Using conceptual vectors to get Magn collocations (and using contrastive properties to get their translations)

Vincent Archer

MTT 2007, Klagenfurt, May 21 – 24, 2007
Wiener Slawistischer Almanach, Sonderband 69, 2007

Using Conceptual Vectors to get Magn Collocations (and using contrastive properties to get their translations)

Vincent Archer

Laboratoire d'Informatique de Grenoble – Université Joseph Fourier
385, rue de la Bibliothèque - B.P. 53 - 38041 Grenoble Cedex 9
vincent.archer@imag.fr

Abstract

This paper presents a semi-automatic approach for extraction of collocations from corpora which uses the results of Conceptual Vectors as a semantic filter. First, this method estimates the ability of each co-occurrence to be a collocation, using a statistical measure based on the fact that it occurs more often than by chance. Then the results are automatically filtered (with conceptual vectors) to retain only one given semantic kind of collocations. Finally we perform a new filtering based on manually entered data. Our evaluation on monolingual and bilingual experiments shows the interest to combine automatic extraction and manual intervention to extract collocations (to fill multilingual lexical databases). It proves especially that the use of conceptual vectors to filter the candidates allows us to increase the precision noticeably.

Keywords

Collocations, semantic filter, conceptual vectors, semi-automatic, contrastive extraction

1 Introduction

Natural language processing needs linguistic knowledge, especially in machine translation: current systems have bad results because of parsing errors and collocations. In fact, some expressions can not be translated word-for-word because the meaning of a whole expression is not necessarily the combination of the meanings of its components. This problem could be easily solved for recognized locutions by considering them as an unique lexical object, but it is really more difficult for collocations (expressions where one term is chosen in function of the other one, like *driving rain* for the intensification of *rain*). It is more difficult to know and recognize collocations than idioms, because they are more numerous and their meaning is not fully independent from the components. In the Meaning-Text Theory, the lexical functions (Mel'čuk et al., 1995) provide a good representation of the collocations, for example Magn(*rain*)={*heavy, driving*} for the intensification. If a translation system knows that *heavy rain* is an intensification of *rain*, and that, in French, the intensification of *pluie* (translation of rain) is *pluie battante*, it will be able to translate this expression correctly.

How to build a database containing those informations ? We can neither do it manually (it will last very long) nor proceed automatically (we need precision). The idea is to allow man and computer to combine their abilities to fill the base. There are several tracks: machine

learning, interaction with non-specialists, etc. In this paper, we will present the use of linguistic knowledge (conceptual vectors) to filter the results of an extraction of collocations.

2 Modeling – the Lexical Functions

Before starting extraction, we have to define what we exactly consider as a collocation: different researchers who worked on collocations covered different notions. (Sinclair, 1970) defined it as "*the occurrence of two items in a context within a specified environment. Significant collocation is a regular collocation between two items, such as they co-occur more often than their respective frequencies*", with no remarks on the dependency between the items. Here we use the definition given in (Kahane & Polguère. 2001): a collocation is "*a linguistic expression made up of at least two components: 1. the* base *of the collocation: a full lexical unit which is "freely" chosen by the speaker; 2. the* collocate*: a lexical unit or a multilexical expression which is chosen in a (partially) arbitrary way to express a given meaning and/or a grammatical structure contingent upon the choice of the base*".

There are many models of the lexicon, and some have been implemented to create lexical database. Some contain informations about co-occurrence, like the co-occurrence dictionary in EDR, the qualia structure in the Generative Lexicon (Pustejovsky, 1998) or troponyms in Wordnet, which may sometimes be collocations, but there is no manner to distinguish a collocation from another co-occurrence. The only representation of the lexicon that really aims to model collocations is the *Lexical Functions* (LF), a part of Mel'čuk's Meaning-Text Theory: a given lexical function links a lexical unit with a set of lexical units which have a particular relation with it. These relations can be paradigmatic (like synonymy, derivation, etc.) or syntagmatic (combinatory links, collocations). Moreover, this theory (Mel'čuk et al., 1995) has been implemented by the realization of explanatory and combinatory dictionaries for French (DEC, on paper), and in automatic database (DiCo[1], a simplification of the DEC). Our research is related to the *Papillon* project. Its aim is to build a multilingual lexical database (Mangeot et al., 2003) which may be consulted at http://www.papillon-dictionary.org, and be edited in a collaborative way. This database can be viewed as a dictionary made of several volumes: one for each language, and one for the interlingual pivot structure. The macro-structure links the entries from the volumes using an abstract pivot made up of interlingual acceptions, or *axies* (Sérasset, 1994), modeling differences of semantic refinement[2]. The structure of the lexical units is similar to the one used in DiCo, using Lexical Functions.

3 Conceptual Vectors

In the theory of Conceptual Vectors, there is a finite set of concepts that could be used to generate the terms of the whole language: each meaning of the language could be considered as a linear combination of those concepts (Schwab et al., 2002). Conceptual vectors represents the language terms, and the dimensions of these vectors are the basic concepts of the language. Using a mathematical representation of meaning like vectors, allows to consider the distance of meaning between terms as the angular distance of the correspondent vectors. An

[1] available online with the interface DicOuèbe at http://olst.ling.umontreal.ca/dicouebe

[2] For instance, *river* can be translated in French by 2 non-synonyms words: *rivière* (flows into a river) and *fleuve* (flows into the sea) ; so the axie for *river* is refined into 2 axies, one for *rivière*, the other one for *fleuve*

experimental implementation of this theory on French is made at the LIRMM (Montpellier, France), where 873 concepts are identified, like *vie*, *mort*, *recherche*, *fin* (respectively life, death, research, end), etc., using a French thesaurus (Larousse, 1992). It uses existing data to refine vectors: new vectors are computed from definitions from different sources (dictionaries, synonym lists, manual indexations, etc.) which are parsed, and from existing vectors. It needs a bootstrap: you must have a kernel made of pre-computed vectors (generally manually indexed) to begin the process.

3.1 Using Conceptual Vectors to filter our results

A collocation can be viewed as made of three components: *base*, *collocate*, and *meaning*. After the extraction of collocations from a corpus, there are lots of candidates with supposed base and collocate but no meaning. The use of existing lexical resources, like (Pearce, 2001) made with Wordnet, allows to improve the quality of extraction: those resources contain semantic informations that can be really interesting in such tasks. Here we need a semantic filter to consider only collocates that express intensification: we want to get a class of such collocates. We can use *Conceptual Vectors* to get it: we believe that the set of nearest conceptual vectors (according to angular distance) from the concept *intensité* (intensity) could be this wanted class of intensifiers. We assume the fact that this filter would find the more transparent collocations (where the collocate has always a meaning close to intensification) but will miss the less decodable collocations: we will increase precision but decrease recall. But even decodable collocations are of great interest because of their unpredictability, a great problem in machine translation: the meaning is generally insufficient to generate the whole collocation from the base ; for instance, *gravement* and *grièvement* are French synonyms, you can generate *gravement malade* (seriously ill) and *grièvement blessé* (seriously hurt), but not **grièvement malade*. So, even if we get decodable collocations, it will be useful to know that we should use one particular collocate and not another one to express intensification.

4 Collocation acquisition

In our research, our aim is to get intensification collocations. (Claveau & L'Homme., 2006) showed the interest of inferring rules from the contexts of known collocations to find other collocations. (Wanner et al., 2006) also used machine learning techniques to label collocations with semantic tags. We do not have a learning base to perform such a task. Furthermore, we want to propose a method that could be easily implemented. That's why we propose an acquisition of collocations by extraction. The *X-tract* system (Smadja, 1993), even if it did not aim to extract the same things as we do (Smadja considers that ``*a collocation is an arbitrary and recurrent word combination*", there is nothing about the fact that the use of a term depends on the other term), showed the interest of using an hybrid method that combines a linguistic (syntactic) analysis and a statistical filter. That's why our approach is also hybrid.

4.1 Syntactic and semantic aspects: Contexts, Conceptual Vectors

The co-occurrence of two terms in the same phrase is not sufficient to recognize a collocation: base and collocate must have a particular relation (like modification). So we use syntactic analysis, considering different kinds of context: the first one is the fact that a term is immediately followed by an other term (*linear context*) ; the second one is the fact that an analyzer says that there is a relation of modification between the two terms (*dependency*

context): (Lin, 1998) obtained quite good results using that kind of context. We also use a stoplist to eliminate stative verbs which make noise because they can never be intensified.

As we explained before, we decided to use conceptual vectors to filter our candidates, in order to get *Magn* collocations. We download at http://www.lirmm.fr/~lafourcade the 500 nearest words from *c4.intensité* (according to the angular distance), this set will be used to filter our results: we will only keep the co-occurrences of which the supposed collocate is part of the set. (Léon & Millot, 2005) acquire bilingual lexical relations using a simple manual validation of English lexical relations to increase the precision of their final results from 7,5% to 83,3%: it shows that it is really interesting to have a human intervention to complete automatic extraction. Our method is based on the same idea: we want to apply a simple manual filter (and so use human knowledge about language) to our automatic acquisition (based on the co-occurrence of terms in corpora) and to our automatic filter (conceptual vectors).

4.2 Collocability

An essential property of collocations is that these co-occurrences are more frequent than by chance. The *mutual information* $MI(x,y)=\log(P(x,y)/[P(x).P(y)])$ [3] seems to be convenient to model that. (Lin, 1998) proposed an adaptation of this measure to triples (2 terms and 1 relation): $MI(w\ r,w\ )=\log[P(w\ r,w\ )/P(r).P(w\ |r).P(w\ |r)]$ [4]. However, *mutual information* has a drawback for its use in NLP tasks: it tends to overestimate the association between two words with low frequencies. That's why (Fung & McKeown, 1997) introduced the *weighted mutual* information, with ponderation: $wMI(w\ w\ )=P(w\ w\ ).\log[P(w\ w\ )/P(w\ ).P(w\ )]$. (Wu & Zhou, 2003) proposed the adaptation of this last measure to triples (with the relation) which we use in our approach:

$$WMI(w\ r,w\ )=P(w\ r,w\ ).\log\frac{P(w\ r,w\ )}{P(w\ |r).P(w\ |r).P(r)}$$

4.3 Bi-collability

As multilingual information on collocations is very useful for translation systems, we are also interested to extract bi-collocations (two collocations which are translations of each other): we shall use bilingual corpora to get such an information. As the choice of the collocate depends on the base, it is frequent that the collocate chosen for the translation of the base is not the translation of the collocate chosen for the base. As a bi-collocation is not really useful when collocates are translations (the translation by MT systems would be correct), we want to extract *contrastive bi-collocations*, where the bases are translations but the collocates are not necessarily translations. We want to express the fact that a bi-collocation is a couple of collocations which often appears in similar (comparable, aligned) documents ; we adapt the cos measure $\cos(x,y)=|X\cap Y|/\sqrt{|X|.|Y|}$ (where *X* and *Y* are the documents where *x* and *y* occur) to similar distinct sets: $\cos_{bilingual}(c_{fr},c_{en})=|BI-DOCS(c_{fr},c_{en})|/\sqrt{|C_{FR}|.|C_{EN}|}$ [5]. It is insufficient to compute the association between the two parts of the bi-collocation candidates:

---

[3] where *P(w)* is the probability of the occurrence of *w*, *P(x, y)* the co-occurrence of the terms in a given context

[4] where $P(w_1|r)$ and $P(w_2|r)$ are the respective probabilities of the occurrence of w_1 as the first element of a relation *r*, and the occurrence of w_2 as the second element of a relation *r*, and $P(w_1, r, w_2)$ the probability of co-occurrence of w_1 and w_2 to be in relation *r*.

the final measure must also model that these two parts are collocations. So we use a ponderation which should be maximized when the collocability of each monolingual property is high and minimized when it is low. Our final measure to rank bi-collocations candidates is: $Bicollocability(c_{fr}, c_{en}) = (WMI(c_{fr}) + WMI(c_{en})) \times \cos_{bilingual}(c_{fr}, c_{en})$.

5 Evaluation

5.1 Experiments

We choose to extract candidates for Magn collocations with a verbal base and a adverbial collocate. We conduct three different experiments in order to evaluate the effectiveness of filtering using conceptual vectors. The first one is *monolingual* (acquisition of Magn collocations for French) ; the other ones are *bilingual* (acquisition of French-English *Magn* bi-collocations - couples of Magn collocations which are translations): one task is made using *comparable corpora*, the other one is made using *parallel corpora*.

	Experiments	Documents	Sentences	Words
LeMonde95 (FR)	Monolingual+Comparable Bilingual	47 646	1 016 876	24 730 579
GH95 (EN)	Comparable Bilingual	56 472	1 321 323	28 122 780
Europarl-FR	Parallel Bilingual	495	1 089 670	31 115 677
Europarl-EN	Parallel Bilingual	491	1 064 462	25 089 232

Table 1: Characteristics of corpora

The monolingual task uses the corpus LeMonde95 ; the first bilingual task uses LeMonde95 and GH95 (two newspapers corpora) as comparable corpora ; finally we use French and English parts of *EuroParl* (proceedings of the debates at the European Parliament) aligned by sentences. We supposed that GH95 and LeMonde95 were comparable, but we did not have any correspondance at the level of documents, so we computed a comparability measure between French and English documents. The criteria we used to determine if documents speak about the same topic were: the proximity in time (less than 2 days between the publications of the articles), the same named entities in the two documents, and the fact that nominal syntagms (which express the thema of a document) are translated. That's why we compute a very simple "comparability measure" between every potential couple of documents $comp(D_{fr}, D_{en}) = (overlap[NS(D_{fr}), NS(D_{en})] + overlap[trans_{en}(NS(D_{fr})), NS(D_{en})])/$ [6]
(we use the overlap measure to allow a short document and a long one to have a great

[5] C_{FR} and C_{EN} are the sets of documents where c_{fr} and c_{en}, appear ; BI-DOCS(c_{fr},c_{en}) is the set of bi-documents (comparable or aligned documents) where c_{fr} appears in the French document and c_{en} in the English document

[6] Where *Dfr* and *Den* are French and English documents, *NS*(*D*) is the set of nominal syntagms in document D, and *trans*(*NS*(*D*)) is the set of the translations of the nominal syntagms in document D

comparability value if their topics are similar). Using a minimal threshold of 0.2, we obtained 63 621 associations (1.34 per French document, 1.13 per English document).

5.2 Evaluation method

We compute *precision* for each experiment. We can not compute *recall* because we do not have a reference base at our disposal: there is no standard evaluation measure for collocation extraction. Moreover, we do not aim to extract the same things as other researchers, because we consider a different definition of collocation: (Smadja, 1993) tried to extract all recurrent co-occurrences, (Lin, 1998) aimed to obtain all "habitual word combinations", etc., whereas we consider collocations like co-occurrences with particular linguistic properties. In addition there is one more difficulty to present an objective comparison with existing works: we tried to extract collocations that express one particular meaning. For each monolingual experiment, we evaluate the 1000 first produced couple candidates, ranked by their WMI value. For the "comparable" bilingual experiment we evaluate the 200 first candidates ; for the "parallel" bilingual one we evaluate the 43 candidates we get using Conceptual Vectors (plus the 200 first candidates without using Conceptual Vectors).

5.2.1 Monolingual

Filtering	No	Conceptual Vectors	Conceptual Vectors	Conc. Vectors + Manual
Context	Dependency	Dependency	Linear	Linear
Precision	17%	41%	44%	83%

Table 2: Evaluation of the monolingual experiments (top 1000 candidates)

The first experiment was a statistical extraction with no filtering, so we got low precision (17%). Using an automatically produced list of adverbs to filter the results, the precision is multiplied by 2,5 (41% or 44%, depending the context). Moreover, we can already retrieve more collocations in top candidates, like *régner sans partage* or *réduire considérablement* (intensifications of *rule* and *reduce*). But we still have candidates in which the adverb never expresses intensification. We can observe that the kind of context seems not to be determining: informations on dependency do not allow to increase precision, we even obtain slightly better results with linear context, because dependency analysis retrieves more adverbs far from the verbs (we increase recall) but is more sensible to noise (we decrease precision). At last, a simple operation, the introduction of a new filter on adverbs (manually defined from the results of the precedent experiment: we remove adverbs like *trop* (too much), *très* (very), *tant* (so much), etc.: the last ones can express intensification but are so frequent that they are not interesting) allows us to eliminate 47% of candidates: then the precision increases from 44% to 83%. This shows the effectiveness of manual intervention on collocation extraction. Even with manual filtering, we get 17% of noise in the results because some adverbs may be intensifiers with a given verb and express another meaning with another one. The filtering allows a gain in precision but implies a loss of recall ; we loose in this case the less decodable collocations like *défendre bec et ongles* or *reprendre de plus belle* (intensifications of *defend* and *resume*): as the locutions do not express clearly the notion of intensification by

themselves, they are unfortunately not enough near to the concept of intensification (in the conceptual vectors) to be kept here.

5.2.2 Bilingual

Experiment on comparable corpora	Experiment on parallel corpora
mettre beaucoup / take seriously jouer pleinement / work hard vouloir particulièrement / want really accomplir particulièrement / perform strongly regretter énormément / regret deeply	soutenir pleinement / support wholy jouer pleinement / work together changer radicalement / change radically modifier radicalement / modify radically jouer pleinement / play right

Table 3: Top 5 candidates for bicollocations, in the two bilingual experiments

Is the English couple a Magn collocation ?	Yes		No
Are the 2 couples in translation ?	Yes	No	Yes
1-100 / 101-200	17% / 5%	23% / 18%	60% / 77%

Table 4: Evaluation of the bilingual comparable experiment (top 200 candidates)

The precision of produced candidates is higher in the experiment using parallel corpora, and this is not surprising: the chance to get translations of co-occurences is higher using aligned documents than using comparable ones. In the comparable experiment (and in a lesser extent in the parallel one), we can have non-collocations candidates where their two components are really collocations ; it could be a problem of polysemy (the intensified verb is used in different acceptions) or the fact that the intensification is not made on the same argument of the predicate. At this point, we should comment the number of candidates produced: in the first experiment (comparable corpora), we get 80 298 candidates before any filter, 3 973 after applying Conceptual Vectors, and 201 after manual intervention. In the second one (parallel corpora), we get 15 583 candidates before any filter, 1995 after applying Conceptual Vectors, and 43 after manual interventions. This last number is very low and insufficient. It seems not the best way to get bi-collocations: it is interesting to filter monolingually to reduce the number of candidates to several thousands, but the bilinguality combines English and French filters and we finally obtain to few collocations. If we want to get more collocations, we have to know that even big bilingual corpora contain few bi-collocations.

Which collocations are correct ?	French+English	French	English	No
Translation	19 / 36,5%	0 / 1%	1 / 1,5%	0 / 13,5%
No translation	2 / 2,5%	5 / 21%	2 / 12%	10 / 12%

Tableau 5: Evaluation of the bilingual parallel experiment

6 Conclusion and Future Work

In this paper, we have described a semi-automatic method of collocation extraction that uses Conceptual Vectors to produce a semantic filter (which is then refined manually). We proved that a human intervention on such a process is necessary to obtain high-quality results, and that the results of conceptual vectors are a good semantic filter for the extraction of one particular kind of collocation, especially when they are completed by a manual intervention. We also showed that this method is more efficient in the monolingual case because it is much harder to find bi-collocations than collocations in corpora (so the recall is much lower). We will make experiments to find the best corpus size to extract collocations. Our current objective is to realize programs that could be easily used by people who are not computer scientists (especially by linguists) to produce candidates for collocations from the corpora they have at their disposal, allowing them to guide the process manually. We also want to explore other ways to get collocations, like machine learning (learn the characteristics of collocations, and retrieve co-occurrences with these characteristics), expansion of results with thesaurus (by instance, retrieve *driving snow* from *driving rain*). Another track is to interact with non-specialists (every native speaker of English can say that a driving rain is an intense rain) using games: questions with known answers allow to evaluate the player, and we keep the answers from good players for the other ones ; we can also have a two-player party (it is a good proof of collocability when two different players give the same answer).

Bibliography

Claveau, V. & M.-C. L'Homme. 2006. Discovering and Organizing Noun-Verb Collocations in a Specialized Corpora Using Inductive Logic Programming, *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, 11(2), 209–243.

Fung, P. & K. McKeown. 1997. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1-2), 53–87.

Kahane, S. & A. Polguère. 2001. Formal foundation of lexical functions. *Proceedings of COLLOCATION: Computational Extraction, Analysis and Exploitation*, Toulouse, 8–15.

Larousse. 1992. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse

Léon, S., & C. Millot. 2005. Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du web. In *Proceedings of RECITAL 2005*, 595–604.

Lin, D. 1998. Extracting Collocations from Text Corpora. In *Proceedings of First Workshop on Computational Terminology*.

Mangeot, M., G. Sérasset & M. Lafourcade. 2003. Construction collaborative de données lexicales multilingues, le projet Papillon. *TAL, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux,* Vol. 44:2/2003, 151–176.

Mel'čuk, I., A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve: Duculot.

Pearce, D. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Pustejovsky, J. 1998. *The Generative Lexicon*. MIT Press

Schwab, D., M. Lafourcade & V. Prince. Antonymy and Conceptual Vectors. In *Proceedings of COLING'02*, 904-910.

Sérasset, G. 1994. Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA. In *Proceedings of COLING'94*. 278–282.

Sinclair, J., S. Jones & R. Daley. 1970. *English Lexical Studies: Report to OSTI on Project C/LP/08*. Internal report, Departement of English, University of Birmingham.

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

Wanner, L., B. Bohnet, M. Giereth & V. Vidal. 2005. The first steps towards the automatic compilation of specialized collocation dictionaries. *Terminology*, 11(1), 137–174.

Wu, H. & M. Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceedings of the 41st Annual Meeting of the ACL*, 120–127.