Induction of Syntactic Collocation Patterns from Generic Syntactic Relations

Violeta Seretan
University of Geneva
Language Technology Laboratory
2 Rue de Candolle, Geneva, CH-1211, Switzerland
Violeta.Seretan@lettres.unige.ch

Abstract

Syntactic configurations used in collocation extraction are highly divergent from one system to another, this questioning the validity of results and making comparative evaluation difficult. We describe a corpus-driven approach for inferring an exhaustive set of configurations from actual data by finding, with a parser, all the productive syntactic associations, then by appealing to human expertise for relevance judgements.

1 Introduction

The term *collocation*, often used in different senses in the literature, is understood here as in the following statement: "The term collocation refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure." [Fontenelle, 1992, 222]. Cross-lingual examples, such as "*heavy* smoker" and "*pay* attention", equivalent to "*grand* fumeur" and "*faire* attention" in French, show why collocations are crucial from the text encoding perspective: the lexical choice of the modifier and support verb is restricted by the conventional usage, while the alternatives are generally perceived as unnatural or "weird"[1].

Typically, collocation acquisition from corpora relies on statistical significance tests for pairs of words occurring close to each other. The recent developments in syntactic parsing allowed the move towards more linguistically-informed methods of extraction. Current systems rely increasingly on the syntactic pre-processing of source text (such as POS tagging, chunking, shallow or deep parsing) intended to support the identification of collocation candidates, prior to the statistical analysis.

Despite the advantages brought by syntax (well-formedness of results; drastic reduction of combinatorial complexity; partitioning of candidates into syntactically homogeneous classes), a serious problem arises: the arbitrariness in the choice of syntactic configurations (*patterns*) for collocation candidates. The rest of the paper discusses the implications of this problem and presents the solution we proposed in order to cope with it.

[1]As, for instance: "*big* smoker", "*make* attention", or "*lourd* fumeur", "*payer* attention".

2 Syntactic Patterns for Collocations

In the most permissive case (when no extraction patterns are defined), any pair of words is regarded as a valid collocation candidate. This results in much noise, since the most frequent word combinations are also the least interesting (e.g., "of the", "in the"). Therefore, many extraction systems perform the linguistic analysis of text and apply a linguistic filter on collocation candidates (very often, they perform POS tagging in order to filter out the pairs involving function words[2]).

There is unfortunately much disagreement with respect to the accepted syntactic configurations for collocations (often as a consequence of the lack of consensus in the understanding of the notion of collocation). There is much divergence, first, with respect to the POS of participating words. Some authors consider open-class words only ([Justeson and Katz, 1995; Hausmann, 1989]), while most of them allow for function words too (as in "agree *on*"). Second, the syntactic relations proposed are almost always different. The following list shows the diversity of the syntactic patterns used, for English, in various works:

1. Lexical collocations[3] in BBI dictionary [Benson *et al.*, 1986]: V-N, N-A, N-V, N-P-N, A-Adv, V-Adv;
2. Hausmann's collocation definition [1989]: N-A, N-V, V-N, V-Adv, A-Adv, N-[P]-N;
3. Xtract collocation extraction system [Smadja, 1993]: N-A, N-V, V-N, V-P, V-Adv, V-V, N-P, N-D;
4. WordSketch concordance system [Kilgarriff *et al.*, 2004]: N-A, N-N, N-P-N, N-V, V-N, V-P, V-A, N-Conj-N, A-P;
5. FipsCo system [Goldman *et al.*, 2001]: N-A, N-N, N-P-N, N-V, V-N, V-P, V-P-N.

These examples show how much the perspectives adopted by various authors differ from the initial view, in which the collocation is seen as independent of the syntactic structure [Fontenelle, 1992, 222].

The arbitrariness in pattern choice questions the quality of the results and makes comparative evaluation difficult. Besides, a set of patterns established for a language may not completely fit another one. It is therefore necessary to find

[2]This restriction is seen as too strong [van der Wouden, 2001].

[3]The BBI dictionary also includes a wide range of "grammatical collocations", such as: N-P, N-Conj, P-N, A-P, A-Conj, etc.

a means to establish, for each language, an exhaustive set of collocation patterns to be used as reference.

3 Pattern Induction Experiments

In order to overcome the problem of arbitrariness in pattern choice, we propose to make a corpus-driven investigation aimed at the discovery of all possible and interesting collocation configurations.

Rather than relying on linguistic prescriptions, we try to induce these configurations from actual data. We do not commit ourselves to pre-defined patterns; instead, we consider any POS combination as a priori possible, and we only require the items of a pair to be syntactically related. We use Fips, a GB-based parser [Wehrli, 2004], to extract such generic relations among words. Then we analyze the obtained results and infer the syntactic patterns from them.

More specifically, we consider the following generic relations: head-modifier, head-complement, verb-argument (both subject and objects). We extract the word pairs represented by the combination of a head with the lexical head of its specifier or of its complement (cf. GB theory).

Two experiments have been performed on English and French corpora of newspaper articles. Several statistics are shown in Table 1: size of corpora; number of word pairs in a generic relation (tokens); number of distinct pairs (types)[4]; and the number of POS combinations detected.

Experimental data	English	French
size (words)	0.5 M	1.6 M
word pairs (tokens)	0.18 M	0.75 M
word pairs (types)	0.07 M	0.17 M
POS combinations	60	57

Table 1: Generic relations extracted

The last row shows that many POS combinations are actually productive, from the total of 98 combinations (with V, N, A, Adv, P, D, Conj in either specifier and complement position). By manually inspecting the obtained POS associations we discovered new collocation patterns that were commonly ignored - especially the patterns involving non-content words. We here provide some examples:

1. English: N-P (*alliance between*), P-N (*across border*), V-Conj (*judge whether*), A-Adv (*mature enough*), Adv-Adv (*much more*), Adv-P (*together with*);
2. French: P-N (*sous pression*) N-P (*débat sur*), N-Conj (*temps que*), A-P (*prêt à*), V-A (*rester impassible*), P-Adv (*comme jamais*).

A more in-depth analysis of results is under way, which aims to identify the relevant configurations from the POS combinations found with our method.

4 Conclusion and Related Work

The performance of collocation extraction systems (in terms of accuracy and coverage) is highly dependent on the degree of syntactical permissiveness. Too much permissiveness leads to the problems of noise and combinatorial explosion; too strong constraints risk not to capture the whole range of collocational phenomena. We proposed a trade-off between these extremes: we only maintained a minimal syntactic constraint (the presence of a syntactic link between collocation's items), then we induced the collocation patterns in a data-driven fashion. The experiments conducted revealed several new collocation patterns that involve closed-class words, and led us to support the claim made by van der Wouden [2001, 17], that "lexical elements of almost any class may show collocational effect".

We compare our work with that of Dias [2003], which attempts to overcome the pattern pre-definition problem by using combined statistics on sequences of adjacent words and their POS. Since this method ignores the sentence structure, it is still affected by combinatorial explosion. In contrast, our method is computationally tractable, and, in addition, allows us to capture long distance collocational pairs[5].

References

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

[Dias, 2003] Gaël Dias. Multiword unit hybrid extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, pages 41–48, Sapporo, Japan, 2003.

[Fontenelle, 1992] Thierry Fontenelle. Collocation acquisition from a corpus or from a dictionary: a comparison. *Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere*, pages 221–228, 1992.

[Goldman *et al.*, 2001] Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations*, pages 61–66, Toulouse, France, 2001.

[Hausmann, 1989] Franz Iosef Hausmann. Le dictionnaire de collocations. In F. I. Hausmann et al., editor, *Wörterbücher: ein internationales Handbuch zur Lexicographie. Dictionaries, Dictionnaires*, pages 1010–1019. de Gruyter, Berlin, 1989.

[Justeson and Katz, 1995] John S. Justeson and Slava M. Katz. Technical terminology: Some linguistis properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[Kilgarriff *et al.*, 2004] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. The Sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France, 2004.

[Smadja, 1993] Frank Smadja. Retrieving collocations form text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[van der Wouden, 2001] Ton van der Wouden. Collocational behaviour in non content word. In *Proceedings of the ACL Workshop on Collocations*, pages 16–23, Toulouse, France, 2001.

[Wehrli, 2004] Eric Wehrli. Un modèle multilingue d'analyse syntaxique. In A. Auchlin et al., editor, *Structures et discours - Mélanges offerts à Eddy Roulet*, pages 311–329. Éditions Nota bene, Québec, 2004.

[4] The parser identifies all the instances of a lexeme pair, irrespectively of the surface realization.

[5] As for instance *"submit proposal"* in "a *proposal* which addressed various topics has been *submitted*".