Semantic Space Models for Profiling Reputation of Corporate Entities

Notebook for RepLab at CLEF 2013

Jussi Karlgren, Linus Ericsson

Gavagai
Skånegatan 97, 116 35 Stockholm
www.gavagai.se
jussi@gavagai.se

Abstract Gavagai used its commercially available system for the filtering and polarity tasks in the evaluation campaign for online reputation management systems at CLEF 2013. The system is built for large scale analysis of streaming text and as part of the services Gavagai provides, it measures the public attitude visavi targets of interest. This mechanism — with no adjustment for this specific task — was used for polarisation and the experiments performed this year was to test a number of settings for testing how an attitude might be learned from the data rather than given by editorial intervention.

1 Semantic spaces

Gavagai provides through its Ethersource suite of services tools for monitoring *targets of interest* for some commercial purpose in streaming data of any scale and editorial quality in any language with respect to *semantic poles* of some permanence. Ethersource is based on distributional semantics [7] represented in a semantic space [6], and realised through a proprietary implementation of the Random Indexing processing framework [2] as described in our position paper at the recent Online Reputation Management workshop [5]. Ethersource is under constant development and the results from this evaluation are being fed back into the system quality assurance cyle.

A target in Ethersource is defined through manual entry of a number of representative terms. Here, the targets were given through the entity description by the query, entity name, and category strings.

2 Poles and polarisation

A semantic pole in Ethersource is likewise defined through a larger and more permanently selected number of terms. This term set can be extensive or limited, depending on if recall or precision is crucial for the task at hand and if typical expression of this pole is wide-ranging or more exact [8]. For typical sentiment analysis purposes, the poles can be defined through a list of positive and negative terms; for other purposes other word lists can be used — in our commercial context we have a large number of poles and

do not generalise to simple positive or negative [3]. For this task, we utilised Gavagai's canonical poles for positive and negative sentiment for English and Spanish, each of some few hundred editorially selected terms, semi-automatically augmented through the semantic space model built from previous large scale analysis of streaming text and static textual collections in each language. This year, we did not use consumer statisfaction poles which yielded very useful results last year [4].

The polarisation score is normally aggregated by the Ethersource system over streaming data into a time series and monitored by our customers for change, varies between 0 and 1 and is not designed to make decisions for text items in isolation from their context. In this case these scores were used in some of the experimental settings given below.

3 Real life experimental data

The profiling task was defined to be based on real data, using a set of microblog posts from Twitter filtered to contain an entity name. There were four tasks defined for this year's Replab [1]. We participated in the filtering task without any mechanism of note, using simple keyword spotting, and we did not participate in the topic detection task.

4 Settings and submissions

We submitted five runs for the polarisation task. One (GAVKTH 2) was based on the dimensionally reduced lexical space used as the starting point for our learning process, one (GAVKTH 6) on the standard poles described above without any learning component, three (GAVKTH 3, 4, and 7) were based on regions in the learned semantic space. The run based on poles (GAVKTH 6) was of no relevance for the priority to task, but the other four were used for the priority task also.

The pole-based run (GAVKTH 6) output a polarisation score for the two semantic poles; the raw index space run (GAVKTH 2) and the semantic space runs (GAVKTH 3, 4, and 7) output a position in a two-thousand dimensional space. The index vector run made no use of the space itself and is a close approximation to a bag-of-words experiment; the other runs made use of the semantic space we have trained on millions of other documents for English. The Spanish semantic space is not as well trained, and as a comparison we submitted one run which combined the English semantic space with the Spanish raw index space.

GAVKTH 2 raw random index vectors
GAVKTH 3 position in semantic space
GAVKTH 4 position in semantic space with frequent terms filtered out
GAVKTH 6 scores from position with respect to positive and negative lexical poles in semantic space
GAVKTH 7 combination of approach 3 for English, 2 for Spanish

5 Results for priority task

Run	Reliability	Sensitivity
GAVKTH 2	0.36	0.19
GAVKTH 3	0.67	≈ 0.0
GAVKTH 4	0.49	0.02
GAVKTH 7	0.37	0.09

This task did not benefit from the semantic space.

6 Results for polarity task

Run	Reliability	Sensitivity
GAVKTH 2	0.37	0.21
GAVKTH 3	0.54	0.10
GAVKTH 4	0.50	0.10
GAVKTH 6	0.30	0.21
GAVKTH 7	0.41	0.18

These runs show consistently mediocre sensitivity scores but for some of them quite useful reliability.

7 Analysis

This sort of analysis is by necessity very subjective, and many of the target posts in this experiment might arguably be interpreted to be neutral, negative, or positive depending on one's perspective. The commercial services provided by Gavagai hinge on aggregation of large numbers of tweets rather than searching for individual items in the text stream: in commercial application of our system we work on time series and sequences rather than single posts. For our purposes we find that the semantic space boosts results quite well (GAVKTH 3 $>>$ GAVKTH 2), that filtering out frequent terms is not worth the trouble (GAVKTH 3 $>$ GAVKTH 4) and that retreating to raw terms rather than index vectors for material with little training is necessary (GAVKTH 3 $>$ GAVKTH 7). Also, most interestingly, these results give us reason to think hard about the editorial effort put into building semantic poles, since the the reliability of the semantic space results are higher than those from the polarisation results (GAVKTH 3 $>$ GAVKTH 6).

References

1. Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: Fourth International Conference of the CLEF initiative, CLEF 2013. Springer LNCS (Sep 2013)
2. Kanerva, P.: Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive Computation 1(2), 139–159 (2009)

3. Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., , Hamfors, O.: Usefulness of sentiment analysis. In: ECIR 2012, 34th European Conference on Information Retrieval. Barcelona (2012)
4. Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., Hamfors, O.: Profiling reputation of corporate entities in semantic space : Notebook for replab at clef 2012. In: CLEF 2012 Evaluation Labs and Workshop Online Working Notes (2012)
5. Olsson, F., Karlgren, J., Sahlgren, M., Espinoza, F., Hamfors, O.: Technical requirements for knowledge representation for attitude mining on a realistic scale. In: Proceedings of the Workshop on Reputation Management in Social Media at LREC'12. Istanbul (2012)
6. Sahlgren, M.: 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), http://soda.swedish-ict.se/437/
7. Sahlgren, M.: The distributional hypothesis. Rivista di Linguistica (Italian Journal of Linguistics) 20(1), 33–53 (2008), http://soda.swedish-ict.se/3941/
8. Sahlgren, M., Karlgren, J., Eriksson, G.: SICS: Valence annotation based on seeds in word space. In: Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics (2007), http://soda.swedish-ict.se/2593/