Sentiment Lexicon Cross-Generation

Chris Jenkins

25/9/16

Introduction

This project consists of the implementation of a technique to generate a lexicon of sentiment-tagged[1] words in German through the use of an existing sentiment-tagged lexicon in English. This is achieved through the use of a Markov random walk model over a graph consisting of similarity relations between English words, similarity relations between German words, and translations between English and German words. Some English nodes contain sentiment labels from a pre-existing lexicon. The task is to assign sentiment labels to German nodes.

Methodology

The methods that I have used are taken from two papers which describe the approach of using a random walk model *Identifying Text Polarity Using Random Walks* by Hassan and Radev,[1] and then the technique's application toward generating a sentiment-tagged lexicon in a foreign language in *Identifying the Semantic Orientation of Foreign Words* by Hassan, Abu-Jbara, Jha, and Radev. [2]

Our first step is to generate a network that connects an English network (with some sentiment labels) to a German network (which contains no sentiment labels). I combined the following resources into an undirected graph structure.

Resources Used

1. GermaNet - a German lexical-semantic network from Uni-Tübingen. Used for German-German edges.[3, 4]
2. WordNet - an English lexical-semantic network from Princeton University. Used for English-English edges.[8, 9]
3. dict.cc - an English-German translation dictionary. Used for English-German edges.[5]
4. Lists of positive and negative words in English by Hu and Liu (2004)[6]

The two lexical-semantic networks were not copied in their entirety. I retained edges between words if they were in the same *synset*[2], had a hypernym-hyponym relationship, or had a *similar-to* relationship.

[1] consisting of two categories: positive and negative
[2] set of synonyms

Algorithm

I implemented the Random Walk algorithm described in Hassan, Radev (2010) to search over the network that I constructed with the resources described above. The transition probability from one node to another is uniformly distributed over all neighboring nodes. k random walks are performed, with a maximum of m steps. A walk ends when a word with the target polarity (either positive or negative) is reached, returning the number of steps taken to reach that node, or else m if no node with the target label is found. This returned value is called the *first-passage time*, and its value is averaged over the k walk attempts. For each word, k walks are performed with both the negative and positive target labels. The resulting mean first-passage times are compared. If the ratio between the negative and positive mean first-passage times is sufficient (greater than r), the German word is classified with the appropriate label. Otherwise it receives a neutral label.

Parameter Values

I used the following values

- $k = 100$ (number of walks)
- $m = 20$ (maximum steps)
- $r = 0.5$ (ratio $\frac{positive}{negative}$ or $\frac{negative}{positive}$ of first passage time required for + or - label)

Evaluation

I compared my system results of German words with predicted positive and negative annotations with a collection obtained from GermanPolarityClues.[7] There are some difficulties in the comparison between my results and this collection, as GermanPolarityClues includes some annotation referring to the probability of a word conveying positive or negative sentiment. This finer-grained judgment is absent from my implementation of this project, and from the English sentiment lexicon that I used as a source. Despite this limitation, I report precision and recall scores for sentiment classification of all (lemmatized) terms against GermanPolarityClues.[3]

	Total Classified	Total True Positive	Precision	Recall	F1
Positive	703	467	0.664	0.128	0.215
Negative	2251	1304	0.579	0.219	0.318

This result indicates that my implementation is working roughly as intended. I believe that better results could be obtained through a more thorough reconciliation of the two different formats used in the sentiment lexicons that I used.

References

[1] Hassan, A., & Radev, D. (2010). Identifying text polarity using random walks. In ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. (pp. 395-403)

[3] Perhaps it would have been preferable to limit the gold-set by words that were available in the German source data used.

[2] Hassan, A., Abu-Jbara, A., Jha, R., & Radev, D. (2011). Identifying the semantic orientation of foreign words. In ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. (Vol. 2, pp. 592-597)

[3] Hamp, Birgit and Helmut Feldweg. "GermaNet - a Lexical-Semantic Net for German." Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid, 1997.

[4] Henrich, Verena and Erhard Hinrichs. "GernEdiT - The GermaNet Editing Tool". Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta, May 2010, pp. 2228-2235. [Download: http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf]

[5] Paul Hemetsberger 2002-2016 http://www.dict.cc

[6] Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 168–177.

[7] Ulli Waltinger (2010). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. In Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10), April 7-10, 2010, Valencia, 2010

[8] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[9] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.