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).

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

1Words 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.
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; Fahrm 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-

---

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

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

⁴Available 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.56

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 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 \(\chi^2\) 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. \(\chi^2\) test assures that no biased word should be selected as the pivot (Oakes and Farrow, 2007). 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, \(\chi^2\) 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 an intensity score to

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


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, or 3-star reviews. In the WNPI formula (Algorithm 1), the value of \( i \) ranges from 1 to 5, where 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 \( \chi^2 \) 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. \( \mu^w \) 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.

---

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. \[ \chi^2(w_i) = ((c_p^w - \mu^w)^2 + (c_n^w - \mu^w)^2)/\mu^w \]
3. \[ WNPI(w_i) = \sum_{i=1}^{5} \frac{1}{C_i} \sum_{w \in \text{dictionary}} (w_i, \chi^2(w_i), WNPI(w_i)) \]
4. Select word from the dictionary with the highest \( \chi^2 \) and WNPI score as pivot.
5. **for each word** \( w_i \) \( \in \) \( W_{sc} \) **do**
6. Calculate Cosine-Similarity between \( (SSWE(w_i), SSWE_{(Pivot)}) \)
7. 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 \( \rho \) and Kendall’s \( \tau \). Spearman’s \( \rho \) measures the degree of association between the two rankings. Kendall’s \( \tau \) 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 \( \rho \) and \( \tau \) as per t-test.

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>Spearman’s ( \rho )</th>
<th>Kendall’s ( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>0.525</td>
<td>0.425</td>
</tr>
<tr>
<td>Word2vec</td>
<td>0.71</td>
<td>0.565</td>
</tr>
<tr>
<td>SSWE</td>
<td><strong>0.82</strong></td>
<td><strong>0.70</strong></td>
</tr>
</tbody>
</table>

Table 1: Spearman’s \( \rho \) and Kendall’s \( \tau \) 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...
the intensity ordering sequence where consecutive similarity scores differ the most. 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.

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


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