Multi-Word Collocation Extraction by Syntactic Composition of Collocation Bigrams

VIOLETA SERETAN, LUKA NERIMA & ERIC WEHRLI
Language Technology Laboratory, University of Geneva

Abstract

This paper presents a method of multi-word collocation extraction, which is based on the syntactic composition of two-word collocations previously identified in text. We describe a procedure of word linking that iteratively builds up longer expressions, which constitute multi-word collocation candidates. We then present several measures used for candidates ranking according to the collocational strength, and show the results of a trigram extraction experiment. The methodology used is particularly suited for the extraction of flexible collocations, which can undergo complex syntactical transformations such as passivization, relativization and dislocation.

1 Introduction

Collocations, defined as "arbitrary and recurrent word combinations" in (Benson 1990) or "institutionalized phrases" in (Sag et al., 2002), represent a subclass of multi-word expressions that are prevalent in language and play an important role in its naturalness. The *collocate*, i.e., the "right word" used in combination with a given word (usually called *base word*) is unpredictable for non-native speakers. The preference for a specific word instead of another is dictated by the conventional usage in a specific language, dialect, domain (or even a time period), rather than by syntactic or semantic criteria. A French speaker, for instance, needs to be aware of the conventional usage of an expression, e.g., "*encounter difficulties*", in order to avoid unnatural paraphrases such as *"*feel difficulties*". Collocations constitute a big concern for non-native speakers faced with the task of producing proficient text. In NLP, the collocational knowledge is indispensable for major applications such as the machine translation and the natural language generation.

The phenomenon of words collocability has been given particular attention since Firth (1957), who made the statement that a words is characterized by "the company it keeps". Two main approaches have been followed for the collocational knowledge acquisition: a lexicographic one, oriented towards the creation of dictionaries encoding words' combinatorial possibilities (notably (Benson et al. 1986, Mel'cuk et al. 1984)), and a statistical one, aimed at automatically extracting relevant word associations from text corpora (e.g., Choueka et al. 1983, Sinclair 1991, Church & Hanks 1990, Smadja 1993).

As stated by Harris (1988) in the "likelihood constraint" ("*each word has a particular and roughly stable likelihood of occurring as argument, or operator,*

with a given word"), the collocating words are syntactically bound. The collocation is, in fact, a well-formed expression. Unfortunately, the existing collocation extraction methods usually ignore the linguistic structure and rely almost completely on the text's surface, while it is generally agreed that the extraction should ideally be done from analyzed rather than from raw text (Smadja 1993:151).

More recent work (e.g., (Grishman & Sterling 1994, Lin 1998, Krenn & Evert 2001)) performs a shallow text analysis (POS tagging, lemmatization, syntactic dependency test) in order to syntactically filter the candidate expressions. Still, it is insufficient to account for the flexible collocations, in which the constituent words may appear inverted, arbitrarily distant from each other, and may not be directly related syntactically (as, for instance, the words "*overcome*" and "*difficulties*" in the expression "*the **difficulties** that the country tried to **overcome***").

Some major drawbacks of the classical methods are: the possible ungrammaticality of the candidates considered, the combinatorial explosion when considering all possible words combinations, the limitation (of the majority of stochastic tests) to two-word combinations. These drawbacks can only be overcome by performing a deep syntactic analysis that takes into account the complex grammatical processes underlying the text form.

The performance of extraction systems is essential for the subsequent treatment of collocations in important NLP applications such as machine translation, information retrieval, word sense disambiguation. Therefore we propose an approach to multi-word collocation extraction which focuses on the use of syntactic analysis and syntactic criteria for defining collocation candidates. Our approach is supported by the strong increase, over the last few years, of the availability of computational resources and software tools dedicated to large-scale and robust syntactic parsing[1].

The paper is organized as follows. Section 2 briefly presents some of the existing methods of multi-word collocation extraction and their main features. Section 3 outlines the method of collocation bigram extraction on which our work relies. In Section 4 we describe in detail the method we propose for extracting multi-word collocations using the collocation bigrams. In Section 5 we present the experimental results obtained by applying this method on a collection of English newspaper articles. The last section draws the conclusion and points out directions for further development.

2 Existing methods of multi-word collocation extraction

Traditional approaches to automatic collocation extraction from text corpora rely on stochastic measures, which range from the simple word co-occurrence frequency to more sophisticated statistical tests (e.g., log-likelihood ratios test

[1] See (Ballim & Pallotta 2002) for recent advances in robust parsing.

(Dunning 1993), Student's *t*-test, Pearson's χ^2 test[2]) or on information theoretic measures (e.g., the mutual information (Church & Hanks 1990)).

One feature these methods share is that they use the textual proximity as the main criterion for the selection of candidate collocations, instead of syntactic criteria. Since they consider any word combination as a valid candidate, they are forced to limit to a text window of fixed size (usually 5 words). Moreover, they usually do not take into account collocations made up of more than two words, as the lexical association measures are generally designed for pairs of items.

Only few methods, e.g., (Choueka et al. 1983, Smadja 1993), are also concerned with *n*-grams ($n>2$). The method proposed by Choueka et al. (1983) for finding *n*-word collocations considers the frequency of consecutive word sequences of length *n* (with *n* from 2 to 6). The limitation to $n=6$ is due to the rapid increase of the number of all possible *n*-grams, for *n* bigger than 6.

The Xtract system of Smadja (1993) retrieves, in a first stage, word bigrams that are not necessarily contiguous in text, but can be separated by several words. It then looks, in the second stage, at the words in the bigrams' surrounding positions and identifies *n*-grams as the repetitive contexts of already identified bigrams. These repetitive contexts can form either "rigid noun phrases", or "phrasal templates" (phrases containing empty slots standing for parts of speech).

Both methods rely only on a superficial text representation, while the authors point out that the selection of terms should ideally be done following linguistic criteria.

Since robust large-scale parsers became in the meantime available, the more recent methods focus on using parsed rather than raw text for bigram extraction (e.g., Lin 1998, Goldman et al. 2001). Our work relies to a large extent on the features of the method of (Goldman et al. 2001), which we will briefly present in the next section.

3 Collocation bigrams extraction with FipsCo

FipsCo (Goldman et al. 2001) is a term extractor system that relies on Fips (Laenzlinger & Wehrli 1991), a robust, large-scale parser based on an adaptation of Chomsky's "Principles and Parameters" theory. The system extracts, from the parsed text, all the co-occurrences of words in given syntactic configurations: noun-adjective, adjective-noun, noun-noun, noun-preposition-noun, subject-verb, verb-object, verb-preposition, verb-preposition-argument. It thus apply a strong syntactic filter on the candidate bigrams. Subsequently, it applies the log-likelihood test (Dunning 1993) on the sets of bigrams obtained, in order to rank them according to the degree of collocability.

[2] For a rather comprehensive overview see chapter 5 of (Manning & Schütze 1999).

The strength of this approach comes from the combination of the deep syntactic analysis with the statistical test. The sentences are normalized (the words are considered in their lemmatized form and in their canonical position). The system is able to handle complex cases of extraposition, such as relativization, passivization, raising, dislocation.

To illustrate this, let us consider the following sentence fragment (a real corpus example we have encountered): "*the **difficulties** that the country tried to **overcome***". Extracting the collocation of verb-object type "*overcome — difficulty*" requires a complex syntactic analysis, made up from several steps: recognizing the presence of a relative clause; identifying the antecedent ("*the difficulties*") of the relative pronoun "*that*"; and establishing the verb-object link between this pronoun and the verb of the relative clause. This collocation will simply be overlooked by classical statistical methods, where usually the size of the collocational window is 5. Such situations are quite frequent for example in Romance languages[3], in which the words can undergo complex syntactic transformations.

4 Multi-word collocation extraction by bigrams composition

The system presented above is able to extract syntactically bound collocation bigrams, which may occur unrestrictedly with respect to the textual realization. The system relies on a deep syntactic analysis that can handle complex cases of extraposition. We will take advantage of these features for the multi-word collocations identification, since they guarantee both the grammaticality of results, and the unconstrained search space and realization form.

Since the FipsCo system actually returns not only the best scored collocations, but all the candidate bigrams, we generate all the possible multi-word associations from text. Our goal is to build up, using the set of extracted bigrams, the sequences of bigrams sharing common words. The obtained collocate chains represent well-formed multi-word associations. The configuration of their syntactic structure is defined by the syntactic relations in the bigrams involved. The shared term must be the same not only lexically, but also indexically (the very same occurrence in the text). Due to the syntactic constraint, the shared term will actually appear in the same sentence as the other collocates.

For instance, given two bigrams (w_1 w_2), (w_2 w_3) with, we can construct the trigram: (w_1 w_2 w_3), as in the case of the following collocations: "*terrorist attack*", "*attack of September*"; we obtain the trigram collocation "*terrorist attack of September*".

[3] Goldman et al. (2001) report a high percentage of cases in which the distance base-collocate is more than 5 words in a French corpus.

Note that the condition of indexical identity avoids combinations with different readings in case of polysemy, e.g., "*terrorist attack*" with "*attack of coughing*".

Repeating the same procedure we can add further words to the obtained trigrams, thus obtaining multi-words collocations of arbitrary length. Moving on to n-grams will conserve the inclusion of all terms in the same sentence. We impose no default restrictions on the syntactic configuration of the resulting expression.

In what follows, we present the word linking procedure that allows the construction of longer multi-word collocations (henceforth MWCs) from shorter collocations and the measures proposed for ranking them.

4.1 *Iterative word linking procedure*

The procedure of linking new words to existing collocations in order to discover longer collocations makes use of the criterion of the existence of a syntactic link between the new words and one of the existing collocation's words. Recursively applied to the set of generated collocations in each step, this procedure allows the incremental composition of longer collocation from shorter subparts. In thus leads to the identification of all collocation candidates in a text; the distance between the composing words is only limited by the sentence's boundaries.

Building up all the possible trigrams from a set of bigrams can be done, for example, by considering all the pairs of bigrams that share terms. We call "pivot" the term shared by two bigrams. There are three possibilities to construct a trigram, that correspond to the position of the pivot in the two bigrams. In the first case, the pivot is the last term in one bigram, and the first in the another. It occupies the middle (internal) position in the new trigram, as in "*terrorist **attack** of September*". In the other two cases, the pivot occupies an external position, either on the left (as in "***have** impact on*", derived from the bigrams "***have** impact*" and "***have** on*"), or in the right (as in "*round [of] presidential **election***", derived from "*round [of] **election***" and "*presidential **election***").

For the general case, the following procedure is used to incrementally build up longer n-grams:

$\mathcal{C} := \mathcal{D}$;
repeat
  $\mathcal{N} := \emptyset$;
  for each MWC_i **in** $\mathcal{C}$
    for each MWC_j **in** $\mathcal{C}$, i$\neq$ j
      if combine(i, j) **then**
        add($\mathcal{N}$, combination(i, j));
        remove($\mathcal{D}$, MWC_i);
        remove($\mathcal{D}$, MWC_j);
  $\mathcal{C} := \mathcal{N}$;
  $\mathcal{D} := \mathcal{D} \bigcup \mathcal{C}$;
until $\mathcal{C} = \emptyset$;

where $\mathcal{D}$ is the set that will contain the results, initially containing all the bigrams; $\mathcal{C}$ - a temporary set currently used in an iteration; and $\mathcal{N}$ - the newly generated n-grams in the current iteration.

The following criterion is considered for combining two multi-word collocations (MWCs) into a larger one: two MWCs can combine if they have at least one term that is different and one that is shared (i.e., that has the same position in text). In the procedure above, the predicate *combine(i, j)* checks whether this criterion is satisfied by the expressions MWC_i and MWC_j, and *combination(i, j)* is the resulting MWC (obtained by merging the terms involved).

At each step, the procedure tries all the possible combinations among already generated MWCs using the above stated composition criterion. It adds the new combinations to the results set $\mathcal{D}$, from which it then eliminates the participating (subsumed) MWCs.

The process is repeated as long as new MWCs can be constructed from the MWCs generated in the previous step. After a finite number of iterations, the procedure terminates since the set of new expressions that can be generated is finite (it is localized within a sentence). It is easy to check that the time complexity of the algorithm is polynomial in the size of the initial bigrams set.

4.2 *Association measures*

The MWCs extracted with the algorithm described above represent all the syntactically bound co-occurrences of terms in the corpus. In order to identify the good collocation candidates among them we proposed 4 methods for quantifying their degree of collocability.

The first method simply computes the MWCs frequency in the corpus. The second uses the log-likelihood values initially assigned to bigrams and considers the sum of participating bigrams' score as a global score for a MWC. The third method tries to find MWCs whose global score is balanced and considers the harmonic mean as an association measure.

Finally, as a fourth method, we apply the log-likelihood test, the same test that FipsCo applied to words in order to score collocation bigrams. We instead apply it to bigrams, in order to score the trigrams build from these bigrams. The contingency table (which is used to compute the log-likelihood values) contains the joint and marginal frequencies for each two bigrams, i.e., the corpus frequency of the two bigrams together (as a trigram), and respectively the corpus frequency of each of the two bigrams.

In order to apply this measure to arbitrarily long MWCs, we can apply it recursively to the sub-MWCs composing a given MWC. Let MWC_1 and MWC_2 be two MWCs that compose a larger MWC (as described in 4.1). The log-likelihood score is computed using a contingency table for the pair (MWC_1, MWC_2), listing co-occurrence frequencies related to each of the two sub-expressions.

5 The experiment. Results and discussion

We applied the method of identifying multi-word collocations as presented in the previous section on a corpus of 948 English articles from the magazine "The Economist". The collection of texts, amounting to about 870,000 words, was first parsed and about 142,000 syntactically bound bigrams were extracted with FipsCo (no frequency filter was applied). About 7.00% of the extracted bigrams involved more than two words[4].

We then extracted trigrams using the word linking method presented in subsection 4.1. We obtained a number of 54,888 trigrams, divided in 13,990, 27,121, and 13,777 for each pivot position case (i.e., left, middle, and right respectively). Table 1 shows the 10 most frequent trigrams in the whole set, and the top 10 trigrams according to the log-likelihood measure.

trigram	**freq**	**trigram**	**log**
weapon of mass destruction	38	weapon of mass destruction	579.03
have impact on	17	have impact on	214.35
go out of	15	move from to	126.10
pull out of	14	turn blind eye	124.01
make difference to	11	rise from in	120.57
rise in to	10	play role in	110.07
move from to	10	make difference to	109.46
rise from in	10	rise in to	105.43
play role in	9	second world war	105.42
be to in	8	rise from to	99.08

Table 1: *Top 10 trigrams according to frequency and log-likelihood*

We consider that both the frequency and the log-likelihood measures are appropriate for scoring collocations, with the log-likelihood slightly more precise.

The measure based on the sum yields uninteresting results, as an expression may obtain a good score if it happen to contain a top scored bigram (as "*prime minister*"), even if it is not a collocation (e.g., "*prime minister promise*").

The measure based on the harmonic mean allows for the identification of good multi-word collocations, like "weapons of mass destruction" that received the best score. Still, we judge its results less satisfactory than those obtained with the log-likelihood measure.

However, to estimate the efficiency of these measures a solid evaluation should be performed against a gold standard, possibly by adopting a n-best strategy, as in (Krenn & Evert 2001).

[4] FipsCo is able to extract some multi-word collocations as bigrams involving a compound, idiom or collocation already present in the lexicon.

We were interested in the syntactic configurations of the multi-word collocations obtained, as they could suggest syntactic patterns to use for the extraction of multi-word collocations directly from parsed text. The most frequent association types found in the corpus are listed in Table 2, together with an example for each type[5].

rel1	**rel2**	**frequency**	**example**
Noun-Prep-Noun	Adjective-Noun	5607	round of presidential election
Verb-Object	Verb-Prep	5364	have impact on
Subject-Verb	Verb-Prep	4904	share fall by
Subject-Verb	Verb-Object	4659	budget face shortfall
Verb-Object	Adjective-Noun	4622	turn blind eye
Adjective-Noun	Subject-Verb	3834	main reason be
Verb-Prep	Verb-Prep	3232	move from to
Verb-Object	Compound	2366	declare state of emergency
Verb-Object	Subject-Verb	1693	want thing be
Noun-Noun	Noun-Prep-Noun	1627	world standard of prosperity

Table 2: *The 10 most frequent association types for trigrams*

As mentioned earlier, during the extraction no predefined syntactic patterns were used (we imposed no restriction on the configuration of newly built expressions). Nevertheless, the trigram patterns which are discovered are dependent on the bigram patterns used by FipsCo extraction system, therefore their coverage depend on how exhaustive the initial patterns are.

6 Conclusions and future work

We have presented a method for the multi-word collocation extraction that relies on the previous extraction of collocation bigrams from text, and is based on iteratively associating already constructed collocations using a syntactic criterion. We have used several measures for estimating the strength of the association. In particular, we applied the log-likelihood ratio statistical test (initially used for word bigrams) to the extracted multi-word collocations. This test appears to be, together with the frequency, the relatively best measure for evaluating the collocational strength.

The methodology used is based on a hybrid (linguistic and statistical) approach aimed at improving the coverage and the precision of multi-word collocation extraction. Unlike purely statistical approaches, the method presented can handle collocations whose terms occur in text at a long distance from each other

[5] The frequency counts refer to the distinct collocations extracted, and do not take into account how many instances a given collocation may have in the corpus.

because of to the various syntactic transformations the collocations can undergo. At the same time, the results are grammatical, due to the syntactically based filter of candidates and to the syntactic nature of the criterion used for the composition of longer multi-word collocations.

Another important advantage over the multi-word collocation extraction methods ignoring the text syntactic structure is that there is no limitation on the length of candidates that can be build. Classical methods (Choueka et al. 1983) were forced to limit to 6-word collocations, and even more recent methods that do not apply a syntactic filter recognize the same limit (Dias 2003).

Further developments of the method include finding criteria for the delimitation of *n*-grams within the sentence, that is, for settling a limit between subsumed and subsuming collocations.

The method will be integrated into a concordance and alignment system (Nerima et al. 2003), which will allow the visualization of extracted multi-word collocations in the source text, and in the parallel (translated) text, when available.

Acknowledgements. This work was carried out in the framework of the research project "Linguistic Analysis and Collocation Extraction" adopted by the Geneva International Academic Network (GIAN) for 2002-2003.

REFERENCES

Ballim, Afzal & Vincenzo Pallotta, eds. 2002. *Robust Methods in Analysis of Natural Language Data* (= Special Issue of *Natural Language Engineering*, 8:2/3). Cambridge: Cambridge University Press.

Benson, Morton, Evelyn Benson & Robert Ilson. 1986. *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.

Benson, Morton. 1990. "Collocations and General-Purpose Dictionaries". *International Journal of Lexicography* 3:1.23-35.

Choueka, Yaacov, S. T. Klein & E. Neuwitz. 1983. "Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus". *Journal of the Association for Literary and Linguistic Computing* 4:1.34-38.

Church, Kenneth W. & Patrick Hanks. 1990. "Word Association Norms, Mutual Information and Lexicography". *Computational Linguistics* 16:1.22-29.

Dial, Gaël. 2003. "Multiword Unit Hybrid Extraction". *Proceedings of the Workshop on Multiword Expressions at the 41th Annual Meeting of the Association for Computational Linguistics* (*ACL'03*), 41-48. Sapporo, Japan.

Dunning, Ted. 1993. "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19:1.61-74.

Evert, Stefan & Brigitte Krenn. 2001. "Methods for the Qualitative Evaluation of Lexical Association Measures". *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188-195. Toulouse, France.

Firth, John R. 1957. "Modes of meaning". *Papers in Linguistics* ed. by J. R. Firth, 190-215. Oxford: Oxford University Press.

Goldman, Jean-Philippe, Luka Nerima & Eric Wehrli. 2001. "Collocation Extraction Using a Syntactic Parser". *Proceedings of the Workshop on Collocation at the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL'01*), 61-66. Toulouse, France.

Grishman, Ralph & John Sterling. 1994. "Generalizing Automatically Generated Selectional Patterns". *Proceedings of the 15th International Conference on Computational Linguistics* (*COLING'94*), 742-747. Kyoto, Japan.

Harris, Zelig S. 1988. *Language and Information*. New York: Columbia University Press.

Laenzlinger, Christopher & Eric Wehrli. 1991. "Fips, un analyseur interactif pour le français". *TA Informations* 32:2.35-49.

Lin, Dekang. 1998. "Extracting Collocations from Text Corpora". *First Workshop on Computational Terminology*, 57-63. Montreal, Canada.

Manning, Christopher & Heinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Mel'cuk, Igor A. et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I, II, III, IV*. Montreal: Presses de l'Université de Montréal.

Nerima, Luka, Violeta Seretan & Eric Wehrli. 2003. "Creating a Multilingual Collocation Dictionary from Large Text Corpora". *Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL'03*), 131-134. Budapest.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. "Multiword Expressions: A Pain in the Neck for NLP". *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (*CICLING 2002*), 1-15. Mexico City.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smadja, Frank. 1993. "Retrieving Collocations form Text: Xtract". *Computational Linguistics* 19:1.143-177.