Le poids des entités nommées dans le filtrage des termes d'un domaine

Nouha Omrane, Adeline Nazarenko, Sylvie Szulman

Le poids des entités nommées dans le filtrage des termes d'un domaine

Nouha Omrane Adeline Nazarenko Sylvie Szulman
LIPN UMR 7030 (Université Paris 13 & CNRS)
99, av. J.B. Clément, 93430, Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Abstract

L'extraction automatique de termes est utilisée pour des tâches variées comme l'analyse terminologique, la détection des mots clés pour la recherche d'information et la construction d'ontologies. Les outils de traitement automatique de la langue (TAL) ont la charge d'extraire les termes d'un domaine à partir de corpus spécialisés, mais ces outils n'extraient pas que des termes pertinents. Notre objectif est d'améliorer la sélection des termes pour un domaine donné. Nous proposons des méthodes de filtrage et de pondération de termes qui tiennent compte de la distribution des termes au voisinage des entités nommées et nous montrons qu'elles aident à détecter les termes représentatifs d'un domaine.

Mots clés : mesure de pertinence, entité nommée, terminologie, potentiel terminologique.

1 Introduction

Les méthodes de construction d'ontologies à partir des textes proposent un processus de conceptualisation d'un domaine sous forme d'un modèle formel (taxonomie, ontologie) à partir de l'acquisition des termes. La difficulté de la tâche de conceptualisation, le prisme de l'application visée ainsi que l'ambiguïté et la plasticité de la langue font qu'on ne peut pas extraire automatiquement une ontologie à partir de textes. Pourtant, si l'ontologie doit être *construite* à partir de textes, l'acquisition des termes en constitue généralement la première étape parce qu'elle aide à identifier les notions à modéliser. L'acquisition de termes est assurée par des outils de TAL qui extraient une liste de termes candidats d'un texte. Selon la méthode de construction d'ontologies (automatique, semi-automatique ou manuelle guidée), un traitement (apprentissage, regroupement distributionnel, regroupement manuel) est effectué sur la liste de termes extraits par les outils de TAL pour identifier des classes sémantiques et construire les concepts qu'elles reflètent. Mais cette étape suppose que la liste soit nettoyée au préalable afin de réduire le bruit. Généralement, le filtrage des termes pertinents pour le domaine est assuré soit par un expert du domaine, soit à l'aide d'une ressource externe (WordNet, une taxonomie du domaine, etc.). Cependant l'expert n'est pas toujours disponible et la liste des termes est parfois ingérable humainement (grande taille de données). De plus il n'existe pas, du moins pas souvent, de ressource disponible qui décrive les objets du domaine.

Les outils de TAL extraient des termes appelés *termes candidats* [1] qui sont souvent des syntagmes nominaux composés de plusieurs mots susceptibles de désigner par leurs propriétés syntaxiques et grammaticales une *notion du domaine*. Les outils de TAL s'appuient sur la structure linguistique des termes candidats et leurs distributions en corpus pour détecter les termes d'un domaine mais le résultat n'est pas toujours probant. *Il n'existe pas de critère formel générique permettant de déterminer si un mot ou un syntagme a une valeur terminologique.*

Il nous semble cependant que, le domaine cible étant choisi, on peut s'appuyer sur les entités nommées qui relèvent de ce domaine et sont mentionnées dans les textes pour repérer les termes qui

1. Généralement, les termes candidats sont extraits à partir du texte en appliquant des patrons linguistiques basés sur des structures morpho-syntaxiques.

sont eux-mêmes les plus pertinents pour ce domaine. Nous faisons en effet l'hypothèse que l'ancrage référentiel des entités nommées confère une valeur sémantique particulière à leurs contextes et "déteint" sur les termes avoisinants.

Les entités nommées sont des unités textuelles qui ont suscité beaucoup d'intérêt en TAL (à travers les campagnes d'évaluation MUC et ACE, notamment) et elles ont la particularité de renvoyer à des entités du monde. Les outils de reconnaissance d'entités nommées permettent de repérer les entités nommées d'un texte et de leur attribuer un type sémantique qui dépend du domaine considéré.

Cet article montre comment le voisinage des entités nommées permet de filtrer et pondérer une liste de termes en fonction d'un domaine particulier. L'approche proposée est testée sur deux cas d'usage qui viennent confirmer notre hypothèse de départ.

L'article se compose de 4 sections. La section 2 rappelle les principaux travaux concernant le filtrage des termes et le rôle des entités nommées. La section 3 présente les deux types d'unités textuelles (termes et entités nommées) sur lesquelles repose notre approche. La méthode de filtrage et de pondération de termes proposée est décrite dans la section 4 et la section 5 présente les résultats obtenus sur deux cas d'usage différents.

2 Etat de l'art

2.1 Filtrage des termes

On a souvent cherché à déterminer le potentiel référentiel des termes pour un domaine spécifique dans le but d'extraire des termes pertinents. Souvent, les travaux s'appuient sur le calcul de la fréquence des termes candidats comme l'un des critères de pertinence. Nous distinguons trois grandes familles d'approches.

Les approches linguistiques s'appuient sur la structure linguistique des termes candidats et/ou de leurs contextes pour détecter ceux du domaine. (Frantzi & Ananiadou, 1997) s'intéressent qu'aux termes ayant comme voisins dans le texte des noms ou des verbes. (Maynard & Ananiadou, 1999) enrichit cette approche en prenant en compte les relations sémantiques (taxonomiques) entre termes qui ont été détectées à partir d'une ressource externe (taxonomie).

Les approches statistiques reposent sur le calcul de la fréquence et de la distribution des termes dans le corpus. (Wong *et al.*, 2007) propose un nouveau indicateur, appelé *TermHood*[2], qui décrit la distribution d'un terme candidat dans un domaine par contraste avec d'autres domaines. D'autres travaux se sont inspirés de la mesure *tf.tdf* pour pondérer les termes en fonction de leur distribution dans les documents constituant le corpus. (Drouin, 2003) propose une méthode d'extraction de termes pertinents pour des corpus de spécialité dont le but est de déterminer les éléments lexicaux qui forment les termes candidats spécifiques à un domaine. La méthode s'appuie sur deux mesures statistiques appliquées à deux corpus l'un de spécialité et l'autre générique : le calcul de la fréquence des éléments lexicaux et la probabilité d'observer une fréquence d'un élément égale ou supérieure à celle du corpus de spécialité. L'outil qui supporte la méthode s'appelle *TermoStat*[3]. Le résultat obtenu est une liste de termes spécifiques à un domaine donné.

Les approches mixtes combinent des critères linguistiques et statistiques. (Daille, 1994) extrait les termes pertinents à partir d'une analyse statistique du texte et d'un filtrage linguistique des termes candidats. *Acabit*[4] (Daille, 2003) produit une liste ordonnée des termes les plus représentatifs d'un domaine spécifique ainsi que les variations morphologiques sous la forme de groupes de termes candidats en s'appuyant sur les cooccurrences des termes candidats.

Bilan (Drouin & Langlais, 2006) ont évalué les approches linguistiques et statistiques en utilisant un corpus annoté à la main par un expert et les mesures de précision et rappel. Les auteurs en concluent que la fréquence est un bon indice de pertinence des termes. Ils notent que certaines mesures statistiques privilégient les termes courts et d'autres les termes composés et que leur combinaison permet d'augmenter le rappel.

2.2 Exploitation des entités nommées

Les entités nommées sont exploitées dans plusieurs domaines pour différentes tâches liées

2. Le potentiel terminologique d'un terme en français.
3. http ://olst.ling.umontreal.ca/ drouinp/termostat_web/
4. Automatic Corpus Based Acquisition of Binary Terms

au travail de conceptualisation. Elles sont utilisées dans la construction d'ontologies (Bendaoud *et al.*, 2007; Omrane *et al.*, 2011), pour leur peuplement (Giuliano & Gliozzo, 2008) mais aussi pour la

détection des relations de domaine (Toru *et al.*, 2010).

Dans cet article, nous nous intéressons aux entités nommées en tant qu'unités textuelles permettant de mettre l'accent sur des éléments du domaine. Notre propos est de montrer que les entités nommées peuvent aider à la détection de termes pertinents pour la conceptualisation d'un domaine donné.

3 Des types d'unités textuelles contrastés

Les termes et les entités nommées sont deux types d'unités textuelles qui sont généralement intéressantes à détecter mais qui n'ont pas la même histoire ni la même valeur sémantique.

3.1 Terme

(Lerat, 2009) définit la notion de terme comme : « le nom donné dans une langue à une entité conceptualisée par une communauté de travail. Cette dénomination est souvent un nom ou un groupe nominal, mais elle peut aussi appartenir à une nomenclature alphanumérique, une unité définie dans les textes de spécialité ».

Un terme est donc une unité syntaxique qui peut être composée d'un ou plusieurs mots mais qui est relative à un domaine spécifique dont elle décrit une notion. Les termes ont généralement un sens précis dans le domaine auquel ils appartiennent, ils sont moins ambigus que les mots courants.

Cette stabilité sémantique est intéressante pour la modélisation conceptuelle et la construction d'une ontologie de domaine : comme le terme reflète une notion pertinente du domaine étudié, il sert souvent à construire un concept formel.

3.2 Entité nommée

Les entités nommées sont des unités textuelles d'un autre type. Création du TAL comme le souligne (Ehrmann, 2008), elles ont une histoire moins ancienne que les termes. On les assimile généralement à des noms propres (noms de personnes, de lieux, par ex.) mais aussi à des valeurs numériques, des dates, des adresses http, etc.

Si l'on s'intéresse au repérage des entités pour la construction d'ontologies, toutes les mentions [5] sont potentiellement intéressantes.

En pratique, on met l'accent sur les noms propres qui sont plus faciles à identifier dans le flux textuel, en sachant qu'on n'identifie ainsi qu'un sous-ensemble des entités mentionnées dans le texte.

L'exploitation des entités nommées est relative à un cadre applicatif et une tâche (Ehrmann, 2008) qui permettent de déterminer quels sont les types d'entités nommées pertinents à prendre en compte. On s'intéresse ainsi davantage aux noms de gènes et de maladies dans le domaine biomédical qu'aux noms d'organisations ou de personnes.

En TAL, on s'intéresse généralement aux entités nommées du fait de leur valeur référentielle : ce sont des expressions linguistiques qui renvoient de manière autonome et non ambigue, dans un contexte donné [6], à des entités du monde (Ehrmann, 2008).

3.3 Différence et lien

Les termes et les entités nommées sont donc des types d'unités textuelles qui jouent chacun un rôle particulier par rapport au domaine mais qui ont un fonctionnement sémantique différent : les termes reflètent des notions alors que les entités nommées renvoient à des objets ou référents. Les méthodes d'extraction des termes et des entités nommées sont également différentes. Alors qu'on part généralement du texte (propriétés linguistiques et statistiques) pour extraire des termes, le repérage des entités nommées est guidé par les types sémantiques considérés comme pertinents et sélectionnés en fonction du domaine considéré.

Sous l'hypothèse que « les termes du domaine n'apparaissent pas seuls ou d'une manière arbitraire » (Maynard & Ananiadou, 1999), nous proposons d'exploiter ce contraste des termes et des entités nommées. Nous considérons que la valeur référentielle des entités nommées "déteint"

5. Les "mentions d'entités nommées" sont des unités textuelles qui renvoient à des "entités" du domaine qui peuvent relever de différentes catégories linguistiques : des noms propres ("Air France"), mais aussi des pronoms ("elle"), et plus largement des descriptions définies ("cette compagnie", "la principale compagnie aérienne française").

6. Les cas d'ambiguïté d'entités nommées ne sont pas rares mais, dans un contexte donné, il s'agit essentiellement de métonymie.

sur leur contexte, qu'elles jouent le rôle de marqueur de domaine pour les termes qui figurent à leur voisinage et peuvent ainsi aider à repérer les termes les plus pertinents pour un domaine donné.

Dans la littérature, les entités nommées, avec leur sémantique référentielle, sont considérées comme des éléments clés du domaine et les relations qu'elles entretiennent sont aussi considérées comme relations de domaine. Beaucoup de travaux se sont ainsi intéressés à l'identification des relations de domaine en explorant les relations sémantiques existant entre entités nommées (Zhu *et al.*, 2009; Hirano *et al.*, 2007) ou entre termes et entités nommées (Ohta *et al.*, 2010).

Notre approche s'appuie sur la même idée que les entités nommées sont des marqueurs de domaine mais pour la détection des termes plutôt que des relations.

4 Entités nommées et extraction de termes

Comme indiqué en section 2, les outils d'extraction de termes, qui sont généralement indépendants de tout domaine, produisent des résultats bruités et nécessitent souvent un travail manuel pour sélectionner les termes les plus pertinents pour un domaine particulier. Nous proposons de nous appuyer sur les propriétés de domaine des entités nommées pour améliorer cette sélection. Etant donné des types sémantiques pertinents pour un domaine donné, nous faisons l'hypothèse que les termes qui figurent au voisinage de ces types d'entités nommées ont plus de chance d'être pertinents pour le domaine que les autres.

Nous définissons une relation de « voisinage » entre des occurrences de terme et d'entités nommées et nous nous proposons de nous appuyer sur cette relation de voisinage pour filtrer les termes d'un domaine et les pondérer. En prenant la phrase comme contexte, nous posons qu'une occurrence de terme t figure au voisinage d'une occurrence d'entité nommé e ($vois(t, e)$) si et seulement si elles figurent dans la même phrase.

4.1 Filtrer les termes

Les listes de termes candidats des extracteurs de termes étant souvent longues à valider manuellement, il s'agit tout d'abord d'éliminer le bruit et de présélectionner – ou filtrer – les termes les plus pertinents pour le domaine considéré.

Notre méthode de filtrage est simple : nous considérons qu'un terme t est pertinent si et seulement si l'une de ses occurrences figure au voisinage d'une occurrence d'entité nommée :

$$Pert(TC) = \begin{cases} 1 & \text{si } \exists t, EN, e / occ(t, TC) \\ & \wedge occ(e, EN) \wedge vois(t, e) \\ 0 & \text{sinon} \end{cases}$$

où TC est un terme candidat, EN une entité nommée et où $occ(x, X)$ et $vois(x, y)$ indiquent respectivement que x est une occurrence de X et que x et et y cooccurrent dans la même phrase.

4.2 Attribuer des poids aux termes

Pour aller plus loin dans la sélection des termes pertinents, nous proposons de trier les termes sur la base de leurs relations de voisinage en considérant que certains voisinages sont plus marqués que d'autres. Cela revient à attribuer plus de poids aux termes qui apparaissent dans des contextes "chargés" en entités nommées. L'objectif est de proposer à l'ingénieur de la connaissance une liste de termes triée autrement que par la fréquence.

Nous définissons le poids d'un terme comme suit :

$$Poids(TC) = \frac{Freq_{vois}(TC)}{FreqTotale(TC)}$$

où $Freq_{vois}(TC)$ est le nombre total de relations de voisinage dans lesquelles entrent les occurrences de TC et $FreqTotale(TC)$ sa fréquence totale (nombre d'occurrences). Le poids d'un terme est donc le nombre moyen de relations de voisinage dans lesquelles entrent ses occurrences.

Nous obtenons une liste de termes candidats *pondérés* en fonction du nombre d'entités nommées figurant dans leur voisinage.

Prenons l'exemple suivant tiré d'un cas d'usage de *American airlines* (décrit dans la section 5) où nous cherchons à calculer le poids du terme candidat *mileage credit* :

> You may request mileage credit for past, eligible transactions up to 12 months from the transaction date. Any claim for uncredited mileage must be received by American Airlines within 12 months after the mileage credit was earned. No mileage credit will be awarded for canceled flights or if you are accommodated on another airline.

En appliquant la mesure, nous obtenons *Poids(mileage credit)* = $\frac{3}{3}$. Le terme *mileage credit* a 3 occurrences dans le texte. La première (non reproduite ici) n'a aucune entité nommée dans son voisinage. La seconde cooccurre avec 1 entité nommée (*12 months*) et la troisième avec 2 (*American Airlines* et *12 months*).

La section qui suit présente les expériences de filtrage et de pondération des termes faites pour deux cas d'usage différents.

5 Expérimentations et évaluation

Après avoir introduit les cas d'usage sur lesquels nous avons testé notre approche, nous présentons les résultats obtenus et évaluons les méthodes de filtrage et de pondération proposées.

5.1 Cas d'usage

Les expériences ont été faites sur deux cas d'usage du projet ONTORULE [7] qui vise à construire des systèmes d'aide à la décision pour des domaines réglementaires et leurs bases de règles métiers.

Dans ce contexte applicatif et pour chacun de ces cas d'usage, nous cherchons à construire une ontologie de domaine à partir de textes qui décrivent les réglementations en vigueur. Notre but est de repérer les termes qui décrivent le mieux le domaine sous-jacent parce qu'ils en reflètent les concepts centraux. Une ontologie ainsi construite sert ensuite de vocabulaire conceptuel pour l'écriture des règles métiers et la formalisation des réglementations.

Nous avons testé nos méthodes de filtrage et de pondération de termes sur deux corpus différents. Le corpus AAdvantage (AA) décrit les règles et conditions d'attribution de « miles » pour des voyageurs [8] et contient 5 744 mots. Le corpus Audi, qui contient 3 704 mots, est un extrait d'une directive européenne décrivant les règles et procédures que les véhicules doivent satisfaire en matière de ceintures de sécurité.

Pour chacun des cas d'usage traités, nous prenons en entrée deux listes d'unités lexicales fournies par des outils de TAL (une liste LTC de termes candidats extraits par l'outil YaTeA [9] et une liste LEN d'entités nommées extraites par la chaîne de traitement Annie de la plateforme GATE [10]), nous calculons une liste de termes filtrés (LTF) ou une liste des termes pondérés (LTP) et nous comparons le résultat obtenu avec une liste de termes de référence [11].

5.2 Filtrage : résultats et évaluation

En appliquant notre méthode de filtrage sur les listes LTC et LEN, nous obtenons d'abord une liste de termes filtrés (LTF) qui contient les termes de la liste initiale figurant au moins une fois au voisinage d'une entité nommée.

Le tableau 1 présente les résultats. On peut noter que l'écart entre le nombre de termes candidats filtrés est moindre dans le cas d'AAdvantage (AA) que pour le corpus Audi. Cela tient au fait qu'un nettoyage préalable a été fait sur la liste des termes candidats de AAdvantage pour éliminer les termes mal formés [12] et mieux apprécier l'apport spécifique de notre filtrage. On observe que le filtrage permet de réduire la liste des candidats termes, même quand elle a été nettoyée (AA) et qu'il compense l'absence de nettoyage préalable (Audi).

Corpus	LTC	LEN	LTF
AAdvantage	680	105	437
Audi	1003	90	436

Tableau 1: Nombres de termes candidats, d'entités nommées et de termes filtrés pour les deux cas d'usage

Pour évaluer ces résultats, nous utilisons les mesures classiques utilisées en recherche d'information, la précision, le rappel et la F-mesure :

$$\textit{Précision} = \frac{\text{UTP}}{\text{UT}} \quad \textit{Rappel} = \frac{\text{UTP}}{\text{UP}}$$

$$\textit{F-Mesure} = \frac{2*\text{Précision}*\text{Rappel}}{\text{Précision}+\text{Rappel}}$$

où UP, UT ou UTP sont respectivement le nombre d'unités pertinentes (*i.e.* figurant dans la référence), le nombre d'unités trouvées (*i.e.* figurant dans le résultat et donc filtrées) et le nombre d'unités à la fois trouvées et pertinentes.

7. Les documents utilisés sont extraits des cas d'usage étudiés dans le cadre du projet ONTORULE (FP7 231875).

8. Nous remercions American Airlines d'avoir mis ce corpus à notre disposition.

9. http ://search.cpan.org/ thhamon/Lingua-YaTeA/

10. http ://gate.ac.uk/

11. Ces listes de référence ont été construites manuellement, au préalable et sans prendre en compte les entités nommées pour modéliser les domaines des cas d'usage.

12. Quelle que soit la qualité des extracteurs de termes, les outils généralistes ne peuvent pas prendre en compte les particuliarités des textes (présence de chiffres, de tableaux, etc.) sur lesquels ils sont appliqués et il reste toujours des scories.

Les résultats figurent dans le tableau 2. Les

Cas	Mesures	LTF	LTC
AA	Précision	71,6%	51%
	Rappel	89,1%	100%
	F-mesure	79,40%	67,55%
Audi	Précision	56,8%	31,4%
	Rappel	78,7%	100%
	F-mesure	65,98%	47,79%

Tableau 2: Impact du filtrage sur les mesures de précision, rappel et F-mesure

résultats sont dans l'ensemble moins bons dans le cas d'Audi mais le filtrage des termes améliore significativement la précision dans les deux cas d'usage. Le fait que le rappel soit moindre sur les listes de termes filtrés montre que le filtrage n'est pas parfait (certains termes pertinents ne figurent pas au voisinage d'entités nommées [13]), mais le gain de plus de 10 points de F-mesure sur la liste des termes filtrés montre que le filtrage est globalement positif. Il l'est même d'autant plus que la liste des termes candidats a été peu nettoyée au départ, comme dans le cas d'Audi.

Le voisinage des entités nommées apparaît donc comme un critère pertinent à prendre en compte pour la détection des termes d'un domaine. Il permet de réduire la part de bruit dans les résultats des outils de TAL et de débroussailler le travail de validation. Il faut seulement prévoir de pouvoir "récupérer" par d'autres méthodes des termes pertinents qui auraient été éliminés lors du filtrage.

5.3 Pondération : résultats et évaluation

L'exploration de longues listes de termes se fait souvent par ordre alphabétique (pour regrouper les termes apparentés) ou par fréquence (pour éviter de passer trop de temps sur les hapax [14] toujours très nombreux). Dans cette section, nous cherchons à évaluer l'intérêt de la pondération des termes comme critère de tri alternatif pour l'analyse d'une longue liste de termes.

Nous évaluons les listes de termes pondérés par rapport aux mêmes références que précédemment. Il s'agit d'apprécier notre capacité à placer des termes pertinents en haut de classement. Nous regardons donc comment la précision évolue quand on considère un nombre croissant de termes bien classés. Le tableau 3 montre les résultats obtenus pour les 100 termes les mieux classés, puis en fixant le seuil aux rangs 200 et 400 et enfin en considérant toute la liste des termes filtrés.

Cas	r=100	r=200	r=400	LTF
AA	78%	75%	71,5%	71,1%
Audi	45%	52,5%	60%	56,8%

Tableau 3: Evolution de la précision en fonction du nombre de termes retenus, le seuil étant fixé en fonction du rang ($r = 100, r = 200, r = 400$)

Les résultats sont contrastés. Dans le cas d'AAdvantage, la précision diminue régulièrement quand le rang augmente mais, dans le cas d'Audi, elle augmente jusqu'au rang 400 avant de s'infléchir au-delà. Le tri proposé est donc globalement pertinent dans le premier cas et moins dans le second. Une analyse détaillée montre que les entités nommées numériques du corpus Audi contribuent à surpondérer des termes dénotant des propriétés de concepts (*length* dans *...maximum length of 840 mm*, par ex.) par rapport aux termes dénotant les concepts eux-mêmes (*strap*) qui sont filtrés eux-aussi mais avec un poids inférieur (*the length of strap... shall be... as close as possible to 450 mm...*).

Comme critère de tri, le voisinage des entités nommées est donc à utiliser avec précaution mais il donne un éclairage complémentaire de la fréquence sur une liste de termes. Nous illustrons ce point sur quelques termes de la référence du cas d'usage Audi (voir tableau 4) qui n'ont pas le même rang dans les deux approches. On a ainsi deux critères de tri qui sont bruités mais pertinents et qui mettent en avant des termes différents.

6 Conclusion

Cet article montre, sur deux cas d'usage, comment les entités nommées peuvent jouer le rôle de marqueurs de domaine, contribuer à l'identification des termes pertinents et ainsi faciliter la validation manuelle de longues listes de termes. Les entités nommées sont des unités textuelles avec une valeur sémantique particulière. Comme elles font référence à des entités du domaine, elles

13. Les valeurs de 100% de rappel s'expliquent du fait que les références ont été construites à partir des termes candidats extraits sans ajout d'autres termes.

14. Termes n'apparaissant qu'une seule fois dans un corpus.

Terme	Rang Poids	Rang Freq.
size class	1	36
Frontal impact test	3	36
diameter	4	33
buckle	53	18
seat	52	11
belt	45	1

Tableau 4: Exemples de termes pertinents et leurs rangs.

sont importantes pour le domaine en question et les termes qui les entourent et qui s'y rapportent tendent à l'être aussi.

Nous avons montré que le critère de voisinage des entités nommées permet de filtrer efficacement la liste des termes fournie par un extracteur de termes générique et d'éliminer une bonne partie de termes faiblement pertinents.

Nous avons également proposé d'utiliser le voisinage des entités nommées pour pondérer les termes. Cela permet de trier la liste des termes candidats sur une base assez différente de la fréquence. Même si l'apport ne semble pas être le même pour tous les corpus, nos expériences montrent que cela donne un éclairage intéressant et complémentaire à la fréquence, qui est souvent utilisée faute de mieux.

Références

BENDAOUD R., HACENE M. R., TOUSSAINT Y., DELECROIX B. & NAPOLI A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'acf. In *Actes IC'2007*.

DAILLE B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. In *Thèse de Doctorat en Informatique Fondamentale*.

DAILLE B. (2003). Terminology mining. In *Information Extraction in the Web Era, In M. Pazienza, Ed, Springer*.

DROUIN P. (2003). Term extraction using non-technical corpora as a point of leverage. In *Terminology, vol. 9, no 1*.

DROUIN P. & LANGLAIS P. (2006). Évaluation du potentiel terminologique de candidats termes. In *Actes des 8e Journées internationales d'analyse statistique des données textuelles (JADT-2006)*.

EHRMANN M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris 7.

FRANTZI K. & ANANIADOU S. (1997). Automatic term recognition using contextual cues. In *Proc. of 3rd DELOS Workshop*.

GIULIANO C. & GLIOZZO A. (2008). Instance-based ontology population exploiting named-entity substitution. In *Proc. of the 22nd Int. on Computational Linguistics (Coling 2008)*.

HIRANO T., MATSUO Y. & KIKUI G. (2007). Detecting semantic relations between named entities in text using contextual features. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

LERAT P. (2009). La combinatoire des termes. exemple : nectar de fruits. In *Hermes. Journal of Language and Communication Studies*.

MAYNARD D. & ANANIADOU S. (1999). Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology*.

OHTA T., PYYSALO S., KIM J.-D. D. & TSUJII J. (2010). A re-evaluation of biomedical named entity-term relations. *Journal of bioinformatics and computational biology*, **8**(5), 917–928.

OMRANE N., NAZARENKO A. & SZULMAN S. (2011). Les entités nommées : éléments pour la conceptualisation. In *Actes IC'2011*.

TORU H., HISAKO A., YOSHIHIRO M. & GENICHIRO K. (2010). Recognizing relation expression between named entities based on inherent and context-dependent features of relational words. In *Proc. of the 23rd Int. Conf. on Computational Linguistics : Posters (COLING '10)*.

WONG W., LIU W. & BENNAMOUN M. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proc. of the sixth Australasian Conf. on Data mining and analytics. Volume 70*.

ZHU J., NIE Z., LIU X., ZHANG B. & WEN J.-R. (2009). Statsnowball : a statistical approach to extracting entity relationships. In *Proc. of the 18th Int. Conf. on World wide web (WWW '09)*.