Using the Web as a Corpus for the Syntactic-Based Collocation Identification

Violeta Seretan, Luka Nerima, Eric Wehrli

Language Technology Laboratory
University of Geneva
Rue de Candolle 2, CH-1211 Geneva, Switzerland
{Violeta.Seretan, Luka.Nerima, Eric.Wehrli}@lettres.unige.ch

Abstract

This paper presents an experiment that uses a Web search engine and a robust parser for the Web-based identification of collocations (statistically significant word associations representing "a conventional way of saying things" (Manning and Schütze, 1999)). We identify the possible collocates of a given word by parsing the text snippets returned by the search engine when querying that word. Then, we rank the list of syntactic co-occurrences retrieved according to the collocational strength of each pair by using different statistical measures.

1. Introduction

The term "collocation" has different acceptions in the literature. In a broad sense, it refers to words co-occurring "within a short space of each other" (Sinclair, 1991) or to "arbitrary and recurrent word combinations" (Benson, 1990). There is a syntagmatic relation holding among the collocating words, as the word likely to occur in a given context is actually suggested by the words already present. Hence the saying "you shall know a word by the company it keeps" (Firth, 1957).

The definitions above do not consider the existence of any syntactic link between words. In a more restricted acception instead, collocations denote well-formed expressions representing "a conventional way of saying things" (Manning and Schütze, 1999). For instance, (Cowie, 1978) defines the collocation as the "co-occurrence of two or more lexical items as realization of structural elements within a given syntactic pattern". The Likelihood constraint stated by (Harris, 1988) expresses the same idea that there exists a syntactic relation between the words typically occurring together: "each word has a particular and roughly stable likelihood of occurring as argument, or operator, with a given word".

In this paper, the term collocation is understood as in the latter definition, i.e., a syntactically sound expression that is statistically significant[1], regardless of its morphological and syntactic realization[2].

Being able to produce collocations, i.e., knowing the right word to use together with another, constitutes an important element for text naturalness. It is indispensable to be aware of the institutionalized, conventional expression in order to avoid unnaturally sounding paraphrases (as the French-like *feel difficulties instead of the correct *encounter difficulties*). The problem arises from the unpredictability of the collocate across languages, dialects, domains and time periods. For instance, the French equivalent of the English collocation *encounter - difficulty* is *éprouver* 'feel' - *difficulté* 'difficulty'. In British English, one usually says **take decision**, while in common English the equivalent is **make decision**.

A popular, ad-hoc means for discovering the collocate of a word is to make a manual search for the possible alternatives on the Web and to compare the number of hits obtained. For instance, in order to find out which word is customarily used with *available*, one compares the Web frequencies of the combinations of *available* with *largely*, *widely* and *highly*, and eventually retains the most frequent combination (**widely** *available*)[3].

The Web offers a huge repository of documents written in a multitude of languages and dialects, of different categories, and constantly changing over time. It is therefore well suited for collocation related tasks.

The Web is being extensively exploited in many fields, such as information retrieval, question answering, terminology extraction, ontology learning, and for different aspects of linguistic research (see (Volk, 2002) for a review). Despite its suitability for collocation discovery, the Web has not been really used for this purpose. Only few references exist in literature, about comparing the co-occurrences hits as evidence for collocability (Manning and Schütze, 1999; Pearce, 2001) or about using Web counts in evaluating the collocations already extracted from static corpora (Inkpen and Hirst, 2002).

2. Motivation

The basic idea in using the Web to discover collocations is to choose the expression that occurs with marked higher frequency than the possible alternatives. For instance, *widely available* is much more frequent than the alternatives *highly/largely available*. It represents therefore a better collocation candidate than the other two expressions.

This test, which is based on the hits number provided by the search engines, can only provide unsure indications about the likelihood of the searched expression to constitute

---

[1]The involved words occur together *more often than by chance*, thus the collocation is restricted with respect to the collocate substitutability.

[2]We consider both *grammatical* and *lexical collocations* (Benson et al., 1986), e.g., *abstain from*, *pay attention*, without limitation on the distance between words in text (apart from the sentence boundaries).

[3]A query with Google on February 2004 returned about 852'000 results for *widely available*, about 236'000 for *highly available* and only 3'620 for *largely available*.

a collocation. In fact, the counts may comprise many cases of the two words not being related syntactically, but only co-occurring in the same context by chance. Also, due to words' ambiguity, many of the hits may concern words of unwanted categories.

Our approach is oriented towards a syntactic collocation search on the Web, in order to filter the Web samples and retain only the true positive ones.

Using a syntactic filter provides many important advantages for the task of collocation identification. First, it reduces the noise in detecting concordance instances. A pair of words is considered only if the two words belong to the same sentence and are related by a syntactic function. Secondly, it can afford the morphological variation and syntactic flexibility: thanks to the sentence normalization performed during parsing (base words form, canonical word order), an instance is detected even if the words are inflected or arbitrarily distant due to complex syntactic operations. Also, the syntactic filter retains only the concordance instances in the pattern the user is interested in, e.g., adjective-noun.

Therefore, the syntactic parsing considerably improves the accuracy in counting the co-occurrences (bigrams) frequency. But our approach goes beyond comparing Web counts of alternative expressions and identifying the most frequent as a possible collocation. We apply different statistical measures on the syntactic bigrams obtained, that quantify the collocational strength more appropriately than the simple frequency counts. Again, the syntactic filter plays an important role, as it allows using collocation tests more adequately. It has been observed (Krenn and Evert, 2001) that statistical methods perform differently when ranking collocations of different syntactic configuration, i.e., some are more appropriate for ranking for instance adjective-noun collocations than verb-object collocations.

In section 3 of this paper we present the approach and the methodology we used for the syntactic-based collocation identification from the Web. Section 4 describes the system we implemented and section 5 presents several experimental data and preliminary evaluation results. The last section points out related work and directions for further development.

3. The approach and methodology

Our approach aims to benefit from the availability and diversity of the Web documents for collocate discovery. It proposes to go beyond the simple statistics of search hits and to ensure that the co-occurring words are actually in a syntactic relation. In particular, the approach aims to build a corpus of Web samples in which a given word is found in specific syntactic relations with other words. The given word represents the base word of a collocation and it is presumably known by the user. The latter might also be specifically interested in a given syntactic configuration and a given part-of-speech for the collocate word.

The methodology we adopted employs the existing Web search technology (search engines, API services), a robust large-scale parser, and several statistical measures (the log-likelihood ratio, the mutual information and the differential *t*-test as presented in (Manning and Schütze, 1999)).

A Web search is performed with a search engine using the base word in the search query. From the snippets returned, i.e. from the contexts of searched word in the documents found, we extract all the words that are in a syntactic relation with the base word. In order to do that, we use the Fips parser ((Laenzlinger and Wehrli, 1991)) for French and English (other languages will be available in the future). We proceed with filtering the types of syntactic patterns detected, and keeping only the patterns the user is interested in.

The user is provided with the list of all syntactic bigrams retrieved that involve the base word and are in the desired syntactic configurations. At the same time, the snippets containing the bigrams detected are displayed and the user can access the originating document by clicking on its link.

Moreover, statistical measures are used to quantify the collocational strength of the syntactic bigrams (word pairs) extracted. The list of bigrams is presented to the user in the order set by these measures, i.e., in the top of the list one can find the word pairs most likely to contain a true collocate of the base word. A enhanced visualization interface allows the user to display the bigram list by using different sort and filter options.

4. Syntactic Web Collocation Extractor

This section presents the system we implemented, that allows to discover the collocate of a given word on the Web by performing a syntactical analysis of the Web text snippets and applying statistical word association measures, as explained in the previous section.

The main components of the systems and their main functionalities are:

1. The *input* module, which defines the user search options: the base word(s), the language, the syntactic pattern(s) desired, and the maximum number of results to retrieve.

2. *Google API*[4] *wrapper*, which is the program interface to the search service of Google search engine which we are employing.

3. The *Fips parser*, used to parse the snippets and to extract all the syntactic bigrams.

4. A module that implements the *statistical measures* used to quantify the collocational strength for each bigram.

5. A *concordance tool*, which enables the user to visualize the collocations extracted from a given text corpus (in this case, the single file containing the search results). It also provides advanced filtering and display options.

The system architecture is pipelined. The data flow and the processing follow the order in which the system's main components have been presented above. Between the second and third components there is a pre-processing module whose role is to extract from Google results the actual text snippets (the results from Google contain various additional information) and to rule out those of them which are empty or do not contain the searched words.

The program that allows using the Google API (the wrapper) is implemented in Java, as the Google API it-

_____

[4]http://www.google.com/apis/

self. The parser, the concordance tool and the main program that co-ordinates the components of the whole system are implemented in Component Pascal under the Black-Box 1.4 environment, a product of Oberon Microsystems Inc. (http://www.oberon.ch). We have chosen the Google search engine, as it is the one with the highest number of indexed web pages[5] offering at the same time API to access the search service. Other search engines having API access have much lower coverage (e.g., Search Hippo, GigaBlast, SearchWho).

Due to the parser's language restriction to English and French, our system can only retrieve collocations in these languages. If we loosen the syntactic constraint, our system could still be employed to find collocations in other languages, by skipping the parsing module and only applying the statistical measure for the collocation test.

One of the system's concerns is to ensure that the system retrieve only documents written in the desired language (either French or English), even if the searched word is homographic across languages (e.g., *obligation*). The search engines provide parametrizable access to their indexed web pages, thanks to which we could ensure that the language of retrieved documents is the same the user specifies. A feature that might be worth implementing is to search through specific directory categories only, in order to maximize the chances to retrieve good-quality documents, at least produced by native language speakers.

A limitation of the system comes from the restricted access to the search results through Google API. A Google API client key allows a maximum of 1'000 queries per day and one cannot access information beyond the 1'000th result for any given query.

Due to the processing time that cumulates the server search time and the parsing time, the system cannot be used online as, for instance, a web application.

Nonetheless, experiments show that interesting results are achieved even with a small amount of query results. Besides, in order to tackle with the time problem we implemented a scripting functionality that allows processing multiple queries from a list, rather than a single query.

5. Experiments and evaluation

This section presents several experimental results obtained by using the system presented above to retrieve collocates of given words.

Table 1 shows the results of an experiment in which we considered a set about 20 base words and extracted the bigrams containing the searched words, by parsing the snippets retrieved. The first row displays the average number of distinct bigrams found, for different levels in the search results list (table's columns). The second row presents the average time, in seconds, for Google API queries[6]. In the third row, we present the average parsing time on the client side, on a Pentium IV, 2.4GHz, 256MB RAM configuration.

Snippets	100	200	300	500	750	1000
Bigrams	35	69.6	103.8	158.6	221.2	262.2
Search time	3.65	9.18	16.82	37.01	114.8	143.17
Parsing time	12.9	32.2	34.96	59.3	72.76	92.58

Table 1: Distinct bigrams obtained at different results strata

For evaluating the proposed methodology, i.e., its effectiveness to find good collocates, three strategies have been adopted which are currently under development.

A first strategy consist in choosing a set of words, each of which constituting the base word of a collocation. The collocates found with our system are ranked using the log-likelihood, mutual information, and the differential *t*-test. A human judge evaluates how many among the top scored bigrams constitute a true collocation. A second experiment evaluates the top scored bigrams against the BBI dictionary (Benson et al., 1986), in order to determine the percentage of overlap between our results and the dictionary. A third experiment is being conducted in a didactic fashion. Non-native English students are asked to solve "cloze exercises", i.e., to fill the missing word in a sentence. This experiment's goal is to make a comparative analysis of the results obtained by students without and by using our tool.

The preliminary results obtained by using our tool have shown a high potential in discovering collocates. To give an idea of the results we get, we provide here two examples, for the words *approach* and *civilization*. Table 2 and 3 display the collocates obtained with our system relatively to those listed in the BBI dictionary.

6. Conclusion. Related and future work

We presented a system aimed at discovering collocations using the Web. It retrieves syntactic co-occurrences by using a search engine and a robust parser, then it applies different statistical measures of lexical associations on the bigrams extracted. The contexts provided by the search engines proved to be an appropriate and sufficient resource for discovering words' collocates. Besides the advantages of using the Web as a corpus resource, the system shows other important features, like the tuning of the statistical measures with the syntactic patterns.

There are many ideas for improving the collocation identification by using this approach. One is to apply a stronger filter on the URL of the web pages searched, by taking for instance into account the page ranking (this may be applied also to the score computation) and by filtering the web pages according to the directory category assigned. The structure of the page could also be taken into account, by looking at which section (title, header) the searched words belong. Another idea is to consider only one occurrence per page, in order to avoid repetitions of similar usages. It would be interesting to asses the impact these factors have in the quality of the results, both syntactically (more well-formed contexts) and lexicographically (more good collocations retrieved).

Other improvements concern the generation of inflected forms of query words (e.g., by adding an inflection tool to the system) and reducing the search time . The current

	Verb-Object	Adjective-Noun	Other types
BBI only	make	creative, holistic, careful, cautious, conservative, judicious, direct, forthright, down-to-earth, pragmatic, realistic, objective, rational, scholarly, easy-going, indirect, simplistic, unrealistic, hard-nosed, inflexible, uncompromising, no-nonsense	$\sim$ from, $\sim$ about, with the $\sim$ of
Common	take	Innovative, fresh, new, novel, scientific	$\sim$ for, $\sim$ with, $\sim$ to
Our tool only	adopt, base, convey, develop, follow, use	unique, comprehensive, complementary, systematic, alternative, multidisciplinary, common, practical, straightforward, reliable, general, traditional, effective	$\sim$ over, $\sim$ a client, $\sim$ directive $\sim$ demonstrate, project $\sim$, introduce to $\sim$, feasibility of $\sim$

Table 2: Collocates of *approach* (obtained with our tool vs. listed in the BBI dictionary)

	Verb-Object	Adjective-Noun	Other types
BBI only	spread, stamp out	advanced	
Common	introduce, create, destroy	ancient, modern	cradle of $\sim$
Our tool only	develop	early, flourishing, human, new, noble	$\sim$ rise, $\sim$ emerge, $\sim$ extend, $\sim$ grow, $\sim$ fall, expansion of $\sim$, fall of $\sim$, collapse of $\sim$, founder of $\sim$, era of $\sim$, development of $\sim$

Table 3: Collocates of *civilization* (obtained with our tool vs. listed in the BBI dictionary)

system version was conceived for experimental purposes rather than for a large-scale use. Still, by its scripting facility it could already be used in lexicography and in NLP applications dealing with collocations, as well as by non-native speakers careful to produce proficient text.

Examples of related work that reflect the same interest in using the Web as a corpus for searching interesting word patterns are the WebCorp[7] and KWiCFinder[8] tools. The first performs a wildcards-based word search and takes into account the frequency counts when looking for collocates. The second performs complex boolean searches for word concordances. Our tool is instead syntactically oriented and uses sophisticated lexical association measures for the collocation identification.

There exist some search interfaces to static corpora, e.g. COBUILD[9] for Bank Of English and SARA[10] for British National Corpus that perform a syntactic-based search, but which is limited actually to the lexical category (POS). The Gsearch tool (Corley et al., 2001) includes also a syntactic patterns search, but it does not account for collocations.

To our knowledge, the system we implemented is the first which is oriented towards the collocation discovery from the Web using the syntactic analysis and statistical measures.

7. References

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

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

Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin, 2001. Finding syntactic structure in unparsed corpora: The gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Cowie, Anthony P., 1978. The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens (ed.), *In Honour of A.S. Hornby*. Oxford: Oxford University Press, pages 127–139.

Firth, John R., 1957. *Modes of Meaning*. Oxford: Oxford Univ. Press, pages 190–215.

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

Inkpen, Diana Zaiu and Graeme Hirst, 2002. Acquiring collocations for lexical choice between near-synonyms. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*. Philadephia, Pennsylvania.

Krenn, Brigitte and Stefan Evert, 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*. Toulouse, France.

Laenzlinger, Christopher and Eric Wehrli, 1991. Fips, un analyseur interactif pour le français. *TA informations*, 32(2):35–49.

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

Pearce, Darren, 2001. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*. Carnegie Mellon University, Pittsburgh.

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

Volk, Martin, 2002. Using the web as a corpus for linguistic research. In R. Pajusalu and T. Hennoste (eds.), *Catcher of the Meaning. A festschrift for Professor Haldur im.* Publications of the Department of General Linguistics 3, University of Tartu, Estonia.

[7] http://www.webcorp.org.uk/index.html
[8] http://miniappolis.com/KWiCFinder/KWiCFinderHome.html
[9] http://titania.cobuild.collins.co.uk/
[10] http://sara.natcorp.ox.ac.uk/lookup.html