SICS: Valence annotation based on seeds in word space

Magnus Sahlgren
SICS
Box 1263
SE-164 29 Kista
Sweden
mange@sics.se

Jussi Karlgren
SICS
Box 1263
SE-164 29 Kista
Sweden
jussi@sics.se

Gunnar Eriksson
SICS
Box 1263
SE-164 29 Kista
Sweden
guer@sics.se

Abstract

This paper reports on a experiment to identify the emotional loading (the "valence") of news headlines. The experiment reported is based on a resource-thrifty approach for valence annotation based on a word-space model and a set of seed words. The model was trained on newsprint, and valence was computed using proximity to one of two manually defined points in a high-dimensional word space — one representing positive valence, the other representing negative valence. By projecting each headline into this space, choosing as valence the similarity score to the point that was closer to the headline, the experiment provided results with high recall of negative or positive headlines. These results show that working without a high-coverage lexicon is a viable approach to content analysis of textual data.

1 The Semeval task

This a report of an experiment proposed as the "Affective Text" task of the 4th international Workshop on Semantic Evaluation (SemEval) to determine whether news headlines are loaded with preeminently positive or negative emotion or *valence*. An example of a test headline can be:

DISCOVERED BOYS BRING SHOCK, JOY

2 Working without a lexicon

Our approach takes as its starting point the observation that lexical resources always are noisy, out of date, and most often suffer simultaneously from being both too specific and too general. For our experiments, our only lexical resource consists of a list of eight positive words and eight negative words, as shown below in Table 1. We use a medium-sized corpus of general newsprint to build a general *word space*, and use our minimal lexical resource to orient ourselves in it.

3 Word space

A word space is a high-dimensional vector space built from distributional statistics (Schütze, 1993; Sahlgren, 2006), in which each word in the vocabulary is represented as a *context vector* $\vec{v}$ of occurrence frequencies: $\vec{v_i} = [f_j, \cdots, f_n]$ where f is the frequency of word i in some context j.

The point of this representation is that semantic similarity between words can be computed using vector similarity measures. Thus, the similarity in meaning between the words w_1 and w_2 can be quantified by computing the similarity between their respective context vectors: $\text{sim}(w_1, w_2) \approx \text{sim}(\vec{v_1}, \vec{v_2})$.

The semantics of such a word space are determined by the data from which the occurrence information has been collected. Since the data set in the SemEval Affective Text task consists of news headlines, a relevant word space should be produced from topically and stylistically similar texts, such as newswire documents. For this reason, we trained our model on a corpus of English-language newsprint which is available for experimentation for participants in the Cross Language Evaluation Fo-

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 296–299,
Prague, June 2007. ©2007 Association for Computational Linguistics

rum (CLEF).[1] The corpus consists of some 100 000 newswire documents from Los Angeles Times for the year 1994. We presume any similarly sized collection of newsprint would produce similar results. We lemmatized the data using tools from Connexor,[2] and removed stop words, leaving some 28 million words with a vocabulary of approximately 300 000 words. Since the data for the affective task only consisted of news headlines, we treated each headline in the LA times corpus as a separate document, thus doubling the number of documents in the data.

For harvesting occurrence information, we used documents as contexts and standard tfidf-weighting of frequencies, resulting in a 220 220-dimensional word space. No dimensionality reduction was used.

4 Seeds

In order to construct valence vectors, we used a set of manually selected seed words (8 positive and 8 negative words), shown in Table 1. These words were chosen (subjectively) to represent typical expression of positive or negative attitude in news texts. The size of the seed set was determined by a number of initial experiments on the development data, where we varied the size of the seed sets from these 8 words to some 700 words in each set (using the WordNet Affect hierarchy (Strapparava and Valitutti, 2004)).

As comparison, Turney and Littman (2003) used seed sets consisting of 7 words in their word valence annotation experiments, while Turney (2002) used minimal seed sets consisting of only one positive and one negative word ("excellent" and "poor") in his experiments on review classification. Such minimal seed sets of antonym pairs are not possible to use in the present experiment because they are often nearest neighbors to each other in the word space. Also, it is difficult to find such clear paradigm words for the newswire domain.

The seed words were used to postulate one positive and one negative point (i.e. vector) in the word space by simply taking the centroid of the seed word points: $\vec{v}_S = \sum \vec{v}_{w \in S}$ where S is one of the seed sets, and w is a word in this set.

[1] http://www.clef-campaign.org/

[2] http://www.conexor.fi/

Positive	Negative
positive	negative
good	bad
win	defeat
success	disaster
peace	war
happy	sad
healthy	sick
safe	dangerous

Table 1: The seed words used to create valence vectors.

5 Syntagmatic vs paradigmatic relations

Our hypothesis is that words carrying most of the valence in news headlines in the experimental test set are *syntagmatically* rather than paradigmatically related to the kind of very general words used in our seed set.[3] As an example, consider test headline 501:

TWO HUSSEIN ALLIES ARE HANGED, IRAQI OFFICIAL SAYS.

It seems reasonable to believe that this headline should be annotated with a negative valence, and that the desicive word in this case is "hanged." Obviously, "hanged" has no paradigmatic neighbors (e.g. synonyms, antonyms or other 'nyms) among the seed words. However, it is likely that "hanged" will co-occur with (and therefore have a syntagmatic relation to) general negative terms such as "dangerous" and maybe "war." In fact, in this example headline, the most negatively associated words are probably "Hussein" and "Iraqi," which often co-occur with general negative terms such as "war" and "dangerous" in newswire text.

To produce a word space that contains predominantely syntagmatic relations, we built the distributional relations using entire documents as contexts (i.e. each dimension in the word space corresponds to a document in the data). If we would have used words as contexts instead, we would have ended up with a paradigmatic word space.[4]

[3] Syntagmatic relations hold between co-occurring words, while paradigmatic relations hold between words that do not co-occur, but that occur with the same *other* words.

[4] See Sahlgren (2006) for an explanation of how the choice of contexts determines the semantic content of the word space.

6 Compositionality and semantic relations

The relations between words in headlines were modeled using the most simple operation conceivable: we simply add all words' context vectors to a compound headline vector and use that as the representation of the headline: $\vec{v}_H = \sum \vec{v}_{w \in H}$ where H is a test headline, and w is a word in this headline.

This is obviously a daring, if not foolhardy, approach to modelling syntactic structure, compositional semantics, and all types of intra-sentential semantic dependencies. It can fairly be expected to be improved upon through an appropriate finer-grained analysis of word presence, adjacency and syntactic relationships. However, this approach is similar to that taken by most search engines in use today, is a useful first baseline, and as can be seen from our results below, does deliver acceptable results.

7 Valence annotation

To perform the valence annotation, we first lemmatized the headlines and removed stop words and words with frequency above 10 000 in the LA times corpus. For each headline, we then summed — as discussed above — the context vectors of the remaining words, thus producing a 220 220-dimensional vector for each headline. This vector was then compared to each of the postulated valence vectors by computing the cosine of the angles between the vectors.

We thus have for each headline two cosines, one between the headline and the positive vector and one between the headline and the negative vector. The valence vector with highest cosine score (and thus the smallest spatial angle) was chosen to annotate the headline. For the negative valence vector we assigned a negative valence value, and for the positive vector a positive value. In 11 cases, a value of -0.0 was ascribed, either because all headline words were removed by frequency and stop word filtering, or because none of the remaining words occurred in our newsprint corpus.

Our method thus only delivers a binary valence decision — either positive or negative valence. Granted, we could have assigned a neutral valence to very low cosine scores, but as any threshold for deciding on a neutral score would be completely arbitrary, we decided to only give fully positive or negative scores to the test headlines. Also, since our aim was to provide a high-recall result, we did not wish to leave any headline with an equivocal score. We scaled the scores to fit the requirements of the coarse-grained evaluation: for each headline with a non-zero score, we multiplied the value with 100 and boosted each value with 50.[5] By this scaling operation we guaranteed a positive or a negative score for each headline (apart from the 11 exceptions, in effect unanalyzed by our algorithm, as mentioned above).

8 Results

The results from the fine-grained and coarse-grained evaluations are shown in Table 2. They show, much as we anticipated, that the coarse-grained evaluation was appropriate for our purposes.

Fine-grained	Coarse-grained		
	Accuracy	Precision	Recall
20.68	29.00	28.41	60.17

Table 2: The results of the valence annotation.

8.1 Correlation coefficients, normality assumptions, and validity of results

The fine-grained evaluation as given by the organisers and as shown in Table 2 was computed using Pearson's product-moment coefficient. Pearson's correlation coefficient is a parametric statistic and assumes normal distribution of the data it is testing for correlation. While we have no idea of neither the other contributions' score distribution, nor that of the given test set, we certainly do know that our data are not normally distributed. We would much prefer to evaluate our results using a non-parametric correlation test, such as Spearman's ρ, and suggest that the all results would be rescored using some non-parametric method instead — this would reduce the risk of inadvertent false positives stemming from divergence from the normal distribution rather than divergence from the test set.

[5] The coarse-grained evaluation collapsed values in the ranges $[-100, -50]$ as negative, $[-50, 50]$ as neutral, and $[50, 100]$ as positive.

8.2 Use cases

Evaluation of abstract features such as emotional valence can be done within a system oriented framework such as the one used in this experiment. Alternatively, one could evaluate the results using a parametrized use case scenario. A simple example might be to aim for either high recall or high precision, rather than using an average which folds in both scenarios into one numeric score — easy to compare between systems but dubious in its relevance to any imaginable real life task. There are metrics, as formal as the simple recall-precision-framework in traditional adhoc retrieval, that could be adapted for this purpose (Järvelin and Kekäläinen, 2002, e.g.).

9 Related research

Our approach to valence annotation is similar to the second method described by Turney and Littman (2003). In short, their method uses singular value decomposition to produce a reduced-dimensional word space, in which word valence is computed by subtracting the cosine between the word and a set of negative seed words from the cosine between the word and a set of positive seed words.

The difference between our approach and theirs is that our approach does not require any computationally expensive matrix decomposition, as we do not see any reason to restructure our word space. Turney and Littman (2003) hypothesize that singular value decomposition is beneficial for the results in valency annotation because it infers paradigmatic relations between words in the reduced space. However, as we argued in Section 5, we believe that the headline valency annotation task calls for syntagmatic rather than paradigmatic relations. Furthermore, we fail to see the motivation for using singular value decomposition, since if paradigmatic relations are what is needed, then why not simply use words as dimensions of the word space?

10 Concluding remarks

Our results show that a resource-poor but data-rich method can deliver sensible results. This is in keeping with our overall approach, which aims for as little pre-computed resources as possible.

At almost every juncture in our processing we made risky and simplistic assumptions — using simple frequencies of word occurrence as a semantic model; using a small seed set of positive and negative terms as a target; postulating one semantic locus each for positive and negative emotion; modelling syntactic and semantic relations between terms by vector addition — and yet we find that the semantic structure of distributional statistics yields a signal good enough for distinguishing positive from negative headlines with a non-random accuracy. Despite its simplicity, out method produces very good recall (60.17) in the coarse-grained evaluation (the median recall for all systems is 29.59). This speaks to the power of distributional semantics and gives promise of improvement if some of the choice points during the process are returned to: some decisions can well benefit from being made on principled and informed grounds rather than searching under the street lamp, as it were.

References

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD Dissertation, Department of Linguistics, Stockholm University.

Hinrich Schütze. 1993. Word space. In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93*, pages 895–902, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04*, pages 1083–1086.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Conference of the Association for Computational Linguistics, ACL'02*, pages 417–424.