FipsCo : A Syntax-Based System for Terminology Extraction

Jean-Philippe Goldman & Eric Wehrli
LATL - Dept. of Linguistics
University of Geneva
CH-1211 Geneva 4
{Jean-Philippe.Goldman, Eric.Wehrli}@lettres.unige.ch

July 27, 2001

Abstract

This paper describes FipsCo, a system developed to extract collocations and multi-word expressions from large text corpora, using a powerful syntactic parser. We will show that syntactically analysed corpora constitute a more reliable basis for MWEs extraction than raw (or shallow parsed) texts. Preliminary results supporting this view will be presented based on both French and English corpora.

1 Introduction

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[1], while, according to Manning (1999), the combined words have the properties of non-compositionality, non-substitutability and non-modifiability. In spite of the lack of agreement over what exactly counts as collocation – how do they differ from compounds and idiomatic expressions (cf. Gross, 1996; Wehrli, 2000)– there is a wide-spread agreement among computational linguists that collocation (whatever its exact definition) and more generally multi-word expressions (henceforth MWE) play a very important role in many NLP applications such as terminology extraction, translation, information retrieval, multilingual text alignment, etc. This, along with the ever-increasing availability of very large text corpora has triggered an important need for tools to extract collocations.

This paper describes FipsCo, a system, currently under development, to extract collocations (and more generally MWEs) from text corpora, based on a powerful syntactic parser. Preliminary results suggest that the acquisition of collocations from syntactically annotated text corpora

[1] "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*).

is far more precise and reliable than the ones obtained by standard statistical methods (even augmented with POS taggers and lemmatizers). There is no doubt that syntactic dependencies, such as the ones expressed by grammatical functions or by modification relations between two terms constitute a more appropriate criterion of relatedness than simple linear proximity, such as being 2 or 3 words away.

We will first describe the method used to extract linguistically motivated cooccurrences corresponding to predefined abstract syntactic patterns, such as subject-verb, verb-direct object, along with the more standard adjective-noun or noun-prep-noun, and then discuss some preliminary results based on an analysis of a corpus of more than 1 million of words. The system has been tested primarily on French data, the English version is still under development. This abstract will only discussed French data and examples. The English version of FipsCo will be described in the full paper.

2 Extracting collocations with FipsCo

In this section we will first provide some arguments in favor of a syntactic approach to collocation extraction, and then describe in some details how the FipsCo system works.

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. To illustrate this point, consider a few examples of the collocation *éprouver - difficulté* ("*experience - problem*") taken in the corpus of the French newspaper "Libération":

(1)a. **éprouvant** de très sérieuses **difficultés**

b. ont **éprouvé** au premier semestre des **difficulté**

c. **éprouvent** toujours les plus grandes **difficultés**

d. **éprouver**, comme pour d'autres entités plus grandes, ou moins européennes dans leurs caractéristiques, de grandes **difficultés**

e. **difficultés** qu'**éprouve**

f. des **difficultés** que peuvent **éprouver**

g. Les **difficultés** de gestion gouvernementale qu'**éprouve**

Such examples show the variety of contexts in which such a simple "verb-direct object" collocation can be found. The distance between the two lexemes can be important. In fact,

an arbitrary number of words can separate them, as illustrated in (1d)[2]. Furthermore, since direct objects can undergo various grammatical operations, such as passivization, extraposition (interrogation, relativization, topicalization, etc.), the canonical order verb-object cannot even be assumed, as illustrated in examples (1e-g), where the two terms occur in reverse order.

Notice that in order to correctly handle such cases, superficial syntactic tagging would not suffice. What is needed is a more comprehensive analysis, capable of interpreting extraposed elements, such as relativized phrases. As illustrated in the (simplified) example (2a) below, a noun phrase such as *the difficulties that he might have experienced* must be recognized as an occurrence of the verb-object collocation *experience-difficulty*. Specifically, this means that the parser must be able (i) to recognize the presence of a relative clause, (ii) to identify the antecedent of the relative pronoun, and (iii) to establish a link between this antecedent and the verb on which the relative pronoun is syntactically dependent.

As illustrated in (2b), the Fips parser (cf. Laenzlinger & Wehrli, 1991, Wehrli, 1997) does exactly that. A chain is established, first, between the canonical direct object position (represented by the empty element [DP e], which stands for the trace of the extraposed element) of the verb *éprouver* and the relative pronoun *que*, and then between this pronoun and its antecedent *difficulté*. In the structures returned by Fips, chains are expressed by means of coindexation. The double chain connecting *difficultés* and the trace of the relative pronoun is represented in the structure (2b) by the index **i**[3].

(2)a. Les difficultés qu'il aurait pu éprouver...
'the difficulties that he might have experienced...

b. [DP les [NP difficultés]$_i$ [CP [DP qu']$_i$ [TP [DP il]$_j$ aurait [VP pu [TP [DP e]$_j$ [VP éprouver [DP e]$_i$]]]]]]

The need for syntactic normalization in order to properly identify particular types of collocations[4] is not restricted to the case of relative clauses. Similar arguments can be made for all cases of extraposed elements, including passives, topicalization, raising, dislocation, etc.

Given the high frequency of such constructions in text corpora, it is clear that the use of syntactically analyzed texts as the basis for the acquisition of collocations yields more precise results than the ones obtained by standard statistical methods. With the development of robust parsers, such as Fips, text corpora can be relatively easily turned into syntactically annotated corpora, opening the way for a wide-scope treatment of collocations.

[2]In the same corpus ("Libération"), we found several cases where the distance between a (fronted) direct object and its governing verb is over 30 words.

[3]The reason why the link between the antecedent of a relative pronoun and the trace of the pronoun is presented here as a double link, although represented by the same index, lies with technical details of the theory, which are not directly relevant here (wh-chain vs predication).

[4]The use of normalized syntactic structures for the identification of idiomatic expressions is described in Wehrli (2000).

3 The FipsCo system

The primary goal of FipsCo is to extract multi-word terminology out of a text corpus. This includes lexicalized multi-word expressions (compounds, collocations and idiomatic expressions which happened to be listed in the lexical database used by the parser), as well as specific types of cooccurrences of lexical items (simple or complex). Thanks to the syntactic representation, it is no longer necessary to take into account any pair of reasonably closed lexical units, but rather we can focus on the truely relevant pairs, corresponding to the types given in (3). It is clear that in a sentence such as (4), the relevant pair is (*porte-parole,déclarer*) ("*spokesman-declare*"), which is a subject-verb pair, and not the pair (*gouvernement, déclarer*) ("*government, declare*")[5].

(3)a. noun - adjective
marée noire ("*oil slick*")

b. adjective - noun
haute technologie ("*high technology*")

c. noun - noun
thé citron ("*lemon tea*")

d. noun - prep - noun
part de marché ("*market share*")

e. noun - verb (subject-verb)
manne tomber ("*(godsent) money fall*")

f. verb - noun (verb direct object)
caresser espoir ("*entertain hope*")

g. verb - prep - noun (verb subcategorized PP)
vouer (à) échec ("*doom to failure*")

(4) Le porte-parole du gouvernement helvétique déclare que ...
"the spokesman of the Swiss government declares that..."

FipsCo first invokes the Fips parser to analyze the corpus sentence by sentence. For each sentence, the parser returns a normalized structure, which corresponds roughly to a GB-type enriched surface structure. Cooccurrences are extracted from these normalized structures in the canonical positions corresponding to the specific types described in (3), along with lexicalized MWEs (compounds, collocations and idiomatic expressions). Finally, for each occurrence of a MWE, and for each cooccurrence of lexical items, the position of the sentence (and of the lexical

[5]The following is a more complex example of the kind of problems that a syntactic parser can easily solve. In the following sentence, extracted from "Libération" *On trouve également des refuges intimes codifiant un nouvel ordre du plaisir* ("*one can also find intimate refuge codifying a new order of pleasure*"), the cooccurrence of *intimes* and *ordre* could be mistakenly taken as an occurrence of the very frequent collocation *intimer ordre* ("*to order*"), whereas in fact *intimes* in this sentence is not a verbal form but the plural form of the adjective *intime* ("*intimate*").

cooc. type	# item	avg	max	dev
noun-adj	27223	1.08	32	0.41
subject-verb	22896	4.46	38	4.13
verb-object	32747	3.18	47	2.31
verb-P-noun	14193	3.06	37	3.29
noun-P-noun	36414	2.71	31	1.00

Table 1: Statistics for 5 cooccurrence types: number of items, average distance, maximum of distance and standard deviation

items) is stored, as well as (in the latter case) the distance expressed in number of words between the two items. All this information is fed into a large database. At the end of the extraction process, the database contains all the lexicalized MWEs used in the text, as well as all the cooccurrences of the specified types. The fact that for each item the location has been stored makes it easy to display all the occurrences of a given collocation (or MWE) in the original text.

We have stressed so far the fact that the parser provides the necessary information to determine which pairs of lexical items are relevant for extraction. An additional advantage of using syntactically parsed sentences to extract collocations is the possibility of representing cooccurrences at a more abstract level than simple orthographic words. In our research, we are using the level of lexemes, i.e. a particular reading associated with simple words or of multiword expressions. For instance, when FipsCo extracts the verb-object collocation *réserver-droit* ("*to reserve the right*"), or the verb-prep-noun collocation *pencher-sur-question* ("*to look into an issue*"), the lexemes *réserver* in the first case and *pencher* in the second case are specifically the pronominal readings (*se réserver*, *se pencher*). Similarly, we get subject-verb collocations with verbal lexemes corresponding to idioms, such as *tribunal-donner raison à qqn* ("*court-prove someone right*") or *rencontre-avoir lieu* ("*meeting-take place*") or with compound, such as *donner-feu vert* ("*to give the green light*"), and not *rencontre-avoir* ("*meeting-have*") or *donner-feu* ("*give-fire*"), which would be irrelevant. The selection of the proper lexeme associated with an orthographic word or compound is an important aspect of the disambiguation process that a full-scale parser must carry out.

4 Preliminary results

Our first experiment with FipsCo concerned the extraction of coocurrences from an excerpt of articles of the French newspaper Libération. A total of 1 million of words were parsed and about 170'000 cooccurrences identified. The frequencies of words and lexemes were counted as well. The coocurrences were classified in eleven classes, the seven classes described in (3), above, and the three classes of lexicalized MWEs, compounds, collocations and idioms. While a more thorough analysis is still pending, the first observations are very encouraging. In table 1, we show the number of instances and inner distances by type of cooccurrences. That is, the total number of coocurrences for a given type (# item), the average distance (avg) in terms of words separating the two items of the coocurrence, the maximum of distance between the two terms (max), and the standard deviation of the distance (dev).

word 1	word 2	log	N	%LDC
jouer	rôle	299.34	38	5.26
avoir	impression	234.18	44	6.82
avoir	an	214.62	98	4.08
battre	record	141.92	14	28.57
avoir	droit	136.62	23	8.70
signer	accord	133.24	21	14.29
avoir	moyen	118.24	37	2.70
résoudre	problème	101.25	12	25.00
tourner	page	99.85	11	9.09
avoir	air	95.81	30	3.33
laisser	place	90.12	13	7.69
trouver	solution	88.23	15	13.33
avoir	intention	87.87	27	7.41
prendre	mesure	87.32	21	9.52
ouvrir	porte	82.88	12	8.33
prendre	décision	70.88	21	23.81
franchir	pas	69.75	11	9.09
conclure	accord	67.37	11	27.27

Table 2: First 18 cooccurrences ranked by log-likelihood ratio

The values for the categories subject-verb, verb-direct-object and verb-prep-noun clearly show that, in our corpus, the distance between the two elements of the cooccurrence is very frequently much larger than the average 3-5 word window commonly used in collocation extraction systems. The high value of the standard deviation for these 3 types confirms our intuition about the large variety of contexts in which such cooccurrences occur.

In order to better identify collocations among theses cooccurrences, a contingency table and the standard log-likelihood-ratio were used to classify our results. This technique avoids the emphasization of collocations with frequent words or words implied with frequent collocations. We also computed the percentage of "long distance cooccurrences" (henceforth LDC), i.e. cooccurrences for which the distance between the two words is greater than 5. Table 2 shows the first 18 items of the "verb-direct object" class.

The average percentage of long distance coccurrences among the 100 first collocations is 29.26%, which means that the part of information provided by LDC is important and justifies our approach. In this respect, it is interesting to note that occurrences of three items among the 100 first – *mener lutte* ("*to carry on fight*"), *estimer nombre* ("*to estimate number*") and *adopter proposition* ("*to adopt proposal*") – have a 100% LDC. In other words, the distance between the two terms is always greater than the average window used by standard collocation extraction systems, which means that such collocations would simply be overlooked by such systems.

Finally, consider the example (5), which illustrates a rather typical LDC collocation (*adopter amendement* "*to pass amendment*"). Two points are worth mentioning: first, the distance between the two terms is quite remarkable (25 words) (the average LDC for this collocation is near 50%). Second, the two terms of the collocation are not in the canonical order verb-object.

As the sentence is in the passive voice, the direct object argument has been raised to the subject position.

(5) Un **amendement** de la commission des lois prévoyant l'envoi, par le médiateur des enfants, d'un rapport annuel au médiateur de la République a été **adopté**.

5 Conclusion

Proper treatment of collocations is increasingly perceived as an important task for a variety of NLP applications. In this paper, we have described our current research to develop a collocation and MWEs extractor based on a syntactic parser. We have argued that the extraction of collocations out of syntactically parsed sentences provides significant advantages such as disambiguation and lemmatization of lexical units. Furthermore, we have argued that syntactic configurations constitute a much more appropriate criterion for the selection of cooccurrences than linear proximity.

6 References

Benson, M. "Collocations and general-purpose dictionaries" *Internation Journal of Lexicography* 3:1, 1990, 23-35.

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

Dunning, T.: "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics* 19:1, 1993, 61-74.

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

Laenzlinger, C. & Wehrli, E.: "Fips, un analyseur interactif pour le français", *TA informations* 32.2, 1991.

Manning, C. & Schütze, H.: *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press, 1999.

Smadja, F. "Retrieving collocations form text: X-tract", *Computational Linguistics* 19:1, 1993, 143-177.

Wehrli, E. *L'analyse syntaxique des langues naturelles : problèmes et méthodes*, Paris, Masson, 1997.

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