Introducing Linggle: From Concordance to Linguistic Search Engine

Jason S. Chang
Department of Computer Science
National Tsing Hua University
son.jschang@gmail.com

Abstract

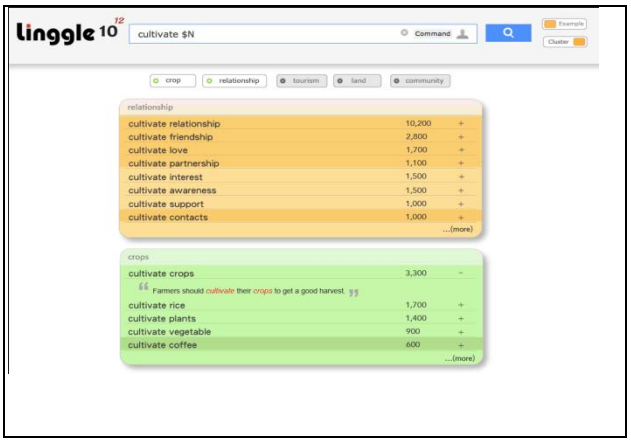
We introduce a Web-scale linguistics search engine, *Linggle*, that retrieves lexical bundles in response to a given query. Unlike a typical concordance, *Linggle* accepts queries with keywords, wildcard, wild part of speech (PoS), synonymous words, and additional regular expression (RE) operators, and returns bundles with frequency counts. In our approach, we argument Google Web 1T corpus with inverted file indexing, PoS information from BNC, and semantic indexing based on Latent Dirichlet Allocation. The method involves parsing the query to transforming the query to several keyword retrieval commands, retrieving word chunks with counts, filtering the chunks again the query as a RE, and finally displaying the results according the count, similarity, and topic. Clusters of synonymous or conceptually related words are also provided. In addition, *Linggle* provide example sentences from *The New York Times* on demand. The current implementation of *Linggle* is the most comprehensive functionally, and is in principle language and dataset independent. We plan to extend *Linggle* to provide a fast and convenient access to a wealth of linguistic information embodied in Web scale datasets including *Google Web 1T* and *Google Books Ngram* for many major languages in the World.

For non-native speakers, doubts concerning the usage of a preposition, the mandatory presence of a determiner, the correctness of the association of a verb with an object or the need for synonyms of a term in a given context are problems that arise frequently when writing in English. Printed collocation dictionaries and reference tools based on compiled corpora offer limited coverage of word usage while knowledge of collocations is vital for the competent use of a language. We propose to address these limitations with a comprehensive system that truly aims at letting learners "know a word by the company it keeps". *Linggle* (**linggle.com**) is a broad coverage language reference tool for English as Second Language learners (ESL). The system is designed to access words in context under various forms.

First, we build inverted file index for the *Google Web 1T Ngram* to support queries with RE-like patterns including PoS and synonym matches. For example, for the query "$V $D +important role", *Linggle* retrieve 4-gram chunks that start with a

verb and a determiner followed by a *important* synonym and the keyword *role* (e. g., *play a key part* 15,900). A natural language interface is also available for users that would be less familiar to pattern based search. For example the question "*How can I describe a beach?"* would retrieve two word chunks with count such as "*sandy beach* 413,300" and "*rocky beach* 16,800". The n-gram search implementation is achieved through filtering, re-indexing, and populating Web 1T ngram in a HBase database and augmenting them with the most frequent PoS for words (without disambiguation) derived from the British National Corpus.

The n-grams resulting from the queries can then be linked to examples extracted from the New York Times Corpus in order to provide full sentential context for more effective learning. In some situations, users might need to search for words in a specific syntactic relation (i. e., *collocates*). Let's consider the example "absorb $N" that queries all the objects of the verb *absorb*. In this case, grouping the words that belong to similar domains together offers a better overview of the usage of the verb than a list of objects ordered by frequency. For example the verb *absorb* takes clusters of objects related to the topic *liquid/energy*, but also to the topics *money*, *knowledge* or *population*.



This tendency of predicates to prefer certain classes is defined by Wilks (1978) as selectional preference and widely reported in the literature. *Linggle* proposes *preferred* clusters of synonymous query arguments of adjectives, nouns and verbs. The clustering is achieved by building on Lin and Pantel (2002)'s large-scale repository of dependencies and word similarity scores and on an existing method for selectional preference induction with a Latent Dirichlet Allocation (LDA) model.

References

Chang, Jason. S. 2008. Linggle: a web-scale language reference search engine. Unpublished manuscript.

Fletcher, William H. 2012. Corpus analysis of the world wide web." In *The Encyclopedia of Applied Linguistics*.

Kilgarriff, Adam, and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of COLLOCTION: Computational Extraction, Analysis and Exploitation workshop*, pp. 32-38.

Kilgarriff, Adam. 2007. Googleology is bad science." *Computational linguistics* 33(1), pp. 147-151.

Lin, Dekang, and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of COLING*.

Lin, Dekang, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil et al. 2010. "New tools for web-scale n-grams." *Proceedings of LREC*.

Potthast, Martin, Martin Trenkmann, and Benno Stein. Using Web N-Grams to Help Second-Language Speakers. 2010. In *Proceedings of SIGIR Web N-Gram Workshop*, pages 49-49.

Wu, Shaoqun, Ian H. Witten, and Margaret Franken. 2010. Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge." *ReCALL* 22(1), pp. 83-102.

Wilks, Yorick. 1978. Making preferences more active. *Artificial Intelligence* 11(3), pp. 197-223.