Creating a Multilingual Collocation Dictionary from Large Text Corpora

Luka Nerima, Violeta Seretan, Eric Wehrli
Language Technology Laboratory (LATL), Dept. of Linguistics
University of Geneva
CH-1211 Geneva 4, Switzerland
{Luka.Nerima, Violeta.Seretan, Eric.Wehrli}@lettres.unige.ch

Abstract

This paper describes a system of terminological extraction capable of handling multi-word expressions, using a powerful syntactic parser. The system includes a concordancing tool enabling the user to display the context of the collocation, i.e. the sentence or the whole document where the collocation occurs. Since the corpora are multilingual, the system also offers an alignment mechanism for the corresponding translated documents.

1 Introduction

Cross-linguistic communication frequently raises the problem of the proper understanding of idiomatic expressions, i.e. multi-word expressions whose meaning differs from the composition of the individual meaning of their parts. The importance of multi-word expressions is widely recognized in the domains of translation and terminology. These expressions can usually not be translated literally, and one must find adequate correspondences in the target language.

This paper describes a system of terminological extraction capable of handling multi-word expressions, based on a detailed linguistic analysis. The originality of our approach comes from the fact that collocations are not extracted from raw texts, but rather from syntactically parsed texts. The linguistic analysis selects potential pairs of words, as only the words occurring in a specific syntactic configuration will be taken into account for further statistical processing. Such a chain of processes significantly increases the quality and the relevance of the extracted collocations.

This system will be applied to textual corpora from the World Trade Organisation (WTO), which consist in parallel documents in three languages: English, French and Spanish. All the examples given in this paper are taken from these corpora. Ultimately, the system will enrich the workbench of translators and terminologists of this organization.

2 Collocations

The notion of "collocation'" is difficult to define in a very precise way. Commonly used to refer to an *arbitrary and recurrent word combination* (Benson, 1990), it is also often taken as a conventional combination of two or more words, with a more or less transparent meaning. "Conventional combinations" means that native speakers recognize such combinations as the "correct" way of expressing a particular concept. For instance, substituting one term of a collocation with a synonym or a near-synonym is usually felt by native-speakers as being "not quite right", although perfectly understandable, e.g. *firing ambition* vs. *burning ambition* or in French *exercer une profession* vs. *pratiquer une profession* (*to practice a profession*). For further discussion on collocations, see (Gross 1996; Manning and Schütze, 1999; Wehrli, 2000).

In spite of the lack of agreement over what exactly counts as collocation, computational linguists agree that collocations and more generally multi-word expressions play a very important role in many NLP applications such as terminology extraction, translation, information retrieval, and multilingual text alignment. This, along with the ever-increasing availability of very large text cor-

pora, has triggered an important need for tools to extract collocations.

3 Collocation Extraction

The problem of extracting collocations from texts has been much addressed in the literature, in particular since the work of Church at al. (1991), and several statistical packages have been designed for this purpose (see for instance, the Xtract system of Smadja (1993)). Although very effective, those systems suffer from the fundamental weakness that the measure of relatedness they use is essentially the linear proximity of two or more words. As pointed out above, grammatical dependencies provide a more appropriate criterion of relatedness than simple linear proximity.

3.1 Cooccurrence Extraction with Fips

Collocations are extracted from syntactically analysed corpora. The analysis is performed by Fips, a large-scale parser based on an adaptation of Chomksy's "Principles and Parameters" theory (Laenzlinger and Wehrli, 1991). Thanks to the syntactic representation, it is not necessary to take into account any pair of reasonably closed lexical units, but rather the relevant pairs bound by syntactic configurations. We consider eight types of configurations: N-Adj, Adj-N, N-N, N-Prep-N, N-V, V-N, V-Prep-N.

Another argument in favour of a full syntactical analysis is that it solves the problem of all cases of extraposed elements, such as passives, topicalisation, and dislocation. To illustrate some of these points, consider a few examples of the collocations *prendre - mesure* (*take - measure*) and *accepter - amendement* (*accept - amendment*):

> "Regular" phrase: *Le Conseil **prendra** les **mesures** qui pourront être convenues ...*
>
> Passive phrase:... *à moins que des **mesures** ne soient **prises** pour s'assurer ...*
>
> The two terms of the following collocation are separated by no less than 39 words!: *Les **amendements** qui auront uniquement pour objet l'adaptation à des niveaux plus élevés de protection des droits de propriété intellectuelle établis et applicables conformément à d'autres accords multilatéraux et qui auront été **acceptés** dans le cadre de ces accords ...*

3.2 Scoring for Collocation Discovery

In order to identify collocations among the cooccurrences, the system achieves an independence hypothesis testing using the Log-Likelihood-ratio (see for instance (Dunning, 1993)).

Based on the contingency table below for the two lexical items w_1 and w_2 that co-occur,

	w_2	$\neg w_2$
w_1	a	b
$\neg w_1$	c	d

Table 1. Contingency table for cooccurrences.

the system computes the cooccurrence score as follow:

$$\log\lambda = 2\,(a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) - (a+c) \log (a+c) - (b+d) \log (b+d) - (c+d) \log (c+d) + (a+b+c+d) \log (a+b+c+d)).$$

The cooccurrences with a high score are good candidates for collocations. It is however difficult to determine a critical value above which a cooccurrence is a collocation and below which it is not.

3.3 Preliminary Results

Our first experiments concerned the WTO corpus on the Uruguay Round trade negotiation of about 10 millions words for each language. About 380,000 cooccurrences were identified. The cooccurrences were classified in eight classes corresponding to specific syntactic configurations. The table below gives the 12 first cooccurrences of type V-N ranked by the Log-Likelihood ratio.

w_1	w_2	$\log\lambda$	a
faire	objet	2599.73	370
atteindre	objectif	1366.59	200
jouer	rôle	1361.40	136
obtenir	résultat	1315.26	249
priver	revenu	983.20	74
appeler	attention	951.49	112
présenter	proposition	833.02	253
tenir	réunion	791.69	183
importer	marchandise	790.36	87
adopter	ordre du jour	745.84	104
avoir	intention	742.48	123
prendre	décision	712.44	188

Table 2. The 12 best collocations of type V-N obtained.

The results clearly show that the combination of an accurate parsing and the use of Log-Likelihood ratio leads to a promising approach. When unable to create a complete analysis of a sentence, the Fips parser returns chunks of partial analyses. If

the collocation is contained in a chunk, it will be correctly identified by the extraction system. Otherwise, if the two terms do not belong to the same chunk, it will be missed. We did not assess yet the number of missed cooccurrences, but we estimate it at about 10%, i.e. less than the number of cooccurrences missed by the mobile window methods.

Actually, it appears that the terms of the collocations of type N-V (subject - verb), V-N (verb - direct object) and V-Prep-N (verb - prep - object) are separated by more than 5 words in about 20% of cases, justifying our approach.

4 Collocation Dictionary

We used the collocations extracted from the French and English corpora for creating a database of knowledge that integrates collocations and instances of their actual use in language. Corpus evidence for each entry in the collocation dictionary is provided, that can be consulted by the user. We display the context of a collocation for all its occurrences in the analysed corpus, and we offer the user the option to consult the entire document, if interested in a larger context.

The collocation context is represented by the sentence in which the collocation occurs (both collocation's keys occur on the same sentence, as they are in a syntactical relation).

When parallel corpora are available, also the translation equivalents of the collocation context are displayed, thus allowing the user to see how a given collocation was translated in different languages, and in different contexts. This is done using a shallow alignment method, without need to parse the documents in the target languages.

4.1 Contexts Alignment Method

The alignment method is aimed at finding, for a given collocation, the translation of its context in the other document's versions. The granularity of text alignment is the sentence level; we are not concerned with a finer, word-level alignment of text that would, for example, put in correspondence the collocations with their translation equivalent (which can be a collocation or not). We focus on sentence alignment since the aim of the dictionary is to provide instances of collocation's actual use in language, that is, coherent text spans found in the corpora resources. At the same time, we intend to provide a quite precise and delimited context, that's why we do not consider a larger context (such as the whole paragraph).

The specificity of our method consists in the fact that the alignment is local and partial. No complete mapping between sentences is done, but only the mapping for the sentence of the currently visualised instance of collocation. It means that the alignment is done "on the fly", for the source sentence that is actually visualised by the user. This is motivated by the big size of the collocation dictionary and corpora.

The sentence alignment method consists of two parts:

1. the alignment of paragraphs;
2. the alignment of sentences inside the aligned paragraphs.

While the second part is limited for now to a simple linear and 1:1 correspondence between sentences, the paragraph alignment method is more complex; it is length-based and integrates a shallow content analysis. It begins by individuating a paragraph in the target text which is a first candidate as target paragraph, and which we call "pivot". The identification of the pivot is based on the documents size proportion. Once the pivot found, we look in its neighbourhood for the optimal candidate as target paragraph.

We perform two kinds of tests on the paragraphs in this span: a test of paragraph content, and a test of paragraphs relative size matching. The first test compares the paragraphs' numbering (if present). The second one determines the paragraph that best matches the rapports of sizes in a context (a sequence of surrounding paragraphs).

Concluding, our approach to sentence alignment follows a length correlation strategy, as most of the existing works do, e.g. (Gale and Church, 1991; Brown et al., 1991). Individuating the pivot is a function of the documents sizes, and selecting the most likely target paragraph is a function of the relative sizes of paragraphs in the neighbourhood of the pivot. Similarly to (Simard et al., 1992), we exploit the text content in order to find word anchors (the paragraph numbering in our case). Like in (Romary and Bonhomme, 2000) and (Catizone et al., 1989), first the macro (paragraph-level) structure of documents is examined, possibly using mark-up from text encoding.

4.2 Method Evaluation

The preliminary results we obtained show that the alignment method outlined above is quite reliable. We performed the test on a sample of 800 randomly chosen collocation instances, half of which extracted from the English corpus, and half from the French corpus. These subsets were further divided in two parts, corresponding to the two target languages. A human judge verified the correctness of alignment in each case. The tables below show the accuracy rating of the alignment method for each test subset. The average precision is 90.87%.

target \ source	French
English	92.5%
Spanish	93.5%

target \ source	English
French	88.0%
Spanish	89.5%

Table 3. Preliminary results of contexts alignment.

5 Conclusion

We presented a system that integrates the extraction of collocations from a large collection of documents with an extensive use of existing translations for creating a tri-lingual collocation dictionary, with samples of actual use in language. Using past translations as reference for the translator's further work was an idea first proposed by Melby (1982). Many concordance tools, such as (Isabelle et al., 1993), allow the user to consult the translations archives. The specificity of our approach lies, on one hand, in using the translations to extract collocations and visualise their context in all the document's versions, and, on the other hand, in relying on syntactically parsed text.

Acknowledgement

This work is supported by *Geneva International Academic Network (GIAN)*, research project "Linguistic Analysis and Collocation Extraction", approved in 2001. Thanks to Olivier Pasteur for the invaluable help in this research.

References

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1), 23-35.

Brown P., Lai J., and Mercer R. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Canada, pp. 169-176.

Catizone R., Russell G., and Warwick S. (1989). Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*, Detroit.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In Zernick, U. (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, pp. 115-164.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

Gale W. and Church K. (1991). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75-102.

Gross, G. (1996). *Les expressions figées en français*. OPHRYS, Paris.

Isabelle P., Dymetman M., Foster G., Jutras J-M., Macklovitch E., Perrault F., Ren X., and Simard M. (1993). Translation Analysis and Translation Automation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, pp. 1133-1147.

Laenzlinger, C. and Wehrli, E. (1991). Fips, un analyseur interactif pour le français. *TA informations*, 32(2): 35-49.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Melby A. (1982). A Bilingual Concordance System and its Use in Linguistic Studies. In *Proceedings of the Eighth LACUS Forum*, Columbia, SC, pp. 541-549.

Romary L. and Bonhomme P. (2000). Parallel alignment of structured documents. Véronis J. (Ed.). *Parallel Text Processing.* Dordrecht: Kluwer.

Simard M., Foster G., and Isabelle P. (1992). Using Cognates to Align Sentences in Parallel Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, pp. 67-81.

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

Wehrli, E. (2000). Parsing and Collocations, in Christodoulakis, D. (ed.), *Natural Language Processing.* Springer Verlag, pp. 272-282.