Extraction de collocations à partir de textes

Béatrice Daille

IRIN
Université de Nantes
daille@irin.univ-nantes.fr

Résumé – Abstract

Les collocations sont intéressantes dans de nombreuses applications du TALN comme la l'analyse ou la génération de textes ou encore la lexicographie monolingue ou bilingue. Les premières tentatives d'extraction automatique de collocations à partir de textes ou de dictionnaires ont vu le jour dans les années 1970. Il s'agissait principalement de méthodes à base de statistiques lexicales. Aujourd'hui, les méthodes d'identification automatique font toujours appel à des statistiques mais qu'elles combinent avec des analyses linguistiques. Nous examinons quelques méthodes d'identification des collocations en corpus en soulignant pour chaque méthode les propriétés linguistiques des collocations qui ont été prises en compte.

Mots clés : collocations, statistiques lexicales, extraction automatique

Collocations are interesting for several NLP applications such as language generation or analysis and monolingual or bilingual lexicography. The first approaches to finding collocations appeared in 1970's and were statistically based. Today, the methods adopted for the identification of collocations still include statistics but also linguistic processing. We introduce a few approaches to finding collocations in corpora. For each method, we precise the linguistic characteristic of collocation which as been taken into account.

Keywords : collocations, automatic identification, lexical statistics

L'extraction ou l'identification de collocations à partir de textes peut s'effectuer dans un cadre monolingue ou bilingue. Les méthodes utilisées sont principalement statistiques mais elles peuvent être combinées à des analyses linguistiques comme l'analyse morphosyntaxique ou l'analyse syntaxique, ou encore être associées à d'autres types de ressources linguistiques comme des dictionnaires de langue générale ou de spécialités, ou des thesaurus comme Wordnet (Fellbaum 1998). Plus la méthode intégrera de connaissances linguistiques plus il sera possible d'obtenir un classement fin des collocations : ainsi, uniquement avec des statistiques, aucune catégorisation n'est permise ; avec une analyse syntaxique, une catégorisation en fonction des parties des discours est possible ; avec une analyse syntaxique et d'autres ressources linguistiques, une catégorisation sémantique est envisageable.

1 Mesures statistiques

De nombreuses mesures statistiques ont été utilisées pour extraire des collocations à partir de textes (cf. Muller (1992), Church et al. (1994) et Manning et Schütze (1999) pour un panorama des différentes mesures statistiques). Nous présentons ci-dessous trois mesures :

1.1 Fréquence

Compter simplement le nombre de fois que des unités textuelles apparaissent ensemble permet lorsque le nombre est important de mettre en évidence des associations qui peuvent se révéler de nature collocative. Cette mesure s'appuie sur le critère définitoire d'"habituel"de la collocation. Le problème réside principalement dans l'identification des éléments à compter. Ainsi si l'on se contente de lister à partir d'un texte les combinaisons de deux mots les plus fréquentes, on obtient des combinaisons de peu d'intérêt intégrant de nombreux mots outils.

1.2 Information mutuelle

L'information mutuelle (Fano, 1961) compare la probabilité d'observer deux éléments ensembles avec la probabilité d'observer ses deux éléments séparément.

$$I(x,y) = \log P(x\ y)/p\ (x)\ p(y)$$

Cette formule peut être estimée par la formule suivante proposée par (Dagan et Church, 1993) :

$$\hat{I}(x,y) = \log_2 N\ freq(x,y)/freq(x)\ freq(y)$$

où freq(x,y) est la fréquence de l'événement (x,y), freq(x), freq(y) sont les fréquences respectivement de x et y et N la taille du corpus.
Ce score évalue donc le lien que peuvent avoir deux éléments entre eux : un score positif indique que deux éléments apparaissent plus fréquemment ensemble que séparément ; un score négatif que les éléments sont en distribution complémentaire. Ces deux éléments peuvent être des mots rencontrés dans un corpus monolingue. La méthode est alors la suivante : sur le corpus, on déplace une fenêtre de taille paramétrable de longueur *n*. Le premier mot m_1 de la fenêtre est associé à tous les mots m_2, m_3, … , m_n qui le suivent à l'intérieur de cette fenêtre. Des couples : (m_1, m_2), … , (m1, m_n) sont ainsi formés. Chaque

couple est accompagné de son nombre d'occurrences et éventuellement de la variance des distances séparant les deux mots pour chaque occurrence. Cette mesure a été utilisée par (Church et Hanks, 1990) sur l'anglais et a permis de détecter lorsque la variance des distances est assez importante des associations sémantiques intéressantes et lorsque la variance des distances est faible des couples pouvant relever de la collocation, mais aussi des expressions figées, des noms propres, etc.

1.3 Z-score

Le z-score défini par (Berry-Rogghe, 1973) s'appuie sur une définition statistique de la collocation : l'association syntagmatique de deux éléments lexicaux quantifiables dans un texte correspond à la probabilité qu'à une distance n de l'élément x, les éléments a, b, c… soient rencontrés. Autrement dit, pour chaque éléments lexicaux d'un texte, il est possible de construire un ensemble ordonné de ses collocatifs significatifs. Le z-score qui correspond à une approximation de la distribution binomiale sert à mesurer ce caractère significatif :

$$z = (K-E) / \sqrt{} Eq \text{ ou } q= (1-p)$$

avec:

- Z : le nombre total d'éléments lexicaux dans le texte
- A : un mot donné apparaissant F_n fois dans le texte
- B : un collocatif de A apparaissant F_c fois dans le texte
- K : le nombre de co-occurrences entre B et A
- S : la taille de la fenêtre, c'est-à-dire la distance maximale autorisée entre A et B
- P : la probabilité d'occurrence de B n'importe où dans le texte sauf avec A ; $p = F_c / (Z - F_n)$
- E : nombre probables de cooccurrences entre B et A ; $E = p F_n S$

Berry-Rogghe (1974) illustre cette formule en calculant l'ensemble des collocatifs de *house* (*maison*) dans le livre de Dickens, "A Christmas Carol". Les collocatifs recevant une valeur élevée du z-score sont successivement : *sold, commons, decorate, this, empty, buying, painting, opposite*, etc.

Si les statistiques permettent de détecter des cooccurrences intéressantes dans les textes, elles ne permettent que de proposer des collocations possibles sans effectuer aucune classification.

2 Mesures statistiques et analyse linguistique

L'utilisation de connaissances linguistiques permet de vérifier la propriété de "syntactiquement bien formée" de la collocation en plus de la propriété d'"habituel" prise en charge par les statistiques.

Nous intéresserons ci-après uniquement aux collocations lexicales composées de deux unités lexicales, la base et le collocatif, qui peuvent être classées en fonction de leurs parties du discours. Les cinq types les plus rencontrés de collocations lexicales sont les suivants :

(a) Nom + Adjectif : *amour platonique*, *colère noire*

(b) Nom + Nom : *bourreau des cœurs*

(c) Verbe + Adverbe : *exploiter efficacement*

(d) Adjectif + Adverbe : *sexuellement transmissible*

(e) Nom + Verbe : *commettre une agression, retirer de l'argent*

2.1 Mesures statistiques et analyse morphosyntaxique

Pour les catégories de (a) à (d), une analyse morphosyntaxique qui assigne à chaque mot d'un texte une étiquette grammaticale, éventuellement accompagné de son lemme, permet soit d'effectuer un filtrage sur les catégories du discours, soit d'écrire des grammaires locales décrivant des syntagmes comme le sous-groupe verbal ou le groupe nominal simplifié. Dans le deuxième cas, la notion de fenêtre textuelle est élargie à la notion de syntagme.

Cette analyse morphosyntaxique a été particulièrement utilisée pour l'identification automatique de combinaisons lexicales spécialisées ou termes complexes. Daille (1996) pour le français et l'anglais, Pazienza (1999) pour l'italien, Heid (1999) pour l'allemand ont défini des grammaires locales permettant d'extraire d'un corpus les séquences nominales caractéristiques des collocations de type (a) et (b). Les statistiques présentées dans la section 1 sont ensuite appliquées sur ces séquences textuelles de manière à obtenir un classement. La simple fréquence donne de bons résultats ; l'information mutuelle est plus discutable car elle donne trop de poids aux associations rares et donc isole plus des expressions figées que véritablement des combinaisons lexicales spécialisées.

2.2 Mesures statistiques et analyse syntaxique

Pour la catégorie (e) qui cherche à mettre en évidence des arguments privilégiés d'un verbe, comme le sujet ou l'objet, une analyse syntaxique est nécessaire. Il faut identifier correctement les différents arguments et pouvoir prendre en compte les différentes constructions phrastiques. Par exemple, la construction à verbe support, *commettre une agression*, peut se rencontrer à l'actif *Jean commet une agression*, au passif, *l'agression a été commise par Jean*, au sein d'une relative, *l'agression que Jean a commise*, etc. L'extraction d'une collocation de type Nom-Verbe où par exemple le nom est l'objet direct du verbe nécessite un parcours de l'arbre syntaxique et une identification correcte des têtes

prédicatives. Là encore les mesures statistiques peuvent être utilisées pour classer ces couples Nom-Verbe. Elles sont cependant moins efficaces du fait d'une moindre représentativité des occurrences sauf dans le cas de constructions verbales figées comme *mettre en œuvre*, *tenir compte*, etc. qui ne sont pas des collocations. Pour extraire certaines collocations, comme celle faisant intervenir un verbe support, Grefenstette et Teufel (1995) ont utilisé la forme concurrente verbale au schéma verbe support et nominalisation en ne retenant que les nominalisations conservant la structure argumentale du verbe d'origine comme dans l'exemple suivant : *He appealed to President Aquino … / He made a public appeal to President Aquino.*

Une analyse linguistique associée aux mesures statistiques permet d'obtenir un classement linguistique des collocations candidates et ainsi d'ignorer les colligations ou les associations thématiques. Cette analyse ne permet pas de trancher sur le statut collocatif d'un type de combinaison syntaxique.

3 Mesures statistiques, analyse et ressources linguistiques

Dans la section précédente, le classement des collocations est syntaxique. Si on veut prendre en compte le critère "arbitraire" des collocations, il faut utiliser un classement sémantique. Les fonctions lexicales de Mel'cuk I. (1984) fournissent un tel cadre : elles permettent de décrire certaines des relations sémantiques existantes entre la base et le collocatif comme les modificateurs d'intensité, les constructions à verbe support ou encore les verbes de "réalisation". Les fonctions lexicales *Magn, Oper1, Real3* en sont des exemples.

Pour extraire des collocations, des premières expériences ont été menées à l'aide de Wordnet, soit pour l'anglais (Pearce, 2001), soit pour l'espagnol (Wanner et Alonso Ramos, 2000). L'idée est d'utiliser les liens de synonymie et d'hyperonymie présents pour chaque entrée du thesaurus pour calculer une distance sémantique entre le collocat et la base. (Pearce, 2001) utilise les liens de synonymie pour extraire les collocations en prenant comme hypothèse que le statut non compositionnel du sens de la collocation correspond à la non-substitution du collocat par l'un de ses synomymes. (Wanner et Alonso Ramos, 2000) exploite les liens d'hyperonymie pour extraire des collocations et les classer sémantiquement selon le formalisme des fonctions lexicales. Pour cela, ils examinent la corrélation entre le sens de la base et le sens du collocatif par rapport au sens exprimé par la fonction lexicale.

Références

Berry-Rogghe, G. L. M. 1971. The computation of collocations and their relevance in lexical studies. *The Computer and Literary Studies*. Eds Aitken A.J. et al. Edinburgh: Edinburgh University Press.

Berry-Rogghe, G. L. M. 1974. Automatic identification of phrasal verbs. *Computer in the Humanities*. p. 16-27. Ed J.L. Mitchell. Edinburgh: Edinburgh University Press.

Church K.W. et Hanks P. 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1) :22-29.

Church, K., Gale, W., Hanks, P., Hindle, D. & Moon, R. 1994. Lexical Substitutability, in Atkins & Zampolli (eds) *Computational Approaches to the Lexicon*, Oxford University Press, pp.153-177.

Daille, B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, in P. Resnik et J. Klavans (éds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language,* ISBN 0-262-11218-3, p. 49-66, MIT Press, Cambridge, MA, USA, 1996.

Dagan, I et K. Church. 1993. Identifying and Translating Technical Terminology, IJCAI'1993.

Fano, R. 1961. *Transmission of Information*, Cambridge, MA: The MIT Press.

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

Grefenstette, G. et Teufel, S. 1995. Corpus-based Methods for automatic identification of support verbs for nominalizations. *EACL'1995*.

Heid, U. 1999. A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology*. 5(2):161-182.

Manning et Schütze. 1999. Collocation in *Foundations of Statistical Natural Language Processing*.

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

Muller, C. 1992. *Initiation aux méthodes de la statistique linguistique*. Collection Unichamp. Paris: Champion

Pazienza, M. T. 1999.A domain-specific terminology-extraction system. Terminology. 5(2):183-202.

Pearce, D. 2001. Synonymy in collocation extraction. In *NAACL 2001 Workshop: WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Carnegie Mellon University, Pittsburgh.

Wanner, L. et Alonso Ramos, M. 2000. Vers une approche sémantique pour l'identification des collocations en corpus, B. Daille et G. Williams (eds), In *Actes de la Journée d'études de l'ATALA du 13 janvier 2001 - La collocation*, IRIN, Université de Nantes, n° 00.13.