Abstract

Near-synonyms are words that have very similar meanings, but cannot generally be substituted by each other in a text. These words can differ in register (*drunk*, *inebriated*), affection (*dad*, *father*), attitude (*stingy*, *thrifty*), grammatical usage (*ajar*, *open*) or collocations (*task*, *job*), among others. Choosing the appropriate near-synonym can be trivial for a native speaker, but not necessarily for L2 learners or for automatic systems. This thesis is mainly focused on modelling the register, attitudinal and collocational dimensions of difference across near-synonyms. Formality and attitude scores for words are induced from their co-occurrence similarity with manually chosen word seeds, and collocational preferences of a word are extracted based on their statistical association. The power of this information is tested in a lexical choice task, the Fill-in-the-Blanks task. In addition, formality information is used to build an automatic paraphraser based on word and phrase substitution that intends to paraphrase sentences changing their level of formality. Results show that, whereas it is possible to obtain valuable information for formality and collocations even with a limited amount of text, this is not the case for attitude. With respect to the lexical choice task, our system does not beat a language model, but it does gain a small but significant improvement over a majority class baseline when combined with frequency information. A manual evaluation of the paraphraser's performance shows that, while it is still far from the quality of human-created paraphrases in terms of naturalness and grammaticality, its performance in terms of formality change is not significantly different.