A New Tool for Quality Control of Text Corpora

Zygmunt Vetulani, Marta Witkowska
Adam Mickiewicz University in Poznań
Poland
vetulani@amu.edu.pl, martusiazielinska@gmail.com

Umut Canbolat
University of Kocaeli
Turkey
u.canbolat@yahoo.com

Keywords: text corpora, corpora quality control, lexical saturation tests

In the demo we present a new tool designed for quality control of text corpora as well as for extraction of subcorpora of desired size with lexical saturation properties. The tool allows to evaluate lexical saturation of a text corpus, where by lexical saturation of a corpus we mean that it is hard to find new lexemes outside this corpus (cf. e.g. Kittredge, 1983, also Vetulani, 1989) Estimation of the saturation degree for a given corpus contributes in a natural way to the corpus quality evaluation. Although the first application of the tool is the evaluation of the lexical coverage of corpora, it may be equally useful to study corpora representativeness for various phenomena, and their usefulness for corpora-based research, both theoretical and practical. It may be useful for cost evaluation of engineering tasks in language competence modelling for AI purposes as well as in literary research (study in progress). The presented system TSCC (Text SubCorpora Creator) is highly language independent, i.e. it may be applied directly or easily adapted to any language in which the text units may be represented in alphabetic scripts. TSCC is derived from our earlier application OCASSC (Opinion Corpora Acquisition Software for Subcorpora Creation) (Vetulani et al., 2015) tested on a corpus of booking.com hotel guests' opinions. We intend to free distribute the TSCC v.1. for beta testing.

TSCC

Recently (2017, summer) we have designed and implemented the system TSCC operating in open (unformatted) text as a tool for extracting from a large corpus its subcorpora of any desired size. In situations where the representativeness of the corpus is closely related to its lexical completeness, evaluation of the degree of lexical saturation using TSCC may help to fix the stop criterion for creation subcorpora with desired properties (appropriate lexical/conceptual coverage).

Experiments

TSCC generates data necessary to draw the saturation graphs. Figure 1 presents the graph for a corpus composed of 8920 lines randomly extracted from a larger corpus of hotel opinions (Vetulani et al., 2015). Although the corpus appears not large enough to be considered as lexically saturated because the increase of numbers of adjectives is quasi linear after the first 4000 lines, still this increase is not high and may be acceptable for many purposes as its local speed equals only 12 words per 1000 text lines (see Fig. 1).

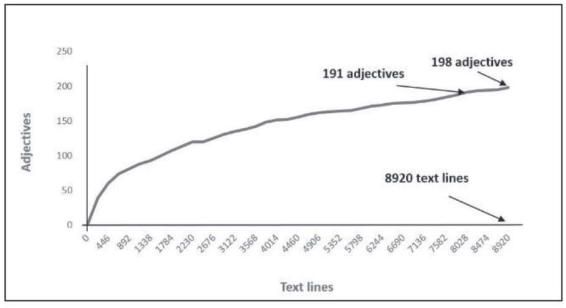

Figure 1: Saturation graph representing increase of number of adjectives observed in the initial segments of the hotel opinions corpus obtained using TSCC

Further research

We intend to further develop the TSCC system from the point of view of literary research. In particular its utility will be beta tested in the research on vocabulary structure of particular authors and particular literary works. We hope to prove its utility for stylometry (cf. Muller 1975). The beta prototype of the TSCC tool will be released for LREC 2018 and distributed under an open license.

References

Kittredge, R. (1983). Semantic processing of texts in restricted sublanguage. *Computers & Mathematics with Applications,* Vol. 9, Issue 1, pp. 45–58.

Muller, Ch. (1975). Peut-on estimer l'étendue d'un lexique? *Cahiers de Lexicologie,* no 27,1975-II, pp. 3-29.

Vetulani, Z. (1989). *Linguistic problems in the theory of man-machine communication in natural language. A study of consultative question-answering dialogues. Empirical approach.* Brockmeyer, Bochum.

Vetulani, Z., Witkowska, M. and Canbolat, U. (in preparation, 2017). "TSCC: a New Tool to Analyze Lexical Saturation of Text Corpora".

Vetulani, Z., Witkowska, M. and Menken, S. (2015). "Corpus Based Studies on Language Expression of Opinions". *LTC 2015 Proceedings*, Poznań.