Word spaces as input to categorisation of attitude

Jussi Karlgren
Swedish Institute of Computer Science
Box 1263
SE-164 29 Kista
Sweden
jussi@sics.se

ABSTRACT

SICS starting points are that given a semantic word space trained on general purpose text, where distance and nearness are measures of semantic similarity, we can represent sentences by the *centroid* of the words that occur in it, that *constructional features* contribute to the organisation of this semantic space, and *attitude is a semantic dimension of variation* in that sentences with similar attitudinal qualities can be expected to occupy space in the vicinity of each other.

This year's simplistic experiment did not yield useful results. Parameter tuning is a necessary step in any categorisation excercise; this year we failed to devote the necessary effort to achieve results worth noting.

Word spaces for opinion analysis

This paper describes briefly the SICS attempt to participate in NTCIR-8 [6]. We have in previous experiments, among them ones performed in NTCIR-7, successfully used constructional features in conjunction with lexical features in a word space, achieving high recall for attitudinal utterances even in cases where the lexical features alone would have yield equivocal evidence on the utterance character[3, 2].

Our approach takes as its starting point the observation that lexical resources always are noisy, out of date, and most often suffer simultaneously from being both too specific and too general. Not only are lexical resources inherently somewhat unreliable or costly to maintain, but they do not cover all the possibilites of expression afforded by human linguistic behaviour: we believe that attitudinal expression in text is not solely a lexical issue. For our present experiments reported here no attitudinal lexical resources were used — only general purpose linguistic analysis was employed to establish the constructions used in the further processes.

A basis for our approach is the *Word Space Model*[5, 4], a data structure based on a general multi-dimensional vector space model, where distance and nearness are used as estimates of semantic similarity, where those distances are computed from distributional data collected from sizeable amounts of general purpose text, and where computation of similarity is made using geometric computations in a multi-dimensional space.

Our starting points are that given a word space to represent semantic relations between terms, we represent sentences by the *centroid* of the words that occur in it but we add *constructional features* to contribute to the organisation of this semantic space, and posit that *attitude is a semantic dimension of variation* in that sentences with similar attitudinal qualities can be expected to occupy space in the vicinity of each other. This worked well for NTCIR-7, given that we put some fair effort into parameter tuning and selecting the most appropriate background text collection.

NTCIR 8 MOAT experiment

This year's experiment was performed as simply as possible, without new parameter tuning, as a simplified version of the more successful experiment performed the year before. This proved insufficient — we were not able to regain the same level of accuracy as we did in previous and other similar experiments[2] where we put more time into tuning the mechanisms for the corpus at hand.

1. We built a background semantic word space using random indexing from several years of newsprint material.
2. We transformed both the training set and the test set by surface syntactic analysis as described in our previous reports, including the attitude tag for the training set.
3. We projected the training and the test set, sentence by sentence, into the background space.
4. We exported the context vectors of the centroids of the training and test sets.
5. We used LIBLINEAR[1] to categorise the test set based on the training set.

Results

Our results were decidedly underwhelming. We only achieved a precision of 14 per cent, a recall of 31 percent, yielding a F1-score of 20 per cent. We expected our simplistic method to be less precise than most but did expect a better recall than given here.

Conclusions

We know from our previous experiments that choice of background model makes a difference. Also, this year, we did not experiment with different models for the semantic space — we used the same parameter settings as we did the previous year. We have in the meanwhile achieved a fair understanding of the parameter space and how it varies across data sets — in the future we will devote more effort to tuning the right process steps appropriately.

Acknowledgments

This work was supported by the Swedish Research Council.

1. REFERENCES

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[2] J. Karlgren, G. Eriksson, M. Sahlgren, and O. Täckström. Between bags and trees - constructional patterns in text used for attitude identification. In *Proceedings of ECIR 2010, 32nd European Conference on Information Retrieval*, Milton Keynes, UK, 2010.

[3] J. Karlgren, G. Eriksson, and O. Täckström. Sics at ntcir-7 moat: constructions represented in parallel with lexical items. In *Proceedings of The 7th NTCIR Workshop (2007/2008) - Evaluation of Information Access Technologies*, page 4, Tokyo, Japan, 2008.

[4] M. Sahlgren. *The Word-Space Model*. PhD thesis, Stockholm University, Department of Linguistics, 2006.

[5] H. Schütze. Word space. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann Publishers, 1993.

[6] Y. Seki, L.-W. Ku, L. Sun, H.-H. Chen, and N. Kando. Overview of multilingual opinion analysis task at ntcir-8 - a step toward cross lingual opinion analysis. In *Eighth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Japan, June 2010. NII.