FipsCoView: On-line Visualisation of Collocations Extracted from Multilingual Parallel Corpora

Violeta Seretan
School of Informatics
University of Edinburgh
violeta.seretan@gmail.com

Eric Wehrli
Language Technology Laboratory
University of Geneva
eric.wehrli@unige.ch

Abstract

We introduce FipsCoView, an on-line interface for dictionary-like visualisation of collocations detected from parallel corpora using a syntactically-informed extraction method.

1 Introduction

Multilingual (parallel) corpora—e.g., Europarl (Koehn, 2005)—represent a valuable resource for tasks related to language production that is exploitable in a wide variety of settings, such as second language learning, lexicography, as well as human or automatic translation. We focus on lexicographic exploitation of such resources and present a system, called FipsCoView,[1] which is specifically aimed at supporting the work of lexicographers who compile multilingual collocation resources.

Collocation, a rather ill-defined linguistic concept referring to a large and heterogeneous sub-class of multi-word expressions, is understood here as a combination of words that produces natural-sounding speech and writing (Lea and Runcie, 2002) and that has syntactic and semantic properties which cannot be entirely predicted from those of its components and therefore has to be listed in a lexicon (Evert, 2004). Collocations are particularly interesting from a translation point of view, and our system can also be used to facilitate the task of translators looking for the right translation of a word in context.

The usage scenario is the following. Given a word, like *money*, our system provides a concise and intuitive presentation of the list of collocations with that word, which have previously been detected in the source language version of the parallel corpus. By selecting one of the items in this list, e.g., *money laundering*, users will be able to see the contexts of that item, represented by the sentences in which it occurs. In addition, users can select a target language from the list of other languages in which the multilingual corpus is available[2] and visualise the target language version of the source sentences.

This presentation enables users to find potential translation equivalents for collocations by inspecting the target sentences. Thus, in the case of French, the preferred equivalent found is *blanchiment d'argent*, lit., 'money whitening', rather than the literal translation from English, **lavage d'argent*. In the case of Italian, this is *riciclaggio di denaro*, lit., 'recycling of money', rather than the literal translation *?lavaggio di soldi*, also possible but much less preferred. Access to target sentences is important as it allows users to see how the translation of a collocation vary depending on the context. Besides, it provides useful usage clues, indicating, *inter alia*, the allowed or preferred morphosyntactic features of a collocation.

In this paper, we present the architecture of FipsCoView and outline its main functionalities. This system is an extension of FipsCo, a larger fully-fledged off-line system, which, in turn, is integrated into a complex framework for processing multi-word expressions (Seretan, 2009). While the off-line system finds direct applicability in our on-going projects of large-scale multilingual syntac-

[1] Available at `http://tinyurl.com/FipsCoView`.

[2] Europarl includes 11 languages: French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek, Finnish. Note that our tool is not tailored to this specific corpus.

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 125–127,
Portland, Oregon, USA, 23 June 2011. ©2011 Association for Computational Linguistics

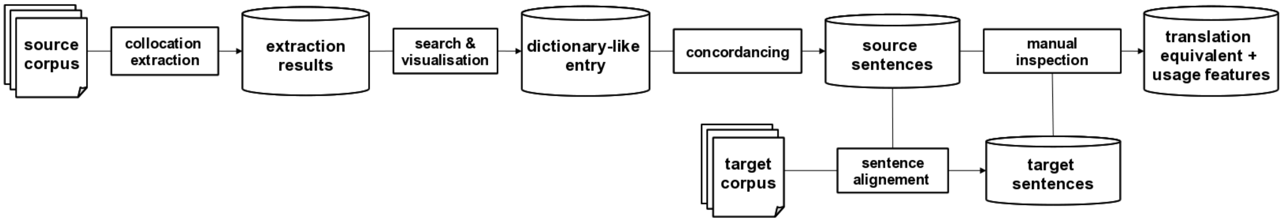

Figure 1: FipsCoView: System architecture.

tic parsing (Wehrli, 2007) and syntax-based machine translation (Wehrli et al., 2009), the on-line version is designed to offer access to the derived collocation resources to a broader community.

2 Architecture and Main Functionalities

Figure 1 shows the architecture of FipsCoView. The main system modules are the *collocation extraction* module, the *search & visualisation* module, the *concordancing* and the *sentence alignment* modules.

The processing flow is pipelined. The key module of the system, *collocation extraction*, relies on a syntax-based methodology that combines lexical statistics with syntactic information provided by Fips, a deep symbolic parser (Wehrli, 2007). This methodology is fully described and evaluated in Seretan (2011). In principle, the extraction takes place only once, but new corpora can be processed later and results are cumulated. The *sentence alignment* (Nerima et al., 2003) is performed partially, i.e., only for the sentences actually displayed by the concordancing module. It is done on the fly, thus eliminating the need of pre-aligning the corpora.

The role of the *concordancing* module is to present the sentence contexts for a selected collocation (cf. scenario described in §1). The words in this collocation are highlighted for readability. The list of sentences is displayed in the order given by the syntactic variation of collocations, that is, the collocation instances for which the distance between the components is larger are displayed first. This functionality is designed to support the work of users inspecting the syntactic properties of collocations.

The *search & visualisation* module takes as input the word entered by the user in the system interface, performs a search in the database that stores the collocation extraction results, and provides a one-page presentation of the collocational information related to the sought word. Users can set visualisation parameters such as the minimal frequency and association score, which limit the displayed results according to the number of occurrences in the corpus and the "association strength" between the component words, as given by the lexical association measure used to extract collocations. The measure we typically use is log-likelihood ratio (Dunning, 1993); see Pecina (2008) for an inventory of measures.

Depending on these parameters, the automatically created collocation entry is more or less exhaustive (the output adapts to the specific user's purpose). A different sub-entry is created for each part of speech of the sought word (for instance, *report* can either be a noun or a verb). Under each sub-entry, collocations are organised by syntactic type, e.g., adjective-noun (*comprehensive report*), noun-noun (*initiative report*), subject-verb (*report highlights*), verb-object (*produce a report*). To avoid redundancy, only the collocating words are shown. The sought word is understood and is replaced by a tilde character, in a paper dictionary style. Unlike in paper dictionary presentations, the online presentation benefits from the HTML environment by using colours, adapting the font size so that it reflects the association strength (the most important combinations are more visually salient), displaying additional information such as score and frequency, and using hyper-links for navigating from one word to another.

With respect to similar systems (Barlow, 2002; Scott, 2004; Kilgarriff et al., 2004; Charest et al., 2007; Rayson, 2009; Fletcher, 2011), our system uniquely combines parallel concordancing with collocation detection based on deep syntactic processing. It is available for English, French, Spanish and Italian and it is being extended to other languages.

Acknowledgement

This work is partly supported by the Swiss National Science Foundation (grant no. PA00P1_131512).

References

Michael Barlow. 2002. Paraconc: Concordance software for multilingual parallel corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research*, pages 20–24, Las Palmas, Spain.

Simon Charest, Éric Brunelle, Jean Fontaine, and Bertrand Pelletier. 2007. Élaboration automatique d'un dictionnaire de cooccurrences grand public. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 283–292, Toulouse, France, June.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

William H. Fletcher. 2011. Phrases in english: Online database for the study of English words and phrases. `http://phrasesinenglish.org`. Accessed March, 2011.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, September.

Diana Lea and Moira Runcie, editors. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, Oxford.

Luka Nerima, Violeta Seretan, and Eric Wehrli. 2003. Creating a multilingual collocation dictionary from large text corpora. In *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 131–134, Budapest, Hungary.

Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Charles University in Prague.

Paul Rayson. 2009. Wmatrix: a web-based corpus processing environment. `http://ucrel.lancs.ac.uk/wmatrix`. Accessed March, 2011.

Mike Scott. 2004. *WordSmith Tools version 4*. Oxford University Press, Oxford.

Violeta Seretan. 2009. An integrated environment for extracting and translating collocations. In Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith, editors, *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, UK.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, Dordrecht.

Eric Wehrli, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 90–94, Athens, Greece. Association for Computational Linguistics.

Eric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic.