Collocation Extraction for Machine Translation

Brigitte Orliac[1] and Mike Dillinger[2]

[1]Département de linguistique et de traduction
Université de Montréal
Montreal, Canada
brigitte.orliac@umontreal.ca

[2]Spoken Translation, Inc.
Berkeley, CA USA
mike.dillinger@spokentranslation.com

Abstract

This paper reports on the development of a collocation extraction system that is designed *within* a commercial machine translation system in order to take advantage of the robust syntactic analysis that the system offers and to use this analysis to refine collocation extraction. Embedding the extraction system also addresses the need to provide information about the source language collocations in a system-specific form to support automatic generation of a collocation rulebase for analysis and translation.

1 Introduction

Collocations describe the habitual word combinations of a language, expressions such as En. *fierce battle*, *strong resistance*, *pay attention* or Fr. *lutte acharnée*, *faire attention*. Ever since their introduction in the works of the British scholar J.R. Firth (1957), collocations have been the subject of numerous descriptions, the most complete of which was put forward by I. Mel'čuk within the framework of the Meaning-Text Theory (for detailed descriptions, see Mel'čuk 1996, 1998, 2003).

Collocations, like idioms, belong to the set phrases of a language. Their interpretation differs from that of idioms in one crucial respect: while the meaning of an idiom is mostly incomprehensible if not previously heard (cf. *to kick the bucket*), the meaning of a collocation is only partially obscure, owing largely to the element which keeps its meaning inside the phrase (and is considered the "base" of the collocation). The relative opacity of a collocation is caused by the "collocate", the word arbitrarily selected to express a particular meaning in relation to the base (consider, e.g., the adjective *fierce* or the verb *pay* in the examples above).

While they are often realized idiomatically, collocational meanings can also be found in regularly constructed (i.e. compositional) phrases: the expression *to brush one's teeth* is semantically transparent, yet considered a collocation of English. To express the same meaning, one would not say **to wash one's teeth* (either verb is used to produce the equivalent collocation in French)[1]. Although a common property with collocations, non-compositionality alone does not define them. The defining property is the non-compositionality of selection of the collocation's components, a phenomenon most readily observable in language production.

There is a clear consensus in the literature that collocation data is very useful for parsing, word sense disambiguation, machine translation, text generation, and other applications. This is particularly true for commercial machine translation (MT) systems where collocation data can contribute to improved performance in almost every component of processing. We have observed in practice that adding new collocations often has an impact on translation quality that is more obvious to end users than incremental additions to the dictionary. In this sense, collocations are the key to producing more acceptable output from commercial systems.

The automatic extraction of collocations can have a dramatic impact on the time and cost

[1] Example borrowed from Mel'čuk, 2003.

involved in developing the linguistic resources needed for higher quality translation output. Automating extraction provides a practical means for augmenting general vocabulary resources with collocations and for supplementing specialized or technical terminologies – made up mostly of noun phrases – with larger, cross-categorial phrases (particularly verb + noun combinations) that are less often present in technical glossaries. Collocation data, then, plays a significant role both in system development and in customization of a system to a particular client's needs. There is also another, less obvious, but equally important application of automated collocation extraction: automation makes it practical to generate and prioritize project-specific lists of collocations to serve as requirements documents for systematic planning and project management, which contribute significantly to reducing development costs (see Dillinger, 2001).

This paper describes a collocation extraction system designed within the Logos machine translation system from GlobalWare AG. After briefly describing how the Logos system represents collocations internally, we review the principal methods that have been applied to the automatic extraction of collocations from texts. We then present a method for acquiring and validating collocation candidates to support automated development of a collocation rulebase. We finish with a discussion of preliminary results based on an analysis of collocations extracted from a half-million-word corpus of computer science texts.

2 Collocations in the Logos MT system

In the Logos MT system, collocations are represented as rules in a separate database, the Semantic Table database or "SemTab". SemTab is essentially a transfer module tailored specifically for collocations. Rules in SemTab are usually indexed by the head of the phrase they describe and they specify a source-language combination together with its translation. Examples of SemTab rules between English and French are given below:

- Adj(firm) N(conviction) = N(conviction, fem) Adj(inébranlable)
- V(go, vi) Prep(into) N(effect) = V(entrer) Prep(en) N(vigueur, masc)
- V(pay, vt) N(attention) = V(faire) N(attention, fem)

Besides the lexical items themselves, SemTab rules represent grammatical information such as part of speech, gender, number, voice, etc. SemTab rules encode deep syntactic relations (between a noun and its modifiers, a verb and its arguments) so that the same rule will be activated in a variety of grammatical contexts (viz. active, passive, and nominalized constructions). Whenever possible, the rules use semantic classes to represent the noun in the combination (i.e., the base of the collocation), allowing a high degree of generalization in the description and translation of collocations (a rule combining the verb *raise* with the semantic class <CROP> will match the words *corn*, *wheat*, *beans*, etc.). The majority of SemTab rules represent verb + noun combinations, and a smaller number exist for adjective + noun phrases.

SemTab rules are used to inform source-sentence parsing by identifying the argument structure and government pattern of a verb and to set target-sentence grammatical features, implementing structural transfer. Most importantly, though, they are the main mechanism for context-dependent selection of target-language lexical items (for example *raise* in "raise corn" is *cultiver* in French, while in "raise children" it is *élever*), which achieves significant improvement in readability and perceived quality of the translation produced.

Currently, these rules are produced manually by lexicographers using a menu-driven tool (based on rule templates) that generates the system code in the background. Our goal is to augment this process with automatically drafted rules based on collocations that are extracted from general and client-specific corpora. Once we can automate the link between extraction and rule generation in compatible formats, we can implement a degree of autonomous learning by the system.

3 Acquiring collocations

Our focus here is on acquiring source-language collocations for use in MT systems, which is the same problem that many researchers have investigated in other application contexts. Early approaches to collocation extraction focused on word strings in linear proximity to each other (Berry-Rogghe, 1973; Choueka et al., 1983;

Church & Hanks, 1990) and developed an array of reliable techniques for assessing the statistical strength of association among the elements of multi-word expressions (see reviews and/or comparisons in Kilgarriff, 1996; Evert & Krenn, 2001; Pearce, 2002).

Goldman et al. (2001), however, argue that a significant proportion of collocations occur outside the arbitrary 5- to 10-word radius that has been studied, because of the wide range of sentential structures that can occur between a base term and its collocate(s). They also raise the problem that searching all positions within a given radius is wasteful because only a limited number of the positions to the left and right of a base term are linguistically motivated. To avoid these problems and also increase the relevance of the collocations extracted, researchers have added linguistic filters to the statistical measures used in identifying collocations. In one of the first programs to combine linguistic and statistical methods, Smadja (1993) uses linguistic criteria to validate combinations extracted on purely statistical grounds. More recent approaches have applied linguistic knowledge first, using richer, syntactically annotated input for the extraction algorithm, in essence performing extraction over (shallow) parse trees rather than over sentences. In these approaches, statistical measures of association are then used to filter out uncommon combinations (Daille, 1996; Lin, 1998; Goldman et al., 2001; Kilgarriff & Tugwell, 2001).

Using richer input makes even more sense when extracting collocations for an MT system, for several reasons. For one, a parser is already available in the system at no additional cost: morpho-syntactic and semantic information is readily available. Collocations are extracted on the basis of word *types* rather than word tokens, making extraction more efficient. Another reason is that it is advantageous for the collocation acquisition to respond to on-going improvements in the parser and dictionaries. This allows the development of collocation data to be coordinated with and respond to the development of other parts of the MT system.

Finally, in the context of MT development it is particularly important that the collocations identified have morpho-syntactic annotations in a format that is system specific. Using external collocation extraction systems based on different dictionaries, tokenization algorithms, and/or parsers may generate collocations that the MT system will not be able to 'absorb' without additional manual work. We need to guarantee that the collocations will be system-specific to support the next step: generating directly the rules that the MT system will use to implement collocation data for translation.

These constraints point to the significant advantages of a collocation acquisition module *embedded* in the MT engine, rather than the use of some external, dedicated system, which leads to the attendant difficulties of data reformatting, system maintenance, and compatibility of analyses. This reasoning motivated our development of the present system for the Logos machine translation system, with the goal of automatically generating corpus-specific collocation requirements for customization, as well as candidate collocation rules for parsing and translation.

4 Colex

Our method for extracting collocations for use in machine translation also combines linguistic and statistical techniques. First, we use linguistic filters in the form of morpho-syntactic rules to match and extract all the verb phrases of a corpus of technical texts. Then, we apply statistical filters to distinguish collocations from accidental combinations. The current version of our collocation extraction tool (Colex) is specialized for verb + noun combinations, which make up the bulk of the collocation data in Logos (a separate *TermSearch* utility exists for noun phrases). Furthermore, Colex is designed to identify all of the verbal collocates for a user-specified noun. In our approach, nouns form the base of verbal collocations, semantically controlling the choice of verb while filling one of three possible (deep) syntactic roles: subject, direct/main object, or indirect/second object.

We begin the process by extracting from the specialized corpus all of the contexts of a base term T (i.e., all of the sentences containing T), using a concordancer. The concordances are then parsed using the Logos machine translation system. This produces a system-specific parse tree for each concordance, which represents its surface syntactic structure (in the form of semantically

labeled governing nodes that have morpho-syntactic annotations). Using syntactically annotated trees as the basis for extraction allows us to target the subtrees that represent the verb and its arguments over a much wider range of structures than would be possible with a raw-text-based tool.

To extract the verbal collocates of a base term T from the parse trees produced by the Logos system, Colex uses morpho-syntactic rules that represent fully-specified verb argument structures. The rules use optional elements and grammatical features to extract verb + noun combinations from a variety of syntactic contexts (e.g., active, passive, infinitive, and gerund sentences). There are three types of rules, one for each of the expected positions of T inside the verb phrase. In the example rule below, T is represented by the variable "string", and the rule extracts a verb + direct object (obj1) combination with an optional indirect object (obj2), auxiliary verbs and adverbs present:

```
If (n_subj (aux) v_active (adv) string_obj1 (n_obj2))
    then {v + string;}
endif
```

Colex rules for each argument position inventory the morpho-syntactic patterns that characterize verb phrases as instances of a single collocation pattern, for identification, normalization, printing, and counting. The search of parsed concordances is activated upon specifying T (which instantiates the variable in each rule). The smallest subtree containing both the verb and T is extracted along with the strings and grammatical features associated with each node. Each pair of verb and T is printed in canonical order (based on deep syntactic relations) and counted.

In extracting the verbal collocates of a noun, Colex discriminates between the intransitive and transitive readings of the verb (based on Logos verb classes): transitive verb pairs are identified and counted separately from the intransitive ones. Colex rules also extract governed prepositions along with the verb and noun strings. This grammatical information about transitivity and preposition governance, together with part of speech and grammatical function, is necessary for generating the appropriate notation for SemTab rule generation.

To measure rule coverage, we randomly selected 183 sentences from a corpus of computer texts. Of the 220 verb + noun combinations present in the sample, 132 were automatically extracted, representing a recall of 60%. Almost half of the missed combinations (18%) were found in relative clauses, a structure not yet covered in Colex, while 10,5% of the possible pairs were missed due to parser errors.

5 Results and evaluation

Using the method described above, we extracted all the verb + noun combinations for the ten most frequent (single-word) terms in our test corpus (*computer, file, program, data, server, software, drive, user, disk, information*). To illustrate, Table 1 shows the first 15 verb + object combinations for *file*. The table lists, along with each verb + object pair, its frequency in the corpus.

Table 1. The first 15 verbal collocates of *file,* ranked by frequency.

f	*verb*	*object*
24	open	file
22	save	file
18	create	file
18	download	file
17	copy	file
14	locate	file
14	send	file
9	store	file
8	access	file
8	delete	file
8	edit	file
8	find	file
8	load	file
8	read	file

Initial analysis of the pairs extracted by Colex showed that frequency, an easily-computed property of terminological units, could not be used to reliably establish the collocational status of a particular pair. First, an evaluation based on frequency alone did not retain low frequency pairs. Second, as illustrated in Table 1, frequency does not discriminate clearly enough among candidates (a number of pairs appear with the same frequency). Finally, and most importantly, high frequency also characterizes atypical associations like "find file."

To more reliably identify the verbal collocates of a term and prioritize the extracted combinations for easier integration into SemTab, we computed two widely used association measures for each extracted pair: mutual information (MI) and log-likelihood (Logl). Intuitively, both measure the probability of observing two words together, while correcting for their frequencies alone. High scores with either of these measures have been shown to correlate with collocational properties.

In order to obtain the frequency data necessary for developing the contingency tables used in computing MI and log-likelihood, we ran a separate extraction for each verb that was paired more than once with T, instantiating the "string" variable with the verb this time and extracting all of its arguments. We then elaborated a contingency table with the frequencies of the following co-occurrence pairs:

(a) verb + T

(b) verb + ¬T

(c) ¬verb + T

(d) ¬verb + ¬T (¬T represents an argument of the same category that is *not* T).

We calculated the sum N of the pairs extracted for a given argument category ($a + b + c + d$) and computed the verb's mutual information and log-likelihood scores with T, using the following definitions for the two association measures:

$$MI = \log_2 \frac{aN}{(a+b)(a+c)}$$

$$\begin{aligned} Logl = {} & a\log a + b\log b + c\log c + d\log d \\ & - (a+b)\log(a+b) - (a+c)\log(a+c) \\ & - (b+d)\log(b+d) - (c+d)\log(c+d) \\ & + N\log N \end{aligned}$$

The following table displays the two association scores, and the raw frequency, of the top 15 verbal collocates of *file* (in direct object position), ranked according to the log-likelihood ratio.

Table 2. The first 15 verbal collocates of *file,* ranked by log-likelihood.

	Logl	*MI*	*f*
open	16.05	3.77	24
copy	14.15	4.44	17
download	13.03	3.98	18
save	11.62	3.19	22
locate	8.87	3.61	14
back up	6.36	4.75	7
edit	5.74	3.94	8
upload	4.99	5.15	5
delete	3.90	3.00	8
attach	3.69	3.16	7
name	3.40	3.79	5
send	3.29	1.91	14
load	2.73	2.38	8
create	2.72	1.49	18
update	2.50	3.05	5

The log-likelihood ratio seems to give the better results for linguistically extracted pairs and has been used as the measure of choice in other systems that base the extraction of collocations on morpho-syntactic criteria (Daille, 1996; Goldman et al., 2001). These researchers also found that mutual information tends to emphasize rare combinations while penalizing frequent ones (MI had ranked the verbs *drag*, *erase*, and *restore* among the top 15 collocates of *file*, whereas log-likelihood selected the more characteristic *create*, *load* and *send*).

Our baseline for evaluation in this study is lexicographers' performance: we want to know what advantage (semi-)automatic methods offer over traditional lexicographic methods. Smadja (1991) calculates that for lexicographers developing the Oxford English Dictionary, for example, the ratio of proposed collocation candidates to good candidates is 4% and "even a precision rate of 40% [by automatic extraction] would be helpful". The usefulness of automatic extraction was borne out in practice by the use of *WASPBench* (Kilgarriff & Tugwell, 2001) in the development of collocations for inclusion in the MacMillan English Dictionary for Advanced Learners (2002) [http://www.hltcentral.org/page-937.shtml]. In our application context, then, precision is much more important than recall and we focus here on the former.

As a first measure of the precision of our system, then, we checked the top combinations for the ten most frequent terms in each of the three argument relations (a total of 150 candidate collocations)

against existing collocation rules in the SemTab database. There were several reasons to take this approach: the existing rules were developed manually by experienced lexicographers; collocations cannot be translated literally, so need to be listed separately in the lexicon (or in a special rulebase as with Logos); and the Logos system has particularly well-developed computer terminology. Moreover, there are few alternatives available: it is very difficult to use published terminologies as an evaluation standard, for example, because they emphasize noun phrases, and verbs, when present, are included without specification of their arguments.

As measured against this human-verified standard, we found that more than half of the domain-specific collocations extracted with Colex (52.5%) matched rules already in the SemTab database.

Further consideration suggested that the remaining candidate collocations (not present in the SemTab database) might include: 1) valid collocations that were not yet present in the database, 2) collocations that were represented as domain-specific specialized senses and placed directly in the Logos dictionary rather than as SemTab rules, or 3) invalid candidates. As a further evaluation step, then, we consulted the Logos dictionary to find out how many of the top verbs in the remaining collocations had been included with a domain-specific sense under the Computers domain (to keep them distinct from the general sense in the General dictionary). This lexical representation strategy accounted for another 22.5% of the candidate collocations, bringing the total of valid collocations extracted with our method to 75%. The remaining 37 collocation candidates not found in Logos were verified by hand and all but four of them represented characteristic verb + object combinations of computer terminology (e.g., "name file," "input data," "load something from disk"), bringing our precision to 97%.

6 Conclusion

We have successfully extracted, from a domain-specific corpus, a significant number of source-language collocations for use in developing and customizing a machine translation system. The use of the MT system's parser ensured that the extraction was linguistically motivated and that processing resources would not be spent on spurious word associations. The Colex tool we developed allowed us to extract candidate collocations from a broader-than-usual array of linguistically motivated surface-syntactic contexts, leading to increased sampling from the corpus and to greater representativity and productivity of the collocations found.

The collocations extracted show a high degree of precision, which suggests that the method described will in fact contribute to more efficient, more cost-effective development and customization of machine translation systems. The collocations also preserve a wide array of grammatical features from the original parse in a system-specific format, so they can be used to speed the development of bilingual transfer rules that will significantly enhance the readability of machine translation output. Finally, by embedding the collocation extraction module in the MT engine, extraction performance improves automatically with ongoing development of grammars and dictionaries and takes us another step closer to automating more elements of rule generation.

7 Acknowledgments

The authors wish to thank GlobalWare AG for their support throughout the development of this project. We would also like to thank P. Drouin and B. Robichaud for helping us make sense of the statistical formulas. Finally, we thank the anonymous reviewers for their valuable comments.

8 Bibliographical References

Berry-Rogghe, G. 1973. The computation of collocations and their relevance to lexical studies. In A.J. Aitken, R.W. Bailey, and N. Hamilton-Smith (Eds.), *The Computer and Literary Studies* (pp. 103-112). Edinburgh: University Press.

Choueka, Y., Klein, S.T. & Neuwitz, E. 1983. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. *Journal of the Association for Literary and Linguistic Computing*, 4, 34-38.

Church, K. & Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22-29.

Daille, B. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J.L. Klavans and P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* (pp. 49-66). Cambridge, Massachusetts: MIT Press.

Dillinger, M. 2001. Dictionary development workflow for MT: Design and management. In B. Maegaard (Ed.), *Machine Translation in the Information Age* [MT Summit VIII] (pp. 83-87). Geneva: European Association for Machine Translation.

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

Evert, S. & Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Firth, J.R. 1957. Modes of Meaning. In *Papers in linguistics 1934-1951* (pp. 190-215). Oxford University Press.

Goldman, J.P., Nerima, L. & Wehrli, E. 2001. Collocation Extraction Using a Syntactic Parser. *Workshop proceedings: COLLOCATION: Computational Extraction, Analysis and Exploitation*, Toulouse, France.

Kilgarriff, A. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex, UK.

Kilgarriff, A. & Tugwell, D. 2001. WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. In B. Maegaard (Ed.), *Machine Translation in the Information Age* [MT Summit VIII]. Geneva: European Association for Machine Translation.

Lin, D. 1998. Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology*, Montreal, Canada.

Mel'čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37-102). Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, I. 1998. Collocations and Lexical Functions. In A.P. Cowie (Ed.), *Phraseology. Theory, Analysis, and Applications* (pp. 23-53). Oxford University Press.

Mel'čuk, I. 2003. Collocations dans le dictionnaire. In T. Szende (sous la dir. de), *Les écarts culturels dans les dictionnaires bilingues* (pp. 19-64). Paris: H. Champion.

Pearce, D. 2002. A Comparative Evaluation of Collocation Extraction Techniques. *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

Smadja, F. 1991. From N-grams to Collocations: An Evaluation of Xtract. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA.

Smadja, F. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1), 143-177.