Vers un diagnostic d'ambiguïté des termes candidats d'un texte

Gaël Lejeune, Béatrice Daille
LINA, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France
prenom.nom@univ-nantes.fr

Résumé. Les recherches autour de la désambiguïsation sémantique traitent de la question du sens à accorder à différentes occurrences d'un mot ou plus largement d'une unité lexicale. Dans cet article, nous nous intéressons à l'ambiguïté d'un terme en domaine de spécialité. Nous posons les premiers jalons de nos recherches sur une question connexe que nous nommons le diagnostic d'ambiguïté. Cette tâche consiste à décider si une occurrence d'un terme est ou n'est pas ambiguë. Nous mettons en œuvre une approche d'apprentissage supervisée qui exploite un corpus d'articles de sciences humaines rédigés en français dans lequel les termes ambigus ont été détectés par des experts. Le diagnostic s'appuie sur deux types de traits : syntaxiques et positionnels. Nous montrons l'intérêt de la structuration du texte pour établir le diagnostic d'ambiguïté.

Abstract.

Towards diagnosing ambiguity of candidate terms
Researches in the field of Word Sense Disambiguation focus on identifying the precise meaning of a lexical unit found in a text. This article tackles another kind of problem : assessing the ambiguity of a lexical unit. In other words, we try to identify if a particular unit is ambiguous or not, we define this task as ambiguity diagnosis. Our evaluation dataset contains scientific articles where ambiguous words have been tagged by experts. In order to give an ambiguity diagnosis for each term, we use two types of features : POS tags and positions in the text. We show that the position of an occurrence in the text is a strong hint for such a task.

Mots-clés : diagnostic d'ambiguïté, extraction de mot-clés, terminologie.

Keywords: ambiguity diagnosis, keyword extraction, terminology.

1 Introduction

La désambiguïsation sémantique est un verrou important pour le Traitement Automatique des Langues. Ce problème a souvent été abordé dans une perspective de résolution. Étant donné les sens possibles d'une unité lexicale (mot ou groupe de mots) en contexte, il s'agit de déterminer lequel de ces sens est activé pour une occurence particulière. Ce champ de recherches a été principalement investi à la suite des travaux de Yarowski (1992,1995) bien que les recherches dans le domaine soient bien plus anciennes avec notamment les travaux de Lesk et l'algorithme éponyme (Lesk, 1986). Dans cet article, nous abordons l'ambiguiïté sémantique d'un terme, simple ou complexe, en domaine de spécialité. Nous nous intéressons au diagnostic d'ambiguïté, c'est à dire que nous cherchons à déterminer si, dans un contexte particulier, le sens d'un terme est difficile à appréhender. Par exemples, si l'on a le mot « classe » dans un texte relevant de la linguistique, il s'agit de savoir si l'on a un emploi terminologique (biunivoque) ou non (susceptible d'être ambigu). Il peut revêtir son sens général ou servir d'équivalent référentiel pour un terme plus complexe qui serait son expansion (Jacques, 2003).

Pour l'unité lexicale non-ambiguë, e.g. dont le sens est clair, le choix du sens à activer est trivial : si son emploi relêve d'un domaine de spécialité, il s'agit d'un cas de monosémie. Autrement dit, le nombre d'inférences à effectuer pour déterminer le sens est minimal pour le récepteur du texte (Sperber & Wilson, 1998; Coursil, 2000; Wilson & Sperber, 2004). Si nous nous replaçons dans le domaine de la désambiguïsation sémantique, cela signifie que parmi tous les sens possibles de l'unité lexicale considérée, c'est le plus terminologique qui doit être activé. Détecter les cas d'emploi terminologique permet donc de guider le processus d'analyse. Nous pensons que le diagnostic d'ambiguïté permet de limiter la combinatoire des sens à explorer. Identifier s'il y a une réelle ambiguïté favorise alors la résolution de cette ambiguïté en permettant de savoir si l'on peut ou non se référer au domaine de spécialité concerné. Ce peut aussi être

un indice pour déterminer quels sont les mots-clés pertinents pour décrire un document. En effet, du point de vue de la terminologie en tant que discipline, les termes d'un document sont non-ambigus.

Pour poser les premiers jalons de nos recherches sur le diagnostic d'ambiguïté, nous exploitons ici un corpus de textes en sciences humaines dont les termes candidats ont été classés selon leur degré d'ambiguïté. Dans ce corpus, nous utiliserons des indices syntaxiques et positionnels pour donner pour chaque terme un diagnostic d'ambiguïté. Nous comparons ce diagnostic automatique avec le jugement humain de manière à évaluer la pertinence des indices choisis. Nous détaillerons la problématique de l'ambiguïté dans la section 2 puis nous décrirons le corpus utilisé pour nos expériences ainsi que notre méthodologie dans la section 3.2. Dans la section 4 nous montrerons nos premiers résultats de nos recherches avant de proposer quelques conclusions et pistes pour des recherches futures (section 5).

2 Problématique de l'ambiguïté

La désambiguïsation sémantique (*Word Sense Disambiguation*) est un champ de recherches très actif dans le domaine du TAL. Cette tâche relève de la classification :il s'agit pour chaque occurrence d'un terme de déterminer le sens le plus approprié parmi tous ceux que ce terme peut revêtir. Ce sens de l'occurrence est une étiquette qui selon les ressources exploitées peut revêtir différentes formes : une définition exprimée en langue naturelle, la position dans une ressource de type ontologie ou encore les traductions possibles de ce terme dans différentes langues. Résoudre l'ambiguïté des termes candidats d'un texte permet par exemple d'améliorer les performances des systèmes de traduction automatique. Pour mesurer l'intérêt de cette classification, nous pouvons également donner en exemple le service *Linguee* [1] qui permet de voir en contexte les différentes acceptions d'une unité lexicale.

D'un point de vue méthodologique, la désambiguïsation sémantique a suivi l'évolution du TAL en général. Les travaux répertoriés les plus anciens (Bar-Hillel, 1960; Wilks, 1975; Small & Rieger, 1982) ont traité la désambiguïsation sémantiquecomme un problème de sélection que l'on pourrait résoudre à l'aide de systèmes experts. L'approche la plus emblématique du domaine est due à (Lesk, 1986) qui a exploité les premiers dictionnaires électroniques à large couverture pour utiliser les relations entre les définitions pour raffiner les connaissances sémantiques sur chaque mot. Puis, c'est l'apprentissage automatique qui est devenu en vogue (Gale *et al.*, 1992) ce qui a permis d'améliorer considérablement les résultats et et autorisé l'extension vers de nouveaux domaines et des langues autres que l'anglais. D'autre part, d'autres recherches ont amené de nouvelles problématiques pour le domaine comme la limitation des ressources impliquées (de Loupy & El-Bèze, 2000; Jin *et al.*, 2009) ou l'interprétabilité des modèles générés (Navigli & Velardi, 2005) Les traits exploités dans ces travaux et leurs successeurs sont principalement de deux ordres : classes sémantiques et étiquettes morpho-syntaxiques. Sont considérés les termes à désambiguïser ainsi que leurs voisins selon une certaine fenêtre (n termes avant et/ou après). Une des principales contraintes rencontrées est la largeur de cette fenêtre, plus elle est grande et plus la complexité de calcul est élevée. Avec une fenêtre de taille n et en moyenne m sens par terme à observer, on a une complexité exponentielle en la largeur de la fenêtre. Le choix de cette largeur ne peut donc être qu'un compromis entre efficacité et temps de calcul.

Donner un diagnostic d'ambiguïté permet, par exemple, de limiter le nombre de combinaisons à envisager. Chaque mot non-ambigu permet de réduire la combinatoire pour le calcul du sens de ses voisins ou encore d'élargir à moindre coût le contexte exploré pour améliorer les résultats. Par ailleurs, le diagnostic d'ambiguïté permet d'identifier plus finement les candidats qui sont véritablement des termes pour le document considéré. Pour le terminologue, le mot terminologique est par définition non-ambigu. Diagnostiquer l'ambiguïté revient alors à distinguer en contexte les emplois terminologiques des emplois non-terminologiques. Nous décrivons dans la section suivante, le corpus et la méthode déployée pour aboutir à un diagnostic d'ambiguïté.

3 Description du corpus et de la méthodologie

3.1 Le corpus

Le corpus que nous avons utilisé est constitué de textes scientifiques (articles et communications) relevant des sciences humaines collectés dans le cadre du projet SCIENTEXT [2]. La portion de SCIENTEXT utilisée est composée uniquement

1. `http://www.linguee.org` (consulté le 1er juin 2015)
2. `http://scientext.msg-alpes.fr` (consulté le 1er juin 2015)

de textes en français relevant de la linguistique, de la psychologie des sciences de l'éducation et du traitement automatique des langues. Dans ces textes, des candidats termes ont été identifiés automatiquement en utilisant l'extracteur de termes TERMSUITE, outil librement disponible et *Open Source* [3]. Chacun des candidats a été évalué par un annotateur humain ce qui a permis d'obtenir 4 classes de candidats (DM signifiant Désambiguïsation Manuelle, les modalités précises d'annotation sont disponibles en ligne [4]) :

DM0 Candidat terme rejeté au niveau syntaxique.

DM1 Candidat terme validé au niveau syntaxique. La validation repose sur des critères propres à chaque discipline.

DM3 Candidat terme validé au niveau disciplinaire. La validation repose sur l'appartenance effective du terme au champ scientifique dont relève les textes.

DM4 Candidat terme validé au niveau terminologique. La validation repose sur un emploi véritablement terminologique dans le contexte du document où l'on retrouve le candidat.

La classe DM4 correspond à un usage purement terminologique, et par conséquent non-ambigu, du terme. Pour chaque document, chaque candidat terme est identifié par une classe parmi les quatre décrites ci-dessus. Les documents utilisés sont disponibles au format XML, les données structurelles (sections, paragraphes, listes et légendes) y sont identifiées. Par contre, les informations sur la mise en forme matérielle (graisse, italique…) sont absentes.

3.2 Notre méthode : exploiter la mise en saillance

Nous faisons l'hypothèse que la position des candidats est un indicateur fort de leur ambiguïté. L'idée est que le texte forme un écosystème dans lequel certains candidats termes sont plus mis en valeur que d'autres. C'est une manière pour l'émetteur du texte de faciliter le travail de son lecteur en plaçant ce qui est pertinent pour la compréhension à des positions remarquables. Le nombre de configurations permettant de mettre en valeur les termes est limité (e.g. tout ne peut pas être pertinent). Ceci permet de limiter le nombre d'inférences que doit faire le lecteur pour discriminer ce qui est important de ce qui est secondaire. Ce qui est important doit alors être non-ambigu (Wilson & Sperber, 2004). Nous faisons de plus l'hypothèse que les vrais termes sont globalement « grégaires », c'est à dire que nous les retrouvons souvent ensemble. À l'opposé, ce qui est ambigu est distribué plus uniformément au sein de l'écosystème texte.

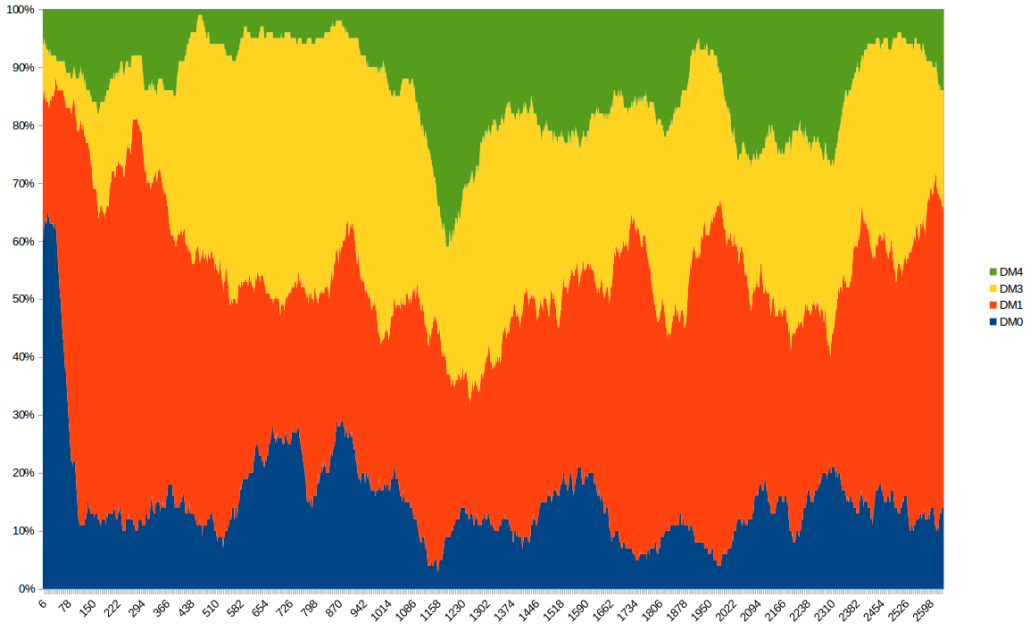

FIGURE 1 – Exemple de distribution des quatre classes au sein d'un texte du corpus

3. https://logiciels.lina.univ-nantes.fr/redmine/projects (consulté le 1er juin 2015)
4. https://apps.atilf.fr/smarties/GuideSmarties.pdf (consulté le 1er juin 2015)

La figure 1 présente la distribution des candidats termes dans chaque classe au fil d'un des textes du corpus. En ordonnée figure le pourcentage de candidats affecté à chaque classe sur chaque série de 100 candidats. En abscisse figure le début de la série considérée. Ainsi, au point d'abscisse zéro figurent la proportion d'appartenance à chaque classe parmi les 100 premières occurrences dans l'ordre du document. Ici, nous pouvons observer que la proportion de candidats appartenant à la classe DM4 (validés au niveau terminologique) est en augmentation dans différentes zones. Si nous alignons dans cet exemple, les parties avec leurs correspondants dans le modèle Introduction-Matériel-Résultats-Discussion ou IMRAD (Sollaci & Pereira, 2004; Bertin & Atanassova, 2014), cela correspond à :

La fin de l'introduction : de 150 à 366 en abscisse ;

La partie méthode : de 870 à 1878 en abscisse ;

Le cœur des résultats : de 1878 à 2310 en abscisse ;

La fin de la discussion : à partir de 2526 en abscisse.

Ce sont donc des zones où sans être majoritaires, les termes véritables du document sont plus fréquemment présents. Autrement dit, un candidat dont les occurrences seraient régulièrement placées dans ces zones aurait une plus forte « propension terminologique ». Les documents que nous étudions sont de deux types différents : articles et communications. Nous décrivons dans la table 1 un certain nombre de statistiques sur le corpus. Nous observons que dans chacun des deux types de textes scientifiques présents dans le corpus, il y a une relative proximité structurelle ainsi qu'une densité proche en termes non-ambigus (DM4). Parmi les candidats extraits par l'extracteur de termes, les occurrences validées au niveau terminologiques ne représentent qu'une minorité : dans les articles moins d'un candidat sur trois est terminologique. Cette proportion est toutefois plus forte dans les communications. Nous proposons dans la section 4 une première série d'expérience visant à identifier au sein de ce corpus des indices positionnels pour identifier automatiquement ces termes non-ambigus.

	Articles	Communications	Corpus combiné
#textes	11	42	53
#parties	251	509	760
parties /texte, moy.(écart-type)	22.82 (± 10.7)	12,12 (± 4.47)	14,34 (± 7.64)
#paragraphes	1350	2318	3668
paragraphes/texte, moy.(écart-type)	122,73 (± 49,93)	55,19 (± 29,14)	69,21 (± 44,05)
paragraphes/parties, moy.(écart-type)	5,66 (± 2,01)	5,08 (± 3,18)	5,2 (± 2,99)
#mots	99942	171119	271061
mots/texte, moy.(écart-type)	9085,64 (± 3116,82)	4074,26 (± 688,25)	5114,36 (± 2553,84)
mots/parties, moy.(écart-type)	420,03 (± 91,35)	378,95 (± 154,33)	387,44 (± 144,51)
mots/paragraphes, moy.(écart-type)	79,16 (± 18,73)	84,23 (± 26,49)	83,18 (± 25,16)
#DM4	**1938**	5785	7723
DM4/texte, moy.(écart-type)	176,18 (± 23,63)	137,74 (± 36,0)	145,72 (± 37,23)
#occurrences DM4	6384	13415	19799
occurrences DM4/texte, moy.(écart-type)	580,36 (± 187,0)	319,4 (± 115,6)	373,57 (± 170,43)
#candidats	**6080**	12447	18527
candidats/texte, moy.(écart-type)	552,73 (± 127,3)	296,36 (± 43,91)	349,57 (± 125,3)
candidats/DM4, moy.(écart-type)	**3,14** (± 0,67)	**2,23** (± 0,39)	2,42 (± 0,59)
#occurrences	22535	31709	54244
occurrences/texte, moy.(écart-type)	2048,64 (± 733,82)	754,98 (± 153,9)	1023,47 (± 637,01)
occurrences/candidat, moy.(écart-type)	3,64 (± 1,0)	2,54 (± 0,34)	2,77 (± 0,7)

TABLE 1 – Statistiques sur différents grains d'analyses disponibles dans le corpus.

4 Résultats

Dans un but exploratoire, nous utilisons ici les indices positionnels de manière brute. L'objectif est de ne pas utiliser de connaissance *a priori* sur le balisage XML exploité. Pour chaque occurrence d'un candidat nous cherchons pour chaque type de balise (en distinguant les fermantes et les ouvrantes), celle qui est la plus proche de celui-ci. Cette proximité est

	Sans étiquettes morpho-syntaxiques								Avec étiquettes morpho-syntaxiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$
Bayésien naïf	0,50	0,52	0,51	0,50	0,50	**0,52**	0,51	0,50	0,50	0,52	0,51	0,50	0,51	0,52	0,51	0,51
Régression logistique	0,47	**0,63**	0,54	0,49	0,75	0,42	0,54	0,65	0,57	**0,66**	0,61	0,59	0,67	0,48	0,56	0,62
Decision Stump	**0,99**	0,41	**0,58**	**0,77**	0,99	0,41	**0,58**	**0,77**	**0,99**	0,41	0,58	**0,77**	**1**	0,41	0,58	0,78
Arbre C4.5 (J48)	0,93	0,40	0,56	0,74	**1**	0,40	0,57	**0,77**	0,74	0,60	**0,66**	0,71	0,94	**0,54**	**0,68**	**0,82**
Baseline	0,34	1	0,51	0,39	0,34	1	0,51	0,39	0,34	1	0,51	0,39	0,34	1	0,51	0,39

TABLE 2 – Résultats de la classification pour la classe DM4 (34,44% des candidats) sur notre fichier initial, la *baseline* classe chaque candidat en DM4.

mesurée en caractères et normalisée en la taille du document [5].Les principales balises existant dans le corpus exploité sont les suivantes :

text l'intégralité du texte avec **title** son titre et **body** l'ensemble de ses sections (hors résumé) ;

div une section (ou sous-section) avec **head** son titre et **p** ses paragraphes ;

list les listes à puces dont les items sont signalés par **item** ;

keywords la liste des mots-clés attribués par les auteurs ;

ref les appels à références.

Les autres balises rencontrées (encodingDesc, addrLine, editor…) présentent un certain nombre de méta-données, donc de « l'extra-texte ». Ces balises ne sont pas exclues du processus d'apprentissage de manière à rester fidèle à l'objectif de ne pas introduire de connaissance *a priori*.

La stratégie que nous employons vise à situer chaque occurrence candidat dans sa position relative avec les éléments qui structurent l'écosystème texte (e.g. les balises ouvrantes et fermantes). Nous avons utilisé les implémentations de classifieurs disponibles dans l'outil WEKA [6]. Nous présentons ici des résultats sur la détection des occurences appartenant à la classe DM4 (cf. Section 3.2) en termes de Rappel, Précision, F_1-mesure et $F_{0,5}$-mesure (de manière à pénaliser les faux positifs). Les vrais positifs sont ici les candidats étiquetés comme non-ambigus (DM4) par les annotateurs et effectivement classés comme tels par notre méthode.

La table 2 [7] présente les résultats des premières expériences menées sur le texte mentionné dans la figure 1. Ce premier test a été mené en effectuant une validation croisée en dix strates. Il s'agit à partir de l'observation d'un seul texte de mesurer si les indices positionnels constituaient de bons traits pour la classification. Nous comparons ces résultats avec ceux obtenus en ajoutant les étiquettes morpho-syntaxiques obtenues à l'issue de la phase de détermination des candidats. Pour les termes complexes, l'étiquette utilisée est la concaténation des étiquettes des éléments lemmatisés qui le composent. Par exemple, le terme complexe « Science du Langage » sera étiqueté « NOM–PRP–NOM ». Nous ne présentons pas ici les résultats obtenus avec les étiquettes seules car ceux-ci sont faibles. Enfin, nous testons deux cas : celui où les 4 classes sont concernées par la phase d'apprentissage et celui où il n'y a que deux classes (DM4 VS le reste). Nous pouvons remarquer d'une part que les traits que nous avons identifiés sont bien adaptés aux arbres de décisions (y compris *Decision Stump* qui n'utilise qu'un règle par classe) et ce d'autant plus que les informations morpho-syntaxiques sont également présentes.

Nous avons réparti les 10 articles scientifiques restants de manière à disposer d'un corpus d'apprentissage de 9 articles et nous avons gardé le dernier pour constituer le jeu de test. La table 3 présente les résultats obtenus sur ce jeu de test à partir du modèle appris. Nous pouvons observer que la régression logistique et le classifieur bayésien ont une nouvelle fois des résultats équilibrés entre rappel et précision. Nous pouvons voir que les arbres de décisions offrent toujours les meilleurs résultats notamment pour ce qui est de la précision. Toutefois l'arbre *Decision Stump* obtient une précision excellente au prix d'un rappel très faible. En comparant avec les résultats présentés dans le tableau 2, nous pouvons remarquer un cas patent de surapprentissage. À l'opposé, l'arbre de décision J48 est plus robuste et offre un équilibre plus intéressant. Ainsi, nous obtenons un premier résultat intéressant pour l'identification des emplois terminologiques. De nombreuses règles utilisées par ces arbres valident notre hypothèse initiale : la position dans la structure du document permet de détecter les termes non ambigus avec une forte précision. Pour donner un exemple de règle extraite : la proximité du candidat avec un début de partie et avec un indicateur de référence bibliographique détermine à plus de 90% la non-ambiguïté du candidat. Autrement dit, la probabilité d'emploi terminologique est d'autant plus forte que l'occurence est proche du début ou de la fin d'une section et d'une référence bibliographique. L'utilisation des étiquettes morpho-syntaxiques accroit de manière significative les résultats.

5. Une distance en mots aurait également pu être exploitée.
6. `www.cs.waikato.ac.nz/ml/weka` (consulté le 1er juin 2015)
7. Nous avons choisi d'écarter ici les classifieurs dont les résultats étaient les moins significatifs, par exemple les séparateurs à vaste marge (SVM).

	Sans étiquettes morpho-syntaxiques								Avec étiquettes morpho-syntaxiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$
Bayésien naïf	0,43	**0,46**	**0,45**	0,44	0,44	**0,45**	**0,44**	0,44	0,49	0,49	0,49	0,49	0,47	0,47	0,47	0,47
Régression Logistique	0,49	0,37	0,42	**0,46**	0,83	0,24	0,37	**0,56**	0,50	**0,62**	**0,55**	0,52	0,73	0,23	0,35	0,51
Decision Stump	**1**	0,14	0,25	0,45	**1**	0,14	0,25	0,45	**1**	0,14	0,25	0,45	**0,82**	0,19	0,30	0,49
Arbre C4.5 (J48)	0,55	0,27	0,36	0,45	**1**	0,16	0,28	0,49	0,55	0,53	0,53	**0,55**	0,75	**0,74**	**0,71**	**0,75**
Baseline	0,32	1	0,48	0,37	0,32	1	0,48	0,37	0,32	1	0,48	0,37	0,32	1	0,48	0,37

TABLE 3 – Résultats de la classification pour la classe DM4 (32,3% des candidats) sur le corpus de test, la *baseline* classe chaque candidat en DM4.

Nous avons appliqué les modèles appris sur les articles scientifiques à la seconde partie de notre corpus, composée uniquement de communications. Les résultats obtenus figurent dans la table 4. Nous observons que le changement de sous-genre scientifique affecte fortement les performances des arbres de décisions. À l'opposé, le classifieur bayésien naïf obtient des résultats plus réguliers. Par ailleurs, la plus-value obtenue en ajoutant les traits syntaxiques est moins nette sur ce sous-corpus. Les indices positionnels pertinents différents entre les deus sous-corpus, les conditions de l'emploi terminologique ne sont pas tout à fait les mêmes. Par exemple, les règles exploitant les références bibliographiques sont moins efficaces, à l'opposé, la proximité avec des items de liste est plus significative dans les communications que dans les articles.

	Sans étiquettes morpho-syntaxiques								Avec étiquettes morpho-syntaxiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$	P	R	F_1	$F_{0,5}$
Bayésien naïf	0,56	**0,50**	**0,53**	**0,55**	0,55	**0,52**	**0,54**	**0,54**	0,62	**0,50**	**0,56**	0,59	0,58	**0,52**	**0,55**	**0,57**
Régression logistique	0,51	0,15	0,23	0,35	0,43	0,49	0,46	0,44	0,55	0,24	0,33	0,44	0,43	0,49	0,46	0,44
Decision Stump	1	0,18	0,31	0,52	**1**	0,17	0,28	0,51	**0,80**	0,30	0,44	**0,60**	**0,86**	0,18	0,30	0,49
Arbre C4.5 (J48)	**0,67**	0,23	0,34	0,48	0,98	0,15	0,26	0,47	0,68	0,26	0,38	0,51	0,85	0,21	0,33	0,53
Baseline	0,39	1	0,56	0,44	0,39	1	0,56	0,44	0,39	1	0,56	0,44	0,39	1	0,56	0,44

TABLE 4 – Résultats de la classification pour la classe DM4 (39,37% des candidats) sur le corpus de communications scientifiques, la *baseline* classe chaque candidat en DM4.

5 Conclusion et perspectives

Nous avons présenté dans cet article quelques pistes pour diagnostiquer l'ambiguïté de candidats termes dans des textes scientifiques. Nous avons fait l'hypothèse que les emplois terminologiques étaient repérables par leur présence à des positions remarquables (ou saillantes) dans les documents. Nous avons défini la saillance comme une mesure proximité vis-à-vis des balises de structure présentes dans les documents XML que nous avons étudié. Nous avons ainsi obtenu un profil des termes non-ambigus que nous avons pu projeter sur de nouveaux documents. En combinaison avec les patrons syntaxiques, ceci nous a permis d'obtenir des premiers diagnostics assez prometteurs en particulier pour détecter avec confiance une certaine proportion des emplois terminologiques. Nous avons identifié des différences significatives entre les règles efficaces pour les articles et les règles efficaces pour les communications. Certains phénomènes observés (proximité avec des débuts et des fins de section par exemple) sont réguliers dans les deux genres tandis que d'autres traduisent de véritables différences entre les deux sous-genres. Il s'agit par exemple de la récurrence des emplois terminologiques dans les items de liste qui est plus marquée dans le sous-genre des communications. Cette première étude devra être approfondie pour mieux combiner les indices positionnels et les indices concernant les candidats termes en eux-mêmes. L'utilisation des lemmes ou des formes prises par chaque terme pourrait ainsi permettre d'améliorer le diagnostic.

Pour approfondir ce diagnostic, il serait intéressant de l'évaluer en fonction d'une tâche particulière. Ce peut-être la désambiguïsation sémantique, l'extraction de termes ou encore de l'aide à l'écriture (e.g. les termes clés sont ils bien placés dans le texte). Pour aller plus loin, une piste serait de traiter des documents utilisant d'autres jeux de balises. Il s'agirait d'identifier automatiquement en contexte quelles balises déterminent des positions remarquables et quelles balises sont à écarter. Le modèle serait de ce fait plus robuste à la variation en genre que les modèles obtenus dans nos expériences. Ainsi, nous pourrions mettre la méthode décrite ici en liaison avec les travaux portant sur l'utilisation du modèle IMRAD dans les textes scientifiques, notamment dans les communications qui est un sous-genre que l'on pourrait considérer comme moins normé. Enfin, si le genre est une variable importante pour l'efficacité de la méthode, il serait intéressant d'étudier cette fois le même genre mais dans différentes langues.

Références

BAR-HILLEL Y. (1960). *Automatic Translation of Languages*. Academic press, New York.

BERTIN M. & ATANASSOVA I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. In *Bibliometric-enhanced Information Retrieval Workshop at the 36th European Conference on Information Retrieval (ECIR-2014)*, Amsterdam, Netherlands.

COURSIL J. (2000). *La fonction muette du langage*. Ibis Rouge.

DE LOUPY C. & EL-BÈZE M. (2000). Using few clues can compensate the small amount of resources available for word sense disambiguation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* : European Language Resources Association (ELRA).

GALE W. A., CHURCH K. W. & YAROWSKY D. (1992). Using bilingual materials to develop word sense disambiguation methods.

JACQUES M.-P. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. Doctorat Nouveau Régime, Université Toulouse II Le Mirail, Toulouse. 3.

JIN P., MCCARTHY D., KOELING R. & CARROLL J. (2009). Estimating and exploiting the entropy of sense distributions. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, NAACL-Short '09, p. 233–236, Stroudsburg, PA, USA : Association for Computational Linguistics.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, p. 24–26, New York, NY, USA : ACM.

NAVIGLI R. & VELARDI P. (2005). Structural semantic interconnections : A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**(7), 1075–1086.

SMALL S. & RIEGER C. (1982). Parsing and comprehending with word experts (a theory and its realization). In W. G. LEHNERT & M. H. RINGLE, Eds., *Strategies for Natural Language Processing*, p. 89–147. Hillsdale, NJ : Erlbaum.

SOLLACI L. & PEREIRA M. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the Medical Library Association : JMLA*, **92**(3), 364–367.

SPERBER D. & WILSON D. (1998). *Relevance : Communication and cognition*. Blackwell press, Oxford U.K.

WILKS Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, p. 53–74.

WILSON D. & SPERBER D. (2004). *Relevance theory*. Blackwell press, Oxford U.K.

YAROWSKY D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the International Conference on Computational Linguistics, COLING 1992*, p. 454–460.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, p. 189–196.