Building Sentiment Lexicons for All Major Languages

Yanqing Chen
Computer Science Dept.
Stony Brook University
Stony Brook, NY 11794
cyanqing@cs.stonybrook.edu

Steven Skiena
Computer Science Dept.
Stony Brook University
Stony Brook, NY 11794
skiena@cs.stonybrook.edu

Abstract

Sentiment analysis in a multilingual world remains a challenging problem, because developing language-specific sentiment lexicons is an extremely resource-intensive process. Such lexicons remain a scarce resource for most languages.

In this paper, we address this lexicon gap by building high-quality sentiment lexicons for 136 major languages. We integrate a variety of linguistic resources to produce an immense knowledge graph. By appropriately propagating from seed words, we construct sentiment lexicons for each component language of our graph. Our lexicons have a polarity agreement of 95.7% with published lexicons, while achieving an overall coverage of 45.2%.

We demonstrate the performance of our lexicons in an extrinsic analysis of 2,000 distinct historical figures' Wikipedia articles on 30 languages. Despite cultural difference and the intended neutrality of Wikipedia articles, our lexicons show an average sentiment correlation of 0.28 across all language pairs.

1 Introduction

Sentiment analysis of English texts has become a large and active research area, with many commercial applications, but the barrier of language limits the ability to assess the sentiment of most of the world's population.

Although several well-regarded sentiment lexicons are available in English (Esuli and Sebastiani, 2006; Liu, 2010), the same is not true for most of the world's languages. Indeed, our literature search identified only 12 *publicly available* sentiment lexicons for only 5 non-English languages (Chinese mandarin, German, Arabic, Japanese and Italian). No doubt we missed some, but it is clear that these resources are not widely available for most important languages.

In this paper, we strive to produce a comprehensive set of sentiment lexicons for the worlds' major languages. We make the following contributions:

- *New Sentiment Analysis Resources* – We have generated sentiment lexicons for 136 major languages via graph propagation which are now publicly available[1]. We validate our own work through other publicly available, human annotated sentiment lexicons. Indeed, our lexicons have polarity agreement of 95.7% with these published lexicons, plus an overall coverage of 45.2%.

- *Large-Scale Language Knowledge Graph Analysis* – We have created a massive comprehensive knowledge graph of 7 million vocabulary words from 136 languages with over 131 million semantic inter-language links, which proves valuable when doing alignment between definitions in different languages.

- *Extrinsic Evaluation* – We elucidate the sentiment consistency of entities reported in different language editions of Wikipedia using our propagated lexicons. In particular, we pick 30 languages and compute sentiment scores for 2,000 distinct historical figures. Each language pair exhibits a Spearman sentiment correlation of at least 0.14, with an average correlation of 0.28 over all pairs.

The rest of this paper is organized as follows. We review related work in Section 2. In Section 3, we describe our resource processing and design decisions. Section 4 discusses graph propagation methods to identify sentiment polarity across languages. Section 5 evaluates our results against

[1] https://sites.google.com/site/datascienceslab/projects/

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 383–389,
Baltimore, Maryland, USA, June 23-25 2014.

each available human-annotated lexicon. Finally, in Section 6 we present our extrinsic evaluation of sentiment consistency in Wikipedia prior to our conclusions.

2 Related Work

Sentiment analysis is an important area of NLP with a large and growing literature. Excellent surveys of the field include (Liu, 2013; Pang and Lee, 2008), establishing that rich online resources have greatly expanded opportunities for opinion mining and sentiment analysis. Godbole et al. (2007) build up an English lexicon-based sentiment analysis system to evaluate the general reputation of entities. Taboada et al. (2011) present a more sophisticated model by considering patterns, including negation and repetition using adjusted weights. Liu (2010) introduces an efficient method, at the state of the art, for doing sentiment analysis and subjectivity in English.

Researchers have investigated topic or domain dependent approaches to identify opinions. Jijkoun et al. (2010) focus on generating topic specific sentiment lexicons. Li et al. (2010) extract sentiment with global and local topic dependency. Gindl et al. (2010) perform sentiment analysis according to cross-domain contextualization and Pak and Paroubek (2010) focus on Twitter, doing research on colloquial format of English.

Work has been done to generalize sentiment analysis to other languages. Denecke (2008) performs multilingual sentiment analysis using SentiWordNet. Mihalcea et al. (2007) learn multilingual subjectivity via cross-lingual projections. Abbasi et al. (2008) extract specific language features of Arabic which requires language-specific knowledge. Gînscă et al. (2011) work on better sentiment analysis system in Romanian.

The ready availability of machine translation to and from English has prompted efforts to employ translation for sentiment analysis (Bautin et al., 2008). Banea et al. (2008) demonstrate that machine translation can perform quite well when extending the subjectivity analysis to multi-lingual environment, which makes it inspiring to replicate their work on lexicon-based sentiment analysis.

Machine learning approaches to sentiment analysis are attractive, because of the promise of reduced manual processing. Boiy and Moens (2009) conduct machine learning sentiment analysis using multilingual web texts. Deep learning approaches draft off of distributed word embedding which offer concise features reflecting the semantics of the underlying vocabulary. Turian et al. (2010) create powerful word embedding by training on real and corrupted phrases, optimizing for the replaceability of words. Zou et al. (2013) combine machine translation and word representation to generate bilingual language resources. Socher et al. (2012) demonstrates a powerful approach to English sentiment using word embedding, which can easily be extended to other languages by training on appropriate text corpora.

3 Knowledge Graph Construction

In this section we will describe how we leverage off a variety of NLP resources to construct the semantic connection graph we will use to propagate sentiment lexicons.

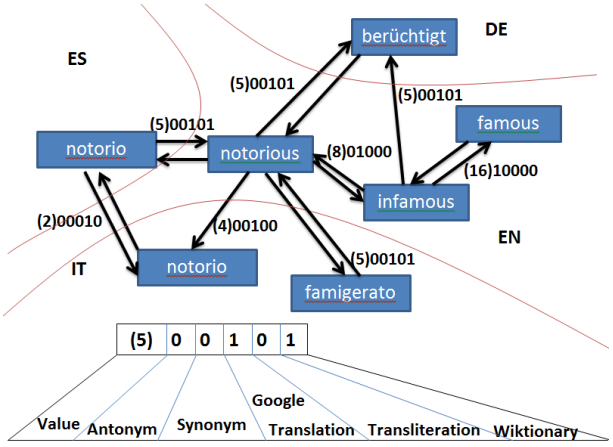

Figure 1: Illustration of our knowledge graph, showing links between words and edge representation to preserve source identity. For each edge between corresponding words, a 5-bit integer will record the existence of 5 possible semantic links.

The Polyglot project (Al-Rfou et al., 2013) identified the 100,000 most frequently used words in each language's Wikipedia. Drawing a candidate lexicon from Wikipedia has some downsides (e.g. limited use of informal words), but is representative and convenient over a large number of languages. In particular, we collect total of 7,741,544 high-frequency words from 136 languages to serve as vertices in our graph.

We seek to identify as many semantic links across languages as possible to connect our network, and so integrated several resources:

- *Wiktionary* – This growing resource has en-

tries for 171 languages, edited by people with sufficient background knowledge. Wiktionary provides about 19.7% of the total links covering 382,754 vertices in our graph.

- *Machine Translation* - We script the Google translation API to get even more semantic links. In particular we ask for translations of each word in our English vocabulary to 57 languages with available translators as well as going from each known vocabulary word in other languages to English. In total, machine translation provides 53.2% of the total links and establishes connections between 3.5 million vertices.

- *Transliteration Links* – Natural flow brings words across languages with little morphological change. Closely related language pairs (i.e. Russian and Ukrainian) share many characters/words in common. Though not always true, words with same spelling usually have similar meanings so this can improve the coverage of semantic links. Transliteration provides 22.1% of the total links in our experiment.

- *WordNet* – Finally, we gather synonyms and antonyms of English words from WordNet, which prove particularly useful in propagating sentiment across languages. In total we collect over 100,000 pairs of synonyms and antonyms and created 5.0% of the total links.

Links do not always agree in a bidirectional manner, particularly for multi-sense words, thus all links in our network are unidirectional. Figure 1 illustrates how we encode links from different resources in an integer edge value.

4 Graph Propagation

Sentiment propagation starts from English sentiment lexicons. Through semantic links in our knowledge graph, words are able to extend their sentiment polarities to adjacent neighbors. We experimented with both graph propagation algorithm (Velikovich et al., 2010) and label propagation algorithm (Zhu and Ghahramani, 2002; Rao and Ravichandran, 2009). The primary difference between is that label propagation takes multiple paths between two vertices into consideration, while graph propagation utilizes only the best path between word pairs.

We report results from using Liu's lexicons (Liu, 2010) as seed words. Liu's lexicons contain 2006 positive words and 4783 negative words. Of these, 1422 positive words and 2956 negative words (roughly 64.5%) appear among the 100,000 English vertices in our graph.

Dataset	Propagation	Acc	Cov
Arabic	Label	0.93	0.45
	Graph	0.94	**0.46**
German	Label	0.97	0.31
	Graph	0.97	**0.32**
English	Label	0.92	0.55
	Graph	0.90	**0.69**
Italian	Label	0.73	0.29
	Graph	0.72	**0.32**
Japanese	Label	0.57	0.12
	Graph	0.56	**0.15**
Chinese-1	Label	0.95	0.62
	Graph	0.94	**0.65**
Chinese-2	Label	0.97	0.70
	Graph	0.97	**0.72**

Table 1: Graph propagation vs label propagation. *Acc* represents the ratio of identical polarity between our analysis and the published lexicons. *Cov* reflects what faction of our lexicons overlap with published lexicons.

Our knowledge network is comprised of links from a heterogeneous collection of sources, of different coverage and reliability. For the task of deciding sentiment polarity of words, only antonym links are negative. An edge gains zero weight if both negative and positive links exist. Edges having multiple positive links will be credited the highest weight among all these links. We conducted a grid search on the weight of each type of links to maximize the best overall accuracy on our test data of published non-English sentiment lexicons. To avoid potential overfitting problems, grid search starts from SentiWordNet English lexicons (Esuli and Sebastiani, 2006) instead of Liu's.

5 Lexicon Evaluation

We collected all available published sentiment lexicons from non-English languages to serve as standard for our evaluation, including Arabic, Italian, German and Chinese. Coupled with English sentiment lexicons provides in total seven different test cases to experiment against, specifically:

Language	‖lexicon‖	+/- Ratio	Language	‖lexicon‖	+/- Ratio	Language	‖lexicon‖	+/- Ratio
Afrikaans	2299	0.40	Albanian	2076	0.41	Amharic	46	0.63
Arabic	2794	0.41	Aragonese	97	0.47	Armenian	1657	0.43
Assamese	493	0.49	Azerbaijani	1979	0.41	Bashkir	19	0.63
Basque	1979	0.40	Belarusian	1526	0.43	Bengali	2393	0.42
Bosnian	2020	0.42	Breton	184	0.42	Bulgarian	2847	0.40
Burmese	461	0.48	Catalan	3204	0.37	Cebuano	56	0.54
Chechen	26	0.65	Chinese	3828	0.34	Chuvash	17	0.76
Croatian	2208	0.40	Czech	2599	0.41	Danish	3340	0.38
Divehi	67	0.67	Dutch	3976	0.38	English	4376	0.32
Esperanto	2604	0.40	Estonian	2105	0.41	Faroese	123	0.43
Finnish	3295	0.40	French	4653	0.35	Frisian	224	0.43
Gaelic	345	0.50	Galician	2714	0.37	German	3974	0.38
Georgian	2202	0.40	Greek	2703	0.39	Gujarati	2145	0.44
Haitian	472	0.44	Hebrew	2533	0.36	Hindi	3640	0.39
Hungarian	3522	0.38	Icelandic	1770	0.40	Ido	183	0.49
Interlingua	326	0.50	Indonesian	2900	0.37	Italian	4491	0.36
Irish	1073	0.45	Japanese	1017	0.39	Javanese	168	0.51
Kazakh	81	0.65	Kannada	2173	0.42	Kirghiz	246	0.49
Khmer	956	0.49	Korean	2118	0.42	Kurdish	145	0.48
Latin	2033	0.46	Latvian	1938	0.42	Limburgish	93	0.46
Lithuanian	2190	0.41	Luxembourg	224	0.52	Macedonian	2965	0.39
Malagasy	48	0.54	Malayalam	393	0.50	Malay	2934	0.39
Maltese	863	0.50	Marathi	1825	0.48	Manx	90	0.51
Mongolian	130	0.52	Nepali	504	0.49	Norwegian	3089	0.37
Nynorsk	1894	0.39	Occitan	429	0.40	Oriya	360	0.51
Ossetic	12	0.67	Panjabi	79	0.63	Pashto	198	0.50
Persian	2477	0.39	Polish	3533	0.39	Portuguese	3953	0.35
Quechua	47	0.55	Romansh	116	0.48	Romanian	3329	0.39
Russian	2914	0.43	Sanskrit	178	0.59	Sami	24	0.71
Serbian	2034	0.41	Sinhala	1122	0.43	Slovak	2428	0.43
Slovene	2244	0.42	Spanish	4275	0.36	Sundanese	476	0.50
Swahili	1314	0.42	Swedish	3722	0.39	Tamil	2057	0.40
Tagalog	1858	0.44	Tajik	97	0.62	Tatar	76	0.50
Telugu	2523	0.41	Thai	1279	0.51	Tibetan	24	0.63
Turkmen	78	0.56	Turkish	2500	0.39	Uighur	18	0.44
Ukrainian	2827	0.41	Urdu	1347	0.39	Uzbek	111	0.57
Vietnamese	1016	0.38	Volapuk	43	0.70	Walloon	193	0.32
Welsh	1647	0.42	Yiddish	395	0.43	Yoruba	276	0.50

Table 2: Sentiment lexicon statistics. We tag 10 languages having most/least sentiment words with blue/green color and 10 languages having highest/lowest ratio of positive words with orange/purple color.

- *Arabic*: (Abdul-Mageed et al., 2011).
- *German*: (Remus et al., 2010).
- *English*: (Esuli and Sebastiani, 2006).
- *Italian*: (Basile and Nissim, 2013).
- *Japanese*: (Kaji and Kitsuregawa, 2007).
- *Chinese-1*, *Chinese-2*: (He et al., 2010).

We present the accuracy and coverage achieved by two propagation model in Table 1. Both models achieve similar accuracy while slightly more words in graph propagation can be verified via published lexicons. Performance is not good on Japanese because of mismatching between our dictionary and the test data.

Table 2 reveals that very sparse sentiment lexicons resulted for a small but notable fraction of the languages we analyzed. In particular, only 20 languages yielded lexicons of less than 100 words. Without exception, they all have very small available definitions in Wikitionary. By contrast, 48 languages had lexicons with over 2,000 words, another 16 with between 1,000 and 2,000: clearly large enough to perform a meaningful analysis.

6 Extrinsic Evaluation: Consistency of Wikipedia Sentiment

We consider evaluating our lexicons on the consistency of Wikipedia pages about a particular individual person among various languages. As our candidate entities for analysis, we use the Wikipedia pages of 2,000 most significant people as measured in the recent book *Who's Bigger?* (Skiena and Ward, 2013). The sentiment polarity of a page is simply computed by subtracting

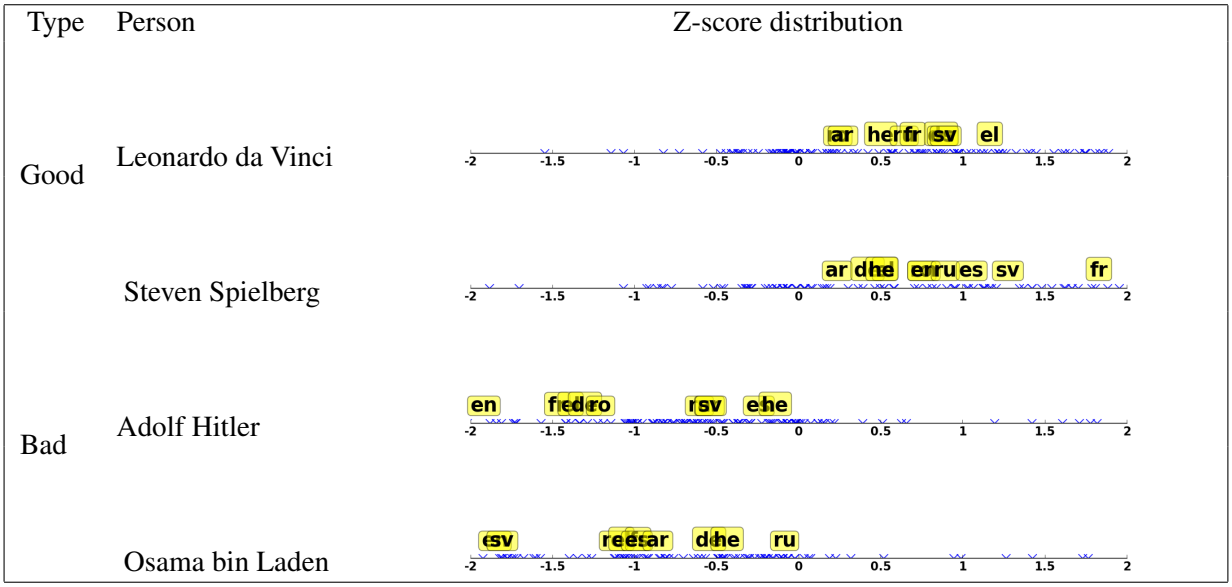

Table 3: Z-score distribution examples. We label 10 languages with their language code and other using tick marks on the x-axis.

the number of negative words from that of positive words, divided by the sum of both.

The differing ratio of positive and negative polarity terms in Table 2 means that sentiment cannot be directly compared across languages. For more consistent evaluation we compute the z-score of each entity against the distribution of all its language's entities.

We use the Spearman correlation coefficient to measure the consistence of sentiment distribution across all entities with pages in a particular language pair. Figure 2 shows the results for 30 languages with largest propagated sentiment lexicon size. All pairs of language exhibit positive correlation (and hence generally stable and consistent sentiment), with an average correlation of 0.28.

Finally, Table 3 illustrates sentiment consistency over all 136 languages (represented by blue tick marks), with the first 10 languages in Figure 2 granted labels. Respected artists like *Steven Spielberg* and *Leonardo da Vinci* show as consistently positive sentiment as notorious figures like *Osama bin Laden* and *Adolf Hitler* are negative.

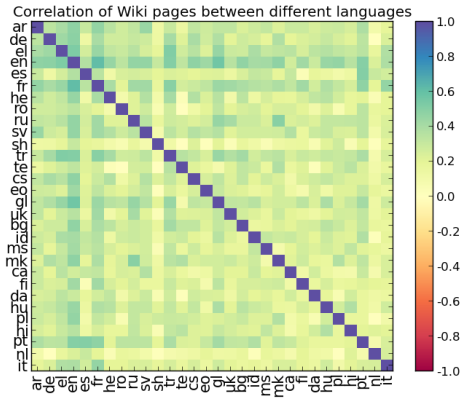

Figure 2: Heatmap of sentiment correlation between 30 languages.

7 Conclusions

Our knowledge graph propagation is generally effective at producing useful sentiment lexicons. Interestingly, the ratio of positive sentiment words is strongly connected with number of sentiment words – it is noteworthy that English has the smallest ratio of positive lexicon terms. The phenomenon possibly shows that many negative words reflecting cultural nuances do not translate wel. We believe that this ratio can be considered as quality measurement of the propagation. Similar approaches can be extended to other NLP tasks using different semantic links, specific dictionary and special seed words. Future work will revolve around learning modifiers, negation terms, and various entity/sentiment attribution.

Acknowledgments

This research was partially supported by NSF Grants DBI-1060572 and IIS-1017181, and a Google Faculty Research Award.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Muhammad Abdul-Mageed, Mona T Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *ACL (Short Papers)*, pages 587–591.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135. Association for Computational Linguistics.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. *WASSA 2013*, page 100.

M. Bautin, L. Vijayarenu, and S. Skiena. 2008. International sentiment analysis for news and blogs. Second Int. Conf. on Weblogs and Social Media (ICWSM 2008).

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558.

Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512. IEEE.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Stefan Gindl, Albert Weichselbraun, and Arno Scharl. 2010. Cross-domain contextualisation of sentiment lexicons. *19th European Conference on Artificial Intelligence (ECAI)*.

Alexandru-Lucian Gînscă, Emanuela Boroş, Adrian Iftene, Diana TrandabĂţ, Mihai Toader, Marius Corîci, Cenel-Augusto Perez, and Dan Cristea. 2011. Sentimatrix: multilingual sentiment analysis service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 189–195. Association for Computational Linguistics.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7.

Yulan He, Harith Alani, and Deyu Zhou. 2010. Exploring english lexicon knowledge for chinese sentiment analysis. *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. In *EMNLP-CoNLL*, pages 1075–1083.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *AAAI*.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:568.

Bing Liu. 2013. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 976.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC*.

Steven Skiena and Charles Ward. 2013. *Who's Bigger?: Where Historical Figures Really Rank*. Cambridge University Press.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.