Un classifieur bayésien pour la résolution des anaphores

Davy Weissenbacher[1] Adeline Nazarenko[1]
(1) Université Paris-Nord - Laboratoire d'Informatique de Paris-Nord, 99 av. J-B. Clément 93430 Villetaneuse, FRANCE
dw@lipn.univ-paris13.fr, nazarenko@lipn.univ-paris13.fr

Résumé On oppose souvent les systèmes de résolution d'anaphores à base de connaissances linguistiques et ceux qui reposent sur des indices de surface. Chaque approche a ses limites et ses avantages. Nous proposons dans cet article une nouvelle approche reposant sur les réseaux bayésiens qui permet de combiner au sein d'une même représentation ces deux types d'informations hétérogènes et complémentaires. Nous montrons l'intérêt de notre approche en mesurant les performances du réseau bayésien qui sont supérieures à celles des systèmes de l'état de l'art.

1 Introduction

Historiquement, on oppose les systèmes de résolution d'anaphores qui exploitent des connaissances linguistiques et ceux qui reposent sur des indices de surface. Les premiers systèmes exploitent des connaissances complexes qui ne sont pas toujours fiables lorsqu'elles sont calculées automatiquement et peuvent faire défaut ou être incomplètes lorsqu'elles sont produites par l'homme. Les seconds systèmes exploitent souvent des méthodes d'apprentissage automatique plutôt que des connaissances. Malheureusement, aujourd'hui, l'apprentissage n'est raisonnablement possible que sur des indices de surface, qui ne permettent souvent pas de résoudre parfaitement la résolution.

Dans cet article nous proposons une nouvelle approche reposant sur le formalisme des réseaux bayésiens (RB) qui permet de dépasser cette opposition entre systèmes "pauvres" et systèmes "riches" en connaissances.

La section suivante revient sur les raisons de cette opposition. La section 3 décrit notre système reposant sur un classifieur bayésien. Enfin, dans la dernière section 4 nous validons notre approche en mesurant les performances de notre classifieur sur un corpus de génomique.

2 La complémentarité des connaissances linguistiques et des indices de surface

L'anaphore est une relation linguistique entre deux entités textuelles définie lorsqu'une entité textuelle (l'*anaphore*) renvoie à une autre entité du texte(l'*antécédent*). Comme la présence d'anaphores dégrade considérablement les performances des systèmes de TAL, la question de leur résolution est étudiée depuis longtemps. Ce travail se limite à la résolution de l'anaphore

du pronom *it* dans les textes anglais, probablement l'anaphore la mieux connue et la plus facile à résoudre.

Les premiers systèmes proposés dans la littérature exploitaient des connaissances linguistiques complexes traduisant les contraintes syntaxiques et sémantiques qui régissent l'anaphore. Comme le calcul automatique de ces connaissances était considéré comme impossible ou comme trop peu fiable pour être utilisable, ces connaissances linguistiques étaient produites manuellement, ce qui présupposait un important travail d'analyse préalable des textes.

Durant les années 1990, devant le besoin de systèmes de résolution robustes et peu coûteux à mettre en place, un nombre important de systèmes à bases d'indices de surface ont été proposés (Mitkov *et al.*, 2001). Ces systèmes abandonnent les connaissances linguistiques complexes des premiers systèmes. Ils approchent les connaissances nécessaires par des indices plus simples et que l'on suppose plus fiables. Leurs apports et leurs limites étaient mal connus (Mitkov, 2002) mais des travaux récents, par exemple (Kehler *et al.*, 2001), commencent à en mesurer les limites.

Ces limites nous renvoient au problème initial. Nous avons besoin de connaissances sémantiques et syntaxiques complexes pour la résolution de l'anaphore pronominale. Ces connaissances linguistiques, lorsqu'elles sont disponibles, ne sont pas fiables. Les précédents travaux ont cherché à remplacer ces connaissances par des indices de surfaces. Si leur calcul est toujours réalisable et plus fiable, ces indices peuvent ne pas exprimer, ou seulement de manière imprécise, les connaissances nécessaires à la résolution et ainsi rendre impossible une décision correcte dans les cas ambigus.

Nous proposons une modélisation reposant sur les RB. Ce modèle a été conçu pour raisonner sur des informations incertaines et incomplètes. Cette approche probabiliste offre la possibilité d'unifier dans une unique représentation connaissances linguistiques et indices de surface. Cette unification permet de corroborer les connaissances linguistiques grâce aux indices de surface qui sont observés en corpus. A l'inverse, l'exploitation de connaissances linguistiques permet de corriger certaines des erreurs des systèmes à base d'indices de surface.

3 Une approche intégrée : le modèle bayésien

Un RB est composé d'une description qualitative des dépendances des informations, un graphe orienté sans circuits, et d'une description quantitative, un ensemble de probabilités conditionnelles où chaque variable aléatoire (VA) est associée à un noeud du graphe. Une première étape de paramétrage permet de représenter les connaissances *a priori* pour chaque VA et chaque probabilité conditionnelle. L'étape suivante, l'étape d'inférence, consiste à réviser certaines probabilités *a priori* pour obtenir des probabilités *a posteriori* et à modifier en conséquence les valeurs des VA correspondantes à partir d'observations faites en corpus. Ces nouvelles informations sont propagées au travers du réseau et permettent de réviser les valeurs *a priori* des informations encore inconnues.

Notre classifieur, présenté par la figure 1, recherche l'élément le plus saillant du discours qui précède un pronom anaphorique. Cet élément est celui qui a la plus forte probabilité d'être l'antécédent du pronom. Pour calculer l'élément saillant nous avons conservé une partie des indices approchés du système MARS (Mitkov, 2002) et nous avons ajouté une série d'autres

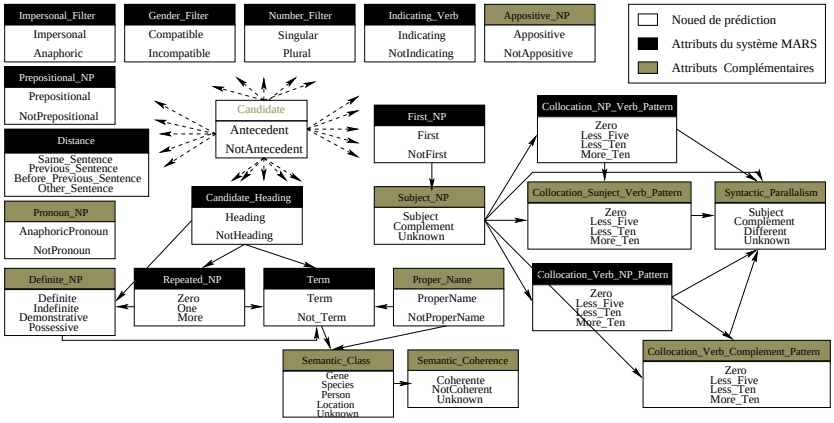

FIG. 1 – Un Réseau Bayésien pour la classification des candidats (le noeud `Candidate` est lié à tous les noeuds du réseau).

indices proposés par plusieurs travaux de l'état de l'art[1]. Nous avons complété ces indices, lorsque cela était possible, par les connaissances linguistiques qu'ils renforcent. Par exemple, le sujet d'une phrase est souvent l'élément saillant mais comme le calcul du rôle grammatical peut être erroné, il est intéressant d'exploiter en parallèle l'information concernant un indice de surface (*First_NP* : le premier groupe nominal (GN) de la phrase est très souvent le sujet du verbe) qui peut confirmer ou infirmer l'hypothèse du rôle grammatical.

4 Expérience et résultats

Nous avons travaillé sur un corpus de 2209 résumés d'articles de recherche de génomique (environ 800 000 mots). Nous avons identifié 697 occurrences du pronom *it*. Deux annotateurs humains ont analysé et classé chaque pronom soit comme impersonnelle soit comme anaphorique puis identifié les antécédent[2].

Notre corpus étant de taille moyenne, nous avons procédé à une validation croisée pour valider nos résultats. Nous sélectionnons aléatoirement deux tiers du corpus pour calculer les probabilités conditionnelles *a priori*. Nous appliquons ensuite notre classifieur paramétré grâce à ces probabilités sur le tiers restant. Nous réitérons 20 fois ces opérations pour obtenir une moyenne de ses performances sur le corpus.

Nous avons réalisé la résolution avec 4 systèmes différents. Trois systèmes servent de systèmes de comparaison. Le système *Aléatoire* choisit un antécédent au hasard dans la liste des candidats. Le système *Premier GN* sélectionne toujours le premier GN de la phrase précédant le pronom comme antécédent et le système MARS que nous avons implémenté. Le dernier système est le classifieur bayésien (CB) que nous cherchons à évaluer.

Pour les 3 derniers systèmes nous donnons deux mesures différentes des performances, un taux

[1] Les attributs décrivant les indices du système MARS sont colorés en noir. Les attributs qui les complètent sont en gris. Le noeud de prédiction est le noeud `Candidat`, au centre. Il estime la probabilité pour une occurrence d'un candidat d'être l'antécédent d'un pronom donné.

[2] Le second annotateur n'ayant pas entièrement terminé l'annotation, le taux d'accord n'a pu être calculé

de succès strict et partiel[3] La dernière colonne donne les scores maximum possibles pour la résolution, compte tenu des erreurs de l'analyse syntaxique.

Systèmes	**Resultats**	
	Strict	*Partiel*
Aléatoire	6%	-
Premier GN	36.3%	51%
MARS	26.7%	43%
Classifieur Bayésien	44.0%	61%
MAX	93.3%	97.8%

TAB. 1 – Comparaison des résultats (taux de Succès)

La comparaison des scores des systèmes MARS et CB permet d'établir l'apport des connaissances linguistiques complexes dans la résolution en dépit de leur qualité imparfaite. Ces connaissances supplémentaires rendent possible la désambiguïsassions entre différents candidats.

Parmi les erreurs restantes du CB, 47% sont dues à un calcule erroné de l'élément saillant. Un nombre plus important d'indices sont trouvés pour un candidat différent de l'élément saillant, l'antécédent, et le classifieur calcule une plus grande probabilité pour ce premier candidat. 21% des erreurs s'expliquent par la mise en échec de la théorie de la saillance. Les erreurs restantes proviennent de la mauvaise segmentation et des candidats manquant totalement ou partiellement.

5 Conclusion

Dans cet article nous avons présenté un système de résolution des anaphores reposant sur un réseau bayésien dont les premiers résultats sont encourageant. Ce modèle permet de dépasser l'opposition historique des systèmes à base de connaissances linguistiques et d'indices de surface. Une opposition qui apparaît infondée : les connaissances linguistiques sont nécessaires mais souvent indisponibles et jamais fiables ; les indices de surface sont généralement calculables et de bonne qualité mais il reste des problèmes d'ambiguïté. En unifiant ce deux types de connaissances au sein d'une unique représentation, le modèle offre un mécanisme de raisonnement dont nous nous servons pour corriger et suppléer les connaissances linguistiques en les complétant des indices de surface.

Références

KEHLER A., APPELT D., TAYLOR L. & SIMMA A. (2001). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference*, p. 289–296.

[3]Strict Success rate = $\frac{Anaphorecorrectementrsolue}{Touteslesanaphores}$

Partial Success rate = $\frac{Anaphorecorrectementetpartiellementrsolue}{Touteslesanaphores}$

Le taux de succès est strict lorsque l'antécédent exacte a été annoté par le système et partiel lorsque seule une partie de l'antécédent à été annotée.

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

MITKOV R., LAPPIN S. & BOGURAEV B. (2001). Introduction to the special issue on computational anaphora resolution. *Computational Linguistics*, p. 27(4) :473–477.