Finding infrequent phenomena in large corpora using distributional semantics

Maria Skeppstedt[a,b,*], Carita Paradis[c], Andreas Kerren[a], Magnus Sahlgren[b]

[a]*Computer Science Department, Linnaeus University, Växjö, Sweden*

[b]*Gavagai AB, Stockholm, Sweden*

[c]*Centre for Languages and Literature, Lund University, Lund, Sweden*

[*]*Corresponding author: maria@gavagai.se*

The abundance of text published on the Internet makes it possible to harvest a large number of samples of linguistic phenomena, also of phenomena that are less frequently occurring. This is useful for qualitative and quantitative linguistic studies, as well as for training machine-learning models for automatic detection of these phenomena.

We investigate cases in which the speaker expresses personal feelings and opinions, i.e., the phenomenon known as stance taking [1]. We have divided the expression of stance into a number of sub-classes, which has the effect that each sub-class becomes an example of a(n at least moderately) infrequently occurring phenomenon. Initial annotation experiments of 100 blog utterances showed, for instance, that the stance class *Concession and Contrariness* occurred in 10% of the utterances, *Uncertainty* in 7% and *Hypotheticals* in 6%. Although we limit the study to instances of stance that are explicitly expressed in the text, e.g., 'perhaps' signalling *Uncertainty* and 'even though' signalling *Concession and Contrariness*, there is no exhaustive list of such constructions and thereby no simple method for automatically harvesting a large number of representative samples. Manual annotation for harvesting samples of infrequent phenomena is, however, a time-consuming task, as large amounts of text must be scanned.

A method used in the machine-learning field for minimising the amount of annotated text required for training a model is to apply active learning for selecting data to annotate. Most active learning systems applied on textual data actively select the samples that are most informative — i.e., most useful — for the machine-learning model that is to be trained [2, pp. 25–28]. Using such sampling based on informativeness reduces the amount of training data required, and, thereby also the amount of data that needs to be manually labelled. Active learning has been shown to reduce classifier problems stemming from unbalanced data, i.e. when one or several of the classes that are to be detected are less frequently occurring than others in the data studied. For larger imbalances, however, there is a risk that not enough samples of the minority class are ever selected for the standard approach of active learning to perform effectively, leading to a lower performance than random sampling [3].

An alternative active learning sampling approach is to use inherent properties of the data that are to be classified for actively selecting suitable training samples [4]. We apply this technique by using distributional semantics properties of words and constructions, derived from semantic models built on very large text corpora, in which the similarity of pairs of words/word constructions is measured according to how often they occur in similar contexts [5]. This distributional semantics information could, for instance, be incorporated in the active learning process by: (a) from a limited set of manually annotated utterances, automatically extract words and constructions that are typical of the minority classes of interest—e.g. "probably" for the class Uncertainty [6]—, (b) select new utterances to annotate using the criterion of whether they contain words and constructions that are distributionally similar to those that have been extracted as typical of the minority classes—e.g. "presumably" is similar to "probably"[1]—.

Selection of utterances to annotate could then, in addition to being based on whether they are deemed as informative by the trained machine learning classifier, be based on the distributional properties of included words and constructions. We hypothesise that such a mixed approach will boost the selection of training samples from the minority class, thereby being more successful for active learning on unbalanced data than an approach based solely on classifier informativeness.

References

[1] D. Biber, Stance in spoken and written university registers, Journal of English for Academic Purposes 5 (2006) 97–116.

[2] F. Olsson, Bootstrapping Named Entity Annotation by Means of Active Machine Learning, Ph.D. thesis, University of Gothenburg. Faculty of Arts, 2008.

[3] J. Attenberg, S. Ertekin, Active learning for imbalanced learning, in: H. He, Y. Ma (Eds.), Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley-IEEE, 2013

[4] S.-j. Huang, R. Jin, Z.-h. Zhou, Active learning by querying informative and representative examples, in: J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems 23, Curran Associates, Inc., 2010, pp. 892–900.

[5] M. Sahlgren, The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, Ph.D. thesis, Stockholm University, 2006.

[6] M. Skeppstedt, T. Schamp-Bjerede, M. Sahlgren, C. Paradis, A. Kerren, Detecting speculations, contrasts and conditionals in consumer reviews, in: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Lisboa, Portugal, 2015, pp. 162–168.

[1] http://lexicon.gavagai.se