Evaluating Lexical Similarity to Build Sentiment Similarity

Grégoire Jadi*, **Vincent Claveau**†, **Béatrice Daille***, **Laura Monceaux-Cachard***
* LINA - Univ. Nantes
{gregoire.jadi beatrice.daille laura.monceaux}@univ-nantes.fr
† IRISA–CNRS, France
vincent.claveau@irisa.fr

Abstract

In this article, we propose to evaluate the lexical similarity information provided by word representations against several opinion resources using traditional Information Retrieval tools. Word representation have been used to build and to extend opinion resources such as lexicon, and ontology and their performance have been evaluated on sentiment analysis tasks. We question this method by measuring the correlation between the sentiment proximity provided by opinion resources and the semantic similarity provided by word representations using different correlation coefficients. We also compare the neighbors found in word representations and list of similar opinion words. Our results show that the proximity of words in state-of-the-art word representations is not very effective to build sentiment similarity.
Keywords: opinion lexicon evaluation, sentiment analysis, spectral representation, word embedding

1. Introduction

The goal of opinion mining or sentiment analysis is to identify and classify texts according to the opinion, sentiment or emotion that they convey. It requires a good knowledge of the sentiment properties of each word. This properties can be either discovered automatically by machine learning algorithms, or embedded into language resources.

On the one hand, there exist lexical resources (opinion lexicons) in which words are associated with their valence (e.g. positive or negative opinion) or regrouped in sentiment classes (e.g. { positive, negative, objective/neutral }, { fear, domination, sympathy, ... }). Many sentiment analysis systems exploit such resources, but one major issue is to complete or adapt them to the domain/context.

On the other hand, rich word representations such as word embeddings are also widely used in sentiment analysis. These word representations encode a semantic similarity meaning that two semantically close words are close in the representation space. They can be used by classification systems for sentiment analysis, or to build and enrich opinion lexicons. In any case, the underlying hypothesis is that the semantic similarity as encoded by the word representation is correlated to the sentiment or emotional similarity.

In this paper, our main goal is to question this hypothesis by evaluating the convergence between lexical similarity and opinion or sentiment-based similarity, that is: if two words are considered as similar from a semantic point-of-view, are they similar in terms of valence or sentiment values.

2. Related work

Lexical similarity has been used in opinion mining systems in different ways. Some studies have proposed to use existing resources to build or extend opinion lexicons. Among them, WordNet is certainly one of the most known semantic lexicon for English. Words are regrouped into *synsets* (sets of synonyms) which are linked by different semantic relations such as hypernymy, or meronymy. SentiWordNet is an extension of WordNet that assigns to each synsets three sentiment scores : positivity, negativity and objectivity. Those resources have been used extensively to build opinion analysis system. They can be used as it, or to extend or build sentiment lexicons. For example, (Toh and Wang, 2014) uses the syntactic categories from WordNet as features for a CRF to extract aspects and WordNet relations such as *antonymy* and *synonymy* to extend an existing opinion lexicons. Similarly, (Agarwal et al., 2011) look for synonyms in WordNet to find the polarity of words absent from an opinion lexicon. Going further, (Badaro et al., 2014) build an Arabic version of SentiWordNet (an extension of WordNet that assigns to each synsets three sentiment scores : positivity, negativity and objectivity) by combining an Arabic version of WordNet and the English version of WordNet and SentiWordNet.

While the previous studies build on a manually constructed lexicon, many systems have been proposed to build resources automatically based on lexical similarity for opinion mining. For that purpose, rich word representations have been proposed to compute the lexical similarity. In (Castellucci et al., 2013), the authors combine three kernels into a SVM model where each kernel tries to capture an aspect of the opinion. A Bag of Word Kernel is used to compute the similarity of ngrams between tweets; A lexical semantic kernel build a Word Space from a matrix of co-occurrence context scores; And a smoothed partial tree kernel handles the syntactic information. (Faruqui et al., 2015) use an ontology to improve the effectiveness of word vectors and evaluate their work on word similarity and sentiment analysis tasks. They refine word vectors by minimizing the distance of words within the vector space and between vectors projected in the ontology. In (Maas et al., 2011), the authors build a vector space using a continuous mixture distribution over words in an unsupervised fashion. Then, they use annotated data to find an hyperplane for which the polarity of words in the vector space depends on its position relative to the hyperplane. They evaluate their model by predicting the polarity of movie reviews.

The most common way to evaluate sentiment analysis systems, is by comparing the prediction of the systems against a gold corpus. For example, SEMEVAL (Nakov et al., 2013; Rosenthal et al., 2014) proposes a task in which participants are asked to predict the polarity (positive, negative, neutral) of tweets and SMS. Another task focuses

Model	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
Ferret 2013 *base*	5,6	7,7	22,5	14,1	10,8	5,3	3,8
Ferret 2013 *best rerank*	6,1	8,4	24,8	15,4	11,7	5,7	3,8
Ferret 2014 *synt*	7,9	10,7	29,4	**18,9**	**14,6**	**7,3**	**5,2**
Spectral (Claveau et al., 2014)	**8,97**	**10,94**	**31,05**	18,44	13,76	6,46	4,54
W2V dim=50 w=5	2,89	3,89	13,48	7,36	5,44	2,58	1,82
W2V dim=100 w=5	3,65	4,84	18,49	9,62	7,04	3,16	2,17
W2V dim=200 w=5	3,92	5,44	22,18	11,39	8,32	3,61	2,59
W2V dim=300 w=5	5,25	6,25	18,67	10,72	7,73	3,49	2,38
W2V dim=400 w=5	5,06	6,43	20,37	11,44	8,29	3,66	2,50
W2V dim=50 w=9	3,12	4,11	13,11	7,80	5,68	2,59	1,87
W2V dim=100 w=9	4,14	5,55	17,18	9,25	6,79	3,21	2,21
W2V dim=200 w=9	4,42	5,60	17,69	10,71	7,47	3,40	2,32
W2V dim=300 w=9	4,07	5,53	20,50	11,13	8,02	3,62	2,52
W2V dim=400 w=9	4,39	5,51	17,81	9,95	7,43	3,24	2,21
W2V Google news	5,82	7,51	13,28	11,60	8,94	3,93	2,54

Table 1: Performance of different lexical representation on the WN+Moby reference

on the aspects have been proposed ((Pontiki et al., 2014)) where participants are asked to identify and summarize the opinions expressed towards all aspects of an entity. An aspect is a constituent of an entity targeted by an opinion, for example, an aspect of the entity *laptop* is its *battery*. Yet, these task-based evaluations do not allow for a direct evaluation of the lexical similarity to represent the sentiment properties of the words, as provided by word embeddings or spectral representations.

3. Lexical and distributional semantic similarities

3.1. From distributional semantics to word embeddings

Since the pioneering work of (Grefenstette, 1994) and (Lin, 1998), building distributional thesauri has been widely studied. They all rely on the hypothesis that each word is semantically characterized by all the contexts in which it appears. Many techniques, implementing this hypothesis has been proposed, and recently, (Claveau et al., 2014) proposed to use Information Retrieval metrics to build distributional thesauri represented as a (weighted) graphs of neighbors. A word is thus represented by its links with other words. This distributional method, which gives state-of-the-art results on lexical similarity tasks, is called Spectral representation hereafter.

In the recent years, other word representation techniques have been proposed to represent words as vectors such that two (semantically) close words are (spatially) close in the vector space. The proximity of words in the vector space is obtained with a distance measure such as the L2 distance or cosine. A vector space of words is a space in which each word is represented by a vector usually built from its context. Since those representations are often very sparse, methods to reduce the dimension were proposed such as Latent Semantic Indexing, Non-negative Matrix Factorization, Singular Value Decomposition. The semantic of words is encoded by context vectors into the topology of the vector space so that, for example, words that are closed to each other in the space are semantically close (e.g. synonym). Another use of vector spaces are analogy resolution, for example, we can determine that *Man* is to *Woman* what *King* is to *Queen* by analyzing the spatial relation between *Man* and *Woman* and *King* and *Queen*. Among the existing techniques to represent words in vector spaces, the word embeddings produced by Word2Vec are very popular (Mikolov et al., 2013).

3.2. Evaluation of the lexical similarity

In the remaining of the paper, we use these two word representation techniques (Word2Vec and Spectral), both trained on the 4 million pages of Wikipedia$_{EN}$[1]. In order to asses their quality as semantic representations, we compare their results (as well as other published results) over one dataset used for the evaluation of lexical similarity.

This dataset, used as reference, is a collection of words with their semantic neighbors as encoded in WordNet (Miller, 1995) and Moby (Ward, 1996). Table 1 shows the results with the usual IR performance score (Mean Average Precision, R-Precision, Precision on the top-k nearest words). We report results of the literature, results with different parameters of Word2Vec: number of dimensions (dim), size of the window (w). For comparison purposes, we also provide the results obtained with a freely available model of Word2Vec trained on the 100-billion word Google News corpus [2].

The results show that the Spectral representation is more accurate than Word2Vec on this task which aims at detecting very close words (synonyms or quasi-synonyms), as in the latest evaluation task. Hereafter, in the experiments reported below, we keep the Word2Vec representation with the parameters yielding the best results, that is the pre-trained Word2Vec model built with the Google News corpus.

[1]The Spectral word representation will be available from `http://people.irisa.fr/Vincent.Claveau/`

[2]The pre-trained vectors are available at `https://code.google.com/archive/p/word2vec/`.

	Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
Spectral	0.1264	0.1128	0.0838
Word2Vec	0.1080	0.0952	0.0636

Table 2: Lexical similarity vs. the proximity in terms of valence on ANEW Lexicon.

	Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
Spectral	0.0484	0.0528	0.0299
Word2Vec	0.0456	0.0456	0.0304

Table 3: Lexical similarity vs. the proximity in terms of arousal on ANEW Lexicon.

4. Comparing opinion lexicons with lexical similarity

As said in the state-of-the-art, sentiment analysis is traditionally evaluated against messages or reviews whereas lexical similarity is evaluated against thesaurus or or other manually crafted tests. In this paper, we follow the evaluation procedure used to evaluate lexical similarity to sentiment analysis. That is, we propose to evaluate our word vector spaces against opinion lexicon and opinion thesaurus.

To our knowledge, we are the first to propose an evaluation of word vectors on both, word similarity tasks and opinion lexicons. In addition to the aforementioned contribution, we also provide the word vectors we have produced to the community.

4.1. Correlation between sentiment proximity and semantic similarity: ANEW

ANEW is an opinion lexicon in which each word is described by three (real-valued) properties: valence, arousal and domination (Bradley and Lang, 1999). Here, our goal is to check whether the proximity in terms of sentiments given by the lexicon is correlated to the semantic proximity obtained with the word representations. Given a word in ANEW, we build a reference list as the ordered list of its neighbors in terms of sentiment (valence, arousal or domination). Then, we build a second the ordered list with the closest semantic neighbors of the word. Finally, we use different correlation coefficients (Pearson's r, Spearman's ρ, Kendall's τ) to compare the two ordered list. This process is repeated for several words, and the average results are given in Table 5. Tables 2, 3, 4 respectively show the correlation coefficients for valence, arousal, domination.

Concerning the learned semantic similarity (Spectral and Word2Vec), it appears clearly that for every coefficient considered, the correlation is very low, which means that the semantic similarity that is obtained from this built resources is not clearly related to the sentiment proximity, for any dimension.

In the experiment reported hereafter, these three properties (valence, arousal, domination) are interpreted as dimensions in a $\mathbb{R}^3$ vector space. Thus, two words are considered as close in terms of sentiment if they are close in this vector space (L2 distance). We also provide the results obtained by comparing directly the SimLex999 reference (Leviant and Reichart, 2015) with ANEW; SimLex999 is a resource which encodes how similar two words are, based on human evaluation. More strikingly, even with the reference similarity list provided by SimLex999, the correlation seems slightly better but remains very low.

4.2. Correlation between sentiment proximity and semantic similarity: SentiWordNet

SentiWordNet is another lexicon that associates positive and negative values (between 0 and 1) to synsets. Note that for a given synset, these two scores are independent: one synset may have a non-zero score for both positive and negative values. SentiWordNet also defines an 'objective' score for each synset as: $1 - (positive\_value + negative\_value)$. Table 6 reports the results with the same experiment settings as for the ANEW lexicon.

Here again, the very low coefficients tend to show that there are almost no correlation between the sentiment-based proximity and the semantic one, either computed (Spectral and Word2Vec) or manually assessed (SimLex999).

4.3. Building classes of similar words: NRC emotion lexicon

The NRC emotion lexicon (Mohammad et al., 2013) is a large list of words associated with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). These emotions and sentiments are encoded as binary properties (the word has or not the emotional property), that were manually obtained through Amazon's Mechanical Turk.

Given a word with its emotion and valence properties, we want to know if the lexical similarity helps finding words sharing the exact same emotion and valence properties. In order to evaluate that, we set up the following experiment: given a query word (expressing at least one of the 10 emotional properties), we consider every word having exactly the same profile (same absence or presence of the 10 emotional properties) as the reference set. For this query word, the ranked lists of similar words computed with the Spectral and Word2Vec representations are compared with this reference list. Table 7 reports the results with usual scores (mean average precision, R-precision, precision on the 1, 5 ,10, etc. closest words). As a baseline, we also report the results obtained with a random ranking of the words.

Here again, the results are very low, not far from those obtained at random, which means that the complete emotional properties are not captured by the semantic similarity of our

	Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
Spectral	0.0805	0.0682	0.0503
Word2Vec	0.0883	0.0777	0.0519

Table 4: Lexical similarity vs. the proximity in terms of domination on ANEW Lexicon.

	Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
Spectral	0.1185	0.1073	0.0713
Word2Vec	0.1214	0.1046	0.0699
SimLex999	0.3896	0.3252	0.2209

Table 5: Lexical similarity vs. the proximity of combination of valence, arousal and domination on ANEW Lexicon.

		Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
Spectral	positive	0.07657	0.0287	0.0233
	negative	0.0880	0.1098	0.0870
	objective	0.1351	0.1241	0.0958
Word2Vec	positive	0.04964	0.0219	0.0173
	negative	0.1214	0.1046	0.0699
	objective	0.0499	0.02816	0.0215
SimLex999	positive	0.0328	0.0159	0.0122
	negative	0.0217	0.0219	0.0174
	objective	0.0467	0.0392	0.0300

Table 6: Lexical similarity vs. negative/positive/objective score proximity on SentiWordNet.

	MAP	R-prec	P@1	P@5	P@10	P@50	P@100
Random	1.09	2.55	3.67	2.25	2.25	2.37	2.52
Spectral	2.75	5.22	15.30	13.57	12.77	11.20	9.58
Word2Vec	1.34	5.24	13.38	11.52	10.59	8.33	7.51

Table 7: Lexical similarity vs. sentiment/opinion proximity on NRC dataset considering all the dimensions (eight emotions + the positive/negative sentiment).

word representations. Hereafter, in Table 8 we report the results by considering only the positive/negative axis.

The better results are not surprising since the emotional complexity (number of sentiment properties) considered is less. Yet, in the best case, only half of the 10 closest neighbors have the same valence than the query word.

In addition to the previous experiment, Figure 1 presents the results (P@1, P@10) obtained for different other sentiment properties (fear, joy, anger, surprise, anticipation) with the Word2Vec model.

It appears clearly that some of these sentiment properties are more easily captured by the lexical similarity.

Of course, when the goal is to extend an opinion lexicon, several seed words of a same class can be used (not just one as in the previous experiment) to find new words sharing hopefully the same sentiment properties. Therefore, it is also interesting to examine how the performance evolves according to the number of words used as queries. In Figure 2, we report the MAP with respect to the number of query words. Unsurprisingly, using lexical similarity from several words helps to get better results, but here again, the results remain low.

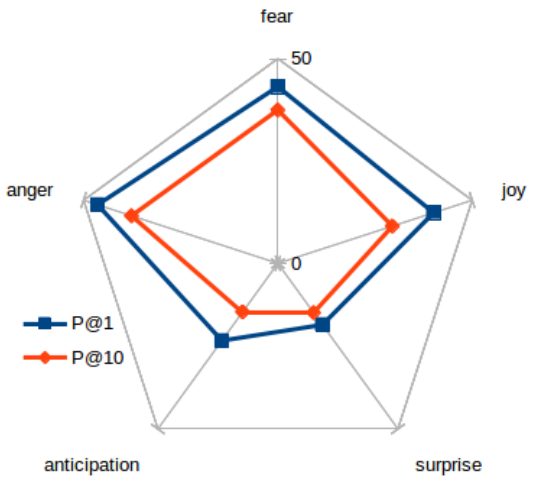

Figure 1: P@1 and P@10 of the Word2Vec model for retrieving different sentiment properties from the NRC opinion lexicon.

	MAP	R-prec	P@1	P@5	P@10	P@50	P@100
Random	6.87	18.73	18.21	19.00	18.15	18.39	18.61
Spectral	18.71	33.59	58.40	55.60	53.42	49.29	46.79
Word2Vec	15.56	28.83	50.00	47.80	46.02	41.98	39.95

Table 8: Lexical similarity vs. positive/negative sentiment proximity in the NRC emotion lexicon.

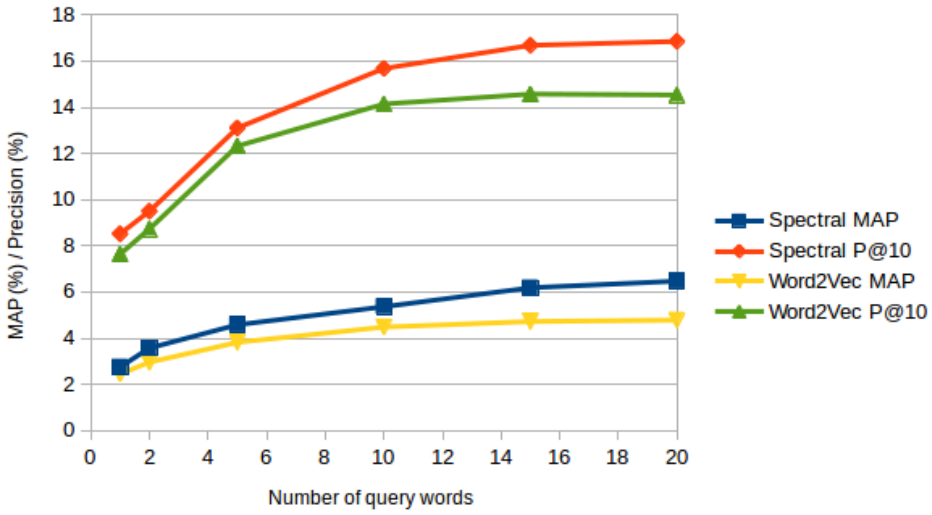

Figure 2: MAP and P@10 according to the number of query words with the same 10 sentiment properties from the NRC opinion lexicon.

5. Discussion and conclusive remarks

Several results presented in this article are noteworthy. First, lexical similarity and sentiment similarity are not fully compatible. Without refinement, we have shown that techniques devoted to lexical similarity (like word embedding or spectral representation) do not perform well to retrieve words with similar sentiment properties in general, even though some differences have been shown according to the specific sentiment property considered. Globally, it is an issue to use word representations to build or to extend opinion resources but also to misuse them as opinion resources in classification systems for sentiment analysis. Word representations can enhance classification systems for sentiment analysis but cannot be considered as good option *per se* to build/enrich opinion resources. Moreover, this negative result also holds with manually built lexical resources such as SimLex999. Furthermore, it has to be noted that opinion lexicons themselves do not fully agree, as it is shown in the SentiWN vs. ANEW comparison illustrated in Table 9. Here again, the correlation between proximity of sentiments in the two opinion resources are above random but remains very low.

In order to improve the effectiveness of word representations, (Dasgupta and Ng, 2009) proposed to use a human oracle to select the best possible dimension when they perform a matrix factorization. In a future work, we plan to use simple techniques exploiting the analogy solving capabilities of our word representations to adapt the semantic similarity to better match the expected sentiment properties.

6. Acknowledgements

This work was funded via the CominLabs excellence laboratory financed by the National Research Agency under reference ANR-10-LABX-07-01 (project LIMAH `limah.irisa.fr`).

7. References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38. Association for Computational Linguistics, June.

Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173. Association for Computational Linguistics, October.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings.

Castellucci, G., Filice, S., Croce, D., and Basili, R. (2013). Unitor: Combining syntactic and semantic kernels for twitter sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 369–374. Association for Computational Linguistics, June.

Claveau, V., Kijak, E., and Ferret, O. (2014). Improving

	Mean Pearson's r	Mean Spearman's ρ	Mean Kendall's τ
ANEW/SentiWN	0.3703	0.3595	0.2807

Table 9: ANEW valence score vs. SentiWN positive-negative score.

distributional thesauri by exploring the graph of neighbors. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 709–720. Dublin City University and Association for Computational Linguistics, August.

Dasgupta, S. and Ng, V. (2009). Topic-wise, sentiment-wise, or otherwise? identifying the hidden dimension for unsupervised text classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 580–589.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. Association for Computational Linguistics, May–June.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Leviant, I. and Reichart, R. (2015). Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. abs/1508.00106.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *17^{th} International Conference on Computational Linguistics and 36^{th} Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, pages 768–774, Montréal, Canada.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, June.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. abs/1310.4546.

Miller, G. A. (1995). Wordnet: A lexical database for english. 38(11):39–41, 11.

Mohammad, S., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327. Association for Computational Linguistics, June.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. Association for Computational Linguistics, June.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35. Association for Computational Linguistics and Dublin City University, August.

Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80. Association for Computational Linguistics and Dublin City University, August.

Toh, Z. and Wang, W. (2014). Dlirec: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240. Association for Computational Linguistics and Dublin City University, August.

Ward, G. (1996). Moby thesaurus.