Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intégral

José Camacho-Collados[1], Mokhtar Boumedyen Billami[1],
Evelyne Jacquey[1], Laurence Kister[2]

[1] UMR ATILF CNRS Université de Lorraine, Prénom.Nom(-Nom)@atilf.fr
[2] Université de Lorraine UMR ATILF CNRS, Prénom.Nom@univ-lorraine.fr

Abstract

Following (L'Homme, 2004), this paper focuses on terms variations in full text in French and more precisely it highlights the semantic ambiguity of terms occurrences with regards to a very high leveled distinction between terminological and general uses. This issue is very present especially in Humanities. For instance, we are interested in distinguishing between the terminological meaning of the term "sujet (*subject*)" in the phrase "le sujet de la phrase (*the subject of the sentence*)" (Linguistics) or "les réponses du sujet (*subject's answers*)" (Psychology), and the general meaning of the noun "sujet (*topic*)" that we may find in a phrase like "le sujet de cet article (*the topic of this article*)". In order to solve this problem, we assume that textual contexts around term occurrences give us relevant information on the kind of use we face, terminological or general. Our research is based on a statistical approach of the textual contexts. The proposed metrics are based on the hypergeometric distribution and the lexical specificity calculus as described in (Lafon, 1980). By using a manually annotated corpus as the training set, we build lexical profiles for each high leveled meaning of the term candidates. We use two methods which were compared to a baseline metric based on term frequency. The results we obtained are analyzed from both a quantitative and a qualitative point of view.

Résumé

A la suite de (L'Homme, 2004), nous nous intéressons à la variation des termes en texte intégral et, en particulier à l'ambiguïté de leurs occurrences entre usage terminologique relevant d'un domaine de spécialité et usage non terminologique. Cette question est particulièrement sensible dans le cas des sciences humaines et sociales où il s'agit de pouvoir différencier un sens terminologique de "sujet", dans "le sujet de la phrase" (linguistique) ou dans "les réponses du sujet" (psychologie), et un sens général dans "le sujet de l'article", "le sujet de la conversation", etc. Pour contribuer à répondre à cette question, nous faisons l'hypothèse que ce sont les contextes autour des occurrences des candidats (et plus spécifiquement les paragraphes où elles se trouvent) qui donnent des indices sur le type d'usage, terminologique ou non terminologique, dont elles relèvent. L'exploitation des contextes autour des occurrences s'appuie sur une approche statistique.
La méthode statistique employée est basée sur la distribution hypergéométrique et la notion de spécificité lexicale de Lafon (1980). Afin de déterminer le type d'usage d'une occurrence, on établit, à partir d'un corpus préalablement annoté manuellement, un profil statistiquement fondé de toutes les occurrences terminologiques et de toutes les occurrences non terminologiques. Ensuite, l'algorithme compare les profils statistiques établis, terminologique vs. non terminologique, avec les éléments du contexte de chaque occurrence.
Les expériences ont été faites sur un corpus d'articles complets en linguistique extraits de la base Scientext. Les résultats obtenus sont évalués sur le plan quantitatif et qualitatif.

Mots-clés : terminologie, variantes terminologiques, ambiguïté sémantique, filtrage statistique

1. Introduction

Dans le domaine de l'extraction terminologique à partir de textes intégraux (Bourigault et al., 2001), sur la question de la validation des candidats termes proposés, l'usage courant consiste à présenter des listes de candidats, parfois contextualisés à l'aide d'exemples d'usage, à des experts du domaine. Les travaux que nous présentons se positionnent de manière

complémentaire en présentant à l'évaluation un ensemble de contextes d'utilisation du candidat terme. Autrement dit, notre approche consiste à déterminer le caractère terminologique de chaque occurrence de candidat terme au sein d'un corpus de textes intégraux relevant d'un domaine spécialisé ou d'un domaine scientifique. À partir du corpus ainsi analysé, notre objectif général est de contribuer à l'automatisation de l'évaluation terminologique de chaque occurrence des candidats termes.

Comme l'ont montré parmi d'autres (Jacquemin, 1996), (L'Homme, 2004), les occurrences de candidats termes peuvent varier sur différents plans. Les variations morpho-syntaxiques et syntaxiques sont suffisamment bien connues aujourd'hui pour avoir été intégrées dans différents outils d'extraction terminologique. C'est ainsi le cas de Fastr (Jacquemin, 1997), Yatea (Aubin et Hamon, 2006), d'Acabit (Daille, 1994, 2003 ; Toussaint et al., 1998), de Termostat (Drouin, 2003)[1]. En revanche, les variations sémantiques des occurrences de candidats termes représentent une question encore largement ouverte.

Dans ce cadre, l'ambiguïté sémantique des candidats termes est un des vecteurs importants de variation. En effet, toutes les occurrences de candidats termes ne relèvent pas nécessairement d'un emploi terminologique, ni toujours du même domaine scientifique[2]. Ceci est particulièrement vrai en sciences humaines et sociales où il faut pouvoir différencier deux grands types d'emploi pour le candidat *sujet* par exemple.

- Emplois terminologiques

le sujet de la phrase (terminologique en linguistique)

les réponses du sujet (terminologique en psychologie)

- Emplois non terminologiques

le sujet de cet article

Relativement à la question de la différenciation entre emploi terminologique ou non terminologique, nous faisons l'hypothèse que les contextes peuvent fournir des indices utiles. À partir d'occurrences de candidats termes annotées manuellement selon la distinction terminologique vs. non terminologique, nous avons développé plusieurs méthodes d'analyse qui visent à déterminer, pour chaque type d'emploi, les éléments de contextes qui sont statistiquement significatifs. A cette fin, nous avons comparé les résultats obtenus à l'aide de deux méthodes, et leurs variantes. La première méthode calcule deux scores de spécificité (construits à partir de la spécificité lexicale de (Lafon, 1980)). La seconde méthode s'appuie sur le théorème de Bayes.

L'ensemble des travaux présentés s'inscrit dans le cadre du projet TERMITH[3] et s'appuie sur l'utilisation du corpus libre SCIENTEXT[4] et de l'extracteur libre TTC-TERMSUITE[5].

---

[1] Yatea : http://taln09.blogspot.fr/2009/03/description-lextracteur-de-termes-yatea.html ;
Acabit : http://taln09.blogspot.fr/2009/03/acabit-acquisition-de-termes-partir-de.html ;
Fastr : http://perso.limsi.fr/jacquemi/FASTR/ ;
Termostat : http://termostat.ling.umontreal.ca/ [pages consultées le 21/11/2013]

[2] Il est connu que l'ensemble des sciences et plus particulièrement les sciences humaines et sociales sont perméables entre elles, fait qui se perçoit aisément si l'on interroge une base terminologique comme TermSciences (http://www.termsciences.fr/) où le même terme est présent dans plusieurs disciplines scientifiques avec une acception spécifique à chacune d'elles.

[3] Le projet TermITH bénéficie d'une aide de l'Agence Nationale de la Recherche (ANR-12-CORD-0029).

[4] http://scientext.msh-alpes.fr/scientext-site/spip.php?article1 [pages consultées le 21/11/2013]

2. Méthodologie

Notre méthodologie est appliquée sur des données textuelles enrichies en candidats termes dont le caractère terminologique des occurrences a été évalué manuellement. À l'issue de cette étape préparatoire, un décompte des occurrences validées et rejetées est établi pour chaque candidat terme : ce décompte permet de calculer un taux d'ambiguïté et de positionner le candidat sur une échelle allant de [très peu terminologique] à [très terminologique]. Par ailleurs, à partir des données textuelles annotées et pour chaque candidat, il est possible de constituer deux sous-corpus : un premier sous-corpus SC_{on} contenant les contextes (ici les paragraphes) des occurrences validées (jugées terminologiques) du candidat et un second sous-corpus *SCoff* contenant les contextes des occurrences rejetées (jugées non terminologiques) du candidat. En nous fondant sur différents types d'analyse statistique de chaque sous-corpus de chaque candidat, il est possible de construire différents profils lexicaux, statistiquement fondés et supposés caractéristiques des emplois terminologiques pour les uns et des emplois non terminologiques pour les autres. Enfin, pour analyser chaque occurrence d'un candidat, nous comparons le contexte de cette occurrence à chaque profil statistiquement fondé, respectivement caractéristique d'un usage terminologique et d'un usage non terminologique, et déduisons automatiquement une évaluation du caractère terminologique de l'occurrence analysée. Nous reproduisons ce type d'analyse pour tous les contextes de dix candidats termes dont le choix est expliqué dans la section (3.1). De plus, nous évaluons l'adéquation des différentes méthodes d'analyse statistique définies (section 2.2) pour la désambiguïsation des occurrences de candidats termes à l'aide des mesures courantes que sont les mesures de rappel et de précision, la F-mesure et le taux d'exactitude *(accuracy*).

2.1. Données de travail

2.1.1. Corpus

Le corpus utilisé rassemble 62 articles appartenant au domaine scientifique des Sciences du Langage. Ce corpus est extrait du corpus libre de droits mis à la disposition de la communauté scientifique par le projet ANR SCIENTEXT sous licence *Creative Common*[6]. Le corpus utilisé, au format XML-TEI, comporte 397 695 occurrences. L'ensemble des textes se répartit en 47 articles de conférences, soit 75,81% des articles et 57,06% des occurrences, et 15 articles de revues, 24,19% des articles et 42,94% des occurrences. Ainsi, en nombre d'occurrences, le corpus utilisé est assez équilibré entre conférences et revues. Les conférences représentées sont le Cédil (Colloque international des Étudiants chercheurs en Didactique des Langues et en Linguistique), Euralex (Conférence de *European Association of Lexicographie*) et le colloque EID (Émotions, Interactions, Développements). Les revues sont Tal (Traitement automatique des langues), Les cahiers de grammaire et LiDil (Revue de Linguistique et de Didactique des langues).

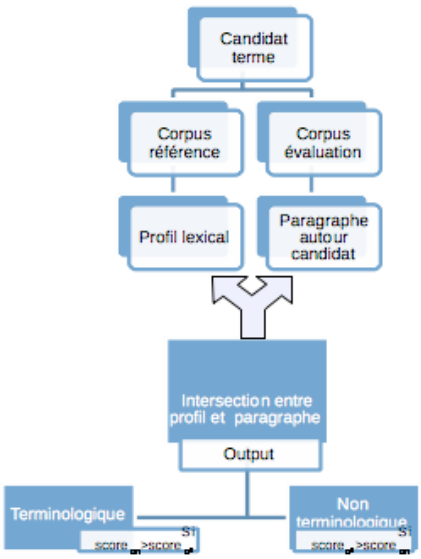
[5] https://code.google.com/p/ttc-project/ [pages consultées le 21/11/2013]

[6] http://scientext.msh-alpes.fr/scientext-site/?article8 [pages consultées le 21/11/2013]

2.1.2. Annotation terminologique manuelle

Le corpus d'articles en Sciences du Langage est traité par l'extracteur automatique de termes *TTC-TermSuite*, qui est librement utilisable et open-source[7], afin d'obtenir une liste de candidats termes. Les candidats termes obtenus sont projetés dans le corpus d'articles, aboutissant ainsi à une version enrichie des données. Enfin, les données enrichies sont évaluées manuellement au sein d'une interface d'annotation librement consultable[8], c'est-à-dire que l'ensemble des occurrences de candidats termes sont matérialisées par le biais d'une mise en couleur et des crochets afin de représenter les bornes du candidat et par le biais d'une puce représentant le choix de l'annotateur (la couleur verte correspond à une validation et la couleur rouge correspond à un rejet). A l'issue de cette annotation manuelle, l'ensemble des évaluations sont stockées et décomptées globalement. Parmi les 77 014 occurrences de candidats termes, correspondant à 23 199 candidats termes différents, 18 142 occurrences sont validées par l'annotation manuelle, soit 23,56 % des occurrences candidates. Ces informations sont, par ailleurs, stockées pour chaque candidat. Ainsi, pour le candidat *structure*, ce décompte permet de mesurer, d'une part, le taux d'ambiguïté de ses occurrences prises dans leur ensemble et, d'autre part, de situer ce candidat comme relevant à 6,32 % d'un usage terminologique. Ces deux mesures sont deux interprétations possibles du même résultat obtenu par un ratio du nombre d'occurrences validées sur le nombre total d'occurrences apparaissant dans le corpus. Pour aller plus loin et comprendre comment est effectué l'évaluation des candidats termes depuis leur état initial vers l'état de leur validation terminologique, nous renvoyons à (Kister et al., 2012).

2.2. Méthodes implémentées d'analyse statistique



L'objectif général est de déterminer automatiquement si une occurrence d'un candidat terme donné est terminologique ou non. Les méthodes proposées suivent le cheminement ci-contre qui passe par la définition de deux profils lexicaux pour chaque candidat terme, l'un pour l'usage terminologique, l'autre pour l'usage non terminologique. La définition des profils lexicaux repose sur l'utilisation de la mesure de spécificité lexicale de Lafon (1980).

Figure 1. Schéma général des méthodes d'analyse statistique implémentées

[7] https://code.google.com/p/ttc-project/downloads/detail?name=ttc-term-suite-1.4.jar [page consultée le 21/11/2013]

[8] https://arcas.atilf.fr/smarties [page consultée le 21/11/2013]

2.2.1. Établissement des profils lexicaux pour chaque candidat terme

Pour représenter les contextes et les indices lexicaux qu'ils contiennent, la mesure de spécificité lexicale (Lafon, 1980 ; Drouin, 2007) semble particulièrement adaptée parce qu'elle est moins sensible que d'autres aux variations de taille des ensembles comparés entre eux. Cette mesure, basée sur la distribution hypergéométrique, permet de connaître la sur/sous-représentation d'éléments lexicaux dans une partie donnée au sein d'un corpus. Autrement dit, si tous les contextes dans lesquels les occurrences d'un candidat terme relèvent bien d'un usage terminologique étaient rassemblés en un sous-corpus et qu'un calcul de spécificité y était appliqué, alors ce calcul devrait permettre de connaître les éléments lexicaux sur-représentés et sous-représentés dans ce sous-corpus. Plus précisément, pour un élément lexical E dont on veut calculer le taux de spécificité lexicale dans un sous-corpus donné, quatre paramètres sont nécessaires : la taille du corpus de référence (T), la taille du sous-corpus (t), le nombre d'occurrences de E dans le corpus de référence (f), le nombre d'occurrences de E dans le sous-corpus (k). Le calcul de spécificité définit une probabilité en considérant le sous-corpus comme un échantillon aléatoire du corpus de référence. Si X est une variable aléatoire qui suit une distribution hypergéométrique, la sur-représentation correspond à un taux de spécificité positif (probabilité de (X>=k)) et celle de la sous-représentation correspond à un taux de spécificité négatif (probabilité de (X<=k)). Enfin, à la suite de (Heiden et al., 2010), la valeur du taux de spécificité est le résultat d'une transformation logarithmique en base 10 de la probabilité calculée. Pour résumer, un taux de spécificité de 1 indique que la probabilité d'observer au moins la fréquence *k* est de 0,1. Plus les taux de spécificité positifs sont élevés, plus la sur-représentation de l'élément *E* dans le sous-corpus est forte.

En utilisant le corpus de référence et en faisant tourner l'algorithme de comptage et de repérage des occurrences terminologiques et non terminologiques pour chaque candidat terme, il est possible de regrouper respectivement tous les paragraphes contenant une occurrence terminologique d'un candidat dans un sous-corpus (SC_{on}) et tous les paragraphes contenant une occurrence non terminologique d'un candidat dans un autre sous-corpus (SC_{off}). En appliquant un algorithme de calcul du taux de spécificité à la Lafon (1980) dont les résultats ont été vérifiés par comparaison avec ceux obtenus à l'aide du logiciel textométrique TXM (Heiden, 2010), nous produisons une liste de mots supposés caractéristiques d'un usage terminologique (appelée LS_{on}) et une liste de mots supposés caractéristiques d'un usage non terminologique (appelée LS_{off}). Tous les mots présents dans chacune de ces listes sont représentés avec leurs taux de spécificité. Par exemple, pour le candidat terme *« définition »*, on obtient de cette manière ses deux profils lexicaux (Figure 2).

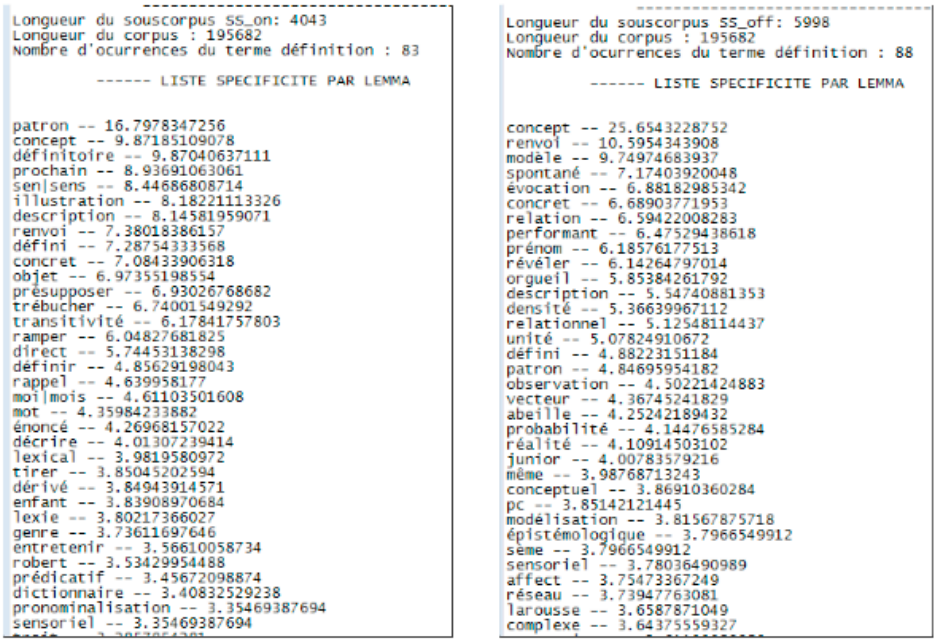


Figure 2. Listes de spécificités du candidat "définition"

2.2.2. Méthode de décision fondée sur la spécificité lexicale

Pour déterminer automatiquement si une occurrence donnée d'un candidat terme est terminologique ou non, nous comparons les mots sémantiquement pleins du contexte de cette occurrence avec les listes d'éléments spécifiques caractéristiques d'un emploi terminologique LS_{on} ou non terminologique LS_{off} pour ce candidat. Cette comparaison a pour objectif d'établir deux scores. Le score "score_on", par exemple, est obtenu en additionnant[9] les taux de spécificité des mots pleins qui sont présents à la fois dans le contexte de l'occurrence à désambiguïser et dans la liste LS_{on} du candidat correspondant. Le score "score_off" est le résultat du même calcul en utilisant la liste LS_{off} correspondante. L'occurrence à désambiguïser est considérée comme terminologique si le "score_on" est supérieur au "score_off".

$$-\log_{10}\left(\prod_{i=0}^{n} P(X_{on} \geq k_i)\right) = \sum_{i=i}^{n} -\log_{10}(P(X_{on} \geq k_i)) = \sum_{i=0}^{n} (scores_i)$$

$$-\log_{10}\left(\prod_{i=0}^{m} P(X_{off} \geq k_i)\right) = \sum_{i=0}^{m} -\log_{10}(P(X_{off} \geq k_i)) = \sum_{i=0}^{m} (scores_i)$$

avec n, le nombre de mots du contexte courant qui appartiennent à la liste de spécificité LS_{on} et m, le nombre de mots du contexte courant qui appartiennent à la liste de spécificité LS_{off}.

Quatre variantes de la méthode de calcul des scores ont été mises en œuvre en fonction du seuil minimal de spécificité choisi pour les listes de spécificité LS_{on} et LS_{off} (1 ou 1,5 en valeur absolue) et de la mise en place ou non d'une normalisation. La normalisation des taux de spécificité consiste simplement à diviser chaque taux par la somme de tous les taux de la liste.

[9] Comme nous nous sommes placés dans le contexte d'une transformation algorithmique pour représenter les taux de spécificité, rendre compte de l'ensemble des taux de spécificité, qui correspondent à des probabilités d'événements indépendants, suppose de les additionner au lieu de les multiplier.

	|spécificité| >=1	|spécificité|>=1,5
taux bruts	(1a)	(1c)
taux normalisés	(1b)	(1d)

Tableau 1. Variantes définies sur la méthode des scores de spécificité

2.2.3. Méthode Bayes

Selon cette méthode, l'idée est de calculer la probabilité d'obtenir la distribution des mots spécifiques vus dans le contexte. On ne se limite pas seulement aux mots communs entre le contexte du candidat terme à désambiguïser et les listes de spécificités LS_{on} et LS_{off} mais on calcule les probabilités d'observer la distribution D du contexte lorsque le candidat terme est supposé être terminologique ou non terminologique. En appliquant le théorème de Bayes, on obtient les formules suivantes :

$$P(on/D) = \frac{P(D/on)P(on)}{P(D/on)P(on) + P(D/off)P(off)}$$

$$P(off/D) = \frac{P(D/off).P(off)}{P(D/on).P(on) + P(D/off).P(off)}$$

Le maximum de ces deux probabilités donne une clé pour décider si le candidat a un usage terminologique ou non. Le maximum de ces probabilités est obtenu à partir d'une formule plus simple que l'application du théorème de Bayes :

$$\max\{P(on/D) , P(off/D)\} = \max\{P(D/on).P(on) , P(D/off).P(off)\}$$

Pour mettre en œuvre ce calcul, on parcourt successivement les deux listes de spécificités lexicales LS_{on} et LS_{off}. On commence à parcourir la liste LS_{on} mot par mot et on calcule $P(Y_{on} \leq k'_i)$ en considérant « Y » une variable aléatoire suivant une distribution hypergéométrique sur le contexte, et en considérant que « k'_i» représente le nombre d'occurrences du mot spécifique « i » dans le contexte. Dans cette distribution, on prend la distribution du sous-corpus SC_{on} (tous les paragraphes autour d'un candidat terme donné dans le corpus Scientext) comme référence et le contexte (paragraphe) autour de ce candidat terme comme échantillon (voir section 2.2.1 pour une explication détaillée avec le candidat terme *définition*). On applique cela pour tous les mots de la liste LS_{on} et de cette façon on obtient P(D/on) comme suit :

$P(D/on) = \prod_{i=1}^{n'} P(Y_{on} \leq k'_i)$	n' est la taille de la liste LS_{on} et k'_i est le nombre d'occurrences de l'élément « i » de la liste LS_{on} dans le paragraphe.

La même procédure est appliquée pour calculer P(D/off). À partir de cette seconde méthode, quatre variantes ont été élaborées.

	P(on) = P(off) = 0,5 hypothèse d'indépendance de la désambiguïsation manuelle	P(on) et P(off) pondérées par les taux d'ambiguïté déduits de la désambiguïsation manuelle
Listes de spécificités non normalisées	(2a)	(2b)
Listes de spécificités normalisées[10]	(2c)	(2d)

Tableau 2. Variantes définies sur la méthode Bayes

2.2.4. Méthode basée sur les fréquences relatives (Baseline)

Cette méthode prend les mots pleins d'un contexte (paragraphe) du candidat terme à désambiguïser et calcule le nombre d'occurrences dans le sous-corpus SC_{on}. Si le nombre d'occurrences est supérieur à 5 dans le sous-corpus SC_{on}, on additionne toutes ces fréquences. Par la suite, on divise le total par la taille de SC_{on}. On répète le même processus avec le sous-corpus SC_{off} et on choisit la valeur supérieure. Cette valeur fournit la décision qui serait prise automatiquement en fonction des fréquences relatives et joue donc le rôle de *baseline* pour notre expérience. On cherchera donc à voir si les deux méthodes proposées et leurs variantes apportent une amélioration par rapport à cette *baseline*.

3. Expériences menées

3.1. Données de l'expérience

Le décompte exhaustif des occurrences validées et rejetées pour chaque candidat sur l'ensemble du corpus de travail a permis de positionner chaque candidat sur un continuum allant de [très peu terminologique] à [très terminologique]. Ce positionnement dépend du taux d'ambiguïté observé qui correspond, pour chaque candidat terme, au ratio du nombre d'occurrences validées par rapport au nombre total d'occurrences. Sur cette base, nous avons appliqué les différentes exploitations statistiques définies dans la section (2.2) ci-dessus, sur cinq catégories établies en fonction de quatre seuils d'ambiguïté. Pour chaque catégorie, deux candidats termes ont été choisis en fonction de la fréquence de leurs occurrences (faible ou forte). De cette manière, dix candidats ont été choisis pour l'évaluation des différentes méthodes d'exploitation. Ils sont représentés dans le tableau ci-dessous.

Candidat	Taux d'ambiguïté	Fréquence	Catégorie
collocation	97,94	97	[très term]
diglossie	93,55	31	
mot	80,11	543	[assez term]
syntagme	88,37	43	
défintion	48,54	171	[amb]
vocable	68,18	22	
modèle	23,9	272	[peu term]
entrée	18,03	61	
structure	6,32	269	[très peu term]
statut	1,96	51	

Tableau 3. Candidats choisis pour l'évaluation

[10] Pour normaliser les listes de spécificités LS_{on} et LS_{off}, nous supprimons les mots pleins de plus faible spécificité dans la liste la plus grande jusqu'à ce que les deux listes soient de même taille.

3.2. Résultats

3.2.1. Analyse quantitative

La première étape d'analyse consiste à rassembler l'ensemble des mesures d'évaluation (taux d'exactitude ou *accuracy*, précision, rappel et F-mesure) pour l'ensemble des candidats sélectionnés pour l'expérience. Le tableau 4 ci-dessous résume les résultats pour le candidat terme *modèle*, qui a un taux d'ambiguïté de 23.9% et apparaît 272 fois dans le corpus (65 occurrences sont évaluées manuellement comme terminologiques et 207 comme non terminologiques).

Modèle (23,9%)	Nbre d'occurrences validées	Nombre d'occurrences rejetées	Taux d'exactitude (%)	Précision (%)	Rappel (%)	F-Mesure (%)
1a	78	194	**93,8**	80,77	96,92	**88,11**
1b	78	194	93,01	79,49	95,38	86,71
1c	151	121	68,38	43,05	**100**	60,19
1d	153	119	67,65	42,48	**100**	59,63
2a	126	146	77,57	51,59	**100**	68,06
2b	0	272	76,1		0	0
2c	33	239	88,24	**100**	50,77	67,35
2d	0	272	76,1		0	0
3 (baseline)	132	140	75,37	42,24	**100**	65,99

Tableau 4. Résumé des résultats par méthode pour la désambiguïsation de « modèle »

Comme on peut le voir dans le tableau 4, il y a des différences notables dans la désambiguïsation des occurrences du candidat terme pour chaque méthode. Même la pondération joue un rôle important : pour cet exemple la pondération est devenue excessive pour les méthodes (2b) et (2d), même si les taux d'exactitude ne sont pas très différents selon les méthodes. Le cas de *modèle* est un cas extrême mais ça nous conduit à penser que la méthode (2c) (avec la normalisation des listes de spécificité) est peut-être une pondération plus adéquate. Une normalisation des listes de spécificités plus fine entre la méthode (2a) et (2c) semble être une idée intéressante, puisque la méthode (2a) tend à avoir un rappel assez proche de 100 %, tandis que la méthode (2c) tend vers une précision très élevée. Dans la suite, nous donnons les taux d'exactitude par méthode pour l'ensemble des candidats termes analysés dans l'expérience.

Le tableau 5 ci-dessous résume les résultats pour les cinq candidats termes peu fréquents dans le corpus de référence. On voit, parmi les méthodes (1), (1a) et (1b), lesquelles sont les moins sensibles au taux d'ambiguïté établi à partir des désambiguïsations manuelles. Ceci pourrait être expliqué par le fait que la normalisation réalisée avec les méthodes (1c) et (1d) diminue leur performance avec les candidats termes ambigus (terminologique ou non terminologique). Cependant, les méthodes (1c) et (1d) se comportent très bien avec les candidats termes ambigus, avec des mesures de précision et de rappel très différents de ceux qu'on obtient avec les méthodes (1a) et (1b). Ceci suggère que la normalisation n'est pas assez systématique et ne permet de détecter des occurrences terminologiques avec une précision supérieure que dans quelques cas.

	statut	entrée	vocable	syntagme	diglossie
1a	**98,04**	**81,97**	86,36	**90,69**	**93,55**
1b	**98,04**	**81,97**	86,36	**90,69**	**93,55**
1c	49,02	80,33	86,36	79,06	35,48
1d	49,02	78,69	86,36	79,06	54,84
2a	17,65	**81,97**	**95,45**	79,06	16,13
2b	80,39	80,33	81,82	**90,69**	**93,55**
2c	80,39	**81,97**	**90,91**	**90,69**	83,87
2d	92,16	**81,97**	81,82	**90,69**	**93,55**
3 (baseline)	92,16	78,69	86,36	**90,69**	**93,55**

Tableau 5. Taux d'exactitude (accuracy) des méthodes pour les candidats termes peu fréquents dans le corpus (ordre croissant par rapport au taux d'ambiguïté)

Le tableau 6 résume les résultats obtenus pour les cinq termes fréquents à analyser. Les résultats reflètent à nouveau la robustesse des méthodes (1a), (1b), (2b), (2c) et (2d) par rapport au taux d'ambiguïté établi sur la base des désambiguïsations manuelles. Par ailleurs, on peut noter les faibles résultats des méthodes (1c), (1d) et (2a) principalement du fait d'un manque de pondération. La méthode (3), basée sur des fréquences relatives (*baseline*) a une robustesse significative mais ne permet pas de désambiguïser clairement une grande partie des occurrences des candidats termes.

	structure	modèle	définition	mot	collocation
1a	95,17	**93,8**	**84,21**	80,85	**97,94**
1b	94,05	93,01	83,04	80,66	**97,94**
1c	64,31	68,38	80,7	50,28	67,01
1d	65,06	67,65	80,7	48,62	67,01
2a	75,46	77,57	63,74	30,57	5,15
2b	94,8	76,1	55,56	79,92	92,78
2c	**95,54**	88,24	61,99	**81,03**	84,54
2d	94,8	76,1	54,39	80,11	94,85
3 (baseline)	88,48	75,37	70,76	**83,79**	**97,94**

Tableau 6. Taux d'exactitude des méthodes pour les candidats termes fréquents dans le corpus (classement par ordre croissant d'ambiguïté)

Par la suite, il pourrait être utile de répéter les expériences avec d'autres corpus dans le domaine des Sciences du Langage et en augmentant aussi le nombre de données analysées, afin de connaître la robustesse réelle de chaque méthode et de trouver des normalisations et pondérations optimales. Puisque les calculs faits dans cet article pour la désambiguïsation terminologique ne sont pas spécifiques à un domaine particulier, ce serait intéressant de répéter les expériences sur d'autres corpus annotés relevant de disciplines différentes.

3.2.2. Analyse qualitative

Pour analyser les résultats obtenus qualitativement, nous les avons classés par ordre décroissant à partir du meilleur taux d'exactitude (*accuracy*) obtenu. Dans le tableau 7 ci-dessous, les taux d'exactitude supérieurs à la *baseline* sont en vert, ceux qui sont égaux sont en rouge. Pour expliquer les résultats obtenus, à la suite de (Jacquey et al., 2010) sur la question de l'ambiguïté terminologique, nous avons testé si les résultats obtenus pouvaient s'expliquer en fonction du taux d'ambiguïté sémantique et du taux d'ambiguïté terminologique de chaque candidat.

Candidat	Fréquence	Catégorie	Taux d'exactitude	Ambiguïté dans les ressources lexicales ou terminologiques				
				Ambiguïté sémantique			Ambiguïté terminologique	
				WOLF	WIK	TLFi	TS	GDT
statut	faible	[très peu term]	98,04	2,0,0	4,0,0	8,1,0	8,1	6,0
collocation	forte	[très term]	97,40	0,0,0	4,2,1	9,4,1	2,2	2,0
structure	forte	[très peu term]	95,54	3,0,0	7,2,0	15,2,1	8,0	50,1
vocable	faible	[amb]	95,45	0,0,0	2,1,0	2,2,2	0,0	0,0
modèle	forte	[peu term]	93,80	11,0,0	10,3,0	11,8,0	15,1	72,0
diglossie	faible	[très term]	93,55	0,0,0	1,1,1	0,0,0	4,2	2,1
syntagme	faible	[assez term]	90,70	1,1,1	2,2,1	4,4,3	5,2	3,1
définition	forte	[amb]	84,21	1,1,1	4,3,1	7,3,1	6,1	13,1
mot	forte	[assez term]	83,79	5,1,1	14,4,1	5,2,1	5,1	10,9
entrée	faible	[peu term]	81,97	4,0,0	16,4,1	15,7,1	1,0	57,2

Tableau 7. Recueil d'informations sémantiques et terminologiques sur les candidats désambiguïsés automatiquement et classés par ordre décroissant de taux d'exactitude

Pour recueillir ces deux informations qualitatives (ambiguïté sémantique et ambiguïté terminologique), nous avons consulté trois ressources lexicales (*WOLF*, le *wordnet* libre pour le français ; le *Wiktionnaire* (WIK) ; le dictionnaire du trésor de la langue française informatisé (TLFi)) pour établir le taux d'ambiguïté sémantique et deux ressources terminologiques (la base de données terminologiques *TermSciences*, TS dans le tableau, et le Grand Dictionnaire de Terminologie, GDT dans le tableau) pour estimer le taux d'ambiguïté terminologique. Le relevé des informations recueillies est indiqué dans le tableau 7. Pour les ressources lexicales, on trouve 3 chiffres séparés par des virgules : le premier correspond au nombre total de définition, le second au nombre de définitions techniques, le troisième au nombre de définitions relevant des Sciences du Langage. Pour les ressources terminologiques, seuls les deux derniers chiffres sont donnés attendu que tous les termes relèvent par définition d'un domaine de spécialité.

Une première conclusion est qu'aucune des informations qualitatives recueillies ne permet d'expliquer de manière globale le classement des candidats par ordre décroissant de taux d'exactitude. *Modèle* qui est plus ambigu que *statut*, sémantiquement et terminologiquement, est moins bien classé que ce dernier. Il en est de même avec *syntagme, définition* et *mot* qui sont moins bien classés que *modèle*. Le même constat peut être fait pour les informations quantitatives recueillies dans le corpus. Une fréquence importante ne garantit pas une meilleure désambiguïsation. Ainsi, *collocation* est moins bien classé que *statut* alors que *statut* a une fréquence plus faible que *collocation*. De la même manière, le classement d'un candidat parmi les « peu ambigus », c'est-à-dire appartenant aux catégories [très peu terminologique] ou [très terminologique] ne garantit pas un meilleur taux d'exactitude. On peut l'observer pour *vocable* qui, bien que n'appartenant à aucune des catégories non ambiguës, est mieux classés que le candidat *diglossie* pourtant classé comme appartenant à la catégorie [très terminologique].

En revanche, si on combine les différentes informations recueillies, on peut observer que les cinq meilleurs candidats du point de vue du taux d'exactitude sont majoritairement peu terminologiques et que les deux candidats restants, qui sont soit très terminologique (*collocation*), soit ambigu (*vocable*), sont considérés dans deux ressources comme appartenant exclusivement au domaine des Sciences du Langage : toutes les définitions du TLFi relèvent de ce domaine, de même que tous les termes correspondant présents dans *TermSciences*. Mais, un tel constat combiné ne peut pas être établi pour les cinq candidats les moins bien classés selon le taux d'exactitude de la désambiguïsation automatique. Parmi ces cinq candidats cependant il faut souligner que *définition, mot* et *entrée* apparaissent très souvent dans des expressions polylexicales comme *par définition*, *en deux mots* ou encore *en entrée*. Cette dernière observation incite clairement à envisager la détection automatique de ces expressions polylexicales avant de procéder à la désambiguïsation terminologique. Enfin, les deux candidats restants parmi les moins bien classés, *diglossie* et *syntagme*, sont aussi ceux pour lesquels les différentes exploitations statistiques testées n'ont pas apporté de gain par rapport à la *baseline*. Tous deux ont une fréquence faible, sont classés comme assez ou très terminologique lors de la désambiguïsation manuelle et sont peu ambigus que ce soit sur le plan sémantique ou terminologique. Dans le cas de *diglossie*, ce constat montre tout d'abord que les différentes méthodes d'exploitation statistique ne sont peut-être pas suffisantes pour désambiguïser mieux qu'on ne le ferait avec la *baseline*. Dans le cas de *syntagme*, ce constat montre ensuite qu'on peut légitimement s'interroger sur le classement du candidat en fonction de la désambiguïsation manuelle dans la catégorie [assez terminologique]. En effet, ce candidat est peu ambigu dans les ressources lexicales et terminologiques. On pourrait donc s'attendre à un classement dans la catégorie [très terminologique].

Une explication plausible tient aux consignes de désambiguïsation manuelle qui précisent que seules les occurrences syntaxiquement correctes peuvent être considérées comme valides terminologiquement. Or, le candidat *syntagme* apparaît de manière très fréquente dans une configuration de type Nom-Adjectif, par exemple, *syntagme nominal*, configuration que l'extracteur de terme ne reconnaît pas forcément dès lors que le composant adjectif de cette configuration varie de manière importante.

Par conséquent, parmi les cinq candidats les moins bien classés, on peut souligner qu'une piste d'explication tient à la fréquence des cas où les occurrences sont dans une configuration polylexicale : locutions adverbiales avec *définition, mot* et *entrée* ou groupe nominal étendu trop varié avec *syntagme.*

5. Conclusion et Perspectives

Dans cet article, nous avons présenté des travaux qui visent l'exploitation statistique des contextes autour d'occurrences de candidats termes que l'on aimerait pouvoir désambiguïser automatiquement du point de vue terminologique. Les occurrences de candidats termes sont détectées automatiquement par la plateforme d'extraction terminologique *TTC-TermSuite* puis sont désambiguïsées manuellement. A l'issue de l'enrichissement des données textuelles en occurrences désambiguïsées de candidats termes, le corpus annoté joue le rôle d'un corpus de référence. Pour évaluer notre hypothèse selon laquelle les contextes fournissent des indices lexicaux utiles pour la désambiguïsation, deux méthodes d'analyse statistique sont mises en œuvre. De plus, quatre variantes par méthode sont examinées. Enfin, les performances des huit méthodes sont comparées à celles d'une méthode *baseline* fondée sur les fréquences relatives des candidats termes. Les résultats obtenus montrent que les performances des

différentes méthodes sont peu sensibles au fait que les occurrences de candidats termes soient majoritairement terminologiques ou non. Autrement dit, même un petit nombre d'occurrences terminologiques dans le corpus permet d'obtenir de bons résultats avec les méthodes de désambiguïsation automatique. Cependant, pour cela, il est nécessaire d'utiliser les méthodes s'appuyant sur une pondération établie à partir des annotations manuelles.

Remerciements

Nous remercions les relecteurs de cet article dont les remarques et suggestions ont augmenté la qualité. Nous souhaitons aussi remercier l'équipe de soutien à la recherche en développement informatique de l'Atilf et en particulier Benjamin Husson, Bertrand Gaiffe, Jean-Marc Humbert et Etienne Petitjean.

Références

Aubin S. et Hamon T. (2006) Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL)*.

Bourigault D., Jacquemin C. et L'Homme M-Cl. (2001). *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam.

Daille B. (1994) Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse en informatique fondamentale, Université Paris 7

Daille B. (2003). Conceptual structuring through term variations. Bond F., Korhonen A., MacCarthy D. and Villacicencio A., editors, *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 9-16.

Drouin P. (2003). Term extraction using on-technical corpora as a point of leverage. In *Terminology.* Vol 9(1) : 99-117.

Drouin P. (2007). Identification automatique du lexique scientifique transdisciplinaire. In *Revue Française de linguistique appliquée*, vol. 12(2) : 45-64.

Jacquemin C. (1996). What is the tree that we see through the window : A linguistic approach to windowing and term variation. *Information Processing & Management*, vol. 32(4) : 445-458.

Jacquemin C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes, Nantes.

Kister L. et Jacquey E. (2012). Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral. In *Actes du Congrès Mondial de Linguistique Française*, Lyon, pp. 909 – 919.

Heiden S., Magué J-P. et Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In *Proceedings of JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome, Italie : ENS-Lyon, 12 pages http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf.

Jacquey É., Kister L., Grzesitchak M., Gaiffe B., Reutenauer C., Ollinger S. et Valette M. (2010). Thésaurus et corpus de spécialité Sciences du Langage : approches lexicométriques appliquées à l'analyse de termes en corpus. In *Proceedings of TALN2010*, Montréal.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. In *Mots* vol. 1 :127-165.

L'Homme M.C. (2004). *La terminologie : principes et techniques*, Montréal : Les Presses de l'Université de Montréal.

Toussaint Y., Namer F., Daille B. Jacquemin C., Royauté J. et Hathout N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. In *Proceedings of TALN'98*, Paris, France.