Book Reviews

Syntax-Based Collocation Extraction

Violeta Seretan
(University of Geneva)

Berlin: Springer (Text, speech and language technology series, volume 44), 2011,
xi+217 pp; hardbound, ISBN 978-94-007-0133-5, $139.00

Reviewed by
Pavel Pecina
Charles University in Prague and Dublin City University

Collocation is a common language phenomenon which has attracted the interest of researchers in many subfields of both theoretical and computational linguistics. Although there is no commonly accepted and precise definition of this phenomenon, collocations are generally understood as complex lexical items, often characterized as unpredictable, idiosyncratic, holistic, mutually selective, and so forth. Together with other types of *multiword expressions* (or phraseological units, such as compound nouns, phrasal verbs, idioms, etc.), collocations form a borderline phenomenon positioned between lexis and grammar: On one hand, they are unpredictable and must be learned in the same way as single words are (as whole units); on the other hand, they often also have internal syntactic structure and their components must then adhere to grammatical rules. Collocations play an important role in applications involving text production (e.g., machine translation and language generation), text analysis (e.g., parsing and word sense disambiguation), and also in other related tasks (such as information extraction, text classification, etc.).

The book *Syntax-Based Collocation Extraction* by Violeta Seretan is based on her doctoral dissertation defended in 2008 at the Department of Linguistics, University of Geneva, under the supervision of Eric Wehrli, and refers to a number of their previous publications. The main text is divided into six chapters (amounting to 128 pages) and six appendices (70 pages).

The first chapter can be regarded as a motivation for the whole work. It introduces the notion of collocation, explains its relevance (and importance) for natural language processing, specifies the aims of the work, and most importantly, it presents arguments for syntax-based collocation extraction as a more appropriate alternative to the traditional syntax-free *n*-gram and window-based techniques.

The second chapter focuses on various aspects of collocation, mainly from the theoretical point of view. It provides a very comprehensive (and readable) survey of numerous and heterogeneous definitions, descriptions, and discussions of this topic appearing in the literature in the past 90 years. The author (correctly) points out that although this phenomenon has an implicit linguistic aspect, the majority of definitions of collocation rather adopt a statistical view. But other perspectives—lexicographic, pedagogical, contextualist, and lexical-semantic—are very well addressed, too. Special attention is paid to the role of collocations in text cohesion, colligation, semantics, metaphoricity, and a lexis-grammar interface. The author also reviews various semantic and morpho-syntactic properties of collocations considered in the literature and attempts to define the "core" of the concept of collocation.

The third chapter focuses on practical issues of collocations, namely, methods for their automatic extraction from text corpora. The author first reviews a general extraction procedure, which consists of two steps: collocation candidate identification using specific criteria and candidate ranking with a given association measure. Both the extraction steps are discussed in detail, including linguistic pre-processing, construction of contingency tables, and application of association measures. A significant part of the chapter is devoted to the survey of the state of the art in this task. An exhaustive compendium of relevant works is practically organized by the languages that the experiments were carried out on.

The fourth chapter forms the core of the book. It presents the author's own contribution—a method of collocation extraction based on syntactic parsing. The author again advocates the need for syntax-based extraction methods (supported by several examples and citations from the literature) and reviews other relevant papers tackling this issue. The remainder of the chapter is devoted to the description of the proposed method (which identifies collocation candidates using the Fips parser and ranks them according to their log-likelihood ratio) and its evaluation in two scenarios. In the monolingual experiments, the extraction procedure is applied to a part of the Hansard corpus containing 1.2 million words in French. In the crosslingual case, the extraction procedure is simultaneously applied to French, English, Spanish, and Italian parts of the Europarl corpus—each containing about 3.8 million words. The proposed syntax-based method is compared against a standard window-based extraction technique (with a sliding window of five consecutive words). The evaluation is performed using manually judged candidates in terms of precision measured for the 500 top-ranked candidates (in Experiment 1) and 5 test sets each containing 50 contiguous items situated at different levels in the ranked list spanning the top 10% for each language (in Experiment 2). In all cases, the evaluation experiments showed that by applying a syntax-based candidate identification "a considerable improvement is obtained over the standard sliding window method" (p. 97). The results are presented in a very thorough way, including qualitative analysis and error analysis with lots of concrete examples.

The fifth chapter of the book contains three interesting extensions of the work presented in the previous chapter. First, the author applies a method based on the simple concept of "collocation of collocations" for inferring longer collocations from the binary combinations extracted. The second extension is a technique for data-driven induction of syntactic patterns which are adequate for identifying collocations. The candidate patterns (part-of-speech combinations) are ranked using the log-likelihood ratio and subsequently undergo a process of manual analysis. The third extension addressed is the automatic acquisition of collocation equivalents from parallel corpora. However, no thorough evaluation (or comparison with alternative approaches) of these methods is presented (e.g., a comparison of the translation equivalents extraction method with a simple look-up of equivalents in a translation phrase table of a phrase-based statistical machine translation system).

The sixth chapter reviews the main contribution of the research described in the book and sketches directions for the future work. The appendices contain an overview of published collocation dictionaries (Appendix A), a list of various definitions of a collocation (Appendix B), some mathematical notes on association measures, and detailed results of the experiments presented in Chapter 4: the monolingual evaluation experiment (Appendix D) and the crosslingual evaluation experiment (Appendix E).

The book is generally well written and comprehensively addresses both theoretical and applied work on collocations and their extraction from text corpora. It does not require any special previous experience in the field. Most of the technical passages

(e.g., in Chapter 3) are presented in an intuitive and self-contained way, suitable even for non-expert users. However, certain parts (mostly in the experimental chapters) lack sufficient detail and/or are somewhat arguable. For example, the author does not provide (or refer to) any evaluation of the parser's accuracy and (more importantly) it is not clear how the accuracy can affect the quality of the collocation extraction procedure. From the figures in the book, one can guess that the parser provides a complete parse for about 50% of the sentences but the accuracy on syntactic pairs is unknown. Further, it seems that the parser features an internal collocation dictionary and collocations detected by this dictionary automatically receive (by convention) maximum association score by default. It is not very clear whether those cases are also used in the evaluation and comparison with the baseline (window-based) method or not. If they are, the comparison might be unfair; if they are not, we would lose information on how these cases would be ranked. In any case, the evaluation is probably biased by this dictionary, although the author does not report on its size and coverage of the corpora and how the dictionaries differ for the different languages used.

The decision on the window size in the baseline window-based method seems arbitrary. This hyper-parameter was set to five, but this choice was not supported by any discussion or evidence (e.g,. distribution of the surface distance of collocation components) although it has a substantial effect on performance of the method. The window-based technique is known to pollute the list of candidates with noise (words without any syntactic relation), which decreases precision. The author reports that the syntax-based method outperforms the baseline quite substantially. The difference, however, can be probably increased by extending the window to six or even more words. On the other hand, the window can also be shortened and the difference in precision might diminish or even be reversed. The composition of the evaluation data raises a concern too: The items which were not agreed upon by the judges were excluded from the evaluation. There were not that many such cases, but this approach (ignoring borderline cases) is not very rigorous.

Some other details of the experiments are either missing or not very clear, either. For example, it is not mentioned which test of statistical significance was used in the experiments, whether sampling of the evaluation data was really stratified, or what is meant by "partitioning the candidate data into syntactically homogeneous classes" and how it affected evaluation.

Overall, *Syntax-Based Collocation Extraction* is an interesting book and will certainly be appreciated by researchers interested in collocations and related phenomena. There is no doubt that collocation extraction should be based on syntactic preprocessing of the source corpora (simply because collocations often have syntactic structure), but the evaluation presented in the book is not very convincing. However, the parts surveying the theoretical aspects of collocations and describing the current state of the art and related work can easily serve well as a handbook for students entering the field of collocation extraction.

Pavel Pecina is a postdoctoral researcher at the Institute of Formal and Applied Linguistics, Charles University in Prague, and Centre for Next Generation Localisation, Dublin City University. His research interests include machine translation, lexical association measures, syntax-driven information retrieval, speech retrieval, and machine learning in natural language processing in general. Pecina's address is Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Malostranské náměstí 25, 118 00 Prague 1, Czech Republic; e-mail: pecina@ufal.mff.cuni.cz, URL: http://ufal.mff.cuni.cz/∼pecina/.