Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling

Svetlana Kiritchenko and **Saif M. Mohammad**
National Research Council Canada
{svetlana.kiritchenko,saif.mohammad}@nrc-cnrc.gc.ca

Abstract

Access to word–sentiment associations is useful for many applications, including sentiment analysis, stance detection, and linguistic analysis. However, manually assigning fine-grained sentiment association scores to words has many challenges with respect to keeping annotations consistent. We apply the annotation technique of Best–Worst Scaling to obtain real-valued sentiment association scores for words and phrases in three different domains: general English, English Twitter, and Arabic Twitter. We show that on all three domains the ranking of words by sentiment remains remarkably consistent even when the annotation process is repeated with a different set of annotators. We also, for the first time, determine the minimum difference in sentiment association that is perceptible to native speakers of a language.

1 Introduction

Word–sentiment associations, commonly captured in sentiment lexicons, are useful in automatic sentiment prediction (Pontiki et al., 2014; Rosenthal et al., 2014), stance detection (Mohammad et al., 2016a; Mohammad et al., 2016b), literary analysis (Hartner, 2013; Kleres, 2011; Mohammad, 2012), detecting personality traits (Grijalva et al., 2015; Mohammad and Kiritchenko, 2015), and other applications. Manually created sentiment lexicons are especially useful because they tend to be more accurate than automatically generated lexicons; they can be used to automatically generate large-scale lexicons (Tang et al., 2014; Esuli and Sebastiani, 2006); they can be used to evaluate different methods of automatically creating sentiment lexicons; and they can be used for linguistic analyses such as examining how sentiment is composed in phrases and sentences.

The sentiment of a phrase can differ significantly from the sentiment of its constituent words. Sentiment composition is the determining of sentiment of a multi-word linguistic unit, such as a phrase or a sentence, from its constituents. Lexicons that include sentiment associations for phrases as well as for their constituent words are useful in studying sentiment composition. We refer to them as *sentiment composition lexicons (SCLs)*. We created SCLs for three domains, and all three were used in recent SemEval shared tasks. We refer to the lexicon created for the English Twitter domain as the *SemEval-2015 English Twitter Sentiment Lexicon*; for the general English domain as the *SemEval-2016 General English Sentiment Modifiers Lexicon*; and for the Arabic Twitter domain as the *SemEval-2016 Arabic Twitter Sentiment Lexicon*. Note that the English Twitter lexicon was first described in (Kiritchenko et al., 2014), whereas the other two are novel contributions presented in this paper.

Most existing manually created sentiment lexicons tend to provide only lists of positive and negative words with very coarse levels of sentiment (Stone et al., 1966; Hu and Liu, 2004; Wilson et al., 2005; Mohammad and Turney, 2013). The coarse-grained distinctions may be less useful in downstream applications than having access to fine-grained (real-valued) sentiment association scores. Only a small number of manual lexicons

Proceedings of NAACL-HLT 2016, pages 811–817,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

capture sentiment associations at a fine-grained level (Bradley and Lang, 1999; Warriner et al., 2013). This is not surprising because obtaining real-valued sentiment annotations has several challenges. Respondents are faced with a higher cognitive load when asked for real-valued sentiment scores for terms as opposed to simply classifying terms as either positive or negative. Besides, it is difficult for an annotator to remain consistent with his/her annotations. Further, the same sentiment association may map to different sentiment scores in the minds of different annotators; for example, one annotator may assign a score of 0.6 and another 0.8 for the same degree of positive association. One could overcome these problems by providing annotators with pairs of terms and asking which is more positive (a comparative approach), however that requires a much larger set of annotations (order N^2, where N is the number of terms to be annotated). Best–Worst Scaling (BWS) is an annotation technique, commonly used in marketing research (Louviere and Woodworth, 1990), that exploits the comparative approach to annotation while keeping the number of required annotations small.

In this work, we investigate the applicability and reliability of the Best–Worst Scaling annotation technique in capturing word–sentiment associations via crowdsourcing. Our main contributions are as follows:

1. We create fine-grained sentiment composition lexicons for Arabic Twitter and general English (in addition to our earlier work on English Twitter) using Best–Worst Scaling. The lexicons include entries for single words as well as multi-word phrases. The sentiment scores are real values between -1 (most negative) and +1 (most positive).

2. We show that the annotations on all three domains are reliable; re-doing the annotation with different sets of annotators produces a very similar order of terms—an average Spearman rank correlation of 0.98. Furthermore, we show that reliable rankings can be obtained with just two or three annotations per BWS question. (Warriner et al. (2013) and Graham et al. (2015) have shown that conventional rating-scale methods require a much higher number of responses (15 to 20)).

3. We examine the relationship between ‘difference in the sentiment scores between two terms’ and ‘agreement amongst annotators’ when asked which term is more positive. We show that agreement grows rapidly and reaches 90% when the difference in sentiment scores is about 0.4 (20% of interval between -1 and 1).

4. We calculate the minimum difference in sentiment scores of two terms that is perceptible to native speakers of a language. For sentiment scores between -1 (most negative) and 1 (most positive), we show that the perceptible difference is about 0.08 for English and Arabic speakers. Knowing the least perceptible difference helps researchers better understand sentiment composition. For example, consider the task of determining whether an adjective significantly impacts the sentiment of the noun it qualifies. This can be accomplished by determining whether the difference in sentiment scores between the combined phrase and the constituent noun alone is greater than the least perceptible difference.

The data and code created as part of this project (the lexicons, the annotation questionnaire, and the code to generate BWS questions) are made available.[1]

2 Capturing Fine-Grained Sentiment Associations By Manual Annotation

We now describe how we created three lexicons, through manual annotation, that each provide real-valued sentiment association scores.

2.1 Best–Worst Scaling Method of Annotation

Best–Worst Scaling (BWS), also sometimes referred to as Maximum Difference Scaling (MaxDiff), is an annotation scheme that exploits the comparative approach to annotation (Louviere and Woodworth, 1990; Cohen, 2003; Louviere et al., 2015). Annotators are given four items (4-tuple) and asked which item is the Best (highest in terms of the property of interest) and which is the Worst (least in terms of the property of interest). These annotations can then be easily converted into real-valued scores of association between the items and the property, which eventually allows for creating a ranked list of items as per their association with the property of interest.

[1] www.saifmohammad.com/WebPages/BestWorst.html

Given n terms to be annotated, the first step is to randomly sample this set (with replacement) to obtain sets of four terms each, *4-tuples*, that satisfy the following criteria:

1. no two 4-tuples have the same four terms;
2. no two terms within a 4-tuple are identical;
3. each term in the term list appears approximately in the same number of 4-tuples;
4. each pair of terms appears approximately in the same number of 4-tuples.

In practice, around $1.5 \times n$ to $2 \times n$ BWS questions, where n is the number of items, are sufficient to obtain reliable scores. We annotated terms for the three lexicons separately, and generated $2 \times n$ 4-tuples for each set.

Next, the sets of 4-tuples were annotated through a crowdsourcing platform, CrowdFlower. The annotators were presented with four terms at a time, and asked which term is the most positive (or least negative) and which is the most negative (or least positive). Below is an example annotation question.[2] (The Arabic data was annotated through a similar questionnaire in Arabic.)

Focus terms:
1. th*nks   2. doesn't work   3. w00t   4. #theworst

Q1: Identify the term that is associated with the most amount of positive sentiment (or least amount of negative sentiment) – **the most positive term**:
1. th*nks   2. doesn't work   3. w00t   4. #theworst

Q2: Identify the term that is associated with the most amount of negative sentiment (or least amount of positive sentiment) – **the most negative term**:
1. th*nks   2. doesn't work   3. w00t   4. #theworst

Each 4-tuple was annotated by ten respondents.

The responses were then translated into real-valued scores and also a ranking of terms by sentiment for all the terms through a simple counting procedure: For each term, its score is calculated as the percentage of times the term was chosen as the most positive minus the percentage of times the term was chosen as the most negative (Orme, 2009; Flynn and Marley, 2014). The scores range from -1 (the most negative) to 1 (the most positive).

2.2 Lexicons Created With Best–Worst Scaling

SemEval-2015 English Twitter Lexicon: This lexicon is comprised of 1,515 high-frequency English single words and simple negated expressions commonly found in tweets. The set includes regular English words as well as some misspelled words (e.g., *parlament*), creatively-spelled words (e.g., *happeee*), hashtagged words (e.g., *#loveumom*), and emoticons.

SemEval-2016 Arabic Twitter Lexicon: This lexicon was created in a similar manner as the English Twitter Lexicon but using Arabic words and negated expressions commonly found in Arabic tweets. It has 1,367 terms.

SemEval-2016 General English Sentiment Modifiers Lexicon aka Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA): This lexicon consists of all 1,621 positive and negative single words from Osgood's seminal study on word meaning (Osgood et al., 1957) available in General Inquirer (Stone et al., 1966). In addition, it includes 1,586 high-frequency phrases formed by the Osgood words in combination with simple negators such as *no*, *don't*, and *never*, modals such as *can*, *might*, and *should*, or degree adverbs such as *very* and *fairly*. More details on the lexicon creation and an analysis of the effect of different modifiers on sentiment can be found in (Kiritchenko and Mohammad, 2016).

Table 1 shows example entries from each lexicon. The complete lists of modifiers used in the three lexicons are available online.[3] Details on the use of these lexicons in SemEval shared tasks can be found in (Rosenthal et al., 2015; Kiritchenko et al., 2016).

3 Quality of Annotations

3.1 Agreement and Reproducibility

Let *majority answer* refer to the option chosen most often for a BWS question. The percentage of responses that matched the majority answer were as follows: 82% for the English Twitter Lexicon, 80% for the Arabic Twitter Lexicon, and 80% for the General English Lexicon.

[2] The full sets of instructions for both English and Arabic datasets are available at:
http://www.saifmohammad.com/WebPages/BestWorst.html

[3] www.saifmohammad.com/WebPages/SCL.html#ETSL
www.saifmohammad.com/WebPages/SCL.html#ATSL
www.saifmohammad.com/WebPages/SCL.html#NMA

Table 1: Example entries from the three lexicons.

Lexicon, Term	Sentiment
SemEval-2015 English Twitter Lexicon	
yummm	0.813
cant waitttt	0.656
#feelingsorryformyself	-0.547
:'(	-0.563
SemEval-2016 Arabic Twitter Lexicon	
#السعادة_الزوجية (marital happiness)	0.800
#يقين (certainty)	0.675
لا امكن (not possible)	-0.400
ارهاب (terrorism)	-0.925
SemEval-2016 General English Lexicon	
would be very easy	0.431
did not harm	0.194
increasingly difficult	-0.583
severe	-0.833

Annotations are reliable if similar results are obtained from repeated trials. To test the reliability of our annotations, we randomly divided the sets of ten responses to each question into two halves and compared the rankings obtained from these two groups of responses. The Spearman rank correlation coefficient between the two sets of rankings produced for each of the three lexicons was found to be at least 0.98. (The Pearson correlation coefficient between the two sets of sentiment scores for each lexicon was also at least 0.98.) Thus, even though annotators might disagree about answers to individual questions, the aggregated scores produced by applying the counting procedure on the BWS annotations are remarkably reliable at ranking terms.

Number of annotations needed: Even though we obtained ten annotations per BWS question, we wanted to determine the least number of annotations needed to obtain reliable sentiment scores. For every k (where k ranges from 1 to 10), we made the following calculations: for each BWS question, we randomly selected k annotations and calculated sentiment scores based on the selected subset of annotations. We will refer to these sets of scores for the different values of k as S_1, S_2, and so on until S_{10}. This process was repeated ten times for each k. The average Spearman rank correlation coefficient between S_1 and S_{10} was 0.96, between S_2 and S_{10} was 0.98, and S_3 and S_{10} was 0.99. This shows that as few as two or three annotations per BWS question are sufficient to obtain reliable sentiment scores. Note that with $2 \times n$ BWS questions (for n terms), each term occurs in eight 4-tuples on average, and so even just one annotation per BWS question means that each term is assessed eight times.

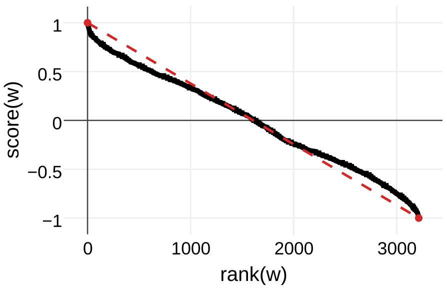

Figure 1: Rank vs. sentiment scores in SCL-NMA.

3.2 Distribution of Sentiment Scores

Figure 1 gives an overview of the sentiment scores in SCL-NMA. Each term in the lexicon is shown as a dot in the corresponding plot. The x-axis is the rank of each term in the lexicon when the terms are ordered from most positive to least positive. The y-axis is the real-valued sentiment score obtained from the BWS annotations. Observe that the lexicon has entries for the full range of sentiment scores (-1 to 1); that is, there are no significant gaps—ranges of sentiment scores for which the lexicon does not include any terms. The dashed red line indicates a uniform spread of scores, i.e., the same number of terms are expected to fall into each same-size interval of scores. Observe that the lexicon has slightly fewer terms in the intervals with very high and very low sentiment scores. Similar figures (not shown here) were obtained for the other two lexicons.

3.3 Perception of Sentiment Difference

The created lexicons capture sentiment associations at a fine level of granularity. Thus, these annotations can help answer key questions such as: (1) If native speakers of a language are given two terms and asked which is more positive, how does human agreement vary with respect to the amount of difference in sentiment between the two focus terms? It is expected that the greater the difference in sentiment, the higher the agreement, but the exact shape of this increase in agreement has not been shown till now. (2) What least amount of difference in sentiment is perceptible to native speakers of a language?

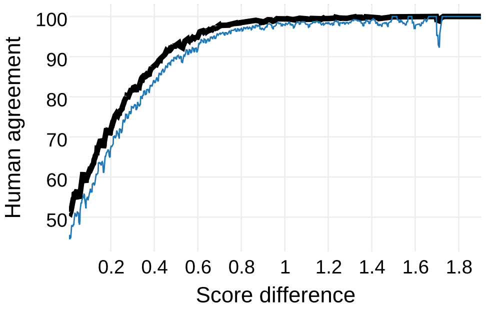

Figure 2: SCL-NMA: Human agreement on annotating term w_1 as more positive than term w_2 for pairs with difference in scores d = $score(w_1)$ - $score(w_2)$. The x-axis represents d. The y-axis plots the avg. percentage of human annotations that judge term w_1 as more positive than term w_2 (thick line) and the corresponding 99.9%-confidence lower bound (thin blue line).

Agreement vs. Sentiment Difference: For each word pair w_1 and w_2 such that $score(w_1) - score(w_2) \geq 0$, we count the number of BWS annotations from which we can infer that w_1 is more positive than w_2 and divide this number by the total number of BWS annotations from which we can infer either that w_1 is more positive than w_2 or that w_2 is more positive than w_1. (We can infer that w_1 is more positive than w_2 if in a 4-tuple that has both w_1 and w_2 the annotator selected w_1 as the most positive or w_2 as the least positive. The case for w_2 being more positive than w_1 is similar.) This ratio is the human agreement for w_1 being more positive than w_2, and we expect that it is correlated with the sentiment difference $d = score(w_1) - score(w_2)$. To get more reliable estimates, we average the human agreement for all pairs of terms whose sentiment differs by $d \pm 0.01$. Figure 2 shows the resulting average human agreement on SCL-NMA. Similar figures (not shown here) were obtained for the English and Arabic Twitter data. Observe that the agreement grows rapidly with the increase in score differences. Given two terms with sentiment differences of 0.4 or higher, more than 90% of the annotators correctly identify the more positive term.

Least Difference in Sentiment that is Perceptible to Native Speakers: In psychophysics, there is a notion of *least perceptible difference* (aka *just-noticeable difference*)—the amount by which something that can be measured (e.g., weight or sound intensity) needs to be changed in order for the difference to be noticeable by a human (Fechner, 1966). Analogously, we can measure the least perceptible difference in sentiment. If two words have close to identical sentiment associations, then it is expected that native speakers will choose each of the words about the same number of times when forced to pick a word that is more positive. However, as the difference in sentiment starts getting larger, the frequency with which the two terms are chosen as most positive begins to diverge. At one point, the frequencies diverge so much that we can say with high confidence that the two terms do not have the same sentiment associations. The average of this minimum difference in sentiment score is the least perceptible difference for sentiment. To determine the least perceptible difference, we first obtain the 99.9%-confidence lower bounds on the human agreement (see the thin blue line in Figure 2). The least perceptible difference is the point starting at which the lower bound consistently exceeds 50% threshold (i.e., the point starting at which we observe with 99.9% confidence that the human agreement is higher than chance). The least perceptible difference when calculated from SCL-NMA is 0.069, from the English Twitter Lexicon is 0.080, and from the Arabic Twitter Lexicon is 0.087. Observe, that the estimates are very close to each other despite being calculated from three completely independent datasets. Kiritchenko and Mohammad (2016) use the least perceptible difference to determine whether a modifier significantly impacts the sentiment of the word it composes with.

4 Conclusions

We obtained real-valued sentiment association scores for single words and multi-word phrases in three domains (general English, English Twitter, and Arabic Twitter) by manual annotation and Best–Worst Scaling. Best–Worst Scaling exploits the comparative approach to annotation while keeping the number of annotations small. Notably, we showed that the procedure when repeated produces remarkably consistent rankings of terms by sentiment. This reliability allowed us to determine the value of the psycho-linguistic concept—least perceptible difference in sentiment. We hope these findings will encourage further use of Best–Worst Scaling in linguistic annotation.

References

Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Steven H. Cohen. 2003. Maximum difference scaling: Improved measures of importance and preference for segmentation. Sawtooth Software, Inc.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 417–422.

Gustav Fechner. 1966. *Elements of psychophysics. Vol. I.* New York: Holt, Rinehart and Winston.

T. N. Flynn and A. A. J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, pages 178–201. Edward Elgar Publishing.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the Annual Conference of the North American Chapter of the ACL (NAACL)*, pages 1183–1191.

Emily Grijalva, Daniel A. Newman, Louis Tay, M. Brent Donnellan, P.D. Harms, Richard W. Robins, and Taiyi Yan. 2015. Gender differences in narcissism: A meta-analytic review. *Psychological bulletin*, 141(2):261–310.

Marcus Hartner. 2013. The lingering after-effects in the reader's mind – an investigation into the affective dimension of literary reading. *Journal of Literary Theory Online*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, New York, NY, USA.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.

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

Svetlana Kiritchenko, Saif M. Mohammad, and Mohammad Salameh. 2016. SemEval-2016 Task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, San Diego, California, June.

Jochen Kleres. 2011. Emotions and narrative analysis: A methodological approach. *Journal for the Theory of Social Behaviour*, 41(2):182–202.

Jordan J. Louviere and George G. Woodworth. 1990. Best-worst analysis. Working Paper. Department of Marketing and Economic Analysis, University of Alberta.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, Submitted.

Saif M Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.

Charles E Osgood, George J Suci, and Percy Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, Dublin, Ireland.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 73–80, Dublin, Ireland, August.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 450–462, Denver, Colorado.

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General*

Inquirer: A Computer Approach to Content Analysis. The MIT Press.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale Twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 172–182.

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.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Joint Conference on HLT and EMNLP*, pages 347–354, Stroudsburg, PA, USA.