Sentiment Intensity Ranking among Adjectives Using Sentiment Bearing Word Embeddings

Raksha Sharma, Arpan Somani, Lakshya Kumar, Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
IIT Bombay, India
raksha,somani,lakshya,pb@cse.iitb.ac.in

Abstract

Identification of intensity ordering among polar (positive or negative) words which have the same semantics can lead to a fine-grained sentiment analysis. For example, *master, seasoned* and *familiar* point to different intensity levels, though they all convey the same meaning (semantics), *i.e.*, *expertise: having a good knowledge of*. In this paper, we propose a semi-supervised technique that uses sentiment bearing word embeddings to produce a continuous ranking among adjectives that share common semantics. Our system demonstrates a strong Spearman's rank correlation of 0.83 with the gold standard ranking. We show that sentiment bearing word embeddings facilitate a more accurate intensity ranking system than other standard word embeddings (word2vec and GloVe). Word2vec is the state-of-the-art for intensity ordering task.

1 Introduction

The interchangeable use of semantically similar words stimulates sentiment intensity variation among sentences. To understand the phenomenon, let us consider the following example:

1. (a) We were *pleased* by the beauty of the island. (Positively low intense)
   (b) We were *delighted* by the beauty of the island. (Positively medium intense)
   (c) We were *exhilarated* by the beauty of the island. (Positively high intense)

Pleased, Exhilarated and *delighted* are the positive words bearing the same semantics, *i.e.*, *directing the emotion*, but their use intensifies the positive sentiment in the sentences 1(a), 1(b) and 1(c) respectively. Identification of intensity ranking among the words which have the same semantics can facilitate such a fine-grained sentiment analysis as exemplified in 1(a), 1(b) and 1(c).[1]

In this paper, we present a semi-supervised approach to establish a continuous intensity ranking among polar adjectives having the same semantics. Essentially, our approach is a refinement of the work done by Sharma et al., (2015). They also built a system that generates intensity of the words that bear the same semantics; however, their system considers only three discrete intensity levels, *viz., low*, *medium* and *high*. The important feature of our approach is that it uses Sentiment Specific Word Embeddings (SSWE). SSWE are an enhancement to the normal word embeddings with respect to the sentiment analysis task (Tang et al., 2014). SSWE capture syntactic, semantic as well as sentiment information, unlike normal word embeddings (word2vec and GloVe), which capture only syntactic and semantic information.

Our Contribution: We propose an approach that generates a continuous (finer) intensity ranking among polar words, which belong to the same semantic category. In addition, we show that SSWE produce a significantly better intensity ranking scale than word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which do not capture sentiment information of the words.

The remaining paper is organized as follows. Section 2 describes the previous work related to intensity ranking task. Section 3 describes the different word embeddings explored in the paper. Section 4 gives the description of the data and the resources. Section 5 provides details of the gold

[1] Words which have the different semantic concepts cannot be used interchangeably. For example, *master (expertise)* and *delighted (directing the emotion)* cannot be a replacement of each other. Hence, our understanding is that a comparison between words belonging to different semantic categories does not give any meaningful information.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 556–561
Copenhagen, Denmark, September 7–11, 2017.

standard data. Section 6 elaborates the proposed intensity ranking approach. Section 7 presents the results and experimental setup. Section 8 concludes the paper.

2 Related Work

Sentiment analysis on adjectives has been extensively explored in NLP literature. However, most of the work addressed the problem of finding polarity orientation of the adjectives (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Wilson et al., 2005; Fahrni and Klenner, 2008; Dragut et al., 2010; Taboada and Grieve, 2004; Baccianella et al., 2010).

The task of ranking polar words has received much attention recently due to the vital role of word's intensity in several real world applications. Most of the literature on intensity ranking consists of manual approaches or corpus-based approaches. Affective Norms (Warriner et al., 2013), SentiStrength (Thelwall et al., 2010), SoCAL (Taboada et al., 2011), and LABMT (Dodds et al., 2011), Best–Worst Scaling (Kiritchenko and Mohammad, 2016) are a few such publicly available sentiment intensity lexicons which are manually created.

Corpus-based approaches follow the assumption that the polarity of a new word can be inferred from the corpus (Hatzivassiloglou and McKeown, 1993; Kiritchenko et al., 2014; De Melo and Bansal, 2013). Corpus-based approaches require a huge amount of data, otherwise they suffer from the data sparsity problem. None of the these approaches considers the concept of semantics of adjectives, assuming one single intensity scale for all adjectives. Ruppenhofer et al., (2014) made the first attempt in this direction. They provided ordering among polar adjectives that bear the same semantics using a corpus-based approach. On the contrary, Sharma et al., (2015) used publicly available embeddings (word2vec) of words to assign intensity to words. Learning of word embeddings does not require annotated (labeled) corpus.

The embeddings used in our work are sentiment specific word embeddings. Integration of sentiment information of a word with syntactic and semantic information makes our approach more accurate for fine-grained sentiment intensity ranking of words.

3 Word Embeddings

In recent years, several models have been proposed to learn word embeddings from large corpora. In this paper, we have explored three types of word embeddings, *viz.*, word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014) and SSWE (Tang et al., 2014). The word embeddings given by word2vec are the distributed vector representation of the words that capture both the syntactic and semantic relationships among words. The Global Vector model, referred as GloVe, combines word2vec with ideas drawn from matrix factorization methods, such as LSA (Deerwester et al., 1990). Word2vec and GloVe model the syntactic context of the words but ignore their sentiment information. For sentiment analysis task, this is problematic as these word embeddings map words with similar syntactic context but opposite polarity, such as *love* and *hate* closer to each other in the vector space.

Sentiment Specific Word Embeddings (SSWE) encode sentiment information along with the syntactic and semantic information in word vector space. These word embeddings are able to separate the words like *love* and *hate* to the opposite ends of the spectrum. Tang et al., (2014) proposed a method to learn sentiment specific word embeddings from tweets with *emoticons* as distant-supervised corpora without any manual annotation. Specifically, they developed three neural networks to effectively incorporate the supervision from sentiment polarity of text in their loss functions.

4 Data and Resources

In this work, we have used the 52 polar semantic categories from the FrameNet data.[2] FrameNet-1.5 (Baker et al., 1998) is a lexical resource which groups words based on their semantics.[3] We also used a star-rated movie review corpus of 5006 files (Pang and Lee, 2005) to extract the pivot for each semantic category.[4] Though our approach uses a corpus, its use is limited to identification of pivot. Intensity ranking of other words of the semantic category is derived by exploiting the cosine-

[2]Sharma et al., (2015) also have presented their results using the same 52 polar semantic categories of the FrameNet data.

[3]Available at: https://framenet.icsi.berkeley.edu/fndrupal/about.

[4]Available at: http://www.cs.cornell.edu/people/pabo/movie-review-data/.

similarity between word embeddings of the pivot and the other words of the semantic category. For all three types of word embeddings, we have used precomputed 300 dimensional vectors of words.[5][6]

5 Gold Standard Data Preparation

The objective of our work is to obtain a continuous ranking among words having the same semantics as per FrameNet data. We asked 5 annotators[7] to rank words in each semantic category on a scale of -50 to $+50$. Here, -50 represents the most negatively intense point and $+50$ represents the most positively intense point on the scale. 0 represents a neutral (neither positive nor negative) point on the scale. It is hard to get any neutral word in the data as we have used only polar semantic categories of the FrameNet. The final ranking scale in a category is obtained by averaging the score assigned by all 5 annotators. For example, for a word, if annotator-1 gave ranking r1, annotator-2 gave ranking r2, annotator-3 gave ranking r3, annotator-4 gave ranking r4 and annotator-5 gave ranking r5, then final ranking is ((r1+r2+r3+r4+r5)/5).

To check the agreement among 5 annotators, we computed Fleiss' kappa. It is a statistical measure of inter-rater reliability. Fleiss' kappa is chosen over Scott's pi and Cohen's kappa, because these measures work for two raters, whereas Fleiss' kappa works for any number of raters giving categorical ratings to a fixed number of items (Fleiss, 1971). We obtained a Fleiss' kappa score of **0.64** by dividing words of the semantic category into six levels (high-positive, medium-positive, low-positive, low-negative, medium-negative, high-negative).

6 Approach: Derive Intensity Ordering Among Words

Our approach establishes a continuous intensity ranking among words based on the following hypotheses:

Hypothesis-1 The classic *semantic bleaching theory* states that a word which has fewer number of senses (possibly one) tends to have a higher intensity in comparison to words having more senses.

Hypothesis-2 Semantically similar words that have fewer number of senses exhibit higher cosine-similarity with each other in comparison to words having many senses. Essentially, fewer number of senses cause fewer number of context words or vice versa.

Considering hypothesis-1 and 2 as a base, Sharma et al., (2015) claimed that the word embeddings (context vectors) of high intensity words depict higher cosine-similarity with each other than with low or medium intensity words. However, they used word embeddings which capture only syntactic and semantic similarity among words (Mikolov et al., 2013). Our approach uses SSWE, which integrate sentiment information with the normal word embeddings. Use of SSWE in place of normal word embeddings provides a more accurate cosine-similarity scores, which in turn leads to a more accurate continuous intensity scale. Section 6.1 describes how a high intensity word (pivot) for each semantic category is extracted from an intensity annotated corpus. Section 6.2 presents the algorithm that assigns intensity ordering to words of a semantic category using the pivot (high intensity) word.

6.1 Pivot Selection Method

An amalgamation of χ^2 test and Weighted Normalized Polarity Intensity (WNPI) formula extracts a high intensity word as pivot for each semantic category from the 5 star-rated review corpus. χ^2 test assures that no biased word should be selected as the pivot (Oakes and Farrow, 2007).[8] By biased word we mean that a word which has very few occurrences in the corpus, but these occurrences are in the high star-rated reviews. For example, in our corpus, the word *lame* occurs only 3 times in the corpus, and these occurrences happen to be in 1-star (negatively high intense) reviews only. In addition, χ^2 test derives polarity orientation of the pivot from the corpus as it associates a class (positive or negative) label with the word (Sharma and Bhattacharyya, 2013).

The WNPI formula assigns a intensity score to

[5] In this work, we have opted for precomputed word embeddings, because they are trained on sufficiently large corpora and widely tested for NLP applications.

[6] Embeddings are available here: word2vec (trained on news corpus of 320M words): https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit, GloVe (trained on Wikipedia 2014): http://nlp.stanford.edu/projects/glove/, SSWE (trained on 91M tweets): http://ir.hit.edu.cn/~dytang/.

[7] Two of the annotators are professional linguists of English language and other three are post graduate students.

[8] The details about the *goodness of fit chi*^{2} test: http://stattrek.com/chi-square-test/goodness-of-fit.aspx?Tutorial=AP.

words based on their frequency count in different star ratings. It is defined based on the concept that a high intensity word would occur more frequently in high star-rated reviews, for example, *outstanding* would occur more frequently in 5-star or 4-star reviews in comparison to 1,2,3-star reviews. In the WNPI formula (Algorithm 1), the value of i ranges from 1 to 5, here star rating is used as intensity of the review. The algorithm extracts two pivots for each category, one positive pivot for positive words and one negative pivot for negative words. For the sake of simplicity, we have used the term 'pivot' only in the Algorithm 1. For a positive word '5-star' is treated as 'i=5' (highest positive intensity) and for a negative word '1-star' is treated as 'i=5' (highest negative intensity) in the WNPI formula. A word which gets the highest score by the χ^2 test and the WNPI formula is set as positive (or negative) pivot.

6.2 Algorithm

Algorithm 1 illustrates the sequence of steps carried out to obtain the intensity ordering of words within a semantic category. c_p^w and c_n^w are the counts of a word w in the positive and negative documents respectively. μ^w is an average of c_p^w and c_n^w. To obtain the values of c_p^w and c_n^w, we divided the 5 star-rated review corpus in two equal parts as the positive corpus and the negative corpus. C_i is the count of a word in i intensity documents. Polarity of the words other than the pivot words is inferred by computing the cosine-similarity between SSWE of other words with the SSWE of the pivot word. Since SSWE have sentiment information, a positive pivot gives *positive* cosine-similarity with the positive words and *negative* cosine-similarity with the negative words.[9] Cosine-similarity order between SSWE of the pivot and other words establishes intensity ranking among words of a semantic category.

7 Results and Experimental Setup

To evaluate the efficacy of our SSWE-based approach over word2vec-based system (state-of-the-art) (Sharma et al., 2015) and GloVe-based system, we compute rank correlation and Macro-F1 between the intensity ranking produced by the embeddings and the gold standard intensity ranking.

[9] Sharma et al., (Sharma et al., 2015) used Bing Liu's lexicon in their approach to identify polarity orientation of words. The use of SSWE in our approach helped us to remove the need of a sentiment lexicon to identify polarity of words.

Algorithm 1: Generating an Intensity ordering of words within a semantic category

Input: Set of words within a semantic category W_{sc} ;
Intensity (i) annotated corpus C ;
Pre-trained Sentiment embeddings $SSWE$.

Output: Ranking of words based on intensity.

1 **for** *each word* $w_i \in W_{sc}$ **do**
2 $\quad \chi^2(w_i) = ((c_p^w - \mu^w)^2 + (c_n^w - \mu^w)^2)/\mu^w$
3 $\quad WNPI(w_i) = \frac{\sum_{i=1}^{5} i*C_i}{5*\sum_{i=1}^{5} C_i}$
4 $\quad$ Store in dictionary $(w_i, \chi^2(w_i), WNPI(w_i))$
5 Select word from the dictionary with the highest χ^2 and WNPI score as pivot.
6 **for** *each word* w_i *in* W_{sc} **do**
7 $\quad$ Calculate Cosine-Similarity between
8 $\quad (SSWE_{(w_i)}, SSWE_{(Pivot)})$
9 Words arranged in increasing order of their cosine-similarity is the Intensity Ordering.

7.1 Rank Correlation

Table 1 shows the average rank correlation coefficient obtained for 52 polar semantic categories of the FrameNet data using Spearman's ρ and Kendall's τ. Spearman's ρ measures the degree of association between the two rankings. Kendall's τ finds the number of concordant and discordant pairs in the rankings to measure association. We observed that SSWE-based system results in a significantly better ρ and τ as per t-test.

Embeddings	Spearman's ρ	Kendall's τ
Glove	0.525	0.425
Word2vec	0.71	0.565
SSWE	**0.82**	**0.70**

Table 1: Spearman's ρ and Kendall's τ rank correlations.

7.2 F1 Measure

In order to compare our work with the state-of-the-art (Sharma et al., 2015), the intensity ordering of words within a semantic category is divided into 3 levels, i.e, *low*, *medium* and *high* for both the positive and negative words respectively. In order to create three levels, we placed 2 break points in

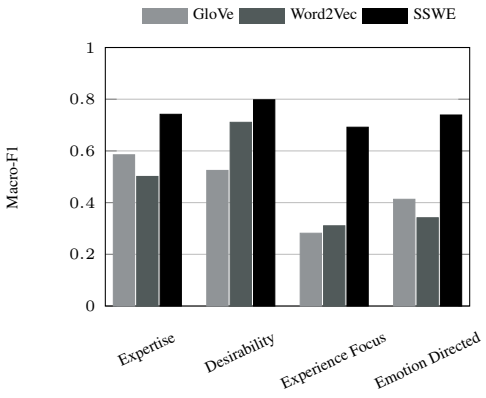

Figure 1: Individual Macro-F1 score for the 4 semantic categories.

the intensity ordering sequence where consecutive similarity scores differ the most.[10] Comparison of Macro-F1 scores for 4 different categories is shown in Figure 1. SSWE outperforms word2vec and GloVe by a big margin in all 4 cases. In addition, we obtain an average Macro-F1 score of 74.32% with SSWE, 54.38% with word2vec and 45.10% with GloVe for the 52 semantic categories.

7.3 Error Analysis

In a few semantic categories of the FrameNet data, words are not confined to any one sentiment and to say that one kind of sentiment has a higher intensity than the other is difficult at times. For example, it is difficult to compare *sadness* and *embarrassment* relatively in terms of intensity, whereas both the words belong to the same semantic category, that is, *emotion directed* as per FrameNet data. In addition, annotators mutually agreed on the fact that when there are limited number of words then it is easier and logical to scale them. More separation based on the finer semantic property within the existing semantic category of the FrameNet data may bring on improvement in the performance of automatic intensity ranking systems.

8 Conclusion

In this paper, we have given a technique that uses Sentiment Specific Word Embeddings (SSWE) to produce a fine-grained intensity ordering among polar words which bear the same semantics. In addition, the use of sentiment embeddings reduces the need of sentiment lexicon for identification of polarity orientation of words. Results show that SSWE are significantly better than word2vec and GloVe, which do not capture sentiment information of words for intensity ranking task. Sentiment intensity information of words can be used in various NLP applications, for example, star-rating prediction, normalization of over-expressed or under-expressed texts, *etc*.

[10]The same break point convention is followed by Sharma et al., (2015) to assign intensity levels to words.

Acknowledgments

We heartily thank English linguists Rajita Shukla and Jaya Saraswati from CFILT Lab, IIT Bombay for giving their valuable contribution in the gold standard data creation.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Eduard C Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. 2010. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1761–1764. ACM.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, pages 60–63.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 172–182. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Michael P Oakes and Malcolm Farrow. 2007. Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. *Literary and Linguistic Computing*, 22(1):85–99.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. *EACL 2014*, 117.

Raksha Sharma and Pushpak Bhattacharyya. 2013. Detecting domain dedicated polar words. In *IJCNLP*, pages 661–666.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. *EMNLP2015, Lisbon, Portugal*.

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

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press*.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.