Détection de coréférences de bout en bout en français

Elisabeth Godbert Benoit Favre
Aix-Marseille Université, LIF-CNRS UMR 7279, 163 avenue de Luminy, 13288 Marseille Cedex 9
Benoit Favre@lif.univ-mrs.fr, Elisabeth.Godbert@lif.univ-mrs.fr

RÉSUMÉ

Notre objectif est l'élaboration d'un système de détection automatique de relations de coréférence le plus général possible, pour le traitement des anaphores pronominales et les coréférences directes. Nous décrivons dans cet article les différentes étapes de traitement des textes dans le système que nous avons développé : (i) l'annotation en traits lexicaux et syntaxiques par le système Macaon ; (ii) le repérage des mentions par un modèle obtenu par apprentissage sur le corpus ANCOR ; (iii) l'annotation sémantique des mentions à partir de deux ressources : le DEM et le LVF ; (iv) l'annotation en coréférences par un système à base de règles. Le système est évalué sur le corpus ANCOR.

ABSTRACT

End-to-end coreference resolution for French.

We aim at developing a general coreference resolution system, for processing pronoun anaphora as well as direct coreferences. We describe in this article the different processing steps of the developped system : (i) extraction of lexical and syntactic features using the Macaon system ; (ii) mention detection with a supervised sequence classifier trained on the ANCOR corpus ; (iii) semantic mention tagging with two resources : DEM and LVF ; (iv) coreference prediction with a rule-based system. The system is evaluated on the ANCOR corpus.

MOTS-CLÉS : expression référentielle, mention, détection automatique de relations de coréférence.

KEYWORDS: referential expression, mention, automatic detection of coreferences.

1 Introduction

La résolution des anaphores participe à l'interprétation sémantique des textes : elle met en relation les expressions qui font référence à une même entité du discours. Les expressions référentielles, également appelées mentions, sont les noms propres, les groupes nominaux (définis ou indéfinis) et les pronoms.
Considérons l'exemple : *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ? il (/cet appareil) m'a coûté cher, j'espère le retrouver ; malheureusement l'écran est un peu rayé.* Les mentions peuvent être liées par différents types de relations de coréférence. Nous utilisons la terminologie définie dans (Muzerelle *et al.*, 2013) : anaphore pronominale *mon portable ... le* ; relation directe *mon portable ... ce portable* ; relation indirecte *mon portable ... cet appareil* ; relation associative *mon portable...l'écran*.

Pour identifier les antécédents des pronoms (*portable...le...il...le*) et trouver les coréférences (*portable...portable / appareil...écran*), on utilise classiquement le genre, le nombre, la position syntaxique, la distance des mots et des informations sémantiques sur les éventuelles mentions coréférentes (Lee *et al.*, 2013; Poesio *et al.*, 2010). Les systèmes développés pour traiter les anaphores portent en majorité sur l'anglais, on en trouve une description détaillée dans (Poesio *et al.*, 2010). Dans les systèmes développés pour le français, on peut citer les travaux de Trouilleux (2001) qui traite les anaphores pronominales et RefGen (Longo & Todirascu, 2010).

L'évaluation d'un système de détection des coréférences peut être faite manuellement, ce qui est très coûteux. Elle peut être faite automatiquement si l'on dispose d'un corpus de référence, et, à ce titre, le corpus ANCOR est une ressource très intéressante, constituée d'un ensemble de 460 dialogues transcrits annotés manuellement en mentions et en relations de coréférence et relations anaphoriques, comptant environ 500 000 mots au total (Muzerelle *et al.*, 2013).

Notre objectif est l'élaboration d'un système de détection automatique de relations de coréférence intra- et inter-phrases pour le français, de bout en bout, le plus général possible, pour le traitement des anaphores pronominales et des coréférences directes. Nous décrivons dans cet article les différentes étapes du traitement des textes sur lesquels nous travaillons :

1. annotation en traits lexicaux et syntaxiques par le système Macaon (Nasr *et al.*, 2011) ;
2. repérage des mentions par un modèle obtenu par apprentissage sur le corpus ANCOR ;
3. annotation sémantique des mentions à partir de deux ressources : le DEM et le LVF (Dubois & Dubois-Charlier, 2010) , (Dubois & Dubois-Charlier, 1997) ;
4. annotation en coréférences par un système à base de règles.

Nous décrivons pour finir l'évaluation mise en oeuvre par alignement sur le corpus ANCOR.

Nos corpus d'application sont des corpus de dialogues oraux transcrits ou dialogues en tchat. Dans ce type de textes, les relations anaphoriques pronominales sont nombreuses. Par exemple, sur le corpus de dialogues ANCOR, une étude distributionnelle sur les mentions a compté un total de 115 672 mentions tous types confondus, dont 51,1% de mentions pronominales et 48,9% de mentions nominales (Désoyer *et al.*, 2015).

2 Extraction des traits

L'extraction de traits consiste en l'analyse syntaxique des phrases, la détection de mentions, puis la classification sémantique des noms et verbes.

2.1 Repérage des mentions

Nous avons utilisé le corpus ANCOR pour construire un modèle de repérage des mentions par apprentissage automatique. Les mentions sont les groupes nominaux et les pronoms, hors pronoms réfléchis. Le corpus a été partitionné en *train, dev* et *test*, le développement du système étant réalisé grâce aux deux premières parties et l'évaluation sur la dernière.

La chaîne de traitement Macaon (Nasr *et al.*, 2011) génère une analyse en parties de discours et en dépendances syntaxiques à partir du corpus brut. Le Lefff permet ensuite de faire une analyse

morphologique des formes et d'en déduire leur lemme (Sagot, 2010). Ces traits servent de base au repérage des mentions.

Les mentions du corpus ANCOR sont projetées sur les mots à l'aide d'un encodage IOB, en applatissant les entités enchassées selon la segmentation maximale. Puis un modèle de séquence CRF (Kudo, 2005) est entraîné à partir des traits suivants :

— 1,2,3-grammes de mots, lemmes et parties de discours ;
— étiquette de dépendance du mot ;
— gouverneur (mot, lemme, partie de discours et étiquette de dépendance) ;
— longueur du constituant dont le gouverneur est la tête ;
— profondeur et hauteur dans l'arbre syntaxique du parent commun entre le mot et son suivant, et le mot et son précédent.

Les segments obtenus sont corrigés à l'aide de l'arbre syntaxique de manière à faire réapparaître les entités enchassées.

2.2 Classification sémantique des noms et des verbes

La sémantique des entités dont on parle est un trait important à prendre en compte. Soit les exemples :

1. *Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard*
2. *Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni.*
3. *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?*
4. *Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.*

Si l'on sait identifier quelles expressions dénotent une personne ou un groupe de personnes, et lesquelles dénotent une entité inanimée, certaines anaphores pronominales pourront être résolues facilement : par exemple l'objet de *punir* ne peut être qu'un humain ou un animal, et l'objet de *rapporter* doit être une entité inanimée. Il est donc utile de disposer d'un typage sémantique des expressions nominales, et d'un typage de chaque pronom dans le contexte où il apparaît.
Pour acquérir des données sémantiques sur les noms, la plupart des systèmes développés pour l'anglais utilisent WordNet ; d'autres travaux (par exemple (Ji & Lin, 2009)) mettent en oeuvre des processus d'apprentissage sur des N-grams du Web, sur les catégories de Wikipedia, ou sur des corpus annotés manuellement, mais la fiabilité des résultats d'apprentissage n'est pas totale, et les rares corpus annotés ont une couverture limitée (Poesio *et al.*, 2010). Les pronoms quant à eux peuvent être typés sémantiquement à partir des verbes dont ils sont actants, si l'on applique des préférences sélectives comme celles que l'on trouve par exemple, pour l'anglais, dans VerbNet (Kipper-Schuler, 2005).

Dans le but d'élaborer pour le français un système de traitement des anaphores qui puisse être utilisé quel que soit le domaine d'application, nous faisons le choix du DEM (*Dictionnaire Électronique des Mots*) et du LVF (*Les Verbes Français*) pour l'acquisition de données sémantiques (Dubois & Dubois-Charlier, 2010) (Dubois & Dubois-Charlier, 1997) : en effet leur couverture du français est maximale et des données sémantiques peuvent en être extraites directement, sans apprentissage.

Classification des noms Le DEM est une base de données qui répertorie tous les mots du français (145 198 entrées). On y trouve pour chaque entrée son sens, des propriétés catégorielles, morphologiques, sémantiques, syntaxiques. Chaque nom peut y avoir plusieurs entrées, qui correspondent à ses

différents sens. Nous avons extrait du DEM tous les noms communs qui y sont présents, et choisi de ne garder que les deux premiers sens de chaque nom, pour ne pas trop en brouiller la sémantique. Nous avons extrait pour chaque nom les informations pertinentes qui donnent :

— son appartenance à l'une des cinq classes : *Tout, Animé, NonAnimé, Humain, Animal* ;

— des informations sur le ou les domaines sémantiques dans lesquels ce nom est utilisé.

Plusieurs opérations de filtrage et de croisement de ces informations ont permis d'attribuer à chaque nom une entrée unique, avec sa classe et ses domaines de rattachement (87 037 entrées).
La classification a été établie à partir des cinq classes *Tout, Animé, NonAnimé, Humain, Animal* ; *Tout* est la réunion des classes *Animé* et *NonAnimé*, *Animé* est la réunion des classes *Humain* et *Animal*. Quelques autres classes ont été définies, par exemple *Véhicule*. Pour les noms polysémiques, on utilise si nécessaire la réunion de deux classes.
Ci-dessous un extrait des informations obtenues (LOC = locatif/lieu ; LAN = langue ; LIT = littérature).

nom	classe	domaine
acadien	Humain+NonAnime	LOC, LAN
académicien	Humain	LIT
académie	NonAnime	LIT

Par ailleurs, nous avons établi une liste de noms propres, où chacun est associé à une classe sémantique : villes, pays, prénoms, etc. Cette liste contient entre autres les noms propres de nos corpus d'application. Nous obtenons finalement une classification sémantique dans une taxinomie d'une quinzaine de classes.

Typage sémantique des verbes à partir du LVF Le LVF donne pour chaque verbe ses constructions syntaxiques accompagnées de la nature de ses actants, à l'interface syntaxe-sémantique (26 610 entrées). Chaque verbe peut y avoir plusieurs entrées. L'extraction d'informations dans plusieurs champs, puis le filtrage de ces informations nous ont permis d'obtenir un typage sémantique des actants sujet et complément d'objet direct. Nous ne gardons qu'une entrée pour chaque verbe, qui couvre les usages les plus courants de ce verbe. Nous avons obtenu une table de 12 484 verbes, où chacun est associé au type de ses actants sujet et objet.

verbe	sujet	compl-objet-direct
attendre	Animé+Véhicule	Tout
atterrir	Tout	
attester	Humain	
attraper	Animé	Tout

3 Traitement des anaphores

Les coréférences sont extraites à partir d'un ensemble de règles sur l'arbre syntaxique et l'ensemble des traits sémantiques et morphologiques extraits. Comme dans la plupart des systèmes de résolution de coréférences, nous cherchons pour chaque mention son coréférent le plus proche dans le texte qui

précède.

Résolution d'anaphores pronominales
On commence par identifier, via les dépendances, les pronoms utilisés dans des formes impersonnelles (*il y a*, *il faut*) ou les pronoms qui reprennent un segment phrastique (- *Louis est arrivé.* - *Oui, je le sais*). Ces pronoms sont ignorés dans les traitements qui suivent.
Ensuite, pour chaque pronom en position d'actant sujet ou objet d'un verbe, les dépendances permettent de trouver le verbe dont il est actant ; on en déduit son type sémantique, à partir de la table de typage des verbes. Pour les pronoms en position de complément indirect ou prépositionnel, on affecte le type NonAnimé à *en* et *y*, et le type Animé à *lui, elle, eux, elles*.
Pour chaque pronom on cherche, parmi les mentions candidates, celle qui est la meilleure au vu des contraintes du système, qui prennent en compte le genre, le nombre, la classe sémantique, la distance, ainsi que des éléments de la théorie du liage (Chomsky, 1981). La relation de coréférence peut être établie avec un nom qui est l'antécédent du pronom, ou un autre pronom dont il est coréférent. La recherche se fait en plusieurs passes, chacune d'elles remontant plus ou moins loin dans le dialogue et étant plus ou moins contrainte : dans les premières passes on impose un accord sur le genre, le nombre et le type sémantique, dans les suivantes on assouplit progressivement ces contraintes. Cette méthode est semblable aux méthodes classiques qui utilisent des filtres linguistiques pour écarter des candidats, et un ordonnanceur qui met en oeuvre des préférences.

Le cas des cataphores est très particulier (*Vous pouvez le décrire ce portefeuille*) : si l'annotation syntaxique en dépendances est correcte, la relation de coréférence entre le pronom et son coréférent est directement tirée des dépendances : les deux mots ont même étiquette et même gouverneur.

Recherche de coréférences nominales
On effectue une recherche de coréférents pour chaque expression nominale définie construite sur un nom commun ou un nom propre. Les expressions définies sont de deux types : celles qui introduisent une nouvelle entité et celles qui font référence à une entité déjà introduite. Notre recherche ne porte que sur les coreférences directes entre expressions de même tête lexicale (*mon portable...ce nouveau portable*). Elle se fait en considérant, outre la tête nominale, la présence éventuelle de modifieurs, pour ne pas mettre en relation par exemple *mon nouveau portable* et *mon ancien portable*. Mais notre recherche à ce niveau est très incomplète car il faudrait savoir a priori quels modifieurs sont compatibles et lesquels ne le sont pas. Dans l'état actuel de notre travail, nous considérons que deux expressions peuvent être mises en relation si l'ensemble des modifieurs de l'une est inclus dans l'ensemble des modifieurs de l'autre ; le cas le plus simple étant que l'une des deux expressions n'ait pas de modifieur (*mon portable...ce nouveau portable*).

4 Évaluation

L'évaluation et la comparaison des systèmes de recherche de coréférences sont complexes car il faut prendre en compte (Poesio *et al.*, 2010) :

- les performances des éventuels pré-traitements pour l'annotation en genre, nombre, syntaxe, etc ;
- les performances de l'éventuel reconnaisseur d'entités nommées et de l'identification des mentions ;

— le type de texte que l'on traite : manuel technique, article de journal, dialogue, etc.

Différentes métriques ont été définies pour les campagnes d'évaluation. Nous utilisons la métrique Blanc (M.Recasens & Hovy, 2010), qui est la plus récemment définie et dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence. Les calculs se font sur les chaînes de coréférences de chaque mention. L'évaluation de notre système a été faite en utilisant comme référence la partie *test* du corpus ANCOR, qui contient 23 079 mentions.

Le corpus ANCOR contient des coréférences *directes*, *indirectes* et *associatives*. Notre système ne traite que les coréférences *directes*, mais l'évaluation a été faite sur les coréférences directes et indirectes pour que les résultats soient comparables [1].

Tout d'abord, sur le *test*, le détecteur de mentions obtient un F-score de 88.94% comme précisé dans la table 1a. Ces performances sont données sur les arbres syntaxiques prédits (ainsi que tous les autres traits). Ensuite, la prédiction des liens de coréférence obtient un score Blanc de 65.70% comme décrit dans la table 1b. Ces performances sont données sur les mentions prédites par le détecteur susmentionné. Pour mettre ce résultat en perspective, on peut se référer aux travaux sur BART dans le projet SENSEI (Kabadjov & Stepanov, 2015) qui a obtenu un score Blanc de 15.04% sur le corpus ANCOR (provoqué par un faible rappel).

Metrique	**Perf.**
Précision	90.05%
Rappel	87.86%
F-score	88.94%

(a) Détecteur de mentions

Système	**F-score**
Toutes les mentions	65.70%
Pronoms uniquement	64.05%

(b) Détecteur de coréférences

TABLE 1: Performances sur l'ensemble de *test* du corpus ANCOR.

Pour explorer la généralité du système créé, nous avons effectué une évaluation manuelle supplémentaire sur le corpus DECODA de service client dans le domaine des transports (Bechet *et al.*, 2012), uniquement sur les pronoms clitiques. Les prédictions du système ont donné un taux de réussite de 0.70, ce qui tend à montrer que l'approche peut être appliquée à de nouveaux domaines.

5 Discussion

Dans les prédictions de notre système, la plupart des erreurs viennent des éléments suivants :

— les cas où le pronom ne peut pas être typé de façon fine : en particulier s'il est sujet de *avoir* ou *être* ou s'il n'a pas de lien de dépendance vers un verbe (problème souvent dû à une disfluence ; les disfluences sont très nombreuses dans les dialogues) ;
— des expressions figées, qui ne sont pas repérées dans les prétraitements ;
— des cas d'homonymie sur lesquels une erreur de lemmatisation produit une erreur d'analyse syntaxique.

Nous pensons que notre système est général dans le sens qu'il a une grande couverture lexicale et que les procédures de recherche de coréférences sont classiques et devraient s'appliquer à tous les types de textes. Néanmoins, ces procédures ont été définies en partant de l'étude de corpus de dialogues ;

1. Les références non supportées par notre systèmes sont donc considérées fausses

elles sont donc probablement plutôt adaptées à ce type de textes.
Notre travail sur le traitement des relations directes peut encore être amélioré. Quelques critères de sélection d'antécédents n'ont pas encore été exploités et vont être étudiés, en particulier le parallélisme dans les rôles syntaxiques des mentions. D'un autre côté, l'ajout d'une mesure de saillance n'a pas été probant, alors qu'il est mis en avant dans de nombreux travaux, par exemple dans (Lappin & Leass, 1994) et (Mitkov, 2002) qui utilisent la saillance mais aucune connaissance sémantique de sens commun. Par ailleurs, nous étudions actuellement l'ajout de la sémantique des adjectifs, mais cela semble n'apporter qu'une faible amélioration.

Le traitement des relations indirectes serait envisageable en intégrant plus de données sémantiques, en particulier pour prendre en compte la synonymie ou l'hyperonymie. Actuellement, nous traitons quelques cas de synonymie spécifiques à nos corpus (*le bus 42... le 42*). Par contre le traitement des relations associatives semble hors de portée d'un système de règles simple.

6 Remerciements

Ce travail a été financé par l'Agence Nationale pour la Recherche au sein des projets suivants : ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et ANR-11-IDEX-0001-02 (A*MIDEX).

Références

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC*, p. 1343–1347.

CHOMSKY N. (1981). *Lectures on Government and Binding*. Foris Publications.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les Verbes Français.* Larousse-Bordas.

DUBOIS J. & DUBOIS-CHARLIER F. (2010). La combinatoire lexico-syntaxique dans le dictionnaire électronique des mots. les termes du domaine de la musique à titre d'illustration. *Langages 3/*, p. 31–56.

DÉSOYER A., LANDRAGIN, F. TELLIER I., LEFEUVRE A. & ANTOINE J. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *TAL*, **55 (2)**, 97–121.

JI H. & LIN D. (2009). Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, p. 220–229.

KABADJOV M. & STEPANOV E. (2015). *D4.2 – The SENSEI Discourse Analysis Tools*. Rapport interne.

KIPPER-SCHULER K. (2005). *VerbNet : A broad-coverage, comprehensive verb lexicon*. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

KUDO T. (2005). Crf++ : Yet another crf toolkit. *Software available at http ://crfpp. sourceforge. net*.

LAPPIN S. & LEASS H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics vol. 20 (4)*, p. 535–561.

LEE H., CHANG A., PEIRSMAN Y., CHAMBERS N., SURDEANU M. & JURAFSKY D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics vol. 39 (4)*, p. 885–916.

LONGO L. & TODIRASCU A. (2010). RefGen : a Tool for Reference Chains Identification. In *International Multiconference on Computer Science and Information Technology*, p. 447–454.

MITKOV R. (2002). *Anaphora Resolution*. Longman.

M.RECASENS M. & HOVY E. (2010). Blanc : Implementing the rand index for coreference evaluation. *Natural Language Engineering*, **17 (4)**, 485–510.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR : premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 555–563.

NASR A., BECHET F., REY J., & LE ROUX J. (2011). Macaon :a linguistic tool suite for processing word lattice. In *49th Annual Meeting of the Association for Computational Linguistics : demonstration session*.

POESIO M., PONZETTO S. & VERSLEY Y. (2010). Computational models of anaphora resolution : A survey. *http ://wwwusers.di.uniroma1.it/ ponzetto/pubs/poesio10a.pdf*.

SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse. Université Blaise Pascal, Clermont-Ferrand.