Comprendre les effets des erreurs d'annotations des plates-formes de TAL

Davy Weissenbacher* — **Adeline Nazarenko****

* *COIN Laboratory, Toyota Technological Institute*
2-12-1 Hisakata, Tempaku - Nagoya – 468-8511 JAPAN
Davy.Weissenbacher@toyota-ti.ac.jp

** *LIPN-Université Paris-Nord*
99 av. J-B. Clément – 93430 Villetaneuse FRANCE
nazarenko@lipn.univ-paris13.fr

RÉSUMÉ. *Les résultats des analyses des outils de TAL sont souvent des annotations qui caractérisent les séquences des textes analysés. Ces annotations sont dites erronées lorsque leurs valeurs diffèrent des valeurs attribuées par un expert. Des architectures innovantes sont aujourd'hui proposées pour annoter et corriger simultanément des annotations de différentes catégories. Mais la complexité des calculs requis limite le nombre d'annotations réellement intégrées. Nous étudions ici une alternative conservant l'architecture standard de traitement en cascade. Nous montrons, sur la résolution des anaphores, que la modélisation de l'incertitude des annotations permet de limiter l'impact des annotations erronées, d'intégrer toutes les annotations nécessaires à l'inférence et de différer la révision des erreurs à un post-traitement.*

ABSTRACT. *Outputs of NLP tools can be regarded as sets of annotations of predefined characters included in a processed document. These annotations are said erroneous if they disagree with the annotations given by the experts. Innovative architectures have been proposed to annotate and revise annotations by processing errors in various levels of linguistic annotations simultaneously. But the computational complexity limits the number of the annotations they can effectively handle. We study an alternative way which keeps the standard cascade pipeline architecture. We show, on the anaphora resolution, that modeling the reliability of the annotations helps to attenuate the impact of the noisy ones, to integrate into the pipeline all annotations needed to fulfill the tasks, and to postpone the correction of errors in a latter stage.*

MOTS-CLÉS : *anaphore, réseau bayésien, annotation erronée, plate-forme d'annotation.*

KEYWORDS: *anaphora, bayesian network, erroneous annotation, annotation pipeline.*

TAL. Volume 52 – n°1/2011, pages 161 à 185

1. Introduction

Selon la définition de Grishman (1997), une annotation est un ensemble de propriétés prédéfinies qui caractérise une séquence continue ou discontinue d'un document, ou le document lui-même. Ces propriétés sont souvent de différents types (lexicales, syntaxiques, sémantiques ou ontologiques, par exemple). L'annotation d'un document relève d'une approche pragmatique du traitement automatique des langue (TAL) : le but n'est pas de produire une analyse complète du sens d'un document mais une représentation partielle de son contenu en vue de faciliter son traitement par d'autres applications, par exemple de résumé automatique, de traduction ou encore de question-réponse.

Une annotation est dite erronée lorsque sa valeur diffère de la valeur attendue, qui peut être déterminée par une expertise humaine[1] ou provenir de corpus étalons. Une annotation peut être erronée pour différentes raisons. Outre le cas où elle est simplement absente, ses frontières peuvent être mal positionnées. On observe souvent ce problème avec les entités nommées qui peuvent être imparfaitement reconnues par l'étiqueteur qui y inclut trop ou pas assez de mots. Lorsque l'annotation est correctement placée, c'est la valeur d'une de ces propriétés qui peut ne pas avoir pu être calculée, être incorrecte ou encore mal spécifiée. Ainsi sans information additionnelle sur le contexte, l'entité nommée *11 septembre 2001* peut être considérée comme une simple date et comme un événement [2].

La plupart des chaînes de traitement de TAL disponibles à l'heure actuelle sont conçues comme des plates-formes d'annotation et reposent sur une architecture incrémentale. Différents modules d'analyse sont appliqués séquentiellement aux documents. Chaque module ajoute un nouveau niveau d'annotation sur le texte analysé et il repose, pour ce faire, sur les annotations produites par les modules de la plate-forme précédemment appliqués sur le même texte. Par exemple, l'annotation des mots est nécessaire pour segmenter les phrases d'un document, de même que l'annotation des mots et des phrases est souvent requise pour l'étiquetage grammatical (*POS tagging*). Même si l'on sait qu'aucun module d'analyse ne réalise parfaitement l'annotation d'un corpus générique, peu de plates-formes sont conçues pour prendre en compte les erreurs d'annotations dans le processus d'analyse. En raisonnant sur des annotations erronées imprévues, les modules ajoutent ainsi de nouvelles annotations erronées sur les anciennes.

Cette dernière décennie, la communauté du TAL s'est cependant intéressée à ce problème des annotations d'entrée incertaines ou erronées. Plusieurs stratégies ont été proposées pour limiter leur impact sur les performances globales des plates-formes d'annotation. Ces stratégies ont mené à une réflexion plus fondamentale sur la ma-

1. Nous ne prenons pas en considération ici le cas où les annotations ne font pas consensus entre les experts, car il s'agit là d'un problème de définition de l'annotation et non d'une erreur à proprement parler.

2. Notons que les deux valeurs `Événement` et `Date` peuvent être simultanément vraies et que l'auteur peut vouloir entretenir cette ambiguïté dans la suite du texte.

nière d'exploiter les dépendances qu'entretiennent les annotations linguistiques entre elles et ont conduit à des projets d'architectures intégrées prometteuses mais difficiles à mettre en œuvre en raison d'une explosion combinatoire qui limite leurs performances.

Cet article étudie une alternative où l'architecture incrémentale peut être conservée grâce à la modélisation de l'incertitude de toutes les annotations d'entrée utiles. Les expériences ont été réalisées sur un problème de TAL particulier, la résolution des pronoms anaphoriques *it* dans les textes de génomiques anglais. L'anaphore est une relation linguistique entre deux entités textuelles. Elle est définie lorsqu'une unité textuelle, l'anaphore, fait référence à une unité précédente, l'antécédent. La résolution des anaphores consiste donc à établir une relation entre les anaphores et leurs antécédents. Cette tâche d'annotation est complexe. Elle fait appel à des connaissances de natures variées qui sont parfois elles-mêmes difficiles à calculer et dont la fiabilité peut être douteuse. La résolution des anaphores pronominales est donc un cas intéressant pour l'étude de l'impact des annotations erronées sur une plate-forme de TAL.

La section 2 analyse les stratégies proposées pour limiter l'impact des annotations erronées sur les performances des plates-formes et les conséquences sur leurs architectures. Dans la section 3, nous présentons le modèle des réseaux bayésiens utilisé pour nos expériences sur la résolution des anaphores pronominales. La section 4 décrit les expériences que nous avons réalisées, leurs protocoles et les résultats obtenus.

2. Une réalité incontournable : les erreurs d'annotations

Les tâches d'annotation consistant à étiqueter certaines séquences d'un document se modélisent bien comme des tâches de classification, mais le bruit des données ainsi que les modalités de l'apprentissage artificiel pour les problèmes de TAL viennent perturber le mécanisme d'induction.

Les exemples utilisés pour apprendre un module d'annotation sont souvent eux-mêmes bruités. Le bruit peut porter sur l'étiquette des exemples d'apprentissage ou sur la valeur des attributs qui les décrivent. Lorsque les corpus d'acquisition font l'objet d'un étiquetage manuel attentif, on peut supposer qu'il y a peu d'erreurs parmi les étiquettes des exemples, mais il en va tout autrement de la qualité de leurs attributs. En effet, ces attributs ne sont souvent pas directement observés dans le texte. Ils sont calculés par d'autres modules de classification, des modules de TAL dont on ne peut pas attendre des performances parfaites. Le cadre d'apprentissage des classifieurs doit donc prévoir la présence de plusieurs attributs erronés pour chaque exemple. Les études théoriques de ce domaine montrent qu'il est possible, en respectant les conditions fixées par le cadre PAC (Valiant, 1984), de garantir, avec une certaine probabilité, une qualité de classification fixée des modules en sortie d'apprentissage. Ces conditions portent sur le nombre d'exemples, la richesse du langage de description des exemples et du temps disponible pour l'apprentissage. Malheureusement, l'estimation de la qualité de l'apprentissage est difficile en TAL. Le cadre PAC en présence d'at-

tributs erronés a déjà été étudié théoriquement par Goldman et Sloan (1995) et évalué sur différents jeux de données réels dans (Zhu et Wu, 2004) mais les hypothèses faites sur le cadre d'apprentissage sont encore trop fortes pour le TAL.

L'ensemble des attributs nécessaires pour apprendre le concept cible est souvent inconnu et difficile à identifier *in abstracto*. Le choix des attributs discriminants est guidé par les exemples trouvés dans un corpus d'acquisition qui reflète rarement de manière exhaustive les séquences à annoter dans le corpus de test (Wen-tau, 2005). Il faut par conséquent envisager que le concept cible puisse ne pas être apprenable.

De surcroît, même si l'on sait un attribut pertinent pour l'apprentissage, son rôle peut varier d'un corpus à l'autre. Un attribut fortement corrélé au concept cible pour un corpus d'un domaine est un très bon indicateur du concept, mais il peut se révéler trop général ou même erroné pour un corpus d'un autre domaine. Par exemple le point, qui est un marqueur fiable car directement observable, indique avec une grande probabilité la fin d'une phrase dans un roman, mais il doit être utilisé avec plus de précautions dans l'analyse des articles de génomique où il est aussi utilisé dans les références bibliographiques, les noms de gènes, les unités de mesure, etc.

Enfin, de par la nature même de la langue naturelle, les annotations sont fortement dépendantes les unes des autres. Or, la plupart des cadres d'analyse supposent que les annotations sont, sinon indépendantes, du moins faiblement dépendantes. L'apport des dépendances, dans l'estimation de la qualité de l'apprentissage comme dans les procédures de détection et de correction du bruit, est donc encore mal connu.

2.1. *Faire au mieux avec des annotations erronées*

La complexité d'une annotation peut être définie par le nombre d'annotations nécessaire à son calcul : plus le nombre d'annotations d'entrée requis est important, plus l'annotation produite est complexe. Par exemple, seuls les mots, la ponctuation et parfois les entités nommées (EN) sont nécessaires pour segmenter les phrases d'un texte alors que la résolution d'une anaphore s'appuie non seulement sur un découpage en phrases mais aussi sur des annotations syntaxiques, sémantiques et pragmatiques. Comme une annotation complexe arrive tard dans la chaîne d'analyse du texte, le risque d'utiliser des annotations d'entrée erronées dans son calcul est très important et sa fiabilité d'autant plus faible. Ainsi une segmentation incorrecte des phrases peut-elle conduire à une résolution erronée des anaphores et même en rendre le calcul impossible.

Devant la difficulté d'obtenir des annotations fiables, une première méthode consiste à ignorer les annotations les plus complexes, qui sont aussi les plus bruitées, et à travailler uniquement avec les annotations les plus simples. Mitkov *et al.* (2001) montrent que cette solution est devenue dominante pour la résolution des anaphores au début des années 2000 avec le besoin de systèmes effectifs et robustes. Cette solution catégorique limite, selon nous, les performances des systèmes puisque les annotations complexes modélisent les connaissances linguistiques qui sont sou-

vent utiles pour discriminer les séquences ciblées. On sait notamment que la plupart des relations anaphoriques ne peuvent être résolues sans connaissance sémantique ou pragmatique.

Bunescu (2008) choisit de conserver les annotations complexes à condition que leur fiabilité puisse être connue et soit accessible à l'ensemble des modules de la plate-forme. Il montre comment sauvegarder le rang de chaque annotation tel qu'il est attribué initialement par un module en réencodant l'ordre relatif de différentes annotations par une distribution de probabilités. Les résultats expérimentaux démontrent une amélioration des performances de la plate-forme modifiée par rapport à celles de la plate-forme classique.

2.2. *Prévenir l'utilisation des annotations erronées*

Une stratégie opposée consiste à détecter et à corriger automatiquement les erreurs d'annotations au moyen de post-traitements. La méthode dominante recherche les annotations atypiques (Zhu et Wu, 2004) et, en TAL, on peut citer Dickinson (2005), qui détecte dans les corpus étalons les variations de suites d'étiquettes morphosyntaxiques attribuées à des séquences identiques, ou Brill (1995), qui recourt à l'apprentissage pour constituer un ensemble de règles logiques dédiées à la correction des annotations d'un module initial. Ces méthodes demeurent locales, en ce sens que seules les annotations produites par le module sont utilisées dans son post-traitement. Hirakawa et Yoshimura (2000) emploient une méthode de révision plus sophistiquée. L'échec d'un analyseur syntaxique signale que le jeu d'étiquettes morphosyntaxiques d'entrée était probablement incorrect. Dans ce cas, l'analyseur doit essayer d'autres jeux d'étiquettes jusqu'à ce qu'une analyse réussisse. L'originalité de cette méthode est qu'elle exploite l'interdépendance entre les annotations du niveau syntaxique.

2.2.1. *Vers des architectures intégrées*

La conception de plates-formes à l'architecture intégrée généralise l'idée d'une révision des erreurs d'annotations grâce à leur dépendance. Ces plates-formes calculent simultanément un ensemble d'annotations de catégories différentes dans une tâche de multiclassification. L'intégration des annotations au sein d'un modèle unique permet d'exprimer les relations qu'elles entretiennent et de les faire intervenir comme des contraintes globales que le classifieur doit satisfaire lors de sa recherche de la solution optimale (Chang *et al.*, 2008). L'apparition d'une annotation erronée dans une relation avec une autre annotation peut rendre cette dernière incohérente et ansi forcer le système à rechercher une autre valeur pour cette annotation. Le modèle proposé par Roth et Wen-tau (2002) cherche à étiqueter les entités nommées et leurs rôles sémantiques. Si le système considère *J.F. Kennedy* à la fois comme un nom de lieu et comme l'agent de la relation *être_assassin_de*, il détecte une contradiction et doit réviser l'annotation de l'entité nommée ou de la relation. Collobert et Weston (2008) analysent l'amélioration permise sur une tâche d'extraction des rôles sémantiques avec une architecture capable de résoudre six tâches d'annotation simultanément.

2.2.2. *Une alternative reposant sur une architecture incrémentale*

Le succès des plates-formes intégrées provient de la révision des annotations erronées mais ce succès à un coût : un important effort d'ingénierie est nécessaire pour adapter et intégrer dans la plate-forme des outils existants, généralement conçus comme autonomes pour des architectures incrémentales. Surtout, la révision n'est possible dans ces architectures que si toutes les tâches d'annotation sont réalisées simultanément. Or, aucun modèle d'inférence existant ne peut répondre à cette dernière contrainte : le nombre de variables et de relations impliquées dans l'inférence pour chaque annotation qui est ajoutée entraîne des calculs dont la complexité est vite prohibitive, ce qui limite en pratique le nombre d'annotations que l'on peut intégrer au sein de ces architectures.

Une alternative possible consiste à fournir à chaque module toutes les annotations d'entrée nécessaires pour le calcul de son annotation cible et à différer la révision des annotations dans un post-traitement final. Le fait de retarder la révision des annotations permet tout d'abord de s'appuyer sur un plus large ensemble d'annotations, et donc de s'appuyer sur les annotations les plus importantes, puis de cibler la révision en fonction de l'application visée. Cette approche repose sur deux hypothèses. Il faut dans un premier temps que les annotations d'entrée les plus complexes puissent être intégrées avec bénéfice dans le calcul de l'annotation cible. Comme elles ne sont que partiellement connues en cours de traitement, il est nécessaire de tenir compte de leur incertitude. Il faut ensuite développer des modules de post-traitement efficaces.

Nous discutons dans cet article de la première hypothèse sur laquelle repose cette alternative et qui doit être établie avant d'aborder le problème de la révision. Calculées avec des annotations d'entrée erronées, les annotations complexes seront, elles aussi, erronées. Alors, est-ce que prendre en compte toutes les annotations améliore le calcul de l'annotation cible ? Plus précisément, comment modéliser leur incertitude et en faire le meilleur usage possible dans l'inférence, avant l'étape de révision ? Quel est l'impact de l'utilisation des annotations complexes les plus erronées sur les performances du module ? Existe-t-il un seuil de qualité minimal requis pour que l'intégration des annotations soit profitable ?

Cet article explore ces questions sur un problème particulier : celui de la résolution des anaphores pronominales. Il apporte des éléments de réponse grâce à l'analyse détaillée des résultats obtenus par un module de résolution des pronoms *it* anaphoriques lors d'une série d'expériences réalisées sur des textes de biologie génomique et dans le contexte réel du projet Européen *ALVIS*[3].

3. La résolution des pronoms *it* : présentation de l'approche

Nous proposons de modéliser l'ensemble des connaissances nécessaires à une tâche d'annotation sous la forme d'un réseau bayésien. Ce modèle, qui apparaît de

3. Références du projet : IST-1-002068-STP.

plus en plus comme un modèle adapté au TAL, permet de donner une représentation unifiée d'un ensemble de connaissances hétérogènes du fait de leur nature, de leur mode de calcul et de leur fiabilité. Nous décrivons ici la résolution des anaphores pronominales comme une tâche de classification des syntagmes candidats à l'antécédence. Nous présentons le réseau bayésien que nous avons construit et les annotations sur lesquelles repose le classifieur.

3.1. *Les réseaux bayésiens, un modèle adapté au TAL*

3.1.1. *L'expression de l'incertitude*

Les connaissances manipulées sont incertaines lorsqu'il est impossible d'affecter de manière univoque la valeur des attributs décrivant les objets considérés. Un monde unique ne pouvant être décrit, plusieurs mondes possibles doivent être envisagés dans l'inférence. Plus il y a d'attributs inconnus plus le nombre de mondes devant être considérés est important. Dans certaines situations, toutefois, il est possible de mesurer un degré de confiance pour certaines valeurs et donc de privilégier un sous-ensemble de mondes possibles. Différentes mesures de confiance ont été proposées pour définir au mieux ce sous-ensemble, parmi lesquelles les probabilités, les mesures duales de possibilité et de nécessité ou encore les mesures de croyance et de plausibilité. Néanmoins la mesure de probabilité est la mesure de confiance la plus précise quand il y a suffisamment d'informations disponibles pour établir des distributions de probabilités fiables (Halpern, 2003). Comme c'est généralement le cas en TAL, où l'on dispose de corpus annotés ou d'expertises, nous avons travaillé dans le cadre de la théorie des probabilités.

3.1.2. *Présentation du modèle*

Pour raisonner à partir d'annotations d'entrée probabilisées nous utilisons le modèle des réseaux bayésiens (RB). Ce modèle repose sur l'idée intuitive de représenter graphiquement les relations d'influence entre plusieurs variables. Un arc unit deux variables lorsqu'il existe une relation d'influence entre elles et l'absence d'arc marque leur indépendance. Le sens de l'arc précise le sens de l'influence : est-ce que A influence B ou l'inverse ? L'adjonction au graphe des distributions des probabilités mesure la « force » de cette influence.

Formellement, un RB est composé d'une description qualitative des dépendances d'un ensemble de variables aléatoires (VA), encodé par un graphe orienté sans circuit où chaque VA est associée à un nœud du graphe, ainsi qu'une description quantitative de leurs dépendances, mesurées par un ensemble de probabilités conditionnelles. Chaque VA modélise une connaissance utilisée par le système, qui est donnée par une annotation d'entrée.

Une première 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, dite d'inférence, consiste à réviser les valeurs de certaines VA à par-

tir d'observations faites en corpus et des probabilités conditionnelles du réseau. Ces nouvelles informations sont propagées au travers du réseau et permettent de réviser les valeurs *a priori* des VA qui sont dépendantes, y compris pour les variables non observées (Weissenbacher, 2008).

3.1.3. *Limitations du modèle*

Le formalisme des RB est un modèle expressif mais il souffre de deux contraintes principales. D'une part, la structure du réseau doit être connue et rester figée durant toute l'inférence. D'autre part, le réseau ne doit pas contenir de circuit en raison de la sémantique causale associée aux arcs : si A est cause de B, B ne peut pas causer A. Différentes extensions du modèle ont été considérées pour lever ces contraintes, notamment les réseaux bayésiens dynamiques, les *conditional random fields* (CRF) ou encore les réseaux d'inférence, mais au prix de calculs plus complexes qui ne sont pas immédiatement nécessaires pour nos expériences.

3.1.4. *Application au TAL*

Le modèle des RB a récemment suscité un certain intérêt en TAL (Goyal *et al.*, 2008). Cette approche probabiliste a le mérite d'unifier dans une représentation toutes les annotations pertinentes pour une tâche donnée, par exemple le contenu textuel mais aussi la structure logique ou encore des informations statistiques (Denoyer et Gallinari, 2004). Cette unification est utile puisque l'annotation d'un document suppose d'exploiter conjointement un grand nombre d'annotations d'entrée hétérogènes. Cette unification permet également d'exprimer les dépendances entre les attributs décrivant les données à annoter et, ainsi, de corroborer les annotations les moins fiables, par l'observation, des plus sûres. Cette propriété des RB a été mise en évidence dans nos travaux préliminaires (Weissenbacher et Nazarenko, 2007) sur la reconnaissance des pronoms impersonnels, une tâche qui, en comparaison de la résolution des anaphores, fait appel à des annotations peu complexes et moins nombreuses.

Enfin, il est possible d'apprendre les probabilités *a priori* ou la structure du réseau sur un corpus d'acquisition d'un nouveau domaine. Cet apprentissage garantit que les annotations d'entrée engagées dans l'inférence du module seront toujours discriminantes. Cette propriété du modèle a été démontrée sur le problème du filtrage des *pourriels* dans (Sahami *et al.*, 1998) et sur deux tâches d'extraction d'information pour l'apprentissage de la structure (Bouckaert, 2002).

3.2. *Un classifieur bayésien pour la résolution des anaphores*

Les nombreux systèmes robustes de résolution d'anaphores existants exploitent plus ou moins le même ensemble d'indices de surface pour sélectionner l'antécédent. Parmi ces systèmes nous avons choisi le système MARS[4] qui est désormais considéré

4. Pour ce travail, nous avons utilisé la première version du système MARS (Mitkov, 2002).

comme une référence. Ce choix est justifié en détail dans (Weissenbacher, 2008). Le fait que son architecture modulaire puisse facilement être enrichie par de nouvelles annotations en fait un bon prototype pour nos expériences. Notre but n'est en effet pas de proposer une stratégie de résolution d'anaphores innovante mais de comparer les performances d'un système utilisant uniquement des annotations simples avec les performances d'un système similaire disposant de connaissances complexes mais potentiellement erronées. L'analyse de leurs résultats devrait mettre en lumière la contribution des annotations complexes et incertaines à une tâche de classification.

Le système MARS repose sur une hypothèse simple : le syntagme nominal (SN) le plus saillant dans le contexte d'un pronom anaphorique est souvent l'antécédent. En choisissant le SN le plus saillant, dont la mesure de la saillance est bien plus facile à calculer automatiquement que l'antécédent, le système a une forte probabilité de résoudre l'anaphore correctement. Pour associer une saillance à un candidat, un poids est attribué aux différentes annotations qui caractérisent le candidat. Le candidat qui obtient le plus haut score est retenu comme antécédent. En cas d'égalité entre plusieurs candidats, plusieurs heuristiques d'arbitrage sont appliquées séquentiellement pour choisir le meilleur candidat. L'implémentation du système et son évaluation sont détaillées dans (Mitkov, 2002).

Nous nous sommes inspirés de la stratégie de résolution du système MARS pour créer notre système. Pour chaque pronom d'un document notre système liste l'ensemble des SN qui apparaissent dans la fenêtre de recherche du pronom[5]. Le système calcule les valeurs des annotations qui caractérisent le pronom et les candidats de la liste. Ces valeurs sont ensuite affectées aux VA correspondantes d'un réseau bayésien qui estime la probabilité pour chaque candidat d'être l'antécédent. Le candidat qui obtient la meilleure probabilité est sélectionné. En cas d'égalité, les heuristiques d'arbitrage du système MARS sont employées pour départager les concurrents.

3.3. *Modélisation des connaissances en jeu dans la résolution des anaphores*

Pour réaliser notre système (figure 1) nous avons conservé la majorité des annotations d'entrée du système MARS (nœuds coloriés en noir sur la figure) que nous avons enrichies par les annotations complexes dont elles dépendent et qui sont utilisées dans différents systèmes de l'état de l'art (nœuds coloriés en gris sur la figure). La liste suivante décrit les informations d'entrée de notre RB, un classifieur qui identifie l'antécédent dans la liste des syntagmes nominaux candidats. Chaque information correspond à une annotation précise et est modélisée par un nœud du réseau.

5. Dans nos expériences, la fenêtre comporte trois phrases : la phrase contenant l'occurrence du pronom et les deux phrases précédentes.

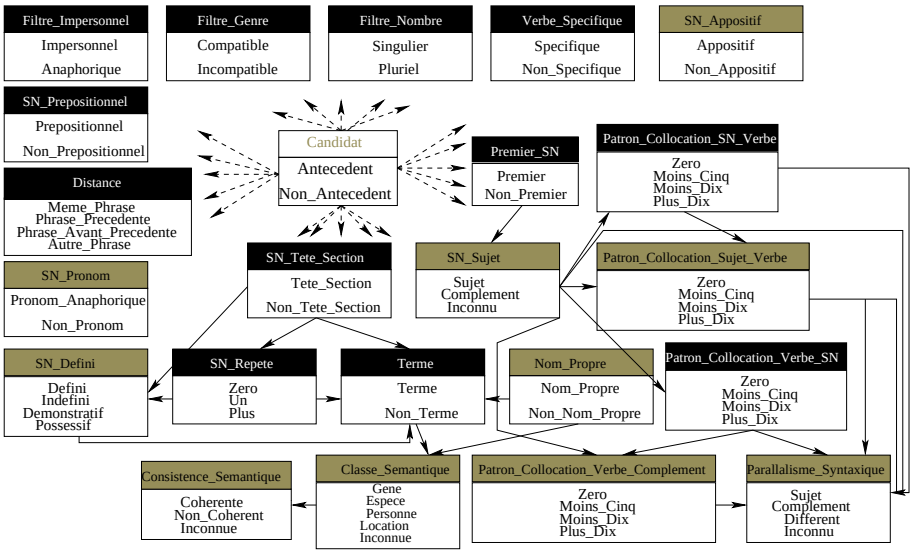

Figure 1. *Un classifieur bayésien pour la sélection de l'antécédent. Le nœud de prédiction* `Candidat` *est lié à tous les nœuds du réseau.*

– `Filtre_Genre/Nombre` : le pronom et l'antécédent doivent être morphologiquement compatibles. Comme c'est le neutre qui domine dans notre corpus d'évaluation, le filtre du genre se contente d'éliminer les pronoms féminins et masculins de la liste des candidats. Le filtre du nombre rejette tout SN contenant un nom étiqueté comme pluriel[6].

– `Filtre_Pronom_Impersonnel` : les pronoms impersonnels n'ayant pas d'antécédent, ils sont exclus de la résolution. Nous avons appliqué le classifieur décrit dans (Weissenbacher et Nazarenko, 2007) pour les identifier.

– `SN_Sujet` : le sujet est souvent l'élément le plus saillant. Lorsque le rôle syntaxique d'un SN n'est pas fourni par l'analyseur syntaxique de notre plate-forme de prétraitement que nous présentons ci-dessous, section 4.1, nous avons indiqué la valeur *Inconnu*.

– `Premier_SN` : le premier SN de la phrase est fréquemment le sujet. Nous retenons le premier SN de la phrase précédente si le pronom est en début de phrase, le premier SN de la phrase courante sinon.

6. Pour un SN composé, il est nécessaire de trouver la tête du syntagme pour spécifier le nombre. Mais l'heuristique usuelle de Cardie et Wagstaff (1999) s'est révélée très bruitée sur les textes de génomique. Cette heuristique consiste à prendre comme tête du candidat le premier nom ou le premier SN qui le compose s'il est précédé d'un article, et le dernier nom ou le dernier SN sinon. Nous lui avons préféré cette dernière heuristique plus simple.

– `Verbe_Specifique` : certains verbes mettent en valeur leur argument et donc augmentent leur saillance. Nous avons complété la liste de verbes spécifiques publiés dans (Mitkov, 2002) par une liste de verbes extraits manuellement de notre corpus d'entraînement.

– `SN_Repete` : plus un candidat est répété dans le document, plus il est saillant. Les répétitions sont calculées par des comparaisons entre les chaînes de caractères composant les têtes des syntagmes[7].

– `SN_Tete_Section` : les candidats figurant dans les titres de sections sont *a priori* saillants.

– `Patron_Collocation` : le choix d'un antécédent dépend aussi de son accointance sémantique avec le verbe que le pronom précède ou suit. Nous utilisons des patrons de collocations pour mesurer cette « accointance » : ils sont de la forme <*SN/pronom verbe*> et <*verbe SN/pronom*> où SN est le premier syntagme nominal ou le premier pronom *it* qui précède ou suit le verbe. Nous considérons la forme lemmatisée du verbe dans les patrons. Nous calculons les fréquences sur la régularité des têtes des candidats et sur l'ensemble du corpus d'entraînement. Nous complétons ces patrons de collocations par de nouveaux patrons de collocations plus fiables de la forme <*Sujet-verbe*> et <*verbe-complement*> : le candidat partage une collocation avec le pronom s'il a déjà été le sujet (*resp.* le complément) du verbe dont le pronom est lui-même le sujet (*resp.* le complément).

  Dans le contexte ci-dessous, le patron de collocations <*SN/verbe*> retrouve la collocation <*ORF/precede*> car le verbe *preceding* suit immédiatement le candidat *ORF*. Si le système connaît le rôle grammatical du candidat et du pronom, il retrouve la collocation attendue <*ORF-encode*> grâce au patron <*Sujet-verbe*>.

  > *We show that in B. subtilis the ORF preceding the rpsA homologue encodes a protein which is highly similar to the product of the E. coli mssA gene which is located upstream of rpsA.*

– `Parallelisme_Syntaxique` : les candidats dont les rôles syntaxiques sont connus et identiques à celui du pronom sont préférés aux autres candidats.

– `Terme` : les termes du domaine sont généralement plus saillants que les autres SN. Pour identifier les termes nous avons appliqué des ressources terminologiques du domaine : Gene Ontology et le MeSH (Derivière *et al.*, 2006), pour avoir une couverture importante.

– `SN_Defini` : les antécédents indéfinis sont moins saillants que les antécédents définis. Un SN est considéré comme indéfini s'il n'est pas précédé d'un article défini, possessif ou démonstratif.

– `SN_Prepositionnel` : nous considérons qu'un syntagme est prépositionnel s'il suit immédiatement une préposition. Nous identifions les prépositions à partir de l'analyse syntaxique.

– `Distance` : si l'analyse syntaxique n'est pas disponible ou insuffisamment fiable, la distance qui sépare un pronom de son antécédent est un indice intéressant à prendre

7. Nous déterminons la tête des syntagmes grâce à l'heuristique de (Cardie et Wagstaff, 1999).

en compte. Les analyses syntaxiques des phrases complexes n'étant souvent que partielles, nous avons simplifié la mesure en calculant le nombre de phrases séparant le pronom du candidat plutôt que le nombre de propositions.

– `Nom_Propre` : les noms propres sont des éléments du discours importants. Nous utilisons l'analyseur morphosyntaxique et le module de reconnaissance des EN de notre plate-forme pour les reconnaître.

– `SN_Pronom` : pour retrouver les pronoms *it* nous nous fions aux étiquettes de l'analyseur morphosyntaxique utilisé.

– `SN_Appositif` : un candidat qui apparaît dans une apposition, précise le nom auquel il se rattache et il est *a priori* moins saillant que ce dernier. Sont identifiés comme appositions les syntagmes entourés d'un signe de ponctuation double[8] et ne contenant pas de verbe.

– `Classe_Semantique` : pour le domaine de la génomique, les gènes et les protéines ont *a priori* un rôle plus important que celui des personnes et doivent être distingués. Nous identifions la classe sémantique des candidats en nous appuyant sur le module de reconnaissance des EN qui associe une classe sémantique aux entités nommées qu'il reconnaît.

– `Coherence_Semantique` : le pronom doit avoir une classe sémantique compatible avec celle de son antécédent. Pour vérifier cette cohérence nous généralisons les patrons de collocations du type <*sujet-verbe*>. Nous recherchons l'ensemble des classes sémantiques des sujets retrouvés dans un patron de collocations identique à celui du pronom. Si la classe du candidat appartient à cet ensemble, les classes sémantiques du pronom et du candidat sont considérées comme cohérentes. Dans l'exemple ci-dessus le candidat *the ORF* et le pronom sont sujets du verbe *encode*. Nous connaissons la classe sémantique du candidat : c'est un gène. Nous recherchons s'il existe une collocation où un nom de gène est sujet du verbe *encode*. Une telle collocation existe (par exemple, dans la phrase ci-dessous), ce qui montre que le pronom peut être un gène.

  The [sacT]$_{gene}$ gene which controls the sacPA operon of Bacillus subtilis [encodes] a polypeptide homologous to the B. subtilis SacY.

– `Candidat` : Ce nœud est le nœud de prédiction du classifieur, il estime la probabilité pour un candidat d'être l'antécédent du pronom considéré. Il est lié avec tous les nœuds du réseau car chaque nœud intervient dans le calcul de la saillance du candidat.

La structure du réseau n'a pas été apprise automatiquement : elle a été définie par un expert. Nous justifions les liens les moins intuitifs. Le nœud *SN_Sujet* est lié aux nœuds des patrons de collocations pour renforcer (ou diminuer) le poids du parallélisme syntaxique entre le pronom et le candidat. Supposons en effet que le pronom précède immédiatement le verbe dont il est le sujet et que le candidat soit le sujet de

8. À l'exception des parenthèses qui sont souvent utilisées pour marquer les acronymes en biologie.

sa proposition. Le système recherche les collocations <*NP/verbe*> ou <*sujet-verbe*>. S'il trouve dans le corpus plusieurs collocations où le candidat est le sujet du verbe suivant le pronom, la probabilité pour le candidat d'être parallèle syntaxiquement doit être augmentée ainsi que la fiabilité du rôle grammatical du candidat.

Le caractère terminologique d'un syntagme n'est pas une propriété booléenne : un syntagme peut être plus ou moins représentatif du domaine. Pour cette raison, nous avons choisi de renforcer la probabilité pour un candidat d'être un terme par différents critères d'importance discursive du candidat (par exemple, s'il est démonstratif, si c'est un nom propre, etc.).

Les probabilités ont été apprises selon l'approche du *maximum de vraisemblance* sur un corpus d'apprentissage. Le nombre de nœuds qui composent notre réseau étant peu important, nous avons utilisé un algorithme d'inférence exacte pour calculer les probabilités des candidats. Les valeurs de certaines informations sont inévitablement bruitées. Après une étude de ces probabilités bruitées, les experts pourraient corriger les données et les intégrer aux données d'apprentissage avec l'approche du *maximum a posteriori* afin d'atténuer le bruit des paramètres, mais cela n'a pas été fait ici.

4. Corpus et protocole expérimental

Les expériences présentées ici visent à comparer différents classifieurs de candidats à l'antécédence anaphorique en faisant varier la nature des connaissances auxquelles ils ont accès et la manière dont ces connaissances sont modélisées.

4.1. *Corpus*

Nous avons évalué notre système sur le corpus *Transcript*. Ce corpus, initialement créé pour le projet Caderige[9], est composé de 2 209 résumés d'articles scientifiques de génomique (800 000 mots) extraits de la base `Medline` avec la requête : *bacillus subtilis, transcription*. Deux annotateurs ont étiqueté séparément les 429 relations anaphoriques et les 242 pronoms impersonnels, puis discuté une à une l'ensemble des annotations divergentes. Le taux d'accord après la résolution des conflits est de 92,4 %.

Le prétraitement de notre corpus a été réalisé grâce à la plate-forme linguistique *OGMIOS* du projet *ALVIS*. Nous avons utilisé une version de la plate-forme spécialisée pour le domaine de la génomique. Une description complète des outils intégrés dans la plate-forme peut être trouvée dans (Hamon et Nazarenko, 2008).

L'outil de segmentation des phrases de la plate-forme repose sur un algorithme à base d'expressions régulières. Cette étape est cruciale pour notre système mais son calcul est très bruité sur notre corpus de génomique, avec seulement 79,3 % de pré-

9. `http://caderige.imag.fr/`

cision, du fait de la présence de mesures, d'équations, de listes, etc. Nous avons pris le parti de corriger manuellement cette annotation. C'est l'unique correction apportée aux données d'entrée du système. Pour l'étiquetage grammatical nous avons utilisé *Genia Tagger*. C'est un outil spécialisé pour la génomique et évalué sur le corpus *Genia*, un corpus très similaire au nôtre. Nous supposons donc une performance voisine autour de 98,5 % d'exactitude. Le module *TagEN*, (Berroyer et Poibeau, 2004), reconnaît les EN grâce à une grammaire à base de transducteurs reposant sur des ressources linguistiques. Cette technologie dépend principalement de la qualité de ces ressources et obtient rarement des performances optimales : sur notre corpus, nous obtenons une performance de 69 % de précision et de 71 % de rappel, la faible précision étant due à la présence de noms de gènes ou de protéines ambigus (*not*, *All*, *Similar*, *RNA*, etc.). Pour l'étiquetage terminologique nous avons appliqué à notre corpus des ressources terminologiques du domaine, Gene Ontology et MeSH. L'évaluation de l'étiquetage terminologique étant difficile et coûteuse à faire, nous nous contentons de mesurer le taux de couverture des ressources sur notre corpus : 18 % des mots sont annotés comme termes du domaine. Enfin, l'analyse syntaxique est assurée par *BioLG* une version du *Link Parser* adaptée au domaine de la génomique (Aubin *et al.*, 2005). L'impact de la qualité de l'analyse syntaxique pour notre tâche de résolution des anaphores est directement mesuré par les performances du système *MAX* que nous présentons dans les sections ci-dessous.

4.2. *Protocole expérimental*

Notre corpus étant de taille moyenne, nous avons choisi une validation croisée avec 10 itérations. Les deux tiers du corpus sont affectés au corpus d'entraînement pour calculer les probabilités *a priori* des systèmes. Le tiers restant est réservé pour le test. Le prétraitement des documents est identique pour tous les systèmes en compétition, nous comparons ainsi uniquement les algorithmes de sélection des antécédents. Nous limitons l'analyse des erreurs de nos systèmes aux erreurs de la première itération mais nous la présentons en détail.

Nous avons résolu les anaphores avec six systèmes différents. Les trois premiers servent de référence. Le système *Premier-SN* sélectionne toujours le premier SN comme antécédent. Le système *MAX* simule le meilleur système possible : il sélectionne toujours l'antécédent dans la liste des candidats si l'antécédent a été correctement calculé lors de l'analyse syntaxique. Nous avons enfin réimplémenté le système MARS, que nous appelons *Bio-MARS*. Nous avons conservé le poids des scores fixés par les auteurs à partir de l'observation d'un corpus composé de documents techniques qui a servi à évaluer le système d'origine[10].

10. Nous avons vérifié les performances de notre réimplémentation du système MARS sur ce corpus. Les performances des deux systèmes sont comparables.

Les trois autres systèmes ont été retenus pour tester différentes configurations du modèle bayésien. Le système *Naïf-MARS* exploite les mêmes annotations que le système *Bio-MARS* mais la décision finale est prise par un classifieur bayésien naïf et non par le module de score original. Le système *Bayesien-Naïf* calcule aussi le score du candidat avec un classifieur bayésien naïf mais il dispose des annotations complexes décrites dans la section précédente en plus des annotations simples du système *Naïf-MARS*. Le dernier système est notre classifieur bayésien baptisé *Bayaphora*. Ce système exploite toutes les annotations, simples et complexes, et calcule le score du candidat avec le mécanisme d'inférence du RB pour atténuer les erreurs de calcul des valeurs de ces attributs.

Pour ces expériences, nous avons modifié la métrique du taux de succès partiel employée par Mitkov. Pour cet auteur, un antécédent est reconnu « partiellement » lorsque la tête de l'antécédent est annotée mais qu'une partie des mots composant l'antécédent a été oubliée. Nous ajoutons une contrainte supplémentaire : il faut que la partie annotée contienne la tête de l'antécédent et puisse être substituée au pronom anaphorique sans incohérence[11].

5. Analyse des résultats

La figure 2 présente les taux de succès partiels obtenus par nos systèmes sur la totalité des itérations et le tableau 1 leurs moyennes.

Système	**Résultats**	
	Strict	*Partiel*
Premier SN	33,48 %	42,97 %
Bio-MARS	26,41 %	37,05 %
Naïf-MARS	40,32 %	56,46 %
Après correction partielle	46,51 %	63,57 %
Bayesien-Naïf	40,48 %	56,57 %
Bayaphora	39,60 %	56,31 %
Après correction partielle	52,71 %	73,64 %
MAX	63,71 %	87,15 %

Tableau 1. *Résultats de la résolution des anaphores pronominales sur le corpus Transcript (taux de succès strict (partiel) = Anaphores résolues strictement (+ partiellement) / Totalité des anaphores)*

11. Dans la phrase : *[beta-Galactosidase expression from the spl-lacZ fusion]*$_1$ *was silent during vegetative growth and was not DNA damage inducible, but [it]*$_1$ *was activated at morphological stage III*... notre système annote uniquement *beta-Galactosidase expression*, mais c'est un candidat qui peut être substitué au pronom sans perte de sens.

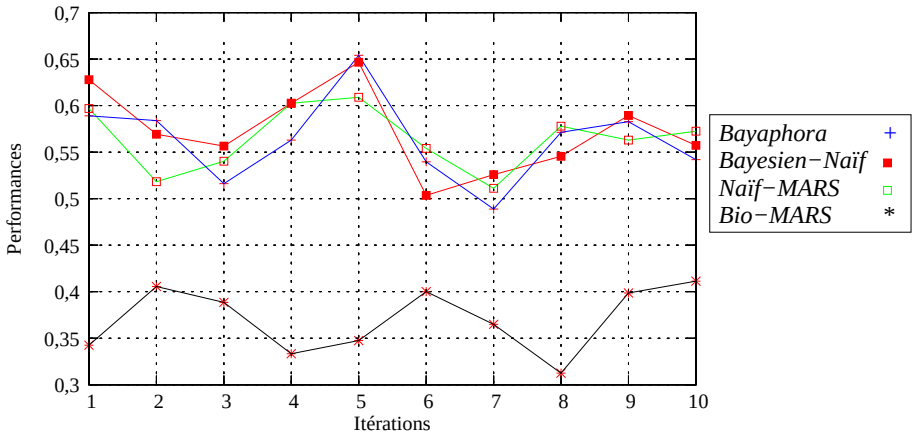

Figure 2. *Détails des taux de succès partiels obtenus par les systèmes sur toutes les itérations*

Le premier fait marquant de ces résultats est une performance de *Bio-MARS* qui est inférieure sur notre corpus à celle obtenue par *Premier-SN*. Les scores de *Bio-Mars*, qui ont été fixés pour un corpus de documents techniques, apparaissent surévalués. Dans la version probabiliste de *Bio-Mars*, *i.e. Naïf-MARS*, le paramétrage adapte les scores des attributs à partir des observations faites sur le corpus d'entraînement et évite les erreurs dues à la surévaluation des scores initiaux. L'écart des résultats montre la sensibilité du système aux variations de domaine des corpus. Le passage aux probabilités permet, en outre, une représentation plus fine de l'incertitude, ce qui réduit le nombre de cas d'égalité entre les candidats et donc les erreurs d'arbitrage dues aux heuristiques.

Le second fait marquant, c'est la proximité des performances entre les trois systèmes bayésiens. Notre objectif consistait à calculer la saillance d'un candidat en employant toutes les connaissances disponibles, aussi complexes soient-elles et quelle que soit leur fiabilité, tout en comptant sur le mécanisme d'inférence du RB pour corriger l'incertitude d'un attribut grâce aux autres attributs dont il dépend. Or, on observe que *Bayesien-Naïf* obtient des résultats à peine meilleurs que *Naïf-MARS* : l'apport des connaissances complexes reste donc marginal dans le contexte de notre expérience. De surcroît, *Bayaphora* a des résultats très proches mais légèrement moins bons que *Bayesien-Naïf*, ce qui laisse penser que les relations de dépendance entre attributs ne sont pas exploitées.

Dans les sections suivantes, nous analysons en détail les raisons de cette contre-performance et nous interprétons ces résultats relativement à la question de l'intégration de connaissances complexes et parfois peu fiables dans le calcul d'annotations.

5.1. *Une utilisation rudimentaire des dépendances entre les annotations*

On constate d'abord que *Bayaphora* n'exploite pas pleinement les dépendances entre les annotations. L'étude des erreurs du système montre que le nombre d'attributs intégrés est encore insuffisant pour améliorer significativement la classification du candidat saillant. De plus la comparaison des performances de *Bayaphora* et de *Bayesien-Naïf* montre que la structure du RB, seule différence entre ces deux systèmes, est trop simple et exprime trop peu de contraintes pour corriger les valeurs erronées des attributs.

5.1.1. *Un nombre insuffisant d'attributs*

Le tableau 2 résume les types d'erreurs de *Bayaphora* que nous avons identifiées.

Causes des erreurs	*Erreurs du système* Bayaphora
Calcul erroné du candidat saillant	27 % (18)
Candidat saillant différent de l'antécédent	15 % (10)
Anomalie du filtre du nombre	6 % (4)
Erreur de prétraitement	9 % (6)
Anaphore clausale	2 % (1)
Candidat partiellement retrouvé mais rejeté	17 % (11)
Candidat partiellement retrouvé et accepté	24 % (16)
Total	100 % (66)

Tableau 2. *Typologie des erreurs de Bayaphora sur le corpus Transcript lors de la 1^{re} itération*

Sur les 39 erreurs du système, 28 erreurs portent sur l'élément saillant. Une partie de ces erreurs pourrait être corrigée par l'ajout de nouvelles informations linguistiques qui viendraient améliorer la discrimination du candidat saillant ou exprimer des contraintes supplémentaires pour rejeter le candidat lorsqu'il diffère de l'antécédent. Nous en donnons quelques exemples dans les paragraphes suivants.

5.1.1.1. Le candidat choisi n'est pas le SN saillant

Ces 18 erreurs sont dues au fait que l'élément saillant, qui est l'antécédent, n'est pas identifié parce qu'une majorité d'attributs favorise un candidat différent de l'élément saillant, que les valeurs des attributs soient correctes ou non. Dans la phrase :

> *[Gel filtration chromatography]$_1$ indicated that [the native enzyme]$_2$ existed as a dimer at high protein concentrations but that [it]$_2$ dissociated to a monomeric form on dilution.*

l'antécédent est clairement l'élément saillant de la phrase (noté 2) : on voit que l'auteur attire l'attention du lecteur sur le SN avec le verbe *indicate* avant de le décrire dans la deuxième proposition. L'élément saillant obtient une bonne probabilité de 74 % mais

notre système lui préfère le candidat 1, qui a une probabilité plus importante (99 %). Le système a correctement identifié que l'antécédent et le candidat sont tous deux sujets d'un verbe et qu'ils ont le même rôle grammatical que le pronom. L'antécédent est reconnu comme un SN défini et il suit un verbe spécifique. Le poids de ces attributs caractérisant l'antécédent aurait pu le qualifier. Mais le poids plus important de la position du candidat 1 comme premier SN de la proposition ainsi que la mauvaise annotation du mot *Gel* confondu avec le gène du même nom, lui donnent l'avantage. Le fait que ce candidat soit indéfini et que le filtre de la cohérence sémantique le refuse ne suffisent pas à le rejeter. La probabilité conditionnelle pour le candidat d'être l'antécédent sachant qu'il est rejeté par le filtre est, anormalement, non nulle en raison des valeurs erronées du corpus d'apprentissage.

5.1.1.2. Le SN saillant choisi n'est pas l'antécédent

Pour 10 autres erreurs, le système trouve bien le candidat que l'annotateur humain juge « intuitivement » être l'élément saillant mais ce candidat n'est pas l'antécédent. Prenons en exemple le contexte suivant :

> *The transcription of spoVE initiated within an hour after the onset of sporulation and coincided with the presence of RNA polymerase associated with a 33-kDa protein. [Amino acid sequence analysis]$_1$ of [the 33-kDa protein]$_2$ revealed that [it]$_2$ is a sigma factor, sigma E. Reconstitution analysis of sigma E purified from the sporulating cell extracts and vegetative core RNA polymerase showed that sigma E recognizes the P2 promoter.*

Dans ce contexte, l'auteur attire l'attention du lecteur sur les méthodes employées pour déterminer la nature de la protéine. Pour cela, il positionne les deux analyses *Amino acid sequence analysis* et *Reconstitution analysis* en position de sujet. Selon nous, ces SN sont les éléments saillants mais, après la première phrase, le centre du discours se positionne sur la protéine *sigma E*. Notre système, qui recherche l'élément saillant ne choisit donc pas l'antécédent. En modélisant des connaissances propres du domaine, nous aurions pu intégrer au réseau une contrainte qui est violée par le candidat (le fait qu'un facteur sigma soit une protéine) et rejeter ce dernier. Ce type d'erreur survient essentiellement lorsque l'antécédent est en position de complément du nom (7 erreurs sur 10).

5.1.1.3. Des annotations d'entrée erronées

Quatre erreurs s'expliquent par une anomalie du filtre du nombre, liée elle-même à une erreur de calcul de notre heuristique ou à une mauvaise annotation morphologique et syntaxique. Prenons l'antécédent *one of the sites*, le SN *sites* est un pluriel correctement annoté et malheureusement filtré par notre heuristique qui rejette tout syntagme contenant un mot au pluriel. De même, pour l'antécédent *the ((G/U) AGCC) 11 RNAs*, le gène *RNAs* n'est pas reconnu par l'analyseur morphologique et est étiqueté pluriel. Le cas de l'antécédent *the spo0B gene* est plus complexe. Nous nous reposons

sur l'analyse syntaxique en constituants pour identifier les SN mais cette analyse est bruitée. Elle regroupe les deux syntagmes *the spo0B gene* et *genes* de la phrase :

> *Transcription of [the spo0B gene and genes] downstream of it was investigated by S1 nuclease protection experiments.*

Notre système est donc contraint d'aligner l'antécédent *the spo0B gene* avec l'unique candidat *the spo0B gene and genes* qui est à son tour exclu par le filtre du nombre. Les rejets du filtre causés par une mauvaise analyse syntaxique des candidats représentent 3 erreurs sur 4.

Les 7 erreurs restantes s'expliquent simplement. Il y a 6 erreurs de prétraitement : 4 antécédents sont absents de la liste des candidats et 2 pronoms anaphoriques apparaissent dans une phrase dont la segmentation n'a pas été corrigée. La dernière erreur est une anaphore clausale : le pronom renvoie à un fait et non à un objet du discours.

5.1.2. *Un mécanisme de renforcement peu efficace*

La correction des annotations d'entrée erronées est normalement dévolue au mécanisme de renforcement du RB de *Bayaphora* mais la comparaison avec les performances de *Bayesien-Naïf* laisse penser que ce mécanisme ne modère pas le bruit des annotations. Le RB apparaît, rétrospectivement, trop simple. Plusieurs modifications dans la structure du réseaux peuvent être envisagées sans toutefois engager de nouvelles connaissances difficiles à obtenir.

La première consiste à modifier les valeurs de certaines VA. Par exemple, les valeurs de l'attribut *SN_Sujet* sont inadaptées pour une analyse syntaxique très silencieuse. Les exemples d'entraînement étiquetés *Sujet* ou *Complément* sont trop peu nombreux et les probabilités *a priori* apprises pour cet attribut sont incorrectes car peu discriminante. Nous envisageons pour de futures expériences de supprimer la valeur *Inconnu* et de distinguer les rôles complément d'objet et du nom. Ces modifications nécessiteront un apprentissage des paramètres plus complexe avec un algorithme capable d'estimer les valeurs manquantes dans les données d'entraînement ou *via* l'interrogation d'experts.

Nous proposons également d'ajouter de nouveaux liens entre les nœuds pour exprimer des renforcements ou des contraintes supplémentaires. Les nœuds `Semantic_Consistence`, `Parallelisme_Syntaxique` et `Patron_Collocation` peuvent être reliés pour renforcer les connaissances syntaxiques grâce aux connaissances sémantiques. Reprenons une dernière fois notre exemple. Si un gène peut être le sujet du verbe *encode* et que le candidat *ORF* est un gène, nous pouvons en déduire qu'il peut, comme le pronom, être le sujet du verbe *encode*. De même, un arc entre les nœuds *SN_Pronom* et *Filtre_Nombre* interdirait le choix d'un pronom personnel pluriel comme antécédent. Sur l'ensemble des itérations, 5 pronoms *we* ont été choisis comme antécédent. La raison est identique à celle des erreurs commises par le filtre de la cohérence sémantique dans l'exemple précédent. La probabilité conditionnelle pour un candidat d'être l'antécédent sachant que c'est un SN pluriel est incorrectement positive. Or, dans ce cas, le système connaît par l'étiquetage morphosyntaxique que ce

candidat est un pronom personnel pluriel. Avec l'ajout d'une valeur mentionnant la nature d'un pronom pluriel dans la variable *SN_Pronom* et le nouvel arc, le système corrigerait ces erreurs.

Enfin, nous avons assigné à chacune de nos observations une certitude maximale alors que des vérifications élémentaires permettraient de la moduler[12]. Par exemple, pour une séquence reconnue comme EN il suffirait de vérifier sa présence dans un dictionnaire de la langue anglaise, la présence de caractères non alphanumériques et celle de marques typographiques. Une telle vérification aurait considérablement diminué la fiabilité des annotations spécifiant que *not* ou *similar* sont des noms de gènes.

5.2. *Le rôle des annotations complexes*

La proximité des performances entre les trois systèmes bayésiens vient aussi du fait qu'il existe un seuil de qualité pour les annotations syntaxiques et sémantiques en dessous duquel elles n'améliorent pas la résolution et la dégradent même un peu. Quelle que soit l'efficacité du mécanisme d'inférence, il ne peut trouver la bonne décision si les informations sur lesquelles il repose contiennent trop d'erreurs ou ne sont pas disponibles. Les sections suivantes expliquent la démarche suivie pour mettre en évidence ce seuil.

5.2.1. *Le rôle négatif des annotations complexes calculées automatiquement*

La figure 2 montre que *Naïf-MARS* n'obtient les meilleures performances que pour 3 itérations et avec très peu d'écart sur ses concurrents. Cette observation suggère que les performances des systèmes *Bayaphora* et *Bayesien-Naïf* sont principalement déterminées par la qualité des annotations complexes qu'ils exploitent.

Pour en rendre compte nous avons mis en regard les erreurs de *Naïf-MARS* et celles de *Bayaphora* lors de la première itération. L'examen détaillé de ces erreurs montre que ce sont essentiellement les erreurs des annotations des connaissances linguistiques clés (la classe sémantique et le rôle grammatical d'un syntagme) qui dégradent les performances de *Bayaphora*. Sur les 5 erreurs propres au système *Bayaphora*, c'est-à-dire les 5 antécédents manqués par *Bayaphora* et strictement ou partiellement retrouvés par *Naïf-MARS*, 4 sont imputables aux valeurs inexactes de la classe sémantique et du rôle grammatical du candidat choisi par le système. En effet, rappelons que le module de reconnaissance des EN commet beaucoup d'erreurs avec 32 % des entités étiquetées incorrectement. Si le calcul du rôle grammatical est réalisé avec une bonne précision (84 % des relations présentes sont correctes), le rappel est très mauvais avec seulement 12 % des antécédents et 8 % des candidats qui ont un rôle associé.

12. Ce type d'évidence est appelé « observation vraisemblable ». Elle est modélisée par un nouveau nœud dans le réseau : la VA est utilisée comme les autres VA mais sa « valeur sémantique » est différente. Elle exprime la confiance dans la probabilité que l'attribut auquel elle est liée prenne une certaine valeur. C'est une probabilité de second ordre.

5.2.2. *Le rôle positif des annotations complexes partiellement corrigées*

L'analyse précédente porte sur un trop petit nombre d'erreurs pour permettre de conclure que seul le bruit des annotations linguistiques est à l'origine des erreurs des systèmes *Bayaphora* et *Bayesien-Naïf*. Pour mettre en évidence son rôle, nous avons procédé à une nouvelle expérience en modifiant certains paramètres.

Nous avons corrigé manuellement les annotations suspectes pour les 50 pronoms résolus incorrectement par un des deux systèmes : seuls les EN, l'analyse syntaxique en constituants et les rôles grammaticaux des phrases de la fenêtre de recherche de l'antécédent ont été corrigés ; aucune correction n'a été apportée aux autres attributs comme les filtres ou les patrons de collocations. Nous avons renouvelé la résolution avec les paramètres du réseau appris sur la totalité du corpus afin de comparer les anciens et les nouveaux taux de succès. En considérant l'intégralité du corpus pour l'apprentissage des paramètres des classifieurs, nous introduisons sciemment un biais : nous réduisons l'effet des attributs bruités sur les paramètres en augmentant le nombre de données d'apprentissage. Les paramètres ainsi obtenus sont donc les meilleurs possibles pour notre corpus compte tenu de la qualité de son annotation.

Revenons sur notre précédent exemple pour voir le processus en détail. Lors de la première itération, les deux systèmes choisissent, comme antécédent, le premier candidat *Gel filtration chromatography*. *Naïf-MARS* choisit le candidat avec une forte fiabilité 82 % car il est le premier SN de la phrase, et rejette l'antécédent *the native enzyme* qui obtient seulement 20 %. *Bayaphora* possède plus d'informations pour décrire les candidats concurrents. Mais, avant correction des annotations, la classe sémantique des candidats est erronée, le mot *Gel* est confondu avec le gène du même nom, ce qui donne l'avantage au premier candidat (avec des scores respectifs de 98 % et 74 % pour l'antécédent). Cet attribut corrigé, *Bayaphora* reconsidère – cette fois correctement – la position de l'antécédent grâce aux attributs du verbe spécifique et de l'article défini qui l'emporte sur le premier candidat avec un score de 98 % contre 97 %. *Naïf-MARS* voit, aux probabilités près, ses entrées et sa réponse inchangées.

Une fois les valeurs des annotations complexes corrigées, *Bayaphora* les exploite correctement et voit ses taux de succès (52,71 % et 73,64 %) se détacher significativement de ceux du système plus pauvre (46,51 % et 63,57 %). Ces résultats sont importants : ils confirment que le bruit des annotations linguistiques limite faiblement les performances des systèmes qui les exploitent et justifie réciproquement l'ajout des annotations complexes dans l'inférence. On voit en effet que, partiellement corrigées, elles apportent un gain réel sans pour autant être très nombreuses. Ces résultats donnent une estimation des performances maximales qu'on peut obtenir quand on dispose d'algorithmes de révision parfaits et optimisés pour une annotation particulière. La différence significative que nous constatons entre les performances obtenues actuellement et les performances possibles après révision légitime et motive la recherche de telles stratégies de révision.

6. Discussion et perspectives

L'imperfection des annotations d'entrée et de sortie des systèmes de TAL semble être une fatalité avec laquelle nous devons composer plutôt qu'un désagrément passager dû à une insuffisante maturité des techniques de TAL que nous pourrions ignorer. Pour traiter le problème de l'imperfection, deux stratégies s'opposent à l'heure actuelle. La première, justifiée par le besoin de systèmes de TAL robustes et effectifs, adapte l'architecture incrémentale des plates-formes d'annotation standard en cherchant à limiter l'usage des annotations erronées ou leur propagation. La seconde stratégie, plus ambitieuse, conçoit des architectures intégrées innovantes pour annoter et corriger simultanément un ensemble d'annotations choisi. Malheureusement, la complexité des calculs requis par les modèles d'inférence de ces plates-formes limite le nombre des annotations qu'elles peuvent réellement prendre en compte.

Dans cet article nous avons étudié les conditions de possibilité d'un troisième type d'architecture. Fonctionnant sur le modèle d'une architecture incrémentale, chaque module de la plate-forme intègre l'ensemble des annotations d'entrée nécessaires pour son calcul et quel que soit leur degré de complexité. La fiabilité et la dépendance des annotations sont données *a priori* par l'apprentissage et utilisées dans l'inférence du module pour limiter l'impact des annotations d'entrée erronées. La révision des annotations peut alors être réalisée *a posteriori* avec des stratégies adaptées aux types d'annotations et optimisées en fonction de l'application visée.

Nous avons établi nos résultats par une série d'expériences menées sur la résolution des pronoms anaphoriques *it* dans les textes de génomique. Ce problème est un bon sujet d'étude car il se traduit aisément en une tâche de classification et met en jeu un grand nombre d'annotations de complexités variées. Nous avons présenté notre système qui est inspiré du système pauvre en connaissance de Mitkov (2002), le système *MARS*. Nous avons conservé la stratégie générale de ce système qui recherche l'élément saillant d'un contexte pour résoudre l'anaphore, mais nous avons enrichi le calcul de cet élément. Le système peut exploiter l'incertitude des annotations d'entrée qui est finement modélisée grâce aux probabilités et il peut combiner des connaissances simples avec les connaissances complexes dont elles dépendent.

Pour isoler le rôle joué par les annotations complexes dans la résolution nous avons comparé les performances de 6 systèmes de résolution, qui ne diffèrent que par les annotations d'entrée qu'ils utilisent et/ou les algorithmes employés pour choisir l'antécédent. Lors de l'analyse des résultats, nous avons été particulièrement attentifs aux cas pour lesquels les systèmes proposent des solutions divergentes. L'analyse détaillée de ces cas nous a permis de mettre en évidence l'existence d'un seuil de fiabilité des annotations complexes qui autorise leur intégration et motive leur révision :

– l'emploi de l'ensemble des annotations pertinentes, même si elles sont erronées ou manquantes, par un système conçu pour en faire usage ne dégrade pas fortement ses performances, si on les compare à celles d'un système travaillant uniquement avec des annotations simples et fiables ;

– au-dessus du seuil, les annotations complexes, bien que toujours imparfaites, apportent une amélioration significative du système.

Pour être généralisés nos résultats doivent toutefois être complétés. Nous avons choisi de travailler dans un cadre applicatif réel avec les outils existants. Ce choix a réduit le nombre d'annotations complexes intégrables dans le module. Pour étayer nos résultats, une expérience supplémentaire doit être menée avec toutes les annotations complexes, qu'elles soient calculées ou données en entrée. De plus, ces résultats ne concernent que la résolution des anaphores. Une série d'expériences similaires doit être réalisée sur les autres modules d'annotations de la plate-forme. Nous pensons, notamment, à la révision de la segmentation des phrase et des étiquettes morphosyntaxiques en suivant une stratégie similaire à celle employée par Hirakawa et Yoshimura (2000). Enfin, même si ces résultats suggèrent que l'intégration des annotations complexes améliorera le calcul de l'annotation cible, cette amélioration est conditionnée par l'existence de stratégies de révision *a posteriori* permettant d'automatiser ce que nous avons fait ici manuellement.

Nous travaillons maintenant à l'amélioration de notre classifieur. Nous avons utilisé peu d'annotations sémantiques dans notre système, alors que ces annotations sémantiques et ontologiques existent et peuvent être facilement ajoutées au classifieur (rôles sémantiques et ontologies spécialisés pour la biologie, par exemple). Avec ces nouvelles annotations il devient possible d'améliorer l'algorithme de résolution en traquant les changements d'antécédent dans le discours pour corroborer le choix de l'élément saillant.

Nous menons en parallèle nos premiers travaux sur la recherche d'une stratégie de révision et sur la mesure du seuil de qualité des annotations. Nous explorons les pistes d'une révision reposant sur la résolution des contradictions dans une théorie logique de révision des croyances. Nous commençons une série d'expériences supplémentaires en introduisant artificiellement des erreurs d'annotations pour mesurer le seuil. Nous posséderons alors un outil statistique très utile : l'intégration dans la plate-forme des annotations dont la qualité est supérieure au seuil sera justifiée car elles amélioreront les performances globales, même s'il y a un coût supplémentaire à supporter pour leur correction automatique ou manuelle.

7. Bibliographie

Aubin S., Nazarenko A., Nedellec C., « Adapting a general parser to a sublanguage », *International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, p. 89-93, 2005.

Berroyer J., Poibeau T., TagEN, un analyseur d'entitées nommées, Technical report, LIPN, 2004.

Bouckaert R., « Low level information extraction, a Bayesian network based approach », *Workshop on Text Learning (TextML-2002)*, 2002.

Brill E., « Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part-of-Speech Tagging », *Computational Linguistics*, vol. 21, p. 543-565, 1995.

Bunescu R., « Learning with Probabilistic Features for Improved Pipeline Models », *Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

Cardie C., Wagstaff K., « Noun phrase coreference as clustering », *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p. 82-89, 1999.

Chang M.-W., Ratinov L., Rizzolo N., Roth D., « Learning and inference with constraints », *Proceedings of the 23rd national conference on Artificial intelligence*, 2008.

Collobert R., Weston J., « A unified architecture for natural language processing : deep neural networks with multitask learning », *25th international conference on Machine learning*, vol. 307, p. 160-167, 2008.

Denoyer L., Gallinari P., « Bayesian Network Model for Semi-Structured Document Classification », *Information Processing and Management*, 2004.

Derivière J., Hamon T., Nazarenko. A., « A Scalable and Distributed NLP Architecture for Web Document Annotation », *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, p. 56-67, 2006.

Dickinson M., Error Detection and Correction in Annotated Corpora, PhD thesis, Ohio State University, 2005.

Goldman S., Sloan R., « Can PAC Learning Algorithms Tolerate Random Attribute Noise ? », *Algorithmica*, vol. 14, p. 70-84, 1995.

Goyal P., Behera L., McGinnity T., « Application of Bayesian Framework in Natural Language Understanding », *IETE Technical Review journal*, vol. 25, n˚ 5, p. 251-269, 2008.

Grishman R., TIPSTER Architecture Design Document Version 3.1, Technical report, DARPA, 1997.

Halpern J., *Reasoning about uncertainty*, MIT press, 2003.

Hamon T., Nazarenko A., « Le développement d'une plate-forme pour l'annotation spécialisée de documents Web : retour d'expérience », *Traitement automatique des langues (TAL)*, 2008.

Hirakawa Hideki K. O., Yoshimura Y., « Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpora », *the 18th International Conference on Computational Linguistics*, 2000.

Mitkov R., *Anaphora Resolution*, Longman(Pearson Education), 2002.

Mitkov R., Boguraev B., Lappin S., « Introduction to the Special Issue on Computational Anaphora Resolution », *Computational Linguistics*, vol. 27(4), p. 473-477, 2001.

Roth D., Wen-tau Y., « Probalistic Reasoning for Entity and Relation Recognition », *COLING'02*, 2002.

Sahami M., Dumais S., Heckerman D., Horvitz E., « A Bayesian Approach to Filtering Junk E-Mail », *Learning for Text Categorization : Papers from the 1998 Workshop*, 1998.

Valiant L. G., « A theory of the learnable », *Communications of the ACM*, vol. 27, 1984.

Weissenbacher D., Influence des annotations imparfaites sur les systèmes de traitement automatique des langues, un cadre applicatif : la résolution de l'anaphore pronominale, PhD thesis, Université de Paris-Nord (Paris XIII), 2008.

Weissenbacher D., Nazarenko A., « A bayesian classifier for the recognition of the impersonal occurrences of the it pronoun », *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'07)*, 2007.

Wen-tau Y., Learning and Inference for Information Extraction, PhD thesis, University of Illinois at Urbana-Champaign, 2005.

Zhu X., Wu X., « Class Noise vs. Attribute Noise : A Quantitative Study of Their Impacts », *Artificial Intelligence Revue*, vol. 22(3), p. 177-210, 2004.