Distinguer les termes des collocations : étude sur corpus du patron <Adjectif – Nom> en anglais médical

François MANIEZ.

Centre de Recherche en Terminologie et en Traduction
Université Lumière Lyon 2
fmaniez@wanadoo.fr

Résumé – Abstract

Un bon nombre des applications de traitement automatique des langues qui ont pour domaine les langues de spécialité sont des outils d'extraction terminologique. Elles se concentrent donc naturellement sur l'identification des groupes nominaux et des groupes prépositionnels ou prémodificateurs qui leur sont associés. En nous fondant sur un corpus composé d'articles de recherche médicale de langue anglaise, nous proposons un modèle d'extraction phraséologique semi-automatisée. Afin de distinguer, dans le cas des expressions de patron syntaxique <Adjectif – Nom>, les termes de la langue médicale des simples collocations, nous nous sommes livré au repérage des adjectifs entrant en cooccurrence avec les adverbes. Cette méthode, qui permet l'élimination de la plupart des adjectifs relationnels, s'avère efficace en termes de précision. L'amélioration de son rappel nécessite toutefois l'utilisation de corpus de grande taille ayant subi un étiquetage morpho-syntaxique préalable.

Abstract

Many of the Natural Language Processing applications that deal with sublanguages are terminological extraction tools. They consequently tend to focus on the identification of noun and prepositional clauses and their modifiers. Using a corpus of English medical research articles, we suggest a semi-automatic phraseological extraction system. In order to separate terms from collocations within the category that fits the <Adjective – Noun> pattern, we experiment with the approach that consists in extracting adjectives that co-occur with adverbs. This method, which makes it possible to eliminate most relative adjectives, proves to provide good precision. However, the use of a larger POS-tagged corpus will be necessary in order to improve the method's recall.

Mots Clés

Termes, collocations, adjectifs, noms, corpus, anglais de spécialité.

Keywords

Terms, collocations, adjectives, nouns, corpus, English for Specific Purposes.

1 Introduction

L'une des principales caractéristiques des langues de spécialité est leur haute densité terminologique. Cette prédominance quantitative de la terminologie fait de l'extraction terminologique un champ d'investigation privilégié en traitement automatique des langues. Mais l'un de ses effets est également de rendre plus délicate l'extraction phraséologique et l'étude des collocations. Les expressions de patrons syntaxiques identiques (par exemple, celle du type <Adjectif – Nom> en anglais)[1] peuvent être repérées automatiquement grâce à un étiquetage morpho-syntaxique, mais ce sont des caractéristiques sémantiques qui distinguent les termes des collocations.

L'étude des collocations en langue de spécialité est une activité qui a été considérablement facilitée par la mise au point de logiciels d'extraction et de concordance. Parmi les travaux les plus récents, on peut citer ceux de G. Williams dans le domaine de la biologie végétale et ceux de M-C. L'Homme dans le domaine de l'informatique. Williams (1998) recherche les cooccurrences significatives entre deux lexèmes, non seulement afin d'extraire des binômes ou des expression polylexicales mais aussi pour déterminer leur « rôle thématique, facteur de cohésion textuelle », ce qui l'amène à utiliser la notion de « réseaux de collocations». L'Homme (1998) concentre principalement son étude sur les collocations à base verbale, en particulier dans le domaine de l'informatique. Elle effectue la description des verbes spécialisés dans une optique de traitement automatique et la situe à différents niveaux (syntaxique, sémantique et combinatoire).

L'intérêt que manifeste L'Homme pour la description des structures verbales se justifie pleinement dans le cadre du TALN, car si les progrès de l'extraction terminologique sont rapides, les problèmes que pose le traitement automatique du groupe verbal, notamment dans le cas de la traduction assistée par ordinateur, rendent cette formalisation indispensable. La description des collocations verbales en langue de spécialité comporte cependant un certain nombre d'écueils. Le premier est la faible fréquence des formes verbales relativement aux groupes nominaux, qui implique l'utilisation de corpus de très grande taille si l'on souhaite atteindre la significativité statistique. Le second est la nécessité du recours à un corpus arboré ou à une analyse syntaxique pour la détection de l'ensemble des structures faisant intervenir des syntagmes prépositionnels, la longueur des groupes nominaux objets n'assurant pas une proximité suffisante entre ces syntagmes prépositionnels et les verbes dont ils dépendent pour qu'une simple recherche de cooccurrents donne des résultats fiables.

[1] Bourigault et Jacquemin (2000), soulignent la relative facilité de cette extraction dans une langue comme l'anglais, qui construit ses termes complexes essentiellement par juxtaposition d'unités lexicales pleines, alors que le français use abondamment des prépositions et des déterminants, phénomène qui rend la distinction entre terme et syntagme libre plus difficile à saisir par les outils automatiques.

2 L'acquisition automatique de collocations

Le repérage de combinaisons lexicales correspondant à un patron syntaxique donné est particulièrement adapté à l'outil informatique, puisque celui-ci est d'une grande puissance pour la génération de collocations tirées de textes numérisés. Le problème de l'automatisation du tri entre termes et collocations est l'une des difficultés rencontrées par les programmes d'extraction terminologique, pour lesquels des suites comme *aspects observés* ou *augmentation localisée* font partie du bruit à éliminer. Nous allons tenter de décrire un modèle inverse, visant à opérer l'extraction des seules collocations. Clas (1994) suggère un classement des collocations lexicales en divers groupes basés sur une fonction syntagmatique qui intègre les six catégories suivantes (p. 578) :

1. verbe et nom, où le verbe a un contenu sémantique très général proche simplement de « faire » (*prononcer un discours*) ;

2. nom et adjectif (*rude épreuve*, *marque distinctive*) ;

3. adverbe et adjectif (*vachement bon*) ;

4. verbe et adverbe (*boire goulûment*)

5. nom (sujet) et verbe (*la cloche sonne*, *le chat miaule*, *l'abeille bourdonne*) ;

6. marquage de la quantité (unité ou collectif) du nom (*essaim d'abeilles*, *troupeau de vaches*, *pincée de sel*, *barre de chocolat*).

Les exemples donnés ici correspondent à des degrés de figement et de lexicalisation divers, et les restrictions s'appliquant à certaines des catégories définies par Clas pourraient être élargies[2]. Par ailleurs, les associations décrites dans le cadre des quatre premières catégories semblent constituer un choix de combinatoire lexicale possible parmi d'autres, alors que celles des catégories 5 et 6 ne peuvent s'accommoder de variations (le marquage de la quantité fait d'ailleurs l'objet d'explications et d'exercices dans la plupart des grammaires de l'anglais). La Figure 1 représente un schéma combinatoire des différentes catégories morpho-syntaxiques, établi à partir des catégories décrites plus haut.

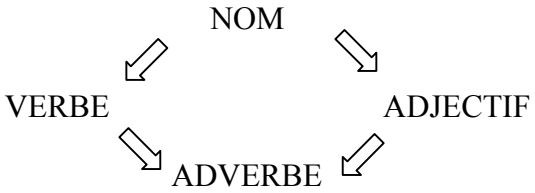

Figure 1 : Combinatoire collocationnelle des catégories morpho-syntaxiques.

[2] Par exemple, faut-il nécessairement limiter les collocations de la première catégorie à celles qui comprennent un verbe support ? Dans le cas des collocations anglaises *set* / *beat (*ou *break) / hold a record* (établir / battre / détenir un record), cela reviendrait à n'accepter que la première et pas les autres dans cette catégorie.

3 Outils informatiques utilisés et extraction des combinaisons <Adjectif-Nom>

Nous avons utilisé pour tester notre technique d'extraction un corpus constitué de 58 articles du *Journal of the American Medical Association*. La longueur totale du corpus est de 134 000 mots. Le calcul de l'indice de probabilité de cooccurrence des collocants a été effectué à l'aide du logiciel Tact[3]. Nous avons testé une méthode d'extraction correspondant à l'un des deux accès possibles aux collocations <Adjectif – Nom> en anglais selon la Figure 1, celle qui utilise comme point de départ les cooccurrents des adverbes (l'ordre suivi est alors Adverbe → Adjectif → Nom). Après avoir évalué divers programmes d'étiquetage automatique, nous avons soumis notre corpus à un étiquetage morpho-syntaxique selon les normes du corpus LOB (Lancaster-Oslo/Bergen)[4].

La technique utilisée pour l'isolation des collocations des termes complexes repose sur une caractéristique connue des adjectifs relationnels, qui rentrent fréquemment dans la composition des termes de patron syntaxique <Adjectif – Nom>. Daille (2001) fait remarquer qu'outre les critères de repérage purement morphologiques, telle la présence majoritaire de certains suffixes de dérivation, on peut également utiliser certaines caractéristiques syntaxiques ou sémantiques des adjectifs relationnels, par exemple le fait qu'ils ne sont pas combinables avec certains adverbes, en particulier les adverbes de degré (on ne dit pas **une production très laitière*). Nous avons donc tenté de repérer automatiquement les seuls adjectifs modifiés par des adverbes dans notre corpus, puis d'extraire les noms entrant en cooccurrence avec ces adjectifs, cette méthode ayant pour but d'éliminer les combinaisons contenant des adjectifs relationnels tels que *coronary* ou *pericardial*, qui constituent le plus souvent des termes de la langue médicale.

Nous avons donc isolé les 1057 combinaisons <Adverbe – Adjectif> de notre corpus médical. Après élimination manuelle des suites contenant certains adjectifs relationnels (*atherogenic*, *bactericidal*, *hyperinsulinemic*, *metastatic*, *occlusive*, *oral, vascular*), de celles dont les adverbes ne modifiaient pas l'adjectif qu'ils précédaient (*therefore*, *still*, *thus*, *likewise, perhaps, also, once, enough, prior*), et de celles contenant des adverbes pouvant précéder des adjectifs relationnels[5], il restait 869 occurrences d'un adjectif précédé d'un adverbe[6]. Ces occurrences regroupaient 596 combinaisons distinctes, contenant 342 adjectifs distincts. Nous avons ensuite extrait du corpus 3121 occurrences du patron syntaxique <Adjectif – Nom> qui contenaient ces 342 adjectifs et formaient 2034 combinaisons distinctes.

Le Tableau 1 donne la liste des 67 séquences de fréquence supérieure à 6 dans notre corpus. La précision de cette méthode d'extraction des collocations est bonne, la liste contenant très

---

[3] TACT (Copyright (c) 1989 John Bradley, *University of Toronto*), logiciel accessible à l'URL suivante : http://www.chass.utoronto.ca:8080/cch/tact.html

[4] Le projet AMALGAM (décrit par Atwell et al., 2000) met à la disposition des internautes plusieurs programmes d'étiquetage grammatical de l'anglais. Il est accessible à l'adresse suivante : http://agora.leeds.ac.uk/amalgam/

[5] Certains adverbes tels que *largely* semblent entrer en cooccurrence principalement avec les adjectifs relationnels, dans des contextes tels que *Ocular rosacea is largely **vascular** in its origin* ou *The residual symptoms are largely **cognitive / vegetative***.

[6] Nous avons retenu les deux classes d'adverbe étiquetées RB (correspondant aux adverbes à dérivation en –ly) et QL (*qualifier,* contenant les adverbes *too*, *least*, *very*, *as*, *so*, *more* et *less*).

peu de termes complets[7] (on relève cependant un certain nombre d'emplois de l'adjectif *low* précédant le premier terme d'une lexie complexe comme *ejection fraction* ou *compliance rate*). Quant à son rappel, il est difficile à évaluer sans une exploitation manuelle intégrale du corpus. La technique utilisée fait que le taux de rappel doit logiquement augmenter avec la taille du corpus, puisqu'il suffit d'une occurrence post-adverbiale d'un adjectif pour que celui-ci soit ajouté à la liste des éventuels cooccurrents des noms qui servent de base à la deuxième passe.

Adjectif	Nom	Freq.
physical	examination	50
increased	risk	23
clinical	trials	22
clinical	probability	19
predictive	value	18
diagnostic	test	17
relative	risk	15
clinical	findings	15
recent	studies	11
high	risk	11
clinical	assessment	11
toxic	effects	10
physical	findings	10
malignant	effusions	10
increased	prevalence	10
high	levels	10
diagnostic	tests	10
higher	levels	9
adverse	effects	9
sensitive	thromboplastins	8
lower	liver	8
lower	levels	8
higher	risk	8
early	detection	8
controlled	trials	8
small	number	7
physical	diagnosis	7
lower	edge	7
low	probability	7
clinical	presentation	7
antimicrobial	therapy	7

Tableau 1 : Combinaisons lexicales <Adjectif – Nom> dont l'adjectif entre en co-occurrence avec les adverbes du corpus médical.

4 Conclusion

L'exploitation d'un corpus étiqueté morpho-syntaxiquement permet des regroupements des phénomènes de cooccurrence par patrons syntaxiques qui peuvent s'avérer utiles dans la démarche consistant à séparer les candidats termes des « candidats collocations », en particulier dans le cas des structures faisant intervenir la prémodification adjectivale du nom. La méthode que nous avons testée obtient une précision correcte, mais demande à être testée sur un corpus de taille supérieure. L'amélioration de son rappel peut être envisagée à partir d'une liste de cooccurrents correspondant au patron syntaxique <Adjectif – Nom> n'ayant pas fait l'objet d'une sélection préalable. La distinction entre termes et collocations de la langue spécialisée reste parfois difficile à opérer, et nécessite à un stade ultérieur la prise en compte de traits sémantiques réglementant la combinatoire des noms et des adjectifs. Dans cette

[7] L'élimination totale du bruit est difficile si l'on souhaite recenser toutes les collocations, car certains adjectifs ont des emplois relationnels sans qu'ils entrent pour autant exclusivement dans la composition de termes. Par exemple, la combinaison la plus fréquente de notre liste, *physical examination*, est un terme, mais ce n'est pas nécessairement le cas de *physical findings* et de *physical diagnosis*.

optique, l'accès à la liste des candidats termes rejetés par les linguistes et les spécialistes du domaine à l'issue du processus de validation terminologique lors d'études antérieures pourrait s'avérer précieux.

Références

Atwell E. et al. (2000), "A comparative evaluation of modern English corpus grammatical annotation schemes" in *ICAME Journal N° 24*, pp 7-23.

Bourigault D., Jacquemin C. (2000), Construction de ressources terminologiques, In J. PIERREL, Ed., *Ingénierie des langues*, Chapitre 9. Hermès.

Clas A. (1994), « Collocations et langues de spécialité », in *Meta*, XXXIX, 4, pp. 576-580.

Daille B. (2001), "Qualitative terminology extraction: Identifying relational adjectives" in Bourigault D., Jacquemin C., L'Homme M.-C., (eds) *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, pp 149-166.

Fellbaum C. (ed.) (1998), *WordNet: An Electronic Lexical Database*. MIT Press.

Gaussier E. (2001), "General considerations on bilingual terminology extraction" in Bourigault D., Jacquemin C., L'Homme M.-C., (eds) *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, pp 167-184

Langlois L. et Plamondon P.(1998), "Le répérage automatique de collocations équivalentes à partir de bitextes" In Fontenelle T., Hiligsmann P., Michiels A., Moulin A., Theissen S. (eds), *Euralex'98: Proceedings of the Eighth Euralex International Congress* Liège, Université de Liège, pp 175-186.)

L'Homme M.C. (1998), « Définition du statut du verbe en langue de spécialité et sa description lexicographique. » *Cahiers de lexicologie 73 (2),* pp. 61-84.

Maniez F. (1999), « The use of electronic corpora and lexical frequency data in solving translation problems », in Altenberg B. & Granger S. (eds), *Lexis in Contrast*, Amsterdam, John Benjamins, 2001.

Maniez F. (2001), « Désambiguïsation syntaxique des groupes nominaux en langue spécialisée : le cas des adjectifs en anglais. » Actes du colloque TALN de Tours, 2-5 juillet 2001, Tome 1, pp. 273-282.

Mel'cuk I. (1984), *Dictionnaire explicatif et combinatoire du français contemporain*, Montréal, Les Presses de l'Université de Montréal.

Williams G. (1998), "Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles". *International Journal of Corpus Linguistics*. *Vol 3/1,* pp. 151-171.