Impact de l'agglutination dans l'extraction de termes en arabe standard moderne

Wafa Neifar[1,2] Thierry Hamon[1,3] Pierre Zweigenbaum[1]
Mariem Ellouze Khemakhem[2] Lamia Hadrich Belguith[2]
(1) LIMSI, CNRS, Paris-Saclay university, F-91405 Orsay, France
(2) MIRACL Laboratory, Sfax university, Tunisia
(3) Paris 13 university – Sorbonne Paris Cité, F-93430, Villetaneuse, France
{neifar,hamon,pz}@limsi.fr, mariem.ellouze@planet.tn, l.Belguith@fsegs.rnu.tn

RÉSUMÉ

Nous présentons, dans cet article, une adaptation d'un processus d'extraction de termes pour l'arabe standard moderne. L'adaptation a d'abord consisté à décrire le processus d'extraction des termes de manière similaire à celui défini pour l'anglais et le français en prenant en compte certaines particularités morpho-syntaxiques de la langue arabe. Puis, nous avons considéré le phénomène de l'agglutination de la langue arabe. L'évaluation a été réalisée sur un corpus de textes médicaux. Les résultats montrent que parmi 400 termes candidats maximaux analysés, 288 sont jugés corrects par rapport au domaine (72,1%). Les erreurs d'extraction sont dues à l'étiquetage morpho-syntaxique et à la non-voyellation des textes mais aussi à complexité de la prise en compte de l'agglutination.

ABSTRACT

Adaptation of a term extractor to the Modern Standard Arabic language

In this paper, we present the adaptation to the Modern Standard Arabic of a term extraction process. The adaptation firstly focuses on the description of extraction process similarly to those already defined for French and English while considering the morpho-syntactic specificity of the Arabic. Then, the agglutination phenomena has been taken into account in the term extraction process. The evaluation has been performed on a medical text corpus. Results show that among 400 maximal candidate terms we analysed, 288 are correct (72.1%). The errors of term extraction are due to the Part-of-Speech tagging and the non voweled texts, but also to the complexity of taking into acount the agglutinative phenomena.

MOTS-CLÉS : Terminologie, Extraction de termes, Langue Arabe, Agglutination, Textes médicaux.

KEYWORDS : Terminology, Term Extraction, Modern Standard Arabic, Agglutination, Medical texts.

1 Introduction

Depuis plusieurs années, de nombreux travaux de recherche ont conduit à la mise au point de méthodes d'acquisition terminologique à partir de textes de spécialité (articles scientifiques, documentations techniques, textes juridiques, etc.) (Cabré *et al.*, 2001 ; Pazienza *et al.*, 2005 ; Marshman *et al.*,

2012 ; Q. Zadeh & Handschuh, 2014). Ces résultats permettent également de prendre en compte la terminologie d'un domaine dans les applications facilitant ainsi l'accès à l'information spécialisée contenue dans les textes (Marshman *et al.*, 2012 ; Cohen & Demner-Fushman, 2013). Toutes les langues ne sont cependant pas outillées de la même manière. Ainsi, bien que quelques approches aient été proposées pour extraire des termes à partir de textes arabes, celles-ci ne sont pas reproductibles, car très partiellement décrites, et les résultats difficilement appréciables (section 2). Les caractéristiques intrinsèques de la langue arabe peuvent également être un frein à la mise en œuvre de méthodes d'analyse terminologique. L'objectif de notre travail est de proposer une première adaptation de l'extracteur de termes YATEA à la langue arabe, qui tienne compte de certaines phénomènes linguistiques comme les proclitiques et les marques morphologiques du cas (section 3). Nous présentons ensuite notre corpus et nous discutons les résultats obtenus à la section 4 avant de conclure.

2 Etat de l'art

L'intérêt croissant pour le traitement automatique de la langue arabe a permis de proposer des méthodes d'extraction de termes sur cette langue. Celles-ci utilisent des approches similaires à ceux réalisés sur l'anglais ou le français (Bourigault, 1993 ; Daille, 2003 ; Drouin, 2002 ; Cabré *et al.*, 2001 ; Pazienza *et al.*, 2005 ; Aubin & Hamon, 2006). Ce choix est justifié par la complexité de cette langue : les approches traditionnelles d'acquisition terminologique ne prennent pas en compte plusieurs phénomènes linguistiques comme l'absence de voyellation, l'agglutination et les ambiguïtés morphologiques et syntaxiques des phrases nominales (Boulaknadel *et al.*, 2008).

A l'instar des méthodes mises en œuvre pour extraire des termes sur le français ou l'anglais, l'extraction de termes en arabe combinent une description linguistique du processus d'extraction et des filtres statistiques pour ordonner les termes extraits. Ainsi, Bounhas & Slimani (2009) proposent d'extraire des termes complexes candidats à l'aide d'une approche hybride composée de deux étapes. Un premier filtre linguistique exploite l'analyse morphologique et l'étiquetage morpho-syntaxique des textes pour identifier des syntagmes candidates. Un second filtre statistique utilisant la mesure d'association LLR (*Log-Likelihood Ratio*) est appliqué sur les résultats ambigus de la première étape pour sélectionner la meilleure solution. AlKhatib & Badarneh (2010) proposent une approche hybride similaire à la précédente mais utilisent deux mesures statistiques : i) le LLR pour identifier le degré de stabilité de la combinaison syntagmatique candidate (*unithood*) ; ii) la C-Value (Maynard & Ananiadou, 2000) pour calculer le degré de liaison du terme à un concept du domaine (*termhood*). Ces approches ont été évaluées sur un corpus de textes en arabe du domaine de l'environnement, collectés sur des sites Web. De même, Abed *et al.* (2013) ont adapté des méthodes destinées à l'analyse de textes de langue générale pour analyser des textes d'un domaine spécifique (des textes religieux) et ainsi en extraire automatiquement les termes simples et complexes. Un corpus contenant l'Arabe Classique (CA) et l'Arabe Standard Moderne (MSA), collecté à partir des archives de journaux islamiques et des sites islamiques, est utilisé pour l'évaluation de l'approche. Dans ce travail, le TF*IDF est utilisé pour trier les termes simples en fonction de leur *termhood*, et plusieurs mesures statistiques sont utilisées pour calculer le degré d'association de leur composants (*unithood*).

Nous pouvons observer qu'à l'exception de (Bounhas *et al.*, 2011), ces méthodes utilisent des approches qui ne tiennent finalement pas compte des particularités linguistiques de l'arabe comme les ambiguïtés morphologiques et syntaxiques, la non-voyellation ou l'agglutination. De plus, comme le souligne (Bounhas *et al.*, 2014), l'évaluation des approches proposées est assez critiquable : seuls quelques centaines de termes classés parmi les premiers sont évalués manuellement, alors que plu-

sieurs milliers ont pu être extraits et que, quelle que soit l'approche, les résultats sont généralement de bonne qualité lorsqu'on ne tient compte que des premiers termes (Korkontzelos *et al.*, 2008 ; Hamon *et al.*, 2014). Notons également que les systèmes présentés ci-dessus ne sont pas librement accessibles : il n'est donc pas possible de reproduire ou de comparer les résultats.

Dans notre travail, nous nous attaquons à ces difficultés de la langue arabe en adaptant le processus d'extraction existant mis en œuvre dans l'extracteur de termes YaTeA[1] (Aubin & Hamon, 2006), et en réalisant une analyse de l'ensemble des termes identifiés dans des textes médicaux en arabe.

3 Extraction terminologique pour le MSA

3.1 Description de l'extraction de termes pour le français ou l'anglais

L'extracteur terminologique YaTeA ayant d'abord été développé pour analyser des textes en français et en anglais, l'adaptation de l'approche à une langue sémitique est un défi. Le processus d'extraction des termes réalise une analyse superficielle de textes étiquetés morpho-syntaxiquement et lemmatisés (Aubin & Hamon, 2006). La première étape consiste à segmenter le texte à l'aide de frontières syntaxiques positives et négatives (prépositions, verbes conjugués, etc.). Les groupes nominaux maximaux obtenus peuvent correspondre ou contenir des termes candidats. Au cours de la deuxième étape, des patrons d'analyse syntaxique prenant en compte la variation morpho-syntaxique des termes sont appliqués récursivement. Cette étape permet de produire des termes candidats complexes, mais il est également possible de disposer de termes candidats simples. Finalement, des mesures statistiques comme la fréquence, la C-Value (Maynard & Ananiadou, 2000), sont ensuite associées à ces termes candidats (Hamon *et al.*, 2014) lors de la troisième étape. L'adaptation que nous avons mise en œuvre porte actuellement sur les étapes 1 et 2 du processus d'extraction de termes.

3.2 Méthode d'adaptation au MSA

Notre adaptation de YaTeA aux textes de spécialité en MSA tient compte des pratiques traditionnelles en constitution de terminologie, mais aussi de certaines particularités de la langue arabe :

- la non-voyellation qui caractérise la plupart des textes en MSA : l'absence de voyellation provoque des erreurs d'étiquetage morpho-syntaxique dues aux ambiguïtés des formes non-voyellées, et par conséquent, une dégradation des résultats de l'extraction des termes.

- l'agglutination, c'est-à-dire les proclitiques et les enclitiques : cette caractéristique élémentaire de la langue arabe consiste à associer des éléments particuliers du lexique appelés clitiques (prépositions, pronoms, articles, conjonctions...), au mot auquel ils se rapportent. Du point de vue automatique, il est parfois difficile de distinguer un proclitique ou un enclitique d'un caractère du mot en question. Par exemple, le mot وسائل peut être analysé de deux manières : il peut représenter un seul mot وسائل (*méthodes*), mais aussi le proclitique و (la conjonction de coordination *et*) suivi du nom سائل (*liquide*). Ainsi, dans notre cas, l'absence de traitement de l'agglutination conduit à considérer abusivement certains mots comme des termes.

[1] Librement disponible à l'adresse suivante `http://search.cpan.org/~thhamon/Lingua-YaTeA/`

- les marques morphologiques de l'état : Nous nous intéressons notamment à l'état construit des noms, appelé aussi *al-'iḍāfah* (possession). Celui-ci permettant d'identifier le cas génitif, nous l'exploitons pour faciliter l'analyse syntaxique en tête/modifieur des groupes nominaux.

De manière similaire au français ou à l'anglais, notre processus d'extraction de termes en MSA s'appuie sur une analyse morphologique et un étiquetage morpho-syntaxique des textes. Dans notre cas, celui-ci est réalisés grâce à l'analyseur MADA+TOKAN (Habash *et al.*, 2010). Contrairement à d'autres systèmes, cet analyseur morphologique permet d'associer un lemme sous sa forme voyellée à chaque mot d'un corpus. Il est ainsi possible d'analyser la plupart des textes en MSA en évitant une détérioration importante des résultats due aux ambiguïtés des formes non-voyellées.

MADA+TOKAN fournit aussi une décomposition des mots du corpus qui permet d'identifier les clitiques, et utilise les marques morphologiques pour identifier les cas et les états. Ces informations sont exploitées dans les différentes étapes de l'extraction de termes. Cependant, la prise en compte des enclitiques étant complexe, nous nous sommes tout d'abord concentrés sur les proclitiques.

L'adaptation de l'étape 1 nous amène à définir les frontières syntaxiques. Comme pour d'autres langues, les pronoms, les ponctuations et les verbes conjugués sont définis comme des éléments ne pouvant pas apparaître dans les termes. Nous considérons également certains éléments spécifiques de la langue arabe, tels que les pseudo-verbes (كان, تكون, إنَّ – *il était*, *elle est*, *certes* pour indiquer une affirmation), les adverbes (ربما, هنا, فقط – *peut-être*, *ici*, *seulement*), ou des expressions lexicales (في بعض الأوقات – *parfois*). Nous utilisons l'analyse morphologique des mots et notamment la présence de proclitiques pour définir des frontières syntaxiques supplémentaires. Enfin, nous avons déterminé les étiquettes morpho-syntaxiques qui ne doivent pas figurer au début ou à la fin d'un terme. Il s'agit surtout des prépositions (من – *de*, إلى – *à*, بَيْنَ – *entre*, عند – *lorsque*) que l'analyseur MADA+TOKAN considère par erreur comme des noms. Ainsi, dans la phrase تكون كمية الأكسجين التي يحملها الدم أقل (*la quantité d'oxygène que transporte le sang est moindre*) les frontières syntaxiques تكون (pseudo-verbe), التي (*que/laquelle*), يحملها (*la transporte*) et أقل (*moins*) permettent d'identifier les syntagmes كمية الأكسجين (*quantité d'oxygène*) et الدم (*le sang*).

Lors de l'étape 2, nous avons défini les patrons spécifiques pour identifier la position syntaxique tête ou modifieur des composants des syntagmes nominaux mais aussi de filtrer les séquences de mots inutiles car ne pouvant pas être analysées à l'aide des patrons définis. Ces patrons prennent en compte les caractéristiques morphologiques telles le genre et le nombre, mais aussi le cas des constituants. En particulier, nous utilisons *al-'iḍāfah* qui marque l'état construit et le génitif. Par exemple, le patron `noun-m-s-g-d(Modifieur) noun-f-s-n-c(Tête)`[2] permet d'analyser le syntagme maximal كمية الأكسجين (quantité d'oxygène). De même les proclitiques sont pris en compte au sein des patrons comme, par exemple, dans le patron `noun-m-s-a-c(Terme1)` و `noun-m-s-n-c(Terme2)`.

4 Évaluation de l'extraction de termes adapté au MSA

4.1 Corpus de travail

Contrairement à la plupart des travaux d'extraction de termes en arabe, nous ne souhaitons pas travailler sur des données textuelles issues de sites Web ou de forums de discussion car la qualité ter-

[2] `noun-m-s-g-d` : nom masculin singulier défini au génitif. `noun-f-s-n-c` : nom féminin singulier construit au nominatif.

minologique est difficilement vérifiable et ne nous permettrait pas d'évaluer correctement notre approche. Aussi, nous avons constitué notre corpus à partir de textes produits par la *National Library of Medicine* (NLM) et disponibles en ligne sur MedlinePlus[3]. Actuellement, nous disposons de 168 textes parallèles arabe/français/anglais au format PDF. Contrairement aux documents en anglais et en français, la conversion au format texte des documents en arabe, en vue de réaliser des traitements automatiques, pose des nombreux problèmes (Habash, 2010) qui rendent la tâche de nettoyage très coûteuse en temps : erreur de forme des caractères, utilisation d'un caractère persan ressemblant graphiquement à un caractère arabe, etc. Aussi, pour les expériences présentées dans cet article, nous avons dû nous limiter à un corpus de 30 textes médicaux en arabe (15 532 mots). Il s'agit pour nous de réaliser une première évaluation de l'adaptation de l'extracteur de termes $_{A}^{Y}T_{E}^{A}$ au MSA.

4.2 Expériences et résultats

Nous avons réalisé deux expériences afin d'évaluer l'apport de la prise en compte de l'agglutination dans le processus d'extraction de termes. Les résultats des différentes expériences sont présentés dans le tableau 1. Nous présentons ensuite une analyse des erreurs. La validation des termes est effectuée manuellement en fonction de leurs pertinence par rapport au domaine.

	Étape 1		Étape 2			
	SNM	TS	TS	TCmax	TC	Total
Pas de prise en compte de l'agglutination	1972	262	262	590	1133	1395
Prise en compte de l'agglutination	1916	298	298	400	824	1122

TABLE 1 – Résultats de l'application de l'extraction de termes sur les textes médicaux en arabe (SNM : syntagmes nominaux maximaux, TS : termes simples candidats, TCmax : termes complexes candidats correspondant aux syntagmes nominaux maximaux, TC : termes complexes candidats).

Tout d'abord, l'extracteur de termes adapté pour l'arabe a été utilisé sans prendre en compte le phénomène d'agglutination. Les proclitiques sont considérés comme des parties du mot en question. Grâce aux frontières syntaxiques définies pour l'étape 1, nous avons pu identifier 1972 syntagmes nominaux maximaux (SNM) mais aussi 262 noms (TS) qui seront considérés comme des termes simples candidats. L'étape d'analyse syntaxique des syntagmes nominaux maximaux(étape 2) permet de retenir 590 syntagmes nominaux maximaux (TCmax). Les constituants des termes complexes étant également considérés comme des termes candidats, nous disposons d'un ensemble de 1395 termes candidats, dont 1133 termes candidats complexes (TC). Par exemple, nous considérons comme termes candidats ارتفاع معدلات الكولسترول (*élévation du taux de cholestérol*), mais aussi معدلات الكولسترول (*taux de cholestérol*), ارتفاع (*élévation*), معدلات (*taux*) et الكولسترول (*cholestérol*).

Nous avons effectué une analyse manuelle des 262 termes simples (TS) et de 590 termes candidats correspondant aux syntagmes nominaux maximaux (TCmax). La qualité de l'analyse syntaxique et la pertinence des termes candidats extraits ont été évaluées par rapport au domaine médical. Ainsi, parmi les 590 syntagmes nominaux maximaux, 388 termes candidats (65,7%) sont jugés correctement analysés et pertinents pour le domaine médical. Il ressort de cette analyse que l'agglutination est la

[3] http://www.nlm.nih.gov/medlineplus/languages/all_healthtopics.html

principale source d'erreurs aussi bien lors de l'utilisation des frontières syntaxiques que au cours de l'analyse syntaxique des syntagmes.

La deuxième expérience a pour objectif d'évaluer l'utilisation des traitements spécifiques visant séparer les proclitiques, des mots auxquels ils sont associés. La première étape permet d'identifier 298 termes candidats simples (TS) et 1916 syntagmes nominaux maximaux (SNM). Parmi ces derniers, 400 sont conservés à la fin de l'étape 2 et permettent de produire 824 termes complexes candidats. Nous avons également analysé les 400 syntagmes nominaux maximaux retenus : 288 (72,1%) sont jugés corrects. La figure 1 présente quelques termes extraits des textes médicaux.

La prise en compte de l'agglutination se caractérise par un nombre moins élevé de termes complexes candidats produits et une augmentation du nombre de termes simples. Ces observations peuvent s'expliquer par le fait que des éléments initialement agglutinés sont alors considérés comme des frontières syntaxiques. Nous observons également que le nombre de syntagmes nominaux maximaux non retenus augmente fortement. Nous expliquons cela par le regroupement, plus ou moins fortuit, des mots privés des proclitiques associés dans des syntagmes maximaux plus grands et, par conséquent, plus difficilement analysables. Une analyse approfondie des syntagmes maximaux non analysés permettra de confirmer cette hypothèse.

سرطان الثدي	(cancer du sein)	سرعة ضربات القلب	(Rythme cardiaque rapide)
الاوعية الدموية	(les vaisseaux sanguins)	الذراع الايمن	(le bras droit)
الزائدة الدودية	(l'appendice)	اخذ عينات انسجة الثدي	(biopsies mammaires)
مرض السكر	(le diabète)	الرئتين	(les deux poumons)
التهاب شعب هوائية	(bronchite)	آلام	(douleurs)
العلاج	(le traitement)	العلاج الاشعاعي	(la radiothérapie)

FIGURE 1 – Exemple de termes extraits des textes médicaux en arabe.

5 Conclusion et perspectives

Nous nous sommes intéressés à l'adaptation d'extracteur de termes de l'état de l'art, YATEA , afin de pouvoir traiter des textes de spécialité en arabe standard moderne. Il s'agissait de définir le processus d'extraction de termes candidats en s'appuyant sur une description linguistique de l'arabe, mais aussi en prenant en compte les particularités morphologiques de cette langue. Ainsi, nous nous sommes intéressés au phénomène d'agglutination et en particulier aux proclitiques. Pour cela, nous avons exploité l'analyse morphologique réalisée par MADA+TOKAN, et nous avons défini des patrons d'analyse syntaxique des termes candidats qui prennent en compte ce phénomène. Les expériences réalisées sur un corpus de 15532 mots montrent une amélioration de la qualité des résultats lorsque les proclitiques sont pris en compte. La précision des termes complexes maximaux augmente de 65,7 à 72,1% et le nombre de termes candidats extraits diminue.

Plusieurs perspectives de travail s'offrent à nous. D'une part, le traitement de la voyellation et de l'agglutination, notamment les enclitiques, doit être amélioré. En effet, l'étiquetage morpho-syntaxique et les informations par l'analyse morphologique de MADA+TOKAN sont parfois erronés, et nécessitent de mettre en place des traitements spécifiques. Les marques morphologiques de cas ou le *masdar*, c'est-à-dire le nom dérivé d'un verbe qui désigne l'action associée (ضمّادة (pansement) /ضمّد (panser)), pourraient être utilisés pour corriger l'étiquetage morpho-syntaxique dans certains cas bien

précis, ou améliorer l'analyse syntaxique des termes candidats. Enfin, nous envisageons d'évaluer notre travail dans d'autres domaines de spécialité en fonction de la disponibilité de corpus.

Références

Abed A. M., Tiun S. & Albared M. (2013). Arabic term extraction using combined approach on islamic document. *Journal of Theoretical & Applied Information Technology*, **58**(3).

AlKhatib K. & Badarneh A. (2010). Automatic extraction of arabic multi-word terms. In *IMCSIT*, p. 411–418.

Aubin S. & Hamon T. (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo & T. Pahikkala, Eds., *5th International Conference on NLP (FinTAL 2006)*, number 4139 in LNAI, p. 380–387 : Springer.

Boulaknadel S., Daille B. & Aboutajdine D. (2008). A multi-word term extraction program for arabic language. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias, Eds., *Proceedings of the LREC'08*.

Bounhas I., Elayeb B., Evrard F. & Slimani Y. (2011). Organizing contextual knowledge for arabic text disambiguation and terminology extraction. *Knowledge Organization Journal*, **38**(6), 473–490.

Bounhas I., Lahbib W. & Elayeb B. (2014). Arabic domain terminology extraction : A literature review - (short paper). In *OTM 2014 Conferences - Confederated International Conferences : CoopIS, and ODBASE 2014*, p. 792–799, Amantea, Italy.

Bounhas I. & Slimani Y. (2009). A hybrid approach for arabic multi-word term extraction. In *Natural Language Processing and Knowledge Engineering, NLP-KE 2009. International Conference on*, p. 1–8 : IEEE IEEE.

Bourigault D. (1993). An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings of the EACL'93*, p. 81–86, Utrecht, The Netherlands.

Cabré M. T., Estopà R. & Vivaldi J. (2001). Automatic term detection : a review of current systems. In *Recent Advances in Computational Terminology*. John Benjamins.

Cohen K. B. & Demner-Fushman D. (2013). *Biomedical Natural Language Processing*. John Benjamins publishing company.

Daille B. (2003). Conceptual structuring through term variations. In F. Bond, A. Kohonen, D. M. Carthy & A. Villaciencio, Eds., *Proceedings of the ACL'2003 Workshop on Multiword Expressions : Analysis, Acquisition, and Treatment*, p. 9–16.

Drouin P. (2002). *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. PhD thesis, Université de Montréal.

Habash N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Habash N., Rambow O. & Roth R. (2010). *MADA+TOKAN Manual*. CCLS-10-01.

Hamon T., Engström C. & Silvestrov S. (2014). Term ranking adaptation to the domain : genetic algorithm based optimisation of the C-Value. In Springer, Ed., *Proceedings of PolTAL 2014 – Advances in Natural Language Processing*, volume 8686 of *LNAI*, p. 71–83.

Korkontzelos I., Klapaftis I. P. & Manandhar S. (2008). Reviewing and evaluating automatic term recognition techniques. In B. Nordström & A. Ranta, Eds., *6th International Conference on NLP (GoTAL 2008)*, number 5221 in LNAI, p. 248–259 : Springer.

Marshman E., Gariépy J. L. & Harms C. (2012). Helping language professionals relate to terms : Terminological relations and termbases. *Journal of Specialised Translation*, **18**.

Maynard D. & Ananiadou S. (2000). Identifying terms by their family and friends. In *Proceedings of COLING 2000*, p. 530–536, Saarbrucken, Germany.

Pazienza M. T., Pennacchiotti M. & Zanzotto F. (2005). Terminology extraction : An analysis of linguistic and statistical approaches. In S. Sirmakessis, Ed., *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, p. 255–279. Springer Berlin Heidelberg.

Q. Zadeh B. & Handschuh S. (2014). The acl rd-tec : A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, p. 52–63, Dublin, Ireland.