TALN 2009, Senlis, 24–26 juin 2009

Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs

Marion Laignelet[1] François Rioult[2]
(1) CLLE-ERSS / Université de Toulouse 2 Le Mirail
(2) GREYC (CNRS UMR6072), Université de Caen Basse-Normandie
marion.laignelet@univ-tlse2.fr, Francois.Rioult@info.unicaen.fr

Résumé. Cet article vise la description et le repérage automatique des segments d'obsolescence dans les documents de type encyclopédique. Nous supposons que des indices sémantiques et discursifs peuvent permettre le repérage de tels segments. Pour ce faire, nous travaillons sur un corpus annoté manuellement par des experts sur lequel nous projetons des indices repérés automatiquement. Les techniques statistiques de base ne permettent pas d'expliquer ce phénomène complexe. Nous proposons l'utilisation de techniques de fouille de données pour le caractériser et nous évaluons le pouvoir prédictif de nos indices. Nous montrons, à l'aide de techniques de classification supervisée et de calcul de l'aire sous la courbe ROC, que nos hypothèses sont pertinentes.

Abstract. This paper deals with the description and automatic tracking of obsolescence in encyclopedic type of documents. We suppose that semantic and discursive cues may allow the tracking of these segments. For that purpose, we have worked on an expert manually annotated corpus, on which we have projected automatically tracked cues. Basic statistic techniques can not account for this complex phenomenon. We propose the use of techniques of data mining to characterize it, and we evaluate the predictive power of our cues. We show, using techniques of supervised classification and area under the ROC curve, that our hypotheses are relevant.

Mots-clés : repérage automatique de l'obsolescence, indices sémantiques et discursifs, textes encyclopédiques, classification supervisée, aire sous la courbe ROC.

Keywords: automatic tracking of obsolescence, semantic and discursive cues, encyclopedic type of documents, supervised classification, area under the ROC curve.

1 Introduction

Cet article s'inscrit dans le cadre d'un projet de recherche visant la description et le repérage automatique de segments obsolescents dans des documents de type encyclopédique. La visée applicative de ce travail consiste en la création d'un outil d'aide à la mise à jour de textes encyclopédiques pour l'édition. C'est à travers la notion de mise à jour que celle d'obsolescence prend son sens : un segment d'obsolescence est un segment textuel qui contient de l'information susceptible d'évoluer dans le temps.

Nous supposons que des indices sémantiques et discursifs peuvent permettre le repérage des segments d'obsolescence. Le système mis en œuvre a pour objectif de repérer des combinaisons

d'indices qui fonctionneront alors comme des marqueurs de l'obsolescence. Ce travail s'appuie sur les recherches menées autour des notions de marqueurs textuels et discursifs comme les *mots-repères* ou les *mots-titres*, notions déjà envisagées par (Edmundson, 1969), les *cue phrases* (Grosz & Sidner, 1986) ou encore les éléments participant de l'analyse de la structure de texte (Marcu, 2000). De plus, considérant le caractère multifonctionnel des marqueurs de surface (Grosz & Sidner, 1986), nous nous focalisons sur leur fonction pragmatique, *i.e.* leur aptitude à déterminer un segment d'information évolutive. Enfin, les aspects discursifs des documents à travers les titres (Ho-Dac *et al.*, 2004), les cadres de discours (Charolles, 1997) ou encore la position dans le document occupent une place importante dans nos recherches. Nous travaillons sur un corpus annoté manuellement par des experts sur lequel nous projetons des indices repérés automatiquement.

Pour évaluer la pertinence de notre approche, nous utilisons une technique d'apprentissage automatique, la classification supervisée. Cette démarche permet à la fois de vérifier si une machine peut détecter les segments obsolescents à l'aide de nos indices, et de valider leur intérêt. De plus, l'apprentissage fournit un modèle des données, par exemple sous forme de règles, que l'expert peut analyser, dans une optique de découverte de connaissances, puis discuter, dans le but d'intégrer sa propre connaissance métier. La classification automatique de textes est une tâche aujourd'hui bien maîtrisée, comme en témoignent les dernières campagnes du DÉfi Fouille de Texte. Les approches les plus courantes sont numériques et fondées sur la détection de n-grammes. Les indices sont dans ce cas homogènes puisqu'ils caractérisent une mesure d'apparition des formes de surface. Ce type de données est naturellement valorisé par des méthodes d'apprentissage à noyaux comme les Séparateurs à Vaste Marge (SVM). Cependant, le modèle fourni sous forme de combinaisons de poids est peu intelligible et n'incite pas à l'interaction avec l'expert.

A contrario, notre approche est orientée vers des indices de type sémantique et discursif car l'obsolescence n'est pas liée à des formes de surface. De plus, nos indices caractérisent des propriétés très hétérogènes, ce qui rend délicat l'usage de SVM. Enfin, nous abordons l'apprentissage automatique selon une démarche exploratoire visant à déterminer les indices potentiellement intéressants et nous souhaitons obtenir un modèle interprétable. C'est pour ces raisons que nous utilisons une méthode de classification supervisée à base de règles d'association. La fouille de motifs permet en effet d'extraire de la connaissance, sous forme de règles, sans faire d'hypothèse sur le modèle des données, ni choisir *a priori* les descripteurs intéressants. Outre sa vocation exploratoire, cette technique offre l'avantage de fournir un modèle interprétable à l'expert. La méthode de décision n'est plus une boîte noire et les interactions entre expert et fouilleur sont bien plus vivantes.

La première section est consacrée à la description de notre corpus annoté manuellement, des segments contenant de l'information obsolescente et de la définition de la notion d'obsolescence (section 2). Dans une seconde partie, nous décrivons quels indices linguistiques (sémantiques et discursifs) sont utilisés (section 3). La section 4 discute de la performance de nos indices pour une tâche de classification supervisée et des connaissances obtenues.

2 Données et Phénomène observé

Cette section rassemble les éléments nécessaires à la compréhension de l'article. Nous décrivons d'abord les données textuelles à notre disposition, puis nous expliquons le phénomène

d'obsolescence.

Notre corpus compte 10 000 phrases. Il est composé de deux sous-corpus : ATLAS, qui regroupe des fiches encyclopédiques éditées par les éditions Atlas, et LAROUSSE qui est constitué par entrées des encyclopédies des Éditions Larousse (le *Grand Larousse Informatisé* et le *Grand Universel Larousse*). Puisque les textes sont de type encyclopédique, ils sont référencés sous des rubriques : géographie, histoire, faune et flore, sciences et techniques, sport, économie, *etc*. Ces deux sous-corpus ont été annotés manuellement en fonction du caractère obsolescent ou non des segments : un expert linguiste[1] a annoté le corpus ATLAS et le corpus LAROUSSE, ce dernier ayant été également annoté par trois experts rédacteurs[2].

L'obsolescence est un phénomène non linguistique, créé par l'usage : les segments d'obsolescence se définissent d'abord par rapport à un besoin réel, à savoir la mise à jour éditoriale. Ils présentent ainsi la particularité de contenir de l'information susceptible d'évoluer dans le temps.

Nous distinguons deux types de segments d'obsolescence : d'un côté ceux dont l'information est devenue fausse du fait de l'évolution temporelle et/ou des connaissances, de l'autre, ceux dont l'information n'est plus pertinente, plus actuelle au moment où elle est lue. Comparons les segments obsolescents dans les deux exemples construits suivants :

(1) *Aujourd'hui, le PIB par habitant de la France est de 27 600 dollars.*

(2) *En 2004, le PIB par habitant de la France est de 27 600 euros.*

Dans l'exemple (1), sachant qu'on est actuellement en 2009, et que le lecteur va naturellement interpréter l'adverbial « aujourd'hui » comme étant l'année en cours, l'information est fausse puisque le PIB de la France le plus actuel (chiffres de 2008) est de 33 800 dollars. À l'inverse, l'exemple (2) montre un cas où l'information restera toujours vraie : en 2004, le PIB par habitant de la France sera toujours de 27 600 dollars. Une mise à jour éditoriale sera cependant nécessaire si l'objectif du rédacteur est de fournir les résultats les plus récents par rapport à la date en cours : il faudra donc actualiser à la fois la référence temporelle (« En 2009 ») et la valeur du PIB. Ce cas est très fréquent dans les fiches/entrées du domaine géographique.

Afin de délimiter et de mieux comprendre la tâche de mise à jour, nous avons demandé à quatre annotateurs de juger de l'obsolescence des phrases du sous-corpus LAROUSSE. Il en ressort que ce sous-corpus est en moyenne composé de 10 à 15 % de segments obsolètes par annotateur. L'*accord observé* entre chacun de ces juges se situe entre 87 et 92 %.

Lorsqu'on utilise le coefficient Kappa, le taux d'accord est situé entre 0.35 et 0.50 : ce score très faible est directement lié au fait que les segments obsolescents ne représentent au mieux que 15 % de l'ensemble du sous-corpus, soit une classe minoritaire dans le corpus. L'accord entre les juges est mieux traduit par le coefficient *r* de Finn[3] (Hripcsak & Heitjan, 2002) car il met l'accent sur la classe minoritaire visée : le score varie de 0.74 pour l'accord le plus bas (les codeurs 2 et 4) à 0.83 pour l'accord le plus haut (codeurs 1 et 3). Ces résultats montrent que l'accord entre nos experts est plutôt bon même s'il reste malgré tout une zone floue, que nous attribuons au caractère parfois subjectif de la notion d'obsolescence.

[1] l'un des auteurs de cet article, Marion Laignelet.
[2] de la société Larousse.
[3] Nous utilisons l'algorithme existant dans le logiciel R.

3 Conception des indices sémantiques et discursifs

Cette section précise la démarche linguistique employée pour concevoir des indices sémantiques et discursifs potentiellement pertinents pour la caractérisation de l'obsolescence.

3.1 Méthodologie

Notre intuition et notre compétence de linguiste nous ont naturellement amené à orienter nos recherches sur les indices de type temporels comme marqueurs potentiels de l'obsolescence. Par ailleurs, notre intérêt de longue date pour les travaux en discours nous a entraîné vers l'exploitation d'indices à plus grande granularité tels que les titres ou la position des paragraphes au sein des documents. Enfin, à la lumière des annotations réelles de nos experts, nous avons élargi nos types d'indices vers la prise en compte d'entités nommées, de valeurs chiffrées ou encore de l'expression du point de vue du locuteur comme indicateurs de l'obsolescence.

Aujourd'hui, nous disposons d'environ 150 types d'indices. Leur repérage et leur annotation sémantique sont effectués de manière automatique avec l'outil ALIDIS (Laignelet, à paraître) développé à l'aide de la plateforme LinguaStream[4] (Widlocher & Bilhaut, 2005). Si l'on considère l'ensemble des indices annotés, leur évaluation indique un taux de précision de 93 % et un taux de rappel de 85 %. La section suivante présente les classes générales des indices linguistiques pris en compte.

3.2 Description des indices potentiels de l'obsolescence

En plus de leur diversité, ces indices présentent la caractéristique d'être multi-échelle (*i.e.* d'apparaître à différents niveaux textuels).

Les indices de type intra-phrastique. Une première classe d'indices concerne les **expressions temporelles**. Le temps joue un rôle prépondérant dans les segments d'obsolescence. Notre analyseur temporel repère et annote sémantiquement les adverbiaux temporels : des syntagmes prépositionnels (« dans les années trente », « de 1980 à 2000 »), des adverbes (« aujourd'hui »), des syntagmes nominaux (« les années 20 »). La sémantique temporelle que nous avons mise en œuvre est très simple et exploite un découpage temporel extrêmement simplifié. En effet, étant donnée notre tâche, il nous semble qu'il n'est pas nécessaire de faire appel à des modèles plus complexes qui reflètent mieux la réalité temporelle tels que celui de (Reichenbach, 1966 1ère édition 1947) ou de (Gosselin, 2005). Ainsi, nous exploitons deux types d'information. Tout d'abord, la nature de l'expression. Elle est évaluée selon les valeurs suivantes : *anaphorique* si l'expression temporelle doit être calculée en fonction du contexte (« trois jour avant »), *déictique* si la date doit être calculée selon le moment d'énonciation (« aujourd'hui »), de *durée* si l'expression exprime une durée (« en trente ans »), de type *itération* si le processus est *itératif* (« tous les ans »), *ponctuel* (« En 2008 ») et enfin elle peut être de type *inachevé* lorsque la frontière finale de l'intervalle n'est potentiellement pas refermée (« depuis 1997 »). Le second trait concerne le découpage temporel. Nous considérons simplement les cinq valeurs suivantes : *antériorité++* pour les dates antérieures à 1949, *antériorité* pour les dates de 1950

[4] http ://www.linguastream.org : c'est une plateforme de développement dédiée au T.A.L.

à 1989, *coïncidence* pour les dates de 1990 à 2008, *postériorité* pour les dates après 2009 et *indéterminé* pour les expressions qu'on ne peut pas calculer comme les anaphoriques.

En relation avec les notions de temps, nous prenons également en compte les **indices aspectuels et modaux** à travers, entre autres, le repérage de périphrases verbales et des temps verbaux. Ainsi, nous exploitons les expressions verbales présentant une action dont l'accomplissement débute, est en cours ou achevé. Par exemple, une expression comme « des recherches sont en cours » sera annotée comme une périphrase verbale dont l'accomplissement est en cours.

Les entités nommées semblent également jouer un rôle important dans les segments d'obsolescence. Ce que nous entendons par *entité nommée* est relativement vaste : nous y englobons des expressions de mesure (« 130 hab./km^2 »), de lieu (« à Paris »), de personne (des noms propres principalement), des sigles.

Enfin, une dernière grande classe d'indices qui semble susceptible de marquer l'obsolescence concerne **les expressions exprimant un point de vue**. Le point de vue peut prendre différentes valeurs selon que l'énonciateur se distancie des propos qu'il tient, se les approprie, les juge importants, nouveaux, *etc*. Par exemple, un syntagme prépositionnel comme « Selon les estimations de l'INSEE » sera annoté comme étant de type *source* puisqu'une source précise est indiquée, une expression comme « on suppose » sera annotée comme étant de type *jugement*, un syntagme nominal comme « les nouveaux formats de compression des données » sera annoté comme étant de type *récence* du fait de la présence de l'adjectif « nouveau ».

Les indices positionnels. Ils sont de deux types. Les indices positionnels phrastiques rendent compte de la position des indices intra-phrastiques au sein de l'unité phrase (début et fin de phrase). Les indices positionnels textuels rendent compte de la position des unités phrases au sein des paragraphes (première phrase ou dernière phrase du paragraphe) et des unités paragraphe au sein du document (premier paragraphe ou dernier paragraphe de la section, sous section, *etc.*) ou encore du niveau de hiérarchie dans le document (par exemple les niveaux des titres).

Les indices hiérarchiques correspondent au fait qu'une phrase peut être sous la dépendance d'une autre unité, le titre. Dans notre protocole d'annotation, un titre ne peut pas faire partie d'un segment d'obsolescence. En revanche, il peut être un bon prédicteur d'obsolescence pour les phrases qui sont sous sa dépendance. Ainsi, les indices que nous projetons sur les phrases sont également projetés sur les titres ; les traits sémantiques des titres sont ensuite hérités par les phrases de la section du-dit titre.

Les indices externes concernent par exemple le type de document ou le domaine pour lequel le document est rédigé. Nous exploitons une dizaine de rubriques différentes (histoire, géographie, faune et flore, *etc*).

3.3 Vers des configurations d'indices

Afin de savoir si les indices que nous prenons en compte apparaissent de manière significative dans des segments obsolescents ou non, nous avons effectué un calcul de Chi Carré (χ^2). Nous constatons que sur 146 indices, 63 sont significatifs à $p < .05$ (41 le sont positivement, 22 le

sont négativement), 36 sont non significatifs et enfin pour 47 d'entre eux, il n'y a pas suffisamment d'occurrences pour que le test du χ^2 soit valide. Comme nous le supposions déjà dans (Laignelet, 2006b), la faiblesse d'occurrence de chaque indice pris isolément invalide l'évaluation de nos indices par des tests simples de corrélation tels que le test du χ^2. Nous avons donc mis au point un système de fouille de données qui nous permet de traiter ces indices en termes de combinaison d'indices.

3.4 Classification supervisée à base de règles d'association

On dispose depuis une douzaine d'années d'algorithmes performants pour extraire les motifs fréquents (*i.e.* les motifs présents dans un nombre minimum d'objets) et construire des règles d'association. Une règle $X \rightarrow c$, de prémisse X (une conjonction d'indices), et de conclusion c (une valeur de classe), indique que les phrases contenant la combinaison d'indice X sont de la classe c. Cette association est mesurée par une *fréquence* (nombre d'objets contenant X et c) et la *confiance* (probabilité conditionnelle d'apparition de c connaissant celle de X).

Pour classer à partir de ces règles, diverses méthodes existent (par exemple (Liu *et al.*, 1998; Li *et al.*, 2001)), aux performances comparables, qui suivent le déroulement suivant :

1. extraction de règles d'association non redondantes (*i.e.* à prémisse minimale au sens de l'inclusion), à support et confiance minimum fixés ;
2. pondération des règles par une mesure, par exemple un χ^2 ;
3. sélection des règles suivant leur mesure et leur représentativité pour les classes minoritaires ;
4. un nouvel exemple est classé à l'issue d'un vote réalisé par toutes les règles qui s'appliquent, selon leur poids.

Pour nos expériences, nous avons implémenté une adaptation de (Li *et al.*, 2001), capable de considérer différents types de règles (Rioult *et al.*, 2008) à des fins exploratoires :
– des règles d'associations classiques ;
– des règles d'association *généralisées* (ou disjonctives), qui couvrent les modèles utilisant des règles de forme quelconque : règles prescrivant une classe, excluant une classe, ou contenant des attributs négatifs en prémisse ;
– des règles construites sur des motifs *émergents* : ces motifs indiquent un contraste entre les classes, car ils sont significativement plus présents dans l'une que dans les autres.

L'évaluation de la performance d'un classifieur automatique se fait généralement à l'aide du *score de classification*, c'est-à-dire la proportion d'objets bien classés. Pourtant, cet indice a peu de sens lorsque les classes ne sont pas équilibrées, comme c'est le cas pour notre problème. Pour chaque classe, les mesures de *rappel* (proportion d'exemples découverts), de *précision* (proportion d'exemples correctement attribués) et leur moyenne harmonique, le *F-score* sont des scores traditionnellement utilisés dans les systèmes de recherche d'information. Dans notre cas, pour la classe obsolescente, le rappel est de 78 % et la précision de 34 % : ce résultat est encourageant, d'autant plus que dans un contexte *encyclopédique*, il vaut mieux favoriser le rappel que la précision, car le fait d'oublier une révision est plus grave que d'indiquer inutilement un paragraphe à réviser. Compte tenu de ces performances et des proportions de classe, notre méthode permet de diminuer de deux tiers la tâche du correcteur humain.

Cependant, le rappel et la précision sont caractéristiques d'une classe et non de l'ensemble. Ces mesures évaluent la performance *statique* du classifieur, qui peut être variablement interprétée.

En faisant évoluer la confiance accordée au classifieur, il est possible de construire une courbe dite de ROC (Fawcett, 2003), qui représente pour chaque seuil de confiance le taux de *vrais positifs* (proportion d'éléments bien classées pour la classe positive) et de *faux positifs* (proportion d'éléments mal classées). Cette courbe permet d'optimiser, au choix, le rappel ou la précision, selon les besoins de l'application. De plus, elle offre l'avantage de fournir une mesure indépendante de la répartition des classes. La mesure de l'aire qu'elle définit est un indicateur fiable de la performance du classifieur.

4 Évaluation - Discussion

Dans cette section, nous évaluons la pertinence de nos indices à l'aide d'une méthode de classification supervisée à base d'associations. La mesure de l'aire sous la courbe ROC du classifieur obtenu est un indicateur fiable de la capacité de ces indices à prédire l'obsolescence. Nous analysons ensuite les connaissances découvertes et discutons de leur intérêt d'un point de vue linguistique.

4.1 Résultats en classification supervisée

Comme indiqué à la section 3.4, nous avons expérimenté trois types de règles pour construire un classifieur à partir de nos indices. L'aire sous la courbe ROC obtenue est la suivante :

- pour les règles d'association classique : $79,8\%$
- pour les règles généralisées (dont les prémisses contiennent entre autres des attributs négatifs) : $79,2\%$
- pour les motifs émergents : $76,5\%$

Les trois courbes correspondantes sont représentées à la figure 1. La diagonale représente la performance d'un classifieur aléatoire.

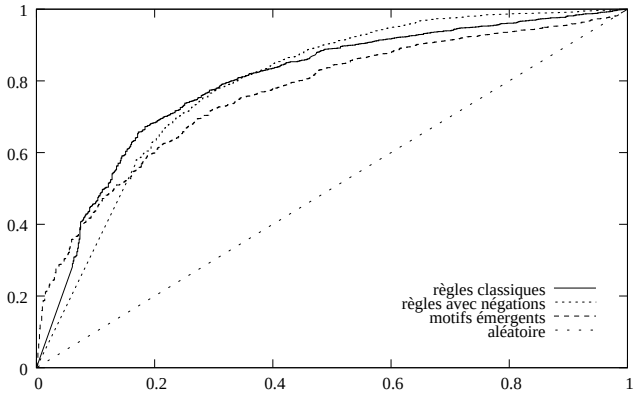

FIG. 1 – Courbes ROC des différents classifieurs. En abscisse le taux de faux positifs, en ordonnée le taux de vrais positifs. Chaque point est obtenu en seuillant la probabilité indiquée par le classifieur.

Ces résultats (autour de 79% quand l'aléatoire donne 50%) montrent que nos indices sémantiques et linguistiques sont pertinents pour la caractérisation de l'obsolescence. On constate que les écarts de performance sont marginaux et que les règles classiques obtiennent les meilleures performances. Dans la suite de la discussion, nous analysons donc la connaissance indiquée par ces règles.

4.2 Analyse des connaissances obtenues

Avant de décrire les règles apprises par le classifieur et pertinentes pour la description et la prédiction de l'obsolescence, en voici un exemple :

entiteNom.classe :mesure ;sousClasse :geopolitique∧title.entiteNom.classe :geopolitique ⟶ classe :obsolescence

Cette règle stipule qu'une entité nommée de type *mesure geopolitique* (« 3000 habitants ») associé à la présence dans un titre d'une entité nommée de type géopolitique (« taux de natalité ») sont importantes pour le repérage des phrases obsolescentes.

Parmi les règles générées, l'**association entre des indices hiérarchiques et des indices intra-phrastiques** est relativement fréquente. Ainsi, les titres comprenant une expression de type géopolitique (« La population ») sont fréquemment associés à un indice de plus bas niveau comme une entité nommée de type *géopolitique* (« 100 000 hab. »), *mesure* (« 78 % ») ou *lieu* (« Madrid », « Barcelone ») ou encore une expression temporelle *déictique* et de type *coïncidence*. La relation est forte également entre des titres contenant un verbe au conditionnel et des phrases dans lesquelles se trouve une entité nommée de type *lieu*.

> (4) ***La population***
> *§ [...] Une quarantaine de villes ont plus de **100 000 hab.**, dominées par les pôles de **Madrid** et **Barcelone**.*

Nous observons également une forte attraction entre des **indices positionnels textuels et des indices intra-phrastiques** : le premier paragraphe d'une division associé à une expression temporelle *déictique* et de type *coïncidence* (« aujourd'hui ») ou à une entité nommée de type *géopolitique* entraîneront souvent l'obsolescence du segment dans lequel l'indice intra-phrastique apparaît. Il en est de même lorsque qu'une entité nommée de type *mesure évolutive* apparaît dans le dernier paragraphe d'une section.

Concernant la **position des phrases au sein des paragraphes**, les premières phrases de paragraphe contiennent souvent des indices temporels *déictiques* de type *coïncidence* ou des entités nommées de type *mesure évolutive* lorsqu'elles sont obsolescentes. De plus, lorsqu'un verbe au conditionnel associé à une entité nommée de type *géopolitique* sont en fin de paragraphe, alors la mise à jour du segment est fortement prévisible.

> (5) *§ L'Union européenne à elle seule se **serait** dépossédée d'un patrimoine de **215 milliards de dollars**.*

L'obsolescence est également mise en valeur par l'association marquée entre plusieurs **indices intra-phrastiques**. Ainsi, une phrase sera obsolescente si une expression temporelle *déictique* de type *coïncidence* cooccurre avec une entité nommée de type *géopolitique* ou de type *mesure évolutive*.

> (6) *§ Les Noirs représentent **aujourd'hui 12 %** de la population ; plus de **50 %** d'entre eux sont **encore** concentrés dans le Sud historique.*

La présence conjointe d'une entité nommée de type *géopolitique* avec une entité nommée de type *mesure évolutive* ou de type *lieu* ou encore de type *géopolitique* est également pertinente pour le repérage des segments obsolescents.

> (7) § Les **industries** de pointe (**11 %** des emplois salariés dans les activités high-tech) sont bien représentées à **Lyon** et à **Grenoble** (électronique, micro- et nanotechnologies).

Des règles mettant en relation **trois niveaux différents d'indices** sont apprises par le système. Ainsi, une phrase sera considérée comme obsolescente si *(i*) elle est la dernière du paragraphe, si *(ii)* le paragraphe est en première position dans la division ou en fin de division et si *(iii)* le titre de la section contient une entité nommée de type *géopolitique*. Il est est de même pour une phrase *(i)* contenant une entité nommée de lieu référant à une ville, *(ii)* qui est située en dernière position de paragraphe et *(iii)* le titre chapeautant la section contient une entité nommée de type *géopolitique*.

Les résultats montrent enfin qu'un **indice de type externe** comme le domaine peut jouer un rôle important. Ainsi, une entrée de type *géographie* va être productive en phrases obsolescentes si la phrase contient une entité nommée de type *mesure géopolitique*.

4.3 Discussion

Les règles que nous venons de décrire confirment nos intuitions concernant la prise en considération d'indices de types différents, de granularité variable ainsi que le fait de les envisager en termes de faisceau, de configuration et non de manière isolée. Nos conclusions vont ainsi dans le sens, entre autres travaux, des résultats de (Marcu, 2000) sur l'importance de la position dans les documents mais également de (HoDac, 2007) et de (Bouffier, 2008) concernant la prise en compte d'indices combinés et à granularité variable. Les travaux de (Teufel, 1999) mettent également en relation des indices de types aussi variés que ceux que nous exploitons dans le but de repérer automatiquement les phrases importantes (*argumentative zoning*) dans des écrits scientifiques à des fins de génération automatique de résumés d'article.

Une dernière remarque concernant la position des indices au sein des phrases nous semble importante à mettre en évidence. En effet, nous sommes surpris de ne pas trouver de relation forte entre la position des indices à l'initiale de la phrase et le fait que les segments soient obsolescents. Contrairement à nos attentes, c'est même la position de fin de phrase qui est pertinente, notamment lorsque cette position est occupée par une entité nommée de type mesure. Ces résultats (quantitatifs) vont à l'encontre des hypothèses que nous avions formulées dans (Laignelet, 2006a) concernant le potentiel structurant des introducteurs de cadre (Charolles, 1997).

5 Conclusion - Perspectives

Le phénomène de l'obsolescence est un phénomène rare (entre 10 et 15 % du nombre de phrases de notre corpus) et relativement flou (le consensus entre les annotateurs n'est pas total). Les indices sémantiques et discursifs que nous repérons et annotons de manière automatique se révèlent, à travers l'aire sous la courbe ROC, être de bons indices pour répondre à notre tâche de recherche de segments obsolescents.

Les règles d'association produites par l'outil de classification automatique nous encouragent dans l'idée que la combinaison d'indices est pertinente pour le repérage automatique de l'obsolescence : d'un côté ces règles s'inscrivent dans la lignée des travaux en linguistique et en T.A.L., de l'autre, les scores fournis (entre 76 et 80 % d'aire sous la courbe ROC selon les types de règles, *cf.* 3.4) valident la pertinence de nos indices. Parallèlement à ce constat, ces résultats nous confirment dans nos choix méthodologiques à considérer des indices de types différents, de niveaux de granularité variés, du mot au discours.

Ce travail ouvre nos perspectives selon deux axes principaux. D'un côté, il nous semble important d'affiner les indices, de les regrouper par *valeurs* sémantiques (par exemple créer un indice *déictique* qui regrouperait des adverbiaux de type *ponctuel* et *inachevé*) dans le but de rendre le modèle plus pertinent et plus efficace. D'un autre côté, il sera nécessaire de projeter les règles apprises automatiquement sur un corpus non annoté manuellement afin de repérer les segments d'obsolescence et de les soumettre à des experts afin qu'ils puissent juger de leur pertinence par rapport à la tâche originelle, à savoir l'aide à la mise à jour de contenus de type encyclopédique.

Références

BOUFFIER A. (2008). *Analyse discursive automatique de textes - Application à la modélisation de textes incitatifs*. PhD thesis, Université Paris Nord - Villetaneuse.

CHAROLLES M. (1997). L'encadrement du discours, univers, champs, domaine et espaces. *Cahiers de Recherche linguistique*, **6**.

EDMUNDSON H. (1969). New methods in automatic abstracting. *Journal of ACM*, **16**(2), p. 264–285.

FAWCETT T. (2003). *ROC Graphs : Notes and Practical Considerations for Researchers*. Rapport interne, HP Laboratories.

GOSSELIN L. (2005). *Temporalité et modalité*. de Boeck.Duculot.

GROSZ J. & SIDNER A. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, **3**(12).

HO-DAC M., JACQUES M.-P. & REBEYROLLES J. (2004). Sur la fonction discursive des titres. Chypre : Actes des 4èmes Journées de Linguistique de Corpus Perspectives.

HODAC M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. PhD thesis, Université de Toulouse 2 - Le Mirail.

HRIPCSAK G. & HEITJAN D. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, **35**(2), 99–110.

LAIGNELET M. (2006a). Les titres et les introducteurs de cadre comme indices pour le repérage de segments d'information évolutive. In *ISDD*, Caen, France.

LAIGNELET M. (2006b). Repérage de segments d'information évolutive dans des documents de type encyclopédique. In *Actes de la 13ème conférence RECITAL*.

LAIGNELET M. (à paraître). *Associer analyse syntaxique et analyse discursive dans un système d'aide à la mise à jour de documents encyclopédiques*. Rapport interne, Université de Toulouse 2 - Le Mirail.

LI W., HAN J. & PEI J. (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining*.

LIU B., HSU W. & MA Y. (1998). Integrating classification and association rules mining. In *International Conference on Knowledge Discovery and Data Mining*, p. 80–86.

MARCU D. (2000). The rhetorical parsing of unrestricted texts : A surface-based approach. *Computational Linguistics*, **26**.

REICHENBACH (1966 (1ère édition 1947)). *Elements of symbolic logic New- york*. Free-Press,.

RIOULT F., ZANUTTINI B. & CRÉMILLEUX B. (2008). Apport de la négation pour la classification supervisée à l'aide d'associations. In F. D'ALCHÉ BUC, Ed., *Conférence francophone d'Apprentissage Automatique (CAP 2008)*, p. 183–196 : Cépaduès.

TEUFEL S. (1999). *Argumentative Zoning*. PhD thesis, Université de Edimbourg.

WIDLOCHER A. & BILHAUT F. (2005). La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conference TALN*, Dourdan, France.