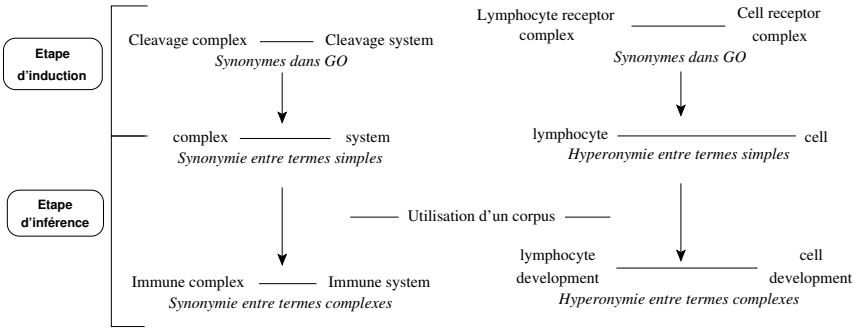


FIGURE 1 – Processus d'acquisition de relations de synonymie (inférence et induction).

sémantiques associées aux termes. Nous allons annoter les textes avec des informations sémantiques (catégories sémantiques ou rôles sémantiques). Pour cela nous allons projeter les catégories sémantiques de l'UMLS et de WordNet sur les corpus de travail ou sur les composants des termes de GO. Nous envisageons d'utiliser des méthodes d'apprentissage par analogie et par la programmation logique inductive pour en déduire les contraintes sémantiques sur les relations élémentaires.

Pour des raisons de facilité de mise en œuvre nous nous sommes pour l'instant concentrée sur la première piste, et plus particulièrement sur le filtrage des relations élémentaires induites.

5 Filtrage des relations élémentaires par l'usage en corpus

Pour filtrer les relations élémentaires en fonction de leur usage en corpus, nous avons travaillé sur le corpus BioNLP. Le corpus a d'abord été segmenté en mots et en phrases. Les mots ont été étiquetés morpho-syntaxiquement et lemmatisés avec Genia Tagger (Tsuruoka *et al.*, 2005). Nous avons ensuite extrait les termes avec YATEA (Aubin et Hamon, 2006). Les différent traitements ont été pris en charge par la plate-forme d'annotation linguistique Ogmios (Hamon et Nazarenko, 2008).

L'acquisition de relations de synonymie entre les termes du corpus a été réalisée à l'aide de SynoTerm. Nous avons utilisé les 3 707 relations élémentaires induites à partir de GO. 277 relations entre termes complexes ont été inférées sur le corpus BioNLP. 104 relations élémentaires ont été utilisées. Les résultats sont en cours d'analyse. La figure 2 présente quelques relations inférées. Nous sommes conscient que la taille du corpus utilisé aura une influence sur le volume de relations inférées. Un filtrage basé uniquement sur le corpus demande une réflexion sur la taille minimale pour appliquer cette approche, ou le recrutement de textes dans lesquels tous les mots ou termes simples en relation apparaissent.

Formation /Development ⟶ *Cell development /Cell formation*
⟶ *Lymphocyte development /Lymphocyte formation*
Contact /Adhesion ⟶ *Cell contact /Cell adhesion*
Embryonic /Fetal ⟶ *Development embryonic /Development fetal*

FIGURE 2 – Exemples de relations inférées

6 Conclusion

L'étude des relations élémentaires induites à partir de Gene Ontology nous conduit à nous intéresser à l'amélioration d'acquisition de relation de synonymie lorsqu'elle est basée sur l'analyse de la structure interne. Les pistes de travail que nous proposons visent à prendre en compte les contextes des termes en corpus de spécialité. Nous proposons d'abord d'utiliser la structure interne des termes pour acquérir des relations, puis de les filtrer ou les trier en contexte (en utilisant des corpus textuels ou les contextes terminologiques que constituent les termes de GO). Les relations élémentaires induites seront ainsi affinées en leur associant des contraintes.

Dans cette perspective, afin de commencer à mettre en œuvre nos pistes de travail, nous avons réalisé une première expérience sur des données textuelles issues du domaine biomédical (corpus BioNLP et Gene Ontology). Les relations élémentaires issues de l'analyse des termes de Gene Ontology ont été exploitées sur le corpus BioNLP pour inférer des relations de synonymie entre termes complexes. Ces premiers résultats vont nous permettre de définir précisément une méthode de filtrage basée sur le contexte et l'usage des termes.

Références

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 380–387. Springer.

AUGER, A. et BARRIÈRE, C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Special Issue on Pattern-based Approaches to Semantic Relation Extraction, Terminology*, 14(1):1–19.

BENDAOUD, R., TOUSSAINT, Y. et NAPOLI, A. (2010). L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des Nouvelles Technologies de l'Information*, E-18:133–164.

BODENREIDER, O., BURGUN, A. et RINDFLESCH, T. C. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *In* URI INIST CNRS, éditeur : *Actes de la conférence TIA-2001*, pages 11–21, Nancy, France.

BOUILLON, P., FABRE, C., SÉBILLOT, P. et JACQMIN, L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *Traitement Automatique des Langues*, 41(2):367–393.

BOURIGAULT, D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, Paris, France.

Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA) – Techniques Informatiques et structuration de terminologiques*, 18(1):87–110.

Bourigault, D. et Jacquemin, C. (2000). Constitution de ressources terminologiques. *In Ingénierie des langues*, chapitre 9, pages 215–233. Hermes Science. Sous la direction de Jean-Marie Pierrel.

Cabré, M. T. (1999). *Terminology. Theory, methods and applications*, volume 1 de *Terminology and Lexicography, Research and practice*. John Benjamins, Amsterdam/Philadelphia.

Caraballo, S. (1999). Automatic acquisition of a hypernym-labeled noun hierarchy from text. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics ACL99*, pages 120–126.

Cimiano, P., Staab, S. et Tane, J. (2000). Automatic acquisition oftaxonomies from text : Fca meets nlp. *In Proceedings of the ECMLPKDD Workshop on Adaptive Text Extraction and Mining CavtatDubrovnik Croatia*, pages 10–17.

Claveau, V. et L'Homme, M.-C. (2004). Discoveringe specific semantic relationships between nouns and verbs in a specialized french corpus. *In Proceedings of the 3rd International Workshop on Computational Terminology, CompuTerm'04*, pages 39–46, Genève, Suisse.

Claveau, V. et L'Homme, M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie-utilisation comparée de ressources endogènes et exogènes. *In Conférence TIA-2005, Rouen, 4 et 5 avril 2005*, Montréal, Canada.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

Daille, B. (2003). Conceptual structuring through term variations. *In* Bond, F., Kohonen, A., Carthy, D. M. et Villaciencio, A., éditeurs : *Proceedings of the ACL'2003 Workshop on Multiword Expressions : Analysis, Acquisition, and Treatment*, pages 9–16.

Fellbaum, C., éditeur (1998). *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferret, O. (2011). Utiliser l'amorçage pour améliorer une mesure de similarité sémantique. *In Actes de TALN 2011*, pages 1–6, Montpellier.

Ganter, B. et Wille, R. (1999). *Formal concept analysis - mathematical foundations*. Springer.

Grabar, N. et Hamon, T. (2006). Terminology structuring through the derivational morphology. *In* Salakoski, T., Ginter, F., Pyysalo, S. et Pahikkala, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 652–663. Springer.

Grabar, N. et Zweigenbaum, P. (2002). Lexically-based terminology structuring : some inherent limits. *In Proceedings of Computerm'2002 (Second Workshop on Computational Terminology)*, Taiwan.

Grefenstette, G. (1994). *Exploration in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA.

Grouin, C., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Ligozat, A.-L., Minard, A.-L., Rosset, S. et Zweigenbaum, P. (2010). Caramba : Concept, assertion, and relation annotation using machine-learning based approaches. *In Proceedings of the workshop I2B2 2010*.

HAMON, T. (2005). Indexer les documents spécialisés : les ressources terminologiques contrôlées sont-elles suffisantes ? *In* 6ème *rencontres Terminologie et Intelligence Artificielle*, pages 71–82, Rouen, France.

HAMON, T. et GRABAR, N. (2008). Acquisition of elementary synonym relations from biological structured terminology. *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing - 9th International Conference, CICLing - Proceedings*, numéro 4919 de LNCS, pages 40–51, Haifa, Israel. Springer-Verlag Berlin Heidelberg.

HAMON, T., GRAÑA, M., RAGGIO, V., GRABAR, N. et NAYA, H. (2010). Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. *In MEDINFO 2010*, pages 964–968. Stud Health Technol Inform. PMID : 20841827.

HAMON, T. et NAZARENKO, A. (2001). Detection of synonymy links between terms : experiment and results. *In Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

HAMON, T. et NAZARENKO, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *Traitement Automatique des Langues*, 49(2):127–154.

HARRIS, Z. (1990). La genèse de l'analyse des transformations et de la métalangue. *Langages*, 99:9–20. A. Daladier (resp.).

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

IBEKWE-SANJUAN, F. (2005). Inclusion lexicale et proximité sémantique entre termes. *In Actes de la conférence TIA 2005*, pages 45–57, Rouen.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

JACQUEMIN, C. (1999). Syntagmatic and paradigmatic representations of term variation. *In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, University of Maryland.

KIM, J.-D., OHTA, T., TATEISI, Y. et TSUJII, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *In ISMB (Supplement of Bioinformatics)*, pages 180–182.

LINDBERG, D., HUMPHREYS, B. et MCCRAY, A. (1993). The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

MARTIENNE, E. et MORIN, E. (1999). Using s symbolic machine learning tool to refine lexico-syntactic patterns. Rapport de Recherche 183, Institut de Recherche en Informatique de Nantes (IRIN).

MCCRAE, J. et COLLIER, N. (2008). Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9:159+.

MCINTOSH, T. et CURRAN, J. R. (2009). Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10:311+.

MEYER, I. (2001). Extracting Knowledge-rich Contexts for Terminography. *In* BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M., éditeurs : *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, Amsterdam/Philadelphia.

MORIN, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes, FRANCE, Université de Nantes, Nantes, FRANCE.

NAMER, F. et BAUD, R. (2007). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *Int J Med Inform*, 76(2-3):226–233.

PYYSALO, S., OHTA, T. et TSUJII, J. (2011). Overview of the entity relations (rel) supporting task of bionlp shared task 2011. *In Proceedings of BioNLP Shared Task 2011 Workshop*, pages 83–88, Portland, Oregon, USA. Association for Computational Linguistics.

RESNIK, P. (1993). *Selection and Information : A Class-Based Approach to Lexical Relationships*. Thèse de doctorat, University of Pennsylvania.

RØST, T. B., AKBAR, S., Øystein NYTRØ et BASGALUPP, M. (2010). Medical relation extraction with semantic grammars. *In Proceedings of the workshop I2B2 2010*.

SAGER, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

SCHUMANN, A.-K. (2011). A case study of knowledge-rich context extraction in russian. *In* KAGEURA, K. et ZWEIGENBAUM, P., éditeurs : *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 143–146, INaLCO.

SNOW, R., JURAFSKY, D. et NG, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. *In* SAUL, L. K., WEISS, Y. et BOTTOU, L., éditeurs : *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA.

SPASIC, I., ANANIADOU, S., MCNAUGHT, J. et KUMAR, A. (2005). Text mining and ontologies in biomedicine : making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology : tool for the unification of biology. *Nature genetics*, 25:25–29.

TSURUOKA, Y., TATEISHI, Y., KIM, J.-D., OHTA, T., MCNAUGHT, J., ANANIADOU, S. et TSUJII, J. (2005). Developing a robust part-of-speech tagger for biomedical text. *In Proceedings of Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, pages 382–392.

VERSPOOR, C. M., JOSLYN, C. et PAPCUN, G. J. (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. *In SIGIR workshop on Text Analysis and Search for Bioinformatics*, pages 51–56.

WEISSENBACHER, D. (2004). La relation de sysnonymie en génomique. *In Actes de la conférence RECITAL'2004*, Fès, Maroc.

YANG, H. et de ROECK, A. (2010). Extraction of medical information using crfs, context patterns, and dependency parse trees. *In Proceedings of the workshop I2B2 2010*.

ZWEIGENBAUM, P. et GRABAR, N. (2000). Liens morphologiques et structuration de terminologie. *In Actes IC'2000*, pages 325–334, Toulouse, France.

État de l'art : l'influence du domaine sur la classification de l'opinion

Dis-moi de quoi tu parles, je te dirai ce que tu penses

Morgane Marchand[1,2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Centre Nano-Innov Saclay, 91191 Gif-sur-Yvette Cedex
(2) LIMSI-CNRS, Univ. Paris-Sud
91403 Orsay Cedex
morgane.marchand@cea.fr

RÉSUMÉ

L'intérêt pour la fouille d'opinion s'est développé en même temps que se sont répandus les blogs, forums et autres plate-formes où les internautes peuvent librement exprimer leur opinion. La très grande quantité de données disponibles oblige à avoir recours à des traitements automatiques de fouille d'opinion. Cependant, la manière dont les gens expriment leur avis change selon ce dont ils parlent. Les distributions des mots utilisés sont différentes d'un domaine à l'autre. Aussi, il est très difficile d'obtenir un classifieur d'opinion fonctionnant sur tous les domaines. De plus, on ne peut appliquer sans adaptation sur un domaine cible un classifieur entraîné sur un domaine source différent. L'objet de cet article est de recenser les moyens de résoudre ce problème difficile.

ABSTRACT

State of the Art : Influence of Domain on Opinion Classification

The interest in opinion mining has grown concurrently with blogs, forums, and others platforms where the internauts can freely write about their opinion on every topic. As the amounts of available data are increasingly huge, the use of automatic methods for opinion mining becomes imperative. However, sentiment is expressed differently in different domains : words distributions can indeed differ significantly. An effective global opinion classifier is therefore hard to develop. Moreover, a classifier trained on a source domain can't be used without adaptation on a target domain. This article aims to describe the state-of-the-art methods used to solve this difficult task.

MOTS-CLÉS : État de l'art, Fouille d'opinion, Multi-domaines, Cross-domaines.

KEYWORDS: State of the art, Opinion mining, Multi-domain, Cross-domain.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 177–190,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Savoir ce que les autres pensent est, depuis toujours, une information très importante pour prendre une décision. Nous consultons des critiques de consommateurs avant d'acheter un appareil photo, des sondages avant des élections ou encore dans le domaine professionnel des lettres de recommandation. Depuis le développement d'Internet, de plus en plus de personnes rendent leurs avis disponibles. Nous avons donc facilement accès à un très large corpus d'opinion en tout genre.
Les applications possibles de la fouille d'opinion sont multiples (Pang et Lee, 2008). Elle peut, par exemple, être utilisée pour agréger des critiques, faire des systèmes de recommandation ou bien des outils de marketing et de business intelligence. Certains moteurs de recherche proposent déjà des applications pour résumer les opinions des consommateurs dans des interfaces dédiées au shopping (Blair-Goldensohn *et al.*, 2008). L'idéal serait de pouvoir disposer de telles fonctionnalités pour des recherches d'ordre général.

La diversité et la quantité de ces témoignages rendent leur traitement manuel long et coûteux. C'est pourquoi l'exploitation automatique de ces données est un enjeu majeur.

La fouille d'opinion se compose de plusieurs tâches, qu'il est utile ou non de mettre en œuvre selon les applications visées. :
– détection de la présence ou non de l'opinion ;
– classification de l'axiologie de l'opinion (positif, négatif, neutre) ;
– classification de l'intensité de l'opinion ;
– identification de l'objet de l'opinion (ce sur quoi porte l'opinion) ;
– identification de la source de l'opinion (qui exprime l'opinion).
L'analyse de l'opinion peut se situer au niveau du texte entier, du paragraphe, de la phrase ou bien du fragment selon les applications envisagées.

Dans cet article, nous nous intéresserons uniquement à la tâche de classification de l'axiologie de l'opinion, dont nous donnons un aperçu des problèmes dans la section 2, de façon générale et dans le cas particulier des domaines différents. En section 3, nous expliciterons brièvement pourquoi les techniques classiques pour la constitution des ressources ou des classifieurs sont moins efficaces lorsque l'on change de domaine d'expression. Nous dresserons alors, dans la section 4, un état de l'art des recherches actuelles visant à améliorer les performances des classifieurs dans ce cas particulier. Dans une dernière partie, nous évoquerons les travaux préliminaires effectués ainsi que les perspectives d'exploration.

2 La subjectivité dans le langage

2.1 Présentation et définition

La terminologie utilisée en fouille d'opinion est multiple : opinion, sentiment, subjectivité, polarité, etc. Nous nous intéressons ici spécifiquement à l'expression de l'opinion, qui peut se classer sur un axe positif/négatif.
On peut distinguer deux niveaux de subjectivité dans le langage (Benveniste, 1966) :

– le premier niveau n'implique pas l'expression d'une évaluation. Il témoigne simplement du degré de présence de l'énonciateur dans son énoncé. Cette présence peut être implicite ou bien explicite en fonction de la présence ou l'absence de certains marqueurs ;
– le second niveau est celui des évaluations exprimées par l'énonciateur. Elles se caractérisent par la présence d'un prédicat exprimant l'évaluation. Ce prédicat peut avoir ou non une valeur axiologique (positif, négatif, neutre...)

C'est ce deuxième niveau qui nous intéresse ici. Il est cependant parfois difficile de distinguer les deux niveaux de subjectivité et cela peut amener à des erreurs de classification.

2.2 Différences d'expression selon les domaines ou les niveaux de langage

Selon le sujet d'un texte, ce ne sont pas les mêmes mots de vocabulaire qui sont employés. On pourrait cependant penser que les expressions d'évaluation sont universelles. En effet, certains mots et certaines structures reviennent avec régularité tels que "j'adore" ou bien "je le déconseille". De plus, les dictionnaires notent que certains mots sont péjoratifs ("avare"), d'autres, au contraire, mélioratifs ("généreux"). Ainsi, selon (Pupier, 1998), il y a des mots à valeur intrinsèquement positive ("généreux, délicieux") et d'autres à valeur intrinsèquement négative ("avare, mauvais"). D'autres mots semblent en revanche neutres : "table" est l'exemple classiquement donné par les linguistes. On parle ici d'orientation *a priori*.

Néanmoins, à côté de mots intrinsèquement positifs ou négatifs, il existe des mots dont l'orientation peut changer selon le contexte dans lequel ils sont employés (Riloff et Wiebe, 2003). Il peut s'agir de mots polysémiques ou bien d'homonymes ayant des axiologies différentes. C'est le cas du "navet" qui est un légume tout à fait ordinaire en cuisine mais un film à éviter dès lors que l'on parle de cinéma. La désambiguïsation lexicale (savoir quel sens est effectivement utilisé) s'appuie justement sur les mots du contexte. Les méthodes existantes utilisent des corpus, annotés ou non, ainsi que des dictionnaires inventoriant les sens existant (Navigli, 2009).
L'orientation d'un mot non polysémique peut également changer à l'intérieur d'un même domaine, selon l'objet qu'il évalue. Par exemple pour un ordinateur portable, une batterie "large" est un inconvénient mais un écran "large" est un atout. L'orientation des mots peut aussi dépendre des préférences et de l'idéologie de l'auteur et c'est alors bien plus difficile à détecter. Les textes politiques sont notamment très sensibles à cela. Par exemple, le mot "bourgeois" est fondé sur une sémantique neutre mais quand il s'agit de préjugé ou d'opinion, ce qui est "bourgeois" est souvent mal vu.

Un problème proche de l'adaptation au domaine est l'adaptation au niveau de langage. On retrouve un vocabulaire différent selon les niveaux mais aussi des mots communs qui changent de polarité ("C'est terrible !, C'est mortel !"). Ces inversions de sens peuvent être extrêmement fortes comme le mot "bad" qui signifie exactement le contraire de son sens littéral dans le domaine du blues à une certaine époque.

Dans la prochaine partie, nous allons voir que les méthodes classiques pour obtenir des lexiques et des classifieurs d'opinions ne sont pas toujours adaptées pour prendre en compte le changement de vocabulaire induit par le changement de domaine.

3 Les problèmes d'adaptation des ressources et des classifieurs classiques

Cette partie se focalise sur les problèmes d'adaptation des ressources et des classifieurs classiques. Pour obtenir plus de détails sur les méthodes de construction classiques, le lecteur se référera à (Pang et Lee, 2008).

3.1 Les ressources

Pour la constitution de ressources, on distingue deux grandes familles d'approches. La première consiste à utiliser des dictionnaires. A partir d'un petit ensemble de mots, appelés mots racines, le lexique est étendu en utilisant les relations de synonymie et d'antonymie (Kim et Hovy, 2005; Esuli et Sebastiani, 2006) ou bien les définitions (Andreevskaia et Bergler, 2006). La seconde consiste à s'appuyer sur un corpus. Le lexique de mots racines est étendu en s'appuyant sur plusieurs indices comme les conjonctions et/mais (Hatzivassiloglou et McKeown, 1997), la cooccurrence entre mots(Turney et Littman, 2002) ou la proximité des contextes d'évaluation (Turney *et al.*, 2003). Il existe également des approches mixtes, combinant l'utilisation de corpus et de dictionnaires (Taboada *et al.*, 2011) ou bien de patrons d'extraction (Riloff et Wiebe, 2003).

Les lexiques obtenus en utilisant des dictionnaires ne sont pas spécifiques à un domaine mais leur couverture est souvent faible et ils sont pour la plupart limités au sens *a priori* des mots, c'est-à-dire hors contexte. Les méthodes à base de corpus sont quant à elles applicables à tous les corpus, quel que soit leur domaine. Cependant, le lexique finalement appris dépendra du domaine du lexique utilisé. Enfin, les patrons d'extraction sont longs et coûteux à créer. De plus, les résultats nécessitent souvent un nettoyage manuel avant d'être réellement exploitables.

3.2 Les classifieurs

En ce qui concerne la création de classifieurs pour l'axiologie positive/négative, on distingue également deux approches principales. La première consiste à utiliser principalement des lexiques et des indices linguistiques (Takamura *et al.*, 2005; Ferrari *et al.*, 2009). La seconde consiste à utiliser des données d'apprentissage afin de construire un classifieur statistique. Le type de classifieur a moins d'importance que les traits utilisées qui peuvent être des n-grammes (Pang *et al.*, 2002), des arbres de relations syntaxiques (Kudo et Matsumoto, 2004), tous les mots ou bien certains mots particuliers comme les adjectifs et les adverbes (Benamara *et al.*, 2007).

Les classifieurs développés à partir de ressources générales ont plusieurs défauts. En effet ces ressources sont trop générales et ne captent pas la spécificité des domaines. Par exemple, (Denecke, 2009) teste les score du lexique général SentiWordNet dans la tâche de classification des opinions sur six corpus différents. Leurs classifieurs statistiques mono-domaines ont de bien meilleurs résultats que les classifieurs à base de règles utilisant uniquement les mots de SentiWordNet. Un autre problème des ressources générales est que certains mots *a priori* positifs ou négatifs peuvent en réalité être employés dans des contextes neutres voire de polarité opposée (Wilson *et al.*, 2009). Quant aux classifieurs développés sur un domaine particulier, les utiliser

directement sur d'autres domaines donne en général de mauvais résultats. Par exemple, dans (Aue et Gamon, 2005), les auteurs comparent des classifieurs entraînés sur quatre domaines différents. Leurs résultats montrent que l'utilisation d'un classifieur entraîné sur un domaine source différent du domaine cible fait perdre entre 2 et 38 % d'exactitude (*accuracy*).

4 Ressources et techniques pour l'adaptation au domaine

Afin de surmonter les défauts de performance des méthodes classiques, la première possibilité est de s'attacher à développer des ressources générales plus performantes (section 4.1). Le but est d'obtenir une performance acceptable sur tous les domaines ou, au moins, sur un grand nombre de domaines. Une autre possibilité est de développer des méthodes permettant, à moindre coût, d'adapter automatiquement une ressource générale à un domaine particulier (section 4.2). Enfin, lorsque l'on dispose déjà de ressources ou d'outils adaptés à un domaine particulier, on peut les adapter à un domaine proche (section 4.3).

4.1 Améliorer les performances des classifieurs généraux

Comme nous l'avons vu précédemment, les lexiques d'opinion généraux donnent des scores de polarité *a priori*. Or cet *a priori* change selon le contexte et il faudrait disposer de lexiques capables d'en rendre compte.
Il existe un flou sur ce qu'on appelle le contexte d'un mot d'opinion : cela peut aller de la cible directe de l'opinion (Jijkoun *et al.*, 2010) à un sac de mots représentant le thème abordé (Li et Zong, 2008). Un lexique donnant des scores différents selon l'étiquette grammaticale du mot, comme SentiWordNet, peut être considéré comme faiblement contextuel (Dang *et al.*, 2010). On peut également imaginer des lexiques d'opinion généraux bien plus fortement contextuels. Par exemple, (Gindl *et al.*, 2010) créent tout d'abord deux lexiques contextuels et valués sur deux corpus A et B. Ils déterminent ensuite pour quels termes l'ajout du contexte a été utile, nocif ou neutre pour A et B. Les résultats obtenus grâce au lexique contextuel sont ainsi comparés à ceux obtenus grâce au lexique non-contextuel. Ils ne gardent ensuite que les termes contextuels qui sont soit utiles soit neutres sur les deux domaines à la fois, créant ainsi un lexique contextuel qui donne de bon résultats sur plusieurs domaines.
(Wilson *et al.*, 2009) ne créent pas un lexique contextuel, mais utilisent les relations déduites d'arbres de dépendance syntaxiques afin de tempérer les informations apportées par les orientations des mots *a priori*.

Une autre carence des lexiques d'opinion généraux classiques est de manquer souvent d'expressions polylexicales. Les mots simples sont les plus faciles à repérer mais ils ne suffisent pas à capter la richesse de l'expression de l'opinion dans la langue. Certaines expressions polylexicales sont même intégralement composées de mots qui ne sont pas eux même évaluatifs, par exemple "un coup de bol" ou bien "une bouffée d'air frais". C'est pourquoi des lexiques exhaustifs sont très difficiles à constituer.
Les travaux de (Vernier *et al.*, 2010) utilisent des marqueurs d'intensité (comme "très") pour pallier ce manque. Ils ont en effet observé que ces marqueurs s'appliquaient le plus souvent à des expressions subjectives. Ils utilisent donc des requêtes Yahoo pour sélectionner les candidats qu'ils séparent ensuite entre objectif et subjectif à l'aide d'un SVM. Ils ont évalué manuellement

l'efficacité de ce nouveau lexique par rapport à un lexique de base sur un corpus de blog qui mélangeait des textes de domaines différents. Ils observent un gain de 15,6% en précision par rapport au lexique de base pour la détection de fragments subjectifs.

Enfin, si on veut utiliser des classifieurs fondés uniquement sur des méthodes d'apprentissage statistique tout en étant les plus généraux possible, il faut des données d'apprentissage venant du plus grand nombre de domaines possible. En effet, quand on a un peu de données annotées dans plusieurs domaines, on peut faire en sorte que les domaines s'aident les uns les autres. C'est ce qu'on appelle de l'apprentissage multitâches. Dans ce cadre, fusionner les classifieurs fonctionne mieux que fusionner directement les données d'apprentissage (Li et Zong, 2008; Li *et al.*, 2011). La fusion la plus efficace dans ces travaux est réalisée par la somme pondérée des résultats des différents classifieurs, les poids de cette somme étant appris sur un petit corpus de développement du domaine cible.
Cette approche donne un classifieur donnant de bons résultats sur plusieurs domaines si l'on dispose d'un peu de données annotées pour tous. Néanmoins, il est impossible de garantir des résultats pour des domaines complètement nouveaux.

4.2 Passer automatiquement du général au particulier

Les lexiques d'opinion généraux peuvent être adaptés à un domaine particulier en utilisant les méthodes d'expansion classiques sur un corpus sélectionné pour être thématique. C'est le cas de (Harb *et al.*, 2008) qui extraient automatiquement du Web un corpus thématique en utilisant des requêtes du type « +opinion +cinema +good -bad -poor -nasty ... ». Ils extraient ensuite les adjectifs porteurs d'opinion en mesurant la cooccurrence dans les phrases entre les adjectifs candidats et les mots racines du lexique initial.
The Double Propagation method, décrite dans (Qiu *et al.*, 2009, 2011), peut être utilisée pour trouver de nouveaux mots d'opinion associés à leur cible sur un corpus particulier. Elle permet à la fois de découvrir les mots d'opinion et leurs cibles grâce à un processus d'amorçage (*bootstrap*). Les travaux se fondent sur la reconnaissance des relations grammaticales reliant les mots d'opinion et leur cible. Ces relations sont décrites au préalable manuellement. Lors de l'expansion, les relations sont détectées à l'aide d'un analyseur en dépendances. Ainsi, à partir d'un lexique d'opinion général on augmente d'une part les cibles détectées et d'autre part le lexique de mots d'opinion en utilisant les relations une fois dans un sens et une fois dans l'autre.

Une autre manière d'adapter un lexique général à un domaine particulier est non pas de l'étendre mais de le restreindre. C'est ce que font (Jijkoun *et al.*, 2010) dans leurs travaux. Ils réalisent une détection de relations syntaxiques afin d'associer à chaque mot du vocabulaire général un certain nombre de candidats pouvant être la cible de l'opinion. Ils font l'hypothèse que les cibles des opinions sont plus diverses que les autres éléments syntaxiquement liés à un terme d'opinion et ne retiennent donc que les mots cibles ayant un fort score d'entropie.

Enfin, sans étendre ou restreindre le vocabulaire, on peut juste vouloir adapter au domaine le score de polarité des mots contenus dans le lexique général. C'est par exemple le cas dans les travaux de (Choi et Cardie, 2009). A l'aide d'une formulation en problème linéaire en nombres entiers, ils exploitent les relations entre les mots d'une même expression et les mots et la polarité des expressions qui les contiennent afin d'adapter la polarité *a priori* des mots.

4.3 Faciliter l'adaptation d'un domaine à un autre

Lorsqu'on utilise des algorithmes d'apprentissage, on présuppose généralement que les données d'entraînement ont la même distribution que les données de test. En pratique, cela n'est pas le cas. On ne peut bien sûr pas espérer obtenir de bons résultats si les distributions des données sources et cibles diffèrent de manière trop importante. Cependant, si elles ne sont que légèrement différentes, l'apprentissage peut être efficace.

4.3.1 Mélanger les corpus ou les traits

Si l'on dispose d'un corpus annoté suffisamment grand, la méthode donnant les meilleurs résultats repose, de façon naturelle, sur un entraînement direct sur les données du domaine. En revanche, dans le cas où l'on ne dispose pas de données annotées, il devient utile de s'entraîner sur d'autres corpus. Dans (Yoshida *et al.*, 2011), les auteurs étudient l'influence du nombre de domaines source et cible, allant jusqu'à quatorze domaines différents. Plus le nombre de corpus source est élevé, plus les résultats sur un corpus cible différent sont bons. De plus, leur modèle probabiliste génératif permet de déterminer si la polarité inférée pour un certain mot dépend ou non du domaine du texte où se trouve le mot. Ainsi, ils construisent automatiquement des dictionnaires valués pour chaque domaine.
Afin de s'adapter plus précisément au domaine cible, des poids peuvent être attribués aux exemples (Bickel *et al.*, 2007) ou aux traits (Satpal et Sarawagi, 2007). Ces méthodes s'appliquent également à l'extraction d'information générale (Gupta et Sarawagi, 2009).

Un problème peut se poser lorsque les corpus sont hétérogènes et couvrent plusieurs domaines. Dans le domaine de la classification d'image, (Hoffman *et al.*, 2011) s'attaquent au problème de plusieurs domaines sources dont on ne connaît pas *a priori* les étiquettes. Ils séparent d'abord les domaines sources à l'aide d'une variante de l'algorithme des *k-means* avant de poursuivre plus classiquement en combinant les classifieurs appris sur les domaines ainsi séparés. A notre connaissance, il n'y a pas de travaux en classification d'opinion traitant ce problème particulier.

4.3.2 Domaine de représentation commune

Une autre approche est d'essayer de détecter des pivots, des structures communes entre deux domaines. La méthode développée dans (Blitzer *et al.*, 2006), le *Structural Correspondance Learning* (SCL) se fonde sur la recherche de pivots entre les deux domaines permettant de comparer les histogrammes de répartition des différents termes des domaines. Elle est motivée par un algorithme d'apprentissage multitâches, ASO (*Alternating Structural Optimization*), proposé par (Ando et Zhang, 2005). Cette méthode a été appliquée à la recherche d'opinion dans (Blitzer *et al.*, 2007), travaux que nous reproduisons dans la partie 5. Les pivots sont ici des mots fréquents utiles à la détermination de l'opinion dans le domaine source annoté. Des classifieurs pivots sont créés qui permettent de comparer les distributions des autres mots par rapport à ces mots pivots. Ce sont les projections de ces distributions qui deviennent les traits représentatifs des textes.
Dans (Blitzer *et al.*, 2011), les auteurs s'intéressent plus spécifiquement au cas où les supports

des domaines source et cibles (l'ensemble des mots qui apparaissent dans chaque domaine) ont peu de mots en commun. Les cooccurrences entre les termes des domaines source et cible ne sont donc pas uniquement apprises par rapport à des mots pivots communs au deux domaines mais également par rapport à des mots spécifiques à un seul domaine.

Un travail plus récent à ce sujet est celui de (Pan *et al.*, 2010). Ils se servent également comme pivots de mots indépendants du domaine sélectionné pour leur fréquence dans le domaine cible et leur information mutuelle par rapport aux étiquettes du corpus source. Ils construisent ensuite un graphe bipartite de corrélation entre les traits pivots et les traits non-pivots. Puis à l'aide d'algorithmes de *clustering* spectral, ils créent des *clusters* entre des traits dépendants des domaines source et cible. Ils obtiennent ainsi un espace de représentation commun aux deux domaines. Les résultats obtenus dans (Pan *et al.*, 2010) montre que la méthode SFA obtient de meilleurs résultats en exactitude que d'autres méthodes, dont SCL.

Plusieurs travaux mettent également en lumière que lorsque l'on peut disposer en plus d'une petite partie annotée du corpus cible, cela permet d'améliorer les résultats de manière conséquente (Daumé, 2007; Blitzer *et al.*, 2007; Aue et Gamon, 2005).

4.3.3 Comment évaluer la transportabilité d'un domaine à un autre ?

Tous les travaux étudiant la portabilité d'un domaine à un autre font état de domaines plus semblables pour lesquels le transfert se passe mieux (Denecke, 2009; Blitzer *et al.*, 2007; Aue et Gamon, 2005). La question de savoir comment mesurer la proximité de deux domaines devient donc centrale.

Dans (Ben-David *et al.*, 2007), les auteurs développent une borne supérieure pour l'erreur de généralisation d'un classifieur entraîné sur un domaine source et testé sur un domaine cible. Cette borne comprend deux termes variables. Le premier est l'erreur effectuée sur le domaine source. Le second est une mesure de la divergence entre les distributions des domaines sources et cibles sous une certaine représentation. Selon la représentation choisie pour les textes (unigrammes, bigrammes, rôles sémantiques...), les distributions des traits seront différentes. Par conséquent, la divergence entre les deux domaines dépend de la représentation choisie. En choisissant une représentation très simplifiée, on peut rendre la divergence entre les deux domaines faible. Mais alors, l'erreur effectuée sur le domaine source sera très grande. Il faut donc choisir avec soin la représentation des textes pour obtenir une divergence faible entre les deux domaines tout en conservant une erreur raisonnable sur le domaine source.
Une fois la représentation définie, se pose le problème de calculer la divergence des deux distributions. Une mesure naturelle serait la distance L_1 ou variationnelle. Cependant, cette distance n'est pas calculable à partir d'un corpus fini pour des distributions à valeur réelle. C'est pourquoi (Ben-David *et al.*, 2007) utilisent ce qu'ils appellent la A-distance. Il s'agit d'une restriction de la distance variationnelle à une collection A d'ensembles de textes issus des corpus de façon à ce que chaque élément de A soit mesurable sous les deux distributions. On obtient ainsi une borne supérieure calculable pour l'erreur de généralisation du classifieur considéré.

D'un point de vue pratique, calculer la A-distance à l'aide de données réelles est comme entraîner un classifieur pour départager les textes selon s'ils appartiennent au domaine source ou cible.

La A-distance fonctionne pour une classification de type 0/1. Les travaux de (Mansour *et al.*, 2009) introduisent la *discrepancy distance* qui peut également être utilisée pour comparer des distributions dans le cadre d'une tâche de régression.

5 Pistes de recherche et travaux préliminaires

Notre thème de recherche concerne l'adaptation au domaine pour la fouille d'opinion et la constitution automatique de lexiques pour ce problème. Les travaux cherchant à projeter deux corpus de domaines différents dans un espace commun semblent prometteurs. Aussi, nous nous sommes attachés à reproduire les travaux présentés dans (Blitzer *et al.*, 2007). Cet article décrit une heuristique pour l'adaptation au domaine appelé *Structural Correspondance Learning* (SCL). SCL utilise des données non-étiquetées provenant de deux domaines différents afin de détecter des correspondances de comportement entre des traits spécifiques au domaine source et des traits spécifiques au domaine cible.

5.1 Description de la méthode SCL

Pour réaliser leur étude, les auteurs ont constitué des corpus thématiques à partir de critiques collectées sur le site internet Amazon. Ils ont utilisé quatre corpus thématiques, *DVDs*, *kitchen*, *electronics* et *books*. Les critiques sont représentées en sac de mots en utilisant les unigrammes et les bigrammes présents. Grâce au nombre d'étoiles attribuées aux critiques, les auteurs se sont assurés que leurs corpus contiennent autant de critiques positives (quatre et cinq étoiles) que de critiques négatives (une et deux étoiles). Les textes ayant obtenus trois étoiles n'ont pas été pris en compte à cause de leur polarité ambigüe.
Les travaux des auteurs cherchent à reproduire la situation réelle où l'on dispose d'un grand nombre de données non annotées à la fois pour le domaine cible et pour le domaine source, mais seulement une petite partie de corpus source annoté. Aussi, lors de chaque expérience, on considère que l'on ne connaît les étiquettes que de 2000 critiques du corpus source : 1000 positives et 1000 négatives.

L'idée de la méthode SCL est d'établir des correspondances entre des mots du domaine source et des mots du domaine cible en fonction de leur comportement par rapport à des mots pivots communs aux deux domaines. Considérons le mot S qui n'apparaît que dans le corpus source et le mot C qui n'apparaît que dans le corpus cible. Un classifieur usuel entraîné sur le domaine source ne saura pas quoi faire de C. Mais si S et C, chacun dans son corpus, co-occurrent avec les mots pivots communs de la même façon, on peut supposer que C équivaut à S dans le domaine cible. Le classifieur devra donc traiter C comme si c'était S.

En pratique, la première étape est donc d'identifier quels mots joueront le rôle de pivots. Les auteurs commencent par sélectionner un ensemble de traits qui apparaissent fréquemment dans les deux domaines. Ces traits sont ensuite classés selon leur information mutuelle par rapport aux classes positive et négative pour les 2000 critiques du corpus source dont on connaît la polarité. Seuls les 1000 plus informatifs sont conservés. Ces traits

pivots sont donc fréquents dans les deux domaines et relativement utile à la tâche de classification de l'opinion pour le domaine source (par exemple "a-must", "loved-it", "weak", "awful", etc.)

Une fois les traits pivots sélectionnés, les auteurs modélisent la corrélation entre tous les traits des deux corpus et les traits pivots en entraînant pour chaque trait pivot un classifieur linéaire appelé classifieur pivot. Ce classifieur, appris sur l'ensemble des corpus source et cible, répond à la question : "Est-ce que le mot pivot considéré à des chances d'apparaître dans ce texte sachant tous les autres mots du texte". Les vecteurs de poids de ces classifieurs pivots sont agrégés en une matrice. Celle-ci est ensuite réduite par décomposition en valeurs singulière. Les auteurs ne conservent que 50 dimensions. Ils obtiennent ainsi une matrice de projection permettant de calculer 50 nouveaux traits (à valeur réelle) pour chaque texte source et cible.
Les textes du corpus source et cible sont représentés par un vecteur contenant à la fois les traits initiaux (les unigrammes et bigrammes) et les nouveaux traits calculés à l'aide de la matrice de projection. C'est sur ces corpus étendus source et cible qu'un classifieur est entraîné et testé. Les auteurs utilisent un classifieur linéaire dont les coefficients sont obtenus par descente stochastique de gradient.

Par rapport à un classifieur entraîné sur un domaine source et testé sur un domaine cible sans rajouter les nouveaux traits, leur approche améliore souvent les performances (10 cas sur 12). En une occasion, ils arrivent même à dépasser les performances d'un classifieur entraîné et testé sur le domaine cible.

5.2 Nos travaux de reproduction

Nous utilisons deux des corpus constitués et utilisés par les auteurs : les corpus *DVDs* et *kitchen* du *Multi-Domain Sentiment Dataset*. Le corpus source *DVDs* contient 5586 critiques et le corpus cible *kitchen* 7945 critiques également réparties entre négatives et positives. Comme dit précédemment, le domaine source contient 1000 critiques positives et 1000 critiques négatives pour lesquelles on connaît les étiquettes. En moyenne, les critiques du corpus *kitchen* contiennent 145 unigrammes et bigrammes, celles de *DVDs*, 269.

Nous avons étudié le sens d'adaptation de *DVDs* vers *kitchen*. Les références que nous utilisons sont les suivantes : un classifieur entraîné et testé sur le domaine source, un classifieur entraîné et testé sur le domaine cible et un classifieur entraîné sur le domaine source et testé sur le domaine cible sans ajouter les traits obtenus par SCL. Nous comparons également nos résultats avec ceux présentés dans (Blitzer *et al.*, 2007).

Les tests effectués ont mis en valeur le fait que le choix des traits pivots influence énormément les performances du classifieur. Les résultats fournis par les auteurs sont des résultats d'exactitude. Il nous a semblé intéressant d'étudier l'influence de la sélection des traits pivots sur la performance en précision pour les deux classes.
Nous avons sélectionné des ensembles de 1000 traits pivots de trois façons différentes :
– sélection uniquement selon l'information mutuelle (MI) par rapport aux étiquettes du domaine source ;
– sélection uniquement selon la fréquence d'apparition dans les domaines source et cible ;
– combinaison des deux critère précédent.

Le tableau 1 présente les résultats obtenus pour un classifieur entraîné sur *DVDs* et testé sur *kitchen* ainsi que les références présentées plus haut.
De plus, dans (Blitzer *et al.*, 2007), les auteurs normalisent les nouveaux traits afin que leur norme moyenne équivaille à α fois celle des anciens traits. Ils obtiennent de cette façon de meilleurs résultats. La dernière ligne du tableau présente donc les résultats avec un seuil α de pondération que nous avons expérimentalement fixé à 0,5.

	Blitzer et al.	Exactitude (*Accuracy*)	Précision classe positive	Précision classe négative	Rappel classe positive	Rappel classe négative
Réf. source->cible	74,0	78,5	79,4	77,6	76,5	80,4
Réf. source->source	82,4	81,8	80,3	83,4	84,6	79,0
Réf. cible->cible	87,7	87,7	88,4	87,0	86,4	88,9
Pivots : fréquence	79,4	79,8	80,9	78,7	77,6	81,9
Pivots : MI	.	79,6	85,0	75,7	71,6	87,6
Pivots : mixte	.	79,9	83,9	76,7	73,6	86,1
Pivots : mixte pond.	81,4	80,7	82,5	79,13	77,6	83,8

TABLE 1 – Résultats pour un classifieur entraîné sur *DVDs* et testé sur *kitchen*

Nous observons quelques différences de résultats entre l'article original et notre implémentation, notamment pour la référence domaine source sur domaine cible. Ces différences s'expliquent par l'utilisation d'un classifieur SVM à noyau linéaire dans notre cas, alors que les auteurs utilisent une descente de gradient stochastique pour déterminer les coefficients de leur classifieur linéaire. Nous observons cependant également une augmentation des résultats grâce à la méthode SCL.

Les pivots sélectionnés uniquement par la fréquence amènent une petite amélioration par rapport à la référence sans toutefois changer l'écart de performance entre la classe positive et la classe négative. Les pivots sélectionnés uniquement par MI, quant à eux, favorisent bien plus la classe positive. En combinant les deux critères de sélection on arrive à réduire un peu cette différence de performance entre les deux classes, d'autant plus si l'on pondère la contribution des nouveaux et des anciens traits.
Nous observons donc une difficulté particulière à la classe négative. Plus de textes positifs sont faussement classés en négatif que l'inverse. Une difficulté similaire a été notée par (Vernier *et al.*, 2009) pour la détection précise de passages subjectifs négatifs. Il faudra donc porter une attention particulière au traitement des opinions négatives.

5.3 Perspectives

L'utilisation de la matrice de projection créée par la méthode SCL est donc utile à la classification des opinions. Cependant, elle peut également réaliser de mauvais alignements. Cela peut notamment arriver lorsqu'un des corpus est plus hétérogène que l'autre. Par exemple le corpus *DVDs*, bien que rassemblant des textes d'un même domaine, fait référence à plusieurs sujets qui sont les sujets des films. Les mots se rapportant aux sujets ne sont pas informatifs pour

notre tâche de classification de l'opinion. Ils apparaissent peu fréquemment en proportion du corpus et risquent d'être mis en corrélation avec des mots du second domaine peu fréquents mais informatifs. Lorsque le classifieur est adapté du domaine hétérogène vers le domaine homogène, il manque donc les informations contenus dans les mots peu fréquents et informatifs du domaine cible. Dans l'autre sens, le classifieur va attribuer des poids à des mots qui ne sont pas informatifs pour la classification d'opinions.

L'utilisation d'une matrice de projection obtenue par une décomposition en valeurs singulières rend l'interprétation des résultats plus difficiles car les traits finaux ne sont plus des unigrammes ou des bigrammes. Nous aimerions pouvoir rendre cette méthode plus interprétable, c'est-à-dire garder des traits liés aux mots de façon directe. Notre idée serait d'utiliser une méthode s'inspirant de (Pan *et al.*, 2010). Une fois les traits sources et cibles projetés dans l'espace commun nouvellement créé, on peut les regrouper en *clusters*. Ce sont ces *clusters* qui seraient alors les nouveaux traits.
Une autre possibilité est d'utiliser l'hyperplan séparateur du classifieur afin de donner des scores d'opinion aux termes cibles qui serait la distance à cet hyperplan séparateur. Nous faisons l'hypothèse que les mots réellement polarisés auront une grande distance à cet hyperplan.

6 Conclusion

Nous nous sommes intéressés dans cet article à la fouille d'opinion et plus particulièrement à la classification de l'opinion et nous avons présenté un état de l'art des différentes méthodes utilisées pour cette tâche, en particulier pour traiter le problème de l'adaptation au domaine. Nous avons vu que l'expression de l'opinion prend des formes très variées et qui dépendent du domaine où l'on se place. Un mot ayant une polarité neutre dans un certain contexte peut avoir une polarité positive dans un autre. C'est pourquoi il est très difficile de mettre au point un classifieur ayant de bonnes performances dans tous les domaines.

Les pistes étudiées pour pallier ce problème sont multiples. On peut tout d'abord améliorer les ressources générales, notamment en créant des lexiques contextuels précis. Une autre approche est de développer des techniques pour particulariser automatiquement des ressources ou des classifieurs généraux à l'aide d'un corpus mono-domaine spécifique. Enfin, une troisième possibilité est de travailler sur l'adaptation entre domaines. Pour cela, on peut projeter l'espace cible sur l'espace source ou bien projeter les deux espaces dans un espace commun. La difficulté réside alors dans la détermination de cet espace de projection.

Nous avons également présenté nos premières pistes de recherche pour définir cet espace de projection de telle sorte qu'il reste lié à un lexique, pour rester interprétable.

Références

ANDO, R. et ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853.

ANDREEVSKAIA, A. et BERGLER, S. (2006). Mining wordnet for fuzzy sentiment : Sentiment tag extraction from wordnet glosses. *In EACL*, volume 6.

AUE, A. et GAMON, M. (2005). Customizing sentiment classifiers to new domains : A case study. *In Recent Advances in Natural Language Processing*.

BEN-DAVID, S., BLITZER, J., CRAMMER, K. et PEREIRA, F. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137.

BENAMARA, F., CESARANO, C., PICARIELLO, A., REFORGIATO, D. et SUBRAHMANIAN, V. (2007). Sentiment analysis : Adjectives and adverbs are better than adjectives alone. *In ICWSM*.

BENVENISTE, E. (1966). *Problèmes de linguistique générale I*. Gallimard.

BICKEL, S., BRÜCKNER, M. et SCHEFFER, T. (2007). Discriminative learning for differing training and test distributions. *In 24th international conference on Machine learning*. ACM.

BLAIR-GOLDENSOHN, S., HANNAN, K., MCDONALD, R., NEYLON, T., REIS, G. et REYNAR, J. (2008). Building a sentiment summarizer for local service reviews. *In WWW Workshop on NLP*.

BLITZER, J., DREDZE, M. et PEREIRA, F. (2007). Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification. *In Annual Meeting of the ACL*.

BLITZER, J., FOSTER, D. et KAKADE, S. (2011). Domain adaptation with coupled subspaces. *Journal of Machine Learning Research - Proceedings Track*, 15:173–181.

BLITZER, J., MCDONALD, R. et PEREIRA, F. (2006). Domain adaptation with structural correspondence learning. *In EMNLP*.

CHOI, Y. et CARDIE, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *In EMNLP*.

DANG, Y., ZHANG, Y. et CHEN, H. (2010). A lexicon enhanced method for sentiment classification : An experiment on online product reviews.

DAUMÉ, H. (2007). Frustratingly easy domain adaptation. *In Annual Meeting of the ACL*.

DENECKE, K. (2009). Are sentiwordnet scores suited for multi-domain sentiment classification ? *In International Conference on Digital Information Management*.

ESULI, A. et SEBASTIANI, F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. *In LREC*.

FERRARI, S., CHARNOIS, T., MATHET, Y., RIOULT, F. et LEGALLOIS, D. (2009). Analyse de discours évaluatif, modèle linguistique et applications. *RNTI*, E-17:71–93.

GINDL, S., WEICHSELBRAUN, A. et SCHARL, A. (2010). Cross-domain contextualisation of sentiment lexicons. *European Conference on Artificial Intelligence*.

GUPTA, R. et SARAWAGI, S. (2009). Domain adaptation of information extraction models. *ACM SIGMOD Record*, 37(4):35–40.

HARB, A., DRAY, G., PLANTIÉ, M., PONCELET, P., ROCHE, M., TROUSSET, F. *et al.* (2008). Détection d'opinion : Apprenons les bons adjectifs ! *Atelier FOuille des Données d'OPinions*.

HATZIVASSILOGLOU, V. et MCKEOWN, K. (1997). Predicting the semantic orientation of adjectives. *In EACL*.

HOFFMAN, J., SAENKO, K., KULIS, B. et DARRELL, T. (2011). Domain adaptation with multiple latent domains. *In NIPS Domain Adaptation Workshop*.

JIJKOUN, V., RIJKE, M. et WEERKAMP, W. (2010). Generating focused topic-specific sentiment lexicons. *In Annual Meeting of the ACL*.

KIM, S. et HOVY, E. (2005). Automatic detection of opinion bearing words and sentences. *In International Joint Conference on Natural Language Processing*, pages 61–66.

KUDO, T. et MATSUMOTO, Y. (2004). A boosting algorithm for classification of semi-structured text. *In EMNLP*.

LI, S., HUANG, C. et ZONG, C. (2011). Multi-domain sentiment classification with classifier combination. *Journal of Computer Science and Technology*, 26:25–33.

LI, S. et ZONG, C. (2008). Multi-domain sentiment classification. *In Annual Meeting of the ACL*.

MANSOUR, Y., MOHRI, M. et ROSTAMIZADEH, A. (2009). Domain adaptation : Learning bounds and algorithms. *In Conference on Learning Theory*.

NAVIGLI, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys*.

PAN, S., NI, X., SUN, J., YANG, Q. et CHEN, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. *In International Conference on World Wide Web*.

PANG, B. et LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrival*, 2:1–2.

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques.

PUPIER, P. (1998). Une première systématique des évaluatifs en français. *Revue québécoise de linguistique*, 26(1).

QIU, G., LIU, B., BU, J. et CHEN, C. (2009). Expanding domain sentiment lexicon through double propagation. *In International Joint Conference on Artificial Intelligence*.

QIU, G., LIU, B., BU, J. et CHEN, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37:9–27.

RILOFF, E. et WIEBE, J. (2003). Learning extraction patterns for subjective expressions. *In EMNLP*.

SATPAL, S. et SARAWAGI, S. (2007). Domain adaptation of conditional probability models via feature subsetting. *In Knowledge Discovery in Databases : PKDD*.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K. et STEDE, M. (2011). Lexicon-based methods for sentiment analysis. *In Computational linguistics*.

TAKAMURA, H., INUI, T. et OKUMURA, M. (2005). Extracting semantic orientations of words using spin model. *In ACL*.

TURNEY, P. et LITTMAN, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Erb-1094, Institute for Information Technology, Canada.

TURNEY, P., LITTMAN, M. *et al.* (2003). Measuring praise and criticism : Inference of semantic orientation from association. *In ACM Transactions on Information Systems*.

VERNIER, M., MONCEAUX, L. et DAILLE, B. (2010). Learning subjectivity phrases missing from resources through a large set of semantic tests. *In LREC*.

VERNIER, M., MONCEAUX, L. et DUBREIL, E. (2009). Catégorisation sémantico-discursive des évaluations exprimées dans la blogosphère. *In TALN*.

WILSON, T., WIEBE, J. et HOFFMANN, P. (2009). Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35:339–433.

YOSHIDA, Y., HIRAO, T., IWATA, T., NAGATA, M. et MATSUMOTO, Y. (2011). Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polaritys. *In Porceedings of AAAI*.

Typologie des questions à réponses multiples pour un système de question-réponse*

Mathieu-Henri Falco
LIMSI-CNRS, Université Paris-Sud, 91403 Orsay, France
falco@limsi.fr

RÉSUMÉ

L'évaluation des systèmes de question-réponse lors des campagnes repose généralement sur la validité d'une réponse individuelle supportée par un passage (question factuelle) ou d'un groupe de réponses toutes contenues dans un même passage (questions listes). Ce cadre évaluatif empêche donc de fournir un ensemble de plusieurs réponses individuelles et ne permet également pas de fournir des réponses provenant de documents différents. Ce recoupement inter-documents peut être necessaire pour construire une réponse composée de plusieurs éléments afin d'être le plus complet possible. De plus une grande majorité de questions formulées au singulier et semblant n'attendre qu'une seule réponse se trouve être des questions possédant plusieurs réponses correctes. Nous présentons ici une typologie des questions à réponses multiples ainsi qu'un aperçu sur les problèmes posés à un système de question-réponse par ce type de question.

ABSTRACT

Typology of Multiple Answer Questions for a Question-answering System

The evaluation campaigns of question-answering systems are generally based on the validity of an individual answer supported by a passage (for a factual question) or a group of answers coming all from a same supporting passage (for a list question). This framework does not allow the possibily to answer with a set of answers, nor with answers gathered from several documents. This cross-checking can be needed for building an answer composed of several elements in order to be as accurate as possible. Besides a large majority of questions with a singular form seems to be answered with a single answer whereas they can be satisfied with many. We present here a typology of questions with multiple answers and an overview of problems encountered by a question-answering system with this kind of questions.

MOTS-CLÉS : question-réponse, questions à réponses multiples, question liste.

KEYWORDS: question-answering, multiple answer questions, list question.

*Ces travaux ont été partiellement financés par OSEO dans le cadre du programme QUAERO.

1 Introduction

Les systèmes de question-réponse (*SQR*) ont pour but de fournir une réponse précise à une question formulée en langue naturelle par un utilisateur : ils peuvent travailler à partir de bases de données et/ou de collections de documents ; nous nous intéressons ici uniquement à ceux interrogeant un corpus de documents. Ces SQR combinent plusieurs domaines dont notamment la recherche d'information et le TAL à travers l'extraction d'informations. En effet, là où des moteurs de recherche renvoient des références de documents (avec éventuellement un extrait) suite à une requête sous forme de mots-clés, les SQR travaillent à partir d'une question en langue dont tous les mots ne sont pas forcément pertinents pour la recherche d'information, sélectionnent un ensemble de documents de la collection puis extraient la réponse précise de ces documents afin de la présenter à l'utilisateur (éventuellement accompagnée de l'extrait contenant cette réponse) .

Les SQR existants utilisent des approches variées, pouvant s'appliquer sur la totalité du système ou seulement certaines parties. Par exemple, des systèmes utilisent une représentation de la question et des documents logique (Moldovan *et al.*, 2007) ou discursive (Bos *et al.*, 2007a). La syntaxe peut être utilisée au niveau de l'extraction de la réponse : par exemple pour la fusion d'information multidocuments à l'aide de dépendances syntaxiques (Moriceau et Tannier, 2010), (Katz et Lin, 2003). Des heuristiques de distance (Fangtao *et al.*, 2008) ou un apprentissage automatique (Grappy, 2011) peuvent être utilisés pour la validation d'un candidat-réponse. Au final, les SQR se trouvent souvent bridés par le processus d'évaluation actuelle des campagnes à savoir fournir, pour chaque question, plusieurs réponses (généralement de trois à cinq) sous la forme d'une paire référence du document/réponse précise éventuellement accompagnée d 'un extrait du document d'où la réponse a été extraite (passage justificatif) comme pour Quaero 2010 (Quintard *et al.*, 2010) ou TREC-2007 (Dang *et al.*, 2007) (toutes les éditions et guidelines des campagnes CLEF[1] et TREC[2] sont en ligne).

Nous avons choisi de nous intéresser plus particulièrement aux questions que nous appellerons "questions à réponses multiples" comme par exemple les questions *Quels sont les sept astres du système solaire visibles à l'oeil nu ?* (le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter, Saturne) ou *Quand le Paris Saint-Germain a-t-il été sacré champion de France de football ?* (1986 et 1994 pour l'équipe professionnelle homme). Ces questions ne sont que peu ou pas évaluées lors des campagnes d'évaluation des SQR. Pourtant, elles présentent un intérêt tant au niveau de l'analyse de la question que de l'extraction des réponses et de leur présentation à l'utilisateur. En effet, un système doit être capable dans le meilleur des cas d'extraire une liste de réponses bien formée d'un document mais, le plus souvent, le système doit reconstituer une liste à partir d'éléments provenant de documents différents. Nous avons choisi de nous intéresser aux SQR qui interrogent le Web car cela nous permet de travailler en domaine ouvert et, étant donné le nombre important de documents, le travail de recoupement des réponses multi-documents s'avère indispensable.

Dans cet article, nous commençons donc par définir le terme **question à réponses multiples** (*question-ARM*) et présentons un état de l'art concernant le traitement de ce type de questions par les SQR ainsi que les élèments structuraux sources de réponses de type liste. Dans la section 3, nous présentons les observations constatées sur les données de deux campagnes d'évaluation proposant des questions-ARM. Les sections 4 et 5 présenteront respectivement notre corpus

[1] `http://nlp.uned.es/clef-qa/`
[2] `http://trec.nist.gov/data/qamain.html`

d'étude et les typologies que nous avons définies. Enfin, la section 6 présentera les perspectives envisagées.

2 Contexte et état de l'art

Une question à réponses multiples est une question pour laquelle plusieurs réponses peuvent être correctes. La forme la plus évidente de réponse à ce type de question est bien sûr la liste, par exemple :

question : *Quels sont les sept astres du système solaire visibles à l'oeil nu ?*
réponse : le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter, Saturne
passage : Les astres visibles à l'oeil nu, **le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter et Saturne** tiennent leur nom du monde romain.

Les éléments composant la liste de réponses peuvent bien sûr être déjà sous la forme d'une liste dans un document mais ils peuvent aussi être répartis dans un document ou même dans plusieurs documents, par exemple :

question : *Quand le Paris Saint-Germain a-t-il été sacré champion de France de football ?*
réponse : 1986 et 1994
document 1 : Le PSG champion de France **1986**, vu par son entraîneur, Gérard Houiller.
document 2 : Les positions resteront les mêmes durant tout le reste de la saison : le PSG sera sacré champion de France **1994**.

Nous présentons ici comment ces questions sont abordées par les SQR ainsi que les éléments structuraux qui permettent d'identifier les réponses dans les textes.

2.1 L'évaluation des SQR

L'évaluation des SQR peut se faire au niveau de la satisfaction utilisateur (point de vue applicatif et qualitatif) ou par l'intermédiaire d'une métrique (point de vue comparatif car quantitatif). Les campagnes d'évaluations de SQR ont pour but de jauger les performances des différentes approches et proposent pour cela un nombre de questions significatif pour les catégories les plus fréquentes : *factuelles*, *booléennes*, *définition*, *complexes*, *liste* et *nil* (questions n'ayant pas de réponses dans la collection de documents). Les systèmes doivent fournir plusieurs réponses pour chaque question (de trois à cinq généralement) et sont le plus souvent évalués grâce à la métrique du *MRR* (Mean Reciprocal Rank) qui favorise ainsi les SQR fournissant une réponse correcte dans les premiers rangs.

Il est très difficile de garantir qu'une unique réponse correcte puisse être obtenue à partir de la collection de documents disponible pour l'évaluation, ce qui serait peu intéressant d'ailleurs, et une évaluation humaine des réponses doit parfois avoir lieu pour juger la réponse ainsi que le passage justificatif.

Dans les campagnes traitant des *question-listes* (questions de type liste), pour indiquer qu'on

n'attend pas une réponse unique, une marque de pluriel est toujours présente mais le nombre de réponses attendues n'est pas toujours mentionné dans la question comme dans les exemples suivants : *Quelles sont les 4 localisations possibles des neuroblastomes ?* (EQueR, Quaero 2008, 2009, TREC 2001, 2002) ou *Quels sont les secteurs qui recrutent ?* (Quaero 2010, TREC 2003 à 2007). La métrique utilisée pour évaluer les réponses à ce type de questions est alors dans le premier cas la précision moyenne (nombre de réponses correctes/nombre de réponses attendues) et dans le second la F-mesure (en considérant l'ensemble des réponses jugées correctes par les évaluateurs.

Par l'utilisation du MRR, les campagnes analysant un triplet question/réponse/passage obligent donc les SQR à faire un choix d'au plus N réponses par question. Une réponse issue d'un recoupement d'informations entre plusieurs documents est donc difficile à justifier dans le cadre d'une campagne d'évaluation. De plus, la réponse et le passage doivent obligatoirement être du texte issu d'un document de la collection alors qu'il peut être parfois plus pertinent de renvoyer un élément structural (un tableau par exemple). Ces éléments structuraux sont très présents dans les documents Web mais, de toutes les campagnes évoquées jusqu'à présent, seule Quaero utilise une collection de documents Web et impose un format de réponse assez identique à celui des autres campagnes (Quintard *et al.*, 2010).

2.2 Le traitement des questions à réponses multiples par les SQR

Les questions-listes sont un cas particulier des questions-ARM : elles attendent comme réponse une liste d'items provenant d'une même entité (phrase ou document). Parmi les SQR ayant participé aux campagnes proposant des questions dont la réponse est de type liste, plusieurs ont adapté leur traitement de questions factuelles aux listes. Cette adaptation consiste à utiliser la liste ordonnée de leurs réponses trouvées dans la collection en renvoyant directement un top-N de leurs réponses, le nombre N pouvant être fixe (par exemple 5 pour (Chu-carroll *et al.*, 2004) et 20 pour (Wu *et al.*, 2003)) ou dépendant d'un nombre de réponses attendues présent dans la question (Bos *et al.*, 2007b) ou d'un seuil déterminé par le SQR selon son système d'ordonnancement (Kaisser et Becker, 2004) (Schlaefer *et al.*, 2007).

Les SQR ayant développé un traitement spécifique pour les listes ont notamment utilisé la détection de doublon pour éviter la redondance de candidat-réponse (Katz *et al.*, 2006) et certains utilisent en plus la réconciliation de référence à l'aide de ressources extérieures (Schlaefer *et al.*, 2007), (Dan I. Moldovan and et Bowden, 2007). À travers l'expansion de requête, la co-occurrence des candidats-réponses, au niveau de la phrase ou du document, sert notamment de critère de validation (Razmara et Kosseim, 2008) (Wang *et al.*, 2008) (Figueroa et Neumann, 2008).

La plupart de ces SQR utilisent donc des ressources extérieures comme des bases de données ou le web, or nous ne souhaitons pas en dépendre et seulement utiliser une collection de documents. De plus l'aspect multi-document n'est vu généralement qu'en phase de validation par la co-occurrence.

2.3 Les éléments structuraux

Nous nous sommes intéressés aux éléments structuraux que sont l'objet tableau et l'objet liste car nous nous attendons à ce qu'ils soient une source de réponses à des questions-ARM (nous ne nous intéressons qu'aux données textuelles et avons donc laissé de côté les images, figures, animations flash, etc.). Nous considérons ici le terme d'objet tableau comme une structure à au moins deux lignes et deux colonnes, et l'objet liste comme une constitution de plusieurs entités disposées horizontalement (énumération dans une phrase) ou verticalement (amorce se terminant par le symbole deux-points et un item par ligne par exemple).

Les listes ont été l'objet d'études approfondies du point de vue discursif afin de mieux cerner la structure d'un document (Ho-Dac, 2007). Les travaux de (Péry-Woodley, 2000), (Luc, 2001), (Bras *et al.*, 2008), (Laignelet, 2009), (Ho-Dac *et al.*, 2010) ont beaucoup traité de cette question et ont ainsi défini l'objet répondant au terme de "structure énumérative" comme étant composé d'une amorce (phrase introductrice), d'une énumération composée d'items (entité co-énumérée caractérisée par diverses marques typographiques, dispositionnelles, lexico-syntaxiques). Plusieurs travaux applicatifs se sont intéressés aux listes dans le cadre du peuplement d'ontologie (Laignelet *et al.*, 2011), les entités nommées (Jacquemin et Bush, 2000) et de l'analyse syntaxique. En effet, l'objet liste est par nature difficile à analyser syntaxiquement au sens où il peut utiliser la verticalité, une autre ponctuation (le point-virgule entre les items), une typographie assez libre (choix des puces) et créer des liens syntaxiques entre les items, l'amorce et la conclusion. Xerox a creusé dans cette direction à travers les travaux de (Aït-Mokhtar *et al.*, 2003) et (Gala, 2003).

Les tableaux ont été traités du point de vue HTML avec pour but de typer les cases, soit à des fins de visualisations ergonomiques, soit pour de l'extraction d'information. Deux types d'approches dominent : à bases de règles (Gatterbauer *et al.*, 2007), (Tajima et Ohnishi, 2008) et par apprentissage automatique sur un corpus annoté manuellement (Wang et Hu, 2002).

3 Premières observations sur des corpus de campagnes d'évaluation proposant des questions-ARM

Nous nous sommes intéressés dans un premier temps aux campagnes pour le français EQueR-Evalda 2004 (Ayache *et al.*, 2006) et QUAERO 2008 (Quintard *et al.*, 2010) qui comportaient des questions-listes et pour lesquelles nous avions accès aux collections (voir tableau 1). La campagne d'évaluation EQueR a proposé deux tâches : une générique sur une collection hétérogène de textes journalistiques (désignée ici par *Eq-Jour*) et une spécifique sur une collection de textes médicaux (désignée par *Eq-Méd*).

3.1 Collecte des données

Dans un premier temps, nous sommes partis des questions typées par les évaluateurs des campagnes et avons étudié les questions-listes. Ces questions portaient toutes une marque de pluriel sous forme de nombre de réponses attendu et se formulaient syntaxiquement sous quatre patrons (voir tableau 2).

	Eq-Méd	Eq-Jour	Quaero
Domaine	médical	presse, politique	ouvert
Format des documents	texte	texte	HTML
Nombre de documents	5 623	557 300	499 736
Taille de la collection	0,135 Go	1,5 Go	5 Go
Nombre de questions-listes	25	30	29

TAB. 1 – Caractéristiques des corpus EQueR et Quaero

	Eq-Méd	Eq-Jour	Quaero
Citez X	13	5	14
Quels sont les X ?	12	22	15
Qui sont les X ?	0	2	0
Comment se prénommaient les X ?	0	1	0
Nombre de questions-listes	25	30	29

TAB. 2 – Nombre de questions-listes par forme syntaxique (X est le nombre de réponses attendu).

Nous avons ensuite effectué une première étude des documents contenant des réponses correctes, documents fournis par les organisateurs des campagnes sous forme d'un fichier de référence. Nous considérons ici qu'une réponse est correcte si elle répond à la question (validation humaine), même s'il existe des réponses correctes plus pertinentes (au sens de plus récentes par exemple, ou bien satisfaisant plus l'utilisateur dans un cadre applicatif), et même si la question attendait plusieurs réponses de façon explicite (nombre déterminé dans la question). Nous utiliserons le terme *réponse-liste* pour désigner le groupe de réponses correctes à une question en attendant un nombre déterminé.

Nous avons utilisé le moteur de recherche Lucene[3] pour rechercher les documents contenant au moins une réponse aux questions-listes puisque les réponses de fichier de référence n'étaient pas forcément exhaustives. Les requêtes ont été formulées soit à partir des termes de la question jugés importants, soit à partir des réponses des fichiers références. Nous avons étudié manuellement jusqu'à 50 extraits de documents par question puis les documents entiers si les snippets étaient pertinents. Les requêtes ont ensuite été reformulées à l'aide de synonymes pour les termes de la question et des réponses nouvellement trouvées afin d'augmenter le nombre de passages-réponses. Nous avons arrêté la collection quand nous n'observions plus de nouvelle réponse ou de nouveaux phénomènes.

3.2 Étude des couples questions-ARM/réponses

L'étude des réponses aux questions-listes a révélé un nombre important de passages-réponses pour les trois collections (voir tableau 3). Les résultats montrent que la forme préférentielle d'une réponse à une question-liste dans ces campagnes est majoritairement la phrase comme par exemple :

[3] `http://lucene.apache.org/core/`

	Eq-Méd	Eq-Jour	Quaero
Nombre de questions-listes	25	30	29
Nombre de passages-réponses	56	112	122
Nombre moyen de passages-réponses par question	2,33	3,73	4,21
Passage sous forme de phrase	**30 (51,85 %)**	**73 (69,67 %)**	**85 (70,25 %)**
Passage sous forme de paragraphe	11 (20,37 %)	37 (33,04 %)	19 (15,57 %)
Passage sous forme de liste	15 (27,78 %)	2 (1,79 %)	17 (13,93 %)
Passage sous forme de tableau	0 (0 %)	0 (0 %)	1 (0,82 %)

TAB. 3 – Forme du passage-réponse.

	Eq-Méd	Eq-Jour	Quaero
Nombre de questions-listes	24	30	29
Nombre de questions avec plusieurs passages-réponses dans un même document	8 (33,33 %)	4 (13,33 %)	12 (41,38 %)
Nombre de questions où un document contient une réponse-liste	12 (50 %)	18 (60 %)	22 (75,86 %)
Nombre de questions où la réponse-liste est obligatoirement inter-document	**6 (25 %)**	**0 (0 %)**	**0 (0 %)**
Nombre de passages-réponses	56	112	122

TAB. 4 – Répartition des réponses dans les documents.

question : *Quels sont les 7 astres du système solaire visibles à l'oeil nu ?*
passage-réponse : *Les astres visibles à l'oeil nu, le Soleil, la Lune, Mercure, Vénus, Mars, Jupiter et Saturne tiennent leur nom du monde romain.*

Cette répartition centrée sur un bloc de texte continu (phrase, paragraphe, liste) contenant toutes les réponses est due aux choix des organisateurs de la campagne. Les questions-listes générées pour Quaero l'étaient notamment sur la base d'un document devant contenir tous les éléments permettant d'y répondre.

Nous avons ensuite étudié la répartition des passages-réponses dans les documents (voir tableau 4). Il en a résulté une confirmation d'une redondance des réponses inter-documents et également intra-document. La redondance inter-documents était totale au sens où chaque document contenant une réponse correcte contenait aussi la réponse-liste : seul un quart des questions de Eq-Méd nécessitait au moins deux documents de Eq-Méd pour pouvoir composer l'ensemble des réponses attendues (il n'existait pas de document unique contenant le nombre de réponse attendu pour 25 % de ces questions de Eq-Méd).

Devant le peu de questions soulevant une nécessité de traitement inter-document dans ces collections, nous avons donc décidé de constituer un autre corpus d'étude pour les questions-ARM.

4 Corpus d'étude pour les questions-ARM

Afin d'étudier en détail la forme des réponses dans le but de mieux les extraire automatiquement, nous avons constitué un corpus d'étude pour les questions-ARM en générant d'abord des questions-ARM puis en récupérant des documents permettant d'y répondre.

4.1 Constitution et caractéristiques du corpus

Nous avons d'abord repris sept questions listes existantes dans EQueR et Quaero en supprimant le nombre de réponses attendu (*Qui sont les ~~huit~~ personnages de "Disney Princess" ?*) ou en changeant des termes (*Quels pays étaient candidats à l'organisation de la coupe du monde 2018 ?* au lieu de *2006*). Nous avons ensuite imaginé des questions propices aux réponses multiples : par exemple, *Qui a incarné Batman ?*, *Quand la France a-t-elle perdu son triple-A ?*. En utilisant trois moteurs de recherche sur Internet (Exalead[4], Bing[5] et Google[6]), nous avons collecté des documents contenant au moins une réponse correcte. Ainsi un document peut ne contenir qu'un seul pays candidat à l'organisation de la coupe du monde 2018 ou qu'un seul nom d'acteur ayant incarné Batman et non pas forcément la réponse-liste. Si ce document contenait une ou des réponses dans un tableau ou une liste, il était ajouté au corpus au même titre que les autres documents.

Le corpus d'étude se compose actuellement de cent questions ayant été générées manuellement sur des thématiques variées (sport, santé, politique, culture, économie, informatique) et sous plusieurs types. Les informations sont présentées de la façon suivante : type de la question (nombre de questions dans le corpus) (nombre de questions necessitant un traitement inter-document pour répondre pertinemment) : exemple.
– factuelle (61) (11) : *Quand la France a-t-elle perdue son triple-A ?* ;
– liste (17) (2) : *Quels pays étaient candidats à l'organisation de la coupe du monde 2018 ?* ;
– complexe (10) (3) : *Comment a évolué la croissance française en 2011 ?* ;
– booléenne (8) (3) : *Pluton est-elle une planète ?* ;
– définition (4) (0) : *Qu'est-ce que la croissance ?*.
Pour ces questions, nous avons récupéré 232 fichiers au format HTML, chacun d'entre eux contenant donc au moins une réponse correcte. Au total, seules 19 questions nécessitent obligatoirement un traitement inter-document pour composer la réponse. Cette basse proportion s'explique notamment par le fait que quelques pages très pertinentes (Wikipédia notamment) contenaient effectivement toutes les réponses. Nous avons décidé de les garder car les réponses étaient réparties sur l'ensemble du document (souvent de très grande taille). De plus leur identification nous servira de baseline pour mesurer les résultats sans traitement inter-document.

4.2 Observations

L'observation des passages-réponses (voir tableau 5) a d'abord montré des problèmes récurrents des SQR pour lesquels il existe déjà une base de travaux s'y intéressant, à savoir la résolution d'anaphore, la réconciliation de référence, le type métaphorique de la formulation de réponse

[4] http://www.exalead.fr/search/
[5] http://www.bing.com/
[6] http://www.google.fr

(*Le triple A, c'est une ligne Maginot.*), le besoin de contexte, les faux candidats (*en France nous avons le quintuple A (amicale des amateurs d'andouillettes authentique*)), pour ne citer qu'eux.

Le recensement précis a montré plusieurs phénomènes dont les plus émergents sont :
– les réponses se trouvant dans des tableaux de données ce qui confirme le besoin de savoir les analyser ;
– la présence d'informations incertaines (par exemple, des rumeurs ou avec l'usage du conditionnel) ;
– les réponses sont réparties dans des chronologies narratives (document retraçant chronologiquement un thème) ;
– le recoupement d'informations réparties dans plusieurs documents.

Nombre OCC	PHÉNOMÈNE
82	critère variant
72	formulation de la réponse
53	ancre référentielle à chercher
46	faux candidats
21	rumeur
20	chronologie narrative
18	tableau de données
17	terminologie dans la question
13	indice d'expansion de requête
12	beaucoup de candidats-réponses du type attendu dans un court passage textuel
12	besoin de contexte
11	type métaphorique de la réponse
10	terminologie dans la réponse

TAB. 5 – Occurrences des phénomènes (non mutuellement exclusifs) recensés les plus fréquents.

Nous présentons ici les 3 phénomènes auxquels nous avons choisi de nous intéresser par la suite car ce sont les plus fréquents dans notre corpus.

Le phénomène le plus fréquent est la variation des réponses selon certains critères : un critère de précision (que nous appellerons **critère variant**) d'un élément permet de créer plusieurs réponses correctes. Ici, la note souveraine de la France dépend de l'agence de notation :

question : *Quelle est la note de la France sur les marchés financiers ?*
passage-réponse 1 : *L'agence de notation américaine Egan-Jones a abaissé aujourd'hui la note attribuée à la dette de la France à "A", cinq crans en dessous du "triple A" des trois grandes du secteur, Standard and Poor's, Moody's et Fitch.(—). La France avait perdu son "AAA" chez cette agence en juillet.*
passage-réponse 2 : *"Moody's a maintenu le triple A de la France, la meilleure note possible", annonçait le matin une dépêche AFP, aussitôt reprise par une partie de la presse française.*
passage-réponse 3 : *Peu après 16 heures, ce vendredi, une source gouvernementale a indiqué que l'agence de notation financière Standard & Poor's avait bel et bien décidé de dégrader la France en lui retirant sa note d'excellence triple A.*

Le problème de **la formulation de la réponse** est un aussi problème habituel des SQR : la réponse, par la synonymie ou la paraphrase, peut prendre plusieurs formes :

question : *Qui a incarné Batman ?*
passage-réponse 1 : *Après avoir usé **Michael Keaton**, **Val Kilmer** et **George Clooney** dans le rôle de Batman, les spéculations sur le prochain vengeur masqué de Gotham City se poursuivent.*
passage-réponse 2 : *Le réalisateur chinois Zhang Yimou a choisi pour son prochain film l'acteur britannique **Christian Bale**, qui a incarné Batman, pour jouer le rôle d'un prêtre héroïque durant le sac de Nankin par les troupes japonaises en 1937.*

L'ancre référentielle est le phénomène nécessitant un besoin de rattachement à une date précise. En effet, le temps est un critère variant et les réponses correctes ne le sont parfois que par rapport à un moment temporel précis. Par exemple dans les trois passages-réponses suivants, les réponses nécessitent de trouver la date absolue à partir des indices temporels relatifs (en gras) pour pouvoir être validées :

question : *Quand la France a-t-elle perdu son triple-A ?*
passage-réponse 1 : *L'agence de notation américaine Egan-Jones a abaissé aujourd'hui la note attribuée à la dette de la France à "A", cinq crans en dessous du "triple A" des trois grandes du secteur, Standard and Poor's, Moody's et Fitch.(—).La France avait perdu son "AAA" chez cette agence **en juillet**.*
passage-réponse 2 : *"Moody's a maintenu le triple A de la France, la meilleure note possible", annonçait **le matin** une dépêche AFP, aussitôt reprise par une partie de la presse française.*
passage-réponse 3 : ***Peu après 16 heures, ce vendredi**, une source gouvernementale a indiqué que l'agence de notation financière Standard & Poor's avait bel et bien décidé de dégrader la France en lui retirant sa note d'excellence triple A.*

5 Expérimentation dans un cadre classique

Nous avons soumis au SQR FIDJI (Moriceau et Tannier, 2010) les cent questions de notre corpus afin d'analyser son comportement prévu pour une campagne d'évaluation classique. L'étude des résultats nous a permis dans un premier temps de mieux catégoriser les questions-ARM afin d'en dresser une typologie et dans un deuxième temps de mieux cibler les difficultés à résoudre pour pouvoir y répondre dans le futur.

5.1 Typologie des questions-ARM

L'étude avait révelé que 47 des 61 questions factuelles se révélaient être potentiellement des questions-ARM. Nous avons donc étudié les phénomènes composant ces questions-ARM en plus des questions-listes afin d'être en mesure de les typer lors de l'analyse des questions (figure 1). La marque du pluriel sur le focus de la question indique explicitement une question-ARM tandis que certains indices (notion temporelle, granularité du pronom *qui* et des adverbes interrogatif *où* et *quand*) peuvent potentiellement indiquer une question-ARM mais seules les réponses permettront au final de trancher. Le critère variant le plus fréquent (53,55 %) est le critère temporel mais il peut être plus général : les questions étant souvent courtes, le sens prototypique des concepts est fréquemment utilisé. Ainsi, parmi les exemples suivants de questions illustrant les phénomènes de

la typologie en figure 1, la question (6) pour un français passionné de football fait communément référence à la ligue des champions de football masculine et européenne alors qu'aucun de ces deux termes n'est présent :

– (identifiant en figure 1) Pourcentage sur les 100 questions du corpus : *Question* explication sur les réponses"
– **(1), 10 %** : *Quels ministères a occupé Alliot-Marie ?* "La Défense, l'Intérieur..." ;
– **(2) 7 %** : *Quels sont les pays de l'UE ?* "France, Finlande, Allemagne..." (27 pays en 2012) ;
– **(3) 1 %** : *Quelles sont les neuf planètes du système solaire ?* "Mercure, Vénus, Terre..." ;
– **(4) 33 %** : *Où/Quand/À qui Sarkozy a-t-il présenté ses voeux 2012 ?* "À Lille le 12 janvier aux fonctionnaire, À Lyon le 19 janvier au monde économique..." ;
– **(5) 11 %** : *Qui sont les Disney Princess ?* "Tiana a été ajoutée en 2009 à la collection, Raiponce en 2010" ;
– **(6) 95,74 %** : *Qui a gagné la ligue des champions en 2011 ?* "Barcelone en UEFA homme, Lyon en UEFA femme, Espérance de Tunis en CAF homme"
– **(7) 4,26 %** : *Quelle superbe victoire a remporté la France en 1998 ?* "1-0 contre la Finlande le 5 juin", "3-0 en France contre le Brésil le 12 juillet..." (onze victoires en tout en 1998) ;
– **(8) 10 %** : *Quand démarra le troisième gouvernement Fillon ?* "le 13/11/10" (annonce par Fillon), "le 16/11/10" (publication au Journal officiel) ;
– **(9) 4 %** : *Quel jour Nicolas Sarkozy est-il devenu président de la République ?* "Élu le 6 mai 2007, investi le 16 mai" ;
– **(10) 1 %** : *Quand fut fêté le bicentenaire de la révolution française ?* $1789 + 200 = 1989$;
– **(11) 11 %** : *Quels JO se sont déroulés il y a 16 ans ?* ;

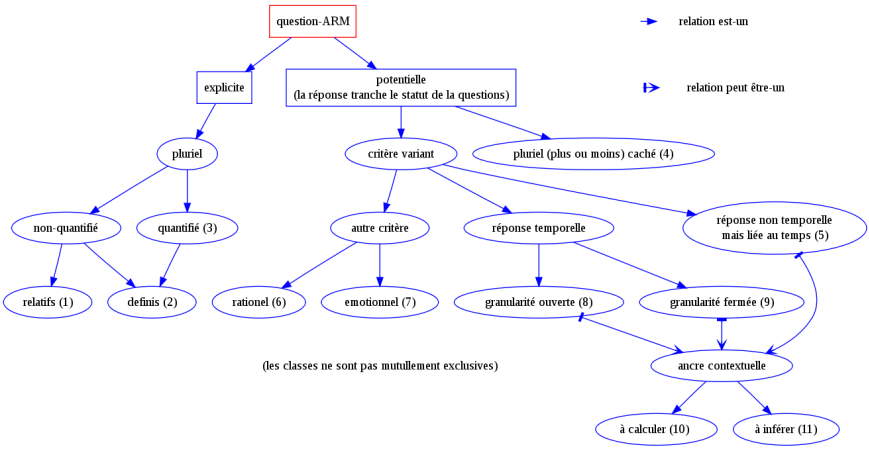


FIG. 1 – Typologie des questions-ARM. Les chiffres correspondent aux exemples précédents.

5.2 Approche classique avec FIDJI

Le SQR FIDJI permet de recouper les informations entre documents en se basant sur la syntaxe (Moriceau et Tannier, 2010) mais n'a pas encore de dispositif fonctionnel concernant les questions-ARM. Un traitement des question-liste existe cependant en recherchant dans un même document un groupe d'éléments consécutifs. Les résultats actuels vont nous servir de premier état des lieux pour implémenter le traitement des questions-ARM. Ainsi nous avons rencontré les phénomènes suivants :

- **(A)** FIDJI choisit de ne renvoyer qu'une seule réponse à fort score de confiance plutôt que plusieurs à faibles scores (même si la question est une question-liste) ;
- **(B)** FIDJI détecte la réponse-liste dans un document mais ne l'extrait pas correctement ;
- **(C)** FIDJI renvoie deux réponses-listes correctes sans les fusionner ;
- **(D)** FIDJI renvoie une réponse correcte ("2005") mais il existait une réponse correcte plus pertinente dans le passage ("octobre 2003") : **Quand est sorti l'Ibook G4 ?** *Avec les nouveaux iBook G4* ***2005****, Apple introduit Bluetooth2 de série (+ERD) (...). Le tableau ci-dessous retrace toute l'histoire de l'iBook G4 de sa sortie en* ***octobre 2003*** *à nos jours.*

Nous voyons donc des pistes concrètes d'améliorations puisque **(A)** est dû à un manque dans l'analyse des questions, **(B)** à une extraction à améliorer, **(C)** à un manque de recoupement entre les documents et **(D)** à une granularité de la pertinence de la réponse à renvoyer.

Le phénomène du critère variant est bien présent dans les résultats et nous montre l'intérêt à dépasser le cadre de la réponse unique à extraire d'un passage-candidat.

6 Conclusion et perspectives

En nous intéressant aux modes d'évaluation des SQR lors des campagnes pour le français, nous avons constaté un bridage nécessaire sur la présentation finale des réponses et relativement peu d'inter-documentalité pour les questions-listes. Après avoir constitué une collection de questions-ARM et de documents permettant d'y répondre, l'expérimentation avec un SQR rodé a confirmé la nécessité de mettre en place un traitement inter-document pour être en mesure de répondre le plus pertinemment possible à une question-ARM.

Nous allons donc implémenter un module de traitement des questions-ARM afin de dépasser le cadre habituel d'évaluation des SQR et se diriger vers un cadre utilisateur. En élargissant nos sources d'informations (HTML, éléments structuraux comme les tableaux), nous espérons bénéficier de plus d'informations pertinentes réparties dans des documents différents.

Un autre aspect intéressant est la présentation des réponses à l'utilisateur. Nous pensons proposer à l'utilisateur des réponses regroupées selon des critères variants s'ils existent, notamment à l'aide d'éléments structuraux (tableau par exemple). De plus, il serait intéressant d'ajouter aux réponses textuelles des données multimedia (URLs, images, etc.) qui permettront de justifier les réponses. L'évaluation des choix de regroupement serait alors faite du point de vue de l'utilisateur.

Plusieus approches applicatives se sont intéressées à la présentation des réponses à l'utilisateur. Par exemple, le SQR WolframQA[7] utilise également les images, les tableaux et les chronologies pour présenter plusieurs réponses à l'utilisateur. On retrouve les tableaux dans *Google squared*

[7] `http://www.wolframalpha.com`

(Crow, 2010) et des chronologies dans *Google News* et ChronoZoom[8] ainsi que dans les travaux de (Llorens *et al.*, 2011) qui s'intéresse à l'annotation temporelle de textes à des fins de visualisations ergonomiques pour l'utilisateur.

Références

AYACHE, C., GRAU, B. et VILNAT, A. (2006). Equer : the french evaluation campaign of question-answering systems. *In Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*.

AÏT-MOKHTAR, S., LUX, V. et BANIK, E. (2003). Linguistic parsing of lists in structured documents. *In Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003), Budapest, Hungary*.

BOS, J., GUZZETTI, E. et CURRAN, J. R. (2007a). The pronto qa system at trec 2007 : Harvesting hyponyms, using nominalisation patterns, and computing answer cardinality. *In TREC-16*.

BOS, J., GUZZETTI, E. et CURRAN, J. R. (2007b). The pronto qa system at trec 2007 : Harvesting hyponyms, using nominalisation patterns, and computing answer cardinality. *In TREC-16*.

BRAS, M., PRÉVOT, L. et VERGEZ-COURET, M. (2008). Quelle(s) relation(s) de discours pour les structures énumératives ? CMLF (Congrès mondial de linguistique française).

CHU-CARROLL, J., CZUBA, K., PRAGER, J. et BLAIR-GOLDENSOHN, S. (2004). Ibm's piquant ii in trec2004. *In TREC-13*.

CROW, D. (2010). Google squared : Web scale, open domain information extraction and presentation. *In ECIR*.

DAN I. MOLDOVAN AND, C. C. et BOWDEN, M. (2007). Lymba's poweranswer 4 in trec 2007. *In TREC-16*.

DANG, H. T., KELLY, D. et LIN, J. (2007). Overview of the trec 2007 question answering track. *In TREC-16*.

FANGTAO, L., XIAN, Z. et XIAOYAN, Z. (2008). Answer validation by information distance calculation. *In Coling 2008 : Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, IRQA '08, pages 42–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

FIGUEROA, A. et NEUMANN, G. (2008). Finding distinct answers in web snippets. *In In the 4th International Conference on Web Information Systems and Technologies*, pages 26–33. INSTICC Press.

GALA, N. (2003). *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires*. Thèse de doctorat, Université Paris-Sud.

GATTERBAUER, W., BOHUNSKY, P., HERZOG, M., KRÜPL, B. et POLLAK, B. (2007). Towards domain-independent information extraction from web tables. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 71–80. ACM.

GRAPPY, A. (2011). *Validation de réponse dans un système de question-réponse*. Thèse de doctorat.

HO-DAC, L.-M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat, Université Toulouse le Mirail.

[8] `http://research.microsoft.com/en-us/projects/chronozoom/`

HO-DAC, L.-M., PÉRY-WOODLEY, M.-P. et TANGUY, L. (2010). Anatomie des structures énumératives.

JACQUEMIN, C. et BUSH, C. (2000). Fouille du web pour la collecte d'entités nommées. *In TALN*.

KAISSER, M. et BECKER, T. (2004). Question answering by searching large corpora with linguistic methods. *In TREC-13*.

KATZ, B. et LIN, J. (2003). Selectively using relations to improve precision in question answering. *In EACL-2003 workshop on natural language processing for question answering*.

KATZ, B., MARTON, G., FELSHIN, S., LORETO, D., LU, B., MORA, F., Özlem UZUNER, MCGRAW-HERDEG, M., CHEUNG, N., RADUL, A., SHEN, Y. K., LUO, Y. et ZACCAK, G. (2006). Question answering experiments and resources. *In TREC-15*.

LAIGNELET, M. (2009). *Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques*. Thèse de doctorat.

LAIGNELET, M., KAMEL, M. et AUSSENAC-GILLES, N. (2011). Enrichir la notion de patron par la prise en compte de la structure textuelle - application à la construction d'ontologie. *In TALN*.

LLORENS, H., SAQUETE, E., NAVARRO, B. et GAIZAUSKAS, R. (2011). Time-surfer : time-based graphical access to document content. *In ECIR'11 : Proceedings of the 33rd European conference on Advances in information retrieval*, pages 767–771, Berlin, Heidelberg. Springer-Verlag.

LUC, C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *In TALN*.

MOLDOVAN, D. I., CLARK, C. et BOWDEN, M. (2007). Lymba's poweranswer 4 in trec 2007. *In TREC-16*.

MORICEAU, V. et TANNIER, X. (2010). Fidji : Using syntax for validating answers in multiple documents. *Information Retrieval, Special Issue on Focused Information Retrieval*, (10791).

PÉRY-WOODLEY, M.-P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. HDR.

QUINTARD, L., GALIBERT, O., ADDA, G., GRAU, B., LAURENT, D., MORICEAU, V., ROSSET, S., TANNIER, X. et VILNAT, A. (2010). Question answering on web data : the qa evaluation in quæro. *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta*.

RAZMARA, M. et KOSSEIM, L. (2008). Answering list questions using co-occurrence and clustering. *In LREC*. European Language Resources Association.

SCHLAEFER, N., KO, J., BETTERIDGE, J., SAUTTER, G., PATHAK, M. et NYBERG, E. (2007). Semantic extensions of the ephyra qa system for trec 2007. *In TREC-16*.

TAJIMA, K. et OHNISHI, K. (2008). Browsing large html tables on small screens. *In UIST*, pages 259–268.

WANG, R. C., SCHLAEFER, N., COHEN, W. W. et NYBERG, E. (2008). Automatic set expansion for list question answering. *In EMNLP*.

WANG, Y. et HU, J. (2002). A machine learning based approach for table detection on the web. *In Proceedings of the 11th international conference on World Wide Web*, pages 242–250. ACM.

WU, M., ZHENG, X., DUAN, M., LIU, T. et STRZALKOWSKI, T. (2003). Questioning answering by pattern matching, web-proofing, semantic form proofing. *In TREC-12*, pages 578–585.

Extraction de PCFG et analyse de phrases pré-typées

Noémie-Fleur Sandillon-Rezer
CNRS, Esplanade des Arts et Métiers, 33402 Talence
LaBRI, 351 Cours de la Libération, 33405 Talence
nfsr@labri.fr

RÉSUMÉ

Cet article explique la chaîne de traitement suivie pour extraire une grammaire PCFG à partir du corpus de Paris VII. Dans un premier temps cela nécessite de transformer les arbres syntaxiques du corpus en arbres de dérivation d'une grammaire AB, ce que nous effectuons en utilisant un transducteur d'arbres généralisé ; il faut ensuite extraire de ces arbres une PCFG. Le transducteur d'arbres généralisé est une variation des transducteurs d'arbres classiques et c'est l'extraction de la grammaire à partir des arbres de dérivation qui donnera l'aspect probabiliste à la grammaire. La PCFG extraite est utilisée via l'algorithme CYK pour l'analyse de phrases.

ABSTRACT

PCFG Extraction and Pre-typed Sentences Analysis

This article explains the way we extract a PCFG from the Paris VII treebank. Firslty, we need to transform the syntactic trees of the corpus into derivation trees. The transformation is done with a generalized tree transducer, a variation of the usual top-down tree transducers, and gives as result some derivation trees for an AB grammar. Secondely, we have to extract a PCFG from the derivation trees. For this, we assume that the derivation trees are representative of the grammar. The extracted grammar is used, via the CYK algorithm, for sentence analysis.

MOTS-CLÉS : Extraction de grammaire, grammaire de Lambek, PCFG, transducteur d'arbre, algorithme CYK.

KEYWORDS: Grammar Extraction, Lambek grammar, PCFG, tree transducer, CYK Algorithm.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 205–218,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cet article décrit les méthodes que nous employons pour transformer les arbres syntaxiques du corpus de Paris VII en arbres de dérivation d'une grammaire AB (Lambek, 1958), pour l'extraction de cette grammaire et son utilisation pour l'analyse de phrases. Les grammaires AB sont utilisées dans des algorithmes d'apprentissage tels que celui de Buskowsky et Penn (Buszkowski et Penn, 1990), qui permet d'apprendre une grammaire AB rigide [1], ou celui de Kanazawa (Kanazawa, 1998), permettant d'apprendre une grammaire k-valuée [2] ; c'est pour cela que nous avons souhaité, dans un premier temps, utiliser de telles grammaires. Les grammaires AB représentent un fragment des grammaires de Lambek, comprenant uniquement des règles de dérivation de type $a \rightarrow a/b\ b$ et $a \rightarrow b\ b\backslash a$. Le corpus de Paris VII (Abeillé *et al.*, 2003) est composé de 12855 phrases tirées du journal *Le Monde*, annotées et analysées par le laboratoire de Paris VII. Les arbres syntaxiques sont planaires, le nombre de fils par noeud et la profondeur ne sont pas fixés. Cela rend l'application d'algorithmes d'apprentissage usuels impossible ; nous avons donc pris le parti d'utiliser un transducteur d'arbres.

Nous avons utilisé, pour notre travail, une sous-partie du corpus, présentée sous forme parenthésée de 12351 phrases, alors que le corpus complet est au format **XML**. Les 504 phrases laissées de côté forment un corpus annexe dont nous nous servons pour l'évaluation.

En premier lieu, nous présenterons le transducteur, utilisé pour transformer les arbres syntaxiques en arbres de dérivation, puis nous nous pencherons sur l'extraction d'une grammaire PCFG. La troisième partie détaillera l'analyse du placement des syntagmes prépositionnels dans la phrase ; tandis que la quatrième présentera les résultats expérimentaux obtenus en utilisant notre grammaire PCFG pour trouver la meilleure analyse possible pour une phrase via l'algorithme CYK (Younger, 1967).

2 Transducteur d'arbres généralisé

Ce n'est pas la première fois que les transducteurs sont utilisés dans le cadre de la linguistique computationnelle ; on peut citer Knight et Graehl (Knight et Graehl, 2005), qui utilisent des transducteurs d'arbres à états finis, dont hélas l'utilisation ne correspondait pas à notre problématique.

Des travaux de recherche plus appliqués, tels que (Hockenmaier et Steedman, 2007; Moot, 2010a,b; Moortgat et Moot, 2001), utilisent des algorithmes spécialisés qui s'appliquent uniquement à un corpus donné, avec un espoir faible de réutilisation. Etant donné les différences d'annotations d'un corpus à l'autre, et les variations grammaticales que l'on peut trouver entre deux langues, adapter un outil pour le corpus de Paris VII est toujours particulièrement laborieux. Etant donné que nous avons totalement séparé, lors de l'implémentation, le fonctionnement du transducteur de l'ensemble des données qui lui sont passées en entrée (telles que les fichiers de règles et le corpus sous forme parenthésée), nous pensons qu'un lissage des données suffit à appliquer notre transducteur à d'autres ensembles d'arbres.

Le transducteur que nous avons créé est le pivot central du processus d'extraction de grammaire.

1. Chaque mot du lexique n'a le droit d'avoir qu'un seul type.
2. Chaque mot du lexique peut avoir jusqu'à k types.

En effet, c'est la binarisation des arbres syntaxiques, fondée sur les règles usuelles de dérivation d'une grammaire AB [3] et les annotations morpho-syntaxiques du corpus (Abeillé et Clément, 2003), qui paramétrise la grammaire extraite.

Nous avons d'abord mis au point une version théorique de notre G-transducteur (G pour généralisé) avant de l'implémenter pour le tester sur le corpus de Paris VII.

La création du transducteur d'arbres est décrite en détail dans (Sandillon-Rezer et Moot, 2011).

En nous fondant sur les transducteurs d'arbres *top-down* décrits dans TATA (Comon *et al.*, 2007), nous avons généralisé les règles de transduction de manière à créer un outil plus adapté au corpus de Paris VII. Ainsi, on peut dire que les trois principales différences entre un transducteur *top-down* classique et notre G-transducteur sont : sa récursivité, sa paramétrisation et son système de règles de priorité.

La récursivité permet d'appliquer un ensemble de règles à un nœud, jusqu'à ce qu'il soit traité en entier, sans pour autant changer l'état du transducteur ni utiliser un nouvel ensemble de règles de transduction (voir figure 1).

La paramétrisation permet de définir des règles avec variables. Ainsi, on peut donner une transduction générale pour les adverbes et les modificateurs (voir figure 2).

Les règles de priorité : assurent le déterminisme de notre transducteur. Ainsi, lorsque deux règles peuvent s'appliquer, on leur donne un ordre d'application qui permet d'avoir toujours les mêmes arbres de sortie (voir figure 3).

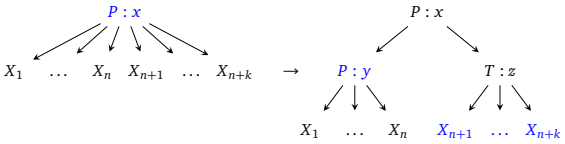


FIGURE 1 – Le nouveau nœud $P : y$ a moins de fils qu'avant la transduction et pour le transducteur, il restera dans le même état et avec le même label que le nœud parent $P : x$. Généralement, la règle sera écrite de manière à ce que le sous arbre $T : z$ soit binaire et les types y et z doivent obligatoirement se combiner pour donner le type x.

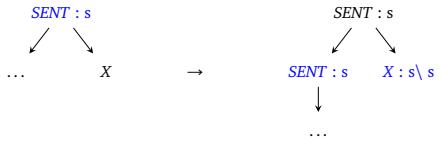


FIGURE 2 – La même règle sera appliquée pour $X \in \{ADV, PP-MOD, AdP-MOD, AP-MOD, ...\}$

3. les groupes nominaux auront le type *np* etc.

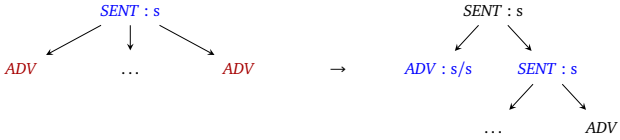


FIGURE 3 – Lorsque plus d'une règle peut s'appliquer à un arbre, le fait de suivre un ordre prédéfini permet d'éviter le non-déterminisme du transducteur.

Les règles ont été déduites d'une analyse systématique des formes présentes dans le corpus. Un exemple de règle est donné dans la figure 4 et un de résultat dans la figure 5. Une fois le corpus transformé en forêt d'arbres de dérivations, nous n'utilisons plus le transducteur, que ce soit pour l'extraction de grammaire ou l'analyse de phrases.

```
(rule
      (SENT:* NP-SUJ (VN tree VPP) PP-OBJ)
      ("SENT:*" "NP-SUJ:np"
            ("VN:np\\*" "VN:(np\\*)/(np\\s_p)"
                  (":np\\s_p" "VPP:(np\\s_)/pp"
            PP-OBJ:pp))))
```

FIGURE 4 – Exemple de règle telle que donnée au transducteur. On note deux points importants, directement dérivés des spécifications de notre G-transducteur : le mot clef *tree*, qui permet de remplacer "un certain nombre de nœuds", qui peut apparaître plusieurs fois dans le motif de départ mais pas dans le motif de remplacement ; et le type *, qui remplace n'importe quel type hérité des étapes précédentes.

3 Extraction de grammaire

Même si le lexique, récupéré à partir des feuilles des arbres de dérivation, suffirait à représenter la grammaire AB, il nous limite aux mots présents dans le corpus. Or, bien que nous puissions avoir des probabilités sur les types des mots, nous voulions une grammaire PCFG. Par conséquent, nous avons pris le parti d'extraire une grammaire probabiliste à partir des arbres.

Les arbres en sortie du transducteur donnent des informations à la fois syntaxiques, car nous gardons les labels donnés par le corpus et, bien sûr, des informations structurelles. Nous avons pris le parti de laisser le choix des informations que nous souhaitons garder, en effectuant une passe de prétraitement, sachant bien sûr que les types sont, de toute façon, obligatoirement conservés. La grammaire extraite sera de toute façon une grammaire hors contexte, avec une probabilité calculée sur les règles en fonction de leur racine. Pour plus de simplicité, on rappelle que les grammaires sont de la forme $\{N, T, S, R\}$:

N l'ensemble des symboles non terminaux, correspondant aux nœuds internes de l'arbre.

T l'ensemble des symboles terminaux, correspondant à l'ensemble des mots typés.

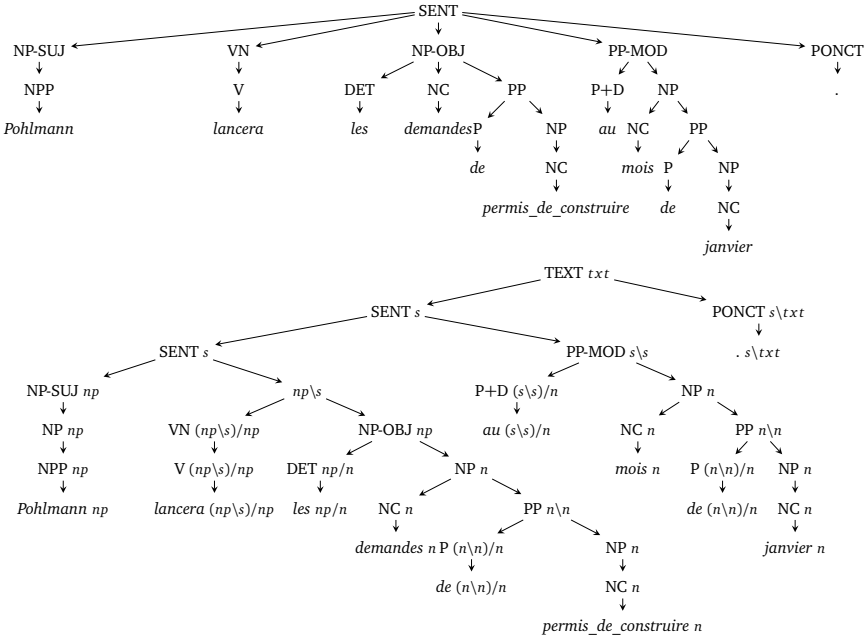


FIGURE 5 – Arbre d'entrée et de sortie du transducteur correspondant à la phrase "Pholmann lancera les demandes de permis de construire au mois de janvier.".

S le symbole initial. On choisira, en fonction de la passe de pré-traitement, *TXT :txt* ou *txt*.

R l'ensemble des règles.

L'algorithme utilisé pour extraire la grammaire consiste à parcourir les arbres donnés en paramètre et stocker les règles de dérivation que l'on rencontre. On considère qu'une règle de dérivation est constituée d'une racine et d'un ou deux fils :

La racine a deux fils : On est dans le cas de figure classique d'une règle d'élimination à droite ou à gauche ($a \rightarrow a/b \quad b$ ou $a \rightarrow b \quad a\backslash b$).

La racine a un seul fils : Il y a simplement transmission de type au fils. Ce cas de figure apparaît, par exemple, lorsqu'un groupe nominal est composé uniquement d'un nom propre, ou encore lorsqu'on est au niveau du nœud pré-terminal, c'est à dire l'étiquette de partie du discours (POS-tag) de la feuille. Dans ce cas, la feuille héritera directement du type du POS-tag.

Chaque règle est accompagnée d'un compteur et les probabilités sur les règles sont calculées par groupe ayant la même racine. On récupère aussi des informations de profondeur minimale et maximale d'apparition de la règle, cependant elles ne sont pas utilisées pour l'instant.

Ainsi, on résume dans le tableau 1 les différentes grammaires que peut générer l'extracteur [4] .Chacune des versions montre un intérêt : autant la première, extraite des arbres juste après transduction, garde les informations syntaxiques données par le corpus ; autant les suivantes sont plus utiles pour appliquer un algorithme d'analyse de phrases, tel que CYK (voir section 4), sur des phrases non typées. Le tableau 2 montre des extraits des différentes grammaires en fonction des arbres donnés en entrée.

Forme des arbres	**Règles extraites**	**Spécifications**	**Nombre de règles**
Arbres de dérivation bruts	$n_1 \rightarrow n_2 \quad n_3$ $n_1 \rightarrow n_2$ $n_1 \rightarrow t_1$	Facilement normalisable en FNC : il suffit d'enlever les chaines unaires	63368
Retrait des chaines unaires et des labels sauf les POS-tag	$n_1 \rightarrow n_2 \quad n_3$ $n_1 \rightarrow t_1$	La grammaire est en FNC.	59505
Retrait de tous les labels et des chaines unaires. Il n'y a plus de différence entre N et T.	$n_1 \rightarrow n_2 \quad n_3$	Les mots n'apparaissent plus, ce qui laisse uniquement le squelette des arbres.	3494

TABLE 1 – Grammaires extraites en fonction des arbres de dérivation donnés en entrée. On précise que $n_i \in N$ et $t_i \in T$.

Arbres de dérivation bruts		
Exemple de règles	$NP:np \rightarrow NPP:np$	$1.01x10^{-1}$
	$NP:np \rightarrow DET:np/n\ NC:n$	$2.02x10^{-1}$
	...	
Retrait des chaines unaires et des labels sauf les POS-tag		
Exemple de règles	$(np\backslash s_i)/(np\backslash s_p) \rightarrow VINF:(np\backslash s_i)/(np\backslash s_p)$	$9.53x10^{-1}$
	$(np\backslash s)/(np\backslash s_p) \rightarrow CLR:cl_r \quad cl_r\backslash((np\backslash s)/(np\backslash s_p))$	$2.88x10^{-2}$
	...	
Retrait de tous les labels et des chaines unaires		
Exemple de règles	$s \rightarrow np\ np\backslash s$	$3,81x10^{-1}$
	$s \rightarrow s\ s\backslash s$	$2,65x10^{-1}$
	$s \rightarrow np\backslash s_p\ (np\backslash s_p)\backslash s$	$1,13x10^{-3}$
	$n \rightarrow n\ n\backslash n$	$7,97x10^{-1}$
	$np \rightarrow np/n\ n$	$8,02x10^{-1}$
	...	

TABLE 2 – Exemples des différentes règles que l'on peut extraire des arbres.

4 Analyse de phrases

La question de l'analyse de phrases en fonction d'une grammaire PCFG se subdivise en deux problèmes. En effet, il faut d'une part trouver les types des mots et d'autre part que les règles

4. Les *POS-tags* sont les étiquettes de parties du discours.

existent dans la grammaire passée en paramètre à l'analyseur.

4.1 Typages des mots

En réunissant les feuilles des arbres de dérivation, nous pouvons collecter un lexique contenant les mots, leur occurrence, les types de ceux-ci et la probabilité du type (*nb_occurrences_du_type*/ *nb_occurrences_du_mot*). Cependant, nous n'utilisons pas encore le lexique pour typer les phrases que nous analysons. Il faudrait pourtant sélectionner les types apparaissant le plus souvent et les lier aux mots. Cette technique n'assurerait pas l'analyse systématique de la phrase, car si le type nécessaire fait partie de ceux écartés, une phrase juste pourrait ne pas avoir d'analyse. Nous avons pris le parti de typer les mots soit en utilisant le Supertagger ((Moot, 2010a,b)), soit en utilisant les phrases typées à la sortie du transducteur. La première méthode nous permet à la fois de valider les types donnés aux mots par le Supertagger et d'analyser des phrases dont les mots n'apparaissent pas dans le corpus de Paris VII, tandis que la seconde méthode nous permet de tester nos différentes grammaires en fonction des arbres de dérivation. On peut aussi utiliser un typage plus manuel, qui utilise le Supertagger pour effectuer une première passe de typage et qui permet ensuite à l'utilisateur de modifier à loisir les types proposés.

4.2 Analyse des phrases typées

Pour l'algorithme de reconstruction des phrases, nous avons décidé d'utiliser l'algorithme CYK (Younger, 1967; Knuth, 1997; Hopcroft et Ullman, 1979) et d'en implémenter une version probabiliste : en effet, étant donné que cet algorithme a déjà été testé et est une référence, il nous a permis de tester l'efficacité de notre grammaire sans avoir à s'inquiéter de l'efficacité de l'algorithme. D'autres algorithmes auraient pu être utilisés, tel que celui d'Earley (Earley, 1973), cependant CYK demandait en entrée une grammaire très proche de celle que nous obtenions après extraction. De plus, l'ajout de l'aspect probabiliste était trivial sur cet algorithme. La seule modification que nous avons effectuée était de retirer la phase de typage des mots, initialement effectuée par CYK grâce aux règles de type $n_1 \rightarrow t_t$. Nous avons donc pu donc utiliser la grammaire la plus simple, de 3494 règles, pour analyser les phrases. Le premier test effectué, pour savoir si l'algorithme fonctionnait correctement, a été d'analyser les phrases extraites des arbres de dérivation avec les règles provenant de ces mêmes arbres. Nous avons ensuite pu tester l'analyse avec des phrases typées par le Supertagger ou notre transducteur et des grammaires extraites soit du corpus de 12351 phrases, soit du corpus de phrases laissées de côté (cf section 6).

Les arbres de dérivation correspondants aux phrases "Pourtant tout n'est pas gagné." et "Ce procès gagné donne au Crédit Lyonnais les coudées franches pour gérer MGM" sont montrés dans la figure 6 et 7. A chaque fois, on a pris les deux arbres les plus probables, typés par le Supertagger et les phrases ont été analysées avec la même grammaire et l'algorithme CYK. Deux informations sont intéressantes pour choisir quel est le meilleur arbre de dérivation sur les phrases : on regarde à la fois la complexité des types et la probabilité. Cependant, nous sommes conscient qu'il est complexe de comparer deux arbres qui n'ont ni la même structure, ni les mêmes feuilles. La préférence que l'on porte à un résultat sera fortement dépendante des critères de sélection donnés. Ainsi, sur la figure 6, on remarque que les deux arbres ont la même probabilité, cependant nous sélectionnons celui qui a l'indexation la plus faible pendant

l'exécution de CYK. Sur la seconde phrase (figure 7), c'est majoritairement l'attachement du groupe prépositionnel final qui modifie la forme de l'arbre. L'attachement de la préposition à un groupe nominal est plus représentatif du corpus d'origine (voir section 5) .

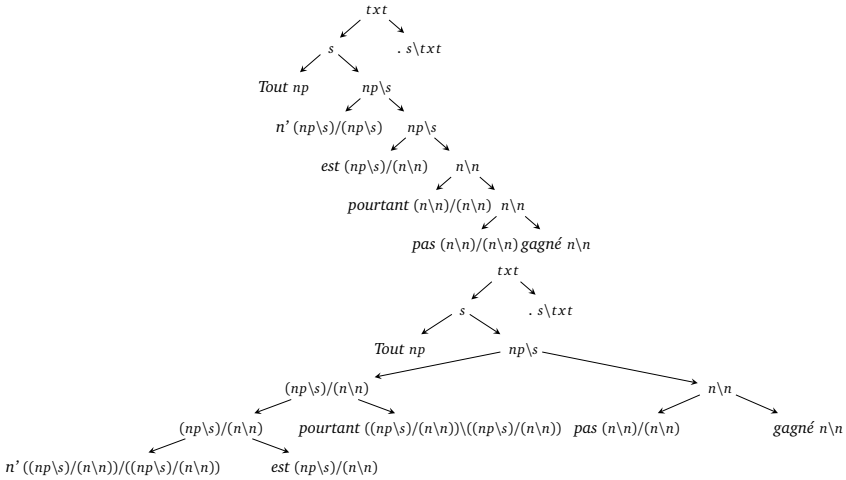


FIGURE 6 – Le premier arbre est généré avec le typage du transducteur et a une probabilité de $9,6x10^{-05}$ et le second est typé avec le Supertagger, avec une probabilité de $1,9x10^{-05}$

5 Analyse des prépositions

Nous allons nous focaliser sur l'analyse des syntagmes prépositionnels (*PP*, *PP-MOD*, *PP-OBJ* etc.) et de l'attachement par rapport à la phrase. Dans un premier temps, nous étudierons l'attachement des groupes prépositionnels dans le corpus d'origine, puis nous nous focaliserons sur les types des prépositions, via le transducteur et le Supertagger pour enfin nous pencher sur l'attachement dans les arbres de dérivation générés via l'algorithme CYK.

5.1 Attachement dans le corpus

Les groupes prépositionnels sont particulièrement nombreux dans le corpus (49039 occurrences). Comme nous pouvons le voir dans le tableau 3, ils sont majoritairement étiquetés *PP*. Leur attachement de départ dans le corpus est aussi particulièrement important, car c'est celui-ci qui définira le type de la préposition. Le tableau 4 résume la répartition des syntagmes prépositionnels dans le corpus, en fonction de leur parent. En effet, la transduction aura tendance à donner un type aux syntagmes prépositionnels qui correspond à leur place dans la structure de la phrase.

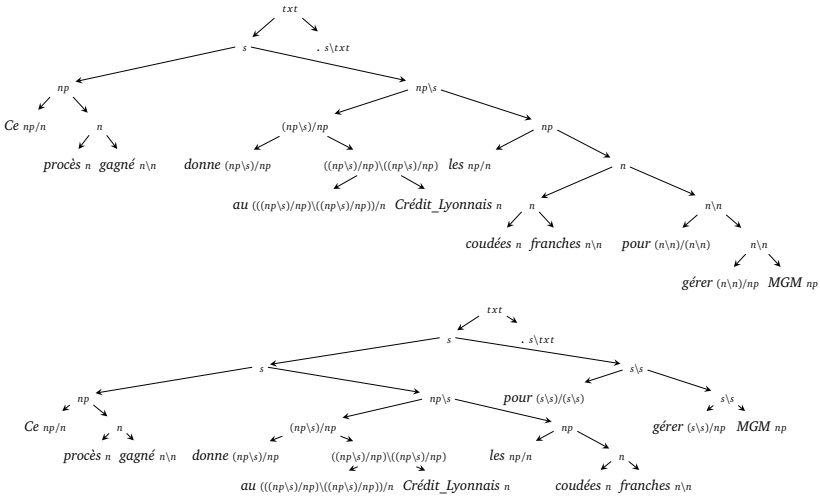


FIGURE 7 – Probabilité du premier arbre : $2,2x10^{-09}$. Probabilité du second arbre : $1,5x10^{-09}$.

Ainsi, dans un groupe nominal, le *PP* aura plus souvent le type $n\backslash n$, alors qu'au milieu d'une phrase le typage sera plus complexe. Lors de la transduction, on ne change pas l'ordre des mots, mais quelques fois leur attachement au sein de la structure. Cependant, on peut dire que les groupes prépositionnels ne bougent pas, sauf s'ils sont à l'extérieur d'un noyau verbal et que celui-ci se termine par un *VPP*, auquel cas on lie plus spécifiquement le participe passé au groupe prépositionnel, comme on peut voir figure 4.

Label	**occurrence**	**Label**	**occurrence**	**Label**	**occurrence**
PP	32023	*PP-MOD*	11899	*PP-DE_OBJ*	1668
PP-A_OBJ	1565	*PP-P_OBJ*	1389	*PP-ATS*	323
PP-OBJ	130	*PP-ATO*	30	*PP-SUJ*	12

TABLE 3 – Distribution des groupes prépositionnels en fonction de leur label.

5.2 Typage des syntagmes prépositionnels

Pour étudier le typage, nous nous sommes focalisés sur les groupes prépositionnels dont, bien sûr, la transduction avait réussi. Cela fait tomber le nombre de syntagmes prépositionnels à 45351 (92,5% du total). Les quatre familles de types les plus donnés (au dessus de 2000 fois) par le transducteur sont résumés dans le tableau 5. Ils couvrent 92,2% des types que l'on peut trouver pour des prépositions. Les types restants, marginaux, correspondent, par exemple, à un syntagme prépositionnel contenant uniquement un pronom relatif, qui prend en argument une

Syntagme parent	occurrence	Label le plus courant	pourcentage
Syntagme Nominal	24817	*PP*	99,2%
Phrase complète	8478	*PP-MOD*	72,2%
Proposition rel. ou sub.	3833	*PP-MOD*	57,9%
Proposition participiale	3552	*PP*	80,1%
Proposition infinitive	3190	*PP-MOD*	63,5%
Syntagme prépositionnel	843	*PP*	89,2%
Noyau verbal	45	*PP*	88,9%

TABLE 4 – Distribution des groupes prépositionnels en fonction de leurs parents. On remarque que les groupes nominaux sont ceux qui regroupent le plus de *PP* , c'est à dire presque la moitié.

subordonnée.

Le typage effectué avant l'analyse via CYK, avec le Supertagger, nous permet de régler la précision que l'on souhaite sur les types : en effet, on peut régler le paramètre β, qui déterminera le nombre de types possibles autorisés par mot. On gardera alors les types ayant une probabilité supérieure ou égale à β fois la plus grande probabilité trouvée [5]. Le tableau 6 résume la justesse des types donnés aux prépositions en fonction de β. On remarque que ce sont des mots difficiles à typer, étant donné que les résultats sont inférieurs aux résultats globaux, bien que les adverbes et les verbes soient encore plus complexes à typer de manière exacte.

Il faut cependant noter qu'il n'est pas nécessaire d'avoir une formule correcte pour que l'attachement du syntagme prépositionnel dans la phrase soit correct.

Famille de type	occurrence
n\n ou np\np ou n\np	23901
a\a (ex. s\s)	8548
pp ou pp_a ou $pp_d e$	6486
a/a (ex. s/s)	2882

TABLE 5 – Les quatre familles de types les plus courants correspondent à un modificateur de groupe nominal, un groupe prépositionnel généralement argument d'un groupe verbal et des modificateurs de phrase, placés au début ou à la fin de la phrase.

β	pertinence des types	pertinence globale
1.0	61,0%	76,9%
0.1	83,1%	87,0%
0.05	86,2%	88,9%
0.01	90,2%	91,7%

TABLE 6 – Justesse du typage via le Supertagger.

5. Plus β est petit, plus il y a de types proposés et plus on a de chance de trouver le type qui se combinera avec ceux des autres mots.

5.3 Attachement des syntagmes prépositionnels dans les arbres reconstitués

Pour cette partie, nous nous sommes focalisés sur 55 *PP*, que nous avons sélectionnés dans le corpus d'origine, de manière à respecter le ratio présenté dans le tableau 3. Cela correspond à 21 phrases, dont l'analyse a réussi. Nous avons généré les types possibles avec $\beta = 0.05$. Ensuite, nous avons étudié la différence de types donnés aux prépositions ainsi que leur attachement. On remarque, dans le tableau 7, que les syntagmes prépositionnels liés aux groupes nominaux sont attachés sensiblement au même endroit. On note une différence faible entre les groupes prépositionnels qui seront arguments d'un verbe, un peu plus importante entre les modificateurs globaux qui agissent sur toute la phrase. Il y a 4 cas, dans les arbres régénérés via CYK, où l'algorithme a jugé plus pertinent de préférer le type *n* ou *np* pour le syntagme prépositionnel ("On ne porte pas impunément atteinte *à des tabous*."), alors qu'on s'attend plutôt à une analyse qui lierait "atteinte" et "à" et qui prendrait en argument le groupe nominal "des tabous" [6].

On peut dire que le typage et l'attachement des syntagmes prépositionnels semblent cohérents avec l'attachement présent dans le corpus d'origine, ainsi que le typage effectué par le transducteur. Cependant, pour pouvoir l'affirmer, il faudrait faire des tests plus poussés, qui prendraient en compte la totalité du corpus.

Type	**occurrence après transduction**	**occurrence après CYK**
pp, pp_{de} ou pp_{a}	6	3
Modificateur de *NP*	35	37
Modificateur de *SENT*	9	4
np	0	4
Modificateur autre	5	7

TABLE 7 – Typage des prépositions dans le cadre d'une transduction comparées à celui effectué via le Supertagger avant reconstitution des arbres de dérivation avec CYK. Les modificateurs autres sont des modificateurs de proposition infinitive ou de syntagme adjectivaux.

Le typage, cependant, n'est pas entièrement lié à l'attachement dans la phrase. Nous avons comparé l'attachement des syntagmes prépositionnels et nous pouvons dire que, sur les 55 cas, il y en a 37 placés de manière identique et 18 non, soit 67,3% de ressemblance. Les différences majeures sont au niveau des prépositions qui sont plus souvent attachées aux groupes nominaux et argument des noyaux verbaux (ceux-ci peuvent alors prendre le type *np* plutôt que *pp*).

6 Évaluation et résultats

L'évaluation des différentes méthodes a été effectuée avec différents ensembles de données. Pour tester la totalité de nos travaux, nous avons utilisé le corpus de Paris VII dans son intégralité, c'est à dire :

– Les 12351 phrases parenthésées que nous avons étudiées en profondeur pour fonder l'ensemble de règles de notre transducteur, que nous appellerons *corpus principal*.

6. L'analyse CYK fait ressortir l'aspect idiomatique de "porter atteinte à".

– Les 504 phrases qui n'existaient pas sous forme parenthésée. Ce *corpus annexe* a été adapté pour être sous forme parenthésée et pour que les étiquettes soient celles utilisées par les règles.

Quel que soit le corpus utilisé, on parlera d'un *corpus partiel* pour dénoter le fragment dont la transduction a réussi. Les grammaires extraites des arbres de dérivation, donc des corpus partiels, auront le même nom que le corpus dont elles sont extraites, soit grammaire principale et grammaire annexe. L'évaluation du transducteur et de l'analyseur de phrases se mesure en pourcentage de phrases sur lesquelles l'opération a réussi. Dans le cadre du transducteur, cette notion correspond à la transformation des arbres syntaxiques en arbres de dérivation et dans le cadre de l'analyseur, elle correspond à la réussite de la combinaison des types donnés aux mots par le Supertagger.

6.1 Transducteur

Le transducteur transforme pour l'instant, avec 1671 règles, 92,6% du corpus principal (soit 11447 phrases) et 87,3% du corpus annexe (404 phrases) en arbres de dérivation d'une grammaire AB. On peut résumer l'utilisation des règles, dans le cadre de la transduction du corpus principal, dans le tableau 8. On remarque que, bien qu'il y ait de nombreuses règles qui sont utilisées peu de fois, elles ont un poids faible sur la totalité des transductions effectuées. Les règles les plus importantes sont exprimées dans le tableau 9, sous forme parenthésée telle qu'utilisée dans la syntaxe de *Tregex* (Levy et Andrew, 2006). La dernière règle, gérant la ponctuation finale, n'est pas utilisée autant de fois qu'il y a d'arbres de dérivation. Cela vient du fait que nous avons souhaité traiter différemment les phrases comprenant uniquement un groupe nominal et que certaines phrases, tels les titres d'articles, n'ont pas de ponctuation finale. De même, la règle qui s'occupe du déterminant au début d'un nom commun devrait être employée plus que ça, vu le nombre de groupes nominaux du corpus. Cependant, une règle prioritaire s'occupe du cas où le groupe nominal est composé d'un déterminant et d'un nom commun et est appelée 8892 fois.

Nombre de règles	Occurrence minimale et maximale	nombre d'applications
1148	entre 1 et 20	005818
303	entre 21 et 100	014174
170	entre 101 et 1000	054266
41	entre 1001 et 10000	**125405**
4	supérieur à 10000	060779

TABLE 8 – Récapitulatif de l'utilisation des règles.

motif de départ	motif d'arrivée	nombre d'applications
(*NP :* NC PP*)	(*NP :* NC :n PP :n*)*	17767
(*NP :* DET tree*)	(*NP :* DET :np/n NP :n*)	16232
(*PP :* P NP*)	(*PP :* P :*/np NP :np*)	16037
(*SENT tree PONCT*)	(TEXT :txt *SENT :s PONCT :s\ txt*)	10819
(*NP :* tree* (*COORD CC NP*))	(*NP :* (:* NP :**	
	(*COORD :** CC :(**)/np NP :np*)))	2511
(*SENT :* NP-SUJ VN NP-OBJ*)	(*SENT :* NP-SUJ :np*	
	(*:np* VN :(np*)/np NP-OBJ :np*))	1820

TABLE 9 – Quelques règles du transducteur, dont les quatre règles les plus utilisées.

Les arbres de dérivation des deux corpus nous permettent d'extraire deux grammaires, sur lesquelles les tests d'analyse que nous avons effectués seront détaillés dans la partie suivante 6.2. En addition des grammaires, nous pouvons créer un lexique, contenant les mots et les différents types qui leur sont associés en fonction des transductions. Le lexique correspondant au corpus principal contient 26765 mots sur les 27589 présents, il couvre donc 96,9% du vocabulaire présent dans le corpus de Paris VII.

6.2 Analyse de phrases

Nous avons effectué de nombreux tests avec notre analyseur de phrases. En effet, nous avons utilisé les deux grammaires différentes et nous avions à disposition des phrases typées par le transducteur ou par le Supertagger, avec $\beta = 0.01$.

Grâce au Supertagger, nous avons pu analyser aussi bien les phrases venant du transducteur que les phrases laissées de côté. Les résultats sont regroupés dans la table 10. On remarque que les résultats sont proportionnellement moins bons, mais que certaines phrases venant de la partie non traitée des différents corpus sont analysées et transformées en arbre de dérivation.

Origine des phrases	**Phrases analysées**	**Grammaire utilisée**	**Résultat**
Transducteur	Corpus principal partiel	Grammaire annexe	54,4%
	Corpus annexe partiel	Grammaire principale	85%
Supertagger	Corpus principal	grammaire principale	89,38%
	Corpus annexe	grammaire principale	83,1%

TABLE 10 – Tableau de résultat.

7 Conclusion et perspectives

Dans cet article, nous avons rapidement rappelé le principe du G-transducteur dont nous nous servons pour transformer les arbres syntaxiques du corpus de Paris VII en arbres de dérivation d'une grammaire AB, puis expliqué la méthode que nous employons pour extraire une PCFG de ces arbres. Les résultats expérimentaux d'analyse de phrase via l'algorithme CYK, en utilisant notre PCFG et des phrases typées au préalable, nous permettent de comparer les annotations produites par le transducteur et la méthode semi-automatique mise en place par Moot.

Cependant, ce travail est loin d'être terminé et nous avons encore plusieurs perspectives à étudier. Bien sûr, nous souhaitons améliorer la couverture du transducteur par rapport au corpus et dépasser les 95% de phrases analysées, bien qu'il ne reste plus que des cas complexes à traiter. Etant donné que les grammaires AB peuvent sembler limitatives lorsque l'on souhaite traiter d'une langue complexe, nous souhaiterions transformer notre transducteur en un transducteur d'arbres vers les graphes. Cela nous permettrait d'utiliser l'ensemble des règles de Lambek et de nous rapprocher de travaux plus modernes sur la question. Par rapport à l'analyseur de phrase, il manque cruellement d'un typage relatif au lexique que nous extrayons des arbres de dérivation. Cette méthode de typage devrait être implémentée rapidement. De même, il pourrait être intéressant d'utiliser d'autres algorithmes que CYK, tel que l'algorithme d'Earley, ou de typer les phrases en utilisant un système tel que SYGFRAN (Chauché, 2011).

Notre travail est disponible à (Sandillon-Rezer, 2012), sous licence *GNU General Public Licence*.

Références

Abeillé, A. et Clément, L. (2003). Annotation morpho-syntaxique.

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *Treebanks, Kluwer, Dordrecht*.

Buszkowski, W. et Penn, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*, 49(4):431–454.

Chauché, J. (2011). Une application de la grammaire structurelle : L'analyseur syntaxique du francais sygfran.

Comon, H., Dauchet, M., Gilleron, R., Löding, C., Jacquemard, F., Lugiez, D., Tison, S. et Tommasi, M. (2007). Tree automata techniques and applications. Available on : `http://www.grappa.univ-lille3.fr/tata`. release October, 12th 2007.

Earley, J. (1973). An efficient context-free parsing algorithm.

Hockenmaier, J. et Steedman, M. (2007). CCGbank : a corpus of CCG derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, page 355–396.

Hopcroft, J. E. et Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Adison-Wesley Publishing Company, Reading, Massachusets, USA.

Kanazawa, M. (1998). *Learnable Classes of Categorial Grammars*. Center for the Study of Language and Information, Stanford University.

Knight, K. et Graehl, J. (2005). An overview of probabilistic tree transducers for natural language processing.

Knuth, D. E. (1997). *The Art of Computer Programming Volume 2 : Seminumerical Algorithms (3rd ed.)*. Adison-Wesley Professional.

Lambek, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, 65(3).

Levy, R. et Andrew, G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures.

Moortgat, M. et Moot, R. (2001). CGN to Grail : Extracting a type-logical lexicon from the CGN annotation. *In* Daelemans, W., éditeur : *Proceedings of Computational Linguistics in the Netherlands CLIN 2000*.

Moot, R. (2010a). Automated extraction of type-logical supertags from the spoken dutch corpus. *Complexity of Lexical Descriptions and its Relevance to Natural Language Processing : A Supertagging Approach*.

Moot, R. (2010b). Semi-automated extraction of a wide-coverage type-logical grammar for french. *Proceedings TALN 2010, Monreal*.

Sandillon-Rezer, N. (2012). Syntab : http ://www.labri.fr/perso/nfsr/.

Sandillon-Rezer, N.-F. et Moot, R. (2011). Using tree tranducers for grammatical inference. *Proceedings of Logical Aspects of Computational Linguistics 2011*.

Younger, D. (1967). Context free grammar processing in n^3.

Analyse automatique de discours en langue des signes : Représentation et traitement de l'espace de signation

Monia Ben Mlouka
IRIT -TCI , UMR5505, 31000 Toulouse
mlouka@irit.fr

RÉSUMÉ

En langue des signes, l'espace est utilisé pour localiser et faire référence à certaines entités dont l'emplacement est important pour la compréhension du sens. Dans cet article, nous proposons une représentation informatique de l'espace de signation et les fonctions de création et d'accès associées, afin d'analyser les gestes manuels et non manuels qui contribuent à la localisation et au référencement des signes et de matérialiser leur effet. Nous proposons une approche bi-directionnelle qui se base sur l'analyse de données de capture de mouvement de discours en langue des signes dans le but de caractériser les événements de localisation et de référencement.

ABSTRACT

Automatic Analysis of Discourse in Sign Language : Signing Space Representation and Processing

In sign language, signing space is used to locate and refer to entities whose locations are important for understanding the meaning. In this paper, we propose a computer-based representation of the signing space and their associated functions. It aims to analyze manual and non-manual gestures, that contribute to locating and referencing signs, and to make real their effect. For that, we propose an approach based on the analysis of motion capture data of entities' assignment and activation events in the signing space.

MOTS-CLÉS : Langue des signes, Espace de signation, gestes de pointage, capture de mouvement, suivi du regard.

KEYWORDS: Sign language, Signing space, pointing gestures, motion capture, gaze tracker.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 219–232,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'étude de l'aspect gestuel dans les langues naturelles fait l'objet de plusieurs travaux. Plusieurs études ont porté sur l'analyse des gestes manuels et non manuels en situation de dialogue. L'une d'entre elles a apporté une classification fonctionnelle des gestes manuels et non manuels (Cosnier, 1997). Celle de (Montredon, 2001) établit une relation entre les caractéristiques spatio-temporelles des gestes manuels et leurs rôles dans l'énoncé. En langue des signes, le canal étant visuo-gestuel, l'analyse d'énoncés est en premier lieu une analyse d'un signal visuel. Elle peut être entièrement réalisée à ce seul niveau ou être complétée par des analyses du geste 3D si l'on sait reconstruire cette information à partir de données visuelles.

Nous proposons dans cette étude une représentation informatique de l'espace de signation comme étant un élément important pour la compréhension d'un discours dans la langue des signes. Pour cela, nous introduirons, en premier lieu, l'espace de signation et les gestes qui contribuent à la localisation de signes. Dans un second lieu, nous décrirons le corpus sur lequel se base notre analyse. Par la suite, nous présenterons notre approche d'analyse géométrique 3D suivie de quelques résultats.

2 L'espace de signation

L'espace de signation est défini comme étant l'espace qui entoure le signeur et qui est atteignable par ses deux mains. L'espace de signation sert à localiser les entités ou notions associées à certains signes, éventuellement à spécifier leurs propriétés de forme et de taille et à établir des relations spatiales entre les entités (Cuxac, 2000).

2.1 Evénements liés à l'espace de signation

Notre représentation informatique de l'espace de signation étant un graphe d'entités spatialisées dans un espace 3D. Elle dispose des fonctions classiques de création, de modification, de suppression et d'accès. Chacune de ces fonctions est déclenchée par un événement survenu dans l'espace de signation. L'image (1a) est un exemple d'un signe [TABLE] qui occupe une zone de l'espace. L'image (1b) illustre une association spatiale d'une action [S'ASSOIR]. L'image (1c) est un exemple d'un pointage manuel vers une zone particulière de l'entité [TABLE]. La zone est spécifiée par la direction de la main dominante.

2.2 Aspect multilinéaire

L'étude de (Fusellier-Souza, 2004) s'est intéressée aux gestes manuels et non manuels qui contribuent aux changements d'états de l'espace de signation. Une étude similaire, celle de (Thompson *et al.*, 2006) a porté sur la relation entre le regard et la réalisation d'actions spatialisées. Les deux études soulignent l'aspect multilinéaire dans la réalisation de gestes. Dans cette étude nous nous focaliserons sur les gestes de création et de référencement d'entités dans l'espace de signation. Comme nous l'avons cité précédemment, les gestes sont manuels et non manuels que ce soit pour les événements de création ou de référencement. Les gestes manuels de création d'entités dans

FIGURE 1 – a : Un objet spatialisé [TABLE], b : Une action spatialisée [S'ASSOIR], c : Un pointage vers une partie de la portion de l'espace occupé par le signe[TABLE]

l'espace de signation représentent les signes effectués à un emplacement spécifique ou les signes localisés sur le corps puis situés dans l'espace de signation par un pointage :

– L'orientation du regard et l'inclinaison de la tête permettent d'associer une zone de l'espace de signation au signe en cours de réalisation.
– Les gestes manuels peuvent associer une forme à l'entité et préciser la taille qu'elle occupe dans l'espace.

Les images (1a), (1b) et (1c) représentent des exemples de réalisation de l'aspect de multi-linéarité :
– Dans (1a) et (1c), on observe que le regard du signeur est orienté vers la même zone d'espace occupée par l'entité.
– On observe également une posture particulière dans (1b) qui se manifeste par une inclinaison de la tête vers l'emplacement de la main qui effectue le signe [S'ASSOIR].
– Dans (1c), on ne peut pas déterminer si le regard fixe cet endroit de l'espace car la tête du signeur est baissée mais on peut le déduire grâce à la position de la tête légèrement inclinée vers le bas.

Par ces exemples, nous avons illustré l'importance de l'aspect multilinéaire de gestes qui contribuent aux fonctions de création, de référencement de l'entité [S'ASSEOIR] : fixation du regard, gestes manuels et mouvements de la tête. Dans la littérature, peu de travaux ont porté sur la représentation informatique de l'espace de signation. Nous citons l'étude de (Lenseigne, 2005) qui a porté sur la représentation informatique de la structure de l'espace de signation dans un discours en langue des signes française. L'aspect gestuel a été pris en compte dans l'étude de (Braffort, 1996) qui consistait à modéliser les gestes dans les verbes directionnels et déictiques à partir de données fournies par un gant numérique. (Lu et Huenerfauth, 2011) a développé une technique qui, à partir de données de capture de mouvement, permet de modéliser les gestes dont la réalisation est influencée par la localisation spatiale des entités dans l'espace de signation.Dans cette dernière étude, (Lu et Huenerfauth, 2011) s'est intéressé aux gestes exprimant des verbes. Notre travail s'inscrit dans le même cadre et s'intéresse à l'aspect multilinéaire (combinaisons de gestes manuels et non manuels) et se base à la fois sur des données tri-dimensionnelles de capture de mouvement et sur des données de suivi du regard synchronisées avec les enregistrements vidéos.

3 Acquisition du Corpus

Le corpus sur lequel nous avons appliqué notre approche d'analyse a été enregistré dans le cadre du partenariat franco-québécois (Marqspat) [1]. Les sessions de capture ont été réalisées avec un système de capture de mouvement (VICON) [2], une caméra vidéo pour filmer le cadre complet de la scène et un système de capture du regard (FaceLab) [3] . Lors de l'enregistrement, le signeur commence et termine sa production par un "clap" manuel qui permettra d'effectuer ultérieurement une synchronisation de la vidéo et des données de la capture de mouvement et commence à répondre à des questions sous forme de vidéos projetées. Les questions concernent des détails à propos de scènes enregistrées préalablement. Les données de capture de mouvement et celles de suivi du regard feront l'objet d'une analyse automatique (Ben Mlouka *et al.*, 2010) dont on détaillera les étapes dans la section suivante.

4 Annotations et représentations informatiques

Nous avons adopté une méthode composée de plusieurs étapes :

4.1 Grille d'annotation

L'annotation [4] du corpus a été réalisée et vérifiée par plusieurs annotateurs québécois de compétences variées en Langue des Signes Française, Québécoise et Américaine. La grille d'annotation se compose de :

Une annotation en gloses Les annotateurs ont transcrit les signes effectués par les deux mains et qui peuvent avoir ou non une association spatiale spécifique. La capture d'écran de la grille d'annotation (2) montre un exemple de valeurs attribuées aux pistes : la main droite (MD) : [S'ASSEOIR], la piste (MG) transcrit les signes effectués par la main gauche, la piste (2M) inclut ceux à deux mains.

Une annotation de gestes et de signes en liaison avec l'espace Les gestes manuels et non manuels transcrits sont nécessairement associés à une zone de l'espace de signation. A chaque geste ou signe spatialisé, on associe une étiquette (ex. x, y, z, etc.) pour étiqueter la zone à laquelle est associé ce geste ou ce signe. Dans l'exemple (2), la piste (MC) mentionne le nom de la zone associée au signe [S'ASSEOIR], cela veut dire qu'une zone est occupée par une entité "y" dont le signifiant est l'action [S'ASSEOIR]. Le reste des pistes visibles inclut les noms d'entités spatiales vers lesquelles un composant corporel fait référence :

1. Lien vers le site web du projet : http ://www.irit.fr/marqspat/index.html
2. Il s'agit d'un système de capture composé de 8 caméras infrarouges qui enregistrent les positions 3D de marqueurs réfléchissants posés sur les membres du signeur
3. Un système non invasif composé de deux caméras et d'un émetteur infra-rouge. Il fournit sous forme de données 3D, l'orientation du regard. Une caméra de scène qui permet de synchroniser les données vidéos et données 3D
4. Ces annotations manuelles ont été effectuées par l'équipe "Groupe de recherche sur la langue des signes québécoise et le bilinguisme sourd" qui pilote le projet (Marqspat)

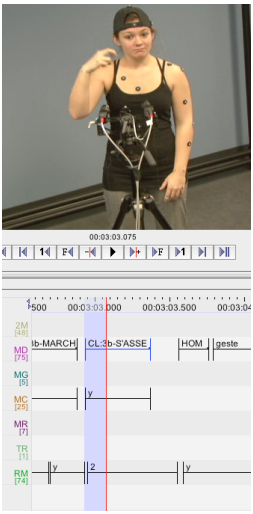


FIGURE 2 – Un exemple de valeurs d'annotation

1. La piste (MR), représente les entités ciblées par les pointages manuels

2. La piste (TR) celle des entités référencées par la tête.

3. La piste (RM) celle de la direction du regard où on nomme l'entité sur laquelle se focalise le regard.

Dans l'image (3a), la zone de l'espace occupée par le signe [S'ASSOIR] est étiquetée x. L'image (3b) illustre un exemple où la main dominante effectue le signe [FILLE], le regard et la tête se dirigent vers une même cible (x)

La lecture transversale de la grille d'annotation est un moyen qui permet de grouper les mêmes étiquettes d'entités (x dans les deux exemples précédents). Une interprétation simple de cette lecture transversale : Le signeur associe une zone de l'espace à une entité (x) grâce à l'orientation du regard et de la tête et associe le référent [FILLE] à la zone (x) grâce au signe effectué près de la tête.

Cette mise en correspondance entre les gestes et leurs interprétations sera généralisée grâce à l'extraction et la synchronisation automatique des annotations manuelles avec les données de capture de mouvement qui leurs correspondent dans le but de mettre en place des modèles géométriques propres aux gestes qui contribuent à la création ou au référencement d'une même entité.

FIGURE 3 – Une lecture transversale d'annotation d'un événement de référencement : a- Localisation manuelle et non manuelle (regard) du signe [S'ASSOIR], b- Localisation non manuelle (tête et regard) du signe [FILLE] car celui-ci se réalise par la main dominante à un emplacement spécifique (près de la tête)

4.2 Représentation géométrique

Comme nous l'avons mentionné dans 4.1, les articulateurs représentées sont la main dominante, la tête et le regard.

Mesure de l'enveloppe de la main La main dominante est représentée par une sphère de centre milieu des bases de l'index et celui de l'auriculaire, de rayon la longueur du majeur(voir figure 4).

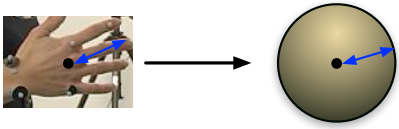


FIGURE 4 – Représentation géométrique de la main

La cible du regard La cible du regard est représentée par un point dont les coordonnées sont fournies [5] par le système de suivi du regard. Il fournit la position 3D de la cible du regard à un instant donné, exprimée dans le même repère que celui des positions des marqueurs de capture de mouvement [6]. Nous avons noté que le taux de points de vergence reconstruits par (FaceLab) est relativement faible par rapport aux données enregistrées. Ceci est dû au fait que à plusieurs moments, les directions du regard calculées ne sont pas convergentes et par conséquent ne permettent pas de calculer les positions des points de vergence.

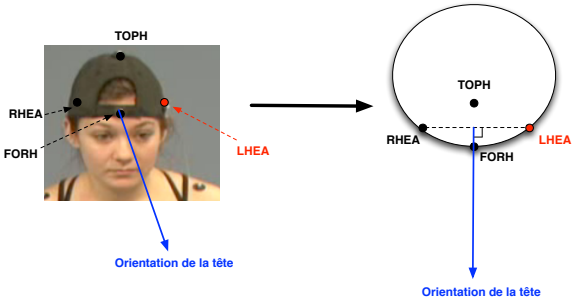


FIGURE 5 – Représentation géométrique de l'orientation de la tête (à gauche : vue de face, à droite : vue de dessus)

Mesure de l'orientation de la tête L'orientation de la tête est mesurée comme étant la normale à la droite passant par les marqueurs "RHEA" et "LHEA" et passant par le marqueur "FORH" (voir figure 5). Le vecteur $\vec{n}$ est le vecteur normal au plan formé par les marqueurs : RHEA, LHEA et TOPH. La figure (5) indique la position de ces marqueurs en vue de dessus. L'équation du plan étant :

$$P : a.x + b.y + c.z + d = 0 \tag{1}$$

Le vecteur normal est le résultat du produit vectoriel suivant :

$$\vec{n} = \vec{AB} \wedge \vec{AC} \tag{2}$$

Tels que : A, B et C représentent la position géométriques des marqueurs RHEA, TOPH et LHEA respectivement. Dans la suite nous allons caractériser l'aspect multi-composant entre les différents modèles géométriques.

5. (FaceLab) fournit une liste de mesures sur : la position des globes oculaires, les pupilles, le degré de fermeture des yeux, l'angle d'orientation du regard, etc. Dans ce travail, nous nous sommes intéressés aux positions de points de vergence seulement

6. Nous avons fusionné les données fournies par le système de capture de mouvement et celles fournies par (FaceLab) dans un même repère

4.3 Mesures et relations

On se propose dans cette phase de prendre en compte le sens comme cela a été détaillé dans 4.1 et d'extraire les positions géométriques correspondantes de chaque composant corporel. Par la suite nous caractériserons la convergence des composants corporels comme étant l'intersection simultanée ou différée des représentations géométriques de l'orientation de la tête, la position de la main dominante et la cible du regard. La notion d'intersection géométrique inclut deux différents composants corporels, on parle ainsi d'une relation entre composants. On qualifie l'intersection différée de deux positions géométriques d'un même composant de relation intra-composant.

4.3.1 Relations entre composants

Mesures de l'intersection orientation tête et main dominante On se propose de mesurer la distance d entre la droite portant l'orientation de la tête D et le centre de la main S :

$$d = \frac{|| \vec{n} \wedge S\vec{M}_D ||}{|| \vec{n} ||} \tag{3}$$

Tels que M est un point appartenant à la droite D. Cette même formule s'applique également pour la mesure de l'intersection entre la cible fixée ou pointée et l'orientation de la tête.

Mesures de l'intersection cible du regard et main dominante On se propose de mesurer la distance d entre le point qui représente la cible fixée par le regard et P et le centre de la main S :

$$d = \sqrt{(x_C - x_p)^2 + (y_C - y_p)^2 + (z_C - z_p)^2} \tag{4}$$

Tels que C est le centre de la main dominante et p et le point cible du regard. Cette formule s'applique également pour la mesure de l'intersection entre la cible fixée et espace référencé par la main dominante.

4.3.2 Relations intra-composant

Mesures de la convergence des espaces occupées par une seule entité On se propose de mesurer la distance d entre deux sphères représentatives de la position de la main dominante à deux instants distincts :

$$d = \sqrt{(x_{C1} - x_{C2})^2 + (y_{C1} - y_{C2})^2 + (z_{C1} - z_{C2})^2} \tag{5}$$

Tels que $C1$ et $C2$ sont les centres des sphères qui représentent la main dominante à deux instants différents. Cette formule s'applique également pour la mesure de convergence entre espace occupé et espace référencé par la main dominante.

Mesure de la variation de l'orientation de la tête Dans le but de mesurer la variation de cette orientation au cours d'un événement de référencement, nous nous proposons de mesurer l'angle

formé par deux vecteurs porteurs des droites D_1 et D_2 d'orientation de la tête correspondante à deux instants distincts.

L'angle θ est mesuré selon cette formule :

$$cos(\theta)\frac{\vec{n_1}.\vec{n_2}}{\sqrt{|| \vec{n_1} || . || \vec{n_2} ||}} \tag{6}$$

Tels que $\vec{n_1}$ et $\vec{n_2}$ sont les vecteurs directeurs de D_1 et D_2 respectivement dont les coefficients sont calculés selon la formule énoncée en (4.2).

4.4 Premiers résultats

4.4.1 Objectifs de l'analyse

On se propose d'apporter des éléments de réponses par rapport à l'état de l'espace de signation. En particulier, on veut déterminer si l'espace de signation, tel qu'il est perçu par le signeur, subit des transformations géométriques (translation et/ou rotation) au cours des événements de référencement. Pour cela, on étudiera la relation entre l'emplacement de l'entité qui occupe une partie de l'espace de signation et la position de la main dominante du signeur quand celui-ci pointe vers cette même entité.

Pointages manuel vers une même entité Comme cela a été détaillé dans 4.3.2, nous avons mesuré la distance qui sépare deux positions de la main dominante lors de la réalisation d'un signe spatialisé et lors d'un pointage vers cette même entité 6.

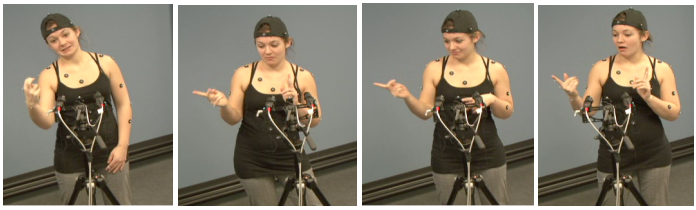


FIGURE 6 – a- Création d'une entité - X, b, c et d- Pointages vers X

	Figure 6b	Figure 6c	Figure 6d
Moyenne de la distance (mm)	685, 6	611, 5	569, 7
Ecart-type de la distance (mm)	3, 2	1, 6	25, 2

TABLE 1 – Mesure de la distance qui sépare deux positions de la main dominante lors de la réalisation d'un signe spatialisé et lors d'un pointage vers cette même entité

Les valeurs moyennes des distances du tableau 1 sont significatives car elles vérifient la règle 68-95-99.7 de la loi normale [7]

$$\mu - 3 * \sigma < 99.7\% * N < \mu + 3 * sigma \tag{7}$$

$$\mu - 2 * \sigma < 95\% * N < \mu + 2 * sigma \tag{8}$$

$$\mu - \sigma < 68\% * N < \mu + sigma \tag{9}$$

Tels que : μ est la valeur moyenne et σ est l'écart-type de l'ensemble des valeurs de distance N

De ce fait, on déduit que les distance entre positions de la main dominante illustrées dans les images (6b, c et d) et celle de l'image (6a) peuvent être générées par une même loi de distribution normale de moyenne : 622,3 et d'écart-type :58,7 (en mm). Cela veut dire que dans ces trois exemples (Création-Référencement), les distances qui séparent deux positions différentes de la main dominante (la première en phase de création et la deuxième en phase de référencement) ne sont pas exactement les mêmes mais varient autour d'une même moyenne.

Pointages manuels vers deux entités différentes La même formule 4.3.2 a été appliquée sur les séquences 7.

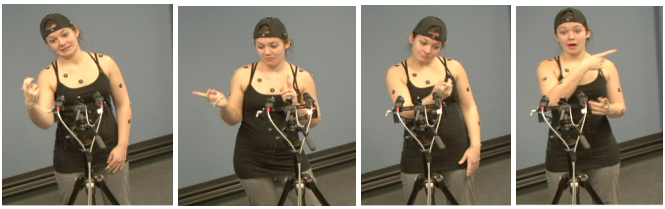


FIGURE 7 – a- Création d'une entité - X, b- Pointage vers X, c- Création d'une entité - Y, d- Pointage vers Y

TABLE 2 – Mesure de la distance qui sépare deux positions de la main dominante lors de la réalisation d'un signe spatialisé et lors d'un pointage vers une même entité

	Figure 7b	Figure 7d
Moyenne de la distance (mm)	685,6	194,3
Ecart-type de la distance (mm)	3,2	10,6

Les valeurs moyennes des distances du tableau 2 ne vérifient pas la règle 68-95-99.7.
Bien que la variation des distances n'est pas importante car l'écart type est de l'ordre de (6.9)mm, les distances mesurées ne vérifient pas une distribution normale.

7. Loi normale : 68% de la population se trouve entre $\mu - \sigma$ et $\mu + \sigma$, 95% de la population se trouve entre $\mu - 2 * \sigma$ et $\mu + 2 * \sigma$,99.7% de la population se trouve entre $\mu - 3 * \sigma$ et $\mu + 3 * \sigma$

Pointages Par la tête La même formule 4.3.1 a été appliquée sur les séquences de la figure (8).

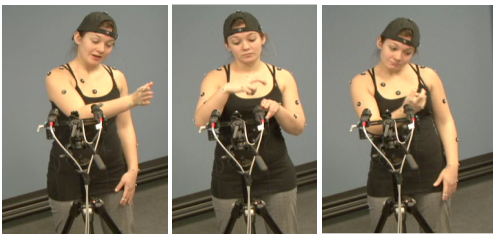


FIGURE 8 – a- Localisation d'une entité - X, b- création d'une entité X, c- création et pointage non manuel vers X

TABLE 3 – Mesure de la distance entre la droite portant l'orientation de la tête et la position de la main dominante

	Figure 8a	Figure 8b	Figure 8c
Moyenne de la distance (mm)	168,1	221,6	106,1
Ecart-type de la distance (mm)	8,2	2,7	2,1

- La distance moyenne 168,1 mm est la distance qui sépare la droite portant l'orientation de la tête (voir 8c) et la position de la main dominante illustrée dans (8a).
- La distance moyenne 221,6 mm est la distance qui sépare la droite portant l'orientation de la tête (voir 8c) et la position de la main dominante illustrée dans (8b).
- La distance moyenne 106,1 mm est la distance qui sépare la droite portant l'orientation de la tête et la position de la main dominante illustrées dans (8c).

Dans le paragraphe 4.4.1, nous avons mesuré la distance qui sépare la position de la main dominante à deux instants différents, lorsque la main réalise un signe spatialié et lorsqu'elle le pointe. Dans ce paragraphe, nous avons appliqué la même méthode en remplaçant le référencement manuel par le référencement réalisé par la tête (comme l'illustre la figure la figure 9c). Nous avons calculé non pas la distance entre deux positions de la main mais la distance entre une position de la main et la droite qui porte l'orientation de la tête. D'après les mesures du tableau 3, n'appartiennent pas à une même loi de distribution normale. Bien qu'il s'agisse de la même entité, les mesures de distances (Tête-main) varient différemment pour chaque cas.

Pointages par le regard La même formule 4.3.1 a été appliquée sur les séquences 10.

- La distance moyenne 545,3 mm est la distance qui sépare la position de la cible du regard et la position de la main dominante dans les figures (10a) et (10b) .

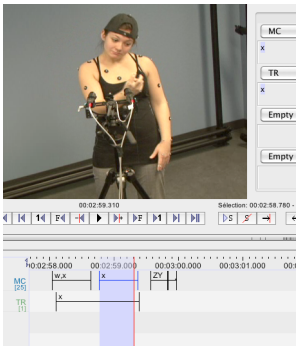


FIGURE 9 – Posture de la tête et position de la main dominante qui réalisent la localisation d'une même entité

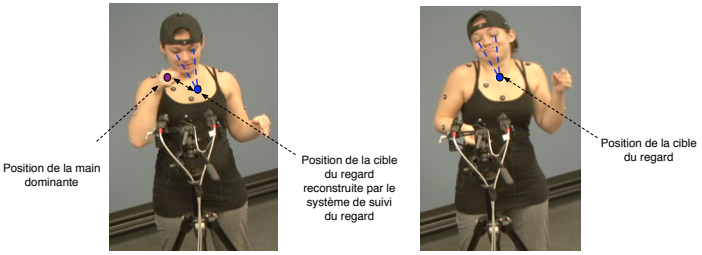


FIGURE 10 – a- Localisation d'une entité (Y) par le regard et par la main dominante , b- Localisation de la même entité par le regard

TABLE 4 – Mesure de la distance entre la position de la cible du regard et la position de la main dominante

Moyenne des distances (mm)	545, 3
Ecart-type des distance (mm)	9, 1

4.5 Retour sur résultats

Pointages manuels D'après les mesures de distances réalisées sur une session de capture, les trois pointages manuels qui pointent vers une même zone spatiale (6b, c et d) présentent un même comportement spatial par rapport à la position de l'entité créée dans l'espace de signation (6 a). Nous en déduisons que la position spatiale de la zone occupée par l'entité [S'ASSOIR] perçue par le signeur est conservée au cours des trois pointages manuels. Les séquences de pointages manuels illustrés dans (7 b et d) font référence à deux entités différentes (X) et (Y) respectivement. Les mêmes mesures de distance indiquent une évolution différente de la distance. Ceci est en relation avec l'entité pointée non pas avec la notion de pointage en tant que notion linguistique qui ne dépend pas de la cible vers laquelle pointe le signeur.

Pointages non manuel Les mesures du tableau (3) montrent que les distances entre la droite portant l'orientation de la tête et la position de la main dominante illustrées dans (8a) et (8b) ne sont pas similaires. Nous en déduisons que l'orientation de la tête est étroitement liée à la position de la main dominante courante pour le signe [S'ASSOIR]. En d'autres termes, la position qu'occupe l'entité [S'ASSOIR] dans l'espace telle qu'elle est perçue par le signeur n'est pas conservée lors des deux différents pointages par la tête. Les mesures du tableau (4) montrent que la distance entre la mire (cible du regard) et la position de la main dominante garde une valeur quasi constante ce qui signifie que la position de l'entité [S'ASSOIR] perçue par le signeur est la même lors des deux pointages distincts par le regard.

5 Perspectives

L'approche que nous avons présentée concerne l'analyse de gestes manuels et non manuels liés à la localisation d'entités dans l'espace de signation. Cette approche pourrait apporter des éléments de réponses par rapport aux propriétés spatiales des entités qui occupent l'espace de signation au moment du discours. Les premières interprétations 4.5 des mesures effectuées sur une base de données de capture de mouvement révèlent que la zone occupée par une entité telle qu'elle est perçue chez le signeur ne change pas au cours d'un discours continu. Nous avons abouti à cette conclusion grâce aux positions relatives de la main dominante, de la cible du regard et de l'orientation de la tête. Cela veut dire que l'espace de signation dans sa globalité ne subit pas de changement (translation ou rotation) au cours des séquences d'enregistrements sur lesquelles nous avons effectué nos mesures. Ceci écarte l'hypothèse d'un éventuel changement de position de l'espace de signation au cours d'un discours continu et apporte des précisions pour sa représentation informatique. Toutefois, il serait intéressant d'analyser l'évolution de l'état de l'espace de signation lors des pseudo-transferts de rôle [8] où l'hypothèse de changement de positions de l'espace de signation est fortement appuyée.

8. ou semi-transfert personnel : Un court moment où le signeur émet une action (verbe) et devient le personnage qui fait l'action à travers sa posture et son expression faciale(Cuxac, 2000)

6 Conclusion

L'approche proposée se base sur l'interprétation linguistique d'un discours en langue des signes et exploite les données tridimensionnelles fournies afin d'extraire des comportements répétitifs des gestes liés à l'espace de signation. Cette analyse prend en compte la multi-linéarité des gestes effectués à la fois par la main dominante et la tête. Le regard contribue par des fixations vers des emplacements spécifiques de l'espace de signation. Cependant, nous nous sommes focalisés seulement sur l'analyse de deux fonctions linguistiques, celles de création et de référencement d'entités dans l'espace de signation.

L'analyse que nous avons menée visait à apporter des précisions sur l'évolution de la structure de l'espace de signation. En particulier, nous avons pu déduire que le signeur perçoit les zones occupées par des entités comme étant des zones fixes. Par conséquent, l'espace de signation reste figé au cours de pointages manuels et de pointages par le regard.

Remerciements

Le corpus 3D a été réalisé dans Le cadre du projet (Maqspat) qui porte sur le thème du marquage spatial dans les langues des signes française, américaine et québécoise. Le projet est soutenu par le CRSH, dans le cadre d'un partenariat stratégique soutenu par le CFQCU. Le corpus d'illustration (1a et c) a été réalisé en interne par Juliette Dalle, assistante ingénieur de l'équipe Traitement et compréhension de d'Image de l'Institut de Recherche en Informatique de Toulouse (IRIT).

Références

Ben Mlouka, M., Albaret-Lefebvre, F., Dalle, J. et Dalle, P. (2010). Annotation automatique d'une vidéo en lsf à partir de données de capture de mouvement. *In TALS*, Montréal, Canada.

Braffort, A. (1996). *Reconnaissance et compréhension de gestes, application à la langue des signes*. Thèse de doctorat, Université de Paris XI.

Cosnier, J. (1997). Sémiotique des gestes communicatifs. *Nouveaux actes sémiotiques*, 52:7–28.

Cuxac, C. (2000). *Faits de Langues - La langue des signes française (LSF) - Les voies de l'iconicité*. Faits Des Langues : Ophrys, Paris.

Fusellier-Souza, I. (2004). *Sémiogenèse des langues des signes : étude de langues des signes primaires (LSP) pratiquées par des sourds brésiliens*. Thèse de doctorat, Université Paris 8.

Lenseigne, B. (2005). Modélisation de l'espace discursif pour l'analyse de la langue des signes. *In TALN*, Dourdan.

Lu, P. et Huenerfauth, M. (2011). Synthesizing American Sign Language Spatially Inflected Verbs from Motion-Capture Data. *In SLTAT*, Dundee, UK.

Montredon, J. (2001). De la gestualité co-verbale, dimensions cognitives et symboliques. *In* Presses Univ. Limoges, ., éditeur : *Sémio*, pages 15–18.

Thompson, R., Emmorey, K. et Kluender, R. (2006). The Relationship between Eye Gaze and Verb Agreement in American Sign Language : An Eye-tracking Study. *Natural Language & Linguistic Theory*, 24(2):571–604.

ResTS : Système de Résumé Automatique des Textes d'Opinions basé sur Twitter et SentiWordNet

Jihene Jmal
LARODEC, ISG, Université de Tunis, 2000, Le Bardo, Tunisie
fer.jmal_jihene@hotmail.fr

RESUME

Comme le E-commerce est devenu de plus en plus populaire, le nombre de commentaires des internautes est en croissance constante. Les opinions sur le Web affectent nos choix et nos décisions. Il s'avère alors indispensable de traiter une quantité importante de critiques des clients afin de présenter à l'utilisateur l'information dont il a besoin dans la forme la plus appropriée. Dans cet article, nous présentons ResTS, un nouveau système de résumé automatique de textes d'opinions basé sur les caractéristiques des produits. Notre approche vise à transformer les critiques des utilisateurs en des scores qui mesurent le degré de satisfaction des clients pour un produit donné et pour chacune de ses caractéristiques. Ces scores sont compris entre 0 et 1 et peuvent être utilisés pour la prise de décision. Nous avons étudié les opinions véhiculées par les noms, les adjectifs, les verbes et les adverbes, contrairement aux recherches précédentes qui utilisent essentiellement les adjectifs. Les résultats expérimentaux préliminaires montrent que notre méthode est comparable aux méthodes classiques de résumé automatique basées sur les caractéristiques des produits.

ABSTRACT

System of Customer Review Summarization using Twitter and SentiWordNet

As E-commerce is becoming more and more popular, the number of customer reviews raises rapidly. Opinions on the Web affect our choices and decisions. Thus, it is more efficient to automatically process a mixture of reviews and prepare to the customer the required information in an appropriate form. In this paper, we present ResTS, a new system of feature-based opinion summarization. Our approach aims to turn the customer reviews into scores that measure the customer satisfaction for a given product and its features. These scores are between 0 and 1 and can be used for decision making and then help users in their choices. We investigated opinions extracted from nouns, adjectives, verbs and adverbs contrary to previous research which use only adjectives. Experimental results show that our method performs comparably to classic feature-based summarization methods.

MOTS-CLES : Fouille d'opinion, Classification, Intensité de l'Opinion, Résumé de texte d'opinion, Popularité.

KEYWORDS: Opinion mining, Sentiment Classification, Opinion Strength, Feature-based Opinion Summarization, Feature Buzz Summary.

1 Introduction

Dans le Web 2.0 (Web social ou participatif), l'utilisateur est un acteur principal qui par-

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 233–246,
Grenoble, 4 au 8 juin 2012.

tage des documents, des informations, des avis. Il interagit, collabore avec autrui, s'exprime et donne son opinion. Il a des services à sa disposition tels que les réseaux sociaux (twitter, facebook, etc.), les blogs, les forums, les wikis, les sites de partages de vidéos, de photos, de musiques, etc. L'utilisation fréquente de ces services fournit un contenu généré par l'utilisateur (UGC : User Generated Content) qui représente de nos jours une quantité de données qui se mesure en yotaoctets (10^{24}). Ce contenu est composé généralement de données textuelles qui sont porteuses d'opinions et de sentiments. L'accès au contenu sémantique des ces données, préalable à la connaissance des opinions qu'elles véhiculent, représente un enjeu pour de nombreux acteurs. Par exemple :

- le consommateur, c'est-à-dire chacun de nous, qui veut s'informer avant toute décision qu'elle soit d'achat ou autre;
- les fournisseurs de biens et de services qui cherchent à se positionner les uns par rapport aux autres dans un univers hautement compétitif et face à une demande de plus en plus complexe à identifier;
- les chercheurs : économistes, sociologues,… ou simplement les responsables publics qui cherchent à comprendre le comportement individuel ou collectif pour anticiper, réguler ou ajuster les rapports entre les différents agents socio-économiques.

C'est dans ce contexte que s'introduit la fouille d'opinion (Opinion Mining, Sentiment Analysis ou Subjectivity Analysis) qui est un sous domaine de la fouille de texte. Son but étant de ressortir les marques d'opinions et de sentiments des documents textuels. Une opinion peut être définie comme l'expression des sentiments d'une personne envers une entité (Liu, 2010). En outre, l'e-commerce devient de plus en plus populaire. Les marchands et les fabricants de produits permettent aux clients de donner leurs avis et opinions sur les produits ou services qu'ils ont vendus (par exemple amazon.com, epinions.com). De plus, les opinions disponibles sur le Web influent sur nos choix et décisions. En effet, d'après une étude menée en 2009 par le CRÉDOC (Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie), 57% des internautes français ont cherché des avis des autres sur le Web et 66% d'entre eux font confiance en ces commentaires (Lehuédé, 2009). La fouille d'opinion peut être divisée en trois sous domaines qui sont la classification de la subjectivité (subjectif/objectif) (Riloff et al, 2003), la classification des sentiments (positif/négatif ou positif/négatif/neutre)(Pang et Lee, 2002), (Wilson et al, 2004) et (Blitzer et al, 2007) et le résumé d'opinions (Hu et Liu, 2004), (Popescu et Etzioni, 2005) et (Gamon et al, 2005).

Nous proposons une nouvelle approche de résumé automatique des textes d'opinions basée sur les commentaires des utilisateurs. Cette approche vise à transformer ces commentaires en des scores qui mesurent l'intensité de l'opinion. Ces scores peuvent être utilisés pour la prise de décision et aident les utilisateurs dans leurs choix. Pour ce faire, nous avons commencé par extraire les caractéristiques des produits à partir des critiques des utilisateurs (exemple batterie, écran, son, image, etc.). Ensuite, nous avons attribué à chaque caractéristique un score calculé à partir de sa fréquence d'apparition dans le corpus pondérée par sa popularité dans le Web 2.0, en particulier sur Twitter[1] ; la plateforme de microblogage la plus populaire. Nous avons par la suite identifié les phrases d'opinion et affecté

[1] www.twitter.com

à chaque verbe et adjectif un score de SentiWordNet (Baccianella et al, 2010). Si la phrase contient un adverbe, ces scores sont pondérés par l'intensité de l'opinion véhiculée par cet adverbe en se référant à la liste de modificateurs (en anglais *intensifier* et *diminisher*) que nous avons préparé. Nous avons enfin calculé le score de tout le produit qui mesure la satisfaction globale des clients. Voici un exemple de résumé généré par notre système pour le produit *iPod* :

Produit : iPod

Satisfaction Client = 60%

Caractéristique 1 : *Player* : Popularité = 70%

Satisfaction Client = 83%

Caractéristique 2 : *Ecran* : Popularité = 54%

Satisfaction Client = 62%

….

Les caractéristiques des produits sont classées en fonction de leurs popularités sur le web 2.0. Dans notre conception, un produit n'est pas simplement considéré comme recommandé ou non recommandé, au contraire, nous laissons l'utilisateur libre de faire son choix en se référant aux différents scores que nous mettons à sa disposition traduisant la satisfaction des clients pour l'ensemble du produit et encore pour chacune de ses caractéristiques. Lors du calcul de ces scores, nous avons étudié l'opinion véhiculée par les noms, adjectifs, verbes et adverbes, contrairement aux autres recherches qui utilisent principalement les adjectifs.

2 Etat de l'art

Nous proposons dans cet article deux types de résumés qui sont le résumé d'opinion basé sur les caractéristiques des produits (Feature-based Opinion summarization), et le résumé de leurs popularités qui montre aux entreprises ce qu'intéresse réellement leurs clients (Feature Buzz Summary). Nous avons également fusionné deux axes de recherche à savoir le résumé basé sur les caractéristiques (Hu et Liu, 2004) (Liu et Ding, 2008) (Zhang et Liu, 2011) et l'identification de l'intensité de l'opinion (Wilson et al, 2004). Nous nous sommes basées essentiellement sur l'approche de Hu et Liu (Hu et Liu, 2004). Les deux auteurs utilisent les règles d'association pour extraire les caractéristiques fréquentes des produits. Pour identifier les mots d'opinion (les adjectifs seulement), ils ont eu recours à WordNet[2] en conjonction avec une liste de mots clés subjectifs (seed words) manuellement préparée. Leur système extrait uniquement les caractéristiques explicites. Une année plus tard, ces auteurs ont mis en œuvre Opinion Observer (Liu et al, 2005), un système offrant une comparaison visuelle entre produits en tenant compte des critiques des utilisateurs sur le Web. Ils identifient les caractéristiques des produits à partir des rubriques *Pros* destinée aux avis positifs et *Cons* celle des avis négatifs.

Plusieurs recherches ont étudié le problème de la détection de mots d'opinion. Il y a des

[2] http://wordnet.princeton.edu/

approches fondées sur le corpus (Corpus-based Approach) (Hatzivassiloglou et McKeown, 1997), (Wiebe, 2000), (Kanayama et Nasukawa, 2006) et (Qiu et al, 2009), d'autres basées sur le dictionnaire (Dictionary-based Approach) (Hu et Liu, 2004), (Kim et Hovy, 2004), (Kamps et al, 2004), (Esuli et Sebastiani, 2005), (Takamura et al, 2005), (Andreevskaia et Bergler, 2006), (Dragut et al, 2010) et (Bouchlaghem et al, 2010). Hu et Liu utilisent seulement les adjectifs pour la détection des opinions. Ils construisent manuellement une liste d'adjectifs qu'ils utilisent pour prédire l'orientation de la phrase et utilisent WordNet pour alimenter la liste par les synonymes et les antonymes des adjectifs dont on connait la polarité. Ils assignent 1 à chaque adjectif positif et 0 à chaque adjectif négatif. Toutefois, dans notre conception, les adjectifs, les verbes et les adverbes jouent un rôle important dans l'analyse des sentiments. Ils sont tous utilisés pour exprimer une opinion ou une émotion dans le texte, par exemple, le verbe apprécier dans «J'apprécie ce produit" inspire un sentiment positif, même si la phrase ne contient ni adjectif, ni adverbe. Dans (Liu et al, 2005), les auteurs comptent le nombre d'occurrences de chaque entité dans la rubrique *Pros* exprimant un avis positif et *Cons* celle des avis négatifs. Dans (Zhang et Liu, 2011), les auteurs ont montré que les syntagmes nominaux et le substantif peuvent aussi enfermer des opinions. Ils comptent le nombre de phrases positives et négatives pour chaque fonctionnalité du produit en utilisant le lexique d'opinion préparé par (Ding et al, 2008). Leur approche permet d'atteindre une précision moyenne d'environ 0,44. Dans notre conception, nous rejoignons l'avis de ces auteurs. Nous considérons également que les noms peuvent exprimer une opinion. En outre, déceler la polarité de l'opinion n'est toujours pas suffisant. La force (intensité) de l'opinion est également nécessaire. En effet, la subjectivité est exprimée de différentes manières ; «*good battery* » est différent de «*great battery* » et de «*excellent battery* ». (Wilson et al, 2004) et (Pang et Lee, 2005) mettent l'accent sur la détection de la force de l'opinion. (Wilson et al, 2004) utilisent les techniques de boosting, rule learning et support vector regression. (Pang et Lee, 2002) et (Turney, 2002) classent les documents comme « thumbs up » ou « thumbs down », selon l'opinion qu'ils véhiculent. Cependant, (Pang et Lee, 2005) exploitent les techniques d'apprentissage automatique pour donner un score de 1 à 5 aux passages d'opinions.

3 Approche proposée

Notre approche est basée sur les travaux de (Hu et Liu, 2004). La figure 1 présente le modèle proposé. Elle a été mise en œuvre dans notre système ResTS. Nous commençons par recueillir les commentaires des internautes à partir du Web et procédons par l'opération de prétraitement du corpus collecté. Notons que notre système effectue toutes les étapes suivantes d'une manière automatisée et sans aucune intervention humaine. Rappelons que l'opinion est une expression des sentiments d'une personne envers une entité ou un aspect de l'entité (Liu, 2010). Une entité peut être un produit, une personne, un événement, une organisation ou un sujet. Elle est représentée comme une hiérarchie de composants, de sous-composant et ainsi de suite où chaque nœud représente un composant et est associé à un ensemble d'attributs (Liu, 2010). Par conséquent, l'entité elle-même peut également être considérée comme une caractéristique. Une critique de l'entité elle-même est appelée une opinion générale comme dans «*I like this iPod* ». Une critique d'une de ses caractéristiques est appelée une opinion spécifique comme dans « *the battery is really good*». Comme Hu et Liu, notre tâche est loin d'être un résumé traditionnel de texte. A partir des critiques des utilisateurs, nous proposons un résumé structuré qui donne une vue globale et concise des

opinions des clients. Hu et Liu ne présentent que le nombre de passages jugés positifs et ceux négatifs pour chacune des caractéristique du produit. Notre système offre plus de détails. Nous fournissons un score révélant le degré de satisfaction des clients pour un produit donné et pour chacune de ses caractéristiques. Notre système n'est pas seulement basé sur le corpus puisque nous avons eu recours au Web 2.0 à chaque étape.

3.1 Prétraitement

Selon Liu, les commentaires des utilisateurs sont en trois formats (Liu, 2005):

- Format 1 - Pros et Cons : les consommateurs sont invités à décrire les avantages et les inconvénients séparément dans les rubriques Pros et Cons.
- Format 2 - Pros, Cons et détail : Les consommateurs décrivent les avantages et les inconvénients séparément dans les rubriques Pros et Cons et écrivent de plus des commentaires détaillés.
- Format 3 - Format libre : Les consommateurs écrivent des avis en format libre, sans séparation entre les avantages et les inconvénients.

Dans ce papier nous utilisons les critiques du troisième format. Tous les exemples qui suivent portent sur le produit *iPod* et toutes les critiques sont en anglais. Le tableau 1, ci-dessous, présente quelques exemples de commentaires des internautes.

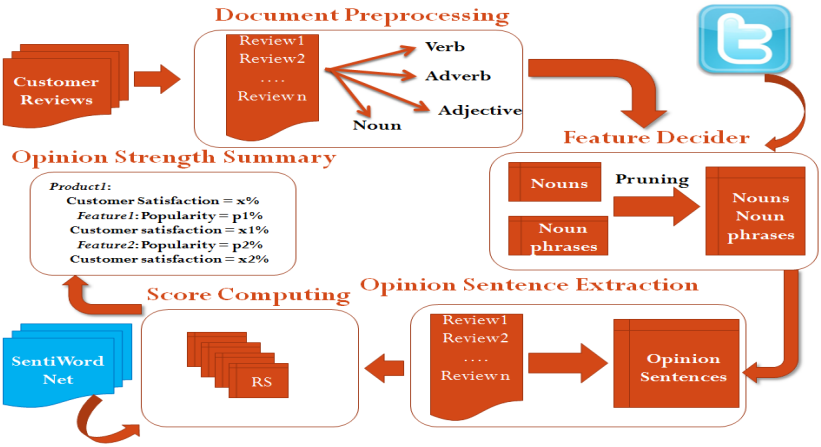


FIGURE 1 – Modèle proposé.

There isn't much features on the iPod at all, except games.
##The Click Wheel is a great design, something no one else came up with (however, the iRiver has a touchpad).

TABLE 1 – Exemples de critiques utilisateurs

Nous avons en entrée une base de données d'opinions recueillies à partir de 2 sites marchands (amazon.com et c|net.com) qui constitue notre corpus. Étant donné un nom de

produit, notre système ResTS choisit les documents correspondants dans la base de données et procède à leur segmentation en phrases. Ensuite, il les convertit en minuscule et supprime les caractères non littéraux du début et de la fin de chaque mot (par exemple « ##ipod## » devient « ipod »). Nous mettons également en relief la négation pour l'utiliser plus tard dans la phase de classification (par exemple « don't » ou « dont » devient « do not »). En outre, Hu et Liu (Hu et Liu, 2004) révèlent que les syntagmes nominaux et le substantif dans la phrase sont susceptibles d'être la caractéristique du produit sur laquelle les clients commentent. Par ailleurs, les adjectifs véhiculent l'opinion et le jugement. Nous avons donc effectué l'étiquetage de l'ensemble du corpus en utilisant TreeTagger[3] pour identifier les classes grammaticales de chaque mot.

3.2 Extraction des caractéristiques des produits

Nous avons extrait tous les syntagmes nominaux (noms) à partir des critiques des utilisateurs. Ces noms seront considérés comme des caractéristiques des produits. Notons qu'une caractéristique peut être un nom simple ou un terme composé (exemple « picture quality »).

3.2.1 Construction des termes composés

Après avoir collecté les différents noms à partir des critiques des utilisateurs, nous avons procédé à la construction des termes composés qui sont formés de deux noms successifs. Prenons un exemple : « *The Click Wheel is a great design*». « *Click Weel* » est considéré comme un terme composé. Nous avons construit de la même manière tous les termes composés mais nous n'avons gardé que ceux qui apparaissent au moins 3 fois dans le corpus.

3.2.2 Caractéristiques fréquentes

Nous avons calculé la fréquence d'apparition des différents noms dans le corpus et nous n'avons gardé que ceux dont la fréquence est supérieure à 0,01. Le Tableau 2 présente quelques résultats.

Caractéristiques	Nombre d'occurrence	Fréquence
Click wheel	9	0.07853403
Battery	30	0.2617801

TABLE 2 – Exemples de caractéristiques fréquentes

La colonne 1 présente les caractéristiques. La colonne 2 donne le nombre d'occurrences de la fonction et la colonne 3 est la fréquence des occurrences de cette caractéristique dans le corpus.

3.2.3 Popularité dans Twitter

Twitter est le service de microblogage le plus populaire. Les gens peuvent publier et lire de courts messages de 140 caractères maximum appelés tweets[4]. Les textes d'opinion suivent un style particulier (texte libre ou dialecte). On parle de nos jours de Discours Electronique Médié (DEM) qui comporte des fautes d'orthographes, des émoticônes (des smileys), des

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] Exemple de tweet : « i neeed an ipod! i have a mill at my house but of course none of them work ☹ ».

acronymes (Exemple : lol), des étirements de mots, etc. Ce type d'écriture est peu étudié par la littérature. Twitter est devenu un domaine attractif pour le traitement automatique de la langue naturelle (NLP). Dans cet article, nous montrons comment les réseaux sociaux, en particulier Twitter, peuvent être utilisés pour détecter la popularité d'un produit donné. Pour ce faire, nous commençons par l'opération de crawling. Nous cherchons seulement les tweets *populaires* parlant d'un produit donné. Notre but étant de déceler les caractéristiques populaires que les gens en montre le plus d'intérêt pour un produit donné en comptant le nombre de personnes qui s'y intéressent. Nous avons utilisé twitter4j[5], une librairie Java qui permet d'accéder au contenu de Twitter, pour recueillir près de 5000 tweets pour chaque produit posté au cours des derniers jours. Nous avons ensuite calculé le nombre de tweets évoquant chacune des caractéristiques. Le tableau 3 montre une comparaison entre le nombre d'occurrences de certaines caractéristiques dans le corpus et dans Twitter. Après avoir calculé le nombre d'occurrences de chaque caractéristique extraite dans Twitter, nous n'avons gardé que celles dont le nombre d'occurrences est supérieur à 1 ; celles qui sont mentionnées par au moins un tweet.

Caractéristiques	Occurrences corpus	Occurrences Twitter
Player	35	480
Reputation	3	0
Storage space	2	0

TABLE 3 – Nombre d'occurrences dans le corpus Vs nombre d'occurrences dans Twitter

3.3 Extraction des phrases d'opinion

L'un des objectifs de notre système est de détecter les passages subjectifs des commentaires des utilisateurs, de déterminer leur polarité et de mesurer la force de l'opinion exprimée. En utilisant la liste des caractéristiques déjà détectées, notre système ResTS a extrait toutes les phrases qui contiennent au moins une caractéristique. Voici un exemple: « *iPod is brilliant, but service was awful.* ». Cette phrase présente deux caractéristiques qui sont « *iPod* » et « *service* ». Les mots d'opinion sont « *brillant* » et « *awful* ».

3.4 Calcul des scores

Dans cette section, nous expliquons comment nous avons procédé pour mesurer l'intensité de l'opinion pour chaque caractéristique, puis pour l'ensemble du produit. Rappelons que nos scores sont compris entre 0 et 1. Le score négatif appartient à l'intervalle [0, 0.5] et le score positif appartient à l'intervalle [0,5, 1]. Pour l'identification de l'intensité de l'opinion nous adoptons l'hypothèse suivante : plus le score est proche de 0, plus le mot est négatif, et vice-versa.

3.4.1 Score d'une caractéristique

Le score d'une caractéristique est sa fréquence d'apparition dans le corpus pondérée par sa popularité sur Twitter. Nous attribuons à chaque caractéristique un score en utilisant la

[5] http://twitter4j.org/en/index.html

formule suivante[6] :

$$score_f = \alpha\, freq_f + (1-\alpha)\,\frac{nbreTweetpf}{nbreTweetp}$$

Avec : $freq_f$ est la fréquence d'apparition de la caractéristique dans le corpus, $nbreTweetpf$ est le nombre de tweets mentionnant à la fois le produit et la caractéristique, $nbreTweetp$ est le nombre total de tweets collectés pour le produit.

Ce poids mesure l'importance que les gens ont pour une caractéristique d'un produit donné. Il mesure également sa popularité. Prenons l'exemple de la caractéristique « battery », son score est égal à 0.3442 (0.6x0.543 + 0.4x0.046).

3.4.2 Opinion sur Twitter

La contrainte de taille des tweets encourage l'utilisation des émoticônes pour exprimer les opinions et les sentiments. Ces émoticônes résument souvent la polarité de toute la phrase. Nous avons construit notre propre liste d'émoticônes (voir exemples dans le tableau 4) et avons divisé les différents tweets collectés en des tweets positifs et d'autres négatifs selon la polarité de l'émoticône qu'ils contiennent.

Polarité	Emoticône
Positif	:-) :) :o) :] :3 :c) :^)
Extrêmement Positif	<=3 <=8 \o/
Négatif	--!-- :-(:(:{
Extrêmement Négatif	:-9 q(;^;)p

TABLE 4 – Exemple d'émoticônes avec polarité

Nous avons compté par la suite le nombre de tweets positifs et négatifs pour chaque caractéristique. Notre hypothèse est qu'une caractéristique doit avoir un score élevé si elle appartient plus à des tweets positifs. Donc, si une caractéristique donnée apparait plus dans des tweets positifs, on doit augmenter son score, sinon on doit le diminuer. Comme nos scores sont entre 0 et 1, nous avons choisi la racine carrée et le carré pour augmenter et diminuer le score des caractéristiques des produits comme le montre l'algorithme suivant.

Algorithm Feature_Score

Input: $score_f$, nbtweetpos, nbtweetneg //nombre de tweets positifs et négatifs

---

Begin Feature_Score

If ($score_f \geq 05$) **then**

If (nbtweetpos > nbtweetneg) ***then***

$$score_{ft} = \sqrt{score_f}$$

[6] Pour les expérimentations $\alpha = 0.6$

Else

$$score_{ft} = score_f^{\,2}$$

EndIf

ElseIf (nbtweetpos < nbtweetneg) ***then***

$$score_{ft} = \sqrt{score_f}$$

Else

$$score_{ft} = score_f^{\,2}$$

EndIf

End If

End **Feature_Score**

Output: $score_{ft}$

Prenons l'exemple de la caractéristique « battery », son score est égal à 0.3442. Comme elle apparait plus dans des tweets négatifs, on doit diminuer son score. Le score devient 0.118.

3.4.3 Score des verbes et des adjectifs

Nous avons utilisé SentiWordNet 3.0 (Baccianella et al, 2006), une ressource lexicale basée sur WordNet 3.0, dans laquelle chaque mot w de WordNet est associé à trois scores numériques ObjScore(w), PosScore(w) et NegScore(w) décrivant à quel point le mot w est objectif, positif ou négatif selon la formule suivante :

$$ObjScore(w) + PosScore(w) + NegScore(w) = 1$$

Par exemple, l'adjectif « great » a six synonymes (synset) et pour chacun un score positif et négatif. Nous ne traitons pas les verbes ou adjectifs objectifs ; ceux dont le score objectif est plus élevé que la somme de leurs scores positifs et négatifs. Étant donné un mot w, et n le nombre de ses synonymes, le score correspondant est calculé en utilisant la formule suivante:

$$score_{w_i} = \frac{\sum_{i=1}^{n} scoreS_{w_i}}{n}$$

Avec : $scoreS_{w_i}$ est le score de SentiwordNet du mot w et donné par l'algorithme suivant.

Algorithm Word_Score_Computing

Input: PosScore, NegScore //les scores de SentiWordNet

Begin Word_Score_Computing

ObjScore = 1-(NegScore + PosScore)

If (PosScore + NegScore ≥ ObjScore) then //non abjectif

If ($\text{PosScore} \geq \text{NegScore}$) ***then***

$score_{w_i}$ = PosScore

ElseIf ($\text{NegScore} \leq 0.5$) ***then***

$score_{w_i}$ = NegScore

Else

$score_{w_i}$ = 1-NegScore

EndIf

End If

End **Word_Score_Computing**

Output: $score_{w_i}$

Prenons un exemple : «*The iPod has one of the worst batteries.* ». La phrase d'opinion est «*worst batteries*». Le mot d'opinion est « *bad* ». Il a 14 synonymes dans SentiWordNet et son score calculé en utilisant l'algorithme énoncé ci-dessus est égale à 0.285.

3.4.4 Les adverbes

Les phrases d'opinions peuvent contenir des modificateurs : *intensifier* comme « Absurdly », « Acutely », « Alarmingly » ou *diminisher* comme « Moderately », « Momentarily », « Improbably » qui peut être utilisé de la même manière dans un contexte positif ou négatif comme « Absolutely great » ou « Absolutely bad ». Nous avons construit notre propre liste d'*intensifier* (192 termes) et *diminisher* (40 termes). Si une phrase contient un modificateur qui précède le verbe ou l'adjectif, nous calculons leurs scores à l'aide de l'algorithme suivant. S'il ya un intensifier précédant un verbe/adjectif positif (score $>=0.5$), nous devons augmenter son score. Cependant, s'il s'agit d'un *diminisher*, nous devons diminuer le score. Dans le cas d'un verbe/adjectif négatif (score $<0{,}5$), s'il est précédé par un *intensifier*, nous devons réduire son score, sinon, nous devons l'augmenter. Prenons un exemple : «The battery is extremely bad. ». Le score de « bad » est égal à 0.285. Comme « Extremely» est un intensifier et bad est négatif (score <0.5), le score de « extremely bad » devient 0.081(= 0.285x0.285).

Algorithm Word_Score_Computing_Modifier

Input: $score_w$, IntensifierG, DiminisherG

Begin Word_Score_Computing_Modifier

If ($score_w \geq 05$) ***then*** // the word is positive

If (Modifier ∈ IntensifierG) ***then***

$$score_w = \sqrt{score_w}$$

ElseIf (Modifier ∈ DiminisherG) ***then***

$$score_w = score_w^2$$

EndIf

ElseIf (Modifier ∈ IntensifierG) ***then***

$$score_w = score_w^2$$

ElseIf (Modifier ∈ DiminisherG) ***then***

$$score_w = \sqrt{score_w}$$

EndIf

End If

***End* Word_Score_Computing_Modifier**

Output: $score_w$

3.4.5 Score des phrases d'opinions

Le score des phrases d'opinions dépend en premier lieu des scores des verbes et des adjectifs qu'elles contiennent. Il dépond également du score de la caractéristique qu'elle contient. Si une phrase contient n caractéristiques, son score est donné par la formule suivante7 :

$$score_s = \frac{\sum_{i=1}^{n} \alpha \times score_{f_i} + (1-\alpha) \times score_{w_i}}{n}$$

Reprenons l'exemple précédent : « *The battery is extremely bad.* ». Le score de « *battery* » est égal à 0.118. Le score de toute la phrase est : 0.3x0.118 + 0.7x0.081 = 0.092. Prenons maintenant un autre exemple qui montre un score positif : « *The sound is pretty good.* ». Ici, la caractéristique est « *sound* ». Son score est 0.354. Elle apparait plus dans des tweets positifs, donc son score devient 0.595. La phrase d'opinion est « *pretty good* ». L'adjectif « *good* » a 21 synonymes. Son score est 0.595. Comme « *Pretty* » est un *intensifier*, le score devient 0.771. Le score de la phrase devient 0.718 (= 0.3x0.595 + 0.7x0.771).

3.4.6 Score du produit

Le score du produit est représenté par le score de tout le corpus relatif à ce produit. Il est donné par la formule suivante:

$$score_r = \frac{1}{n} \sum_{i=1}^{n} score_{fs_i}$$

Avec : $score_{fs}$ est le score d'une phrases d'opinions et n est le nombre de phrases

[7] Pour les expérimentations $\alpha = 0.3$

d'opinions dans le corpus.

3.5 Expérimentations

L'approche proposée a été implémentée en langage Java sous l'environnement Eclipse. Nous avons évalué notre système en utilisant plusieurs corpus de critiques des utilisateurs sur les produits suivant : deux appareils photo numériques, un téléphone cellulaire et un iPod. Ces corpus ont été collectées à partir de 2 sites marchands (Amazon.com et C|net.com) et annotées manuellement[8] par (Hu et Liu, 2004). Le premier objectif de notre système est d'extraire les caractéristiques des produits les plus proches de celles de l'annotation manuelle. Le tableau 4 résume la précision et le rappel de la phase de collecte des caractéristiques des produits. La colonne 1 présente la liste des produits utilisés pour l'évaluation. La colonne 2 donne la précision et le rappel du système de Hu et Liu. La troisième colonne indique la précision et le rappel de notre système. Nous constatons que nos résultats sont très proches de ceux de Hu et Liu ; le F-score moyen du système de Hu et Liu est 0,657, il est de 0,651 pour cette recherche.

Produit	Hu et Liu		Collecte		Collecte (utilisant Twitter)	
	Précision	Rappel	Précision	Rappel	Précision	Rappel
iPod	--	--	0.702	0.697	**0.754**	0.518
A Photo1	0.634	0.658	0.617	0.679	**0.743**	0.55
A Photo 2	0.679	0.594	**0.69**	0.58	**0.727**	0.508
Téléphone C	0.676	0.716	0.556	**0.731**	**0.725**	0.503
Moyenne	0.663	0.656	0.641	**0.671**	**0.737**	0.519

TABLE 4 – Précision et rappel de la méthode proposée Vs Hu et Liu

$$Precision = \frac{NCPR}{NC} \qquad Rappel = \frac{NCPR}{NCP} \qquad F - score = 2 * \frac{Precision * Rappel}{Precision + Rappel}$$

Avec : NC : Nombre de caractéristiques collectées par le système, NCPR : Nombre de caractéristiques pertinentes collectées par le système (qui correspondent à ceux de l'annotation manuelle), NCP : Nombre de caractéristiques de l'annotation manuelle.

L'utilisation de Twitter au cours de la phase de collecte des caractéristiques du produit a amélioré la précision, mais a causé une baisse du rappel. Ce déclin est dû à la suppression d'un certain nombre de caractéristiques qui ne sont pas populaires, c'est à dire qui n'intéressent pas la majorité des utilisateurs de Twitter.

Le deuxième objectif du système est de résumer l'opinion des utilisateurs envers un produit donné. Pour ce faire, nous avons extrait les phrases d'opinions puis calculé leurs scores. Ces scores sont corrélés à **82%** avec ceux de l'annotation manuelle.

[8] Exemple d'une phrase annotée par Hu et Liu : "battery[-2]##This is really stupid to me. 18 months for a battery isn't good," "Battery" est la caractéristique et "-2" est le score de la phrase.

3.6 Conclusion

Cet article présente une nouvelle approche de résumé automatique des textes d'opinions des critiques des utilisateurs. Notre approche vise à transformer les critiques des consommateurs en un score qui mesure l'intensité de l'opinion. Ce score est compris entre 0 et 1 et peut être utilisé pour la prise de décision et aide les utilisateurs dans leurs choix. Dans notre conception, un produit n'est pas simplement considéré comme recommandé ou non recommandé, au contraire, nous laissons l'utilisateur libre de faire son choix en fonction de certains scores que nous mettons à sa disposition traduisant la satisfaction des clients pour l'ensemble du produit et encore pour chacune de ses caractéristiques. Lors du calcul de ces scores, nous avons étudié l'opinion véhiculée par les noms, adjectifs, verbes et adverbes, contrairement aux autres recherches qui utilisent principalement les adjectifs. Nous avons de plus montré que les réseaux sociaux tel que Twitter peuvent être exploité pour mettre en évidence les caractéristiques les plus pertinentes pour l'utilisateur et de détecter leurs popularités. Dans les travaux futurs, nous prévoyons améliorer nos résultats (augmenter le rappel), éventuellement en exploitant les passages négatifs et ironiques et d'expérimenter notre méthode à l'aide d'autres entités, non seulement les produits.

Références

ANDREEVSKAIA, A. AND BERGLER, S. (2006). Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. *In Proceedings of EACL 2006*.

BLITZER, J., DREDZE, M., AND PEREIRA, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *In Proceedings of ACL 2007*.

BACCIANELLA S., ESULI A., SEBASTIANI F. (2010). SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *In Proceedings of LREC'10*.

BOUCHLEGHEM R., ELKHLIFI A., AND FAIZ R. (2010). Automatic extraction and classification approach of opinions in texts. *ISDA 2010*, IEEE Press, 918-922.

DING, X., LIU, B., AND YU, P.S. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. *In Proceedings of WSDM*, Stanford University, Stanford, California, USA.

DRAGUT, E. C., YU, C., SISTLA, P., AND MENG, W. (2010). Construction of a sentimental word dictionary. *In Proceedings of CIKM*.

ESULI A., AND SEBASTIANI, F. (2005). Determining the Semantic Orientation of Terms through Gloss Classification. *In Proceedings of CIKM*.

GAMON, M., AUE, A., CORSTON-OLIVER, S., RINGGER, E. (2005). Pulse: Mining Customer Opinions from Free Text. *In Proc. 6th Int. Symp. Advances in intelligent data analysis*, 121–132.

HU, M., LIU, B. (2004). Mining and Summarizing Customer Reviews. *In Proc. 10th Int. Conf. Knowledge Discovery and Data Mining, Seattle*, WA, 168–177.

HARRIS, Z. S. (1998). Mathematical structures of language. *Interscience tracts in pure and applied mathematics*, no.21, New York: Interscience Publishers. ix,230 p.

HATZIVASSILOGLOU, V., AND MCKEOWN, K. (1997). Predicting the Semantic Orientation of Adjectives. *In Proceedings of ACL 1997*.

KANAYAMA, K., NASUKAWA, T. (2006). Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis. *In Proceedings of EMNLP 2006*.

KAMPS, J., MARX, M., ROBERT J. M., AND RIJKE, M. (2004). Using WordNet to measure semantic orientation of adjectives. *In Proceedings of LREC 2004*.

KIM, S.M., AND HOVY, E. (2004). Determining the Sentiment of Opinions. *In Proceedings of COLING 2004.*

LEHUEDE, F. (2009). L'internet participatif redonne confiance aux consommateurs.

LIU, B., HU, M., AND CHENG, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. *In Proceedings of WWW 2005*.

LIU, B. (2007). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, *Springer 2007*, New York.

LIU, B. (2010). Invited Chapter for the *Handbook of Natural Language Processing*, Second Edition. March, 2010.

MIHALCEA, R., CORLEY, C., AND STRAPPARAVA, C. (2006). Corpus-based and knowledgebased measures of text semantic similarity. *In Proceedings of the 21st national conference on Artificial intelligence* - Volume 1, pages 775—780, AAAI Press.

PANG, B., LEE, L., VAITHYANATHAN, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proc. Conf. Empirical Methods in Natural Language Processing*, 79-86.

PEDERSEN, T., AND PATWARDHAN, S. AND MICHELIZZI, J. (2004). WordNet::Similarity: measuring the relatedness of concepts. *Association for Computational Linguistics*, 2004.

POPESCU, A. M., ETZIONI, O. (2005). Extracting Product Features and Opinions from Reviews. *In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, 339–346.

QIU, G., LIU, B., BU, J. AND CHEN, C. (2009). Expanding Domain Sentiment Lexicon through Double Propagation. *In Proceedings of IJCAI 2009*.

RILOFF, E., JANYCE, W., THERESA, W. (2003). Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *In Proc. 7th Conf. Natural Language Learning*, 25-32.

TAKAMURA, H., INUI, T., AND OKUMURA, M. (2007). Extracting Semantic Orientations of Phrases from Dictionary. *In Proceedings of HLT-NAACL*.

TURNEY, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL". *Machine Learning: ECML 2001*, pages 491–502.

WIEBE,J. (2000). Learning Subjective Adjectives from Corpora. *In Proceedings of AAAI 2000*.

WILSON, T., WIEBE, J., HWA, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. *In Proceedings of AAAI 2004*.

Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe

Ahmed Hamdi
Aix Marseille Université, LIF-CNRS, Marseille
ahmed.hamdi@lif.univ-mrs.fr

RESUME

Ce travail s'inscrit dans le cadre de l'analyse morphologique et syntaxique automatique de la langue arabe. Nous nous intéressons au traitement de la diacritisation et à son apport pour l'analyse morphologique. En effet, la plupart des analyseurs morphologiques et des étiqueteurs morphosyntaxiques existants ignorent les diacritiques présents dans le texte à analyser et commettent des erreurs qui pourraient être évitées. Dans cet article, nous proposons une méthode qui prend en considération les diacritiques lors de l'analyse, et nous montrons que cette prise en compte permet de diminuer considérablement le taux d'erreur de l'analyse morphologique selon le taux de diacritiques du texte traité.

ABSTRACT

Apport of Diacritization in Arabic Morpho-Syntactic Analysis

This work is concerned with the automatic morphological and syntactical analysis of the Arabic language. It focuses on diacritization and on its contribution to morphological analysis. Most of existing morphological analyzers and syntactical taggers do not take diacritics into account; as a consequence, they make mistakes that could have been avoided.
In this paper, we propose a method which process diacritics. We show that doing so reduces considerably the morphological error rate, depending on the diacritics rate in the input text.

MOTS-CLES : diacritisation, traitement automatique, analyse morphosyntaxique, langue arabe.
KEYWORDS : diacritization, computer processing, morpho-syntaxic analysis, Arabic language.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 247–254,
Grenoble, 4 au 8 juin 2012.

1 Introduction

En plus de sa morphologie fortement flexionnelle, dérivationnelle et agglutinante, la langue arabe se caractérise par l'absence des voyelles courtes (diacritiques) dans la plupart des textes écrits. En effet, contrairement au français, les voyelles courtes arabes ne sont pas des lettres de l'alphabet, ce sont des signes diacritiques qui se rajoutent aux consonnes (lettres) et qui jouent le même rôle que les voyelles dans les autres langues. La diacritisation en arabe est l'opération qui consiste à attribuer des diacritiques aux lettres des mots non diacrités. Cet exercice est à la fois classique et important dans le traitement automatique de l'arabe.

Généralement, les écrits en arabe sont non diacrités et c'est au lecteur de deviner les diacritiques des textes au moment de la lecture. En revanche, les textes religieux et quelques ouvrages scolaires sont entièrement diacrités. D'autres ressources, telles que les textes journalistiques, peuvent être partiellement diacrités. Les diacritiques rajoutés dans ces écrits sont utilisés pour lever des ambiguïtés morphologiques, syntaxiques et parfois sémantiques. Les diacritiques casuels, par exemple, servent à lever l'ambiguïté syntaxique. Ces diacritiques s'associent à la dernière lettre d'un mot à valeur nominale et ils marquent le cas. Ils aident à identifier les fonctions syntaxiques des mots dans une phrase. Les diacritiques affectés aux autres lettres sont appelés lexicaux, ils sont employés pour lever les ambiguïtés morphologiques et sémantiques.

On pourrait faire un parallèle entre la diacritisation de l'arabe et l'accentuation du français. Prenons le mot « presse » comme exemple, ce mot peut être reconnu comme un nom « presse » ou bien un participe passé « pressé ». La différence entre accentuation et diacritisation est que, en arabe, cette opération associe à chaque lettre d'un mot un diacritique.

Dans cet article, nous allons utiliser la convention de translittération définie par (Buckwalter, 2004), nous représentons entre crochets les caractères translittérés. La translittération est l'opération qui consiste à utiliser un autre jeu de caractères pour faciliter la lecture du lecteur francophone. Le diacritique est représenté après la consonne à laquelle il est affecté. Les diacritiques arabes sont classés en trois catégories :

- les diacritiques simples qui sont au nombre de quatre ◌َ [a], ◌ُ [u], ◌ِ [i] et ◌ْ [o], tous ces diacritiques se prononcent de la même façon que leurs translittérations sauf le dernier qui indique l'absence de tout son.
- les diacritiques doubles sont ◌ً [F], ◌ٌ [N] et ◌ٍ [K] : il s'agit de diacritiques casuels, ils produisent, respectivement le même son que les trois premières voyelles simples avec l'ajout du son « n » à la fin. Exemple : ◌ً se prononce « an ».
- le diacritique ◌ّ appelé «chadda», qui a pour effet le doublement de la lettre à laquelle il est associée.

Un mot arabe peut être non diacrité, partiellement ou entièrement diacrité. L'absence des diacritiques dans un mot provoque des difficultés dans le traitement automatique. C'est-à-dire, qu'un mot non diacrité est plus ambigu qu'un mot partiellement diacrité. D'après (Debili, 1998), 74% des mots en moyenne acceptent plus d'une diacritisation lexicale, et 89.9% des noms acceptent plus d'un diacritique casuel. La proportion des mots ambigus est de 90.5% si les comptages portent sur leurs diacritisations globales

(lexicales et casuelles).

Pour illustrer ces ambiguïtés de façon plus claire, prenons l'exemple du mot non diacrité ورد [wrd]. Ce mot peut être reconnu comme étant:

- le verbe وَرَدَ [warada] (*apparaître*), troisième personne du singulier, passé, voix active : «*est apparu*»,
- le verbe وُرِدَ [wurida] (*être apparu*), troisième personne du singulier, passé, voix passive : «*a été apparu*»,
- le verbe وَرَّدَ [war∼ada] (*fleurir*), troisième personne du singulier, passé, voix active : « *a fleuri* »,
- le verbe وُرِّدَ [wur∼ida] (*faire fleurir*), troisième personne du singulier, passé, voix passive : « *a fait fleurir* »,
- le nom وَرْد [warod] (*roses*), cette forme peut prendre cinq voyelles casuelles différentes suivant le contexte.

En comptant, aussi, les deux formes agglutinées وَ+رَدَّ [wa+rad∼a] (*et a rendu*) et وَ+رُدَّ [wa+rud∼a] (*et rends/et a été rendu*), la forme ورد [wrd] présente au total 11 diacritisations potentielles, pour 4 lemmes et 2 catégories grammaticales. Cet exemple montre bien que l'ambiguïté vocalique d'un mot produit des ambiguïtés lemmatiques et grammaticales.

Bien que les diacritiques soient destinés à lever les ambiguïtés lors d'un traitement automatique, la majorité des analyseurs morphosyntaxiques de l'arabe comme celui de Buckwalter (Buckwalter, 2004), Xerox (Beesley, 2005) ou l'analyseur MADA (Habash, 2005) n'analysent que des textes non diacrités à cause du manque de ressources arabes diacritées. Par conséquent, si l'entrée est partiellement diacritée, ces analyseurs commencent par éliminer tous les diacritiques, puis ils font l'analyse comme si l'entrée était non diacritée. Les analyseurs morphosyntaxiques de l'arabe ne profitent donc pas des diacritiques présents dans les textes pour désambiguïser les mots. Dans le cadre du travail présenté ici, nous proposons une méthode qui permet de prendre en compte ces diacritiques. Leur prise en compte améliorera naturellement la diacritisation automatique. Nous nous intéressons ici à étudier l'apport de ces diacritiques sur les autres niveaux d'analyse : l'étiquetage grammatical et l'analyse morphologique.

Dans la section 2 de cet article, nous présentons la méthode proposée, ainsi que l'analyseur auquel nous avons intégré notre proposition. Dans la section 3, nous décrivons les expérimentations réalisées pour évaluer notre travail, et nous donnons enfin les résultats de l'analyse morphologique avant et après l'ajout de notre solution.

2 Description de la méthode

Afin d'étudier d'une façon concrète l'influence des diacritiques sur l'analyse morphologique, nous avons introduit un ensemble de modules dans l'analyseur MADA. Nous commençons par présenter cet analyseur, en nous focalisant sur les erreurs provoquées par la non prise en compte des diacritiques lors de l'analyse. Enfin, nous décrivons en détail la méthode avec laquelle nous visons à améliorer les performances de MADA.

2.1 L'analyseur morphosyntaxique MADA

MADA (*Morphological Analyzer and Disambiguator of Arabic*) (Habash, 2005) est un analyseur morphologique de l'arabe. Cet analyseur réalise la segmentation, la diacritisation, la lemmatisation, l'étiquetage grammatical et l'analyse morphologique.

Les données d'apprentissage de MADA proviennent du corpus *the Penn Arabic Treebank* PATB (Maâmouri, 2004), le corpus d'apprentissage contient 120 000 mots alors que 12 000 mots ont été utilisés pour l'évaluation. Nous présentons les résultats de l'évaluation dans la section suivante.

Lors de l'analyse d'un texte, MADA produit pour chaque mot toutes ses analyses possibles, ensuite, le modèle SVM (*Support Vector Machines*) est utilisé pour générer une prédiction de quelques traits morphologiques. Enfin, MADA fait la hiérarchisation des analyses retournées, la meilleure analyse étant celle qui s'accorde le plus avec la prédiction.

Comme nous l'avons évoqué ci-dessus, l'un des inconvénients de cet analyseur est qu'il ne prend pas en considération les diacritiques de l'entrée, et peut donc produire des analyses incompatibles avec l'entrée. Une entrée E est non compatible avec une analyse A de MADA, si les diacritiques de E ne sont pas tous présents dans la diacritisation de A. Prenons comme exemple le mot partiellement diacrité كتبت [ktb**a**t], ce mot possède un diacritique [a] associé à la troisième lettre du mot [ktbt]. Les trois premières diacritisastions renvoyées par MADA sont respectivement :

1. كَتَبْتُ [katab**o**tu] (*j'ai écrit*)
2. كَتَبْتَ [katab**o**ta] (*tu as écrit*)
3. كَتَبَتْ [katab**a**t] (*elle a écrit*)

Le diacritique affecté à la troisième lettre de l'entrée dans les deux premières analyses retournées n'est pas identique avec l'entrée. Ainsi, ces deux analyses sont considérées comme incompatibles avec l'entrée. Elles ont entrainé des erreurs qui auraient pu être évitées au niveau de la diacritisation et aussi dans d'autres traits morphologiques tels que le genre et la personne.

2.2 Méthode proposée

Pour remédier à ce problème d'incompatibilité d'analyses, la méthode que nous proposons consiste à restreindre l'ensemble des analyses des mots retournées par MADA à celles qui contiennent des diacritisations compatibles avec les mots diacrités passés en entrée. Pour ce faire, nous avons eu recours aux automates à états finis. Les mots comportant éventuellement des diacritiques, qui sont fournis à MADA ainsi que les sorties de MADA sont représentés sous la forme d'automates. Ce mode de représentation va permettre de réaliser les tests de compatibilité entre l'entrée et la sortie grâce à des opérations standard sur les automates.

Dans un premier temps, nous représentons l'entrée par un automate à états finis A1. Chaque lettre et chaque diacritique correspond à une transition, la transition du diacritique vient juste après celle de la lettre qui lui a été associée. La figure suivante donne un exemple de représentation d'un mot avec un automate à travers le mot كتبت

[ktb**a**t].

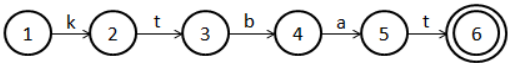


FIGURE 1 – Représentation du mot partiellement diacrité [ktbat] par A1

De la même manière, les diacritisations produites par MADA sont représentées par des automates, sauf qu'on a ajouté une transition vide (ε-transition) à chaque transition qui représente un diacritique. Deux cas sont envisageables, selon que la diacritisation est compatible ou non avec le mot en entrée. Reprenons l'exemple de la section 2.1, la figure 2 présente l'automate de la première diacritisation retournée par l'analyseur كَتَبْتُ [katab**o**tu], alors que l'automate qui correspond à la troisième diacritisation كَتَبَتْ [katab**a**t] est présenté dans la figure 3.

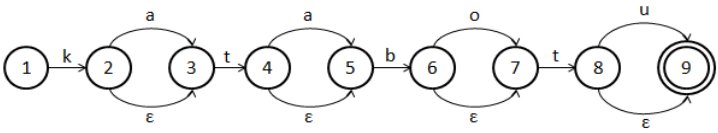


FIGURE 2 – Représentation de la diacritisation [katabotu] par A2

Le choix de l'analyse à retenir passe par la vérification de la compatibilité de l'automate A1 avec les automates A2 et A3. Deux automates sont compatibles si leur intersection est non nulle, c'est-à-dire, s'il existe un chemin commun entre eux de l'état initial à l'état final. Par conséquent, nous pouvons constater que A1 n'est pas compatible avec A2, donc, l'analyse qui contient cette diacritisation devrait être rejetée. En revanche, nous gardons la troisième analyse puisqu'elle contient une diacritisation qui s'accorde avec l'entrée.

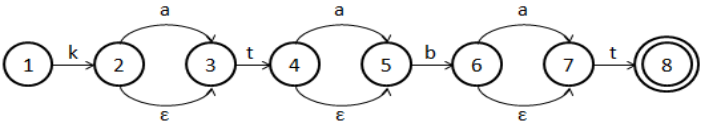


FIGURE 3 – Représentation de la diacritisation [katabat] par A3

3 Expérimentation et évaluation

Pour évaluer les prédictions de MADA avant et après l'ajout de l'option d'analyse des textes diacrités, nous avons eu recours à un corpus de test partiellement diacrité (1.3% de diacritiques) qui contient 25 295 mots préalablement annotés à la main, c'est-à-dire qu'à chaque mot, lui est attribué son analyse morphologique correcte dans le contexte où le mot apparait. Ces analyses contiennent les diacritisations, les étiquettes grammaticales et d'autres valeurs des différents traits morphologiques.

Afin de construire d'autres ressources textuelles partiellement diacritées, nous sommes partis du corpus de test entièrement diacrité, et nous l'avons dépourvu, aléatoirement, d'un taux variable de diacritiques. De cette manière, 10 corpus de test ont été obtenus, ils contiennent un pourcentage de diacritiques qui varie entre 10% et 100%.

Rappelons que MADA ne prend pas en considération les diacritiques présents dans le corpus en entrée, il produit alors les mêmes résultats d'analyse pour chacun des corpus de test. Ses performances au niveau de la diacritisation, de l'étiquetage grammatical et de l'analyse morphologique sont présentées dans le tableau suivant :

Critère	Diacritisation	Etiquetage grammatical	Analyse morphologique
Performance	86.38%	96.09%	84.25%

TABLE 1 – Performances de MADA sur notre corpus de test

Une analyse morphologique est estimée correcte, si toutes les valeurs prédites des traits morphologiques sont conformes avec l'analyse annotée dans le corpus de référence. Ainsi, l'analyseur a produit environ 86% de bonnes diacritisations pour sa meilleure analyse, 96% des catégories grammaticales correctes et 84% de bonnes analyses.

Les tests ont été réalisés, aussi, avec la prise en compte des diacritiques. Tel que nous en faisons l'hypothèse, l'expérience a montré que plus le corpus contient des diacritiques, plus les performances de MADA devraient s'améliorer (cf. table 2).

Taux de diacritisation	Performances MADA		
	Diacritisation	Etiquetage grammatical	Analyse morphologique
1.3%	86.97%	96.41%	84.91%
10%	88.47%	96.79%	86.28%
40%	91.74%	97.12%	89.48%
70%	94.85%	97.33%	92.51%
100%	98.01%	97.49%	95.59%

TABLE 2 – Performances de MADA dans l'analyse des corpus diacrités

Le tableau 2 illustre l'apport de la diacritisation dans l'analyse morphologique de l'arabe. En effet, les performances de l'analyse morphologique passent de 84.25% à 95.59% si MADA prend en considération les diacritiques présents dans les textes entièrement diacrités. L'amélioration est significative, également, dans la diacritisation et l'étiquetage grammatical. Nous remarquons, aussi, que même si le texte est entièrement diacrité, les

performances de la diacritisation de MADA n'atteignent pas 100% et s'arrêtent au niveau de 98%. Cela est dû aux mots non reconnus par MADA.

La courbe ci-dessous décrit l'évolution des performances de MADA en fonction de la proportion des diacritiques présents dans le corpus de test.

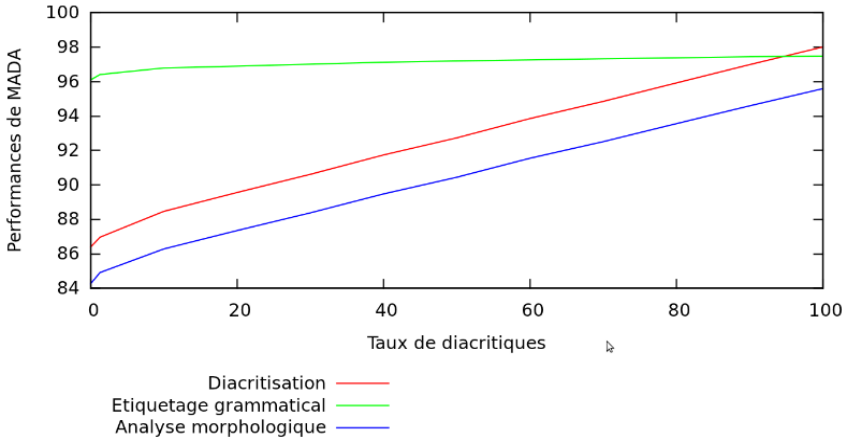


FIGURE 4 – Evolution des prformances de MADA

Ces courbes montrent l'apport des diacritiques dans la désambiguïsation morphologique. Plus il y en a dans l'entrée, plus les performances sont élevées. Nous remarquons, également, que chaque courbe présente deux pentes, avec une pente maximale lorsque le taux de diacritiques de l'entrée est entre 0% et 1.3%. Cela vient du fait que les diacritiques rajoutés à l'entrée sont « naturels », ajoutés par des humains pour désambiguïser des mots considérés comme difficiles à comprendre par le lecteur quand ils sont sous la forme non diacritée. Il est donc normal que l'apport de ces diacritiques soit important. La deuxième pente est constante puisque les diacritiques sont rajoutés d'une façon aléatoire.

4 Conclusion

Dans toute analyse linguistique, la détermination des traits morphologiques d'un mot dans son contexte constitue une étape importante. En arabe, cette détermination est rendue plus difficile par le fait que la majeure partie des mots de la langue sont ambigus. En effet, l'absence des diacritiques en arabe écrit rend les niveaux d'ambiguïté très élevés par rapport aux autres langues. Avec ce travail, nous avons voulu tester l'influence des diacritiques sur les performances pour l'analyse morphologique, l'étiquetage grammatical et la diacritisation automatiques de l'arabe. Les résultats obtenus prouvent que la prise en compte des diacritiques présents dans les textes améliore considérablement toutes ces analyses.

Cette étude pourrait être étendue à d'autres niveaux d'analyse automatique, à savoir syntaxique et sémantique. En effet, seules les diacritiques permettent de distinguer le sujet et l'objet dans ces deux phrases verbales[1] اصطحب الرجلُ الولدَ (*l'homme a accompagné l'enfant*) et اصطحب الرجلَ الولدُ (*l'enfant a accompagné l'homme*). Nous poursuivons nos travaux de thèse dans cette perspective, afin de mesurer l'influence de la diacritisation dans l'analyse syntaxique de l'arabe.

Remerciements

Je tiens à exprimer ma reconnaissance la plus sincère à mes encadrants Mme Nuria GALA et M. Alexis NASR, pour tout le temps qu'ils m'ont consacré, leur directives précieuses et leurs suivis réguliers.

Mes plus vifs remerciements s'adressent aussi à mes collègues de l'équipe TALEP pour leurs sympathies et leur soutien.

Références

BEESLEY, R. (2005). Xerox Arabic Morphological Analysis and Generation Romanization, Transcription and Transliteration.

BUCKWALTER, T. (2004). Buckwalter Arabic Morphological Analyser Version 2.0. *Linguistic Data Consortium (LDC) Catalog Number LDC2004L02, ISBN 1-58563-324-0*.

DEBILI, F. et ACHOUR, H. (1998). Voyellation automatique de l'arabe. *Actes du Workshop on Computational Approaches To Semitic Languages,* Université de Montréal.

DEBILI, F. et SOUISSI, E. (1998). Etiquetage grammatical de l'arabe voyellé ou non. *In Proceedings of the Workshop on Computational Approaches to Semetic Languages, Stroudsburg*.

DEBILI, F., ACHOUR, H. et SOUISSI, E. (2002). La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique. *Correspondances de l'IRMC, N°71, Tunis*.

HABASH, N. et OWEN, R. (2005). Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. *In Proceedings of the Conference of the American Association for Computational Linguistics*, New York.

HAJIC, J., SMRZ, F., BUCKWALTER, T. et JIN, H. (2005). Feature-Based Tagger of Approximations of Functional Arabic Morphology. *Actes de la quatrième conférence sur les Treebanks et les théories linguistiques,* Université de Barcelone.

MAAMOURI, M., BIES, A. et BUCKWALTER, T. (2004). The Pen Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. *In EMAR Conference on Arabic Language Ressources and Tools,* le Caire.

[1] Une phrase verbale dans la langue arabe est une phrase qui contient un verbe.

Pour un étiquetage automatique des séquences verbales figées : état de l'art et approche transformationnelle

Aurélie JOSEPH
LDI, 99 avenue Jean-Baptiste Clément, 93 Villetaneuse
ITESFOT, Parc d'Andron, Le Séquoia, 30470 Aimargues
joseph.aurelie@gmail.com

RESUME

Cet article présente une approche permettant de reconnaitre automatiquement dans un texte des séquences verbales figées (*casser sa pipe, briser la glace, prendre en compte*) à partir d'une ressource. Cette ressource décrit chaque séquence en termes de possibilités et de restrictions transformationnelles. En effet, les séquences figées ne le sont pas complètement et nécessitent une description exhaustive afin de ne pas extraire seulement les formes canoniques. Dans un premier temps nous aborderons les approches traditionnelles permettant d'extraire des séquences phraséologiques. Par la suite, nous expliquerons comment est constituée notre ressource et comment celle-ci est utilisée pour un traitement automatique.

ABSTRACT

For an Automatic Fixed Verbal Sequence Tagging: State of the Art and Transformational Approach

This article presents a resource-based method aiming at automatically recognizing fixed verbal sequences in French (i.e *casser sa pipe, briser la glace, prendre en compte*) inside a text. This resource describes each sequence from the view-point of transformational possibilities and restrictions. Fixed sequences are not totally fixed and an exhaustive description is necessary to not only extract canonical forms. We will first describe some transformational approaches that are able to extract phraseological sequences. The building of the resource will be then addressed followed by our approach to automatically recognize fixed sequences in corpora.

MOTS-CLES : séquences verbales figées, reconnaissance automatique, étiquetage, transformations linguistiques, ressources électroniques.

KEYWORDS: fixed verbal sequences, automatic recognition, tagging, linguistical transformations, electronic resources.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 255–266,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le découpage en mots est la première opération effectuée dans un traitement automatique de la langue. Mais le terme mot est linguistiquement inapproprié car il correspond en informatique à une entité, appelée token, délimitée par des séparateurs graphiques (blancs, retour à la ligne…). Il n'est pas nécessaire de rappeler ici que la notion de mot est beaucoup plus complexe. Et lorsque nous disons complexe nous pensons tout autant à la difficulté de les cerner qu'à leur potentielle polylexicalité. En effet, alors que les traitements informatiques se sont concentrés sur le mot simple, les chercheurs ont prouvé que les mots complexes sont tout aussi importants dans le traitement des langues (Gross, 1982 ; Gross, 1996 ; Mejri, 1997 ; Lamiroy, 2006…). Le traitement automatique de ces séquences devient un enjeu incontournable et se doit d'être traité correctement car la bonne identification et donc l'étiquetage correct de ces séquences dites figées est utile pour de nombreuses applications telles que la traduction, l'extraction d'information, la classification... Notre approche consiste à créer une ressource électronique décrivant les séquences verbales figées (SVF) en termes de possibilités transformationnelles. Nous pensons, que comme les termes simples, elles ont besoin d'être, dans un premier temps, reconnaissables sous toutes leurs formes. Comme un verbe peut être reconnu dans un texte même s'il est conjugué, les SVF doivent être également reconnues malgré leurs modifications. Inversement, pour les séquences avec un dédoublement de sens (Mejri, 2003), les contraintes liées à certaines transformations peuvent nous diriger sur une séquence littérale et non globale (*prendre la veste verte* sera reconnu comme non figé).

2 Séquence figée et extraction automatique : état de l'art

Notre objet d'étude concerne la séquence figée (SF) connue aussi sous le nom d'expression figée, locution, expression idiomatique... Une séquence figée est un groupe de mots non nécessairement contigus, possédant une unité sémantique (sens global), et un figement à la fois morphologique (blocage du nombre), lexical (blocage du paradigme commutationnel) et syntaxique (blocage de la passivation, de la relativation… pour des séquences verbales) (Lamiroy et al., 2008).

A l'instar de Mejri (2011) nous distinguons les séquences figées de deux autres concepts :

- les séquences totalement figées (*au fur et à mesure*) qui n'acceptent aucune modification. L'ensemble est un bloc immuable et dont le traitement nécessite un simple référencement dans un dictionnaire.
- Les collocations : séquences répétées apparaissant fréquemment ensembles (Firth, 1957). Elles peuvent être propres à un domaine (collocation terminologique selon Smadja, 1993) ou typique d'une langue (comme les verbes supports ou les verbes appropriés)

Nous étudions ici, plus précisément la séquence verbale figée (*casser sa pipe, prendre le taureau par les cornes, faire faux bond*…). La problématique des SF et de façon plus importante des SVF, vient du fait qu'elles ne sont pas totalement figées (Gross, 1982 ; Gross, 1996 ; Lamiroy, 2005 ; Abeillé, 1989). En effet, elles autorisent certaines modifications d'ordre syntagmatique et/ou paradigmatique créant ainsi des degrés de

figement (Gross, 1996). Cependant, il n'est apparemment pas possible de définir a priori les transformations réalisables d'une séquence. Villada-Moiròn (2005) remarque que "*there is no uniform presence or absence of syntactic constraints in all fixed expressions since not all fixed expressions exhibit the same syntactic versatility*".[1] (Villada Moiròn, 2005:46). Balibar-Mrabti (2005) va plus loin en postulant que des séquences de même structure syntaxique n'acceptent pas les mêmes libertés transformationnelles (*bruler ses vaisseaux* vs *casser sa pipe*). Comment peut-on alors extraire de telles séquences malgré des modifications évidentes ? Il existe plusieurs approches assez répandues.

La première est purement syntaxique. Laporte et al. (2008) utilisent les patrons syntaxiques productifs en noms composés et vont les proposer à un transducteur (avec l'outil Unitex). En permettant certaines transformations (insertion, coordination…) ils récupèrent ainsi des séquences nominales correspondant syntaxiquement à des noms composés.

La deuxième approche est purement statistique. Ces méthodes utilisent une mesure pour déterminer au mieux la cohésion entre les éléments. Dias (2003) propose ainsi le GenLocalMax qui permet de calculer le degré de figement d'une séquence plus grande que deux mots non obligatoirement contiguë. Cependant l'approche est largement dépendante du corpus et de sa taille.

L'approche la plus utilisée conjugue à la fois syntaxe et statistique. Certains chercheurs (Manning & Schütze, 1999 ; Daille, 1996 ; Watrin, 2008…) commencent par un filtrage linguistique (sélection de lexèmes, patrons syntaxiques) pour ensuite prendre une décision basée sur un calcul probabiliste (*logarithme de vraisemblance, information mutuelle*…). D'autres à l'inverse, génèrent un premier filtrage par critères statistiques pour ensuite effectuer leur choix sur critères linguistiques (Smadja, 1993). Ces méthodes hybrides sont les plus prisées. Toutefois, elles permettent l'extraction de données terminologiques (souvent nominales) plus que l'extraction de séquences figées que nous pourrions appeler 'langagières', c'est-à-dire, qui peuvent se retrouver dans n'importe quel texte quel que soit le domaine. Les possibles modifications intégrées sont de l'ordre de l'expansion de séquence. Peu d'études (Al Haj et Wintner, 2010) testent les transformations morphologiques, lexicales ou même syntaxiques que peut subir une SF afin d'en calculer le degré de figement.

Une autre approche, permettant d'extraire des unités phraséologiques, est basée sur des dictionnaires électroniques : les travaux du LDI (notamment ceux de Ben-Henia Ayat, 2006, 2009 ; Buvet, 2008 ; Cartier, 2008) ou du Lexique-Grammaire (notamment Tolone, 2011). Les dictionnaires électroniques du LDI décrivent les emplois des termes de manière syntactico-sémantique. Les éléments sont catégorisés en prédicat, argument, actualisateur et leur comportement syntaxique lié au sens est déterminé par ces mêmes notions : *prendre(HUM,taureau,corne)*. Les séquences polylexicales sont également

[1] « *observationnel il n'y aucune uniformité dans la présence ou dans l'absence de contraintes syntaxiques dans toutes les expressions figées puisque toutes les expressions figées ne présentent pas la même polyvalence syntaxique* » (Villada Moiròn 2005:46).

traitées de la sorte : *prendre le taureau par les cornes(HUM)*. Cependant leur description est souvent liée à la syntaxe externe (les arguments qu'elles acceptent) et leur traitement interne n'est pas exhaustif. Une description des emplois en termes de prédicat et d'argument peut se révéler essentielle pour extraire, par exemple, des SF analytiquement fausses c'est-à-dire dont le rapprochement syntactico-sémantique n'est pas logiquement correct (*avoir un chat dans la gorge :* un humain ne peut pas avoir littéralement un chat dans la gorge) (Ben-Henia Ayat, 2009) ou pour les désambiguïser d'une séquence homonyme dont le sens est littéral (*prendre une veste*). Cependant, cette description s'avère coûteuse en description sous-jacente car le traitement d'un texte doit être très fin pour pouvoir être analysé.

Abeillé et Schabes (1989) proposent, grâce aux grammaires d'arbres adjoints une méthode pouvant extraire les SF malgré leur discontinuité (insertion, modifieur) et leurs potentiels changements syntaxiques (passivation, clivage…). Cela implique toutefois que la description transformationnelle soit complète.

Finalement, les approches hybrides sont assez rapides et nécessitent peu de ressources et de prétraitement. Toutefois elles incluent dans leur extraction toutes sortes d'unités phraséologiques (souvent des collocations terminologiques). De plus, elles ne prennent pas en compte toutes les possibilités transformationnelles d'une séquence en se limitant souvent à de simples expansions. Les dictionnaires électroniques, plus exhaustifs et précis sont néanmoins coûteux en réalisation et en prétraitement. De plus, même si les chercheurs décrivent la séquence figée comme ayant une double structuration, la description systématique de la structuration interne n'est pas détaillée.

Nous voulons constituer une ressource électronique répertoriant chaque SVF associée à toutes les transformations qu'elle accepte. Cette ressource sera utilisée dans un outil et permettra de reconnaitre toutes les SVF malgré leurs variations possibles.

3 Création de la ressource

3.1 Les transformations

Les transformations 'bloquées' des SVF ont été décrites dans la littérature (notamment Gross, 1982 ; Gross, 1996 ; Mejri, 1997 ; Lamiroy et al., 2006…). Nous l'avons dit précédemment, certaines de ces transformations peuvent être réalisées dans certaines SVF mais elles ne sont pas déterminables a priori. Nous reprenons donc chacune d'elles afin de transformer automatiquement chaque séquence[2] (Cartier et Joseph, 2011). Cette méthode s'apparente à l'utilisation de grammaire d'arbres adjoints (Abeillé, 1989) sans pour autant aspirer à des phrases toujours grammaticales.

- Conjugaison : le verbe est mis à toutes les personnes au présent, imparfait et futur.
- Flexion : chaque séquence est modifiée en changeant le nombre des noms et de leurs actualisateurs associés. Toutes les possibilités sont prises en compte. S'il y a

[2] Les séquences étudiées ont été récupérées dans diverses ressources et correspondent à des séquences particulières (cf Cartier et Joseph 2011)

deux noms dans la séquence nous aurons donc 4 possibilités (*prendre le taureau par les cornes, prendre les taureaux par les cornes, prendre le taureau par la corne, prendre les taureaux par la corne*).

- Substitutions : les noms, les adjectifs, les verbes, les adverbes sont substitués par leurs synonymes et antonymes (pour les adjectifs et verbes) les plus récurrents. Les déterminants sont substitués par d'autres déterminants (*indéfini, défini, possessif, démonstratif*).
- Modifieurs nominaux : seuls les compléments du nom et les relatives peuvent être modélisées pour être requêtés sur des données non étiquetées. (l'adjectif sera un simple mot joker).
- Suppression d'éléments : les adjectifs (*prêter main forte* ➔ *prêter main*), les adverbes, les déterminants, le verbe 'introducteur' sont tour à tour supprimés. Lorsqu'une séquence possède plusieurs syntagmes alors on supprime tour à tour les syntagmes (*prendre le taureau par les cornes* ➔ *prendre le taureau, prendre par les cornes*).
- Insertion : on teste la possibilité d'insérer un déterminant entre le verbe et le nom ou la préposition et le nom s'il n'existe pas (*prendre peur* ➔ *prendre la peur*).
- Négation / affirmation : les séquences affirmatives sont mises au négatif et inversement.
- Inversion : les syntagmes d'une séquence sont inversés s'ils sont au moins deux (*prendre le taureau par les cornes* ➔ *prendre par les cornes le taureau*)
- Passivation : la séquence est modifiée en une phrase au passif (*le taureau est pris par les cornes, le taureau par les cornes est pris*)
- Relativation : la séquence est modifiée en une relative (*le taureau que je prends par les cornes*).
- Clivage : chaque syntagme est clivé (*c'est le taureau que je prends par les cornes, c'est par les cornes que je prends le taureau*).
- Pronominalisation/Détachement : test non effectué actuellement automatiquement.
- Les questions : test non effectué actuellement automatiquement.

3.2 Aide à la création de la ressource

Afin que la décision sur la possibilité transformationnelle ne soit pas uniquement liée à l'intuition du linguiste, un programme (Cartier et Joseph, 2011) va permettre de soumettre chaque séquence transformée en tant que requête à un moteur de recherche (Google, Google Books, Google News). Ce moteur de recherche va nous retourner le nombre de liens trouvés dans le web. Le linguiste pourra finalement valider les possibles transformations en s'appuyant à la fois sur les résultats et les attestations retournées par les moteurs de recherche en vérifiant que la transformation effectuée est bien relative au sens de la séquence figée. Le linguiste reste malgré tout le garant de la validité des transformations en ayant une aide sur l'usage réel de la séquence.

Nous obtenons ainsi une ressource décrivant pour chaque SF les transformations qu'elle peut subir et par conséquent celles qui sont bloquées. Actuellement notre base se restreint à environ 500 séquences transformées et validées. Chaque séquence génère en moyenne 40 transformations. Le nombre de transformations réalisées dépend du nombre

de constituants. Précisons également que les trois quart des SVF de notre base correspondent à la structure 'verbe déterminant nom préposition déterminant nom'. Ce choix totalement arbitraire, avait pour but de trouver des règles transformationnelles potentielles malgré les postulats de départ.

Le peu de données actuelles est essentiellement dû au fait que le programme de transformation a nécessité une attention toute particulière afin de garantir sa robustesse pour qu'il puisse traiter toutes sortes de structures syntaxiques.

Voici comment se présentent les transformations d'une SVF, dans un premier temps celles qui affectent les constituants (Table 1), dans un deuxième temps celles qui affectent la séquence entière (Table 2).

WORD	MODIFIEUR	CONJUGAISON/FLEXION	SUBSTITUTION
prendre	False	True	saisir
taureau	True	True	boeuf
le	un Adj0	False	False
cornes	False	False	False
par	False	False	False
les	False	False	False

TABLE 1 – Transformations des constituants de la séquence *prendre le taureau par les cornes*

	LOCUTION	passivation_1	negation_neg	affirmation_aff	inversion_syntagme_021
▶	prendre le taureau par le cornes	TRUE	TRUE	TRUE	N1+modif

TABLE 2 – Transformations possibles de la séquence *prendre le taureau par les cornes*

Cette ressource nous permet de connaître sous formes de traits définitoires ce qui caractérise les séquences en termes de transformations possibles. De plus, il peut exister des liens entre les transformations. En effet, un déterminant peut se voir substituer si et seulement si le nom qu'il actualise est modifié. C'est le cas de *le* qui devient *un* si *taureau* est modifié par un adjectif placé avant lui (noté Adj0). Il est donc primordial de le répertorier.

4 Outil de reconnaissance des SVF

Nous proposons une méthode, qui devra être améliorée par la suite, permettant de reconnaitre automatiquement des SVF même si celles-ci ont été transformées et de les distinguer partiellement de séquences littérales homographes. Finalement nous pourrons les relier à leur forme canonique créant ainsi une lemmatisation de la SVF.

4.1 Corpus

Nous testons actuellement notre outil sur une base constituée des premières pages retournées par le moteur de recherche lors de nos requêtes sur les différentes transformations de *prendre le taureau par les cornes.* Dans l'état actuel des choses, la base

n'est utile que pour tester visuellement notre outil. Un corpus servant de référence, permettant de tester nos résultats doit être réalisé au plus tôt. Notre corpus actuel, de plus de 32000 mots, est étiqueté morpho-syntaxiquement et lemmatisé par Treetagger à l'exception des noms. En effet, une séquence figée se caractérise très souvent par un blocage flexionnel des noms (*singulier/pluriel*). C'est pour cette raison que nous ne lemmatisons pas les noms. Toutefois, les séquences ayant des noms acceptant les versions singulier et pluriel seront indiquées dans la ressource et ne seront bien évidemment pas écartées de l'analyse.

4.2 Étapes de reconnaissance

Notre programme procède en plusieurs étapes. La première consiste à constituer un dictionnaire d'entrées des composants de la SVF (noms, verbes, adjectifs, adverbes), qui n'acceptent ni substitution ni suppression. Nous les appelons 'invariants'. Ce terme utilisé de manière un peu abusive représente les termes obligatoirement présents dans la séquence mais peuvent toutefois varier en nombre. Ils peuvent s'apparenter à la « zone fixe » de Laporte (1988) correspondant à un ou plusieurs termes obligatoirement présents dans la séquence. Les 'invariants' se rapprochent de la définition de « tête » d'Abeillé (1989) qui correspond à un terme simple déclenchant la zone de recherche d'une potentielle SF.

Locution	Invariant		
mettre la tete a le carre	tete	carre	
mettre le doigt sur la bouche	doigt	bouche	
mettre le doigt sur la plaie	doigt	plaie	
mettre le feu a les poudres	feu	poudres	
mettre le nez a la fenetre	nez	fenetre	
mettre les pieds dans le plat	pieds	plat	
montrer le bout de le oreille	montrer	bout	oreille
occuper le devant de la scene	devant	scene	
prendre la balle a le bond	balle	bond	
prendre la cle de les champs	cle	champs	
prendre la main dans le sac	main	sac	
prendre le air de le bureau	air	bureau	
prendre le taureau par les cornes	cornes		
rater une vache dans un couloir	rater	vache	couloir

TABLE 3 – Échantillon de SVF associées à leurs invariants

Les 'invariants' sont utilisés comme des déclencheurs d'une potentielle SVF. En effet, le texte étudié est découpé en tokens et chaque forme différente (nom, adjectif, adverbe) est recherchée comme un possible invariant. Les SVF associées à cet invariant sont donc récupérées et deviennent les séquences candidates à évaluer de manière plus approfondie. Autour de cet invariant une fenêtre de recherche (que nous appellerons

'capture') est récupérée. Elle correspond à 10 mots de part et d'autre de l'invariant. Ce nombre a été choisi arbitrairement. Il sera par la suite réévalué, pouvant même dépendre du nombre de constituants de la séquence. Il s'en suit à partir de cette capture, une succession de tests, susceptibles d'éliminer la capture et par la même occasion la potentielle SVF associée. Ces tests concernent dans un premier temps, les constituants de la séquence :

- les autres 'invariants' : tous les éléments obligatoires dans la séquence doivent être retrouvés dans la capture.
- les éléments 'variants' : substituables ou supprimables.
  Si l'élément est supprimable sa présence n'est pas requise. Un élément supprimable peut être également le fait d'un syntagme supprimable (*avoir des fourmis dans les jambes, avoir des fourmis*).
  Un élément de la séquence peut être substitué. Dans ce cas les substitutions possibles sont listées sous formes de lemmes ou de classes d'objets (*avoir des fourmis dans <PARTIE DU CORPS>*)
- les modifieurs possibles ou impossibles : les compléments du nom, les adjectifs, les subordonnées relatives sont actuellement les trois modifieurs que nous recherchons dans la capture à partir d'un nom. Si ce nom ne doit pas être modifié et que des éléments représentants les modifieurs sont trouvés (*de, une étiquette <adjectif>, un relatif*) alors la séquence est éliminée.
- les modifieurs possibles selon une contrainte particulière (substitution du déterminant : *prendre une veste, prendre la veste de sa vie*).

Dans un deuxième temps, nous testons les transformations liées à la séquence entière (inversion, passivation, clivage, relativation). Pour ce faire, nous testons tout d'abord l'ordre dans lequel apparaissent les constituants. En effet, ces types de transformations impliquent un changement syntagmatique des éléments. Ainsi, en disant que *prendre le taureau par les cornes* possède 3 composants (*prendre, taureau, cornes*), ces composants constituent l'ordre suivant : 1 2 3. Le clivage de *taureau* modifie alors l'ordre en 2 1 3 (*c'est le taureau qu'il prend par les cornes*). En procédant de cette manière nous validons l'ordre des éléments par rapport à la transformation associée. Cependant, ceci n'est pas suffisant pour savoir si nous traitons bien la transformation cible. Il nous faut ainsi des éléments définitoires de chaque transformation devant être présents dans la capture pour que la transformation soit validée. Ainsi le passif est défini par la présence d'un verbe au participe passé avec l'auxiliaire *être*, la relativation par un pronom relatif et le clivage par la présence de *c'est* et un pronom relatif (Riegel et al., 1994).

Prenons un exemple : la ressource nous renseigne sur le fait qu'une passivation est acceptée (l'indice 1 indique que c'est une passivation de type : *le taureau est pris par les cornes*. Pour affirmer que nous avons ce passif, nous devons donc trouver l'ordre 2 1 3 mais également avoir un participe passé avec l'auxiliaire *être* (ou sans l'auxiliaire selon certaines conditions).

Précisons également que des interdépendances peuvent survenir et qu'il faut les prendre en compte. C'est le cas dans *prendre le taureau par les cornes* où lors d'une inversion *le taureau* doit être modifié par un complément du nom pour que l'inversion soit acceptée (*prendre par les cornes le taureau de la fantaisie à la française* dans *actualites.leparisien.fr*).

4.3 Prenons le taureau par les cornes : exemple

L'outil a été testé sur notre corpus constitué, comme nous l'avons dit précédemment, de différentes phrases incluant *prendre le taureau par les cornes* sous différentes formes (les différentes transformations). Les séquences libres et les séquences figées cohabitent donc dans ce corpus. La figure 1 illustre une partie des résultats retournés par l'outil.

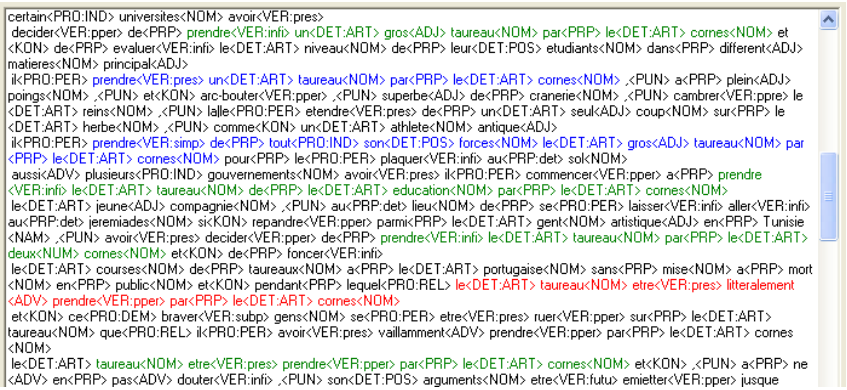


FIGURE 1 – Reconnaissance de *prendre le taureau par les cornes*

Nous pouvons remarquer que les SVF même transformées ont été retrouvées (en vert). Les séquences ne correspondant pas aux possibilités transformationnelles décrites ne sont pas extraites (en bleu). Elles correspondent en effet à la version littérale de la séquence. Cependant, nous ne réglons pas – et nous l'avions prévu – les conflits entre séquence littérale et séquence figée ayant les mêmes transformations ou la forme canonique (en rouge). L'outil aura donc tendance à privilégier la séquence figée. Pour régler de tels conflits, nous ne pourrons pas nous dédouaner d'un traitement de la syntaxe externe par l'analyse des arguments ou encore par un traitement sémantique plus étendu.

Nous présentons Table 4 les résultats de l'étiquetage de la séquence *prendre le taureau par les cornes*. La mesure prend en compte des séquences libres correctement ou incorrectement étiquetées.

	Prendre le taureau par les cornes
Rappel	99.29%
Précision	99.53%
F-score	99.41%

TABLE 4 – Reconnaissance de *prendre le taureau par les cornes*

Les principaux faux négatifs sont dus a des problèmes d'étiquetage morpho-syntaxique.

Précisons que ces données chiffrées sont à titre indicatives car elles ne représentent qu'une petite partie du problème, d'autant plus que même s'il peut avoir un sens littéral,

le sens opaque pour *prendre le taureau par les cornes* est plus fréquent. Le nombre d'attestations de séquences libres en témoignent avec seulement 1% des occurrences dans notre corpus. Ces résultats doivent également être comparés à d'autres méthodes (probabilistes, hybrides…). Toutefois, nous pensons que le début est prometteur et sera compétitif avec d'autres méthodes, notamment par le fait que la reconnaissance est indépendante de la taille du document et que l'importance est donnée aux séquences transformées qui représentent plus de 10% des occurrences, dans notre corpus.

5 Conclusion et Perspectives

Nous venons de présenter une méthode permettant de reconnaitre automatiquement des séquences verbales semi-figées. Cette approche est basée sur une ressource électronique constituée de SVF associées à leurs transformations possibles ou impossibles. Celle-ci répertorie à la fois les transformations liées aux composants de la séquence (modification flexionnelle, substitution, modifieurs…) mais également les changements liés à la séquence entière (passivation, relativation, inversion…) et les dépendances possibles entre les transformations (changement de déterminant en présence d'un modifieur). La ressource permet dans un premier temps, de trouver les mots du texte qui apparaissent comme des éléments obligatoirement présents dans la séquence, permettant de sélectionner des SVF potentielles qui seront testées sur une fenêtre autour de cet élément. Par la suite les composants sont vérifiés aussi bien par leur présence, leur substitution ou leur possible modification. Enfin l'ordre des composants est analysé s'il ne correspond pas à l'ordre canonique. Si aucune transformation syntagmatique n'est réalisable alors la séquence est rejetée. A l'inverse, si l'ordre trouvé correspond à une transformation possible celle-ci doit être validée par rapport à ses éléments définitoires (participe passé pour le passif). C'est ainsi, par l'utilisation d'une ressource électronique et une méthode qui se veut être la plus rapide possible (pour une éventuelle utilisation industrielle), que nous arrivons à extraire des SVF et à désambiguïser certaines de leur double littéral, indépendamment de la taille du corpus. Néanmoins, le travail est loin d'être terminé. La ressource établissant les différentes possibilités transformationnelles doit être complétée. Les SVF peuvent être décomposées en plusieurs types non pas selon leur degré de figement mais selon leur littéralité, leur dédoublement de sens, leur opacité, ou selon le domaine dans lequel on se trouve. En effet, les étapes de reconnaissance peuvent être allégées selon certaines conditions. De plus, la notion d'invariant doit être revue et élargie peut-être même jusqu'à une prise en compte de classes d'objets. L'ajout de la syntaxe externe doit également compléter la description pour une désambiguïsation totale. Enfin, des tests de robustesse doivent être effectués sur un corpus de référence, et les résultats doivent être comparés aussi bien à d'autres mesures qu'à d'autres ressources.

Références

ABEILLE, A. et SCHABES, Y. (1989). Parsing idioms in lexicalized TAGs. *In Proceedings of the the European Chapter of the Association for Computational Linguistics (EACL'89)*. Manchester, Angleterre.

AL-HAJ, A. et WINTNER, S. (2010). Identifying Multi-words Expressions by Leveraging

Morphological and Syntactic Idiosyncrasy. *In Proceedings of COLING 2010 (Conference on Computational Linguistics)*, Beijing, Chine.

BALIBAR-MRABTI, A. (2005). Semi-figement et limites de la phrase figée. *In LINX (53),* pages 35–54.

BEN-HENIA AYAT, I. (2006). Degrés de figement et double structuration des séquences verbales figées. Thèse de doctorat, Université Paris 13, Paris.

BEN-HENIA AYAT, I. (2009). Les séquences verbales figées métaphoriques. *In Synergie (1)*, pages 159–171.

BUVET, P.-A., (2008). Quelle description lexicographique du figement pour le TAL? Le cas des adjectifs prédicatifs à forme complexe. *In* (*Blumenthal et Mejri 2008*), pages 43–54.

CARTIER, E. (2008). Repérage automatique des expressions figées : état des lieux, perspectives. *In (Blumenthal et Mejri 2008)*, pages 55-70.

CARTIER, E. et JOSEPH A. (2011). Repérage automatique des séquences figées pour la classification des documents. *In LTT 2011 (Lexicologie, Terminologie, Traduction)*.

DAILLE, B. (2001). Extraction de collocation à partir de textes. *In TALN 2001 (Traitement automatique des langues naturelles*). Tours.

DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In (Klavans et Resnik 1996)*, pages 29–36.

DIAS, G. (2003). Multiword Unit Hybrid Extraction. *In Workshop on multiword expressions of ACL meeting (Association for Computational Linguistics)*. Sapporo, Japon.

GROSS, G. (1996). Les expressions figées en français noms composés et autres locutions. Ophrys, Paris.

GROSS, M. (1982). Une classification des « phrases figées » du français. *In Revue québécoise de linguistique*.

LAMIROY, B. et al. (2010). Les expressions verbales figées de la francophonie Belgique, France, Québec et Suisse. Ophrys, Paris.

LAMIROY, B. (2008). Les expressions figées : à la recherche d'une définition. *In (Blumenthal et Mejri 2008)*, pages 85–88.

LAMIROY, B. (2005). Le problème central du figement est le semi figement. *In LINX (53)*, pages 135–153.

LAPORTE, E. (1988). Reconnaissance des expressions figées lors de l'analyse automatique. *In Langages 23(90)*, pages 117–126.

LAPORTE, E., NAKAMURA, T. et VOYATZI, S. (2008). A French Corpus Annotated for Multiword Nouns. *In LREC 2008 (International Conference on Language Resources and Evaluation)*. Maroc.

LEPESANT, D. et MATHIEU-COLAS, M. (1998). Introduction aux classes d'objets. *In Langages 32(131)*, pages 6–33.

MANNING, C. et SCHÜTZE, H. (1999). Collocation. *In Draft*, pages 141–177.

MEJRI, S. (2011). Les Dictionnaires électroniques sémantico-syntaxiques. *In (CARDOSO et al. 2011)*, pages 159–188.

MEJRI, S. (2008). La place du figement dans la description des langues. *In (Blumenthal et Mejri 2008), p*ages 117–129.

MEJRI, S. (2003). Polysémie et polylexicalité. *In Syntaxe et sémantique (5)*.

RIEGEL, M., PELLAT, J.-C. et RIOUL, R. (1994). Grammaire méthodique du français. PUF, Paris.

SMADJA, F. (1993). Retrieving Collocations from Text: Xtract. *In Computational linguistics 19(1),* pages 144–177.

TOLONE, E. (2011). Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français. Thèse de doctorat, École des Ponts ParisTech, Paris.

VILLADA MOIRON, M.B. (2005). Data-driven identification of fixed expressions and their modifiability. Thèse de doctorat. Université de Groningen, Pays-Bas.

WATRIN, P. (2007). Collocations et traitement automatique des langues. *In Lexis and Grammar*, Bonifacio.

L'analyse de l'émotion dans les forums de santé

Céline Battaïa (Doctorante en Sciences de l'information et de la communication à l'université Stendhal de Grenoble)

Groupe de Recherche sur les Enjeux de la Communication (GRESEC), Université Stendhal, Laboratoire Gresec, Grenoble 3 celine.battaia@u-grenoble3.fr

RÉSUMÉ

Les travaux sur l'émotion dans les forums sont nombreux en Linguistique et Psychologie. L'objectif de cette contribution est de proposer une analyse de l'émotion dans les forums de santé selon l'angle des Sciences de l'Information et de la Communication mais également selon une approche interdisciplinaire. Il s'agira ici, d'étudier l'émotion comme un critère de pertinence lorsque des personnes malades effectuent des recherches dans les forums.
Ce papier introduit la méthodologie utilisée en traitement automatique de la langue afin de répondre à cette interrogation. Ainsi, le travail présenté abordera l'exploitation d'un corpus de messages de forums, la catégorisation semi-supervisée et l'utilisation du logiciel NooJ pour traiter de manière automatique les données.

ABSTRACT

Analysis of Emotion in Health Fora

Studies about emotion in fora are numerous in Linguistics and Psychology. This contribution approaches this subject from an Information and Communication Sciences point of view, and studies emotion as a criteron of pertinence for patients in a health forum. This paper introduces the empirical step of automatic language processing in order to answer this question, and uses data processing on the corpus of forum messages, semi-supervised categorisation of messages and use of software NooJ for Natural Language Processing.

MOTS-CLÉS : émotion, forum de santé, traitement automatique de la langue, désambiguïsation lexicale.

KEYWORDS: emotion, health forum, automatic language processing, lexical disambiguation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 267–280,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La recherche d'information peut être définie comme « l'activité d'un individu qui vise à localiser et traiter une ou plusieurs informations au sein d'un environnement documentaire complexe, afin de répondre à une question ou résoudre un problème »(Dinet et Rouet, 2002). Dans le contexte de la maladie rare, grave, ou chronique, la recherche d'information est une activité subordonnée à d'autres objectifs tels que rechercher des informations pratiques, des renseignements sur les traitements ou encore des contacts avec d'autres malades (Romeyer, 2008). Pour les Sciences de l'Information et de la Communication, la santé est un domaine illustrant les évolutions du processus de recherche d'information (Boubée et Tricot, 2010). En effet, ce champ « est tellement important qu'il influence même le domaine de la recherche d'information en général [...] sur à peu près tous les aspects : besoin d'information, tâche de recherche d'information, démarche de recherche d'information, évaluation des sources, etc. »(Boubée et Tricot, 2010).
Avec le succès d'Internet (Renahy *et al.*, 2007) et plus précisément des forums de santé (Paganelli et Clavier, 2010), se pose la question de l'évaluation de l'information recherchée. En effet, dans ces dispositifs, il n'est pas possible de mobiliser les critères de pertinence précaunisés par les différents modèles représentant le processus de recherche d'information (auteur, date, sources, etc.). Nous nous intéressons, par conséquent, aux critères de pertinence mobilisés par les malades pour valider les informations médicales recherchées dans des forums. En effet, puisqu'il n'est pas possible de mobiliser les critères traditonnels tels que la connaissance de l'auteur, la source, ou encore l'adéquation avec le domaine, nous nous demandons comment les malades valident les différentes informations trouvées.
Lors d'une maladie grave, rare, ou chronique, la motivation des malades à utiliser un forum de santé est informationnelle mais également émotionnelle (Paganelli et Clavier, 2010). Cela nous amène à nous intéresser à l'émotion non pas seulement comme raison pour rechercher de l'information dans les forums de santé mais également comme critère de pertinence pour évaluer les recherches faites dans dispositifs. Étudier l'émotion comme critère de pertinence est une approche originale car cela permet de mettre en lumière les mutations du processus de recherche d'information et à terme, proposer un nouveaux paradigme de recherche d'information.

Pour répondre à cette question, nous formulons quatres hypothèses : a) les malades utilisent le forum de santé pour des raisons émotionnelles mais également parce qu'il s'apparente à un dispositif de recherche d'information simple à utiliser et non contraignant, b) dans les forums de santé l'évaluation de l'information se fait de manière affective, c) Les malades sont à la recherche d'astuces de vie avec la maladie, d)lorsqu'il y a de l'information médicale, elle est donnée sur un ton émotif.
Afin d'affirmer ou infirmer ces hypothèses, nous avons mobilisé deux méthodes de travail empirique : une méthode d'expérimentation-entretien et une méthode de traitement automatique de la langue. Dans ce travail, nous présenterons d'abord notre posture théorique puis la méthode de traitement automatique de la langue employée. Nous expliciterons, donc, la catégorisation semi-supervisée utilisée afin de classer les différents types de messages et l'utilisation du logiciel NooJ afin d'étudier la proximité entre des termes émotifs et des termes médicaux. Enfin, dans la dernière partie nous présenterons les premiers résultats obtenus.

2 État de l'art

2.1 Cadre

Plusieurs modèles décrivent les différentes étapes adoptées par un utilisateur lors d'une recherche d'information (Barry et Schamber, 1998), (Guthrie, 1988), (Kuhlthau, 1991), etc. Ces modélisations sont majoritairement issues des Sciences de l'Information et de la Communication et de la Psychologie Cognitive.
Les Sciences de l'Information et de la Communication, mettent l'accent sur le fait que la recherche d'information est un processus dans le sens où il s'agit d'une « suite continue de faits présentant une unité »(Boubée et Tricot, 2010). Il s'agit d'une activité dynamique et évolutive car humaine. Cependant, pour la Psychologie Cognitive, l'activité de recherche d'information n'existe pas intrinsèquement mais dépend de contextes d'utilisation différents. L'individu réalise une tâche et s'aperçoit qu'il manque de connaissances pour la mener à bien. Il décide donc de pallier ce manque et de rechercher les informations qui lui font défauts. Mais pour cela il doit posséder des connaissances sur le contenu qui lui manque car, pour le domaine de la Psychologie Cognitive, il n'est pas possible de rechercher une information sans avoir un minimum de connaissances sur le domaine.
Quelque soit la posture disciplinaire de ces modèles, tous accordent une place importante à l'évaluation de la pertinence de l'information ou du document sélectionné. Ils mettent en évidence des critères de confiance que les individus doivent utiliser pour leur recherche. Ainsi, des critères tels que l'adéquation au thème recherché, la clarté des données, la source d'où provient le document, l'auteur, sont autant d'indices qui peuvent être mobilisés pour évaluer l'information (Barry et Schamber, 1998).

Cependant, force est de constater que les nouvelles technologies ont transformé les pratiques de recherche d'information (Boubée et Tricot, 2010). En effet, différentes enquêtes réalisées en Europe et aux États-Unis, mettent en avant le fait qu'Internet est de plus en plus utilisé pour rechercher de l'information médicale (Aube et Thoër, 2010).
Le champ de la santé est un domaine dans lequel ces évolutions sont le plus visibles (Boubée et Tricot, 2010). Le succès de la recherche santé en ligne est tel que « Google Flu »détecte les épidémies de grippe quelques jours avant leur survenue en analysant les requêtes des malades (Ginsberg *et al.*, 2008). Pour certains chercheurs, Internet conduirait à trois types de recherche de la part des personnes malades (Aube et Thoër, 2010) :
– une « logique professionnelle » : le malade cherche à obtenir des informations pour lui permettre d'appréhender la décision des médecins.
– une « logique consumériste » : le malade utilise Internet pour confronter les informations sur sa maladie, les différents traitements et ainsi faire un choix.
– une « logique communautaire » : le malade participe à des réseaux communautaires, tels que des forums, où s'exercent à la fois échange d'informations de santé et informations personnelles.

Nous nous intéressons à cette dernière logique car, dans les pays occidentaux, les forums de santé rencontrent beaucoup de succès (Paganelli et Clavier, 2010).
Un forum peut se définir comme un dispositif de communication médiaté par ordinateur, asynchrone et anonyme (Blanchard, 2007). C'est un lieux d'échange où l'information est

construite collectivement et le contenu auto-structuré. En effet, les participants sont à la fois producteurs et usagers de l'information (Clavier *et al.*, 2010). Par conséquent, l'inconvénient majeur des informations délivrées dans un forum, que ce soit des informations sur un utilisateur ou dans le contenu du message, est l'impossibilité de mobiliser les critères d'évaluation traditionnellement utilisés, même si le forum est hébergé dans un site fiable. Se pose alors la question de savoir pourquoi les forums de santé ont autant de succès.
Il apparaît que ce dispositif présente, un double objectif : un soutien informationnel et émotionnel (Gaglio, 2010), (Paganelli et Clavier, 2010). Avant de poursuivre, il nous semble important de définir ce que nous entendons par émotion car bien que ce concept ait été très étudié ces dernières années, il reste assez difficile à appréhender de façon claire (Novakova et Tutin, 2009a). Il existe plusieurs systèmes de définitions de l'émotion. Ainsi, les spécialistes ne sont pas toujours d'accord sur les définitions à adopter. Dans notre travail, nous avons choisi de traiter de l'émotion au sens large du terme. Cela signifie que nous englobons l'émotion (joie, tristesse, haine, etc.) et les sentiments (amertume, crainte, honte, etc.) (Blumenthal, 2009).

Pour les Sciences de l'Information de la Communication, la prise en compte de l'émotion dans le processus de recherche d'information n'est pas nouveau. Néanmoins, force est de constater que l'émotion n'est étudiée que en tant qu'émotion ressentie par un individu dans les étapes de sa recherche ou dans des lieux documentaires. À titre d'exemple, nous pouvons citer le modèle Information Search Process créé en 1991 (Kuhlthau, 1991). Contrairement aux précédentes modèlisations du processus de recherche d'information, celui-ci mettait en évidence les rôle des émotions et de la confiance en soi dans les étapes d'une recherche documentaire. Le modèle ISP se compose de six phases :

– une phase d'initiation qui correspond à l'apparition d'une sensation d'incertitude et d'attente informationnelle,
– une phase de sélection où l'individu définit ce qu'il veut rechercher et a donc un sentiment d'optimisme,
– une phase d'exploration qui correspond à une phase de réflexion et d'acquisition d'information
– une phase de formulation. Cette étape correspond à une phase de sérénité, d'incertitude diminuée et une augmentation de confiance dans sa recherche,
– une phase de collecte des données qui renforce la confiance de l'individu au fur et à mesure qu'il trouve des informations,
– Enfin, une phase de présentation des résultats. L'individu n'a pas d'incertitude mais se trouve dans un état de satisfaction ou d'insatisfaction.

Au cours de ces étapes, l'individu va chercher à réduire son principe d'incertitude. L'incertitude est un état cognitif vague, difficile à formuler pour l'individu et souvent accompagné de sentiments de confusion ou de frustration (Boubée et Tricot, 2010).

Plus récemment, d'autres travaux, se sont interrogés sur le rôle de l'émotion dans la détermination des prises de décision (Julien *et al.*, 2005) ou encore sur le lien entre contrôle de l'émotion et tâche de recherche en ligne (Kim, 2008). Il semblerait qu'Internet favorise le lien entre recherche d'information et émotion (Abbasi *et al.*, 2008). Leur enquête met en avant la richesse émotionnelle des discours sur Internet. Par conséquent, ces dernières années, l'analyse des sentiments a été appliqué à différentes formes du web et dans différents domaine. Par exemple, l'analyse des sentiments dans les forums a déjà été traité en linguistique mais dans le but d'améliorer la qualité d'un système d'extraction de sentiment grâce au traitement

automatique de la langue naturelle (Maurel *et al.*, 2008). Plus généralement, d'autres travaux en linguistique cherchent à caractériser les émotions, étudient un ou deux champs sémantiques de l'émotion ou encore travaillent sur l'extraction automatique des affects (Novakova et Tutin, 2009a), (Barbé, 2007). Nous n'allons pas étudier l'émotion de cette manière là mais nous aider des précédents travaux pour nous intéresser aux indices émotionnels contenus dans les messages des forums de santé et nous demander si l'émotion peut devenir un critère de pertinence.

2.2 Hypothèses

A/ Les malades ou proches de malades cherchent principalement des astuces de vie avec la maladie
Les travaux réalisés précédemment (Romeyer, 2008), (Lemire, 2009),(Clavier *et al.*, 2010) mettent en avant le fait que les internautes, et plus précisément les malades ou proches de malades, recherchent principalement des informations pratiques concernant la vie avec la maladie ; ceux-ci veulent, par exemple, obtenir des astuces sur la gestion des effets secondaires d'un médicament pour la thyroïde. Or, d'autres études ont des résultats contraires et révèlent que : « *Dans son panorama de la « consommation d'informations »sur l'Internet médical, D. Nabarette (Nabarette, 2002) y relie cinq buts : s'informer pour connaître (maladies, traitements...), choisir (praticiens, plans de soins...), superviser l'action des professionnels, produire (un diagnostic...), se coordonner dans le cadre de la relation patient/médecin. Notre enquête révèle une prééminence de la première catégorie* (Gaglio, 2010) ».
Il nous semble important de pouvoir déterminer quel type d'information vont chercher les malades dans les forums de santé. Or, les résultats des travaux réalisés jusqu'à présent ne vont pas toujours dans le même sens.
Après une première observation et un croisement des lectures qui ont pu être faites, nous émettons l'hypothèse que les malades sont plutôt à la recherche d'astuces de vie avec la maladie. Afin de typer les informations contenues dans notre corpus, nous réalisons, à l'aide d'un partenariat avec le Laboratoire d'Informatique d'Orléans, une catégorisation semi-supervisée.

B/ L'information médicale est donnée sur un ton émotif
Par information médicale nous voulons dire information médicale (professionnelle) et information de santé (grand public, non spécialisée) (Romeyer, 2008). Plusieurs enquêtes telles que celle de Joëlle Kivits (Kivits, 2006), d'Hélène Romeyer (Romeyer, 2008), ou encore Gérard Gaglio (Gaglio, 2010) rappellent que la relation avec le praticien ne change pas dans le sens où les professionnels de santé restent les principales sources d'obtention de l'information médicale(Vercher et Touboul, 2011). Les forums de santé ne sont donc pas des lieux uniquement faits pour échanger de l'information médicale. Étant donné que le forum est un lieux d'échange avec d'autres malades, nous pensons que les informations médicales sont délivrées sur un ton émotif car lié à l'expérience et au ressenti que le malade éprouve face à la maladie. Afin d'infirmer ou affirmer cette hypothèse, nous allons nous servir du logiciel NooJ et observer la proximité entre les termes d'émotion et les termes médicaux.

Les deux hypothèses suivantes ne font pas l'objet d'un traitement automatique de la langue mais d'une expérimentation conmplétée par des entretiens. Cela nous permet à la fois de

dépasser le biais méthodologique puisque les entretiens ne sont finalement que des discours sur leurs pratiques faites par les individus interrogés, et non l'observation du processus en train de se faire (Blanchard, 2007). Néanmoins, nous avons choisi de les évoquer brièvement, afin de présenter de manière approfondie le cadre théorique sur lequel la recherche est fondée.

C/ Motivation des malades à utiliser le forum de santé
Nous pensons que, outre la dimension émotionnelle et le fait de pouvoir s'exprimer entre personnes malades, que l'on peut retrouver dans d'autres dispositifs, le forum de santé donne l'impression d'être un outil de recherche d'information moins contraignant que les autres pour deux raisons. La première est que l'on peut poser les questions que l'on veut sans avoir besoin de naviguer entre plusieurs sources pour obtenir une réponse. Il est également possible de trouver des informations sans avoir besoin de participer, de s'inscrire. La lecture des messages suffit. La deuxième raison, induite par les observations qui ont pu être menées lors de la composition du corpus de travail, est que lorsque les malades obtiennent une réponse à leur question ou n'en ressentent plus le besoin, il est plus facile de ne plus participer au forum. Nous nous appuyons ici sur les travaux d'H. Romeyer (Romeyer, 2008). En effet, son enquête de 2008 a révélé que les malades ne cherchent pas a établir de lien avec les autres malades. C'est la raison pour laquelle nous pensons que dès qu'ils ont obtenu satisfaction ou non avec leur recherche, les malades ne participent plus ou ne lisent plus les messages des forums de santé .

D/ Évaluation affective de l'information faite dans les forums de santé
Nous considérons le forum de santé comme une source d'information à dimension affective dans le sens où les malades partagent aussi bien des informations médicales que personnelles (Clavier *et al.*, 2008), (Clavier *et al.*, 2010), (Paganelli et Clavier, 2010). Les messages sont un mélange d'informations objectives et émotionnelles. Pour nous, l'émotion change la perception qu'un internaute a du message et modifie les critères de sélection habituellement utilisés. Cela signifie que les malades s'attachent aux indices émotionnels contenu dans un message pour l'évaluer. Plus un message contiendra d'indices émotionnels (mots, ponctuation exagérée, smileys), plus les individus auront confiance dans l'information véhiculée. Cela leur donnerait l'impression que le message n'a pas pu être inventé. Par conséquent, nous pensons que plus un message contiendra des termes émotifs, plus les malades le considèreront comme vrai, comme digne de confiance car ils se reconnaissent dans les termes émotifs utilisés.

3 Méthodologie du travail empirique

Dans cette partie, nous présentons la méthodologie de traitement automatique de la langue mobilisée pour répondre à nos deux premières hypothèses (A/ et B/).
Le nombre de forums de santé francophones existant sur Internet est trop élevé pour qu'il soit possible de les recenser de manière exhaustive. Notre parti-pris a donc été de les répertorier à l'aide du Catalogue et Index des Sites Médicaux de langue Française (Cismef [1]), du site Health On the Net Foundation (Honcode [2]) et via une requête sans la zone de recherche de deux moteurs de recherche. [3]. Le Cismef est un projet initié par le Centre Hospitalier Universitaire de Rouen depuis

1. http ://www.chu-rouen.fr/cismef/
2. http ://www.hon.ch/HONcode/French/
3. « Forum+santé »dans Google et Yahoo

1995. Des documentalistes et médecins recensent des sites médicaux répondant au référentiel des critères de qualité de l'information de santé sur le net (netscoring [4]). Le netscoring a été développé par un groupement de professionnels de santé, bibliothécaires médicaux et juristes. Le Honcode quant à lui,est financée par le Canton de Genève, des projets Européens, la Haute Autorité de Santé de France (HAS) et la fondation Provisu. Sa mission est de guider les utilisateurs d'internet vers des sources d'informations médicales et de santé fiables.

Les forums accessibles via ces trois types de sources, ont été catégorisés selon leur spécialisation : forums de santé généralistes, spécialisés dans une maladie, modérés, modérés par un médecin. Nous avons sélectionné des fils de discussion traitant de maladies rares, graves, ou chroniques car nous pensons qu'ils sont plus à même de fournir à la fois des données informationnelles et émotionnelles. D'autre part, il est extrêment difficile de trouver un forum spécialisé sur le rhume par exemple. Or, nous voulons pouvoir effectuer des comparaisons entre forums généralistes et forums spécialisés.

Nous travaillons sur un corpus de 2481 messages répertoriés comme suit :

– Forums généralistes modérés par un médecin= catégorie GD
  Forum Atoute (8 fils de discussion, 631 messages)

– Forums généralistes modérés= catégorie GM
  Forums Au féminin (1 fil de discussion, 101 messages), Doctissimo (1 fil de discussion, 159 messages), E-santé (1 fil de discussion, 172 messages) et Santé médecine (2 fils de discussion, 150 messages).

– Forums spécialisés modérés par un médecin= catégorie SD
  Forums Ligue Cancer (6 fils de discussion et 329 messages) et Traitement du pied bot à la naissance (14 fils de discussions, 292 messages).

– Forums spécialisés modérés= catégorie SM
  Forums maladies Lysosomales (2 fils de discussion, 153 messages), Renaloo (2 fils de discussion, 157 messages), Solhand (3 fils de discussion, 173 messages) Vivre sans thyroïde (1 fil de discussion, 164 messages).

Il aurait été intéressant pour notre enquête de pouvoir également analyser des forums de santé non modérés mais nous n'en avons pas trouvé.
Un traitement semi-automatique (avec le logiciel Python [5]) a été réalisé sur notre corpus. Nous avons en effet, répertorié et classé les informations externes au contenu des messages tel que le titre des fils de discussion, la source, l'url ou encore le nombre de messages. Nous avons souhaité traiter ces informations afin de pouvoir effectuer des comparaisons entre les différents fils de discussion mais également afin d'avoir une trace d'archivage. Nous avons ensuite créé un filtre Python afin de numéroter automatiquement les messages et de ne conserver que le texte. L'inconvénient majeur est la suppression des smileys car le filtre ne conserve que le texte et non les images. Les smileys ont ensuite été intégrés au corpus en tant que signification (par exemple : SMILEY=SOURIRE).

4. http ://www.chu-rouen.fr/netscoring/
5. http ://www.python.org/

3.1 Catégorisation semi-supervisée

La catégorisation semi-supervisée se fait en partenariat avec le Laboratoire d'Informatique Fondamentale d'Orléans [6]. La classification semi-supervisée est une hybridation entre la classification supervisée (ou classification par apprentissage) et la classification par extraction (non-supervisée). La classification par apprentissage permet d'établir des classifications à partir de documents pré-classés tandis que, la classification par extraction consiste en une fouille de données. Cela permet de pallier le nombre trop élevé de documents en les classant de manière automatique.
Faire du semi-supervisé dans le cadre de ce travail, permet d'orienter la classification et de faire ainsi du clustering semi-supervisé avec plusieurs vues : lexique, smileys, ponctuation exagérée. L'objectif est donc de répondre à l'hypothèse concernant le type d'information recherchée par les malades (hypothèse A/) et de pouvoir ainsi catégoriser de manière semi-automatique les messages. L'originalité de cette méthode est également de permettre de travailler en parallèle sur les différentes manières de transmettre l'émotion dans un forum de santé, à savoir les mots, la ponctuation et les smileys. En effet, quelques fois l'émotion n'est pas transmise par les termes émotifs.

```
Je conduis pour aller au boulot , mais c'est mais arriver de
m'endormir en voiture après le travail SMILEY=PEUR fatigue .
Je vais a Créteil environ tout les 18 mois comme la maladie
et stable ( pour le moment ,rien ne dit que sa peut évaluer !!!!!!

Forum spécialisé modére Solidarité Handicap, Fil de discussion Syndrome
de Poems (numéro 1), Message SMSH01-0014
```

Nous avons, dans un premier temps, créé une liste de catégories d'après le compte rendu du colloque : web participatif et santé (Paganelli et Clavier, 2010), d'une exploration de notre corpus, et de forums de santé sur Internet. La raison pour laquelle nous avons décidé de procéder ainsi est que nous ne voulons pas avoir de catégories correspondant uniquement à notre corpus car cela risquerait de fausser les résultats.
Une pré-catégorisation manuelle de 586 messages du corpus a ainsi été faite et envoyé au laboratoire d'Informatique Fondamentale d'Orléans. 9 « items »ont ainsi pu être mis en évidence :

– Astuces face à la maladie

– Autre
  Demande de présentation de soi ou d'un proche sans évoquer son récit de vie, présentation de soi sans évoquer son récit de vie avec la maladie, encouragements sans apports d'informations, demande de contact physique ou virtuel, réponse à une demande de contact, proposition de contact sans demande préalable dans les autres messages.

– Feed Backs

6. le LIFO

– Hors sujet

– Informations ponctuelles

– Informations sur le traitement suivi

– Information sur la maladie en général

– Ressenti vis-à-vis de la maladie

– Témoignage

Ces classes ont ensuite été divisées en sous-catégories. Le typage des messages et le corpus de travail ont ensuite été envoyés au LIFO pour traitement informatique.

3.2 Traitement automatique de la langue avec NooJ

Une phase de lemmatisation précède le traitement du corpus avec NooJ [7]. Elle permet de ramener les termes à leur forme de lemme, mais également de corriger les fautes (nombreuses dans les forums). Lemmatiser le corpus permet, en effet, de mettre en évidence les fautes puisque le logiciel ne reconnaîtra pas les mots mal orthographiés et les catégorisera comme « unknows ». La lemmatisation a été faite à l'aide du logiciel Treetagger [8]. Ne pas corriger les erreurs auraient compliqué l'analyse du corpus puisque les termes mal orthographiés n'auraient pas pu être analysés par NooJ.
NooJ est un environnement linguistique créé en 1993 , d'abord sous le nom d'INTEX, par M. Silberztein (Svetla *et al.*, 2007). Il permet de formaliser les langues naturelles, de développer des applications de traitement automatique du langage mais également de travailler sur de gros corpus (Yamouni-Aoughlis, 2010).
Afin de pouvoir affirmer ou infirmer notre hypothèse sur la proximité de termes émotifs et de termes médicaux (hypothèse B/), nous avons créé des grammaires afin de localiser les adjectifs, noms et verbes émotifs. La figure 1 présente un extrait de la grammaire faite pour les adjectifs.

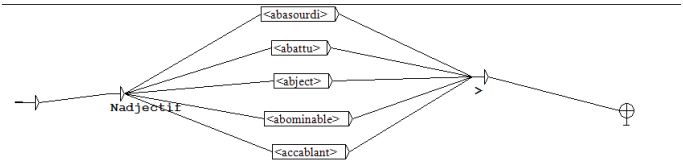


FIGURE 1 – Extait de la grammaire NAdjectif

7. http ://www.nooj4nlp.net/pages/nooj.html
8. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Les affects ne constituent pas une classe homogène (Novakova et Tutin, 2009b). Certains se rapprochent des sentiments (par exemple honte), d'autre des émotions (par exemple peur). Néanmoins, force est de constater que nous ne travaillons pas sur la catégorisation de l'émotion en tant que tel. Pour le moment, nous n'étudions pas la valeur des termes utilisés par les internautes mais nous nous servons de listes de termes émotifs pour analyser la manière dont est transmise l'information médicale. Le but est d'analyser la proximité entre les termes émotifs et les termes médicaux. Les trois grammaires (Nadjectif, Nnom et Nverbe) permettent de mettre en lumière les phrases contenant des affects. Une comparaison est ensuite faite afin de vérifier la présence ou non de termes médicaux transmis aussi bien sous la forme spécialisée (Carcinome à cellules claires) que sur le mode grand public (Tumeur rénale).
L'inconvénient de cette méthode est de ne pas pouvoir mettre en lumière des expressions ne mobilisant pas a priori de'affects mais transmettant néanmoins une émotion : en avoir gros sur la patate. Pour pallier ce désagrément, nous établirons une liste d'expressions reflétant des émotions et nous créerons une quatrième grammaire.

Enfin, il nous semble important de préciser, que les paramètres standarts de Treetagger et NooJ ont été utlisés pour traiter le corpus.

4 Premiers résultats et discussion

4.1 Les malades recherchent des témoignages

Nous sommes en attente des résultats de la catégorisation semi-supervisée. Néanmoins, le typage manuel transmis au LIFO, nous permet d'affirmer, à priori, que les malades ne sont pas à la recherche d'astuce de vie (un saignement du nez provoqué par un médicament pour le foie). Les messages analysés contiennent essentiellement des témoignages, des récits de vie liés à la maladie.
En effet, sur 586 messages traités, 150 proviennent d'échanges de témoignages (17 demandes, 24 apports de témoignages suite à une demande, 109 apports de témoignages sans demande particulière au préalable).
116 messages sont liés au ressenti émotionnel vis-à-vis de la maladie, 95 correspondent à des demandes de contact et/ou encouragments, 93 concernent des échanges d'information sur les traitements, 66 sont hors sujet. 47 messages sont des retours sur les réponses (feedback), 46 concernent des échanges d'informations ponctuelles, d'astuce de vie et 19 échanges traitent de la maladie en général.
Il s'agit majoritairement d'échanges de témoignages et de ressentis avec la maladie. Cela confirme l'affirmation de Céline Paganelli et Viviane Clavier (Paganelli et Clavier, 2010), à savoir que les motivations des malades sont d'ordre émotionnel. Par contre cela remet en cause les travaux faits par (Clavier *et al.*, 2010) puisque les résultats de leur étude montrent que les malades sont plutôt à la recherche d'astuce de vie avec la maladie.
Le traitement entier du corpus doit être attendu pour généraliser les résultats. Néanmoins les premières données montreraient que les malades n'utilisent pas les forums dans un but informationnel. Ces résultats doivent être mis en parrallèle avec la thèse réalisée par (Sénis, 2003), pour qui les malades restent attachés au monde médical pour valider les informations recherchées sur Internet. Cela expliquerait également un autre de ses résultats, à savoir que les

patients sont hostiles à la présence de médecins comme intervenants dans les forums. Ce ne sont pas, pour eux des lieux de recherche mais d'échange de leur vie.

4.2 L'information médicale n'est pas donnée sur un ton émotif

La question de l'émotion dans les forums est majoritairement traitée selon une approche communicationnelle dichotomisante. Ce type de communication favoriserait l'émergence d'une communication émotive entre écrit et oral (ponctuation exagérée, smileys)tandis que, à l'inverse, d'autres études considèrent que la communication dans les forums et plus généralement la communication médiatisée par ordinateur, entraverait la communication émotionnelle (Atifi *et al.*, 2010). Les premiers résultats obtenus avec l'analyse du corpus par NooJ, confirment la catégorisation semi-supervisée du contenu et laissent à penser que la communication émotionnelle exite bien mais elle est réservée aux témoignages, aux expériences de vie avec la maladie et non avec la diffusion d'informations médicales. Ainsi, sur 200 messages analysés, 153 contiennent de l'émotion que cela soit par les smileys, les mots ou la ponctuation exagérée :

```
Ma mère se fait opérer mercredi et depuis 2 jours,
comme vous, son moral est au plus bas.
Elle angoisse et est très négative.
Heureusement elle n'a ni vertige ni maux
de tête mais elle est très fatiguée

Forum Génaraliste Modére par un médecin Atoute,
Fil de discussion Opération d'un méningiome, Message GDAT02-0002
```

Sur les 57 messages restants, 22 messages contiennent des informations objectives comme des informations pratiques par exemple et seulement 35 messages contiennent des informations médicales. D'après nos premiers résultats, les informations médicales ne semblent pas être transmises sur un ton émotionnel :

```
mon médecin  me prescrit un scanner et le 5 Juillet : méningiome
4 x 3,5 x 2,5 avec œdème. Premier contact avec un
chirurgien de la Pitié-Salpétrière (en vacance) RDV pour
consultation le 11/8, diagnostique : opération pour le 30
Août. Je suis entrée à l'hôpital le 29 car j'avais passé les
examens demandés chez moi. Le 29 au soir double douche avec
shampooing à la Bétadine, léger somnifère et le lendemain
matin 6H.

Forum Génaraliste Modére par un médecin Atoute,
Fil de discussion Opération d'un méningiome, Message GDAT02-0006
```

Les données recuillies, donnent l'impression que l'émotion est vraiment réservée à l'apport d'informations personnelles. On peut alors s'interroger sur l'émotion comme critère de pertinence. En effet, cela laisse à penser que les individus ont peur de ne plus être crédibles s'ils délivrent une information médicale avec beaucoup d'émotion. Le forum de santé serait, avant d'être un lieu de recherche d'information, un dispositif de maintien de lien entre des personnes vivant la même expérience de maladie.

5 Conclusion

Ce travail s'intéresse à l'analyse de l'émotion dans l'étude des pratiques de recherche d'information, que les spécificités propres aux forums font évoluer (Paganelli et Clavier, 2010) et tout particulièrement dans le domaine de la santé où la fiabilité des informations médicales recherchée est difficilement vérifiable pour des malades. C'est la raison pour laquelle il est intéressant d'étudier comment l'émotion peut devenir un critère de pertinence.

La méthodologie présentée dans l'article a mis en évidence que les malades semblent plutôt être à la recherche de récit de vie avec la maladie. Les conversations, lorsqu'elles concernent la maladie en elle-même sont plus objectives. L'émotion intervient lorsqu'il s'agit d'échange plus personnels sur la vie avec la maladie. Les premiers résultats nous laissent à penser que l'émotion n'est pas un critère de pertinence en tant que tel car elle n'intervient pas dans les échanges d'informations médicales mais dans l'échange de témoignages, de ressenti vis-à-vis de la maladie. Cela confirme les résultats de l'enquête menée par (Clavier *et al.*, 2008), à savoir que les malades ne viennent pas pour rechercher des informations médicales *stricto sensu* mais pour partager leur expérience avec la maladie. La mutation du processus de recherche d'information interviendrait dans le sens où des personnes recherchent des informations filtrées par l'expérience de l'individu et non plus des informations objectives. Néanmoins, les résultats sont à compléter par la continuation du travail de terrain (catégorisation semi-supervisée et traitement NooJ de tout le corpus), mais également par la phase d'expérimentation-entretien afin d'appréhender le discours des malades mais aussi leur manière de faire.

Remerciements

Un grand merci à Julien Corman qui a lemmatisé le corpus de travail avec Treetagger et à Agnès Tutin pour nous avoir transmis les listes de termes émotifs.

Références

Abbasi, A., Chen, H. et Salem, A. (2008). Sentiment analysis in multiple languages : Feature selection for opinion classification in web forums.

Atifi, H., Gauducheau, N. et Marcoccia, M. (2010). L'expression et le rôle des émotions dans les forums de discussion. *In* Yasri-Labrique, E., éditeur : *Les forums dediscussion : agora du XXIème siècle ? Théorie, enjeux et pratiques discursives*, pages 71–87. Paris.

Aube, S. et Thoër, C. (2010). La construction des savoirs relatifs aux médicaments sur internet : étude exploratoire d'un forum sur les produits amaigrissants utilisés sans supervision médicale. *In* Renaud, L., éditeur : *Les médias et la santé : de l'émergence à l'appropriation des normes sociales*, pages 239–266. Québec.

Barbé, M. (2007). *Analyses linguistiques et modélisations des connaissances en vue d'un traitement automatique des e-mails entrants : vers un système de veille dans le tourisme aérien*.

Barry, L. C. et Schamber, L. (1998). User's criteria for relevance evaluation : a cross-situational comparison. *Information Processing and Management*, 34(2-3):219–236.

Blanchard, G. (2007). *La communication politique partisane sur Internet : des pratiques et des stratégies nouvelles ?*

Blumenthal, P. (2009). Les noms d'émotion : trois systèmes d'ordre. *In* Novakova, I. et Tutin, A., éditeurs : *Le lexique des émotions*, pages 41–79. Grenoble.

Boubée, N. et Tricot, A. (2010). *Qu'est-ce-que rechercher de l'information ? : état de l'art*, page 204. Villeurbanne.

Clavier, V., Paganelli, C., Manes-Gallo, M.-C., Mounier, E., Romeyer, H. et Staii, A. (2010). Dynamiques interactionnelles et rapport à l'information dans les forums de discussion médicale. *In* Millerand, F. e. a., éditeur : *Web social : mutation de la communication*, pages 297–314. Québec.

Clavier, V., Paganelli, C., Manes-Gallo, M.-C., Mounier, E., Romeyer, H. et Staii, A. (Québec 6-7 mai 2008). Web participatif et santé : de nouveaux rapports à l'information ? *In Web participatif : mutation de la communication ?*

Dinet, J. et Rouet, J.-F. (2002). La recherche d'information : processus cognitifs,facteurs de difficultés et dimension de l'expertise. *In* Paganelli, C., éditeur : *Interaction Homme-Machine et recherche d'information*, pages 113–161. Paris.

Gaglio, G. (2010). Consommation d'informations sur internet et modulation de la relation aux médecins. le cas d'aidantes de malades atteints d'une pathologie lourde. *Sociologies Pratiques*, 1(20):63–74.

Ginsberg, J., Mohebbi, M.-H., P. R.-S. B. L., Smolinski, M.-S. et Brilliant, L. (2008). Detecting influenza epidemics using search engine query data.

Guthrie, J. T. (1988). Locating information in documents : Examination of a cognitive model. *Reading Research Quarterly*, 23:178–199.

Julien, H., McKechnie, L. E. F. et Hart, S. (2005). Affective issues in library and information science systems work : A content analysis. *Library and Information Science Research*, 27(4):453–466.

Kim, K. (2008). Effects of emotion control and task on web searching behavior. *Information Processing and Management*, 44(1):373–385.

Kivits, J. (2006). Informed patients and the internet : a mediated context for consultations with health professionnals. *Journal of Health Psychology*, 11(2):269–282.

Kuhlthau, C. C. (1991). Inside the search process : Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371.

Lemire, M. (2009). Internet et responsabilisation : perspective de l'usager au quotidien. *Santé Publique*, 21:13–25.

Maurel, S., Curtoni, P. et Dini, L. (Fontainebleau - 2008). L'analyse des sentiments dans les forums. *In Atelier Fouille des Données d'Opinion*.

Nabarette, H. (2002). L'internet médical et la consommation d'information par les patients. *Réseaux*, 4:249–286.

Novakova, I. et Tutin, A. (2009a). *Le lexique des émotions*. Grenoble.

Novakova, I. et Tutin, A. (2009b). Les émotions sont-elles comptables. *In* Novakova, I. et Tutin, A., éditeurs : *Le lexique des émotions*, pages 65–79. Grenoble.

Paganelli, C. et Clavier, V. (2010). Le forum de discussion : une ressource informationnelle hybride entre information grand public et information spécialisée. *In* Yasri-Labrique, E., éditeur : *Les forums dediscussion : agora du XXIème siècle ? Théorie, enjeux et pratiques discursives*, pages 39–54. Paris.

Renahy, E., Parizot, I. et Chauvin, P. (2007). Whist :a web-based survey on health information seeking on internet in france. Rapport technique, Paris, INSERM.

Romeyer, H. (2008). Tic et santé : entre information médicale et information de santé.

Svetla, K., Maurel, D. et Silberztein, M. (2007). *Formaliser les langues avec l'ordinateur : de INTEX à NooJ*.

Sénis, F. (2003). *Pourquoi accéder à l'information médicale sur Internet par le biais des groupes de discussion ? Qualité, centres d'intrérêt et motivations des participants aux forums médicaux. À propos du forum Usenet Fr.bio.medecine*.

Vercher, E. et Touboul, A.-L. (2011). L'information santé : entre vulgarisation scientifique et circulation communautaire, le cas des sites santé 2.0. *In Internet et santé : regards croisés France/Québec. Lyon : Médiathèque Université Lyon 1, 21 février*.

Yamouni-Aoughlis, Y. (2010). *Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales*.

Peuplement d'une ontologie modélisant le comportement d'un environnement intelligent guidé par l'extraction d'instances de relations

Driss Sadoun[1, 2]
(1) LIMSI/CNRS, B.P. 133 91403 Orsay Cedex, France
(2) Université Paris-Sud, 91400 Orsay, France
driss.sadoun@limsi.fr

RÉSUMÉ

Nous présentons une approche de peuplement d'ontologie dont le but est de modéliser le comportement de composants logiciels afin de faciliter le passage de descriptions d'exigences en langue naturelle à des spécifications formelles. L'ontologie que nous cherchons à peupler a été conçue à partir des connaissances du domaine de la domotique et est initialisée à partir d'une description de la configuration physique d'un environnement intelligent. Notre méthode est guidée par l'extraction d'instances de relations permettant par là-même d'extraire les instances de concepts liés par ces relations. Nous construisons des règles d'extraction à partir d'éléments issus de l'analyse syntaxique de descriptions de besoins utilisateurs et de ressources terminologiques associées aux concepts et relations de l'ontologie. Notre approche de peuplement se distingue par sa finalité qui n'est pas d'extraire toutes les instances décrivant un domaine mais d'extraire des instances pouvant participer sans conflit à un des multiples fonctionnements décrit par des utilisateurs.

ABSTRACT

Population of an Ontology Modeling the Behavior of an Intelligent Environment Guided by Instance Relation Extractions

We present an approach for ontology population, which aims at modeling the behavior of software components, for enabling a transition from natural language requirements to formal specifications. The ontology was designed based on the knowledge of the domotic domain and is initialized from a description of a physical configuration of an intelligent environment. Our method focuses on extracting relation instances which allows the extraction of concept instances linked by these relations. We built extraction rules using elements coming from syntactic analysis of user need descriptions, semantic and terminological resources linked to the knowledge contained in the ontology. Our approach for ontology population, distinguishes itself by its purpose, which is not to extract all instances describing a domain but to extract instances that can participate without any conflict to one of the mutiple operation decribed by users.

MOTS-CLÉS : extraction de relations, peuplement d'ontologie, représentation des connaissances.

KEYWORDS: relation extraction, ontology population, knowledge representation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 281–294,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les ontologies permettent une modélisation formelle d'un domaine, en décrivant ses concepts et les relations qu'ils entretiennent ainsi que les individus (ou instances) qui leurs sont associés. La conception d'une ontologie se fait en deux étapes, d'abord sa *conceptualisation* (Bendaoud *et al.*, 2007) (Wang et Turner, 2009) qui a pour but de faire émerger les concepts et relations représentant le domaine visé, ainsi que les axiomes qui permettront d'ordonner concepts et relations dans l'ontologie et de classer leurs instances, ensuite son *peuplement ou instanciation* qui consiste à associer des instances à des concepts et relations existants. C'est cette deuxième tâche qui sera traitée dans cet article.

La méthode de peuplement choisie est en général conditionnée par l'usage auquel est destinée l'ontologie. Dans le cas d'une ontologie modélisant un domaine de façon descriptive (De Boer *et al.*, 2007) (Maynard *et al.*, 2008), le but est de détecter, d'extraire et de lister systématiquement toutes les instances contenues dans les textes pouvant refléter un concept ou une relation du domaine à décrire. L'ontologie que nous avons conçue a pour vocation de modéliser le comportement dynamique d'un système de composants logiciels, en l'occurrence le comportement d'un environnement intelligent, que nous avons modélisé à partir de l'étude des domaines de réseaux de capteurs, des environnements intelligents et de la domotique (Sadoun *et al.*, 2011). Le processus de peuplement aura pour but d'extraire seulement les instances reflétant un comportement possible de ce système spécifique à des exigences d'utilisateurs.

L'ambition du projet (Envie Verte [1]) dans lequel s'inscrit ce travail, est de permettre le pilotage d'un environnement intelligent en partant de descriptions en langue naturelle du comportement du système vers des spécifications formelles. Les spécifications sont données par des utilisateurs pour exprimer leurs besoins. L'intérêt est de permettre une vérification formelle de la configuration résultat avant de la déployer. A cette fin, nous avons proposé de passer par l'intermédiaire d'une ontologie du comportement de composants logiciels, pour faciliter le passage de l'informel vers le formel (Sadoun *et al.*, 2012). En effet, du fait de leur formalisme basé sur les logiques de description, les ontologies constituent un outil puissant pour l'aide à la vérification d'exigences et de spécifications formulées en langue naturelle. Nous exploitons ontologie et règles pour représenter et vérifier la consistance et la complétude des spécifications décrivant les comportements de composants logiciels. L'intérêt est de rendre ces comportements facilement paramétrables et adaptables.

L'ontologie ayant été définie, cet article en décrit une méthode de peuplement. La méthode proposée est centrée sur la détection d'instances de relations appartenant à la conceptualisation, point de départ pour l'identification d'instances de concepts.

La section suivante décrit l'état de l'art, la section 3 décrit brièvement l'ontologie, la section 4 décrit les descriptions d'environnement et des exigences. En section 5, nous introduisons la notion d'interprétation des exigences. La section 6 décrit la méthode de peuplement d'ontologie, puis en section 7 nous discutons la gestion de l'implicite avant la conclusion et la définition des perspectives.

1. http ://envieverte.limsi.fr/, projet financé par DIGITEO 2010

2 État de l'art

Le peuplement automatique d'ontologie à partir de textes est une problématique qui a donné lieu à différentes approches (Alani *et al.*, 2003) , (Alani *et al.*, 2004) , (Amardeilh *et al.*, 2005), (Maynard *et al.*, 2008).

Afin d'extraire des connaissances relatives aux artistes sur le web, (Alani *et al.*, 2004) applique une analyse syntaxique et sémantique pour la reconnaissance d'entités nommées et l'extraction de relations entre instances. La base de connaissance ainsi peuplée servira à la génération automatique de biographies.

(Amardeilh *et al.*, 2005) propose l'enrichissement d'une base de connaissance (Topics Maps) contrainte par une ontologie du domaine. Pour cela il définit des règles d'acquisition en langage XPath pour mettre en correspondance des extractions linguistiques avec l'ontologie du domaine. Ces règles d'acquisition s'appuient sur le parcours d'arbres conceptuels contenant les informations pertinentes pour le domaine. (Maynard *et al.*, 2008) utilise une méthode linguistique et statistique pour la reconnaissance de termes et utilise des informations contextuelles de trois types : syntaxique, terminologique et sémantique. (Witte *et al.*, 2010) utilise des annotations issues d'analyses syntaxiques pour associer des entité nommées à des instances de concepts et relations.

L'identification de relations conceptuelles dans les textes est un problème majeur dans le processus de peuplement d'une ontologie, cette tâche a suscité plusieurs travaux.

(Makki *et al.*, 2008) décrit l'utilisation d'outils de traitement des langues pour le peuplement d'une ontologie de gestion des risques. En partant du principe que les relations sémantiques entre concepts sont le plus souvent représentées par des verbes, il extrait des relations issues d'une analyse des verbes et de leur entourage afin d'extraire des triplets, en utilisant des listes de verbes synonymes des relations conceptuelles construites à l'aide de WordNet. Dans notre méthode nous construisons des classes sémantiques contenant des termes, et pas uniquement des verbes, dont l'association avec des relations sémantiques ou concepts est propre au domaine d'application.

(Hasegawa *et al.*, 2004) propose une approche fondée sur le regroupement de paires d'entités nommées apparaissant dans des contextes similaires, où chaque paire d'une même classe est considérée comme une instance de la même relation. (Bentibebel *et al.*, 2009) généralise cette approche en construisant des classes d'association de termes représentant des relations conceptuelles potentielles. Une fois ces relations conceptuelles validées, ils y associent des règles d'extraction pour extraire d'autres occurrences.

(Lin et Pantel, 2001) avance que des relations binaires peuvent être représentées par des chemins dans les arbres de dépendances syntaxiques, ces chemins auront des significations similaires s'ils ont tendance à se connecter aux mêmes ensembles de mots. (Nakamura-Delloye et Stern, 2011) exploite ces chemins syntaxiques, pour identifier des relations entre couple d'entité nommées préalablement extraits (Nakamura-Delloye, 2011) et leurs règles. Nous généralisons cette approche pour la définition de règles d'extraction d'instances de relations et de concepts, en exploitant des chemins reliant différent types d'unité textuelles.

3 Ontologie de l'environnement intelligent

L'environnement intelligent consiste en un ensemble d'objets communicants (capteurs, actionneurs et processus de contrôle) qui peut être vu comme un réseau de capteurs. Ces objets influencent le comportement des équipements de l'environnement, sous des conditions bien définies. On peut distinguer une partie matérielle : les différents équipements, leur nombre, leur type, leur localisation etc. et une partie logicielle : la configuration du comportement.

Ces deux aspects permettent de décrire le comportement général d'un environnement intelligent :
- Un capteur détecte l'apparition d'un phénomène ou mesure un phénomène quantifiable dans un espace restreint.
- Un phénomène, pour être détecté ou mesuré par un capteur, doit être localisé dans la zone de capture du capteur et être du type perçu par le capteur (température, mouvement, ...)
- Un actionneur est fixé ou connecté à un appareil de l'environnement.
- Quand un phénomène (ou un ensemble de phénomènes) est mesuré ou détecté, un contrôle des informations collectées est effectué et peut conduire à l'activation d'un ou plusieurs actionneurs pour déclencher une ou plusieurs actions (allumer, éteindre, diminuer, augmenter) sur les appareils auxquels ils sont connectés.
- Un actionneur peut être activé par un capteur (ou un ensemble de capteurs), s'il est localisé dans sa zone (leur zone) de contrôle et gère le(s) même type(s) de phénomène.

La figure 1 représente la conceptualisation du comportement de l'environnement intelligent [2].

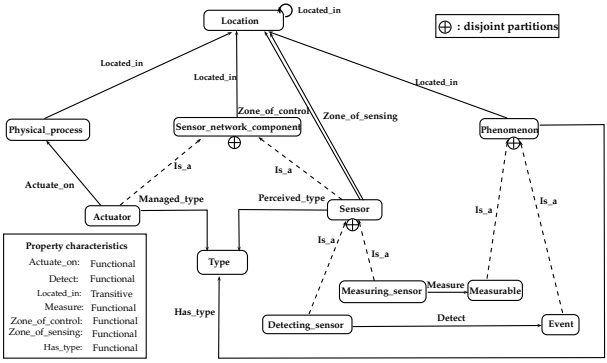


FIGURE 1 – Ontologie d'environnement intelligent

3.1 Rôle de l'ontologie

L'ontologie que nous avons conçue a pour vocation de faciliter le passage de descriptions des besoins à des spécifications formelles, en vue d'une vérification de la consistance et de la

2. L'absence d'un concept utilisateur dans l'ontologie, est dû au fait que les utilisateurs de l'environnement ne sont représenté que par les phénomènes qu'il peuvent engendrer tel qu'une pièce vide ou pleine.

cohérence de la configuration décrite en langage naturel par l'utilisateur de l'environnement intelligent. Elle n'a donc pas vocation à décrire exhaustivement le domaine des environnements intelligents et doit rester à un niveau de description assez élevé pour être la plus générale possible, afin de pouvoir représenter tous les cas de fonctionnement envisageables par un utilisateur. Pour plus de détails sur la conception de l'ontologie voir (Sadoun *et al.*, 2012)

La conceptualisation n'étant pas soumise au changement selon les besoins utilisateurs, la structure de l'ontologie (concepts et relations) a été conçue et figée à partir de l'étude du domaine des environnements intelligents et des réseaux de capteurs. Au contraire, l'instanciation de l'ontologie variera en fonction des descriptions de la configuration physique de l'environnement et des besoins utilisateur, et reflètera pour chaque instanciation un comportement particulier de l'environnement intelligent. La vérification du modèle résultant permettra de détecter d'éventuelles incohérences ou oublis dans les descriptions.

Nous avons décidé d'utiliser comme langage de modélisation, le langage OWL augmenté de SWRL et SQWRL pour sa grande expressivité de représentation des connaissances et ses mécanismes de raisonnement permettant d'effectuer un bon nombre des vérifications nécessaires avant le déploiement du système modélisé.

3.2 Connaissances modélisées

L'ontologie contient deux niveaux : terminologique (*TBox*) et assertionnel (*Abox*). Le premier niveau permet d'exprimer à l'aide d'axiomes des relations entre concepts, ce qui permet par exemple de définir le Domaine et le Range d'une relation ou de marquer la différence entre un capteur qui mesure et un capteur qui détecte.

La relation *Measure* a comme Domaine : $Measuring\_Sensor$ et comme Range : $Measurable$
Un capteur qui mesure est un capteur qui mesure un mesurable.
$Measuring\_Sensor \equiv Sensor \wedge \exists Measure.Measurable$

La relation *Detect* a comme Domaine : $Detecting\_Sensor$ et comme Range : $Event$
Un capteur qui détecte est un capteur qui détecte un évènement.
$Detecting\_Sensor \equiv Sensor \wedge \exists Detect.Event$

Le peuplement portera sur le second niveau de notre ontologie et consistera à créer les assertions identifiées à partir de l'analyse des descriptions. Différents cas de comportement seront modélisés par différentes instanciations, issues des descriptions d'environnement et des besoins utilisateurs. Chaque instanciation est une spécialisation de l'ontologie, qui représente un comportement spécifique d'un environnement donné correspondant aux exigences des utilisateurs.

3.3 Règles de comportement

Des règles SWRL permettent de représenter les règles de comportement dynamique de l'environnement, liant phénomènes et capteurs et capteur et actionneur. Certaines de ces règles sont indépendantes des descriptions, car elles représentent un comportement général du système. D'autres règles sont créées automatiquement lorsqu'elles reflètent un comportement spécifique aux exigences d'utilisateurs.

Les règles SWRL (1),(2),(3) représentent des règles de comportement général de l'environnement, la règle (4) représente une règle spécifique à des spécifications d'exigence. Les deux règles ci-dessous permettent de déduire (ou de créer) les instances de la relation Shared_type(?s, ?e)

Le phénomène p et le capteur s partagent le même type
$Has\_type(?p, ?t), Perceived\_type(?s, ?t) \rightarrow Shared\_type(?p, ?s)$ **(1)**

L'actionneur a et le capteur s partagent le même type
$Managed\_type(?a, ?t), Perceived\_type(?s, ?t) \rightarrow Shared\_type(?p, ?s)$ **(2)**

Ainsi, certaines instances sont issues du processus de peuplement quand d'autres sont déduites lors d'un raisonnement sur les instances présentes dans l'ontologie.

La règle SWRL ci-dessous, stipule que la présence d'un événement *e* et d'un capteur *s* qui partagent le même type, quand *e* est localisé dans la zone de détection de *s*, entraine la détection de *e* par *s* ($Detect(?s, ?e)$)

$Event(?e), Sensor(?s), Shared\_type(s?, ?e), Located\_in(?e, l),$
$Zone\_of\_sensing(?s, l) \rightarrow Detect(?s, ?e)$ **(3)**

Dans la règle spécifique ci-dessous, nous distinguons deux parties, une première fixe et indépendante des textes, et une seconde partie (soulignée dans la règle) générée en fonction des exigences décrites dans les textes.

Lorsqu'un capteur mesure une valeur supérieure à 20 , l'actionneur du même type augmente l'appareil qu'il contrôle.

$Actuator(?a), Physical\_process(?p), Actuate\_on(?a, ?p), Located\_in(?a, ?l), Measure(?s, ?m),$
$Shared\_type(?s, ?a), Zone\_of\_control(?s, ?l), \underline{Has\_value(?m, ?v)}, \underline{lessThanOrEqual(?v, \text{"20"})}$
$\rightarrow \underline{Increase(?a, ?p)}$ **(4)**

4 Descriptions textuelles

Pour piloter son environnement, un utilisateur décrit, selon la configuration physique de son environnement, les fonctionnalités devant être configurées pour satisfaire ses besoins.

1. *Description de l'environnement intelligent :* décrit les composants de l'environnement (capteurs, actionneurs, processus physique, ...), leur nombre, leur type, leur localisation et leur manière d'interagir. Cette partie définit l'état statique de l'environnement et doit être traitée avant les besoins utilisateur.
   Exemple : *L'appartement vert possède un couloir, deux chambres, une salle de bain, et un grand living qui contient une salle à manger et une cuisine. Chaque pièce est équipée de capteurs de mouvement. Chaque ampoule est équipée d'un actionneur.*
2. *Besoins utilisateur* : décrit comment et sous quelles conditions les objets de l'environnement doivent interagir. Cela permet de produire différentes instanciations de l'ontologie selon différents scénarios.
   Exemple : *Quand une personne est dans le living room, y allumer la lumière.*

Le but de notre projet étant la vérification automatique d'exigences, la description de l'environnement qui ne fait qu'énumérer des composants physiques et leurs modes d'utilisation. ne nous

intéressera pas dans la phase d'extraction automatique d'informations pour le peuplement de l'ontologie. Cette description sert néanmoins à initialiser l'ontologie (première instanciation) avec les connaissances liées aux caractéristiques physiques de l'environnement qui seront non soumises aux changements. L'instanciation automatique s'appuiera donc plus particulièrement sur les informations contenues dans les descriptions des besoins utilisateur.

Dans le cadre de notre projet nous partons de descriptions en anglais. Pour la collecte de descriptions de besoins utilisateurs (en anglais et français), nous avons mis en place une plateforme [3] permettant à des participants de décrire un environnement intelligent, en l'occurrence une maison intelligente, dont la configuration physique est représentée par un plan et une description de différents objets de l'environnement et de leur mode d'interaction. Cette description reste néanmoins assez générale de sorte à permettre aux participants d'exprimer différentes idées de configuration correspondant à des besoins variés.

5 Interprétation des exigences utilisateur

Les descriptions utilisateur décrivent le comportement de l'environnement en fonction de la présence de différents phénomènes. Comme les phénomènes possibles ne peuvent généralement pas tous apparaître en même temps, ces descriptions contiennent en fait différents cas de fonctionnement qu'on appelle *interprétation*.

Une interprétation représente l'état de l'environnement intelligent à un instant t, cet état sera déterminé par toutes les instances qui seront présentes dans l'ontologie à l'instant t. Une interprétation ne contiendra pas toutes les instances reconnues dans les textes mais seulement un sous-ensemble cohérent d'instances modélisant une partie des exigences utilisateur.

Ces instances ont une existence conditionnelle, comme cela est illustré par les deux phrases suivantes "*When someone is in the kitchen, turn on the light.*" et "*When nobody is in the kitchen, turn off the light.*". Les instances de relations et de concepts pouvant être extraites de ces phrases devront donc logiquement appartenir à deux interprétations différentes, car il ne peut y avoir au même instant quelqu'un et personne dans la cuisine. Dans le cas contraire les mêmes instances déclencheraient deux règles contradictoires engendrant l'assertion des deux relations Turn_on et Turn_off entre les mêmes instances, ce qui créera une incohérence lors du raisonnement.

L'instanciation doit donc prendre garde à ne pas ajouter toutes les instances extraites en même temps ou du moins celles pouvant être en conflit. Pour le moment, l'instanciation et le raisonnement portent sur les instances d'une phrase à la fois pour éviter les incohérences dues à des instances opposées pouvant déclencher des règles contradictoires. Il y a donc autant d'instanciations que de phrases présentes dans les descriptions. L'identification d'instances en conflit s'inscrit dans nos travaux futurs.

Afin de pouvoir naviguer entre descriptions et ontologie, chaque instance extraite se voit allouer un indice composé de deux numéros, celui de la phrase dont elle est issue ainsi que la position de son nœud dans l'arbre syntaxique.
Exemple : Si l'instance i est issue de l'analyse de la phrase numéro 11 et que le numéro du nœud représentant i dans l'arbre syntaxique est 4, on crée les deux instances de relations $Phrase\_number(i, 11)$ et $Node\_number(i, 4)$.

3. http ://perso.limsi.fr/Individu/sadoun/Application/en/SmartHome.php

Cette référence à l'arbre syntaxique va permettre d'aller dans les descriptions pointer dans les phrases les termes correspondant aux instances qui sont ambiguës ou qui créent des incohérences, pour permettre aux utilisateurs de corriger et d'améliorer leurs exigences.

6 Peuplement de l'ontologie

La figure 2 illustre le processus de peuplement à partir de spécifications en langue naturelle. L'ontologie est d'abord initialisée avec les instances qui décrivent sa configuration physique. L'utilisation des résultats d'une analyse syntaxique des descriptions utilisateurs faite par le Stanford Parser (de Marneffe *et al.*, 2006), de ressources terminologiques et des connaissances déjà présentes dans l'ontologie, permettent d'appliquer des règles pour l'extraction d'instances. Les instances extraites vont peupler l'ontologie et peuvent à leur tour servir à l'identification de nouvelles instances.

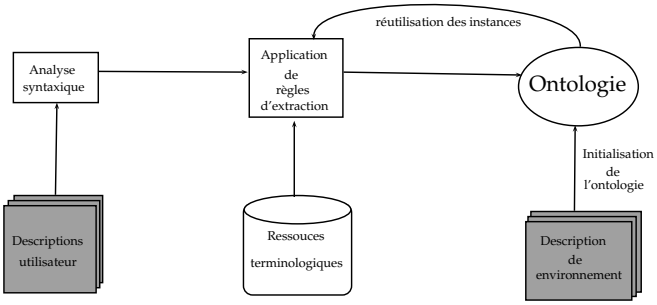


FIGURE 2 – Processus de peuplement de l'ontologie

6.1 Initialisation de l'ontologie

La phase d'initialisation a pour but de peupler l'ontologie avec toutes les instances qui représentent la configuration physique de l'environnement intelligent (instances de capteurs, d'actionneurs, appareils, localisations et relations entre instances). Ces instances sont peu sujettes aux changements, et donc l'initialisation ne se fait qu'une fois via une interface de saisie.

Ci-dessous un exemple d'instanciation du concept $Detecting\_sensor$:

Detecting_sensor(s_1**) :** s_1 **est une instance de** $Detecting\_sensor$**. (**I_1**)**
Perceived_type(s_1**,** $movement$**) :** s_1 **détecte des phénomènes de type** $movement$**. (**I_2**)**
Zone_of_sensing(s_1**,**$kitchen$**) :** s_1 **a comme zone de détection** $kitchen$ **(**I_3**)**
Zone_of_control(s_1**,**$kitchen$**) :** s_1 **a comme zone de contrôle** $kitchen$ **(**I_4**)**

Lors de l'initialisation de l'ontologie, une instance est créée pour chaque capteur avec comme

propriétés son type, sa zone de détection et sa zone de contrôle.

6.2 Extraction des instances

Nous avons choisi de prendre comme point de départ la détection d'instances de relations plutôt que de concepts pour aborder le processus de peuplement, choix qui se justifie, par la liste plus limitée des formulations pouvant représenter des instances de relations dans les corpus. De plus, seule l'extraction des instances de concepts participant à une relation sémantique modélisée dans l'ontologie nous intéresse.

Pour extraire une instance de relation, il faut reconnaitre d'abord des concepts pouvant la refléter, puis vérifier que les concepts auxquels elle est liée dans le texte correspondent aux concepts du Domaine ou du Range de la relation.

Nous distinguons deux types de classes sémantiques, certaines calquées sur les concepts et relations modélisés dans l'ontologie, telle que la classe "Turn on" contenant tous les termes pouvant refléter l'action d'allumer ou de mettre en marche (light on, switch on, open, ...), à l'opposé de la classe "Turn off" (light off, switch off, close, ...) qui reflète l'action d'éteindre ou de mettre à l'arrêt, ou la classe "Location" contenant tous termes correspondant à une instance de localisation issue de la description de l'environnement (hall, kitchen, parent's bedroom, living room, ...). et d'autres classes sémantiques portant sur des connaissances moins spécifiques, telle que la classe Sémantique *Conditional Introducer* (when, if, each time, ...) qui contient les termes introduisant une condition.

L'utilisation de ressources telles que WordNet, VerbNet ou FrameNet s'avère utile pour la construction de ressources terminologiques. Néanmoins, étant basées sur des connaissances générales, les classes sémantiques qu'elles définissent se révèlent être moins pertinentes en domaine spécialisé et ne correspondent pas aux concepts que nous avons définis.

L'ontologie permet de définir un vocabulaire conceptuel sans toutefois contenir le volet terminologique. L'essentiel des connaissances terminologiques que nous utilisons pour l'extraction d'instances de relations et de concepts sont issues de ressources terminologiques extérieures à l'ontologie mais liées aux connaissances qu'elle modélise.

6.3 Extraction des instances

Nous avons choisi de prendre comme point de départ la détection d'instances de relations plutôt que de concepts pour aborder le processus de peuplement, choix qui se justifie par la liste plus limitée des formulations pouvant représenter des instances de relations dans les corpus. De plus, seule l'extraction des instances de concepts participant à une relation sémantique modélisée dans l'ontologie nous intéresse.

Pour extraire une instance de relation, il faut reconnaitre d'abord des termes et structures syntaxiques pouvant la refléter, puis vérifier que les concepts auxquels elle est liée dans le texte correspondent aux concepts du Domaine ou du Range de la relation.

Nous avons donc construit des lexiques permettant de lier termes et concepts. Nous distinguons deux types de classes sémantiques. Certaines sont calquées sur les concepts et relations modélisés

dans l'ontologie, telle que la classe "Turn on" contenant tous les termes pouvant refléter l'action d'allumer ou de mettre en marche (light on, switch on, open, ...), à l'opposé de la classe "Turn off" (light off, switch off, close, ...) qui reflète l'action d'éteindre ou de mettre à l'arrêt, ou la classe "Location" contenant tous termes correspondant à une instance de localisation issue de la description de l'environnement (hall, kitchen, parent's bedroom, living room, ...). D'autres classes sémantiques portent sur des connaissances moins spécifiques, telle que la classe Sémantique *Conditional Introducer* (when, if, each time, ...) qui contient les termes introduisant une condition.

L'utilisation de ressources telles que WordNet, VerbNet ou FrameNet s'avère utile pour la construction de ressources terminologiques. Néanmoins, étant basées sur des connaissances générales, les classes sémantiques qu'elles définissent se révèlent être moins pertinentes en domaine spécialisé et ne correspondent pas directement aux concepts que nous avons définis.

L'ontologie permet de définir le vocabulaire conceptuel sans toutefois contenir le volet terminologique. L'essentiel des connaissances terminologiques que nous utilisons pour l'extraction d'instances de relations et de concepts sont issues de ressources terminologiques extérieures à l'ontologie mais liées aux connaissances qu'elle modélise.

6.4 Construction de règles d'extraction

6.4.1 Chemin syntaxique

La construction de règles d'extraction s'appuie sur une analyse des dépendances syntaxiques. Dans un arbre de dépendances (cf figure 3), les dépendances syntaxiques lient deux nœuds, le premier que nous appellerons nœud *directeur* et le second nœud *dépendant*. Un chemin syntaxique entre deux nœuds est représenté par les dépendances qui les lient.

Par exemple, soit la dépendance syntaxique $subj(is, someone)$

– nsubj : nom de la relation (dépendence sujet nominal)
– is : mot contenu dans le nœud directeur.
– someone : mot contenu dans le nœud dépendant.

Exemple : *When someone is in the kitchen, turn on the light.* La figure 3 représente l'arbre des dépendances syntaxiques de cette phrase [4].

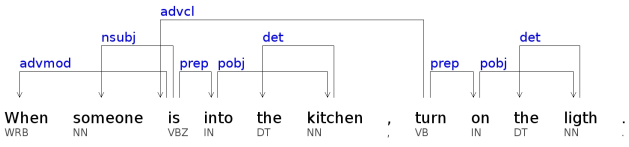


FIGURE 3 – Arbre des dépendances syntaxiques

4. Image produite à l'aide de DependenSee.jar http ://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/

6.4.2 Règles d'extraction

Les règles d'extraction que nous construisons doivent être assez générales pour s'appliquer sur différentes formulations de phrases et assez spécifiques pour n'extraire que les instances de relations et de concepts les plus pertinentes.

L'utilisation de chemins syntaxiques pour reconnaitre la présence d'instances candidates à l'extraction permet de traiter différent types de phrases car les dépendances syntaxiques formant ces chemins peuvent apparaitre dans différentes formulations. La vérification de la pertinence des instances candidates se fait à l'aide d'éléments issus de l'analyse syntaxique des descriptions, des ressources terminologiques, ainsi que des connaissances modélisées dans l'ontologie.

Dans les exemples qui suivent, les termes commençant par des minuscules et contenant des majuscules représentent des fonctions codées en java permettant d'accéder aux ressources terminologiques ou aux connaissances contenues dans l'ontologie.

Règle pour l'extraction d'une instance de la relation Located_in() :

Cette règle identifie tous les chemins syntaxiques existant entre les deux dépendances *prep* [5] et *pobj* [6], en vérifiant qu'il existe un nœud commun p entre elles, que ce nœud est une préposition de lieu ($isPrepLocation(p)$), que le nœud directeur l de la dépendance *pobj* corresponde à une instance du concept *Location* ($isLocation(l)$) existante dans l'ontologie, puis que le nœud dépendant i de la dépendance *prep* a comme catégorie syntaxique un nom ou un verbe.

$$pobj(l, \underline{p}) \wedge prep(\underline{p}, i) \wedge isPrepLocation(p) \wedge isLocation(l) \wedge (isNoun(i) \vee isVerb(i)) \rightarrow Located\_in(i, l)$$

L'application de cette règle à la phrase précédente a pour résultat :

$$pobj(kitchen, \underline{in}) \wedge prep(\underline{in}, is) \wedge isPrepLocation(in) \wedge isLocation(kitchen) \wedge (isNoun(is) \vee isVerb(is)) \rightarrow Located\_in(is, kitchen)$$

Dans le cas où le directeur de la relation Located_in i est issu du verbe auxiliaire *to be* et qu'il existe une relation *nsubj* [7] entre cet auxiliaire et un nom, l'auxiliaire est remplacé par ce nom :

$$Located\_in(i, l) \wedge isToBe(i) \wedge nsubj(i, n) \rightarrow Located\_in(n, l)$$

Ce qui a pour résultat :

$$Located\_in(is, kitchen) \wedge isToBe(is) \wedge nsubj(is, someone) \rightarrow Located\_in(someone, kitchen)$$

L'instance de relation Located_in(someone, kitchen) extraite ne permet pas de typer ou de déduire lors d'un raisonnement sur les instances, le type de *someone* car le domaine de la relation Located_in est égal à tous les concepts (Thing). Aussi, nous décrivons des règles d'extraction d'instances de concepts permettant de désambiguïser des instances qui ne pourraient l'être au sein de l'ontologie.

Règle pour l'extraction d'une instance du concept Event :

Cette règle identifie tous les chemins syntaxiques existant entre les deux dépendances *advmod* [8]

5. pobj : objet de la préposition
6. prep : modificateur prépositionnel
7. nsubj : sujet nominal
8. advmod : modificateur adverbial

et *nsubj*, en vérifiant qu'il existe un nœud commun *v* entre elles, que ce nœud est issu du verbe auxiliaire *to be*, que le nœud dépendant *c* de la dépendance *advmod* appartient à la classe Sémantique *Conditional Introducer* et que le nœud dépendant *n* de la dépendance *nsubj* participe déjà à une instance de la relation *Located_in* dans l'ontologie, car un évènement n'a aucun intérêt si l'on ne connait pas sa localisation.

$advmod(v,c) \wedge nsubj(v,n) \wedge isToBe(v) \wedge IsCondIntroducer(c) \wedge Located\_in(n,l) \rightarrow Event(n)$

Ce qui a pour résultat :

$advmod(is,When) \wedge nsubj(is,someone) \wedge isToBe(is) \wedge IsCondIntroducer(When) \wedge Located\_in(someone,kitchen) \rightarrow Event(someone)$

Règle pour l'extraction d'une instance de la relation Has_type :

La règle ci-dessous identifie la présence des deux instances *Event* et *Located_in* issues de l'ontologie, contenant l'instance commune *n* dénotant le type t (information présente dans nos ressources terminologiques et sémantiques)

$Event(n) \wedge Located\_in(n,l) \wedge denoteType(n,t) \rightarrow Has\_type(n,t)$

Le terme *someone* en tant qu'évènement dénote le type *movement*, on donc a pour résultat :

$Event(someone) \wedge Located\_in(someone,kitchen) \wedge denoteType(someone,movement) \rightarrow Has\_type(someone,movement)$

A l'aide de ces règles, nous obtenons des instances de relations et de concepts qui sont ajoutées à l'ontologie et qui peuvent être réutilisées dans d'autres règles.

L'ordre d'application des règles d'extraction n'a pas de conséquences sur le résultat du peuplement. Néanmoins comme l'application de certaines règles nécessite des résultats (instances) issues de l'application d'autres règles, il est préférable d'ordonner leur application pour optimiser le processus.

Les règles varient en fonction des instances à extraire, mais aussi pour une même instance en fonction de toutes les formulations. Dans un premier temps, nous complèterons celles-ci pour définir les plus courantes, afin de valider notre méthodologie d'instanciation de l'ontologie. Nous étudierons ensuite comment disposer d'un plus large corpus si le nombre de descriptions collectées ne nous permet pas d'avoir recours à des méthodes d'acquisition automatique.

7 Discussion

Dans les descriptions utilisateur, les références aux capteurs ne sont pratiquement jamais présentes, ces références sont souvent implicites, réduisant le capteur à son fonctionnement, ce que l'on voit bien dans la phrase : *When someone is in the kitchen, turn on the light.*

Or, une instance de relation que nous souhaitons extraire est la relation *Detect*, qui représente la détection d'un événement par un capteur.

C'est là que réside tout l'intérêt de notre modèle, qui ne s'appuie pas seulement sur les informations présentes dans les descriptions, mais tient aussi compte de connaissances issues du

domaine, qui, le cas échéant, permettent de lever les ambiguïtés. L'ontologie modélise ainsi le fait qu'un capteur est identifiable par sa zone de détection et le type d'événement qu'il peut détecter. Aussi la reconnaissance d'instances de ces concepts permettra de déduire le capteur concerné.

Nous allons détailler ce processus pour l'exemple :
Les connaissances issues des initialisations I_1, I_2, I_3 (Section 6.1) font que le déclenchement de la règle SWRL (3) (Section 3) n'est plus soumis qu'à l'extraction des instances suivantes : Located_in(someone,kitchen), Event(someone), Has_type(someone,movement), car le capteur devant intervenir dans la règle, sera déduit automatiquement.

Instanciation de la règle SWRL (3)
$Event(someone), Sensor(s1), Shared\_type(s1, someone), Located\_in(someone, kitchen),$
$Zone\_of\_sensing(s1, kitchen) \rightarrow Detect(s1, someone)$

8 Conclusion et travaux futurs

Nous avons décrit une méthode pour le peuplement automatique d'une ontologie modélisant le comportement d'un environnement intelligent, laquelle à pour but de faciliter le passage de descriptions de besoins utilisateurs en langue naturelle vers des spécifications formelles de ces besoins.

Cette problématique est distincte des approches habituelles pour le peuplement d'ontologie à partir de textes qui ont pour but de peupler une ontologie avec toutes les instances extraites des textes permettant de décrire un domaine. L'approche adoptée dans cet article vise à extraire uniquement les instances participant à la modélisation d'un des multiples comportements d'un système de composants logiciels.

Notre méthode est guidée par l'extraction d'instances de relations permettant par là-même d'extraire les instances de concepts liées par ces relations. L'extraction d'instances se fait à l'aide de règles construites manuellement à partir de résultats d'analyse syntaxique, de ressources terminologiques et des connaissances modélisées dans l'ontologie. Ainsi seules les instances les plus pertinentes, car participant à au moins une relation conceptuelle, sont extraites. Le recours aux connaissances modélisées par l'ontologie offre un cadre formel pour la résolution des ambiguïtés.

La modélisation résultante permet la navigation entre descriptions des exigences en langue naturelle et spécifications formelles, pour la vérification de ces exigences et leur amélioration.

Nos travaux futurs porteront sur l'approfondissement de l'acquisition de règles, la création automatique de règles SWRL spécifiques au comportement décrit dans les descriptions utilisateurs et l'identification d'instances pouvant être en conflit et dont la présence simultanée dans une même instanciation serait incohérente.

Références

ALANI, H., KIM, S., MILLARD, D. E., WEAL, M. J., HALL, W., LEWIS, P. H. et SHADBOLT, N. (2004). Using protege for automatic ontology instantiation. *In 7th International Protégé Conference*.

ALANI, H., KIM, S., MILLARD, D. E., WEAL, M. J., HALL, W., LEWIS, P. H. et SHADBOLT, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*.

AMARDEILH, F., LAUBLET, P. et MINEL, J.-L. (2005). Document annotation and ontology population from linguistic extractions. *In Proceedings of the 3rd international conference on Knowledge capture*.

BENDAOUD, R., ROUANE HACENE, M., TOUSSAINT, Y., DELECROIX, B. et NAPOLI, A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. *In IC 2007*.

BENTIBEBEL, R., NAZARENKO, A. et SZULMAN, S. (2009). Mise en lumière de relations sémantiques pour la construction d'ontologies à partir de textes. *In TIA 2009*.

DE BOER, V., VAN SOMEREN, M. et WIELINGA, B. J. (2007). Relation instantiation for ontology population using the web. *In Proceedings of the 29th annual German conference on Artificial intelligence*.

de MARNEFFE, M.-C., MACCARTNEY, B. et MANNING, C. D. (2006). Generating typed dependency parses from phrase structure trees. *In LREC*.

HASEGAWA, T., SEKINE, S. et GRISHMAN, R. (2004). Discovering relations among named entities from large corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

LIN, D. et PANTEL, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*.

MAKKI, J., ALQUIER, A.-M. et PRINCE, Vp, V. (2008). Ontology Population via NLP Techniques in Risk Management. *In ICSWE : Fifth International Conference on Semantic Web Engineering*.

MAYNARD, D., LI, Y. et PETERS, W. (2008). Nlp techniques for term extraction and ontology population. *In Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*.

NAKAMURA-DELLOYE, Y. (2011). Named entity extraction for ontology enrichment. *In IPSJ Special Interest Group - Information Fundamentals and Access Technologies (IFAT)*.

NAKAMURA-DELLOYE, Y. et STERN, R. (2011). Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie. *In TOTh 2011 : Terminologie & Ontologie : Théories et Applications*.

SADOUN, D., DUBOIS, C., GHAMRI-DOUDANE, Y. et GRAU, B. (2011). An ontology for the conceptualization of an intelligent environment and its operation. *Mexican International Conference on Artificial Intelligence*.

SADOUN, D., DUBOIS, C., GHAMRI-DOUDANE, Y. et GRAU, B. (2012). Formalisation en OWL pour vérifier les spécifications d'un environnement intelligent. *In Actes de la conférence RFIA 2012*.

WANG, F. et TURNER, K. J. (2009). An ontology-based actuator discovery and invocation framework in home care systems. *In Proceedings of the 7th International Conference on Smart Homes and Health Telematics : Ambient Assistive Health and Wellness Management in the Heart of the City*.

WITTE, R., KHAMIS, N. et RILLING, J. (2010). Flexible ontology population from text : The owlexporter. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances

Andon Tchechmedjiev
LIG-GETALP
Laboratoire d'Informatique de Grenoble-Groupe d'Étude pour la Traduction Automatique/Traitement Automatisé des Langues et de la Parole
Université de Grenoble
andon.tchechmedjiev@imag.fr

RÉSUMÉ

Dans cet article, nous présentons les principales méthodes non supervisées à base de connaissances pour la désambiguïsation lexicale. Elles sont composées d'une part de mesures de similarité sémantique locales qui donnent une valeur de proximité entre deux sens de mots et, d'autre part, d'algorithmes globaux qui utilisent les mesures de similarité sémantique locales pour trouver les sens appropriés des mots selon le contexte à l'échelle de la phrase ou du texte.

ABSTRACT

State of the art : Local Semantic Similarity Measures and Global Algorithmes for Knowledge-based Word Sense Disambiguation

We present the main methods for unsupervised knowledge-based word sense disambiguation. On the one hand, at the local level, we present semantic similarity measures, which attempt to quantify the semantic proximity between two word senses. On the other hand, at the global level, we present algorithms which use local semantic similarity measures to assign the appropriate senses to words depending on their context, at the scale of a text or of a corpus.

MOTS-CLÉS : désambiguïsation lexicale non-supervisée, mesures de similarité sémantique à base de connaissances, algorithmes globaux de propagation de mesures locales.

KEYWORDS: unsupervised word sense disambiguation, knowledge-based semantic similarity measures, global algorithms for the propagation of local measures.

1 Introduction

Les ambiguïtés font partie intégrante des langues naturelles, mais les humains ont la capacité, dans la plupart des cas et en s'aidant du contexte, à désambiguïser sans trop d'efforts. Cependant, pour le traitement automatique des langues naturelles, cette ambiguïté pose problème, et il est fondamental de trouver des méthodes pour affecter aux mots les sens corrects vis à vis du contexte.

Il existe différentes approches pour résoudre ce problème. Elle se divisent en deux catégories principales : d'une part les approches supervisées, nécessitant des corpus d'entraînement étiquetés manuellement et, d'autre part, des approches non-supervisées (Navigli, 2009).

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 295–308,
Grenoble, 4 au 8 juin 2012.

Le problème avec les algorithmes supervisés est le fait qu'obtenir de grandes quantités de texte annoté en sens est très couteux en temps et en argent, et que l'on se heurte au goulot d'acquisition de données (Wagner, 2008). De plus, la qualité de la désambiguisation de ces approches est restreinte par les exemples utilisés pour l'entrainement.

C'est pourquoi les méthodes non supervisées sont intéressantes. Elles n'utilisent pas de corpus annotés. Il existe là aussi des distinctions : d'une part les approches non supervisées classiques (clustering) qui exploitent les données non annotées ; et d'autre part les approches à base de savoirs qui utilisent des connaissances issues de ressources lexicales.

Nous nous intéressons ici à ces dernières. Il y a différents aspects à considérer dans le cadre des approches non-supervisées à base de connaissances : d'abord la question essentielle des ressources lexicales qu'il est possible d'utiliser, ensuite la question de comment exploiter la, ou les ressources lexicales pour désambiguïser.

Ce dernier aspect se présente sous deux dimensions : la dimension locale où l'on cherche à déterminer la proximité entre les sens possibles des différents mots et, la dimension globale où l'on cherche à affecter les bons sens aux mots à l'échelle d'un texte. Il existe à la fois des méthodes qui *propagent* les mesures locales en les utilisant pour évaluer les combinaisons de sens, mais aussi des méthodes purement globales qui exploitent directement la structure linguistique de l'ensemble du texte sans s'intéresser aux sens individuellement.

Nous présenterons, les principales ressources lexicales (Section 2), les principales mesures de similarité sémantique (Section 3) puis une description de quelques algorithmes globaux qui exploitent ces mesures (Section 4). Nous terminerons par des considérations sur l'évaluation et la comparaison de ces algorithmes (Section 5).

Pour un état de l'art complet et plus détaillé, le lecteur se référera à (Ide et Veronis, 1998) et plus récemment (Navigli, 2009).

2 Ressources lexicales

Une caractéristique des approches à base de connaissances est qu'elles utilisent des ressources lexicales. Un premier type de ressource qui peut être exploitée est l'inventaire de sens, c'est-à-dire une ressource qui, à chaque mot, lie une liste de sens possibles comme par exemple, un dictionnaire (par exemple (Collins, 1998)). D'autre part, des ressources telles que les thésaurus (par exemple (Roget, 1989)) peuvent être utiles pour établir des liens entre les sens des différents mots.

Par ailleurs, des ressources lexicales telles que WordNet (Miller, 1995) sont structurées et jouent le rôle d'inventaires de sens et de dictionnaires, mais donnent également accès à une hiérarchie de sens (en quelque sorte un thésaurus structuré).

La majorité des mesures de similarité que nous allons présenter se basent sur Wordnet [1].

WordNet est structuré autour de la notion de synsets, c'est-à-dire en quelque sorte un ensemble de *synonymes* qui forment un concept. Un synset représente un sens de mot. Les synsets sont reliés entre eux par des relations, soit lexicales (antonymie par exemple) ou taxonomiques (hyperonymie, méronymie, etc).

1. Il est possible de les utiliser sur d'autres ressources également.

3 Mesures de similarité Sémantique à base de connaissances

Parmi les mesures de similarité sémantique on retrouve trois types principaux que nous allons maintenant décrire. Il faut noter que les mesures de de similarité géométriques ne sont pas à base de connaissances et ne seront pas présentées.

3.1 À base de traits

3.1.1 Similarité de Tsversky

Avant d'être abordée en TALN, la notion de similarité sémantique a été traitée dans le domaine de la psychologie cognitive. Un travail souvent cité est (Tversky, 1977) qui propose une nouvelle approche basée sur le recouvrement ou non de traits entre deux objets. Plus précisément, est considéré comme concept ou signification rattachée à un objet toute propriété dudit objet. La similarité entre deux objets est exprimée comme le nombre pondéré de propriétés en commun, auxquelles on retire le nombre pondéré de propriétés spécifiques à chaque objet. Il propose donc un modèle de similarité non symétrique, que l'on appelle «modèle de contraste» (Figure 1).

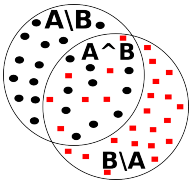


FIGURE 1 – Le modèle de contraste entre deux concepts

Plus formellement, si l'on reprend la notation de (Pirró et Euzenat, 2010) où $\Psi(c)$ est l'ensemble des traits se rapportant à un sens s, alors la similarité de Tsversky peut s'exprimer par : $sim_{tvr}(s_1,s_2) = \theta F(\Psi(s_1) \cap \Psi(s_2)) - \alpha F(\Psi(s_1) \setminus \Psi(s_2)) - \beta F(\Psi(s_2) \setminus \Psi(s_1))$ où F est une fonction qui associe une pertinence aux traits, et où θ, α et β sont des facteurs qui marquent respectivement l'importance relative de la similarité entre les sens, des dissimilarités entre s_1 et s_2 et des dissimilarités entre s_2 et s_1, et où $\setminus$ est l'opérateur de différence ensembliste.

Il est également possible d'exprimer cette mesure en tant que rapport, afin d'avoir une valeur de similarité normalisée (avec $\theta = 1$) :,

$$sim_{tvtr}(c_1,c_2) = \frac{F(\Psi(c_1) \cap \Psi(c_2))}{F(\Psi(c_1) \cap \Psi(c_2)) + \alpha F(\Psi(c_1) \setminus \Psi(c_2)) + \beta F(\Psi(c_2) \setminus \Psi(c_1))}$$

Comme récapitulé dans (Pirró et Euzenat, 2010), différentes valeurs de α et de β mènent à différents types de similarité. Si $\alpha = \beta = 0$, on ne s'intéresse qu'aux points communs entre les deux sens. Si $\alpha > \beta$ ou $\alpha < \beta$ alors on s'intéresse assymétriquement à la similarité de s_1 avec s_2 ou vice versa. Si $\alpha = \beta \neq 0$ on s'intéresse à la similarité mutuelle entre s_1 et s_2. Quand $\alpha = \beta = 1$ la mesure de Tversky est équivalente à la similarité de Tanimoto (Rogers et Tanimoto, 1960). Dans le cas où $\alpha = \beta = 0.5$ alors elle est équivalente au coefficient de Dice (Dice, 1945).

3.1.2 Similarité de Lesk

(Lesk, 1986) a proposé un algorithme de désambiguïsation lexicale très simple, qui considère la similarité entre deux sens comme le nombre de mots en commun dans leurs définitions. Dans la version originale, on ne prend pas en compte l'ordre des mots dans les définitions (sac de mots). Dans ce cadre là, il apparaît que cette méthode puisse être ramenée à un cas particulier de la similarité de Tsversky (en tant que rapport ou non), en considérant que les concepts sont des sens de mots, que les traits sont des mots de la définition des sens, avec $\alpha = \beta = 0$, et avec $\Psi(s) = D(d)$ qui retournant un ensemble contenant les mots de la définition d'un sens de mot s. Quant à la fonction F on la choisit comme la fonction cardinalité d'ensemble. On obtient ainsi : $sim_{lesk}(s_1, s_2) = \mid D(s_1) \cap D(s_2) \mid$

L'avantage de cette mesure de similarité est qu'elle est extrêmement simple à calculer, et ne requiert qu'un dictionnaire. Dans le contexte de l'algorithme de Lesk original, la similarité était calculée de manière exhaustive entre tous les sens de tous les mots du contexte, il existe une variante (Navigli, 2009) utilisée sur une fenêtre de contexte autour du mot auquel appartient le sens. Elle correspond au recouvrement entre la définition du sens et entre un sac de mot contenant tous les mots des définitions des mots du contexte : $Lesk_{var} = \mid contexte(w) \cap D(s_{w_n}) \mid$. Comme le met en avant (Navigli, 2009), un problème important de la mesure de Lesk est qu'elle est très sensible aux mots présents dans la définition, et si certains mots importants manquent dans les définitions utilisées, les résultats obtenus seront de qualité moindre. De plus si les définitions sont trop concises (comme c'est souvent le cas) il est difficile d'obtenir des distinctions de similarité fines.

Cependant, un certain nombre d'améliorations de la mesure de Lesk ont été proposées.

3.1.3 Extensions de la mesure de Lesk

Tout d'abord, (Wilks et Stevenson, 1998) proposent de pondérer chaque mot de la définition par la longueur de celle-ci afin de donner la même importance à toutes les définitions, au lieu de systématiquement privilégier les définitions longues.

Plus récemment (Banerjee et Pedersen, 2002) ont proposé la mesure de "Lesk étendu", qui améliore Lesk de deux façons. La première est l'incorporation des définitions des sens reliés par des relations taxonomiques WordNet dans la définition d'un sens donné. La deuxième est une nouvelle manière de calculer le recouvrement entre les mots des définitions.

Pour calculer le recouvrement entre deux sens, ils proposent de considérer le recouvrement entre les définitions des deux sens mais aussi des définitions issues de différentes relations : hyperonymes (`has-kind`), hyponymes (`kind-of`), meronymes (`part-of`), holonymes (`has-part`), troponymes mais aussi par les relations `attribute`, `similar-to`, `also-see`.

Afin de garantir que la mesure soit symétrique, ils proposent de prendre les combinaisons deux à deux des relations considérées et de ne conserver une paire de relations (R_1, R_2) que si la paire inverse (R_2, R_1) est présente. On obtient ainsi un ensemble $RELPAIRS$. De plus, le recouvrement entre deux définitions A et B se calcule comme la somme des carrés des longueurs de toutes les sous-chaines de mots de A dans B, ce que l'on exprime avec l'opérateur $\Cap$. Nous avons ainsi : $Lesk_{etendu}(s_1, s_2) = \sum_{\forall (R_1,R_2) \in RELPAIRS^2} (\mid D(R_1(s_1)) \Cap D(R_2(s_2)) \mid)$

Le calcul du recouvrement est basé sur le principe relevé par la loi de Zipf (Zipf, 1949), qui met en évidence une relation quadratique entre la longueur d'une phrase et sa fréquence d'occurrence dans un corpus. De ce fait, n mots qui apparaissent ensemble portent plus d'informations que si ils étaient séparés.

3.2 À base de distance taxonomique

Le principe des mesures à base de distance taxonomique est de compter le nombre d'arcs qui séparent deux sens dans une taxinomie.

La Figure 2 (Wu et Palmer, 1994) représente la relation de deux sens quelconques S_1 et S_2 dans une taxinomie par rapport à leur sens commun le plus spécifique S_3 et par rapport à la racine de la taxinomie ; cette figure servira à exprimer de manière homogène les formules des différentes mesures de similarité.

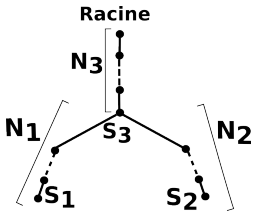


FIGURE 2 – Deux sens et leur sens commun le plus spécifique dans une taxinomie

La mesure de Rada (Rada *et al.*, 1989) est la première à utiliser la distance entre les nœuds correspondant aux deux sens sur les liens d'hyponymie et hyperonymie :

$$Sim_{Rada}(s_1,s_2) = d(s_1,s_2) = N_1 + N_2$$

Les termes se trouvant plus profondément dans la taxonomie étant toujours plus proches que les termes plus généraux, (Wu et Palmer, 1994) proposent de prendre en compte la distance entre l'ancêtre commun le plus spécifique et la racine pour y remédier.

$$Sim_{WuP} = \frac{2 \cdot N_3}{N_1 + N_2 + 2 \cdot N_3}$$

(Leacock et Chodorow, 1998) se basent également sur la mesure de Rada, mais au lieu de normaliser par la profondeur relative de la taxinomie par rapport aux sens, ils choisissent une normalisation par rapport à la profondeur totale de la taxinomie D et normalisent avec un logarithme :

$$Sim_{LCH} = -log(\frac{N_1 + N_2}{2 \cdot D})$$

(Hirst et St-Onge, 1998) adaptent le concept de chaines lexicales développées par (Morris et Hirst, 1991) comme mesure de similarité sémantique en utilisant la structure de WordNet. Cette mesure se base sur l'idée de (Halliday et Hasan, 1976) que dans un texte, des mots ont une forte probabilité de référer à des mots antérieurs ou à d'autres concepts reliés, et que l'enchainement de ces mots forment des chaines cohésives. Par exemple, (Navigli, 2009) `Rome->ville->habitant` et `manger->plat->légume->aubergine`, sont des chaines lexicales.

À chaque relation est associée une direction horizontale, ascendante ou descendante, qui marque respectivement une relation forte, très forte et moyennement forte (par exemple l'hyperonymie

est une relation ascendante, l'holonymie une relation descendante, et l'antonymie une relation horizontale). Les changements de direction constituent un élément de dissimilarité et la proximité dans la taxinomie un élément de similarité. Un changement de direction est défini comme le passage d'un élément A de la taxinomie à un élément B par une relation d'un autre type que celle qui a permis d'arriver sur A. Notons que plus la distance entre les sens est grande, plus il y aura de changements de direction potentiels.

Même si l'algorithme proposé est un algorithme global, il est possible d'utiliser la fonction d'évaluation des chaines lexicales en tant que mesure de similarité.

Soient C et k deux constantes et soit la fonction $virages(s_1, s_2)$ qui retourne le nombre de changements de direction entre les sens s_1 et s_2, alors la mesure de similarité s'exprime :

$Sim_{Hso} = C - (N_1 + N_2) - k \cdot virages(s_1, s_2)$

Il existe également d'autres mesures exploitant la structure taxonomique, mais en pondérant les arcs avec des valeurs de contenu informationnel, notion que nous allons maintenant définir.

3.3 À base de contenu informationnel

L'approche de (Resnik, 1995) est basée sur la détermination du contenu informationnel du concept commun le plus spécifique à deux sens. Dans la figure 2, le concept commun le plus spécifique à S_1 et S_2 est S_3 (notée $lso(S_1, S_2) = S_3$ pour *lowest superordinate*). Quant à la quantité d'information, elle est calculée à partir de la probabilité $p(S_3)$: $IC(s) = -log(P(S_3))$

Les probabilités d'occurrence de chaque concept de la taxinomie sont calculées à partir d'un corpus non-annoté par estimation du maximum de vraisemblance. Ainsi la mesure similarité de Resnik s'exprime :

$$Sim_{Res} = IC(lso(S_1, S_2)) = IC(S_3)$$

(Jiang et Conrath, 1997) partent du constat qu'utiliser seulement l'ancêtre commun n'offre pas une granularité assez fine et proposent de prendre en compte la quantité d'information portée par les deux sens. Cette mesure s'exprime :

$$Sim_{JCN} = IC(S_1) + IC(S_2) + 2 \cdot IC(lso(S_1, S_2))$$

(Lin 1998) propose également un mesure de similarité très proche, qui revient essentiellement à une reformulation sous forme de rapport de la formule de Jiang and Conrath :

$$Sim_{Lin} = \frac{2IC(lso(S_1, S_2))}{IC(S_1) + IC(S_2)}$$

Plus récemment, (Seco *et al.*, 2004) ont argumenté qu'il est possible d'extraire directement de WordNet les valeurs de contenu informationnel sans avoir à passer par un corpus. La taxinomie de WordNet étant structurée à partir de principes psycholinguistiques, on peut faire l'hypothèse que cette structure (liens d'hyperonymie et d'hyponymie) est représentative du contenu informationnel ; c'est-à-dire que les concepts qui ont beaucoup d'hyponymes portent une quantité d'information moins importante que les concepts *feuilles*. Si l'on note $hypo(s)$ une fonction qui retourne le nombre d'hyponymes d'un concept s et max_{wn} le nombre total

de concepts dans la taxinomie, alors on peut exprimer le contenu informationnel intrinsèque : $iIC(s) = 1 - \frac{log(hypo(s)+1)}{log(max_{wn})}$

On peut substituer iIC à IC dans toutes les mesures précédentes pour supprimer le besoin d'apprentissage non supervisé. Cependant, l'iIC est limitée car, elle n'exploite qu'un seul type de relations, alors que d'autres pourraient être intéressantes. C'est pourquoi (Pirró et Euzenat, 2010) proposent une mesure d'iIC étendue qui va prendre en compte les autres relations présentes (meronymie par exemple). Ils l'expriment comme : $eIC(s) = \zeta iIC(s) + \eta EIC(s)$

où ζ et η sont des constantes et où $EIC(s)$ est la somme des iIC moyennes des concepts reliés à s par les autres relations. Si l'on note $Rels(s)$ l'ensemble des relations possibles pour s, et si pour une relation $s_r \in Rels(s)$, $s_r(s)$ est l'ensemble des concepts reliés à s par s_r, alors on obtient : $EIC(s) = \sum_{s_r \in Rels(s)} \frac{\sum_{c \in s_r(s)} iIC(c)}{|s_r(s)|}$

Il est également possible de substituer eIC à IC.

Plus récemment ont commencé à apparaître des mesures hybrides qui, soit essayent de combiner différents types de mesures de similarité, soit essayent d'exploiter la structure de plusieurs ontologies (*cross-ontology similarity*).

3.4 Mesures hybrides

Ici, nous nous focalisons sur l'aspect combinaisons de mesures plutôt que sur les approches à ontologies croisées.

(Li et al 2003) proposent une mesure qui combine à la fois la distance taxonomique ($l = N_1 + N_2$), la profondeur du concept commun le plus spécifique dans la taxinomie ($h = N_3$) ainsi que la densité sémantique locale($d = IC(lso(s_1, s_2))$) , cette dernière étant exprimée en terme de contenu informationnel. Leur mesure est exprimée par : $Sim_{Li}(s_1, s_2) = f(f_1(l), f_2(h), f_3(d))$ où f_1, f_2 and f_3 sont les fonctions de transfert non-linéaires respectives pour chaque type d'information.

Le but des fonctions de transfert est de normaliser dans l'intervalle $[0; 1]$ les mesures pour qu'elles puissent être combinées. $f_1(l) = e^{-\alpha l}$ où α est une constante. $f_2(h) = (e^{\beta h} - e^{-\beta h}) \div (e^{\beta h} + e^{-\beta h})$ où $\beta > 0$ est un facteur de lissage. $f_3(d) = (e^{\lambda d} - e^{-\lambda d}) \div (e^{\lambda d} + e^{-\lambda d})$ avec $\lambda > 0$. Quant à la fonction f, elle constitue n'importe quelle combinaison de ces trois mesures, et est à choisir selon les applications et la nature des sources d'information disponibles.

FaITH, une autre mesure locale qui combine des aspects de différents types, est proposée par (Pirró et Euzenat, 2010) sous la forme de l'extension des mesures à base de contenu informationnel en reprenant le modèle de contraste de Tsversky :

$$Sim_{FaITH} = \frac{IC(lso(s_1, s_2))}{IC(s_1) + IC(s_2) - IC(lso(s_1, s_2))}$$

Ici la fonction F est remplacée par IC, les traits communs aux concepts sont représentés par le contenu informationnel du concept commun le plus spécifique, et les traits spécifiques à un concept sont représentés par la différence entre le contenu informationnel de ce concept auquel on soustrait le contenu informationnel du concept commun le plus spécifique.

4 Algorithmes globaux de désambiguïsation lexicale

Maintenant que nous avons passés en revue les principales de mesures de similarité sémantique, nous allons présenter différents algorithmes qui les utilisent comme heuristiques pour évaluer des combinaisons de sens.

4.1 Approche exhaustive

L'approche originellement adoptée par (Lesk, 1986) pour désambiguïser un texte en entier, est d'évaluer toutes les combinaisons possibles de sens et de choisir la combinaison qui maximise le score du texte – exprimé comme la somme des scores des sens choisis par rapport aux autres mots du texte.

En d'autres termes, si le sens sélectionné d'un mot w dans une combinaison C est S_w et un texte T une liste ordonnée de mots, alors le score de la combinaison est $score(C) = \sum_{w_i \in T} \sum_{w_j \in T} sim(S_{w_i}, S_{w_j})$ et il y a en tout $\prod_{w \in T} N_w$ combinaisons à évaluer, avec N_w le nombre de sens de w (Gelbukh *et al.*, 2005), c'est-à-dire un nombre exponentiel de combinaisons.

Par exemple pour une phrase de 10 mots avec 10 sens en moyenne par mot il y aurait $10^{10^2} = 10^{100}$ combinaisons.

Pour diminuer le temps de calcul on peut utiliser une fenêtre autour du mot afin de réduire le temps d'évaluation d'une combinaison au prix d'une perte de cohérence globale de la désambiguïsation.

Une autre approche est d'utiliser des meta-heuristiques d'optimisation combinatoire pour obtenir des solutions de qualité convenable d'une manière qui soit traitable calculatoirement.

4.2 Recuit simulé

Le recuit simulé est une méthode d'optimisation stochastique classique, et fût appliqué à la désambiguïsation lexicale par (Cowie *et al.*, 1992).

Le principe est de faire des changements aléatoires dans la configuration [2] itérativement de l'espace de recherche puis d'évaluer si le changement est bénéfique. Dans le cas échéant, il est conservé, sinon, il y a une probabilité de le conserver quand même.

Cette évaluation se fait en utilisant une métrique heuristique, et dans le cas de la désambiguïsation, on utilise les mesures de similarité pour jouer ce rôle. Le score d'une configuration se calcule de la même manière que pour l'évaluation d'une combinaison dans le cas de l'approche exhaustive.

Quant à la probabilité de conserver une configuration inférieure, elle se calcule par rapport à la différence des scores entre la configuration modifiée(C') et la configuration avant modification (C) avec $\Delta s = score(C') - score(C)$ et un paramètre de température T : $P(conservation) = e^{\frac{-\Delta s}{T}}$.

Dans le cas d'une descente par gradients où l'on ne garde que les meilleures configurations, on est confronté à un problème de convergence sur des maxima locaux. L'objectif de l'acceptation possible des configurations inférieures pour le recuit simulé est de leur échapper. Cependant cela pourrait mener à une non convergence du système, c'est pourquoi la diminution géométrique de la température T permet progressivement de se ramener à une descente de gradient, dont la convergence est garantie.

2. Une configuration est représentée par un vecteur d'entiers correspondant aux numéros de sens sélectionnés pour chaque mot dans l'ordre d'apparition des mots dans le texte.

4.3 Algorithme génétique

Les algorithmes génétiques sont inspirés de l'évolution génétique des espèces et sont utilisés pour trouver des solutions à des problèmes d'optimisation combinatoire.

(Gelbukh *et al.*, 2003) l'ont appliqué à la désambiguïsation lexicale. La représentation de la configuration utilisée est la même que pour le recuit simulé, cependant, on considère une population de λ configurations (chromosomes). Chaque indice du vecteur d'une configuration est considéré comme un allèle, et les allèles possibles pour un indice sont les différents sens du mot en question.

Le déroulement de l'algorithme est inspiré des cycles reproductifs et de sélection naturelle des espèces. La *qualité* d'un individu est estimée avec une fonction de score heuristique, ici la même que pour le recuit simulé.

À chaque cycle, les scores de tous les individus sont calculés, et un nombre pair d'individus sont sélectionnés de manière probabiliste pour être croisés : $\forall \lambda_i \in \lambda,\ p(crois_{\lambda_i}) = Cr * \left(\frac{score(\lambda_i)}{score_{max}}\right)$, où $score_{max}$ est le meilleur score dans la population et où Cr est un rapport de sélection.

Le croisement s'effectue par une permutation autour d'un ou plusieurs points de pivots choisis au hasard dans la configuration, habituellement un ou deux. Les individus non retenus pour le croisement sont dupliqués. On obtient ainsi une nouvelle population (qui remplace l'ancienne). Sur chaque individu, on applique avec une probabilité $p(M)$, Mn changements aléatoires uniformes. Le score de la nouvelle population est calculé après la phase de mutation, puis un nouveau cycle commence.

Parmi les stratégies de convergence, on trouve un nombre fixe de cycles où encore une stabilisation de la distribution des scores de la population pendant plusieurs cycles successifs.

4.4 Chaines lexicales

Comme décrit dans la Section 3.2 (Hirst et St-Onge, 1998) est un algorithme global qui se base sur la construction de chaines lexicales afin d'évaluer les combinaisons en intégrant des connaissances linguistiques pour réduire l'espace de recherche.

Ils placent tout d'abord des restrictions sur les enchainements de types de liens possibles : il est impossible d'avoir plus d'un changement de direction ; un lien ascendant est terminal, sauf si il est suivi d'un lien horizontal faisant le lien avec un lien descendant.

La chaine lexicale globale est construite dans l'ordre des mots du le texte. Lorsqu'un mot est inséré (présent dans le texte, ou transitivement par une relation), un certain nombre de ses *synsets* lui sont reliés : si c'est le premier mot de la chaine ou si le mot provient d'une relation très forte alors on garde tous les *synsets* ; quand il provient d'un lien fort, alors on inclut seulement les *synsets* qui lui sont reliés par des liens forts ; et quand le mot provient d'une relation moyennement forte alors on ne considère que les *synsets* avec le meilleur score (avec leur mesure locale).

Tous les *synsets* qui ne participent pas à une relation selon les critères précédents sont supprimés. De plus, au fur et à mesure que des mots sont ajoutés à la chaine, et si elle devient illégale, tous ses synsets sont supprimés.

Par ailleurs quand on insère un mot, on cherche d'abord parmi les relations par ordre décroissant de force et dans un contexte de plus en plus petit (respectivement, toutes les phrases, sept

phrases, trois phrases) si une chaine existe déjà, auquel cas le mot y est ajouté ; dans le cas contraire une nouvelle chaine est créée.

Le problème de cette méthode est qu'elle est peu précise à cause de sa nature gloutonne (Navigli, 2009) et différentes améliorations ont été proposées. On peut citer (Barzilay et Elhadad, 1997) qui gardent toutes les interprétations possibles, ce qui augmente la précision au détriment des performances. (Silber et McCoy, 2000) proposent un algorithme de construction de chaines lexicales linéaire, qui permet de résoudre le problème de performance tout en conservant la qualité accrue.

4.5 Algorithme à base d'exploration pseudo aléatoire de graphes

D'autres algorithmes sont ceux qui se basent sur le principe d'une marche aléatoire dans un graphe, ce qui inclut à la fois des algorithmes de type *PageRank*, mais aussi des méta-heuristiques à colonies de fourmis.

4.5.1 Algorithme à base de PageRank

(Mihalcea et al. 2004) ont appliqué l'algorithme de *PageRank* (Brin et Page, 1998) pour la désambiguïsation lexicale. Le principe est d'assigner des poids aux arcs d'un graphe récursivement en exploitant les informations globalement disponibles.

(Mihalcea et al. 2004) utilisent WordNet et ses relations pour construire un graphe dirigé [3] ou non représentant les différentes combinaisons de sens à partir du texte à désambiguïser. Pour chaque mot, l'ensemble des *synsets* qui lui sont liés constituent les nœuds du graphe, alors que les arcs sont les relations issues de Wordnet (ou d'une combinaison de relations) entre les *synsets* des mots du texte. À noter que les *synsets* du même mot ne sont jamais reliés entre eux.

Une fois le graphe construit, un poids est associé à chaque arc. Le choix des poids initiaux peut se faire de deux manières : initialisés à 1, ou par une mesure de similarité sémantique.

S'en suit la marche aléatoire du PageRank. Le marcheur choisit l'arc à emprunter de manière aléatoire suivant la distribution des valeurs de PageRank des nœuds reliés au nœud courant. À chaque passage sur un nœud, le score est mis à jour avec : $S(N_i) = (1-d)+d*\sum_{j \in In(N_i)} \frac{S(N_i)}{|Out(N_i)|}$, où d est un facteur de lissage, N_i un nœud du graphe, $In(N_i)$ l'ensemble des nœuds prédécesseurs de N_i et $Out(N_i)$ l'ensemble des nœuds successeurs. Lorsque le système a convergé sur la distribution stationnaire, le sens de chaque mot correspondant au nœud avec le meilleur score PageRank est sélectionné le sens.

4.5.2 Algorithme à colonies de fourmis

Les algorithmes à colonies de fourmis sont des algorithmes multi-agents réactifs qui cherchent à imiter le fonctionnement d'une colonie de fourmis. Cette approche fut initialement proposée par (Dorigo, 1992), et part de la constatation faite par (Deneubourg *et al.*, 1983) que les fourmis, lorsque plusieurs chemins sont possibles pour atteindre de la nourriture, convergent systématiquement vers le chemin le plus court. Lors de leur passage, les fourmis déposent une substance chimique (phéromone) pour alerter les autres fourmis de la présence de nourriture ;

3. Sachant que les relations dans Wordnet son symétriques, on retiens arbitrairement soit la relation, soit son inverse

cette substance s'évapore si elle n'est pas renforcée par le passage d'autres fourmis. C'est cette communication indirecte au travers de l'environnement (stigmergie) qu'émerge une convergence optimale du système. L'utilisation d'un tel algorithme pour la désambiguïsation lexicale a été proposée par (Guinand et Lafourcade, 2010) en utilisant un modèle de similarité à base de vecteurs pour régir le déplacement des fourmis. Plus récemment (Schwab *et al.*, 2011) ont proposé de remplacer ce modèle par l'utilisation d'une approche basée sur la mesure de Lesk étendu.

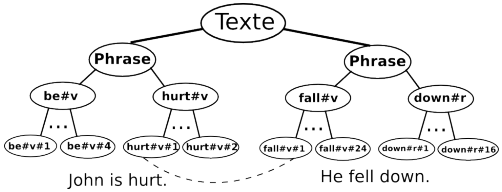


FIGURE 3 – La structure de l'environnement de l'algorithme à colonies de fourmis

Le graphe, contrairement à celui de (Mihalcea *et al.*, 2004), ne lie pas les sens entre eux, mais reprends une structure d'arbre qui suit celle du texte [4].

Nous pouvons voir sur la Figure 3 un exemple de graphe pour une phrase simple. La racine est un nœud correspondant au texte. Les nœuds fils sont des nœuds correspondant aux phrases ; leurs nœuds fils correspondant aux mots les feuilles des mots correspondent aux sens. Ces nœuds produiront les fourmis (fourmilières). Au départ il n'y a aucune connexion entre les nœuds "sens" des différents mots, ce sont les fourmis qui vont créer des "ponts" entre eux. Chaque nœud qui n'est pas une fourmilière possède un vecteur de sens qui lui est attaché (vecteur contenant des identifiants de sens WordNet).

Les nœuds fourmilières possèdent une quantité d'énergie qu'elles peuvent utiliser pour produire des fourmis à chaque itération de l'algorithme.

Les fourmis, partent à l'exploration du graphe de manière pseudo aléatoire. La probabilité de prendre un chemin dépend de la quantité d'énergie sur le nœud, de la concentration de phéromone, et du score entre le nœud (son vecteur de sens) où elle se trouve et sa fourmilière d'origine [5].

Quand une fourmi arrive sur un nœud, elle prélève une quantité d'énergie et a une probabilité dépendant de la quantité d'énergie qu'elle porte de passer en mode retour pour rapporter l'énergie à sa fourmilière.

Lorsqu'une fourmi passe sur un nœud non fourmilière elle va déposer le sens correspondant à sa fourmilière dans le vecteur de sens du nœud ainsi qu'une quantité de phéromone. La phéromone s'évapore en partie à chaque itération.

Si une fourmi arrive sur le nœud d'un autre sens que le sien, il y a une probabilité (dépendant du score avec son sens d'origine) qu'elle construise un "pont" vers sa fourmilière afin d'y revenir directement. Lorsque la fourmi passe par un pont elle dépose également des phéromones, ce qui pourra inciter d'autres fourmis à la suivre.

4. On conserve ainsi la proximité et l'ordre des mots par exemple
5. à l'aide de mesures de similarité locales.

Lorsque de nombreux ponts ont été construits, certains ponts vont se renforcer et d'autres s'évaporer (lorsqu'il n'y aura plus de phéromone) ; cela va mener à une monopolisation des ressources au niveau des fourmilières avec les ponts les plus fréquentés. Les ponts correspondant ainsi à des chemins interprétatifs parmi les combinaisons de sens possibles. À la fin de la simulation les sens qui correspondent aux fourmilières avec le plus d'énergie sont choisis.

5 Critères de choix des algorithmes

Le choix de la mesure de similarité à utiliser dépends d'une part des contraintes sur les ressources lexicales disponibles mais aussi du contexte applicatif : certaines seront plus adaptées car relevant mieux de certains aspects plutôt que d'autres de la similarité sémantique réelle. (Budanitsky et Hirst, 2006) proposent une comparaison empirique de 5 mesures par rapport au jugement humain de manière très détaillée, ce qui peut constituer un élément de choix utile. Plus récemment (Pirró et Euzenat, 2010), entreprennent également de comparer leur mesure à la plupart des mesures classiques.

Quant aux algorithmes globaux, il y a deux aspects à considérer, d'une part l'évaluation de la qualité des solutions, et d'autre part le temps d'exécution de l'algorithme. Les taches de désambiguïsation des campagnes d'évaluation telles que Semeval (anciennement SenseEval), ne sont axées que sur l'évaluation de la qualité par rapport à une désambiguïsation de référence du corpus faite manuellement. Elles fournissent cependant un premier élément de comparaison.

D'une part on peut discuter de la valeur d'une telle évaluation de la qualité dans un système appliqué à un problème réel, et d'autre part de la vitesse d'exécution qui est un facteur très important pour des applications telles que la traduction automatique, surtout si il s'agit de traduction de parole à parole (où il y a un besoin de traitement en temps réel).

(Schwab *et al.*, 2012) ont entrepris de comparer le recuit simulé, l'algorithme génétique ainsi que l'algorithme à colonie de fourmis à la fois en termes de qualité (Semeval 2007 – Tâche 7), mais également en terme de convergence et de vitesse d'exécution en utilisant comme mesure locale Lesk étendu. Ils concluent que les trois algorithmes avec la même mesure de similarité locale offrent des résultats en terme de qualité comparables ; c'est cependant l'algorithme à colonie de fourmis qui s'avère le plus rapide (environ 10 fois plus que le recuit simulé et 100 fois plus que l'algorithme génétique).

6 Conclusions et perspectives de recherche

Nous avons passés en revue les principales méthodes de désambiguïsation lexicale basées sur des connaissances, que ce soit les mesures au niveau local ou les algorithmes au niveau global. D'un point de vue local, les mesures de similarité sémantique sont bien entendu très utiles pour les systèmes de TALN, mais elles jouent également un rôle de plus en plus important pour des applications au web sémantique, et également pour la construction automatisée de ressources lexicales. La recherche se focalise principalement d'une part sur l'hybridation et la combinaison des mesures de similarité, mais également sur la combinaison de sources d'information (multilingues ou non), ou encore des ontologies croisées.

Du point de vue global une possibilité d'amélioration se situe au niveau du temps de convergence et des performances en général. Nous nous intéressons surtout aux combinaisons de mesures de

similarité. Ces combinaisons peuvent se faire en utilisant des mesures différentes pour essayer de capter différents aspects utiles à la désambiguïsation. Par exemple, on peut imaginer utiliser une mesure par catégorie lexicale, utiliser différentes mesures exploitant différentes sources d'information ou adopter une stratégie de vote, ou enfin, au niveau de l'algorithme à colonies de fourmis, utiliser différentes castes de fourmis, chacune utilisant une mesure de similarité différente pour ses déplacements.

Références

BANERJEE, S. et PEDERSEN, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. *In CICLing '02*, pages 136–145, London, UK.

BARZILAY, R. et ELHADAD, M. (1997). Using lexical chains for text summarization. *In In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

BRIN, S. et PAGE, L. (1998). The anatomy of a large-scale hypertextual web search engine. *In WWW7*, pages 107–117, Amsterdam.

BUDANITSKY, A. et HIRST, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

COLLINS (1998). *Cobuild English Dictionary*. Harper Collins Publishers.

COWIE, J., GUTHRIE, J. et GUTHRIE, L. (1992). Lexical disambiguation using simulated annealing. *In COLING '92*, pages 359–365, Stroudsburg, PA, USA. ACL.

DENEUBOURG, J. L., PASTEELS, J. M. et VERHAEGE, J. C. (1983). Probabilistic behaviour in ants : a strategy of errors ? *Journal of Theoretical Biology*, 105:259–271.

DICE, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

DORIGO, M. (1992). *Optimization, Learning and Natural Algorithms*. Thèse de doctorat, Politecnico di Milano, Italie.

GELBUKH, A., SIDOROV, G. et HAN, S.-Y. (2003). Evolutionary approach to natural language word sense disambiguation through global coherence optimization. *WSEAS Transactions on Communications*, 1(2):11–19.

GELBUKH, A., SIDOROV, G. et HAN, S.-Y. (2005). On some optimization heuristics for lesk-like wsd algorithms. *In NLDB'05*, pages 402–405, Berlin, Heidelberg.

GUINAND, F. et LAFOURCADE, M. (2010). *Artificial ants for Natural Language Processing*, chapitre 20, pages 455–492. Artificial Ants. From Collective Intelligence to Real-life Optimization and Beyond. Monmarché, N. and Guinand, F. and P. Siarry.

HALLIDAY, M. A. et HASAN, R. (1976). *Cohesion in English*. Longman Group Ltd, London, U.K.

HIRST, G. et ST-ONGE, D. D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet : An electronic Lexical Database. C. Fellbaum.*, pages 305–332. Ed. MIT Press.

IDE, N. et VERONIS, J. (1998). Word sense disambiguation : The state of the art. *Computational Linguistics*, 24:1–40.

JIANG, J. J. et CONRATH, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In ROCLING X*.

LEACOCK, C. et CHODOROW, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An Electronic Lexical Database. C. Fellbaum. Ed. MIT Press. Cambridge. MA*.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In SIGDOC '86*, pages 24–26, New York, NY, USA. ACM.

MIHALCEA, R., TARAU, P. et FIGA, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. *In COLING '04*, Stroudsburg, PA, USA. ACL.

MILLER, G. A. (1995). Wordnet : a lexical database for english. *Commun. ACM*, 38(11):39–41.

MORRIS, J. et HIRST, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48.

NAVIGLI, R. (2009). Word sense disambiguation : A survey. *ACM Comput. Surv.*, 41(2):10 :1–10 :69.

PIRRÓ, G. et EUZENAT, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. *In ISWC 2010*, volume 6496 de *Lecture Notes in Computer Science*, pages 615–630.

RADA, R., MILI, H., BICKNELL, E. et BLETTNER, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

RESNIK, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In IJCAI'95*, pages 448–453, San Francisco, CA, USA.

ROGERS, D. et TANIMOTO, T. (1960). A computer program for classifying plants. *Science*, 132(3434):1115–1118.

ROGET (1989). *New Roget's Thesaurus*. P.S.I.

SCHWAB, D., GOULIAN, J. et GUILLAUME, N. (2011). Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. *In TALN*, Montpellier (France).

SCHWAB, D., GOULIAN, J. et TCHECHMEDJIEV, A. (2012). Comparaison théorique et pratique d'algorithmes d'optimisation globaux pour la désambiguïsation lexicale non supervisée. *Traitement Automatique des Langues*, 1(53):37 pages. Soumis à la revue Traitement Automatique des Langues.

SECO, N., VEALE, T. et HAYES, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. *In Proceedings of ECAI'2004*, pages 1089–1090, Valencia, Spain.

SILBER, H. G. et MCCOY, K. F. (2000). Efficient text summarization using lexical chains. *In IUI '00*, pages 252–255, New York, NY, USA. ACM.

TVERSKY, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

WAGNER, C. (2008). Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journal*, 19(1):70–83.

WILKS, Y. et STEVENSON, M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. *In COLING '98*, pages 1398–1402, Stroudsburg, PA, USA. ACL.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on ACL*, volume 2 de *ACL '94*, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA).

Compression textuelle sur la base de règles issues d'un corpus de sms

Arnaud Kirsch
UCL - CenTAL - Place Blaise Pascal 1, Louvain-la-Neuve, 1348
arnaud.kirsch@student.uclouvain.be

RÉSUMÉ

La présente recherche cherche à réduire la taille de messages textuels sur la base de techniques de compression observées, pour la plupart, dans un corpus de sms. Ce papier explique la méthodologie suivie pour établir des règles de contraction. Il présente ensuite les 33 règles retenues, et illustre les quatre niveaux de compression proposés par deux exemples concrets, produits automatiquement par un premier prototype. Le but de cette recherche n'est donc pas de produire de "l'écrit-sms", mais d'élaborer un procédé de compression capable de produire des textes courts et compréhensibles à partir de n'importe quelle source textuelle en français. Le terme "d'essentialisation" est proposé pour désigner cette approche de réduction textuelle.

ABSTRACT

Textual Compression Based on Rules Arising from a Corpus of Text Messages

The present research seeks to reduce the size of text messages on the basis of compression techniques observed mostly in a corpus of sms. This paper explains the methodology followed to establish compression rules. It then presents the 33 considered rules, and illustrates the four suggested levels of compression with two practical examples, automatically generated by a first prototype. This research's main purpose is not to produce "sms-language", but consists in designing a textual compression process able to generate short and understandable texts from any textual source in French. The term of "essentialization" is proposed to describe this approach of textual reduction.

MOTS-CLEFS : résumé automatique, compression de texte, sms, lisibilité, essentialisation.

KEYWORDS : summarization, text compression, text messaging, readability, essentialization.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 309–322,
Grenoble, 4 au 8 juin 2012.

1. Introduction

Les locuteurs, en pleine "convergence numérique", sont amenés simultanément à échanger divers types de messages écrits dans de mêmes environnements (c'est le cas de Facebook où la même interface permet l'échange de mails, de messages instantanés et de sms) et de mêmes types de messages écrits dans des environnements divers (des emails peuvent être rédigés depuis un ordinateur, une tablette ou un GSM). Un mode de communication bref et informel traverse ainsi les genres et les *media*, mode mis en évidence par la présente recherche, qui veut proposer une aide à la rédaction de messages courts.

En cela, nous nous sommes naturellement intéressé au résumé automatique et à la compression de données, mais ces deux approches ne rencontrent pas totalement notre objectif : nous avons donc choisi de dériver de nouvelles règles de contraction émanant des pratiques des locuteurs. En effet, le résumé automatique, apparu en linguistique informatique avec Luhn à la fin des années 1950 (Luhn, 1958), à pour but la simplification d'un texte en ne retenant que les idées principales de celui-ci et en les rassemblant dans un nouveau texte grammaticalement acceptable ((Minel, 2004), (Yousfi-Monod, 2007)). La compression de données vise elle à modifier la représentation binaire des données pour en réduire le poids numérique. Elle nécessite décompression du côté du récepteur.

À l'inverse, la proposition qui fait l'objet de cet article possède trois caractéristiques essentielles. (1) Le processus de simplification se situe au **caractère** près : nous cherchons à isoler les caractères les plus essentiels à la transmission de la totalité de l'information contenue dans le message initial, ce qui nous éloigne du résumé automatique *stricto sensu* (Minel, 2004). Nous nous distinguons également de la compression textuelle (Yousfi-Monod, 2007) qui cherche à supprimer les segments non porteurs de sens dans un texte. (2) Le processus de simplification ne doit pas entraver la **compréhension** : quel que soit le taux de compression recherché, nous devons produire le texte le plus décodable possible, soit celui à partir duquel, malgré la disparition d'un certain nombre de caractères, un lecteur pourra le plus aisément retrouver le texte initial. Cela signifie qu'il y aura une borne minimale en deçà de laquelle il ne sera pas possible de descendre, au risque de rendre le texte de base totalement inintelligible (puisque nous entendons transmettre le texte tel que comprimé, sans intention de décompression à l'arrivée). (3) Enfin, il s'agit d'un processus de **compression** : seuls les mécanismes permettant de réduire la taille du message doivent être pris en compte. Pour respecter ces caractéristiques, nous nous sommes basé sur le corpus de sms de (Fairon *et al.*, 2006a), et avons extrait une série de règles de compression. Malgré les caractéristiques du corpus d'entraînement, nous ne projetons pas de reproduire de l'*écrit-sms* ((Cougnon et François, 2010) ; nous le noterons e-sms). Notre intention n'est donc pas d'inverser les systèmes de normalisation tels que ceux développés par (Yvon, 2008) et (Beaufort et al., 2010).

Notre recherche, même si elle participe du résumé automatique, s'en distancie donc. En effet, selon la première caractéristique énoncée ci-dessus, nous ne cherchons pas à extraire les idées principales d'un texte, ni à produire un message grammaticalement acceptable : nous travaillons presque exclusivement au niveau des caractères et tentons de conserver toutes les informations du texte initial. Aussi nous avons préféré parler "d'essentialisation" de texte, dont voici la définition : "sous-procédé du résumé automatique focalisé sur la réduction du nombre de caractères". Le prototype de programme d'essentialisation automatique, ou essentialiseur, a notamment pour objectif

la création d'une application Twitter®[1]. Nous envisageons, à l'instar de ce qui se fait déjà en anglais (140it, 2012), qu'un utilisateur encode un texte plus grand et le réduise automatiquement à l'aide dudit programme.

Dans cet article nous aborderons tout d'abord la méthodologie que nous avons suivie pour établir nos règles. Nous expliquerons ensuite le fonctionnement global d'un premier prototype d'essentialiseur, en détaillant les règles retenues et les classements de celles-ci. Nous terminerons par l'évocation des difficultés et résistances rencontrées et des perspectives de recherche que nous envisageons.

2. Méthodologie

Notre proposition d'essentialisation se base sur le corpus de (Fairon *et al.*, 2006a). Ce corpus a été constitué dans le cadre du projet *sms4science* : 30.000 messages ont été transcrits et annotés en 2004. Ces messages sont des sms véritablement échangés entre usagers et permettent donc d'observer les pratiques réelles de simplification de la population. Nous avons travaillé sur trois sous-corpus : l'un contenant les messages d'exactement 160 caractères [2], un autre ceux de plus de 160 caractères, et le dernier les messages transcrits de plus de 160 caractères[3]. Les messages de 160 caractères ont constitué notre corpus d'entraînement, et les deux autres ont servi à valider nos observations.

Sur la base de ces sous-corpus, nous avons établi une liste de règles de contraction. Celles-ci consistent en la formalisation de certains phénomènes en vue de respecter deux enjeux majeurs : l'**intercompréhension** et la **compression**. Les phénomènes les plus récurrents sont retenus parce qu'ils sont supposés être les plus intelligibles. Parmi ceux-ci, nous ne nous intéressons qu'à ceux permettant de diminuer la taille du texte. À ces observations nous avons ajouté certaines règles qui ne sont pas issues de notre observation du corpus, par exemple la formalisation des dates et des heures (norme ISO 8601, *cf.* (ISO, 2012)). Enfin, nous avons ordonné ses règles et avons établi différents niveaux de contraction.

3. Le système

La deuxième étape de ce travail a consisté à implémenter un premier prototype d'essentialiseur. À ce stade de nos travaux, certains tests statistiques et une évaluation préliminaire sont rendus possibles, comme nous le montrons ensuite.

---

[1] Site de *microblogging*, soit d'échange de courts messages ne pouvant dépasser 140 caractères.

[2] Ancienne limite technique, la limite de 160 caractères est aujourd'hui une limite de facturation : un sms de 161 caractères sera facturé comme deux messages. Des études statistiques menées, notamment, sur notre corpus, ont démontré qu'il y avait un pic de sms de 160 caractères (Cougnon *et* François, 2010).

[3] Le premier corpus contenait 2223 sms, le deuxième 2298 et le troisième 8310.

3.1. Les règles

Nous avons établi 33 règles. Le TABLEAU 1 reprend chacune de ces règles, en donne une explication succincte et l'illustre par un exemple concret. Les codes utilisés renvoient aux classements présentés par la suite (*cf.* 3.3.).

Règle	Description	Exemple
[0101]	Réduction des smileys	• :-) ➔ :) • -___-" ➔ -_-"
[0302]	Remplacement des URL par une forme plus courte	• http://www.uclouvain.be/cental-cahiers.html#langsms ➔ http://tinyurl.com/7lumcf7
[0203]	Réduction des répétitions inutiles de caractères	• Mdrrrrrrr ➔ mdr • Looooool ➔ lol
[0504]	Réalisation des unités lexicales en logogrammes	• Vingt-quatre ➔ 24 • Et ➔ &
[0705]	Remplacement par un synonyme	• Travailler ➔ bosser
[0406]	Normalisation des dates et des heures au format ISO 8601	• Le 14 juillet 1989 vers 12h30 ➔ 1989-07-14 vers 12:30
[1107]	Suppression du pronom sujet quand les formes verbales sont non ambiguës (notamment les il impersonnels)	• Nous allons arriver ➔ allons arriver • Il faut partir ➔ Faut partir
[1208]	Si deux verbes consécutifs utilisent le même pronom sujet, et que la première occurrence du pronom sujet est maintenue, la seconde peut-être supprimée	• Je pense à toi. Je veux te revoir. ➔ Je pense à toi. Veux te revoir.
[0809]	Ellipse des "et" qui retournent une interrogation et des "ne" de la négation	• Je vais bien, et toi ? ➔ Je vais bien, toi ? • Je ne pense pas ➔ Je pense pas

Règle	Description	Exemple
[1010]	Suppression des mots répétés (ne concerne pas les pronoms personnels)	• Hello hello ➔ Hello
[2411]	Réduction aux initiales des noms propres composés et des noms communs composés les plus courants	• Pierre-Yves ➔ P.Y. • Week-end ➔ w.e.
[3012]	Simplification des répétitions de pronoms personnels	• Nous nous reverrons ➔ nous reverrons
[1513]	"tu" et "vous", suivis de voyelles, deviennent, respectivement, "t'" et "z'"	• Tu arrives ➔ t'arrives • Vous arrivez ➔ z'arrivez
[1614]	Réduction des "tu" et "je" à l'initiale	• Tu vas ➔ tvas • Je pense ➔ jpense
[2115]	Fusion des mots composés, locutions, etc. Reconnus par ailleurs	• Porte-monnaie ➔ portemonnaie
[2016]	"bisou(s)" en fin de texte > "x"	• À plus, bisous. ➔ à plus, x
[2317]	Les monosyllabiques courants sont réduits à leur squelette	• Temps ➔ tps
[2218]	Certaines fins de mots courantes sont réduites à leur squelette	• Internement ➔ internemt
[1319]	Certains mots très courants du texte sont réduits à leur squelette	• Beaucoup ➔ bcp
[1420]	Apocopes des mots les plus fréquents	• Anniversaire ➔ anniv
[2721]	Suppression des consonnes finales muettes, sauf les marques du pluriel	• À travers ➔ à traver
[2522]	Si pas d'ambiguïté, suppression des marques muettes du pluriel	• Journaux ➔ journau
[1723]	Suppression des schwas peu prononcés et fusion avec le mot suivant	• Dernièrement ➔ dernièrment • Je ne sais pas -> je nsais pas

Règle	Description	Exemple
[3124]	Suppression de tous les schwas ; les monosyllabiques fusionnent avec les mots qu'ils précèdent	• Je me grouillerai ➔ jmgrouillerai
[1825]	Suppression des "h" muets	• Hérisson ➔ érisson
[2626]	Simplification des doubles consonnes	• Notamment ➔ notament
[2827]	Les phonèmes transcrits par plusieurs caractères sont remplacés par des caractères uniques	• Je voulais que tu viennes ➔ je voulè ke tu vienes
[2928]	Phonétisation des syllabes par la lettre, le signe ou le chiffre	• J'ai envie de toi ➔ G envie 2 toi
[0629]	Simplification des abréviations	• P.-S. ➔ PS
[1930]	Suppression des points finaux, des apostrophes, des traits d'union et simplification des points de suspension (réduits à deux points successifs)	• J'imagine ➔ jimagine • Penses-tu ➔ pensestu
[3231]	Suppression des doubles points et points-virgules	• Je disais : comment vas-tu ? ➔ je disais comment vas-tu ?
[0932]	Les espaces séparant deux systèmes graphiques différents sont supprimés	• J'ai passé 1 bonne journée ➔ j'ai passé1bonne journée • Je suis dégouté ! Et il est là ! ➔ je suis dégouté!Et il est là!
[3333]	Suppression de tous les espaces : *scriptura continua* (sauf entre deux caractères numériques)	• Je ne crois pas ➔ jenecroispas

TABLEAU 1 - *Présentation des règles*

3.2. Remarques sur le tableau

Nous regroupons ici une série de remarques concernant le Tableau 1 :

- L'ordonnancement des règles n'est pas figé : en fonction des résultats d'une enquête qualitative, ou en cas d'ajout ou de suppression d'une règle, l'ordre global peut être repensé : il y a une grande variété de types de compression d'un même texte, nous en

choisissons une, sans prétendre qu'elle soit la meilleure, afin qu'elle serve de ligne de conduite ;

- Certaines règles, comme [0101] et [0203] proviennent typiquement de caractéristiques propres au corpus sur lequel nous nous sommes basé : le but de notre système étant de pouvoir travailler sur n'importe quel type de texte, nous devons donc y intégrer des règles plus spécifiques à certains textes qu'à d'autres ;
- [0203] seule : il est clair que ces répétitions sont porteuses de sens également, et que les supprimer revient à enlever une partie du sens encodé initialement. C'est néanmoins un choix que nous posons de les réduire, afin de gagner quelques caractères ;
- Un logogramme (mentionné en [0504]) est un "[d]essin représentatif d'une notion (logogramme sémantique ou idéogramme) ou d'une suite phonique constituée par un mot (logogramme phonétique ou phonogramme)." (TLFi, 2012) ;
- Les synonymes, tels qu'évoqués par la règle [0705], soulèvent un problème évident : il n'existe que très peu de synonymes parfaits, et une substitution peut donc fréquemment modifier une partie du sens apporté. Pour cette approche préliminaire, nous avons établi manuellement une courte liste de synonymes, dont les variations sémantiques portent plutôt sur le registre de langue ("bosser" pour "travailler", par exemple). Ce choix est certes discutable, mais est en lien avec notre remarque préliminaire : nous cherchons à cadrer avec un mode de communication bref et informel ;
- De nombreuses règles, comme [1107], [1208], [0809], etc. se basent sur le constat que le français est parfois redondant (répétitions, reformulations, etc.). Nous cherchons à définir ces redondances (et d'autres) pour les réduire au maximum et donc gagner de l'espace ;
- La règle [2115] prévoit la fusion des mots composés et autres locutions (puisque plus facilement identifiables/compréhensibles par le lecteur). Si nous avions simplement décidé de fusionner les collocations[4] se serait posée la question de l'évaluation du degré de figement des syntagmes. Nous avons donc tout d'abord repris une liste finie de mots composés que nous avons augmentée de locutions relevées automatiquement dans notre corpus[5] ;
- Les règles relatives aux squelettes consonantiques ([2317], [2218] et [1319]) sont établies, elles aussi, sur la base de notre corpus : nous avons relevé de très nombreux squelettes et en avons dégagé des constantes. La difficulté de ces trois règles est de déterminer quelles unités sont plus facilement compréhensibles lorsqu'elles sont autant réduites. En ce qui concerne leurs fréquences d'apparition, nous nous sommes basé sur une évaluation statistique de notre corpus ;
- Lors de la phonétisation des syllabes [2928], les lettres à lire pour leur valeur phonétique seront encodées en majuscule. Aussi, avant l'application de cette règle, toutes les majuscules orthographiques sont réduites, afin d'éviter tout risque de

---

[4] Au sens de « cooccurrences statistiquement privilégiées ».

[5]Pour ce faire, nous avons mesuré la fréquence d'apparition de chaque mot dans les transcriptions des sms du corpus, ainsi que celle de leurs contextes gauche et droit. Nous avons donc pu établir les fréquences d'apparition de tous les syntagmes du corpus et avons retenu les plus présents. Nous avons utilisé les transcriptions afin d'éviter les problèmes d'instabilité orthographique.

confusion. La question se pose cependant de savoir quel système suivre lors de la fusion des caractères [3333] : maintenir la phonétisation ou concaténer les mots non phonétisés en utilisant des majuscules pour marquer l'emplacement des espaces supprimés ? Cette question devra être tranchée après enquête qualitative ;

- Nous distinguons quatre systèmes graphiques pour établir la règle [0932] : les lettres, les chiffres, les signes de ponctuation et les symboles. Cette distinction est établie par nous-même. La *scriptura continua*[6] peut sembler excessive. Elle est cependant présente dans quelques sms, et apparaît à quelques reprises dans l'histoire de l'écriture (par exemple durant l'Antiquité). Ainsi nous décidons de la maintenir.

3.3. Leur classement

Nous avons ordonné ces règles selon deux classements distincts. L'un dépendant de l'ordonnancement informatique, l'autre prenant en compte l'application du système.

Le premier classement répond à deux exigences : d'une part classer les règles selon leur influence sur la lisibilité (ce classement est représenté par les deux premiers chiffres du code d'une règle) ; d'autre part selon l'ordre dans lequel elles doivent être appliquées pour ne pas se gêner mutuellement : par exemple, il faut formaliser les dates avant de phonétiser les noms des mois (les deux derniers chiffres du code illustrent ce second ordre).

Le second classement, orienté vers l'application finale, envisage deux types d'utilisation : le premier, qualitatif, où l'utilisateur doit choisir entre quatre niveaux d'essentialisation prédéfinis, et le second, quantitatif, où le programme connaît le seuil de caractères sous lequel ramener le texte initial et applique les règles jusqu'à y parvenir (ou non).

3.4. Les quatre niveaux

Revenons plus en détail sur les quatre niveaux d'essentialisation que nous proposons. Chaque niveau permet de définir une séquence de règles à appliquer pour atteindre un certain degré d'essentialisation. Nous envisageons d'abord deux exemples produits par notre prototype, puis nous les commentons et détaillons les choix théoriques sous-jacents.

3.4.1.Définitions

Superficiel : ce niveau touche uniquement à ce que nous jugeons accessoire, aux caractères qui ne sont là que pour fluidifier la lecture, comme les répétitions émotives de caractères ("looooool"), les espaces précédant certains signes de ponctuation ("Non !"), l'utilisation de logogrammes ("vingt-quatre" > "24" ; "et" "&"), la suppression de certains mots redondants, par ailleurs souvent absents du langage oral ("ne" de la négation), etc. Ce niveau ne supprime donc que des caractères que nous pourrions qualifier de sémantiquement moins pertinents.

Conventionnel : Il s'agit d'appliquer une série de règles qui sont des pratiques fréquentes de l'écrit (bien au-delà des seuls sms), ou des transcriptions de l'oral. Nous pensons par exemple aux apocopes de mots fréquents ("anniv"), à l'effacement des schwas toujours silencieux ("effacement"), aux abréviations courantes ("tps", "ds",

[6] Le terme *scriptio continua* est également employé.

"qq"...), à l'élision de certains pronoms sujets, ou à leurs simplifications ("zavez vu", "faut y aller"...). Nous commençons ici à atteindre plus conséquemment l'orthographe et la syntaxe, mais selon des pratiques qui, par leur fréquence dans notre corpus et notre propre perception, semblent rapidement déchiffrables.

Morpho-syntaxe du sms : Ce niveau se concentre sur des phénomènes couramment trouvés dans des sms, notamment la phonétisation de certains phonèmes transcrits par plusieurs caractères, la contraction en squelettes consonantiques de certaines fins de mots, ou de certaines unités lexicales fréquentes. Nous ne recensons ici que les phénomènes les plus fréquents afin que nos messages restent les plus compréhensibles possible : en effet nous partons du principe que les phénomènes les plus courants seront les plus intelligibles.

Cryptage : On parle ici de l'application de toutes les règles observées dans notre corpus, afin de gagner un maximum de caractères. Il ne s'agit plus de tenter de conserver un minimum de décodabilité, le but est la compression : phonétisation de toutes les syllabes, suppression d'un partie de la ponctuation, suppression des espaces, etc.

3.4.2.Application

Les quatre niveaux d'essentialisation que nous avons définis servent à dégager un formalisme. Ils ont été posés arbitrairement, puisque seule la définition des niveaux est importante. Le but est d'obtenir une description stricte des attentes et conditions de ces niveaux, afin de choisir les règles qui y correspondront le mieux. Quel que soit le nombre

Niveau	Forme	Taille
Corpus	Hi tite puce,g pensé a tfèr 1pti sign 2vi,tu m'mank grav.pq pa svoir 2m1?alé pass 1bonnuit	*90*
Forme standard	Hi petite puce, j'ai pensé à te faire un petit signe de vie, tu me manques grave. Pourquoi pas se voir demain ? Allez passe une bonne nuit	*138*
Superficiel	Hi petite puce,j'ai pensé à te faire1petit signe de vie,tu me manquegrave.Pourquoi pas se voir demain?Allez passe1bonne nuit	*124*
Conventionnel	Hi ptit puc,jai pensé à tfair1ptit signde vie,tu mmanqugrav.Pourquoi pas svoir demain?Allez pass1bonnnuit	*105*
Morpho-syntaxe du sms	hi ptit puc G penC à t9èr1pttsign2vi tmmankgrav pourkoi pasvoir2m1alé pas1bOn n8	*80*
Cryptage	hiptitpucGpenCàt9èr1pttsign2vitmemankgravpourkoip asevoir2m1alépas1bOnn8	*71*

TABLEAU 2 - *Essentialisation d'un message illustrant la proximité*

de niveaux définis, l'important réside dans la gradation continue entre ceux-ci, tant du point de vue de l'intercompréhension que de la compression.

Deux exemples permettent d'illustrer les résultats de chaque niveau. Le premier exemple[7] est un texte illustrant l'immédiateté (*cf.* TABLEAU 2), le second[8] exposant la distance[9] (*cf.* TABLEAU 3).

Niveau	Forme	Taille
Forme standard	Je rappelle que les banques ont payé plus à l'État belge que ce qu'on leur a donné. Les garanties sur Dexia ont rapporté bien plus que le milliard qui a été mis dans Dexia. Tout a été payant.	*191*
Superficiel	Je rapelle que lebanques ont payé plus à l'État belge que ce qu'on leur a doné.Les garanties sur Dexia ont rapporté bien plus que le milliard qui a été mis dans Dexia.Tout a été payant.	*185*
Conventionnel	Jrapell qulebanqus ont payé plus à lÉtat belgque cquon leur a doné Les garanties sur Dexia ont rapporté bien plus que lmilliard qui a été mis ds Dexia.Tout a été payant.	*169*
Morpho-syntaxe du sms	jrapèl klébanks on pèyé + à léta bèlj ke ce kon leur a dOné lé garantisur dèxia on rapOrT bi1 + ke lmiliar ki a éT mi ds dèxia tout a éT pèyan	*144*
Cryptage	j r a p è l k e l é b a n k o n p è y é +àlétabèljkecekoleuradOné.légarantisurdèxiaonrapOrT bi1+klmiliarkiaéTmidandèxiatoutaéTpèyan	*110*

TABLEAU 3 - *Essentialisation d'un message illustrant la distance*

3.4.3.Analyse des résultats

Le niveau **superficiel** enregistre un taux de réduction de 14% pour le premier exemple, et de seulement 3% pour le second. Le taux moyen de compression de ce niveau se situe à hauteur de 9%. Les textes produits ne sont pas très réduits, mais restent compréhensibles : la forme des mots n'est pas atteinte. Il reste quelques erreurs produites par notre prototype, notamment "lebanques", qui aurait dû rester "les banques", et la même chose pour "maquegrave". Une assez grande différence que l'on

[7] Il s'agit d'un sms issus du corpus.

[8] (Libre Belgique, 2011 : 5)

[9] L'opposition immédiat *versus* distance est proposée par (Koch et Österreicher, 2001). La distance communicative dénote le degré d'implication (immédiat) ou de détachement (distance) des locuteurs dans le discours.

trouve entre les taux de compression des deux messages est probablement due au fait que le premier message est du même type que notre corpus d'entraînement.

Au niveau **conventionnel**, le sms obtient un taux de réduction de 24% et l'extrait de journal 10,5%. À nouveau, l'écart est assez large entre les deux. Le taux moyen se situe à 18%. Par rapport au premier niveau, la progression est assez marquée du point de vue de la compression. En ce qui concerne la compréhension, le texte devient déjà plus hermétique, principalement lorsque des schwas ont été supprimés, avec l'espace qui les suivait. Cette agglutination de mots, si elle permet de gagner des caractères, gêne le décodage du texte. Cependant, si nous regardons le troisième niveau, l'écart perceptif entre celui-ci et le deuxième qui nous occupe semble plus important qu'entre les deux premiers niveaux. Notre impression est donc que la gradation n'est pas continue entre les trois premiers niveaux.

Les taux de réduction des deux exemples du niveau **morpho-syntaxe du sms** sont respectivement de 42% et 25%. Le taux moyen de ce troisième niveau est de 34%. Les deux textes reprennent certaines caractéristiques des sms et leur aspect est similaire à l'écrit sms. Nous pouvons d'ailleurs apprécier cette ressemblance en comparant le niveau 3 du premier exemple à la version originale de ce sms dans notre corpus. Il y a des différences, (rappelons que nous ne cherchons pas à produire de l'e-sms) mais nous devons aussi garder à l'esprit que les possibilités de combinaison des règles sont très nombreuses, et que notre système est déterministe : il produira toujours la même sortie pour un même texte. À l'inverse, un locuteur n'emploiera pas forcément les mêmes techniques d'un message à l'autre.

Les deux taux de réduction du niveau **cryptage** convergent assez fort, puisque le premier texte atteint un taux de réduction de 49% et le second de 43%. Le taux moyen est de 44%. La compression est excellente, puisque près de la moitié du texte a été supprimée. La compréhension est par contre bien plus délicate. Il semblerait qu'il manque trop de caractères pour que la lecture reste fluide. En effet, l'absence d'espace empêche le lecteur de délimiter les unités lexicales. Les deux textes produits ressemblent à une suite continue de phonèmes, à l'instar de ce qu'est le signal sonore.

3.5. Évaluation

Étant donné que nous sommes à une étape préliminaire de notre recherche, nous n'avons pas encore pu réaliser d'évaluation qualitative[10] de ce travail. Nous proposons donc une évaluation en deux parties : une évaluation quantitative d'une part, et le relevé manuel des limites de notre système, d'autre part.

3.5.1.Évaluation quantitative

Pour établir les taux moyens de compression, nous avons utilisé un corpus de test composé de cent textes courts ou extraits répartis comme suit : cinquante transcriptions de sms tirées de notre corpus, quarante extraits de journaux et dix extraits littéraires[11]. Nous avons mesuré le taux moyen de compression produit par les quatre niveaux mentionnés ci-dessus sur l'ensemble de notre corpus de test. Il nous est impossible de mesurer la divergence de nos résultats par rapport à une référence, celle-ci n'existant

[10] Validation des choix théoriques par un échantillon de testeurs humains.

[11] De dix auteurs différents.

pas. À l'inverse du résumé automatique, nous ne pouvons essentialiser manuellement, ni comparer les résultats d'un système équivalent.

Au mieux, dans le cas des 50 sms, aurions-nous pu comparer nos taux de compression à ceux des locuteurs. Cependant, la variabilité de l'écrit sms nous empêcherait de savoir à quel niveau d'essentialisation comparer le sms réel. Et un taux moyen ne serait pas plus éclairant. Nous devons donc nous limiter à une observation des taux moyens de chaque niveau. Le but est d'évaluer la gradation de la compression entre les quatre niveaux proposés.

Les chiffres ainsi obtenus viennent, d'une certaine façon appuyer notre première impression : la progression n'est pas continue entre les quatre niveaux d'essentialisation tels qu'ils sont définis actuellement. Nous passons en effet de 9% à 14% puis à 34% pour atteindre 44% au dernier niveau. Le deuxième niveau semble donc ne pas être un bon intermédiaire entre le premier et le troisième ; ou alors ceux-ci sont-ils respectivement trop conservateur et trop destructeur. Nous devrons attendre d'avoir obtenu les résultats d'une évaluation humaine pour le déterminer et corriger ces premières propositions.

Cependant, malgré certains soucis de gradation et quelques problèmes d'implémentation, nous obtenons assez rapidement des taux de réduction intéressant. Le dernier niveau peut aller jusqu'à réduire de moitié la taille du texte initial, et le deuxième s'approche de 15% de réduction. Considérant que notre but est de transmettre un texte maintenant toutes les nuances du texte d'origine et générant le moins d'ambiguïté possible, ces premiers résultats sont assez encourageants, même s'ils ne nous permettent pas encore d'évaluer objectivement l'évolution de la lisibilité des textes produits.

3.5.2.Limites du système

Il convient d'être assez critique au regard de nos premiers résultats. Nous rencontrons en effet deux types de problèmes avec notre premier prototype d'essentialiseur : le premier se situe au niveau de l'algorithme général d'application des règles, l'autre au niveau de certaines règles elles-mêmes.

Tout d'abord, bien que nous en ayons conscience et ayons tenté d'éviter ce type d'inconvénients, certaines règles gênent l'application des suivantes, voire annulent leurs résultats. Par exemple, la règle [2031] supprime une série de signes de ponctuation qui pourraient se retrouver dans des adresses web préalablement réduites par la règle [0302]. Ou encore, si la règle [0504], qui remplace, notamment, les nombres écrits en toutes lettres par leur équivalent en chiffres arabes ne s'applique pas correctement, toutes les autres règles qui s'appuient sur des chiffres seront induites en erreur. Ensuite, certaines règles qui semblent évidentes pour l'esprit humain ne sont pas toujours les plus simples à formaliser pour la machine. Il en va ainsi de la règle [0504], que nous avons déjà citée : des trente-quatre, c'est elle qui fut la plus difficile à implémenter, et à optimiser. Il existe en effet une multitude de configurations possibles pour l'énonciation de nombres, et nous n'avons considéré que les plus fréquentes, partant du double constat que nous pourrons améliorer cette règle lors des prochaines étapes de notre réflexion et qu'il y aura probablement fort peu de cas où les utilisateurs entreront des nombres en toutes lettres.

Mais le principal problème que nous rencontrons reste la variabilité de la langue : comme nous l'avons déjà précisé ci-dessus, le français, comme d'autres langues naturelles, offre d'innombrables possibilités de variations que nous ne pourrons pas

toutes envisager, or certaines d'entre elles amèneront notre système à commettre des erreurs et à manquer son objectif (par exemple les dates : nous tentons de les formaliser, mais un utilisateur pour aussi bien entrer "le 2 juillet 1989" que "2 du 7 89" ou encore "7-2-89"). Ce premier prototype doit donc encore être amélioré.

4. Conclusions

La première étape de cette recherche ouvre de très nombreuses perspectives d'optimisation. D'un point de vue technique, pour corriger les problèmes mentionnés ci-dessus, nous envisageons notamment de marquer certains caractères, de sorte qu'ils ne puissent plus être modifiés. Nous continuons par ailleurs à réfléchir à un meilleur algorithme de conversion de nombres. Au niveau du système lui-même, d'autres règles pourraient être ajoutées, s'appuyant éventuellement sur d'autres types de corpus qu'un corpus de sms, pour y observer d'autres mécanismes. Enfin les quatre niveaux que nous avons présentés devront également être affinés, et de nouveaux pourraient être ajoutés au système.

Une évaluation qualitative effectuée par des locuteurs nous permettra d'améliorer notre estimation de la lisibilité des textes essentialisés, éventuellement de repenser l'organisation des règles ou des niveaux du système actuel, mais également de valider nos propositions quant à l'essentialisation en général. Il reste donc de nombreuses pistes de réflexion, et nous n'apportons ici que nos premières propositions.

Si le champ de recherche ici présenté peut sembler restreint ou limité à quelques applications ludiques, nous affirmons qu'il n'en est rien. Il ouvre d'une part la voie à une approche originale de l'étude des phénomènes de contraction présents dans les sms en ce qu'il postule leur application formelle. D'autre part, il propose une nouvelle méthode de compression textuelle, pertinente notamment dans le cadre de sites de microblogging comme twitter® ou dans d'autres situations de communication textuelle soumises à une forte contrainte d'espace (affichage de messages d'alerte, de notifications). Mais il reconnaît surtout l'existence d'un nouveau mode de communication bref et informel, et propose une aide à son utilisation.

Remerciements

Nous tenons à remercier Cédrick Fairon qui nous a soufflé l'idée de cette recherche. Nous souhaitons que Louise-Amélie Cougnon soit également reconnue pour son indéfectible soutien, sa disponibilité et son aide précieuse. Enfin merci à MM Beaufort, Bouraoui et Watrin pour leurs conseils avisés.

Références

140it, 2012-03-20 http://140it.com.

2011-11-05, La Libre Belgique, Bruxelles.

BEAUFORT Richard, COUGNON Louise-Amélie, FAIRON Cédrick *et* ROEKHAUT Sophie, 2010, "Une approche hybride traduction/correction pour la normalisation des sms", *in TALN*, Montréal.

COTTIN Florent, 2011, *Le "pourrisseur de texte" du RALI*, Université de Montréal.

COUGNON Louise-Amélie *et* FRANÇOIS Thomas, 2010, *Étudier l'écrit sms. Un objectif du projet sms4science*, Linguistik Online.

COUGNON Louise-Amélie *et* LEDEGEN Gudrun, 2008, "c'est écrire comme je parle. Une étude comparatiste de variétés de français dans l'écrit sms", *Actes du Congrès annuel de l'AFLS*, Oxford.

FAIRON C., KLEIN J. *et* PAUMIER S., 2006a, *sms pour la science. Corpus de 30.000 sms et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve.

FAIRON C., KLEIN J. *et* PAUMIER S., 2006b, *Le langage sms, étude d'un corpus informatisé à partir de l'enquête "faîtes don de vos sms à la science"*, Cental (Cahier du Cental), Louvain-la-Neuve.

GLOR HOWARD Paul, 1993, *The Design and Analysis of Efficient Lossless Data Compression Systems*, Brown University, Providence.

International Organisation for Standardization, 2012-03-20 www.iso.org.

KOCH Peter *et* ÖSTERREICHER Wulf, 2001, « Gesprochene Sprache und geschriebene Sprache », *in Lexikon der romanistischen Linguistik*, Günter Holtus, Tübingen, pp. 584-627.

LUHN H.P., 1958, "The Automatic Creation of Literature Abstracts", *in IBM Journal of Research and Development*, pp. 159-165.

MINEL Jean-Luc HDR, 2004, *Le résumé automatique de textes : solution et perspectives*, Sorbonne, Paris.

TLFi - Trésor de la Langue Française informatisé, 2012-03-24, www.cnrtl.fr/definition/.

YOUSFI-MONOD Mehdi, 2007, *Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus*, Thèse de doctorat à l'Université de Montpellier II, Montpellier.

YVON François, 2008, *Réorthographier des sms*, LIMSI, Paris.

De l'utilisation du dialogue naturel pour masquer les QCM au sein des jeux sérieux

Franck Dernoncourt[1]
(1) LIP6, 4 place Jussieu, 75005 Paris
franck.dernoncourt@lip6.fr

RÉSUMÉ

Une des principales faiblesses des jeux sérieux à l'heure actuelle est qu'ils incorporent très souvent des questionnaires à choix multiple (QCM). Or, aucune étude n'a démontré que les QCM sont capables d'évaluer précisément le niveau de compréhension des apprenants. Au contraire, certaines études ont montré expérimentalement que permettre à l'apprenant d'entrer une phrase libre dans le programme au lieu de simplement cocher une réponse dans un QCM rend possible une évaluation beaucoup plus fine des compétences de l'apprenant. Nous proposons donc de concevoir un agent conversationnel capable de comprendre des énoncés en langage naturel dans un cadre sémantique restreint, cadre correspondant au domaine de compétence testé chez l'apprenant. Cette fonctionnalité est destinée à permettre un dialogue naturel avec l'apprenant, en particulier dans le cadre des jeux sérieux. Une telle interaction en langage naturel a pour but de masquer les QCM sous-jacents. Cet article présente notre approche.

ABSTRACT

Of the Use of Natural Dialogue to Hide MCQs in Serious Games

A major weakness of serious games at the moment is that they often incorporate multiple choice questionnaires (MCQs). However, no study has demonstrated that MCQs can accurately assess the level of understanding of a learner. On the contrary, some studies have experimentally shown that allowing the learner to input a free-text answer in the program instead of just selecting one answer in an MCQ allows a much finer evaluation of the learner's skills. We therefore propose to design a conversational agent that can understand statements in natural language within a narrow semantic context corresponding to the area of competence on which we assess the learner. This feature is intended to allow a natural dialogue with the learner, especially in the context of serious games. Such interaction in natural language aims to hide the underlying MCQs. This paper presents our approach.

MOTS-CLÉS : Agent conversationnel éducatif, intelligence artificielle, jeu sérieux, questionnaire à choix multiple, système d'évaluation de réponses libres.

KEYWORDS : Educational conversational agent, artificial intelligence, serious game, multiple-choice questionnaire, automatic assessment of free-text answer.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 323–336,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Nous définirons dans cette première partie les concepts clés de l'article, nommément le contexte des jeux sérieux ainsi que les agents conversationnels qui constitue la solution que nous explorons pour répondre à la problématique de masquage des QCM.

1.1 Les jeux sérieux

Les jeux sérieux correspondent à une approche de l'apprentissage qui utilise des moyens ludiques. L'apprentissage peut se situer aussi bien dans le cadre d'une formation que dans un contexte de sensibilisation ou de communication (Thomas, 2004). Le marché des jeux sérieux présente une croissance exponentielle : atteignant déjà 1 milliard de dollars en 2004 (Sawyer, 2004), les spécialistes l'estimaient à environ 10 milliards de dollars en 2010.

Dialoguer avec un agent virtuel contribue à maintenir l'attention et la motivation du joueur dans un jeu sérieux. Actuellement, ce dialogue, que ce soit dans les jeux sérieux ou dans les jeux vidéo de type récit (*storytelling*) ainsi que dans la plupart des environnements informatiques pour l'apprentissage humain, est constitué de QCM : le joueur interagit donc avec le jeu avec des QCM, qui font office de dialogue.

Le dialogue est donc très contraint, réduisant ainsi l'apprentissage du joueur qui peut se contenter de cliquer sur une des possibilités sans véritablement réfléchir. Nous pensons que des systèmes de dialogue davantage flexibles peuvent constituer une réponse pertinente à ce problème.

1.2 Les agents conversationnels

Un dialogue est une activité verbale qui fait intervenir au moins deux interlocuteurs servant à accomplir une tâche ou simplement échanger des mots dans une situation de communication donnée. Il constitue une suite coordonnée d'actions (langagières et non-langagières) (Vernant, 1992).

L'idée d'une interaction homme-machine se basant sur le fonctionnement du langage naturel n'est pas nouvelle : elle a vu le jour dans les années 1950 avec le test de Turing. Néanmoins, cette problématique, aux niveaux conceptuel et pratique, demeure toujours d'actualité. Il existe, par exemple, des compétitions annuelles comme le Loebner Prize (Loebner, 2003) ou le Chatterbox Challenge visant à réussir un test de Turing en imitant l'interaction verbale humaine, mais aucun programme n'est parvenu à ce jour à atteindre le niveau d'un humain (Floridi et al., 2009).

Afin de définir des critères d'efficacité des agents conversationnels, nous allons prendre en compte les quatre critères suivants pré-conditionnant l'élaboration d'un système de dialogue intelligent et proposés par (Rastier, 2001) :

1. apprentissage : intégration au moins temporaire d'informations issues des propos de l'utilisateur ;
2. questionnement : demande de précisions de la part du système ;
3. rectification : suggestion de rectifications à la question posée, lorsque

nécessaire ;

4. explicitation : explicitation par le système d'une réponse qu'il a apportée précédemment.

Les agents conversationnels se divisent en deux classes principales :

- les agents conversationnels non orientés tâche destinés à converser avec l'utilisateur sur n'importe quel sujet avec une relation souvent amicale, tel ALICE (Wallace, 2009) ;
- les agents conversationnels orientés tâche, lesquels ont un but qui leur est assigné dans leur conception.

Les agents conversationnels orientés tâche sont eux-mêmes classés usuellement en deux catégories :

- les agents conversationnels orientés service, par exemple fournir un service de conseil sur un site Internet, telle l'assistante virtuelle Sarah de PayPal[1] ;
- les agents conversationnels éducatifs, dont le but est d'aider l'utilisateur à apprendre.

Notre travail se concentre sur les agents conversationnels éducatifs (*tutor bots*).

2 État de l'art

Après avoir posé les définitions de base dans la partie précédente, nous exposerons ici brièvement l'état de l'art sur l'architecture des agents conversationnels ainsi que sur les systèmes d'évaluation des réponses libres plus en détails.

2.1 Architecture d'un agent conversationnel

La figure 1 montre un exemple d'architecture d'un agent conversationnel. L'utilisateur entre une phrase que l'agent conversationnel convertit en un langage abstrait, ici Artificial Intelligence Markup Language (AIML) : cette traduction permet d'analyser le contenu de la phrase et de faire des requêtes via un moteur de recherche dans une base de connaissances. La réponse est générée via un langage abstrait, ici également AIML, qu'il faut traduire en langage naturel avant de la présenter à l'utilisateur.

Néanmoins, cette architecture est rudimentaire et très rigide. Il faut par exemple souvent mettre à jour la base de connaissances pour y inclure des connaissances sur l'utilisateur, notamment dans le cadre d'une activité de tutorat qui nécessite le suivi des acquis de l'utilisateur ainsi que de sa motivation. Un certain nombre d'agents conversationnels éducatifs ont déjà été conçus et implémentés, comme (Zhang et al., 2009), (De Pietro et al., 2005), (Core et al., 2006), (Pilato et al., 2008) ou encore (Fonte et al., 2009).

Diverses architectures ont été élaborées, voici les éléments communs à la plupart d'entre

[1] https://www.paypal-virtualchat.com/

elles :

- une base de connaissances inhérente au domaine, objet de l'application ;
- un gestionnaire de répliques ;
- des structures de stockage des échanges sous forme d'arborescences surtout dans les agents conversationnels éducatifs conçus dans le cadre d'un jeu vidéo.

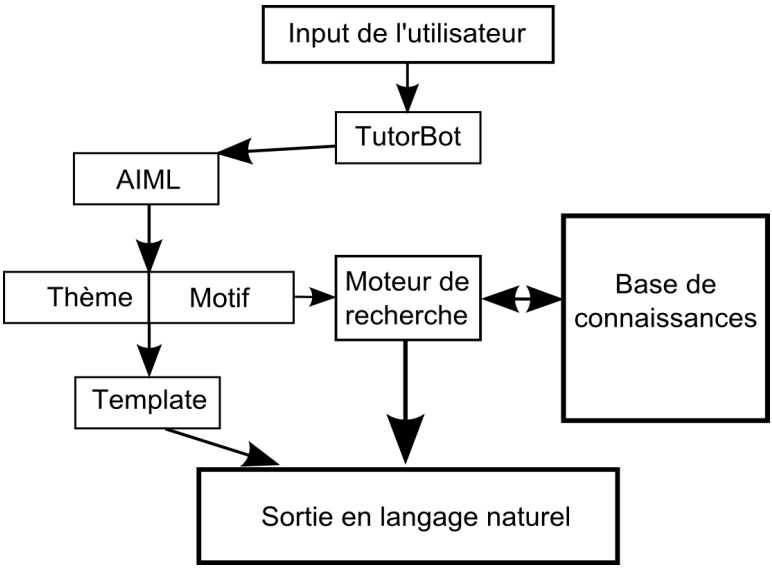


FIGURE 1 – Exemple d'architecture d'un agent conversationnel (TutorBot)
Source : (De Pietro et al., 2005).

Bien que sa simplicité d'utilisation ainsi que la performance relativement bonne des agents conversationnels l'utilisant le rendent attrayant, AIML est un langage très limité qui peut se résumer à un simple filtrage par motif, les motifs des inputs (phrases de l'utilisateur) et des outputs (réponses de l'agent conversationnel) étant définis en grande partie par extension et a priori.

2.2 Systèmes d'évaluation des réponses libres

En parallèle de la recherche sur les agents conversationnels, beaucoup de travaux se sont penchés sur l'évaluation des réponses libres, c'est-à-dire en langage naturel, données par des apprenants. Ces travaux sont motivés par les résultats expérimentaux montrant les limites des QCM, en tant qu'outil d'évaluation de la connaissance des

apprenants (Whittington et Hunt, 1999), ainsi que sa complémentarité avec les réponses libres (Anbar, 1991). Par connaissance, nous entendons ici et dans le reste de l'article non seulement la capacité à retranscrire des informations précédemment apprises, mais également la capacité à opérer des raisonnements de base montrant la compréhension du sujet.

Par exemple, (Anbar, 1991) a montré que les étudiants qui excellent lors des examens oraux auront tendance à avoir des performances médiocres dans les QCM. Inversement, les résultats aux QCM ne permettent pas de prédire les performances de l'apprenant dans le cadre d'un dialogue en langage naturel.

Nonobstant ces limitations bien connues des QCM, ces derniers représentent toujours l'outil le plus utilisé pour les évaluations des apprenants. Ce paradoxe s'explique simplement par le coût beaucoup plus élevé des méthodes alternatives : s'il est trivial de corriger automatiquement les QCM, il n'en va pas de même des autres méthodes, lesquelles nécessitent, étant données les techniques actuelles, des interventions humaines longues, donc coûteuses.

L'évaluation automatique des réponses libres a toutefois également ses détracteurs, qui soulignent que le fait qu'évaluer un essai est une tâche par nature complexe et subjective. Cependant, cette subjectivité ayant pour conséquence une variation de notes non négligeable parmi les correcteurs humains, le système d'évaluation automatique pourra au moins être consistant dans sa subjectivité.

Les premières recherches sur l'évaluation automatique apparurent il y a une cinquantaine d'années. Un des projets remarquables fut le Project Essay Grade, dirigé par Ellis Batten Page à l'université Duke (Page, 1968). Ses travaux se sont basés sur l'utilisation des caractéristiques stylistiques de la réponse de l'apprenant, tels la taille des mots et le nombre de prépositions, pour prédire la note du correcteur humain. Dans ses dernières expériences (Page, 1995), ce système semble prédire la note du correcteur humain plus précisément que ne le fait un second correcteur humain.

À la fin des années 1980, une nouvelle technique a été développée en vue de mieux saisir les concepts sous-jacents à un texte : l'analyse sémantique latente (LSA) (Deerwester et al., 1988 ; Deerwester et al., 1990). Cette technique fut dans un premier temps utilisée dans le cadre de recherche de l'information ; elle ne fut que plus tard appliquée à l'évaluation des réponses libres. La LSA serait aisée à réaliser si un mot ne correspondait qu'à un seul concept, et inversement. Néanmoins, dans les langages naturels, un mot peut avoir différentes significations : un mot peut subséquemment faire référence à différents concepts, faisant ainsi apparaître une ambiguïté à l'échelle du mot. La LSA utilise le contexte dans lequel le mot est utilisé afin de lever l'ambiguïté, autrement dit de comprendre à quel concept le mot fait référence dans le contexte donné. Par exemple, le mot *vol* désigne très certainement le concept de soustraire frauduleusement le bien d'autrui si le mot est utilisé à proximité des mots *butin* et *dérober*. Par contre, si mot *vol* est proche des mots de *ciel* et *oiseau*, *vol* désigne alors probablement un moyen de locomotion aérienne. La figure 2 illustre l'objectif de la LSA.

La LSA ne prend pas en compte l'ordre des mots, ni a fortiori les relations syntaxiques ou logiques. En outre, elle peut s'avérer assez coûteuse d'un point de vue computationnel. Malgré cela, des expériences ont montré que les scores de qualité

globale à un essai donnés par des experts sont moins précis que le score résultant d'une LSA (Landauer, 1998). Ce résultat surprenant est néanmoins à relativiser au vu des limitations de la LSA précédemment mentionnées et dépend évidemment des conditions de l'expérience.

Une approche totalement différente de la LSA a été adoptée par l'Educational Testing Service (ETS). ETS est la plus grande organisation privée à but non lucratif de mesure et d'évaluation éducative au monde. Faisant passer plus de 20 millions d'examens annuellement (TOEFL, GRE, GMAT, etc.), ETS peut ainsi avoir accès à des corpus considérables. Depuis plus d'une vingtaine d'années, son département R&D travaille sur des solutions permettant de noter automatiquement les réponses des candidats. Après avoir essayé d'utiliser la LSA afin de classifier les réponses (Burstein et al., 1996), ETS a décidé de s'en éloigner pour développer la technologie c-rater (Leacock et al., 2003), C pour contenu, qui se focalise sur les réponses de petite taille, allant de quelques à une centaine de mots. C-rater se base sur un pré-traitement de la réponse suivant l'architecture présentée à la figure 3. Ce pré-traitement permet de faire apparaître dans la réponse diverses caractéristiques linguistiques, tels les POS tags, les lemmes de chaque mot ou la présence de négation. Ces caractéristiques linguistiques sont ensuite utilisées pour comparer la réponse du candidat avec une réponse modèle à l'aide d'un algorithme de détection des concepts nommé *Goldmap*. Dans un premier temps, Goldmap était basé sur un ensemble de règles de filtrage par motif déterminées de façon binaire. Bien que cela permettait de comprendre aisément les décisions, la binarité des règles induisait un manque important de flexibilité. Afin de faire face à ce problème, Goldmap adopte à présent une approche probabilistique en se basant sur le principe d'entropie maximale pour la détection des concepts et en intégrant une dizaine de règles ad hoc. Les résultats obtenus semblent prometteurs selon leurs auteurs (Leacock et al., 2003). Cependant, à notre connaissance il n'existe pas à ce jour de test de performance standardisé pour comparer les différents systèmes d'évaluation automatique : il est donc difficile de comparer efficacement les différents systèmes.

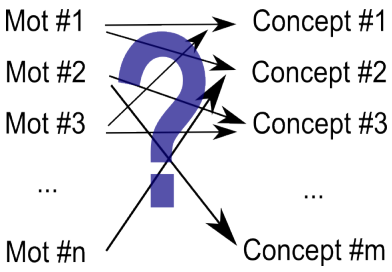


FIGURE 2 – Objectif de la LSA : trouver les concepts auxquels les mots sont associés.

Outre la LSA et c-rater, il est intéressant de noter que beaucoup d'articles soulignent les apports potentiels de la traduction automatique vers l'évaluation de réponses libres. Un des meilleurs exemples est la méthode BLEU (Papineni et al., 2001). Conçue originellement pour évaluer et classer les systèmes de traduction automatique, la

méthode BLEU a été appliquée avec succès à l'évaluation de réponses libres. La méthode repose sur la comparaison entre le texte candidat et un ensemble de textes modèles. Appliquée à la traduction, le texte candidat correspond à la sortie du système de traduction automatique, et les textes modèles correspondent à des traductions réalisées par des experts humains. La note donnée par BLEU au texte candidat se base sur le nombre de N-grammes communs entre le texte candidat et les textes modèles, ce qui s'avère être une mesure efficace malgré sa simplicité, mais est toutefois très sensible au choix d'écriture dans les textes modèles. Lorsque BLEU est appliqué à l'évaluation de réponses libres, le texte candidat correspond alors à la réponse de l'apprenant, et les textes modèles correspondent à des réponses types données par les professeurs. Néanmoins, BLEU présente des limitations importantes, comme par exemple la mauvaise gestion des négations : une phrase niant un fait A aurait par exemple presque le même score qu'une phrase affirmant A.

Au-delà de la méthode BLEU, il est intéressant de remarquer que le domaine de la traduction ainsi que de l'évaluation cherche le même idéal : trouver un formalisme dans lequel les faits pourraient être exprimés indépendamment de langage.

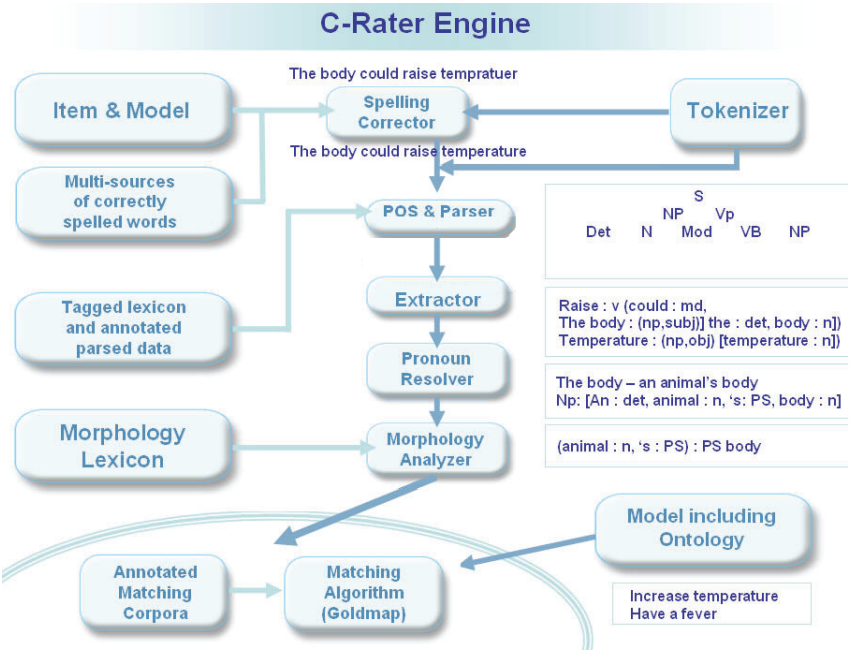


FIGURE 3 – Architecture de c-rater. Source : Sukkarieh et al., 2009.

3 Approche

Nous avons vu dans la partie précédente que beaucoup de travaux se sont penchés sur les systèmes d'évaluation de réponses libres. Dans cette partie, nous mettrons en exergue les particularités de notre approche, en particulier les particularités afférentes à l'évaluation de la réponse libre à l'aune du QCM sous-jacent ainsi qu'à l'environnement de jeu sérieux.

3.1 Particularités des QCM

Nos travaux ont pour but de donner une note à la réponse de l'apprenant. Dans notre approche, nous nous distinguons des systèmes d'évaluation classiques de réponses libres de par deux points majeurs :

- La réponse de l'apprenant n'est pas notée par rapport à des réponses modèles, mais est reliée à un QCM sous-jacent ;
- Une interaction est possible avec l'apprenant, car le système a la forme d'un agent conversationnel.

Ainsi, les recherches se sont penchées sur l'évaluation de réponses libres mais à notre connaissance aucune n'a cherché à évaluer une réponse libre à l'aune d'un QCM sous-jacent. Nous allons donc élaborer des variantes aux techniques habituelles (LSA, BLEU et c-rater) afin de les adapter à l'utilisation de QCM.

L'intérêt d'apparenter la réponse de l'utilisateur à un QCM est multiple. D'une part, de nombreux tests d'évaluation se présentent actuellement sous forme de QCM : nous pourrions ainsi nous baser directement sur les tests existants. D'autre part, la littérature sur la génération automatique de QCM à partir d'ontologie est riche (Papasalouros et al., 2008) : nous pourrons donc à terme avoir un système complet d'évaluation directement à partir des ontologies, voire des supports de cours. Le QCM permet de faire la jonction entre la base de connaissances que constitue le cours et les tests donnés à l'apprenant.

Dans un QCM, l'apprenant choisit une ou plusieurs réponses. Outre les choix corrects, il existe également un certain nombre de choix incorrects. Ces choix incorrects permettent de détecter la présence d'erreur chez l'apprenant de façon active, c'est-à-dire en vérifiant directement si la réponse ne contient pas le choix incorrect. Cette détection active des erreurs est absente de la plupart des systèmes d'évaluation de réponses libres car ces derniers ne reposent que sur la comparaison avec des phrases modèles. Nous pouvons par conséquent identifier ces erreurs alors que les systèmes classiques ont tendance à les ignorer.

Le fait que le système soit sous la forme d'un agent conversationnel nous permet naturellement de faire face plus aisément aux situations où la réponse de l'apprenant ne réussit pas à être directement évaluée par le système : via l'agent conversationnel, une nouvelle question pourra être posée à l'apprenant afin de l'inviter à reformuler ou préciser sa réponse. Cette interaction avec l'agent conversationnel peut être comparée aux tests oraux avec un examinateur humain et permet donc d'éviter les inconvénients émanant des examens écrits classiques qui sont par nature statiques.

3.2 Insertion dans un environnement ludique et sérieux

La simulation d'un dialogue naturel avec le joueur dans un jeu vidéo date d'une trentaine d'années. Le jeu d'aventure *King's Quest I: Quest for the Crown* développé par Sierra On-Line et publié en 1984 figure parmi les pionniers dans le genre. Ce n'est que récemment que la modalité conversationnelle a été utilisée à des fins pédagogiques, notablement dans le jeu Façade (Mateas et al., 2005), que nous allons très brièvement présenter dans le paragraphe suivant.

Dans Façade, le joueur est invité à un dîner où se déroule un conflit marital : l'objectif du joueur est de réconcilier le couple. Pour cela, le joueur entre des phrases de manière écrite, et les deux membres du couples répondent oralement. La figure 4 montre une capture d'écran dans laquelle le joueur demande à la femme Grace si elle se sent énervée vis-à-vis de son mari Trip. En interagissant ainsi avec le couple, le joueur apprend à mieux comprendre les relations de couple.

FIGURE 4 – Capture d'écran de jeu Façade. Le joueur interagit avec le couple.

Néanmoins, jusqu'à présent, ce genre de système de dialogue repose essentiellement sur le repérage de mots clés en fonction desquels le scénario du jeu s'adapte et ne fait pas appel à un QCM sous-jacent. Afin de nous focaliser sur les aspects agent conversationnel et QCM, nous intégrons notre système au sein de la plate-forme Learning Adventure[2] (Carron, 2010).

Learning Adventure est un environnement ouvert en 3D, en ligne et multijoueur où l'apprenant doit réaliser des quêtes en réalisant diverses activités qui le font interagir avec l'environnement et les autres joueurs. L'accent est mis sur le caractère immersif du jeu, à l'instar des MMORPG populaires actuels. L'interaction avec les autres joueurs, autrement dit avec les autres apprenants, est une dimension importante du jeu car elle contribue grandement à la motivation du joueur : le QCM devient pas un jeu solitaire, mais un jeu social, où entrent alors les mécanismes classiques de motivation par les pairs (Dickey, 2007) (Kim et al.,2009).

[2] http://learning-adventure.eu

Outre la motivation résultant de cette collaboration et compétition entre les apprenants, cette dimension multijoueur peut également donner l'occasion pour un tuteur humain d'intervenir dans le jeu. Une telle intervention peut avoir plusieurs objectifs : aider les apprenants dans les tâches réputées difficiles, renforcer les relations élèves-professeur en partageant un moment ludique, etc.

La modalité en ligne du jeu présente quant à elle de nombreux autres intérêts, en particulier s'assurer que le contenu pédagogique est à jour, suivre aisément l'avancement des différents apprenants et faciliter le déploiement de nouveaux contenus.

La figure 5 illustre un QCM qui apparaît dans le cadre du jeu. La figure 6 présente l'éditeur de scénarii, qui permet notamment d'ajouter et modifier aisément des QCM sans avoir aucune compétence informatique particulière. Notre système a pour objectif à terme de rendre le QCM invisible et d'utiliser l'éditeur de scénarii pour permettre à l'enseignant d'inclure les QCM ainsi que les autres éléments du scénario pédagogique.

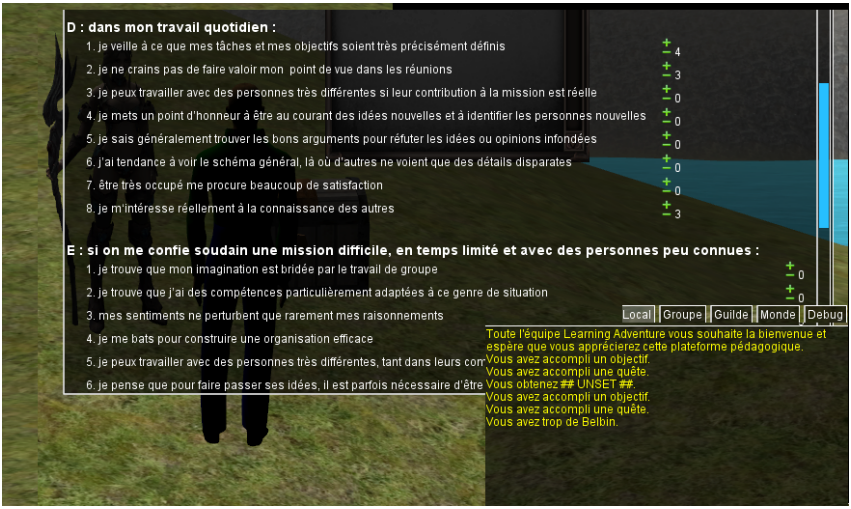


FIGURE 5 – Capture d'écran de la plate-forme Learning Adventure (Carron, 2010)

À l'instar de c-rater d'ETS, nous opérons un pré-traitement de la phrase de l'apprenant, comme le montre la figure 3 : après une étape liminaire de correction de surface (orthographe, ponctuation, etc.), la réponse est analysée lexicalement, puis étiquetée grammaticalement et parsée, avant de passer à l'étape de désambiguïsation référentielle et d'analyse morphologique. Comme ces processus sont supposés être connus chez le lecteur, nous ne les détaillerons pas ici. Sur le plan technique, ces étapes (hormis la première qui utilise le correcteur Hunspell) se basent essentiellement sur la librairie open-source et gratuite NLTK (Bird et al. 2009) qui offre de nombreuses fonctionnalités de traitement des langages.

Une fois la réponse pré-traitée, chaque mot de la réponse est décrit par un ensemble

d'attributs. Nous pouvons alors la comparer avec les différentes réponses du QCM. Les données extraites lors de la phase de pré-traitement contiennent de nombreuses informations essentielles pour la comparaison, telle la présence de négation.

Afin d'accroître la robustesse du système, nous envisageons de regrouper les résultats issus de la comparaison se basant sur la réponse pré-traitée avec les algorithmes statistiques LSA et BLEU. Cette comparaison multiple ainsi que les résultats obtenus seront exposés en détail dans un futur article, le but du présent article étant de présenter globalement les objectifs, le contexte du problème ainsi que la structure générale de notre système.

Les agents conversationnels classiques n'utilisent que très peu les outils usuels du TALN, comme le montre cet exemple suivant issu du site officiel de Cleverbot.com, qui est pourtant un des agents conversationnels les plus réputés :

Human: Hi.

Cleverbot: Are you a boy?

Human: Yes I am a boys

Cleverbot: Are you a girl?

Dans cet exemple, nous voyons que la faute de frappe de l'utilisateur qui a écrit « boys » au lieu de « boy » a manifestement empêché l'agent conversationnel Cleverbot de comprendre la phrase et de proposer une suite cohérente au dialogue.

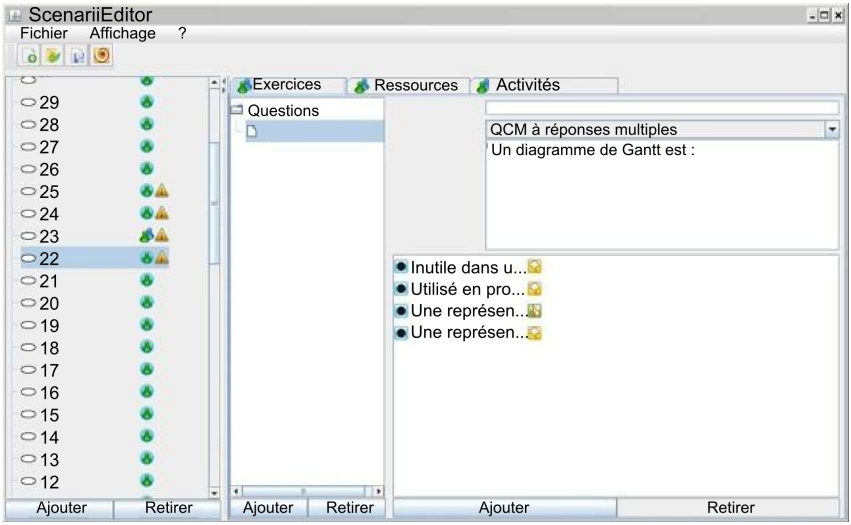


FIGURE 6 – L'éditeur de scénarii pour Learning Adventure

En restreignant le champ sémantique et en précisant son objectif, nous pouvons ainsi intégrer les techniques usuelles du TALN dans notre agent conversationnel afin de rendre transparents les QCM vis-à-vis de l'apprenant.

Enfin, comme le montre (D'Mello et al., 2010), l'apprentissage par agent conversationnel est amélioré lorsque la modalité est orale et non écrite. Par conséquent, nous utilisons Dragon NaturallySpeaking 11, qui est le leader de la reconnaissance vocale et édité par la société Nuance, ainsi que le logiciel AT&T Natural Voices® Text-to-Speech pour transmettre les réponses de l'agent conversationnel sous forme orale. À noter que ces deux logiciels ne sont pas libres.

4 Conclusions et perspectives

Cet article a présenté une approche nouvelle pour évaluer les apprenants en se basant sur des QCM masqués par un agent conversationnel au sein d'un jeu sérieux. La nature interactive du dialogue peut apporter au système d'évaluation une dimension nouvelle, permettant notamment de faire des demandes de clarification (Purver et al., 2003).

Une des difficultés dans la recherche de systèmes d'évaluation de réponses libres est l'absence de benchmarks, absence que certains expliquent par des raisons de propriété intellectuelle (Sukkarieh et Blackmore, 2009). Quelles qu'en soient les raisons, cette lacune est gênante pour la recherche dans le domaine.

Depuis quelques mois, trois initiatives majeures MITx, Coursera et Udacity ont été lancées ; leur objectif est de fournir aux internautes des cours en ligne gratuits, qui ont déjà attiré plus de 100 000 étudiants. Tous trois reposent en grande partie (en plus des tests de programmation dans lesquels le code de l'apprenant est évalué sur un jeu de tests) sur des QCM pour évaluer les apprenants, à défaut de systèmes plus efficaces. Or, ces QCM sont critiqués comme étant une des limites de ce genre de cours en ligne dont l'évaluation est entièrement automatique afin de pouvoir garantir la gratuité vis-à-vis d'un nombre important d'apprenants. La demande de masquage des QCM est donc très importante et continuera de s'accentuer par le nombre croissant de cours en ligne.

Au-delà des contextes d'apprentissage, un tel système pourrait également être utilisé dans d'autres domaines comme l'aide personnalisée, à l'instar de celle fournie par les centres d'appel qui est en général très scriptée, c'est-à-dire suivant des scénarii très peu flexibles, correspondant à un enchaînement de QCM.

5 Remerciements

Je souhaite particulièrement remercier mon directeur de thèse Jean-Marc Labat pour ses précieux conseils, indispensables à la réalisation de ce projet, ainsi que la DGA pour son soutien financier. Je souhaite également remercier Thibault Carron pour ses nombreuses idées ainsi que son aide sur Learning Adventure dont il est un des initiateurs.

6 Références

ALHADEFF, E. (2008). Reconciling Serious Games Market Size Different Estimates. *In Futurlab Business & Games Magazine* - Numéro du 9 avril 2008.

BIRD, S., KLEIN, E. et LOPER, E. (2009). Natural Language Processing with Python. O'Reilly Media.

BURSTEIN, J., KAPLAN, R., WOLFF, S. et LU, C. (1996). Using Lexical Semantic Techniques to Classify Free-Responses. *In Proceedings of SIGLEX 1996 Workshop, Annual Meeting of the Association of Computational Linguistics*, University of California, Santa Cruz.

CARRON T., MARTY JC. et TALBOT S. (2010). Interactive Widgets for Regulation in Learning Games. *The 10th IEEE Conference on Advanced Learning Technologies*, Sousse, Tunisia.

CORE, M., TRAUM, D., LANE, H. C., SWARTOUT, W., GRATCH, J., LENT, M. V. et MARSELLA. S. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation* 82(11):685–701, 2006.

D'MELLO, S., GRAESSER, A. et KING, B. (2010). Toward Spoken Human-Computer Tutorial Dialogues. *Human-Computer Interaction,* (4):289--323.

DE PIETRO, O., M. DE ROSE et G. FRONTERA. (2005). Automatic Update of AIML Knowledge Base in E-Learning Environment. *In Proceedings of Computers and Advanced Technology in Education.*, Oranjestad, Aruba, August (2005): 29–31.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T., HARSHMAN, R., LOCHBAUM K. et STREETER, L. (1988). Brevet (US Patent 4,839,853).

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. et HARSHMAN, R., Indexing by Latent Semantic Analysis. *In Journal of the Society for Information Science*, vol. 41, no 6, 1990, p. 391-407.

DICKEY, M. D. (2007). Game design and learning: A conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development*, 55(3), 253–273.

FLORIDI, L., TADDEO, M. et TURILLI, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. *Minds and Machines*. Springer.

LANDAUER, T.K., LAHAM, D., REHDER, B. et SCHREINER, M.E. (1997). How Well can Passage Meaning be Derived Without Using Word Order? A Comparison of Latent Semantic Analysis and Humans, *in Proceedings of the 19th Annual Conference of the Cognitive Science Society*.

LEACOCK, C. et CHODOROW, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and Humanities.* pp. 389-40.

LOEBNER, H. (2003). Home Page of the Loebner Prize - The First Turing Test. http://www.loebner.net/Prizef/loebner-prize.html [consultée le 03/03/2012].

KIM, B., PARK, H. et BAEK, Y. (2009). Not just fun, but serious strategies: Using meta-

cognitive strategies in game based learning. *Computers & Education*, 52(4), 800-810. doi:10.1016/j.compedu.2008.12.004.

MATEAS, M. et STERN, A. (2005). Structuring Content in the Façade Interactive Drama Architecture. *AIIDE*.

PAGE, E.B. (1968). The Use of the Computer in Analyzing Student Essays. *International Review of Education*, 14, 210-224.

PAGE, E.B. (1995). The Computer Moves into Essay Grading: Updating the Ancient Test, *Phi Delta Kappan*, 76(Mar), 561-565.

PAPASALOUROS, A., KOTIS, K. et KANARIS, K. (2008). Automatic generation of multiple-choice questionsfrom domain ontologies. *IADIS e-Learning*, Amsterdam.

PAPINENI, K., ROUKOS, S., WARD T. et ZHU, W. (2001). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on Association for Computational Linguistics*. 311—318.

PEREZ, D., ALFONSECA, E. et RODRIGUEZ, P. (2004). Application of the BLEU method for evaluating free-text answers in an e-learning environment. *In Proceedings of the Language Resources and Evaluation Conference* (LREC).

PILATO, G., ROBERTO P. et RICCARDO R. (2008). A kst-based system for student tutoring. *Applied Artificial Intelligence 22*, no. 4: 283-308.

PURVER, M., GINZBURG, J. et HEALEY, P. (2003). On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*. Springer. 235—255.

RASTIER, F. (2001). Sémantique et recherches cognitives, *PUF* (2e éd).

SUKKARIEH, J. Z. et BLACKMORE, J. (2009). c-rater: Automatic content scoring for short-constructed responses. *Florida Artificial Intelligence Research Society (FLAIRS) Conference*, Sanibel, FL.

SAWYER, B. (2004). Serious Games Market Size. *Serious Games initiative Forum 01-04-2004*.

THOMAS, P., éditeurs (2010). *Actes de RJC EIAH 2010 (Rencontres Jeunes Chercheurs en Environnements Informatiques pour l'Apprentissage Humain)*, Lyon. ATIEF.

VERNANT, D. (1992). Modèle projectif et structure actionnelle du dialogue informatif. *In Du dialogue, Recherches sur la philosophie du langage*, Vrin éd., Paris, n°14, p. 295-314.

WALLACE. , S. (2009). Parsing the Turing Test, The Anatomy of A.L.I.C.E. Springer.

WHITTINGTON, D. et HUNT, H. (1999). Approaches to the computerized assessment of free text responses. *In Danson, M. (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.

ZHANG, H. L., Z. SHEN, X. TAO, C. MIAO et B. Li. (2009). Emotional agent in serious game (DINO). *In Proceedings of The 8th International Conference on Autonomous Agents and Multi-agent* Systems-Volume 2, 1385–1386.

Extraction d'indicateurs de construction collective de connaissances dans la formation en ligne

Alexandre Baudrillart[1, 2]
(1) Université Stendhal Grenoble 3, BP 25, 38040 Grenoble Cedex 9, France
(2) Université de Lyon, CNRS INSA-Lyon, UMR5205 F-69621, France
alexandre.baudrillart@u-grenoble3.fr

RÉSUMÉ

Dans le cadre d'apprentissages humains assistés par des environnements informatiques, les techniques de TAL ne sont que rarement employées ou restreintes à des tâches ou des domaines spécifiques comme l'ALAO (Apprentissage de la Langue Assisté par Ordinateur) où elles sont omniprésentes mais ne concernent que certaines dimensions du TAL. Nous cherchons à explorer les possibilités ou les performances des techniques voire des méthodes de TAL pour des systèmes moins spécifiques dès lors qu'une dimension de réseau et de collectivité est présente. Plus particulièrement, notre objectif est d'obtenir des indicateurs sur la construction collective de connaissances, et ses modalités. Ce papier présente la problématique de notre thèse, son contexte, nos motivations ainsi que nos premières réflexions.

ABSTRACT

Collaborative Knowledge Building Indicators Extraction in Distance Learning

Natural Language Processing techniques are still not very much used within the field of Technology Enhanced Learning. They are restricted to specific tasks or domains such as CALL (standing for Computer Assisted Language Learning) in which they are ubiquitous but do not match every linguistic aspect they could process. We are seeking to explore possibilities or performances of thoses techniques for less specific systems including a network or community aspect. More precisely, our goal is to get indicators about collective knowledge building and its modalities. This paper presents the problem and the background of our thesis problem, as well as our motivation and our first reflections.

MOTS-CLÉS : TAL, EIAH, formation en ligne, socio-constructivisme, acquisition des connaissances, apprentissage collaboratif en ligne.

KEYWORDS: NLP, TEL, distance learning, socio-constructivism, knowledge aquisition, collaboration, CSCL.

Introduction

Nous abordons dans cet article une problématique liée à un potentiel apport des traitements automatiques de la langue aux Environnements Informatiques pour l'Apprentissage Humain (EIAH) avec une finalité d'amélioration des performances pédagogiques.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 337–350,
Grenoble, 4 au 8 juin 2012.

Les Environnements Informatiques pour l'Apprentissage Humain permettent à des apprenants, à des tuteurs ou à des professeurs d'interagir avec et au travers d'un système. Les apprenants vont grâce à lui pouvoir prendre part à des situations d'apprentissages, participer individuellement ou en groupe à des activités. Le rôle du système est de faciliter l'apprentissage. Une manière de favoriser la réussite d'un apprenant dans son entreprise peut être de guider l'apprenant et de lui fournir des retours (*feedback*), sur les performances d'apprentissage. Ces retours peuvent présenter des informations plus ou moins brutes, ou des conseils précis ou directifs. Ce sont des indicateurs. Afin qu'un apprenant puisse en tirer un bénéfice, ces retours doivent être interprétables et exploitables.

Dans le cadre d'apprentissages collectifs à distance, un des rôles importants de l'environnement est de permettre l'interaction et la discussion entre les différents protagonistes et principalement entre les apprenants. C'est grâce à ce dispositif qu'un partage d'informations peut avoir lieu et que des débats peuvent s'instaurer, permettant ainsi la construction collective de connaissances.

Une des fonctionnalités qu'offre ce genre de système est la conservation de traces d'activités. Ces traces sont donc de deux types : des traces d'interactions avec le système, c'est-à-dire les actions réalisées par les utilisateurs et médiatisées par l'Interface Homme-Machine, et des traces d'interactions entre agents humains, supportées par la langue et donc textuelles, au travers du système (dans un cadre collaboratif). Ces échanges textuels véhiculent des idées, des états d'esprit ou des comportements sociaux qui participent à une construction individuelle et collective de connaissances. Les traces textuelles et d'actions s'organisent sur un axe temporel ce qui permet d'accéder à une dimension séquentielle de l'activité des participants. À ces traces s'ajoute la conservation de documents produits individuellement ou en collaboration qui sont alors le résultat attendu des échanges, qui synthétisent les idées et les cristallisent. Une production collective peut être construite en parallèle des échanges ou *a posteriori*, et est susceptible d'être augmentée, réduite ou refondue, à plusieurs reprises. Elle exprime un état des connaissances collectives.

La problématique que nous voulons aborder consiste à évaluer les possibilités d'applications des traitements automatiques des langues actuels pour extraire des indicateurs sur le niveau, la qualité et les modalités de connaissances construites collectivement et en réseau, dès lors que des traces d'interactions en langue naturelle entre les différents apprenants ou de productions collectives sont conservées et collectées. Cette extraction, qui doit résulter de traitements automatiques sur les différentes traces accumulées, a pour but d'avoir un impact sur les performances d'apprentissage. Cette problématique est encore vague et sujet à discussion, tant en ce qui concerne les situations d'apprentissage précises concernées, le matériau textuel susceptible d'être analyser et les indicateurs à produire.

Cet article a pour but de présenter les motivations, apporter des justifications de notre problématique. Il s'organise de la façon suivante. Nous commencerons par exposer un modèle de notation constituant pour nous une forme canonique de ce que nous pourrions vouloir automatiser : une première motivation directe. Nous présenterons ensuite dans quelle mesure sont utilisées certaines méthodes de TAL pour les EIAH et notamment pour les systèmes ALAO pour éclairer non pas un verrou scientifique à lever, mais une pluridisciplinarité encore embryonnaire : apportant des motivations indirectes. Nous exposerons ensuite en quoi l'Analyse Sémantique Latente, convient pour représenter la connaissance, pour simuler un modèle d'apprentissage ou pour évaluer des notions acquises par l'apprenant en tant qu'individu isolé. Nous essayerons de montrer qu'il est possible d'aller d'une approche statistique et connexionniste pour l'individu à un

modèle plus symbolique et dialogique pour un groupe d'apprenants, et donc à caractère social, ce qui rejoint notre problématique. Nous terminerons sur la nécessité d'amorcer nos recherches en nous focalisant sur un sous-problème et en construisant un corpus d'observation.

1 Motivation directe

1.1 Présentation des travaux

Dans (Hou et Wu, 2011), sont étudiées les caractéristiques des discussions synchrones médiatisées à but pédagogique. L'intérêt se porte sur l'impact d'une discussion synchrone et de sa médiatisation informatique par messagerie instantanée sur la construction collective de connaissances, les types d'interactions et les degrés de coordination. Un autre objectif est de voir émerger des motifs séquentiels de comportements c'est-à-dire de types de message caractérisant une discussion de haute qualité ou une discussion de basse qualité. Pour cela, une expérience à deux facettes est mise en place. Cette étude requiert l'intervention d'experts humains pour catégoriser et évaluer les interactions textuelles individuellement d'une part et la qualité des discussions d'autre part.

1.2 L'expérience

L'expérience consiste à observer à long terme, trois mois, des groupes d'étudiants de moins de dix membres chacun, constituant au total quarante apprenants, discuter et débattre par le biais d'une application de messagerie instantanée. Ces discussions portent sur un sujet fixé par leur professeur ; elles doivent mener les étudiants à des conclusions et des synthèses.

La première facette de l'expérience est une analyse quantitative des messages. Des experts humains, possédant des compétences psychologiques, classent les messages en types d'interaction représentant un certain comportement de l'apprenant au sein de chacune des discussions, comportements eux-mêmes regroupés en quatre catégories : élaboration de la connaissance, organisation et collaboration entre apprenants, interactions sociales et hors-sujet. Puis, ce sont les qualités des discussions entières qui sont évaluées par l'enseignant, expert dans la discipline qu'il enseigne et sur lesquelles portent les discussions. Elles sont notées selon quatre critères : clarification du sujet, collecte d'informations, profondeur d'analyse et conclusion. Ces notes permettent ensuite de construire deux catégories : haute qualité et basse qualité. Peu de précisions sont données à ce sujet.

La deuxième facette de l'expérience concerne les suites de comportements. Afin de faire émerger des continuités, des discontinuités ainsi que des dépendances entre comportements ou catégories de comportement, une analyse séquentielle des suites de comportements statistiquement significatives est réalisée.

1.3 Résultats

Les résultats quantitatifs concernant toutes les discussions montrent qu'entre la moitié et deux tiers des contributions sont hors-sujet.

Une autre part de 30 % est principalement constitué de messages concernant des échanges sur le savoir académique. Le reste concerne la coordination des étudiants ou des messages à caractères sociaux comme des remerciements ou des encouragements.

Les résultats relatifs à chaque qualité de discussion apparaissent et discriminent haute qualité et basse qualité. Le premier, le plus flagrant, met en évidence que les discussions de haute qualité ont généré quatre fois plus de contributions, ce qui souligne un rythme soutenu et une profondeur des discussions, ainsi qu'une motivation de la part des apprenants. Ces discussions sont plus variées dans l'élaboration de la connaissance, et l'identification de désaccords et la négociation du sens sont plus présentes. La dimension organisationnelle est presque inexistante au sein des discussions de basse qualité alors qu'elle apparaît de manière pertinente dans celles de haute qualité indiquant que ces apprenants explicitent la coordination de leurs démarches. Selon les auteurs, il résulte de la comparaison des chiffres que des interactions sociales telles que des encouragements ou des félicitations sont un ciment entre la construction de la connaissances et la coordination. Une autre observation intéressante est l'apparente indépendance des contributions hors-sujet vis-à-vis de la qualité des discussions. Ces discussions peuvent être à l'origine d'un climat propice à une meilleure qualité de discussion.

L'analyse séquentielle permet de mettre en évidence qu'il n'y a pas de véritable continuité d'un comportement particulier de construction de connaissance. Par contre, il existe des suites de comportements variés de construction de la connaissance, garante du maintien du focus des apprenants sur le sujet et de discussions plus approfondies. En outre, ces motifs séquentiels significatifs ne traduisent aucune dépendance des contributions hors-sujet vis-à-vis des autres catégories.

1.4 Discussion sur une automatisation

Nous pouvons nous interroger quant à la capacité de techniques informatiques à automatiser ces protocoles expérimentaux et à aboutir des conclusions similaires. Il faut noter que chaque analyse, classification ou évaluation est réalisée par des experts humains et qu'aucune remarque n'est donnée au sujet de traitements pouvant automatiser ces processus. La classification des messages est effectuée par des experts en psychologie, et les types de comportements sont eux-mêmes regroupés en catégories selon, deux niveaux. La qualité des discussions est évaluée, selon un barème par un expert du domaine : l'enseignant.

Identifier automatiquement la catégorie d'un message pourrait permettre la détection de séquences particulières et émettre des hypothèses sur la direction que prennent les échanges permettant d'inférer des conseils, des pistes ou encore simuler un participant virtuel (proche d'un *tuteur intelligent*), à cette discussion afin d'améliorer discussion et apprentissage par son biais. Ces processus décisionnels ne sont pas triviaux. Cette perspective nous amène à nous interroger sur la pluridisciplinarité qui existe entre EIAH et TAL : a-t-on les moyens de répondre à cette automatisation les méthodes voire les techniques et les ressources dont dispose la communauté ?

2 Motivations indirectes

D'autres motivations à notre problématique proviennent d'une utilisation très modérée et non entière du TAL. Pourtant, les connaissances ou les échanges à traiter au sein d'environnement d'apprentissage sont supportés par un matériau langagier.

Nous allons exposer des situations d'apprentissages assistées par ordinateur afin de montrer qu'il y a encore beaucoup d'opportunités à marier EIAH et TAL par l'intermédiaire d'un état de l'art encore partiel, en nous appuyant sur une classification proposée dans (Gurevych *et al.*, 2009).

L'actuelle utilisation des technologies du TAL au sein du champ des EIAH y est dépeinte et fractionnée en quatre catégories : génération automatique d'exercices, évaluation automatique de dissertation, aide à la lecture et à l'écriture et gestion de contenus et apprentissages collaboratif(s). La suite détaille à quelles activités pédagogiques ou éléments de l'ingénierie pédagogique fait référence chacune de ces catégories, et présenter en quoi est employé le TAL.

2.1 Génération automatique d'exercices

La première catégorie regroupe entre autres la génération automatique d'éléments d'exercices tels que les questions à choix multiples (Karamanis, 2006), ou d'exercices entiers tels que les exercices lacunaires (Lee et Seneff, 2007). Elle réunit la construction des intitulés, des réponses associées, la notation automatique des résultats ainsi que l'évaluation de l'efficacité de ces tests pour juger la qualité d'acquisition de connaissances et pour différencier les « bons » étudiants des « moins bons» .

La démarche générale consiste à extraire automatiquement de textes, à l'aide de motifs syntaxiques, la réponse à une question, notamment au sein de phrases définitoires. Pour les QCM, il faut ensuite générer l'énoncé qui va amener à la bonne réponse, en transformant affirmations en interrogations, en inversant la construction syntaxique pour l'anglais et en choisissant un des fameux pronoms 'WH'. En revanche, pour de simples exercices lacunaires, il suffit de capturer le matériau authentique qui constitue en lui-même l'objet du test. Il ne reste alors plus qu'à ôter l'élément qui constituera la lacune, et d'indiquer l'attente qui est souvent générique.

Dès lors qu'un choix multiple est envisagé, une dernière étape intervient dans la génération de l'exercice : le choix de « *distractors*» . Ces éléments représentent les autres choix possibles que les réponses attendues dont leur rôle et leur choix répondent à des critères particuliers comme une proximité sémantique suffisante sans pour autant installer une ambiguïté. Ces exercices attendent des réponses fermées, simplifiant ainsi la vérification de celles proposées par les étudiants : l'évaluation automatique reste donc relativement aisée.

Il faut noter que ces types d'exercices portent souvent sur des questions de vocabulaire, d'orthographe ou de conjugaison et ne concernent donc que l'apprentissage de la langue elle-même, maternelle ou seconde. Près de 90% des références de cette catégorie citées dans (Gurevych *et al.*, 2009) concernent l'apprentissage de la langue. Leurs emplois ne se limitent pas à cet apprentissage mais peuvent aborder d'autres disciplines, pour des questions de terminologie ou de compréhension.

2.2 Évaluation automatique de dissertation

L'évaluation de résumés ou de dissertations produits par des apprenants est une tâche pouvant éprouver le TAL. Cette seconde catégorie évalue des critères que nous regroupons selon quatre principaux points : lisibilité, focalisation sur le sujet à traiter (en évitant les hors-sujet) et les thèmes ou notions abordés, qualité d'argumentation et validité des propos représentant la réelle compréhension.

L'évaluation de la lisibilité est depuis longtemps traitée et utilise notamment des méthodes numériques qui consistent par exemple à compter des descripteurs de surface. Ces derniers sont par exemple le nombre moyen de syllabes par mot, de mots ou de syntagmes par phrase, ou encore le nombre d'hapax, de termes répétées exactement sans emploi anaphorique ou de synonymes (Burstein, 2009; Gurevych *et al.*, 2009).

La lisibilité dépend aussi d'une cohérence discursive, au moins à courte portée. Pour cela il faut détecter des ruptures thématiques inappropriées entre des unités textuelles adjacentes (Burstein, 2009). Détecter ces ruptures peut être réalisé par le biais de méthodes utilisant une représentation vectorielle et lexicale du sens. Le Text Tiling de Martin Hearst (Hearst, 1997) permet à l'aide de telles représentations de réaliser un découpage des paragraphes en ensembles de phrases cohérents.

Une dissertation n'est bien rédigée que si elle met en valeur une thèse soutenue par suffisamment d'arguments, eux-mêmes étayés par des faits. Une argumentation insuffisamment alimentée et structurée peine à convaincre et ne répond pas non plus à une nécessité d'exposer des notions attendues. Pouvoir construire un discours complet en restituant des connaissances organisées est un moyen de montrer l'acquisition, la compréhension et la maîtrise de notions (Trausan-Matu, 2010b). Cela permet d'aborder un point concernant l'évaluation de la compréhension voire de vérifier un caractère de vérité des propos.

C'est pourquoi, il est possible de considérer que la rédaction d'un essai respectant les points 2 et 3 est un indicateur de compréhension. L'appariement de ces productions avec des textes, fortement similaires, faisant autorité sur la question ou jouant le rôle d'étalons évalués par un groupe de juges humains pour chaque « note » donnée est une manière de pouvoir donner un score à la compréhension. Ce qui reviendrait à un calcul de similarité avec, par exemple, un cours (Dessus, 1999), ou une classification de nouvelles copies dans des catégories correspondant à chaque échelon de notes, par à un calcul de similarité maximale avec les copies représentatives des catégories (Foltz *et al.*, 1999). Nous avons ainsi soit une répartition des copies selon leur similarité avec un « gold-standard » soit une tâche de classification supervisée.

Ces similarités sont calculées sur le contenu (le signifiant) mais doivent rendre compte du sens (le signifié). C'est pourquoi des modèles lexicaux de calcul et de représentation du sens peuvent être adaptés (pour plus de détails, voir 3.1). En outre, l'utilisation d'ontologies, de thésaurus spécifiques à un domaine (GeneOntology [1], UMLS [2]) ou de réseaux sémantiques plus généraux (WordNet [3]) peut permettre la manipulation plus directe de concepts et de sens en faisant abstraction des lexèmes qui les incarnent, permettant notamment d'unifier des synonymes.

1. http ://www.geneontology.org/
2. http ://www.nlm.nih.gov/research/umls/
3. http ://wordnet.princeton.edu/

2.3 Aide à la lecture et à l'écriture

Lire des textes en langue non maternelle ou contenant des termes spécifiques à un domaine, abordant des concepts inconnus (« loin au-delà de la zone proximale de développement») ou rédigés avec un style pompeux peut être une tâche difficile. C'est pourquoi, ces lectures peuvent nécessiter une aide extérieure sous la forme de simplifications de textes, de propositions de synonymes, de glossaires (Gaudio, 2007) ou encore de documents tiers explicitant définitions, concepts ou simplement traitant du même sujet.

Dans l'idéal, ces documents doivent être accessibles à l'apprenant tant en terme de vocabulaire que de connaissances pré-requises, tout en permettant l'acquisition de nouvelles connaissances. Les connaissances qu'ils transportent sont alors présentes dans ce que Lev Vygotsky, père du constructivisme, nomme la zone proximale de développement, « ni trop proches » du modèle de l'apprenant « ni trop éloignés» (Zampa, 2005).

La rédaction, quant à elle, peut être assistée en fournissant des correcteurs automatiques orthographiques et syntaxiques ou encore des dictionnaires de synonymes mais aussi en permettant d'identifier les notions exposées par l'apprenant et abordées dans le cours dans le cadre duquel la production écrite s'inscrit (Lemaire et Dessus, 2001).

L'identification d'une structure des thèmes et des notions couverts par l'écrit, et calquée sur la structure typographique ou logique peut mettre en évidence des problèmes de cohérence. En observant les couvertures respectives d'unités adjacentes, on peut alors identifier des ruptures thématiques (Lemaire et Dessus, 2001). En cela, ces possibilités rejoignent le *Text-Tiling* (Hearst, 1997) permettant d'identifier ces ruptures afin de délimiter des zones textuelles cohérentes.

2.4 Gestion de contenus et apprentissage collaboratif(s)

L'essor du Web a permis l'accès à des ressources en ligne comme des sites spécialisés ou des encyclopédies numériques, mais il a aussi permis de créer un savoir construit socialement dans des forums, des blogs ou encore des wikis. Dans (Gurevych *et al.*, 2009), les auteurs insistent quasi-exclusivement sur des travaux dans lesquels le TAL est utilisé pour organiser et structurer la connaissance notamment dans des wikis. Ces travaux sont dans l'ensemble très proche de l'ingénierie des connaissances et de la recherche documentaire.

Mais il s'agit aussi d'analyser des échanges dans le cadre de débats imposés, de forums de formations ou de discussions. Le projet européen LTfLL a notamment apporté une contribution non négligeable et on peut noter l'existence du module PolyCAFe (Rebedea *et al.*, 2010; Trausan-Matu, 2010a,b) qui fournit des *feedbacks* aux différents protagonistes d'une situation d'apprentissage sous forme de débats. Nous revenons en particulier sur ce module dans la section suivante.

Entre autres, des travaux sur l'analyse automatique de forums de formation à distance alimentent aussi cette catégorie. Dans (Sidir *et al.*, 2006), ce sont les forums libres c'est-à-dire sans limite de temps et sans tâche fixe qui sont ciblés. Ces travaux essayent notamment d'identifier s'il existe « des processus de co-construction de connaissances entre apprenants indépendamment des interventions des tuteurs» .

Une analyse thématique automatique avec le logiciel ThemAgora souligne une progression discursive en rapport avec celle de la formation. Une analyse linguistique et manuelle du discours

fondée sur le modèle d'exposition de Yamada dégage plusieurs mouvements dans le discours correspondant à des phases différentes de cette co-construction de connaissance.

2.5 Discussion

Ces premiers éléments d'un état de l'art nous amènent à deux conclusions. La première est que l'utilisation du TAL dans les EIAH semble se trouver principalement dans l'Apprentissage de la Langue Assisté par Ordinateur (ALAO) et que de surcroît l'apprentissage en groupe n'est peut-être pas ce qui tire le plus parti du TAL. La seconde conclusion réside dans le fait que toutes les axes linguistiques ne sont pas couverts par l'application du TAL aux EIAH. En effet, la dimension rhétorique et l'*argumentative zoning* (Teufel, 1999) semblent délaissés et l'utilisation de modèles discursifs ou dialogiques encore embryonnaires. C'est pourquoi, le TAL a sa place dans les EIAH car il pourrait notamment apporter des indicateurs qualitatifs sur la construction de connaissances, ce qui fait actuellement défaut, laissant donc un espace encore vierge entre TAL et EIAH et de nombreuses opportunités (Antoniadis, 2008; Burstein, 2009; Antoniadis *et al.*, 2009)

3 Du cognitif/connexionniste au socio-constructiviste : de LSA vers le discours et le dialogue au travers d'un réseau

3.1 L'Analyse Sémantique Latente : d'un modèle documentaire, à une représentation des connaissances et un modèle de leurs acquisitions

L'Analyse Sémantique Latente est un modèle statistique et lexical de représentation vectorielle du sens latent des termes porté par les relations de co-occurence locale qu'ils entretiennent au sein de documents d'un corpus (Deerwester *et al.*, 1990) .

Un corpus de documents est représenté par une matrice de co-occurrence qui associe à chacun de ces documents (ou tout autre unité textuelle de grain pertinent) le nombre d'occurrences de chacun des termes du corpus. Une décomposition en valeurs singulières permet de réduire le rang de la matrice aux termes les pertinents et d'obtenir des vecteurs de mêmes dimensions. Ce modèle permet ainsi de calculer la proximité sémantique entre termes ou documents, c'est-à-dire les contextes qu'ils partagent directement ou indirectement, grâce à un simple cosinus entre vecteurs.

D'abord appliqué dans le champ de la recherche documentaire (Dumais, 1991), ce modèle s'est vu employé dans différents autres domaines comme l'apprentissage et l'acquisition de connaissances, et utilisé dans des EIAH fournissant certains indicateurs.

LSA permet de simuler certains processus cognitifs et notamment l'acquisition de vocabulaire par exposition à des textes. (Landauer et Dumais, 1997) ont montré qu'en faisant traiter par LSA autant de mots/textes qu'un jeune entre 2 et 20 ans, le nombre de nouveaux mots acquis par jour par celui-ci est du même ordre que le nombre de nouvelles paires de mots proches sémantiquement construites. De plus, la simulation de réponses à des tests de synonymies du TOEFL par LSA, après exposition au contenu d'une encyclopédie montre des résultats proches de

ceux atteints par une population d'élèves étrangers (Landauer et Dumais, 1997). La simulation de réponses de LSA à des QCM sur des notions de mathématiques après traitement de quelques cours obtient des résultats moins brillants. Dans ces deux cas, les réponses choisies sont celles qui sont les plus proches sémantiquement des énoncés dans le modèle vectoriel calculé à partir du corpus d'apprentissage. La différence de performance met alors en évidence l'importance de ce corpus pour ces processus décisionnels. LSA a aussi été mis à profit dans des tâches de notations et d'évaluation de productions écrites individuelles (Foltz *et al.*, 1999). Dans (Dessus et Lemaire, 2002) les auteurs ont procédé à des expériences similaires mais avec un traitement différent. Une indexation thématique sur deux niveaux hiérarchiques (Sujets et notions) de cours de sociologie permet d'évaluer ces productions tant en termes de cohérence que de couverture du sujet.

En effet, le traitement de ce corpus par LSA selon ces deux niveaux permet de fournir des indicateurs sur la couverture du sujet selon un axe macroscopique et microscopique. Le calcul de similarité par LSA entre la production de l'apprenant et les différents sujets et notions du cours permet de déterminer quel est le contenu du cours couvert par sa copie d'une part, mais aussi d'appréhender le plan, l'organisation de sa copie grâce à un appariement à des grains plus petits. Ces indications de couvertures au niveau microscopique permettent aussi de fournir un retour sur la cohérence textuelle inter-phrastique (Foltz *et al.*, 1999; Dessus, 1999, 2000), d'une manière proche du *Text-Tiling* (Hearst, 1997), et d'identifier des ruptures et des changements thématiques brutaux et inattendus faisant baisser la qualité de la copie.

Ce principe est intégré dans le logiciel Apex (Dessus et Lemaire, 2002) qui permet de guider l'apprenant dans l'exploration de documents ou de cours à des fins d'apprentissage. Ce système propose à un étudiant de lire des textes en rapport avec une requête qu'il fournit puis de dire s'il peut ou non résumer ce texte. Dans le cas affirmatif, il est invité à construire un résumé du texte qui est alors évalué par le processus précédent. Dans l'autre cas ou si le résumé est de faible qualité, un autre texte est proposé. Le choix du texte suivant est crucial et fondé sur LSA. En effet, le texte suivant proposé à l'apprenant est celui qui est le plus proche sémantiquement de l'ensemble des résumés qu'il a pu produire.

Deux choses sont intéressantes dans cette utilisation de LSA. La première est le fait que les modèles de représentation des connaissances de l'apprenant et celui du but à atteindre sont les mêmes : LSA. L'état des connaissances de l'apprenant est alors l'ensemble des résumés qu'il a pu produire et le but réside dans les documents proposés en réponse à sa requête préalable. L'autre aspect intéressant est d'utiliser la réponse de l'apprenant sur sa capacité à résumer un texte pour en réalité savoir s'il l'a compris.

Nous avons essayé de présenter l'Analyse Sémantique Latente, certaines de ses possibilités et de ses utilisations en rapport avec le domaine éducatif. Nous voulions insister sur le fait qu'il est ici principalement utilisé pour représenter l'état de connaissance d'un individu et permet entre autres de comparer ses productions à un matériau qui fait autorité sur cette connaissance, grâce une représentation des relations entre termes. À notre connaissance, LSA ne semble pas employée dans des situations d'apprentissage qui sont le siège d'interactions distantes et informatiquement médiatisées entre apprenants, du moins pas à même escient.

3.2 Vers une construction sociale de connaissances grâce au dialogue et au partage dans un réseau

Nous voudrions présenter des travaux mettant en œuvre différentes techniques du TAL pour restituer certains indicateurs dans la collaboration. Leurs travaux s'appuient sur les transpositions des notions de dialogisme, de polyphonie et d'inter-animation de Bakhtin ainsi que sur l'hypothèse que discours et dialogue jouent un rôle prépondérant dans la construction et l'acquisition des connaisances. (Trausan-Matu, 2010b).

3.2.1 Définitions

Voici les définitions des notions relatives à Bakhtin traduites depuis (Trausan-Matu, 2010b) :

Dialogisme Un concept introduit par Mikhail Bahthin, qui considère que chaque création et activité langagière humaine est un dialogue, incluant non seulement les conversations mais aussi des textes ou même des réflexions.

Inter-animation Un phénomène spécifique à la polyphonie ou à des groupes de personnes collaborant dans lequel plusieurs voix entrent en dialogue et, à cause d'interactions caractérisées par le même ou le différent (centripète ou centrifuge), un thème est développé.

Polyphonie Une réalisation conjointe qui implique plusieurs individus qui construisent en collaborant une structure cohérente et durable à partir d'un thème donné, même si des dissonances délibérées et transitoires peuvent apparaitre. Afin d'atteindre une cohérence, différentes règles assurant l'harmonie se doivent d'être respectées.

3.2.2 Le système

Dans (Trausan-Matu, 2010a) et (Rebedea *et al.*, 2010), les auteurs présentent un système nommé PolyCAFe analysant les échanges entre des étudiants dans une optique de débat, et de synthèse, concernant un domaine bien défini, sur une plateforme informatique dédiée conservant les traces de ces discussions.

Ce système associe à une chaîne de traitement linguistique traditionnelle une ontologie représentant les concepts du domaine, ici les interfaces Homme-Machine. Afin d'éviter des ambiguïtés dues à la polysémie des langues naturelles et dans le but d'identifier les différents fils de discussion, une LSA est réalisée au préalable sur un corpus du domaine comparant les concepts évoqués dans deux messages au sein de l'espace sémantique construit.

Le but est ici d'identifier les dimensions longitudinale et transversale de la polyphonie mais aussi de rendre compte de l'inter-animation : l'entremêlement des propos des uns dans ceux des autres et d'identifier les références des uns aux autres. À cette fin, le système de discussion invite les participants à préciser à quel apprenant ils répondent. Cette information est utilisée pour identifier de premières interactions explicites. Des traitements linguistiques de plus haut niveau prennent place pour identifier les références implicites des uns aux autres. Ces traitements consistent notamment en une identification de répétitions, une résolution de la coréférence, un calcul de similarité grâce à LSA et la prise en compte de connecteurs logiques afin d'identifier des actes de langage et des paires adjacentes. Cette dernière identification peut permettre de détecter des comportements et les différents rôles qu'endossent les apprenants dans la discussion.

Ce point avait déjà été envisagé dans (George, 2004) mais évincé suite à des réserves concernant la faisabilité d'une automatisation de cette détection.

Capter les différentes interactions entre les apprenants permet de construire alors un réseau qui va représenter l'inter-animation amenant à une construction du savoir. Une analyse des réseaux sociaux identifie différents critères significatifs comme la centralité des graphes, les degrés, les participants faisant autorité (au sens du pagerank de Google) ou la cohésion, notamment avec le calcul de composantes fortement connexes et de cliques. Cette étape permet de retourner des indicateurs quant à la participation de chacun dans les différents fils de discussions ou la position plus ou moins centrale dans les débats. S'ajoutent à ces indicateurs des informations sur la lisibilité et la cohérence textuelle des propos.

Ces travaux sont très intéressants car ils mettent en perspective les techniques actuelles du TAL de bas niveau (analyse morphosyntaxique, LSA, ontologie) et de haut niveau (discours et dialogisme) mais aussi les techniques d'analyse de réseaux sociaux pour des situations collaboratif en ligne.

3.3 Bilan

Nous avons voulu montrer qu'en passant de l'individu au groupe, certaines méthodes restaient intéressantes même si leur utilisation n'était pas identique et que les problématiques concernant la construction de la connaissance se déplacent. Chez l'individu, c'est la connaissance construite qui nous intéresse alors que dans une co-construction, ce sont aussi les constructeurs et le chantier. De plus, une constante apparait : la construction du sens grâce, non pas aux éléments (termes ou individus) eux-mêmes, mais grâce aux relations qu'ils entretiennent entre eux.

4 Réel et Focalisation

Nous cherchons à explorer un champ de recherche pluridisciplinaire mariant EIAH et TAL. Cela nous impose de résoudre un conflit méthodologique. En effet, le champ des EIAH aborde l'ingénierie pour répondre à des besoins et permettre des usages par un média et un outillage informatique à des fins didactiques : il met en œuvre des analyses théoriques et généralistes, sans nécessairement de matériau d'observation, des modélisations pédagogiques et informatiques, des expérimentations et des évaluations par les utilisateurs des systèmes produits. Le TAL utilise des méthodologies plus empiriques mettant habituellement à profit l'observation d'un matériau représentatif d'une entrée à traiter, proposant des modèles informatiques et linguistiques éprouvés dans le cadre d'expérimentations sur des corpus d'évaluations. La vastitude apparente de notre problématique nous invite ainsi à amorcer nos recherches en précisant la tâche à traiter c'est-à-dire la situation ou les situations d'apprentissages, les protagonistes, les traces textuelles collectées et les indicateurs à produire, mais aussi à observer un matériau langagier réel à partir d'un corpus de traces.

Les situations d'apprentissage socio-constructivistes auxquelles nous nous intéressons et pour lesquelles nous devons fournir des indicateurs restent floues. Nos discussions ont mis en avant parmi ces situations celles d'apprentissages par projet ou de débat/résolutions de problèmes. Dans ces deux situations, les apprenants sont amenés à discuter, à faire des recherches personnelles, à partager les informations recueillies et leurs points de vue. Il s'agit aussi de critiquer le point

de vue d'autrui ou le sien, de négocier le sens de certaines informations, la validité de propos, d'exclure en accord des informations ou encore de discuter en parallèle de plusieurs notions. Ce type d'exercice attend généralement sa clôture par une étape de synthèse et de conclusion des débats. Cette étape est le siège de consensus ou de l'intégration de plusieurs opinions et a pour vocation de répondre à l'exercice. Cette réponse peut prendre plusieurs formes : clore les échanges ou être rédigée collectivement de manière déportée. Dans ce dernier cas, l'élaboration de la réponse se déroule soit en parallèle des échanges soit *a posteriori* et peut alors être révisée maintes et maintes fois. Des aspects comme la discipline abordée, la durée des exercices et le nombre d'apprenants participant au même exercice restent obscurs. Il nous faudra aussi décider si nous devons traiter oui ou non ces paramètres de manière générique, ce qui paraît ambitieux.

Nous n'avons pas non plus défini ou décrit la population qui constitue les différents utilisateurs, la discipline concernée, leur niveau dans leur discipline la langue dans laquelle ils discutent (maternelle ou seconde), leur niveau de langue. Or ces informations sont nécessaires pour établir un profil, un modèle de l'apprenant et pourraient s'avérer déterminantes dans une chaine de traitement automatique de la langue.

Les traces textuelles que nous allons traiter sont à la fois des productions qui répondent aux exercices auxquels prennent part les apprenants, et des messages échangés au travers d'un réseau de manière synchrone ou asynchrone. Ce type de messages présente des caractéristiques proches de l'oral et peut notamment être bruité. (Sidir *et al.*, 2006; Bouchet et Sansonnet, 2006).

Nous travaillons donc un matériau réel et bruité dans lesquels les usages et les normes ont plus leur place que des règles. Pour appréhender cette « réalité empirique » et capter certains invariants propres à ce genre et à ce type de discours, il est donc nécessaire de s'imprégner d'un échantillon représentatif et de s'atteler à une description. Notre tâche consiste à manipuler du sens, or, selon Rastier (Rastier, 2005), le sens obéit notamment à ces spécificités à causes des problèmes sémantiques que sont l'implicite et la polysémie, caractérisant la langue naturelle.

Ainsi, (Rastier, 2005; Williams, 2003; Bouchet et Sansonnet, 2006) nous incitent donc à construire un corpus d'étude pouvant nous aider à focaliser nos recherches sur le réel d'une situation. Même si ce n'est qu'une amorce, cela nous permettra peut-être par la suite d'atteindre une certaine généricité attendue.

Conclusion

Nous avons présenté la problématique que nous formulons, qui consiste à explorer les capacités du TAL pour l'extraction d'indicateurs sur la construction collective de connaissances au sein d'EIAH, et avons tenté de la justifier en présentant des opportunités et en montrant la couverture partielle des EIAH par le TAL et du TAL par les EIAH. Nous avons montré les possibilités qu'offre LSA pour représenter et évaluer les connaissances d'un individu, mais aussi un système récent mettant en œuvre différentes techniques du TAL pour fournir des indicateurs quantitatifs sur la construction collective de connaissances et qui représentent un point de départ et de repère pour nos travaux. Nous terminons en présentant les particularités liées à la pluridisciplinarité de notre problématique et des choix qu'il nous faudra peut-être faire assez tôt malgré un conflit méthodologique qui s'y oppose.

Nos perspectives concernent la consolidation de l'état de l'art et la justification de notre problé-

matique mais aussi des questions liées à la représentation de la construction et son évolution en prenant en compte ce qui est construit, qui construit quoi, et si le multiple construit du même ou du différent. Il nous faut aussi approfondir les moyens d'identifier des moments du discours et de calculer des différentiels entre l'état de la construction des connaissances entre deux instants ou en un instant et état cible. Les graphes (Zouaq *et al.*, 2000) et les cartes conceptuelles nous semblent un moyen envisageable (Berlanga *et al.*, 2009).

Références

ANTONIADIS, G. (2008). *Du TAL et de son apport aux systèmes d'apprentissage des langues : Contributions*. Habilitation à diriger des recherches en informatique et traitement automatique des langues, Université Stendhal - Grenoble 3.

ANTONIADIS, G., GRANGER, S., KRAIF, O., PONTON, C., MEDORI, J. et ZAMPA, V. (2009). Integrated Digital Language Learning. *In* BALACHEFF, N., LUDVIGSEN, S., JONG, T., LAZONDER, A. et BARNES, S., éditeurs : *Technology-Enhanced Learning*, chapitre 6, pages 89–103. Springer Netherlands, Dordrecht.

BERLANGA, A. J., KALZ, M., STOYANOV, S., van ROSMALEN, P., SMITHIES, A. et BRAIDMAN, I. (2009). Using language technologies to diagnose learner's conceptual development. *Advanced Learning Technologies, IEEE International Conference on*, 0:669–673.

BOUCHET, F. et SANSONNET, J. (2006). Étude d'un corpus de requêtes en langue naturelle pour des agents assistants. *In Actes du Deuxième Workshop sur les Agents Conversationnels Animés (WACA 2006)*, pages 95–104, Toulouse, France.

BURSTEIN, J. (2009). Opportunities for natural language processing research in education. *CICLing 09 Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, 5449:6–27.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. et HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

DESSUS, P. (1999). Apex, un système d'aide à la préparation d'examens. *Sciences et Techniques éducatives*, 6(2):409–415.

DESSUS, P. (2000). Construction de connaissances par exposition à un cours avec LSA. *In Cognito*, 18:27–34.

DESSUS, P. et LEMAIRE, B. (2002). *Using production to assess learning : An ILE that fosters self-regulated learning*, volume 2363, pages 772–781. Springer.

DUMAIS, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods*, 23(2):229–236. 10.3758/BF03203370.

FOLTZ, P. W., LAHAM, D. et LANDAUER, T. K. (1999). Automated essay scoring : Applications to education technology. *In Proceedings of EDMEDIA*, volume 1, pages 939–944. AACE.

GAUDIO, R. D. (2007). Supporting e-learning with automatic glossary extraction : Experiments with portuguese. *In RANLP 2007 workshop : Natural Language Processing and Knowledge Representation for eLearning Environments*.

GEORGE, S. (2004). Analyse automatique de conversations textuelles synchrones d'apprenants pour la détermination de comportements sociaux. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, Vol. 10:p. 165–193.

GUREVYCH, I., BERNHARD, D. et BURCHARDT, A. (2009). Educational natural language processing. Notes for ENLP tutorial held at AIED 2009 in Brighton.

HEARST, M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64.

HOU, H.-T. et WU, S.-Y. (2011). Analyzing the social knowledge construction behavioral patterns of an online synchronous collaborative discussion instructional activity using an instant messaging tool : A case study. *Computers & Education*, 57:1459–1468.

KARAMANIS, N. (2006). Generating multiple-choice test items from medical text : A pilot study. *In In Proceedings of INLG 2006*, pages 104–107.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

LEE, J. et SENEFF, S. (2007). Automatic generation of cloze items for prepositions. *In INTERSPEECH*, pages 2173–2176. ISCA.

LEMAIRE, B. et DESSUS, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24(3):305–320.

RASTIER, F. (2005). *Enjeux épistémologiques de la linguistique de corpus*, pages 31–45. Presses Universitaires de Rennes.

REBEDEA, T., DASCALU, M., TRAUSAN-MATU, S., BANICA, D., GARTNER, A., CHIRU, C. et MIHAILA, D. (2010). Overview and preliminary results of using polycafe for collaboration analysis and feedback generation. *In Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL : from innovation to learning and practice*, EC-TEL'10, pages 420–425, Berlin, Heidelberg. Springer-Verlag.

SIDIR, M., LUCAS, N. et GIGUET, E. (2006). De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, 13.

TEUFEL, S. (1999). *Argumentative Zoning : Information Extraction from Scientific Text*. Thèse de doctorat, University of Edinburgh, School of Cognitive Science.

TRAUSAN-MATU, S. (2010a). Automatic support for the analysis of online collaborative learning chat conversations. *In Proceedings of the Third international conference on Hybrid learning*, ICHL'10, pages 383–394, Berlin, Heidelberg. Springer-Verlag.

TRAUSAN-MATU, S. (2010b). The polyphonic model of hybrid and collaborative learning. *In* WANG, F. L., FONG, J. et KWAN, R., éditeurs : *Handbook of Research on Hybrid Learning Models : Advanced Tools, Technologies, and Applications*, pages 466–486. Information Science Publishing, New York.

WILLIAMS, G. (2003). Texte et Corpus. *In Actes des Troisièmes Journées de la Linguistique de Corpus*, pages 1–307.

ZAMPA, V. (2005). Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité. Volume 8.

ZOUAQ, A., FRASSON, C. et ROUANE, K. (2000). The explanation agent. *In* GAUTHIER, G., FRASSON, C. et VANLEHN, K., éditeurs : *Intelligent Tutoring Systems, 5th International Conference, ITS 2000, Montréal, Canada, June 19-23, 2000, Proceedings*, volume 1839 de *Lecture Notes in Computer Science*, pages 554–563. Springer.

Index