Extraction de patrons sémantiques appliquée à la classification d'Entités Nommées

Ismaïl El Maarouf (1,2) Jeanne Villaneau (2) Sophie Rosset (3)

(1) HCTI UBS-UEB, Centre de Recherche Christiaan Huygens, 56321 Lorient
(2) Valoria UBS-UEB, Rue Yves Mainguy, Campus de Tohannic 56017 Vannes cedex
(3) LIMSI-CNRS, F-91403 Orsay Cedex
ismail.el-maarouf@univ-ubs.fr, jeanne.villaneau@univ-ubs.fr, sophie.rosset@limsi.fr

Résumé **La variabilité des corpus constitue un problème majeur pour les systèmes de reconnaissance d'entités nommées. L'une des pistes possibles pour y remédier est l'utilisation d'approches linguistiques pour les adapter à de nouveaux contextes : la construction de patrons sémantiques peut permettre de désambiguïser les entités nommées en structurant leur environnement syntaxico-sémantique. Cet article présente une première réalisation sur un corpus de presse d'un système de correction. Après une étape de segmentation sur des critères discursifs de surface, le système extrait et pondère les patrons liés à une classe d'entité nommée fournie par un analyseur. Malgré des modèles encore relativement élémentaires, les résultats obtenus sont encourageants et montrent la nécessité d'un traitement plus approfondi de la classe Organisation.**

Abstract **Corpus variation is a major problem for named entity recognition systems. One possible direction to tackle this problem involves using linguistic approaches to adapt them to unseen contexts : building semantic patterns may help for their disambiguation by structuring their syntactic and semantic environment. This article presents a preliminary implementation on a press corpus of a correction system. After a segmentation step based on surface discourse clues, the system extracts and weights the patterns linked to a named entity class provided by an analyzer. Despite relatively elementary models, the results obtained are promising and point on the necessary treatment of the Organisation class.**

Mots-clés : entités nommées, patrons sémantiques, segmentation discursive de surface
Keywords: named entities, semantic patterns, surface discourse segmentation

1 Introduction

Ritel est un Système de Question-Réponse interactif à domaine ouvert (Rosset et al., 2008), permettant à un utilisateur de dialoguer et d'obtenir des réponses à ses questions. Parmi les questions auxquelles doit pouvoir répondre un tel système, certaines (factuelles, définitions) correspondent à une demande d'information sur une Entité Nommée (EN), une valeur ou une date. L'analyse détaillée des composants d'un SQR (Moldovan et al., 2003) montre que la prédiction du type de réponse dans les documents indexés est l'une des causes d'erreurs majeure, en partie due à la Reconnaissances de ces Entités Nommées (REN).

Si la REN semble être une tâche bien maitrisée lorsque l'on se réfère aux résultats des campagnes d'évaluation classiques (Grishman et al., 1995), on connaît moins leurs performances sur des corpus différents de ceux pour lesquels ils ont été développés (Grishman, 2010). La robustesse des systèmes peut être mise à l'épreuve par le degré de granularité de la typologie d'EN (Galliano et al., 2009 ; Markert et al., 2007), la qualité de retranscription du corpus (Galliano et al., 2009) et l'hétérogénéité des corpus (Mota et al. 2008). Les dégradations de performances constatées dans ces trois cas justifie la conception de systèmes de correction qui assurent leur robustesse face à des textes et à des EN inconnus.

C'est dans la perspective d'adaptation d'un analyseur linguistique intégrant la détection d'EN (Ritel-nca) utilisé par le SQR Ritel, que s'inscrit notre recherche. Nous présentons un système destiné à corriger les résultats de l'analyseur, à partir de patrons sémantiques extraits de corpus écrits. Après avoir décrit notre approche et les traitements effectués (section 2), les modèles de construction de patron sont présentés en section 3 puis évalués en section 4. Les premiers résultats obtenus ouvrent la voie à des perspectives de recherche explicitées section 5.

2 Approche et traitements

2.1 Choix de l'approche

Les systèmes de REN sont principalement divisés en deux catégories : d'une part, les systèmes à base de règles, correspondant à des automates qui utilisent des listes d'entités nommées (« gazetteers »), des mots déclencheurs et des indices de surface, et d'autre part, les modèles probabilistes utilisant différents niveaux de représentation (forme, lemme, catégorie morphosyntaxique) entrainés sur des corpus annotés (pour un état de l'art, voir Nadeau et al., 2007). Notre système s'appuie sur un corpus annoté automatiquement par un système de REN. Les EN sur lesquelles nous travaillons correspondent à la triade Personne-Lieu-Organisation car elles cristallisent de nombreux problèmes : la possibilité pour une forme de recouvrir les trois catégories rend une analyse fine du contexte indispensable (voir également 4.1).

L'approche développée dans cet article s'inscrit dans une perspective d'enrichissement linguistique de systèmes à base de règles, en créant des grammaires de patrons sémantiques sur le modèle de celles qui sont employées en Extraction d'Information (Ward et al., 1992). Elle s'inspire de la linguistique de corpus britannique (Sinclair, 1991), dont le but est de décrire les usages des mots en contexte à partir de l'exploration de grands corpus. Le sens d'un mot est défini en fonction des patrons majeurs dans lesquels il est employé. La notion de patron peut recouvrir différentes réalités, des collocations, associations significatives entre lexies (Sinclair, 1991), jusqu'à l'identification de cadres complexes, proches des structures prédicatives (Hanks, 2008).

Les associations sont automatiquement calculées sur de grands corpus et permettent de créer des réseaux sémantiques pour chaque mot. La méthode générale consiste à sélectionner les mots apparaissant de manière significative dans une fenêtre de taille arbitraire autour du mot-clé. Dans notre système, les unités du réseau correspondent aux informations syntaxiques et sémantiques contenues dans des chunks (décrits en 2.2). La fenêtre, quant à elle, peut être de taille variable, car elle correspond à la notion de segment, défini à partir d'indices discursifs de surface tels que les marques de ponctuation et les formes en « qu- » (cf. 2.3).

2.2 Chunking grammatical

Ritel utilise les résultats d'un analyseur linguistique à base de règles, Ritel-nca, dont les sorties sont présentées en structure arborescente (figure 1). Ce système a fait l'objet d'un développement particulièrement approfondi : il permet l'accès à des lexiques catégorisant plus d'un million de mots, dont une grande partie de noms propres, et près de 2000 règles sont actuellement implémentées. La taxonomie utilisée et continuellement augmentée, comprend plus de 300 types, dont les EN classiques (*Personne*, *Organisation* et *Lieu*), affinés et structurés en sous-types et en composants. La f-mesure associée à la classification d'entités classiques est de 0,8 sur l'écrit et à hauteur de l'état de l'art pour les corpus oraux (Rosset et al., 2008).

Un système comme Ritel-nca facilite l'analyse linguistique : la détection des entités classiques, comme les dates, permet de regrouper de nombreuses variantes, faisant ainsi émerger des associations nouvelles et sémantiquement pertinentes. Les mots grammaticaux (déterminants, prépositions) sont cependant rarement rattachés, ce qui nuit à la construction de patrons. Pour réduire les phrases à des groupes plus homogènes, nous avons intégré les entités dans des chunks grammaticaux (dont la version initiale est décrite dans Villaneau et al. 2007).

Les chunks grammaticaux regroupent les nœuds de l'arbre de la phrase à partir de règles exploitant des indices de formes, de type d'entité et de position. Cinq types de chunks sont définis parmi les groupes nominaux (GN, GNP) et les groupes verbaux (GV, GVP, GVADJ). Les chunks sont associés à la catégorie majeure : un verbe dans le cas des groupes verbaux, un lieu dans le cas d'un GNP regroupant les entités [préposition déterminant lieu]. La figure 1 illustre la représentation arborescente de l'exemple (1) après la passe de chunking.

(1) Patricia Highsmith est morte le 4 février 1995 dans un hôpital de Locarno

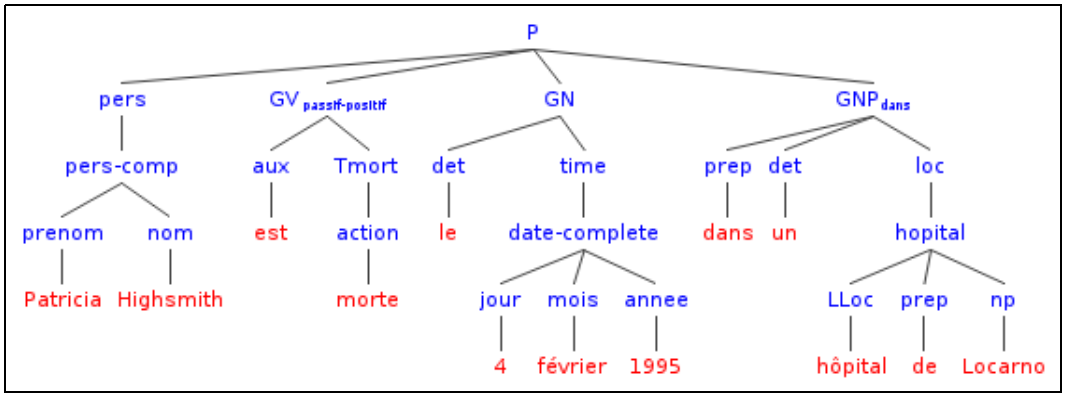


Figure 1- Représentation arborescente après chunking de l'exemple 1.

Dans cet exemple, le nombre de nœuds de l'arbre de la phrase est divisé par deux sans perte d'information : la catégorie et la forme de la tête sont conservées et il est toujours possible de parcourir les fils d'un nœud pour obtenir des informations plus détaillées. Les nœuds de chunks grammaticaux possèdent en sus des attributs concernant la voix et le mode du verbe (pronominal, passif, actif) permettant de distinguer rapidement les différentes réalisations verbales.

Nous avons manuellement cherché à extraire des relations sémantiques en projetant des cadres sémantiques du type de FrameNet (Fillmore et al., 2003) à partir de grammaires régulières de chunks. Si les résultats obtenus sont précis, l'analyse se heurte à des limites qui compliquent la création de règles : rigidité due à l'ordre de détection des éléments et sensibilité à la variation de surface, due principalement à la présence de ponctuation. Plutôt que de multiplier les règles, nous avons intégré une étape de segmentation de surface qui permet d'isoler des séquences de chunks.

2.3 Segmentation de surface

La segmentation consiste à isoler des séquences de chunks en fonction d'indices de frontière surfaciques comme la ponctuation : elle est principalement applicable à des textes écrits, quoique les travaux récents en insertion automatique de ponctuation (Favre et al., 2008) ouvrent des perspectives d'application à l'oral. Les indices de ponctuation jouent un rôle structurant dans la présentation de l'information. Leur prise en compte peut permettre la délimitation de structures appositives, comme en (2), citationnelles en (3) et parenthétiques en (4).

(2) le secrétaire d' Etat aux PME , Renaud Dutreil , s' était ainsi vu convoquer par le directeur du cabinet du premier ministre , Michel Boyon.

(3) "le navire doit être capable de prendre un mauvais coup de chien sans risque", explique Guy Ribadeau-Dumas, l' architecte naval maître d' œuvre du chantier .

(4) Eric Tanguy (né en 1968) passait alors pour un nouveau Dusapin.

Ces structures imposent des contraintes sur leur environnement : dans l'exemple (3), la citation entre guillemets précède un verbe dont le sujet est inversé. On observe également en (4) que les parenthétiques isolent un groupe d'information semi-autonome qui contient une relation sémantique (date de naissance).

Les segments que nous constituons se définissent par leurs frontières gauche et droite et leur taille. Pour première expérience, nous nous limitons à une typologie simple de frontières en distinguant les frontières fortes (point, point d'exclamation, etc.) des frontières faibles (virgules, parenthèses, etc.). Seules les frontières faibles sont franchies pour identifier des relations sémantiques (virgule dans l'exemple 2). Nous avons également considéré les connecteurs et conjonctions comme des frontières faibles pour obtenir des segments proches de propositions. L'algorithme de segmentation applique d'abord les règles associées à la détection de frontières fortes, puis, dans une seconde passe et au sein des unités déjà segmentées, celles associées aux frontières faibles (automates adaptés aux sorties du chunking). Le tableau 1 résume les fréquences des segments obtenus avec cette méthode de segmentation sur un corpus de presse (4,5 millions de mots).

Taille de Segment	Fréquence	Proportion
1	207260	0.29
2	160177	0.22
3	110760	0.15
4	78605	0.11
5	53846	0.07
6	35657	0.05
7	23103	0.03
8	14686	0.02
9	9282	0.01
10	5783	<0.01
>10	*29951*	*0.04*

Tableau 1 – Fréquence des segments en fonction de leur taille (exprimée en nombre de chunks).

En focalisant sur les relations contenues dans les segments de petite taille, on peut ainsi traiter la majorité des segments, sachant que 30% des segments (de taille 1) ne renferment pas de relation. La détection de relations sémantiques peut s'effectuer de deux manières : en analysant les associations entre chunks au sein de segments de taille supérieure à 1, et en analysant les relations entre chunks appartenant à des segments différents. Les segments de taille 1 sont donc réservés à l'analyse entre segments. L'exemple (5) montre le type de relations que l'on peut détecter au sein des segments dits « pluriels » (car contenant plus d'un chunk) : dans l'exemple, ces derniers figurent en rouge, les segments simples, en violet.

(5) A. V. Shinde, né à Goa, en Inde, décédé en 2003 à New York (à l' âge de 86 ans), avait parcouru le monde en quête des plus belles pierres pour le joaillier Harry Winston.

Nous nous concentrons dans cet article sur les relations existant entre chunks au sein des segments pluriels. Il s'agit dans l'exemple (5), d'analyser les relations entre « né » et « à Goa », ou encore entre « décédé », « en 2003 » et « à New York ». Plus précisément, nous utilisons ces chunks comme indices de désambiguïsation des EN figurant dans les mêmes segments. Le système permettant d'extraire les patrons d'association [chunk-EN] est décrit en section suivante.

3 Système et modèles

3.1 Un système multi-niveaux

Le système analyse l'arbre obtenu après la passe de chunking et y ajoute les segments. Quatre classes de règles dédiées à chaque niveau de représentation sont définies : *Forme*, *Entité*, *Chunk Grammatical* et *Segment*. Chaque niveau de représentation possède son propre lexique de contraintes (attribut-valeur), permettant d'identifier un élément ou un segment. Par exemple, un groupe prépositionnel en « à » sera défini par sa classe (Chunk Grammatical), son sous-type (Groupe Nominal Prépositionnel) et la forme de la préposition. Les règles et les lexiques sont externalisés afin de pouvoir appliquer des patrons définis manuellement ou extraits automatiquement.

Les patrons extraits au sein des segments peuvent s'appuyer sur les chunks, les entités ou les formes. À titre d'illustration, la figure 3 décline les caractéristiques internes de chaque segment de l'exemple (6) en fonction du niveau de représentation.

(6) il y a près de cinquante ans déjà , Jacques Monod rappelait au colloque de Caen que 50 pourcent du chiffre d' affaires de la société américaine Du Pont de Nemours provenait de la commercialisation de produits inconnus dix ans plus tôt .

Segment 1

Chunk	-	-	-
Entité	<_pres>	<_annee_dur>	<_adv>
Forme	il y a	50 ans	déjà

Segment 2

Chunk	-	-	GNP_au	GNP_de
Entité	<_pers>	<_action>	<_subs>	<_loc>
Forme	Jacques Monod	rappelait	colloque	Caen

Segment 3

Chunk	GN	GNP_de	-	GNP_de	-	GNP_de	GNP_de	-
Entité	<_subsnn>	<_org>	<_pers>	<_loc>	<_action>	<_subs>	<_subs>	<_time>
Forme	chiffre d'affaires	société américaine	Du Pont	Nemours	provenait	commercialisation	produits	dix ans plus tôt

Figure 3 – Tableaux des niveaux de représentation des segments de l'exemple (6).

Cette représentation nivelée de la phrase permet de concevoir des modèles qui combinent les informations de plusieurs niveaux. Trois modèles sont évalués : la combinaison des niveaux Chunk et Forme (modèle CF), des niveaux Chunk et Entité (modèle CE) et un modèle mixte (modèle CEM) qui combine les niveaux Chunk et Entité, en substituant les entités « substantif », « action » et « adjectif » par les formes correspondantes, les verbes étant lemmatisés. L'existence de ce dernier modèle est motivée par l'hypothèse que ces classes contiennent régulièrement des informations sémantiques pertinentes qui seraient autrement masquées. Le tableau 2 fait figurer les éléments extraits dans le segment 2 en fonction de chaque modèle.

	E1	*E2*	*E3*	*E4*
CF	**Jacques Monod**	**rappelait**	**GNP_au/colloque**	**GNP_de/Caen**
CE	**pers**	**action**	**GNP_au/subs**	**GNP_de/loc**
CEM	**pers**	**rappeler**	**GNP_au_colloque**	**GNP_de/loc**

Tableau 2 - Exemples de patrons extraits du segment 2 en fonction des modèles.

Alors que le modèle CEM cherche à optimiser les informations détenues par chaque élément, le modèle CE est le plus générique. Quant au modèle CF, il peut être plus précis en cas d'erreurs d'analyse des entités.

3.2 Méthode d'extraction et score

Le corpus dont nous disposons correspond à une année du Journal LeMonde de l'année 2003. Il est divisé en deux parties, un corpus de développement, à partir duquel nous avons extrait les patrons (17 millions de mots), un quart (5,5 millions de mots, 10 000 articles) ayant été conservé dans la perspective de l'annoter manuellement (cf. 4.1). Pour chaque modèle, nous avons sélectionné les segments contenant une des entités classiques (« personne », « organisation » ou « lieu ») fournies par l'analyseur Ritel-nca, et puisque notre étude porte sur les patrons trouvés à l'intérieur d'un segment, exclu les segments de taille 1.

Le système évalué s'appuie sur les patrons intra-segment observés dans le corpus de développement. Un patron correspond à un chunk identifié dans un segment contenant une EN, modélisé selon un niveau de représentation. Par exemple, le patron « GNP_de/_loc » du modèle CEM apparaît 12829 fois en corpus, 6411 fois en coocurrence avec un Lieu, 2604 fois avec une Organisation et 3814 avec une Personne. À partir de ces données, nous calculons deux scores d'association d'un patron pour chaque classe : la probabilité de cooccurrence entre un chunk et une classe d'EN donnée (PROBA) et l'information mutuelle (IM) :

$$PROBA(EN \mid Patron) = \frac{P(EN, Patron)}{P(Patron)}$$

$$IM(EN,\ Patron) = P(X=EN, Y=Patron) \times \log \frac{P(X=EN, Y=Patron)}{P(X=EN) \times P(Y=Patron)}$$

Ces scores nous permettent de prédire la classe d'EN la plus probable vis-à-vis d'un patron donné. Pour l'évaluation, nous avons calculé trois scores globaux pour prendre en compte l'ensemble des patrons contenus dans le segment : la moyenne des scores des patrons (Mean) pour tenir compte du nombre de patrons dans le segment, le score du meilleur patron (Max) et le produit des scores, pour atténuer l'importance de patrons fréquents et communs aux différentes classes. Pour exemple, le score PP (Produit de Probabilités) d'un segment pour la classe Personne, se calcule à partir du produit des probabilités de ses patrons :

$$PP(Personne) = \prod_{1}^{n} PROBA(Personne \mid Patron_i)$$

Six scores différents ont ainsi été expérimentés pour chaque type de modèle de représentation vis-à-vis de chacune des trois classes, Personne, Lieu et Organisation.

4 Évaluation

L'évaluation présentée dans cette section compare les performances des modèles CF, CE et CEM, combinés avec les scores PROBA et IM présentés précédemment. Leur performance de classification est comparée à celle du système Ritel-nca qui sert de référence.

4.1 Un corpus d'évaluation établi sur des critères contextuels

200 articles de presse ont été annotés pour obtenir plus de mille instances d'entités de chaque classe (plus exactement 1426 organisations, 1004 lieux et 1377 personnes). Les segments de taille 1 étaient pré-détectés et exclus de l'annotation. Les conventions d'annotation ont réduit la tâche de détection des EN au nom propre (avec ou sans majuscule) lorsque c'était possible, en excluant les titres, fonctions,

déterminants. Lorsque certains éléments (parfois même des entités nommées dans des cas d'imbrication) pouvaient être considérés comme constitutifs du nom d'une EN, ils ont cependant été inclus (exemple 7). En revanche, seule l'EN était prise en compte lorsque sa dénomination ne dépendait pas d'éléments englobants. Ces conventions distinguent ainsi les cas (7) et (8).

(7) l'<org> université de Poitiers </org>

(8) le maire de <loc> Poitiers </loc>

La difficulté majeure consiste à identifier des critères fiables pour résoudre les cas où le type d'une EN diffère de son rôle en contexte. Les conventions d'annotation privilégient dans ces cas l'interprétation contextuelle. Deux types de divergences ont été rencontrées : lorsque cette divergence était explicitée par un déclencheur immédiatement apposé (9) et lorsqu'elle était due à une interprétation globale de la phrase voire du contexte de l'article (10)

(9) c' est la mesure phare de la loi Perben du 9 mars sur la criminalité

(10) l' Italie s' oppose à une réforme du Conseil de sécurité de l' ONU

En (9), l'EN « Perben » a été exclue de l'annotation car elle relève du type « Loi », bien qu'elle soit nommée d'après son fondateur. L'exemple (10) est généralement décrit comme un cas de métonymie (Markert et al., 2007), pour rendre compte de la relation existant entre un lieu (« Italie ») et des individus, l'interprétation étant due au verbe avec lequel elle est employée. Cet exemple ne désigne pas une personne comme les conventions d'annotation de métonymie de la campagne Semeval7 semble l'indiquer à travers la catégorie « Loc-for-People » (Markert et al., 2007) : il s'agit d'une organisation politique, dans ce cas très probablement le gouvernement. D'autres types d'organisations répondent à ce phénomène, comme les équipes de sport (11).

(11) dans les autres rencontres disputées mercredi soir, <org>Auxerre</org> s' est imposé à <loc>Rennes</loc>

Les conventions considèrent ainsi que l'EN « Italie » peut désigner un lieu ou une organisation, comme l'EN « Florence », une personne (12) ou un lieu (13).

(12) ce n' est pas le moindre des mérites de l' essai d' Anton Brender et Florence Pisani

(13) une forte pluie commença à tomber sur la Toscane et Florence

Les lieux ont donc été annotés comme tels lorsque l'interprétation en contexte le justifiait (localisation, destination, origine, etc.), comme en (13).

4.2 Détection et classification brute des EN

Les résultats de l'évaluation du système confirment ce qui a été dit précédemment à propos de l'impact du corpus de développement ainsi que des conventions d'annotations adoptées. 980 des 3807 EN annotées n'ont pas été détectées par le système Ritel-nca, soit 25%, parmi lesquelles 515 sont étiquetées comme noms propres non catégorisés. Les « erreurs » de détection affectent principalement la catégorie *Organisation* et s'expliquent pour les raisons suivantes : EN non retenues, mots inconnus, problèmes de normalisation du texte (suppression de majuscules, encodage), etc.

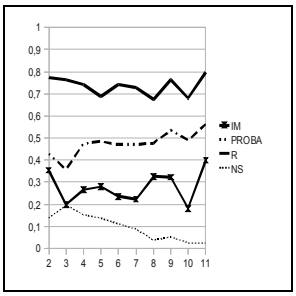
Classe	Correct	Faux Positif	Raté	PRECISION	RAPPEL	FMESURE
PERSONNE	1087	333	290	0,77	0,79	0,78
LIEU	686	508	318	0,57	0,68	0,62
ORGANISATION	523	360	903	0,59	0,37	0,45

Tableau 3 – Rappel, Précision et F-mesure de classification du système Ritel-nca.

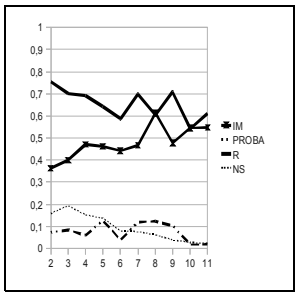
Étant donné que les patrons générés classent les EN à partir des entités fournies par le système Ritel-nca, l'évaluation a uniquement porté sur les EN détectées. Les résultats de ce dernier ont ainsi été recalculés et figurent dans le tableau 4 (modèle R). Le nombre de segments total s'élève à 1712, réduisant le nombre de segments contenant au moins une personne détectée à 943 (les lieux à 818 et les organisations à 659), 72% d'entre eux ne contenant qu'une seule entité.

4.3 Résultats

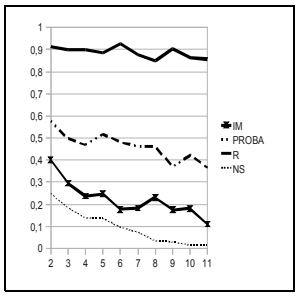
Les diagrammes 1 à 3 présentent les f-mesures des modèles en fonction de la taille des segments pour chaque classe d'EN ; le nombre de segments par taille figure également sur les diagrammes (NS), ainsi que les résultats de Ritel-nca (R), à titre comparatif. Par degré d'importance, le score d'association (IM, PROBA) est la variable qui influence le plus les résultats, suivi par le niveau de représentation (CE, CF, CEM). Quand au calcul global du score (MAX, PROD, MEAN), il n'a qu'une faible influence : le choix du meilleur score d'association (MAX) équivaut globalement à calculer la moyenne ou le produit des scores de tous les patrons. Les diagrammes font uniquement figurer les moyennes des scores en fonction de la mesure d'association (PROBA, IM).



Lieu



Organisation



Personne

Diagramme 1 à 3 : F-mesure pour les Lieux, les Organisations et les Personnes.

Comme on peut l'observer, ces modèles ne rivalisent pas avec le modèle de référence (R). Les meilleurs modèles atteignent 0,62 de F-mesure sur les Personnes, 0,55 pour les Lieux et 0,45 sur les Organisations. On peut retenir globalement que le score PROBA est plus approprié pour la classification des Lieux et des Personnes, alors que l'IM semble plus performante sur les Organisations. L'augmentation de la taille du segment semble avoir un impact négatif sur la classification des Personnes, mais elle est liée à une amélioration des modèles PROBA pour les Lieux et des modèles IM pour les Organisations.

Ces résultats nous permettent de connaître le comportement global des modèles mais ne nous renseignent pas directement sur leur utilité dans le cadre de la correction. Pour expérimenter la tâche de correction, motivés par le fait que la mesure de score global avait une influence minime sur les résultats, nous avons sélectionné tous les modèles MAX, qui, pour chaque EN, nous permet d'extraire un patron (un chunk du segment). Par exemple, le modèle CEM_IM_MAX a classé correctement 17 instances de personnes grâce au patron "expliquer", comme dans l'exemple (14) :

(14) " mon rôle est de bousculer la perception que les gens ont de Burberry ", *explique* Christopher Bailey

L'exemple (15) est un cas d'erreur que ce patron permet de corriger : « Maud » est classé en Lieu par le système de référence.

(15) une sorte de tri sélectif qui " élimine les cellules mortes et rend la peau douce et satinée ", explique *Maud*

En nous basant sur les résultats de l'évaluation, on peut alors assigner un taux de réussite à chaque patron. En ne sélectionnant que les patrons dont le score est sans appel (100%), il est alors possible de corriger les erreurs du modèle de référence, et, ce faisant, de juger de la pertinence linguistique des patrons correcteurs.

Les résultats présentés dans le tableau 4 indiquent les performances obtenues lorsque le système de référence détecte correctement une EN (R) et lorsque les patrons de correction sont appliqués. Les résultats sont organisés en fonction de l'ajout de patrons issus d'un des trois niveaux de représentation, du score d'association, ou tous modèles réunis. Lorsqu'aucun patron n'est identifié pour une instance d'EN donnée, le choix se porte sur la catégorie choisie par le système de référence (R).

Catégorie	Modèle	Précision	Rappel	F-mesure	# Corrigés
LIEU	R+TOUS	0,767	0,945	0,847	91
	R+CF	0,754	0,930	0,833	78
	R+PROBA	0,726	0,939	0,819	86
	R+CE	0,727	0,910	0,808	61
	R+IM	0,739	0,886	0,806	41
	R+CEM	0,724	0,899	0,802	52
	R	0,670	0,836	0,744	*NA*
ORGANISATION	R+TOUS	0,952	0,686	0,797	121
	R+CF	0,941	0,672	0,784	107
	R+IM	0,917	0,666	0,772	101
	R+CE	0,925	0,624	0,745	61
	R+CEM	0,918	0,623	0,742	59
	R+PROBA	0,940	0,606	0,737	45
	R	0,866	0,561	0,681	*NA*
PERSONNE	R+TOUS	0,924	0,978	0,950	31
	R+CF	0,916	0,975	0,944	27
	R+PROBA	0,909	0,977	0,942	30
	R+CE	0,898	0,970	0,932	21
	R+IM	0,897	0,967	0,930	18
	R+CEM	0,894	0,969	0,930	20
	R	0,861	0,951	0,904	*NA*

Tableau 4 - Potentiel de correction du système de référence.

Globalement, la prise en compte de tous les patrons de correction permet d'améliorer les f-mesures de 5% pour les Personnes, et de 10% pour les Lieux et les Organisations. Le niveau de représentation qui permet de corriger le plus d'instances est le niveau CF, en partie du fait qu'il génère un plus grand nombre de patrons, multipliant ainsi les possibilités de désambiguïsation. Le taux de correction par modèle ne dépend pourtant pas uniquement du nombre de patrons extraits : les niveaux CE et CEM ont un taux de correction relativement équivalent alors que le niveau CEM génère un plus grand nombre de patron. Ceci s'explique simplement par le fait qu'une large part des patrons corrects de ce dernier vient confirmer le modèle de référence. La combinaison de tous les modèles permet de corriger 66% d'erreurs pour les Lieux, 55% pour les Personnes et 28% pour les Organisations.

Ces résultats semblent corroborer le lien qui peut exister entre la performance d'un modèle et sa capacité de correction : les modèles ayant permis de corriger un grand nombre d'organisations sont basés sur le score IM, alors que le score PROBA contribue à générer plus de patrons de correction pour les lieux et les personnes. Nous avions constaté des tendances similaires sur les diagrammes 1 à 3.

Travaux similaires

Les systèmes obtenant les meilleurs performances sur la REN en français comme dans d'autres langues s'appuient généralement sur une classification supervisée qui nécessite l'établissement d'un corpus d'entrainement annoté manuellement. Notre système est entrainé sur un corpus automatiquement annoté par un système de REN, qui, par conséquent, comporte nécessairement des erreurs. Dans ce cadre, Petasis et al. (2001) ont initié un travail proche de nos objectifs : leur système est basé sur un corpus

automatiquement annoté par un premier système et les points de désaccord sont considérés comme des indices de défaillance. Leur système ne leur permet cependant pas d'extraire des patrons pour envisager une correction automatique : les erreurs sont manuellement corrigées par un expert. Plus généralement, l'inconvénient des systèmes d'apprentissage automatique est leur manque de transparence sur le lien entre les indices contextuels et la décision de classification. Les travaux rapportent au mieux l'impact de classes d'indices (« feature sets » en anglais) sur les performances : capitalisation, taille de fenêtre, prise en compte d'information syntaxique ou de ressources externes, etc. Notre méthode permet de juger directement de la pertinence d'un patron sur lequel s'appuie la décision de classification et d'en induire des règles de correction.

Conclusion et Perspectives

Cet article présente un système de correction d'EN à partir de patrons sémantiques. L'extraction s'appuie sur l'annotation de l'analyseur linguistique Ritel-nca, une phase de chunking et une segmentation de surface. Plusieurs modèles de patrons combinant différentes dimensions sont évalués : mesure d'association, score global et niveau de représentation. L'évaluation de ce système montre que la mesure d'association a une forte influence sur les performances, même si ces dernières sont en-deça de celles du système de référence. Nous mesurons le potentiel de correction du système de référence par ces modèles et obtenons des améliorations de 10% en F-mesure pour les Organisations et les Lieux et de 5% pour les Personnes. Ces améliorations nous encouragent à tester cette approche sur d'autres classes d'EN.

Les modèles employés dans l'évaluation sont relativement élémentaires : dans les travaux à venir, nous évaluerons l'apport de patrons conçus à partir des probabilités conjointes des éléments d'un segment, en commençant par exemple par la prise en compte du chunk contenant l'EN. Des modèles qui prennent en compte l'ordre (en établissant des contraintes de position droite ou gauche par exemple) méritent également d'être testés. Ces pistes seront évaluées dans le cadre d'analyses intra-segment et inter-segment telles que décrites dans cet article. Le problème majeur de notre approche que nous ne pouvons qu'évoquer ici, réside dans la sélection des patrons de correction parmi la totalité des patrons générés par chaque modèle. L'intervention humaine semble indispensable pour permettre d'y remédier mais l'utilisation de méthodes de filtrage automatique n'est pas exclue.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme QUAERO (financement OSEO).

Références

EL MAAROUF I., VILLANEAU J., SAID F., DUHAUT D. (2009). Comparing Child and Adult Language: Exploring Semantic constraints. Actes de *WOCCI ICMI-MLMI 2009*.

EL MAAROUF I. (2009). Natural Ontologies at Work : Investigating Fairy Tales. Actes de *Corpus Linguistics Conference 2009.*

ERHMANN M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de Doctorat en sciences du langage, Université Paris 7.

Favre B., Grishman R., Hillard D., Ji H., Hakkani-Tür D. & Ostendorf M. (2008). Punctuating speech for information extraction. Actes de *ICASSP 2008* : 5013-5016.

FILLMORE, C. J., JOHNSON C.R., PETRUCK M.R.L. (2003). Background to Framenet. *International Journal of Lexicography,* (16.3) : 235-250.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Thèse de doctorat en informatique, Université Paris-Sud 11, Orsay.

GALIBERT O., QUINTARD L., ROSSET S., ZWEIGENBAUM P., NÉDELLEC C., AUBIN S., GILLARD L., RAYSZ J.-P., POIS D., TANNIER X., DELÉGER L., LAURENT D. (2010). Named and specific entity detection in varied data: The Quaero Named Entity baseline evaluation. Actes de *LREC'10.*

GALLIANO S., GRAVIER G., CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. Actes de *INTERSPEECH-2009*. 2583-2586.

GRISHMAN R. & SUNDHEIM, B. (1995). Design of the MUC-6 evaluation. Actes de *MUC6*.

GRISHMAN R. (2010). The Impact of Task and Corpus on Event Extraction Systems. Actes de *LREC'10*.

HANKS P. **(**2008). Lexical Patterns: From Hornby to Hunston and Beyond. *Actes d'Euralex* 2008.

MARKERT K., NISSIM M. (2007) SemEval-2007 task 08: metonymy resolution at SemEval-2007. Actes de the 4th International Workshop on Semantic Evaluations : 36-41.

MOLDOVAN D., PASCA M., HARABAGIU S., SURDEANU M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *ACM Transactions in Information Systems*.

MOTA C., GRISHMAN R. (2008). Is this NE tagger getting old. Actes de *LREC'2008*.

NADEAU D., SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1) : 3–26.

PETASIS G., VICHOT F., WOLINSKI F., PALIOURAS G., KARKALETSIS V., SPYROPOULOS. C.D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. Actes de *ACL-EACL 2001* .

ROSSET S., GALIBERT O., BERNARD G., BILINSKI E., ADDA G. (2008). The LIMSI participation to the QAst track. Actes de *Working Notes of CLEF 2008 Workshop*.

SEKINE S., NOBATA C. (2004). Definition, Dictionary and Tagger for Extended Named Entities. Actes de *LREC 04*.

SINCLAIR J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.

VILLANEAU J., ROSSET S., GALIBERT O. (2007). Semantic Relations for an Oral and Interactive Question-Answering System. Actes de *SRSL7*.

WARD W., ISSAR S., HUANG X., HON H., HWANG M., YOUNG S., MATESSA M., LIU F., STERN R. (1992). Speech Understanding In Open Tasks. Actes de *the Fifth DARPA Workshop on Speech and Natural Language.*