Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne

Davy WEISSENBACHER, Adeline NAZARENKO
Université Paris-Nord, LIPN, 99 av. J-B. Clément, F-93430 Villetaneuse
{dw,nazarenko}@lipn.univ-paris13.fr

Résumé. On oppose souvent en TAL les systèmes 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 qui repose sur les réseaux bayésiens et qui permet de combiner au sein d'une même représentation ces deux types d'informations hétérogènes et complémentaires. Nous justifions l'intérêt de notre approche en comparant les performances du réseau bayésien à celles des systèmes de l'état de l'art, sur un problème difficile du TAL, celui de la résolution d'anaphore.

Abstract. In NLP, a traditional distinction opposes linguistically-based systems and knowledge-poor ones, which mainly rely on surface clues. Each approach has its drawbacks and its advantages. In this paper, we propose a new approach based on Bayes Networks that allows to combine both types of information. As a case study, we focus on the anaphora resolution which is known as a difficult NLP problem. We show that our bayesain system performs better than a state-of-the art one for this task.

Mots-clés : réseaux bayésiens, résolution des anaphores, connaissance linguistique, indice de surface.

Keywords: bayesian network, anaphora resolution, linguistic knowledge, surface clue.

1 Introduction

On oppose souvent en TAL les systèmes qui exploitent des connaissances linguistiques et ceux qui reposent sur des indices de surface. Les premiers systèmes ne sont pas toujours fiables parce qu'ils exploitent des connaissances complexes qui peuvent être erronées lorsqu'elles sont calculées automatiquement ou incomplètes lorsqu'elles sont produites manuellement. Les seconds systèmes s'appuient généralement sur des méthodes d'apprentissage automatique et sur des indices de surface qui sont plus faciles à obtenir mais qui ne permettent de traiter que les cas simples ou les plus courants de la tâche dévolue au système.

Dans cet article nous proposons une nouvelle approche qui permet de dépasser cette opposition entre systèmes «pauvres» et système «riches» en connaissances. Cette approche repose sur le formalisme des réseaux bayésiens. Ce formalisme est encore peu exploité en TAL mais il repose sur un modèle probabiliste conçu pour raisonner sur des informations incertaines, partielles et manquantes.

Nous validons notre approche sur la tâche de la résolution automatique des anaphores où, en raison de la complexité et du nombre de connaissances nécessaires, l'opposition des systèmes à base de connaissances linguistiques et d'indices de surface est très marquée. Après avoir validé l'approche en développant un premier classifieur bayésien qui permet de distinguer pronoms impersonnels et pronoms anaphoriques, nous analysons les performances d'un second classifieur qui trouve l'antécédent des pronoms anaphoriques.

La section suivante revient sur les raisons de l'opposition précédente dans le cadre de la résolution des anaphores pronominales. La section 3 décrit le modèle des réseaux bayésiens et son intérêt pour le TAL. Dans la section 4 nous validons notre approche en comparant les performances de différents systèmes pour la distinction des pronoms impersonnels et anaphoriques. Enfin, la dernière section présente un classifieur pour la tâche complète de la résolution des anaphores et compare ses résultats par rapport à l'état de l'art.

2 La complémentarité des connaissances linguistiques et des indices de surface

2.1 Le choix 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, l'anaphore la mieux connue et la plus facile à résoudre.

L'approche classique pour sa résolution automatique distingue trois étapes : la distinction des pronoms anaphoriques et impersonnels (*it is known that...* vs *it produced...*), la sélection des candidats possibles à l'antécédence et le choix de l'antécédent. Pour chaque étape, 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 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, 2002). Ces systèmes abandonnent les connaissances linguistiques complexes des premiers systèmes. Ils approchent les connaissances nécessaires par des indices plus simples à calculer et que l'on suppose plus fiables.

Pour la distinction des pronoms anaphoriques, (Husk & Paice, 1987) a ainsi proposé un ensemble d'automates encodant des connaissances linguistiques et permettant de reconnaître les séquences contenant des pronoms impersonnels. Jugeant que ces automates avaient une couverture trop faible, (Evans, 2001) propose une voie alternative reposant sur l'apprentissage automatique des indices de surface pour reconnaître les séquences caractéristiques. Pour le choix de l'antécédent, les connaissances syntaxico-sémantiques sont approchées de la même manière par des méthodes robustes. On sait que les schémas prédicat-argument améliorent les résultats du filtrage (Ponzetto & Strube, 2006), mais comme ces ressources ne sont pas toujours disponibles,

on a cherché à les approcher par un calcul fréquentiel : les régularités des coocurrences entre les sujets, les compléments et les verbes dessinent les contours des classes sémantiques. Les auteurs de (Dagan & Itai, 1990) montrent que les contraintes obtenues peuvent partiellement remplacer les connaissances sémantiques.

2.2 Les limites des indices de surface

Si les indices approchés proposés lors des années 1990 ont permis l'implémentation de systèmes robustes (Mitkov, 2002), leur apport et leurs limites étaient mal connus. Des travaux récents commencent à en mesurer les limites. L'étude de (Kehler *et al*., 2001) montre ainsi que les fréquences de (Dagan & Itai, 1990) n'améliorent pas les performances d'un système qui exploite déjà des informations morpho-syntaxiques. Les auteurs en concluent que l'apport des fréquences tient davantage du hasard que d'une véritable capture du sens sémantique.

Les limites rencontrées par les systèmes à base d'indices de surface 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. On peut chercher à les remplacer par des indices de surface dont le calcul est toujours réalisable et plus fiable mais ces indices peuvent ne pas exprimer, ou seulement de manière imprécise, les connaissances nécessaires à la résolution, ce qui produit des erreurs.

Nous proposons une modélisation reposant sur les Réseaux Bayésiens (RB), conçu pour raisonner sur des données 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

3.1 Des problèmes de classification

La distinction des pronoms impersonnels comme le choix de l'antécédent sont des tâches qui, comme de nombreuses tâches du TAL, se reformulent facilement en problèmes de classification.

Considérons par exemple la classification des pronoms impersonnels et anaphoriques : soit *Corpus* un ensemble de textes d'un même domaine, *Corpus_entraînement* et *Corpus_test* deux sous-ensembles stricts disjoints de *Corpus*, C_1 et C_2 les classes des occurrences des pronoms impersonnels et anaphoriques présents dans *Corpus*. e est une occurrence d'un pronom présent dans *Corpus* décrit par un vecteur $a = v_1, ...v_a$ d'attributs à valeurs dans $\mathbf{R}$. Pour les occurrences de *Corpus_entraînement*, les valeurs des attributs v_i sont obtenues à partir d'une analyse humaine du corpus : elles représentent selon les cas des connaissances linguistiques ou des indices de surface.

Le théorème de Bayes dit comment prédire la meilleure classe d'appartenance pour une occurrence d'un pronom inconnu de *Corpus_test* sur la base d'observations faites sur les occurrences

de *Corpus_entrainement*. La classe sélectionnée doit maximiser la probabilité

$$P(C_i|E) = \frac{P(E|C_i)*P(C_i)}{P(E)}$$

où $C_i \in \{C_1, C_2\}$, E une occurrence du corpus de test et $P(C_i|E)$ la probabilité conditionnelle que E appartienne à la classe C_i sachant la valeur des attributs de E, une probabilité estimée à partir des données d'entraînement. Si nous imposons la contrainte d'indépendance des attributs, le classifieur est un «classifieur bayésien naïf». Les attributs étant indépendants, la probabilité $P(E|C_i)$ se décompose en $P(v_1|C_i) * ... * P(v_a|C_i)$ et la probabilité à maximiser se reformule en

$$P(C_i|E) = \frac{P(C_i)}{P(E)} \prod_{j=1}^{a} P(v_j|C_i)$$

Pour tout E de *Corpus_test*, un classifieur bayésien attribue la classe C_1 à l'exemple E si P(Pronom=Impersonnel|E)≥P(Pronom=Anaphorique|E) et la classe C_2 sinon.

3.1.1 Le choix des attributs pour la classification

L'un des premiers systèmes distinguant les pronoms *it* impersonnels et anaphoriques (Husk & Paice, 1987) s'appuie sur un ensemble de règles de logique du 1^{er} ordre pour reconnaître les séquences qui contiennent une occurrence du pronom impersonnel. Les séquences qui introduisent les *it* impersonnels partagent une forme remarquable : elles commencent par un *it* et se terminent par un délimiteur comme *to, that, whether....* Les règles varient selon le délimiteur. Les tests réalisés par Paice montrent que ces règles réalisent un bon score avec 91,4%Acc [1] sur un corpus technique. Cependant les performances sont dégradées si on applique les règles à des corpus de nature différente. Le nombre de faux positifs (FP) augmente : certains attributs sont discriminants sur les corpus techniques mais ne le sont plus sur des corpus de nature différente.

Afin d'éviter cet écueil, (Lappin & Leass, 1994) décrit entièrement les séquences au moyen d'automates à états finis de la forme *It is not/may be<Modaladj>* ; *It is <Cogv-ed> that <Subject>* où *<Modaladj>* et *<Cogv>* dénotent des classes d'adjectifs modaux et de verbes cognitifs connus pour introduire des *it* impersonnels (par exemple *necessary, possible* et *recommend, think*). Ce système a une bonne précision (il produit peu de FP), mais il a un mauvais rappel (il produit beaucoup de FN) : seules les séquences exactes sont reconnues et il est toujours difficile d'obtenir des classes d'adjectifs et de verbes exhaustives.

(Evans, 2001) renonce à exploiter des connaissances linguistiques aussi complexes et se concentre sur des attributs plus fiables, les indices de surface. Evans considère 35 indices syntaxiques et contextuels (ex. la position du pronom dans la phrase, le lemme du verbe suivant...). Un système d'apprentissage, utilisant la méthode des K plus proches voisins, détermine le poids des attributs discriminants pour le domaine du corpus et classe les occurrences inconnues. Les premiers essais réalisent un score de 71,31%Acc satisfaisant sur un corpus de langue générale. (Litran *et al.*, 2004) reproduit un essai identique avec une Machine à Support de Vecteur (SVM) sur un corpus de génomique et obtient un score de 92,71%Acc.

[1] L'exactitude, en anglais Accuracy : Acc= $\frac{VP+VN}{VP+VN+FP+FN}$, où les faux positifs (FP) correspondent aux occurrences d'un pronom anaphorique étiquetées impersonnelles, les faux négatifs (FN) les occurrences de pronoms impersonnels étiquetées anaphoriques, les vrais positifs (VP) et les vrais négatifs (VN) correctement étiquetées comme impersonnels et anaphoriques, respectivement.

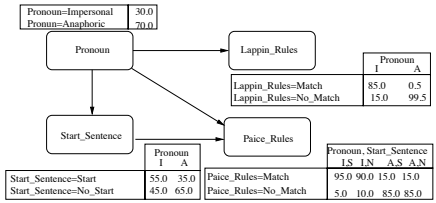

FIG. 1 – Exemple d'un classifieur bayésien modélisé par un réseau bayésien

Ces deux derniers systèmes d'apprentissage reposent donc uniquement sur des indices de surface. Constatant que les connaissances linguistiques sont peu fiables ou incomplètes, les auteurs renoncent à les utiliser comme attributs. Ce choix nous paraît trop radical : dès lors que ces connaissances linguistiques sont pertinentes pour notre tâche, il faut les intégrer dans la décision sous la forme d'attributs mais en se donnant les moyens de raisonner sur des attributs hétérogènes et de qualité variable.

3.1.2 L'inférence sur des attributs imparfaits

Le RB est un modèle conçu pour raisonner sur des attributs incertains et incomplets. Il est composé d'une description qualitative de leurs dépendances, 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 1^{er} étape de paramétrage permet de représenter les connaissances *a priori* pour chaque VA sous la forme d'une table de probabilités conditionnelles. 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* même pour les variables non-observées.

Expliquons sur un exemple très simplifié le mécanisme d'inférence du réseau de la figure 1, un réseau destiné à la classification des pronoms *it*. La 1^{er} étape de paramétrage du réseau, permet de calculer les valeurs *a priori* des probabilités. Sur l'analyse des fréquences d'un corpus d'entraînement ou à partir de l'estimation d'un expert, nous établissons *a priori* qu'environ un tiers des pronoms *it* du corpus sont impersonnels, P(Pronoun=Impersonal)=0,3. Un lien d'influence relie les variables `Pronom` et `Lappin_Rules`, indiquant qu'un *it* a d'autant plus de chance d'être reconnu par une règle de (Lappin & Leass, 1994) qu'il est impersonnel. De même, les liens entre les variables `Pronoun` et `Paice_Rules` d'une part, `Pronoun` et `Start_Sentence` d'autre part indiquent respectivement qu'un *it* a d'autant plus de chance d'être reconnu par une règle de (Husk & Paice, 1987) et d'être en début de phrase qu'il est impersonnel. L'arc (`Start_Sentence, Paice_Rules`) unit les deux variables, car, toujours au regard du corpus d'entraînement ou de l'estimation de l'expert, elles ne sont pas indépendantes. La fiabilité de la règle de (Husk & Paice, 1987) reconnaissant une séquence est augmentée si la séquence est située en début de phrase. Cette influence est mesurée par la table de probabilités conditionnelles associée au noeud `Paice_Rules` de la figure 1.

Une fois l'ensemble des probabilités conditionnelles déterminé, l'étape d'inférence débute.

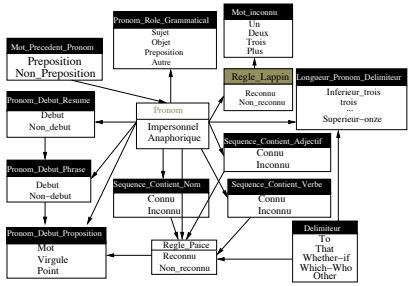

FIG. 2 – Un Réseau Bayésien pour la classification des pronoms *it* impersonnels

Considérons par exemple la phrase *It is well documented that treatment of serum-grown*.... Nous appliquons les règles de (Lappin & Leass, 1994) et les règles de (Husk & Paice, 1987) sur cette séquence. Aucune règle de (Lappin & Leass, 1994) ne reconnaît la séquence, nous posons P(Lappin_Rules = No_Match)=1. Une règle de (Husk & Paice, 1987) la reconnaît, nous posons P(Paice_Rules = Match)=1 et comme la séquence se situe en début de phrase nous posons aussi P(Start_Sentence = Start)=1. En représentant graphiquement l'indépendance conditionnelle des VA, le RB permet de compacter la loi jointe globale. A l'aide des probabilités conditionnelles fournies en paramètres nous pouvons inférer la probabilité qui nous intéresse : P(Pronoun=Impersonal|Lappin_Rules=No_Match, Start_Sentence=Start, Paice_Rules=Match)

Du fait qu'une règle de (Husk & Paice, 1987) a reconnu la séquence et que l'occurrence se trouve en début de phrase, le réseau infère une probabilité de 38,9% pour l'occurrence d'être impersonnelle. Nous pouvons modifier cette conclusion en ajoutant d'autres variables au réseau ou en raisonnant avec des observations incertaines ou manquantes. On peut par exemple indiquer que la fiabilité de l'observation est inférieure à 100% et poser P(Lappin_Rules=No_Match)=0,9 pour tenir compte de l'incomplétude des règles de (Lappin & Leass, 1994).

4 1^{re} expérience : l'identification des pronoms impersonnels

4.1 Le protocole expérimental

L'objectif de cette première expérience est de valider notre modèle (on trouvera dans (Weissenbacher & Nazarenko, 2007) une description précise du système développé et une analyse plus complète des résultats obtenus). Nous avons mesuré les performances du Classifieur Bayésien (CB) de la figure 2[2], ainsi que celles du classifieur bayésien naïf (CBN) associé[3], puis nous les avons comparées avec celles des systèmes de l'état de l'art.

[2]Les attributs représentant le fait qu'une règle de (Lappin & Leass, 1994) ait reconnu une séquence sont colorés en gris, en blanc ceux qui correspondent aux règles de (Husk & Paice, 1987), enfin en noir les attributs de (Litran *et al.*, 2004) et (Evans, 2001). Le noeud de prédiction est le noeud `Pronom`, au centre. Il estime la probabilité pour une occurrence donnée de pronom d'être impersonnel ou anaphorique.

[3]Le classifieur bayésien naïf possède les mêmes attributs mais sa structure est différente : le noeud `Pronom` est lié à tous les noeuds et ces derniers ne sont liés à aucun autre.

Méthode	Résultats		
Règles De (Lappin & Leass, 1994)	88,11%	12,8	169,1
Règles De (Husk & Paice, 1987)	88,88%	123,6	24,2
Machine à Vecteurs de Support	92,71%	-	-
Classifieur Bayésien naïf	92,58%	74,1	19,5
Classifieur Bayésien	**95,91%**	**21,0**	**38,2**

TAB. 1 – Résultats des prédictions (Exactitude/Faux Positifs/Faux Négatifs)

Nous avons travaillé sur un corpus de résumé d'articles de génomique construit à partir de la base *Medline* interrogée avec les mots clés *bacillus subtilis, transcription factors, Human, blood cells, gene and fusion*. Nous en avons extrait 11 966 résumés (environ 5 millions de mots) où nous avons identifié 3347 occurrences du pronom *it*. Deux annotateurs humains ont classé chaque occurrence du pronom soit comme anaphorique soit comme impersonnelle. L'accord des annotateurs fut entier après discussion.

Notre corpus étant de taille moyenne, nous avons procédé à une validation croisée pour valider nos résultats. Nous sélectionnons aléatoirement 2/3 du corpus pour calculer les probabilités conditionnelles *a priori*. Nous appliquons ensuite notre CB, ainsi que le CBN, paramétrés grâce à ces probabilités sur le tiers restant. Nous réitérons 20 fois ces opérations pour obtenir une moyenne des performances de chaque système sur le corpus.

4.2 Résultats

Le tableau 1 résume les moyennes des résultats (en exactitude) obtenus par les systèmes de l'état de l'art décrits plus haut[4] et celles des deux classifieurs. Ces résultats montrent que le CB produit une meilleure classification que les autres systèmes, notamment les systèmes à base de règles. Ces résultats valident notre modèle : le CB exploite tous les attributs pertinents et corrige le bruit d'un attribut par la fiabilité des autres. Privé des relations de dépendance entre les attributs, le CBN ne bénéficie pas du mécanisme de correction et surestime leurs fiabilités. Les systèmes à base de règles sont quant à eux entièrement assujettis à la fiabilité des attributs. Les résultats confirment les craintes soulevées dans la section 3.1.1 : on obtient un faible rappel pour les règles de (Lappin & Leass, 1994) et une mauvaise précision pour celles de (Husk & Paice, 1987).

5 2^{nde} expérience : la résolution des anaphores

Assurés des bonnes performances de notre modèle sur la distinction des pronoms impersonnels, nous proposons un classifieur bayésien pour la résolution d'anaphore.

[4]Nous avons ajouté le score du SVM obtenu par (Litran *et al*., 2004) sur un corpus de génomique similaire pour comparer leurs résultats aux nôtres. Les attributs utilisés par les SVM sont ceux défini par les auteurs. Les valeurs FP et les FN n'ont pas été publiées.

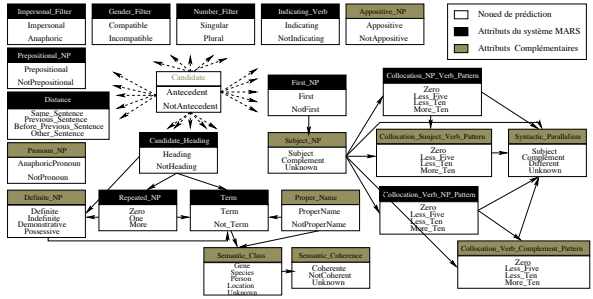

FIG. 3 – Un réseau Bayésien pour la classification des antécédents

5.1 Un classifieur bayésien pour la résolution des anaphores

Nous avons utilisé le système MARS (Mitkov, 2002) comme système de référence pour notre évaluation. Ce système repose sur des indices de surface pour trouver l'élément le plus saillant dans le discours qui précède une occurrence donnée de pronom. Cet élément est celui qui a la plus forte probabilité d'être l'antécédent du pronom. Nous avons ré-implémenté le système en utilisant le même prétraitement des textes que dans notre système bayésien[5] de manière à comparer uniquement les algorithmes des deux systèmes (choix des attributs et mécanisme de prise de décision).

Pour réaliser notre classifieur bayésien (voir figure 3[6]), nous avons conservé tous les indices approchés de MARS (noeuds coloriés en noir sur la figure) mais nous avons ajouté une série d'autres indices (en gris sur la figure) qui sont également pertinents pour le calcul de la saillance et qui sont proposés par plusieurs travaux de l'état de l'art. De notre point de vue, il est en effet utile d'avoir à la fois les indices et les connaissances linguistiques qu'ils approchent. 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 GN de la phrase est très souvent le sujet du verbe) qui peut confirmer ou infirmer l'hypothèse du rôle grammatical.

En suivant un protocole expérimental identique à celui de la section précédente sur le même corpus, nous avons réalisé la résolution avec 4 systèmes différents. Trois systèmes servent de comparaison : le système *Aléatoire* qui choisit un antécédent au hasard dans la liste des candidats, le système *Premier GN* qui sélectionne toujours le premier GN de la phrase précédant le pronom comme antécédent et le système MARS. Le dernier système est le classifieur bayésien (CB) que nous cherchons à évaluer.

Pour les trois derniers systèmes, nous donnons deux mesures différentes des performances, un taux de succès strict et partiel[7]. Le taux de succès est strict lorsque l'antécédent exact a été

[5] Nous avons utilisé dans les deux cas les analyses produites par la plate-forme d'annotation OGMIOS (Derivière *et al*., 2006).

[6] 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. Ce noeud `Candidate` est lié à tous les noeuds du réseau.

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

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

annoté par le système et partiel lorsque seule une partie de l'antécédent à été annotée. En raison des erreurs de l'analyse syntaxique en constituants sur laquelle la liste des candidats est calculée, certains GN candidats ne sont identifiés que partiellement ou font défaut. Les performances de nos systèmes ne peuvent atteindre 100%, la dernière colonne donne les scores maximum possibles pour la résolution.

System	**Results**	
	Strict	*Partial*
Aléatoire	6%	-
Premier GN	36.3%	51%
MARS	26.7%	43%
Classifieur Bayésien	44.0%	61%
MAX	93.3%	97.8%

TAB. 2 – 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 désambiguisation entre différents candidats. Considérons les phrases *[A grpE heat-shock gene]$_1$ was found by sequencing in [the genome of the methanogenic archaeon Methanosarcina mazei S-6]$_2$. [It]$_1$ is the first example of grpE from the phylogenetic domain Archaea*. Le système MARS attribue des scores identiques pour les candidats 1 et 2 et ne les départage que grâce à l'heuristique du candidat le plus récent, ce qui le conduit à choisir le candidat 2. Le classifieur CB évite cette erreur. La connaissance du sujet et du type sémantique *gène* du candidat 1 augmente à 0.73 sa probabilité d'être l'antécédent du pronom et lève l'ambiguité.

Une analyse détaillée des erreurs du CB montre les limites de notre analyse de la saillance. 47% des erreurs sont dues à un calcul erroné de l'élément saillant : le système ne retrouve pas ce que l'annotateur humain juge «intuitivement» être l'élément saillant parce qu'un nombre plus important d'indices favorisent un candidat différent de l'élément saillant auquel le classifieur associe la plus grande probabilité d'antécédence. Dans 21% des cas, le systèmes trouve bien l'élément qui paraît saillant à l'annotateur humain mais cet élément n'est pas l'antécédent, ce qui met en cause soit notre définition de la saillance soit son rôle dans la résolution de l'anaphore. Dans l'exemple suivant *[Amino acid sequence analysis]$_1$ of [the 33-kDa protein]$_2$ revealed that it is a sigma factor, sigma E*. l'élément le plus saillant est le candidat 1 et il est choisi comme antécédent par le système, une décision qui viole les connaissances du domaine, un facteur sigma est une protéine, des connaissances qu'il faut prendre en compte pour choisir le candidat 2 comme antécédent. Les erreurs restantes proviennent des imperfections des prétraitements linguistiques : principalement des erreurs de segmentation en phrase et de l'analyse syntaxique incorrecte qui ne permet pas de repérer tout les GN candidats.

6 Conclusion

Les réseaux bayésiens présentent un véritable intérêt pour les nombreuses tâches de classification du TAL. Ce modèle permet de dépasser l'opposition historique des systèmes à base de connaissances linguistiques et d'indices de surface. De fait, cette opposition apparaît infondée :

les connaissances linguistiques sont nécessaires mais souvent indisponibles et peu 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. Tout l'enjeu consiste selon nous à raisonner sur l'ensemble des connaissances et indices disponibles à un moment donné mais en tenant compte de leur relative fiabilité dans le processus de décision.

Nous avons ensuite validé notre modèle sur le problème de la résolution des anaphores en proposant deux classifieurs, le premier pour distinguer les pronoms impersonnels et anaphoriques, le second pour le choix de l'antécédent. Les résultats de nos classifieurs sont supérieurs à ceux des systèmes de l'état de l'art.

Actuellement seule une expertise linguistique rend compte de la structure des deux classifeurs que nous avons présentés. Nous envisageons de tester les mécanismes permettant d'apprendre la structure même du réseau. Comparer notre structure avec une structure apprise automatiquement devrait permettre de vérifier et d'enrichir la structure du CB actuelle.

Références

DAGAN I. & ITAI A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of COLING'90*, p. 3 :330–332.

DERIVIÈRE J., HAMON T. & NAZARENKO. A. (2006). A scalable and distributed nlp architecture for web document annotation. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, p. 56–67.

EVANS R. (2001). Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, **16**, 45–57.

HUSK G. & PAICE C. (1987). Towards the automatic recognition of anaphoric features in english text : the impersonal pronoun it. *Computer Speech and Language*, **2**, 109–132.

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.

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

LITRAN J. C., SATOU K. & TORISAWA K. (2004). Improving the identification of non-anaphoric it using support vector machines. In *Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, p. 58–61.

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

PONZETTO S. & STRUBE M. (2006). Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of EACL'06*., p. 143–146.

WEISSENBACHER D. & NAZARENKO A. (2007). A bayesian classifier for the recognition of the impersonal occurrences of the *it* pronoun. In *Proceedings of DAARC'07*.