

# A Lexical Semantic and Statistical Approach to Lexical Collocation Extraction for Natural Language Generation

Rita McCardell Doerr

Lexical collocations are frequently occurring word pairs in natural language whose presence are not always predictable by their usage. These collocations are used by native speakers of a language almost without thought, yet they must be learned by nonnative speakers of the language. A native speaker of English might say that he/she drinks "strong coffee," but a nonnative speaker might say either "powerful coffee" or "sturdy coffee." Collocations tend to vary among languages and topic domains. Unfortunately, the task of correctly identifying lexical collocations, even by native speakers of the language, has been shown to be difficult.

Computer systems that translate natural languages, or machine-translation systems, need to know about lexical collocation information to produce natural-sounding or colloquially proper text. Natural language generation is a component of a machine-translation system that automatically produces natural-sounding text in a particular language given a language-independent meaning as input. This dissertation (Doerr 1994)<sup>1</sup> demonstrates how to automatically locate and extract lexical collocations from machine-readable text for use within a machine-translation system's natural language-generation component.

A lexical-semantic and statistical approach is adopted for the location and extraction of lexical collocations. For this approach, a computational definition is provided for lexical collocations that demonstrates that (1) they occur as adjacent word pairs, (2) they occur more often than would be expected by chance, and (3) they

comprise words for which neither word can be substituted by a synonym or hyponym. Potential collocations comprising certain adjacent part-of-speech tags are extracted from text. An online thesaurus and a lexical database of word classes are queried for synonyms and hyponyms, respectively, for each potential collocation. These queries create potential challenger pairs, such as "strong java" and "powerful coffee." A substitution procedure is then applied to determine if any of these challenging word pairs occur more frequently than the potential collocation.

The VERIFY lexical collocation-extraction system has been implemented incorporating these ideas. Results to date have been positive; a system using lexical-semantic knowledge, that is, synonymy and hyponymy, within a lexical collocation-extraction system outperforms a system using purely statistical knowledge. To compare system output to human judgments of training data, a training component was also incorporated into VERIFY. This component is able to adapt to new data. Overall system performance, measured in recall and precision scores,<sup>2</sup> improved using this component. To provide a more flexible system given a user's application, a weighting mechanism was used to produce a range of recall and precision scores. These weights can be adjusted to optimize system performance.

Through an experimental procedure, VERIFY was trained on a set of human-analyzed data and then independently tested on another set of

data. Testing results showed that VERIFY obtained 88 percent (the recall score) of those collocations in the test set that were identified by the humans. Additionally, the overall performance quality of the VERIFY system was evaluated by comparing it to a different set of human subjects ("testers"). VERIFY's performance exceeded the testers' performance on a different data set given a 99-percent confidence interval containing the true population scores for these humans: [43%,53%] for recall and [41%,51%] for precision.

The use of lexical-semantic knowledge has advanced the state of the art for lexical collocation extraction beyond traditional statistical approaches. Incorporation of a training component within an extraction system provides the capability of adapting to any changes within the data. Controlling overall system performance through the use of a weighting mechanism provides flexibility to the user of an extraction system. In an experiment to compare VERIFY's performance to that of human performance on a particular set of data, VERIFY outperformed humans in both recall and precision.

## Notes

1. A copy of this dissertation is available as technical report TR CS-94-12 from the Computer Science and Electrical Engineering Department, University of Maryland Baltimore County, Catonsville, MD 21228.
2. *Recall* is the percentage of correct answers that the system found among the model answers, that is, human judgments. *Precision* is the percentage of correct answers among the total number of answers that the system reported.

## Reference

Doerr, R. M. 1994. A Lexical-Semantic and Statistical Approach to Lexical Collocation Extraction for Natural Language Generation. Ph.D. diss., Dept. of Computer Science, Univ. of Maryland Baltimore County.

Rita McCardell Doerr is a U.S. Department of Defense senior computer scientist and is currently the technical director for the Software Engineering and Support Division at a U.S. government site in Germany.