Résoudre la coréférence à l'aide d'un classifieur bayésien naïf

Olivier Tardif[1,2]

[1]France Télécom, TECH/EASY/LN
[2]Université de Provence, CILSH
o.tardif@francetelecom.com

Résumé

Nous présentons ici les bases d'une méthode de résolution de la coréférence entre les expressions nominales désignant des entités nommées. Nous comptons appliquer cet algorithme sur un corpus de textes journalistiques ; certains aspects de ce que l'on pourrait nommer les « facteurs de coréférence » dans ces textes nous amènent à favoriser l'utilisation de méthodes statistiques pour accomplir cette tâche. Nous décrivons l'algorithme de résolution de la coréférence mis en œuvre, constitué d'un classifieur bayésien naïf.

Mots-clés : classifieur bayésien naïf, coréférence, entités nommées.

Abstract

In this paper we describe a coreference resolution algorithm for nominal expressions denoting named entities. The corpus used consists of French newspaper texts. In these texts, some properties of what we call "coreference factors" lead us to prefer a statistical approach for the task. We describe a coreference resolution algorithm consisting in the implementation of a naive Bayes classifier.

Keywords: naive Bayes classifier, coreference, named entities.

1. Introduction

On dit de deux expressions qu'elles sont coréférentes lorsqu'elles désignent la même entité (Kibble et van Deemter, 2000) :

> **Manuel Antonio Noriega** *revêtu de son uniforme kaki a fait sa première apparition devant un tribunal américain, jeudi 4 janvier, à Miami (Floride). Par l'intermédiaire de son interprète,* **il** *a répondu très brièvement aux questions. [D'ici l'ouverture du procès,] les procureurs chargés de l'accusation auront eu le temps de réunir d'autres documents qui confirmeront, selon eux, la culpabilité de* **l'ancien dictateur panaméen**. (Le Monde 1103)

Dans cet extrait, les trois expressions *Manuel Antonio Noriega*, *il* et *l'ancien dictateur panaméen* sont coréférentes. Nous décrivons dans ce travail le fonctionnement d'un outil de résolution de la coréférence entre expressions désignant des entités nommées[1] dans les textes écrits de style journalistique. L'objet d'étude englobe donc tous les types d'expressions nominales : noms pro-

[1] Nous définissons les « entités nommées » comme toute entité désignée dans un texte par au moins un nom propre, son identifiant unique. Une entité nommée n'est donc pas, de ce point de vue, un objet linguistique.

Facteurs issus d'une comparaison entre les deux expressions	
Morphologie	accord de genre accord de nombre accord de trait HUMAIN (en anglais, permet de distinguer *it* des autres pronoms)
Lexique	même tête syntagmatique deux noms propres mots en commun
Syntaxe	apposition parallélisme (même fonction dans deux propositions distinctes) prédication (sujet et l'objet d'un verbe attributif) même domaine d'arguments (deux arguments du même verbe non attributif)
Sémantique	même classe sémantique relation de synonymie ou d'hypéronymie arguments de verbes ayant les mêmes restrictions sélectionnelles
Texte	Distance en mots, phrases ou paragraphes Collocations chaîne de caractères en commun
Facteurs propres à une seule des deux expressions	
Syntaxe	reprise est en position sujet antécédent est en position sujet
Détermination	antécédent précédé d'un déterminant défini reprise précédée d'un déterminant défini reprise précédée d'un déterminant démonstratif
Texte	Nombre d'occurrences de l'antécédent
Autres	Alias (*e.g.* équivalence entre *Mikhail Gorbatchev* et *Mr Gorbatchev* Acronymes

Tableau 1. Facteurs de coréférence

pres, descriptions définies et indéfinies, et anaphores pronominales. Concrètement, il s'agit de retracer, pour chaque entité nommée, toutes les expressions nominales qui y réfèrent.

2. Les méthodes existantes

On peut classer les différents travaux du domaine selon le type d'algorithme de résolution utilisé. Il existe deux tendances majeures : les méthodes *symboliques*, et les méthodes statistiques. Dans le premier groupe on retrouve notamment les algorithmes de contraintes / préférences (Lappin et Leass, 1994), à base de règles (Harabagiu et Maiorano, 2000), ou fondés sur un calcul de similarité (Cardie et Wagstaff, 1999). Les méthodes statistiques sont pour leur part principalement basées sur les arbres de décision (Soon *et al.*, 2001 ; Ng et Cardie, 2002), mais on retrouve aussi d'autres types d'algorithmes probabilistes (Witte et Bergler, 2003). Ces classes ne sont pas mutuellement exclusives : de nombreux travaux emploient des méthodes mixtes (Hartrumpf, 2001 ; Iida *et al.*, 2003) ; notons également que les recherches mentionnées ci-dessus ne sont que les exemples les plus représentatifs de chaque classe.

Qu'il soit de type symbolique ou statistique, chaque système de résolution dépend de données issues d'une analyse linguistique ; ces données et la façon dont elles sont utilisées par chaque système permettent de définir des *facteurs de coréférence*. Le genre grammatical, par exemple, peut servir à comparer deux expressions afin de déterminer si elles sont coréférentes. Précisons que ces facteurs peuvent aussi bien servir à déterminer que deux expressions sont coréférentes qu'à exclure qu'elles le soient. La liste suivante contient les facteurs les plus souvent mentionnés, regroupés en fonction du niveau d'analyse linguistique auquel ils font appel.

Tous les facteurs utilisés dans tous les travaux ne sont pas représentés ici : nous avons rejeté

ceux que nous ne jugions pas généralisables au français, ceux qui sont trop spécifiques à la méthode de résolution employée (*e.g.* les différentes notions de *focus* employées dans (Azzam *et al.*, 1998)), ou encore ceux qui dépendent fortement d'une ressource spécifique comme le type de corpus utilisé.

3. Les données à traiter

Nous avons sélectionné 79 textes du journal *Le Monde* comprenant entre 500 et 900 mots, et dont les phrases ont en moyenne une trentaine de mots. Ces textes ont ensuite été soumis à la chaîne de traitement linguistique TiLT développée chez *France Télécom*, configurée pour générer une annotation lexicale, morphosyntaxique, syntaxique et sémantique de chaque texte segmenté. À cette annotation automatique, de format xml, des informations de coréférence entre expressions nominales ont été ajoutées manuellement par un annotateur. Une description plus détaillée du corpus et des annotations est faite dans (Tardif, à paraître).

L'extraction de certaines données statistiques de ce corpus nous a permis de constater qu'aucun facteur de coréférence n'est totalement autonome : en effet plusieurs d'entre eux sont corrélés. À titre d'exemple, considérons deux paires de facteurs : *correspondance de classe sémantique* et *catégorie grammaticale*, et *distance en mots* et *catégorie grammaticale*.

La figure 1 regroupe en colonnes les expressions en fonction de certaines caractéristiques grammaticales, soit de gauche à droite : pronom personnel sujet, pronom objet, autres pronoms, groupe nominal défini, groupe nominal indéfini, autres groupes nominaux, et groupes nominaux sans déterminant (incluant les noms propres). On constate que les proportions relatives d'expressions de chaque classe sémantique varient d'une « catégorie » grammaticale à l'autre : par exemple, presque 90 % des pronoms personnels désignent des personnes, alors que ce pourcentage est beaucoup plus faible (environ 55 %) pour les groupes nominaux définis.

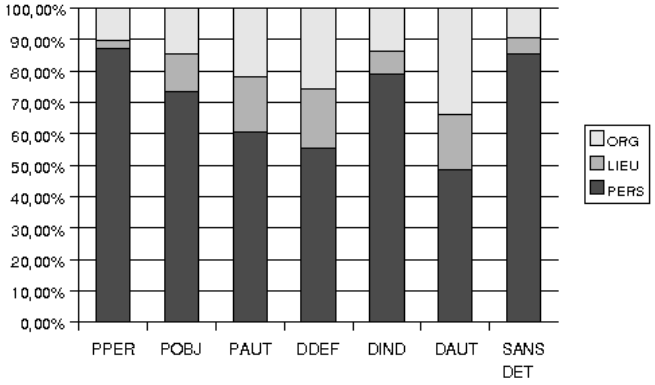

Figure 1. Catégorie grammaticale pour chaque classe sémantique

La figure 2 illustre la relation entre la distance et certaines de ces catégories grammaticales : nous avons calculé la distance moyenne entre les reprises et leur antécédent[2] ; ces valeurs sont regroupées dans des intervalles, dont la taille est de 5 mots, sur l'axe horizontal. On constate donc que les pronoms relatifs sont proches de leur antécédent ; que la probabilité de coréférence entre une expression quelconque et un pronom personnel décroît assez régulièrement en fonc-

[2] On définit l'antécédent de *x* comme l'expression coréférente la plus proche de *x* et qui la précède.

tion de la distance ; et que cette probabilité varie très peu pour les groupes nominaux définis.

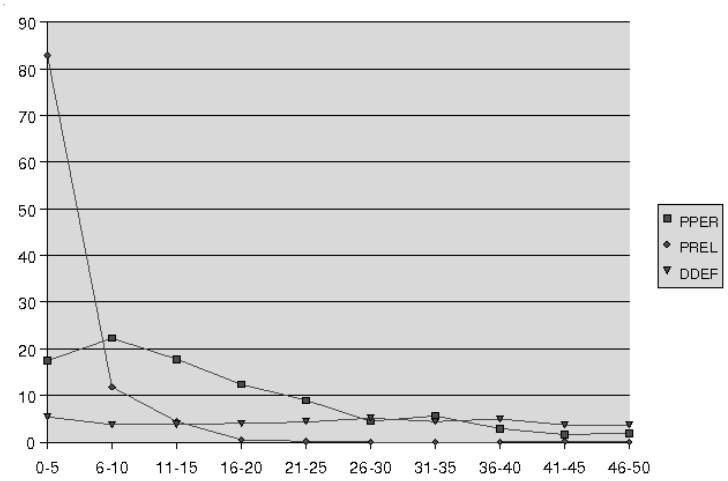

Figure 2. Distance moyenne entre coréférents par catégorie grammaticale

Ces faits nous incitent à diriger notre choix vers les méthodes statistiques plutôt que symboliques. En effet, une approche symbolique, par exemple à base de règles, de contraintes ou de reconnaissance de motifs, nécéssiterait de faire l'inventaire et la description des interactions possibles entre les facteurs de coréférence : la tâche s'annonce fastidieuse. À l'inverse, les méthodes statistiques misent sur un apprentissage automatique des corrélations de ce type.

4. Méthode de résolution

La méthode choisie consiste à implémenter un classifieur bayésien naïf (CBN), une approche adoptée par (Ge *et al.*, 1998) pour la résolution des pronoms en anglais.

La plupart des algorithmes de résolution de nature statistique (notamment (McCarthy, 1996 ; Soon *et al.*, 2001 ; Ng et Cardie, 2002)) se basent sur les arbres de décision (ID3, C4.5 et C5) ; cependant à notre avis les CBN ont deux avantages importants sur ces derniers. Premièrement, ils montrent une certaine robustesse face aux données superflues - la présence de données non-pertinentes lors de l'entraînement n'a aucun impact sur les performances. En second lieu, des stratégies d'optimisation comme le *boosting* (Elkan, 1997), permettent de compenser efficacement l'absence de certaines données durant l'entraînement. Ce dernier aspect est utile en TAL où les informations peuvent être difficiles à obtenir dans certains contextes, en raison de la complexité des traitements requis ou encore de l'ambiguïté « naturelle » qu'on retrouve dans les données.

Comme pour toutes les méthodes d'apprentissage automatique, une période d'entraînement est nécéssaire. La phase d'entraînement du CBN consiste à extraire d'un ensemble de textes annotés toutes les paires d'expressions nominales possibles. Chacune de ces paires constitue une instance d'entraînement dont la classe $c \in \{\text{COREF}, \text{NON COREF}\}$ est connue. Chaque instance est également caractérisée par un vecteur de variables $V = \{v_1 \dots v_n\}$ correspondant aux facteurs de coréférence décrits plus haut ; leurs valeurs sont déterminées par les données annotées (cf. 3).

Lors de la phase de classification il s'agit également d'extraire toutes les paires d'expressions

possibles d'un ensemble de textes et de leur associer les variables V, mais ici on ignore si ces paires sont ou non coréférentes : on détermine leur statut en calculant laquelle des deux classes COREF et NON COREF est la plus probable. Ces probabilités sont calculées suivant la formule (1) :

$$Pr(c|v_1, \ldots, v_n) = Pr(c) \cdot Pr(v_1|c) \cdot \ldots \cdot Pr(v_n|c) \quad (1)$$

ainsi la classe la plus probable est déterminée par la fonction $Coref$, définie comme suit :

$$Coref(v_1, \ldots, v_n) = argmax_c \, P(c) \prod_{i=1}^{n} P(v_i|c) \quad (2)$$

Une des particularités des CBN est que formellement on suppose l'indépendance entre les variables : on ne cherche nullement à déterminer, par exemple, $P(v_1|v_2)$. Ceci peut sembler incompatible avec les données linguistiques observées, dont on a pu constater plus haut qu'elles pouvaient être dépendantes. Cependant il a été démontré (Rish, 2001) que les CBN pouvaient donner en pratique de très bons résultats même lorsque les variables choisies sont corrélées ; l'indépendance des variables est donc simplement un *a priori*.

5. Méthode d'évaluation

La première étape de nos travaux consistera à implémenter un classifieur basé sur les variables utilisées par (Soon *et al.*, 2001), qui nous servira de point de départ ; par la suite l'évaluation des performances pourra nous guider dans les modifications à apporter à cet ensemble de variables. Notre procédure d'évaluation consiste à diviser le corpus en deux afin d'obtenir un sous-corpus d'entraînement et un sous-corpus d'évaluation, respectivement constitués de 3/4 et de 1/4 des textes.

6. Conclusion et perspectives

Le présent travail décrit une méthode de résolution de la coréférence s'appuyant sur un classifieur bayésien. Nous avons dressé une liste des facteurs de coréférence les plus utilisés dans les algorithmes de résolution existants, et avons montré qu'il existe des corrélations entre certains facteurs. Celles-ci laissent croire que les méthodes statistiques sont plus aptes à traiter le problème de la résolution de la coréférence que les méthodes symboliques. Nous avons également présenté les raisons qui nous font préférer les CBN aux arbres de décisions dans le cadre de cette tâche. La suite des travaux consiste évidemment à implémenter et évaluer la méthode proposée.

Références

AZZAM S., HUMPHREYS K. et GAIZAUSKAS R. (1998). « Extending a simple coreference algorithm with a focusing mechanism ». In DAARC (éd.), *Proceedings of the second colloquium on discourse anaphora and anaphor resolution (DAARC2)*. Granada, Spain, p. 15–27.

CARDIE C. et WAGSTAFF K. (1999). « Noun Phrase Coreference as Clustering ». In *Joint SIGDAT conference on empirical methods in natural language processing and very large corpora (EMNLP-VLC99)*.

ELKAN C. (1997). *Boosting and naive bayesian learning*. Rapport interne, Department of Computer Science and Engineering, University of California in San Diego.

GE N., HALE J. et CHARNIAK E. (1998). « A Statistical Approach to Anaphora Resolution ». In *Proceedings of the Sixth Workshop on Very Large Corpora*.

HARABAGIU S. M. et MAIORANO S. J. (2000). « Multilingual Coreference Resolution ». In *Proceedings of ANLP-NAACL00*.

HARTRUMPF S. (2001). « Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics ». In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*.

IIDA R., INUI K., TAKAMURA I. et MATSUMOTO Y. (2003). « Incorporating Contextual Cues in Trainable Models for Coreference Resolution ». In *Proceedings of EACL03*.

KIBBLE R. et VAN DEEMTER K. (2000). « On Coreferring. Coreference in MUC and Related Annotation Schemes ». In *Computational Linguistics*, 26 :4, 629–637.

LAPPIN S. et LEASS H. J. (1994). « An Algorithm for Pronominal Anaphora Resolution ». In *Computational Linguistics*, 20, 4, 537–561.

MCCARTHY J. (1996). *A Trainable Approach to Coreferecne Resolution for Information Extraction*. PhD thesis, University of Massachussetts.

NG V. et CARDIE C. (2002). « Improving Machine Learning Approaches to Coreference Resolution ». In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

RISH I. (2001). « An empirical study of the naive Bayes classifier ». In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

SOON W. M., NG, HWEE T. et LIM D. C. Y. (2001). « A Machine Learning Approach to Coreference Resolution of Noun Phrases ». In *Computational Linguistics*, 27 :4.

TARDIF O. (à paraître). « Annotation de la coréférence entre expressions référentielles ». In JILC (éd.), *Actes des Journées Internationales de Linguistique de Corpus (JILC) 2005*. Lorient, France.

WITTE R. et BERGLER S. (2003). « Fuzzy Coreference Resolution for Summarization ». In ARQAS (éd.), *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*. Venice, Italy.