A Hybrid Multilingual Method to Extract Collocations from Corpora

Amalia Todirascu
FDT, LILPA, Université de Strasbourg
22, rue René Descartes, 67084 Strasbourg cedex
todiras@unistra.fr

WG 3: Statistical, hybrid and multilingual extraction methods

We present a method to automatically extract collocations from monolingual corpora for French and Romanian. The monolingual extraction algorithm combines statistical methods and linguistic information to select relevant candidates. Filters are defined by exploiting some morpho-syntactic properties of collocations (preference for some specific categories of determiners, for number or gender, person or voice). For our experiments we use tagged and lemmatized corpora. We present some evaluation results of monolingual extraction for French and Romanian. First, each module is evaluated on a monolingual corpus (composed of several genres) and the results are compared with existing dictionaries. The extraction method might be adapted for other languages.

Multiword expressions include idioms (*kick the bucket*), compound nouns or verbs (*to get up*), domain-specific terms (*arbre à cames*). In this presentation, we focus on collocations. Collocations are composed of two or several lexical units, with specific syntactic and semantic behavior. Their sense is often non-compositional and they present an important degree of syntactic variability. Due to these properties, collocations are difficult elements for NLP systems but also for human translators or for language learners. No consensual definition is given for collocations, each system adopts various criteria to identify these expressions (Gledhill, 2007):

- statistic criteria : collocations are word cooccurrences (frequent word associations) (Sinclair, 1991);
- linguistic criteria: the words composing the collocations are linked by syntactic relations (Haussmann, 2004) (Tutin, 2010);
- pragmatic : the collocation is know by its appropriate use in context.

Several extraction systems exploit one or several of these aspects. Frequent word associations are exploited by (Evert, 2005), but these methods generate noisy output (simple cooccurrences considered as collocations). Other systems identify syntactic relations between the words (Seretan et Wehrli, 2009) (Savary, 2008), (Constant *et al*, 2013) but these methods require detailed syntactic analysis. To avoid this drawback, hybrid methods such as (Krenn, 2000) (Smadja, 1993), and more recent work such as (Nissim, Zanninello, 2013) were proposed, combining statistical methods and linguistic properties.

In our approach, we consider that collocations are characterized by their usage in context, following Heid and Ritz (2005), Odjik(2013). The context of these expressions (the words composing the collocations but also the neighbors words) are exploited to identify relevant collocation candidates. The context shows strong preferences for some morpho-syntactic properties such as number or gender for nouns or the adjectives, mood, time or person, valency for the verb. For example, V-N collocations are characterized for a set of properties, specific to each language (Gledhill, 2007) :

- noun's properties : preference for definite or zero determiner, preference for singular or for plural number (both Romanian and French), preference for accusative or dative case of the object (only for Romanian);
- verb's properties : preference for some specific arguments (marked by a given preposition), preference for passive voice.

We manually analyze collocations using functional systemic grammar (Halliday, 1985) for French

and Romanian (Todirascu *et al*, 2008). Based on the identified properties, we define a set of linguistic filters to select relevant candidates. These language-specific filters use lemma and morpho-syntactic properties. The results are manually filtered.

Using these linguistic properties, we develop a hybrid method combining statistical and linguistic information to automatically extract collocations from tagged and lemmatized corpus using a multilingual POS tagger, TTL available for Romanian (Ion, 2007) and for French (Todirascu *et al*, 2011). First, we identify frequent word combinations (in a window of 11 words), by ordering using Loglikelihood measure (Dunning, 1990). Then, for each language, we apply the set of linguistic filters used to extract collocation candidates. The evaluation focuses on the VN collocations. We evaluate the results of the module on comparable corpora from the same domain (medicine, computer science, news) and we compare to the results obtained for law texts. For French, we compared the results with existing resources : 96% of the candidates are found in Base Lexicale du Français (Verlinde *et al*, 2003), and only 93% were present in the LADL tables (Gross, 1993, Laporte *et al*, 1998)). For Romanian, we compare the results with a multilingual dictionary (Todirascu *et al*, 2008) containing 250 entries for French and Romanian. 84% of the candidates were available in this dictionary. We present a qualitative and quantitative analysis of the monolingual extraction and we compare the results obtained for the two languages.

Références

Constant, M., Sigogne A., Watrin, P. (2013) Stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un analyseur syntaxique en constituants. Traitement Automatique des Langues. Vol. 54(1). 24 pp.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics, 19,(1).

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart

Gledhill (C.) 2007. « La portée : seul dénominateur commun dans les constructions verbo-nominales. » In *Actes du 1er colloque Res per nomen,* Reims.

Gross, M (1993) Les expressions figées en français, L'Information grammaticale, Vol 59, no 1, pp. 36-41.

HAUSMANN (Franz Josef) : 2004, « Was sind eigentlich Kollokationen? », in STEYER (K), eds., Wortverbindungen – mehr oder weniger fest, pp. 309-334

HALLIDAY (Michael) : 1985, An Introduction to Functional Grammar, (London, Arnold)

Heid, U. et Ritz J. (2005). Extracting collocations and their contexts from corpora, Actes de Conference on Computational Lexicography and Text Research, Budapest

Ion, R. (2007). Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română, Teză de doctorat, Academia Română, Bucureşti, mai 2007, 148 p.

Krenn, B.(2000) *The Usual Suspects : Data-Oriented Models for Identification and Representation of Lexical Collocations*. PhD thesis, Universität des Saarlandes.

Odijk, J., 2013, DUELME: Dutch Electronic Lexicon of Multiword Expressions, In Francopoulo, G. (ed.) LMF: Lexical Markup Framework, ISTE / WILEY, 2013

SAVARY A., (2008), Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches, in Linguistic Issues in Language Technology 1(2), CSLI, pp. 1-53

Seretan, Violeta and Eric Wehrli (2009). Multilingual collocation extraction with a syntactic parser. Language Resources and Evaluation, 43(1), 71–85.

Sinclair, J.(1991) Corpus, Concordance, Collocation, Oxford University Press

Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177

Todirascu A., Gledhill C., Stefanescu D. (2009) Extracting Collocations in Contexts. In Vetulani, Z., Uszkoreit, H. (eds.) Responding to Information Society Challenges: New Advances in Human Language Technologies, LNAI 5603, Springer-Verlag, ISBN 978-3-642-04234-8.

Todirascu, A., Heid, U., Stefanescu, D., Tufis, D., Gledhill, C., Weller M., Rousselot François (2008) « Vers un dictionnaire de collocations multilingue » Cahiers de Linguistique, Université de Louvain.

Tutin A. (2010). Les collocations dans les dictionnaires monolingues spécialisés de collocations. 2e Congrès Mondial de Linguistique Française (CMLF-2010).

Verlinde, S., Selva, T., Binon, J. (2003) « Les collocations dans les dictionnaires d'apprentissage : repérage, présentation et accès», in Grossmann (F.), Tutin, (A.), dir. Les collocations : analyse et traitement, (Amsterdam : De Werelt), pp. 105-115.