Happy Accident: A Sentiment Composition Lexicon for Opposing Polarity Phrases

Svetlana Kiritchenko and Saif M. Mohammad
National Research Council Canada
1200 Montreal Rd., Ottawa, ON, Canada
svetlana.kiritchenko@nrc-cnrc.gc.ca, saif.mohammad@nrc-cnrc.gc.ca

Abstract

Sentiment composition is the determining of sentiment of a multi-word linguistic unit, such as a phrase or a sentence, based on its constituents. We focus on sentiment composition in phrases formed by at least one positive and at least one negative word—phrases like *happy accident* and *best winter break*. We refer to such phrases as *opposing polarity phrases*. We manually annotate a collection of opposing polarity phrases and their constituent single words with real-valued sentiment intensity scores using a method known as *Best–Worst Scaling*. We show that the obtained annotations are consistent. We explore the entries in the lexicon for linguistic regularities that govern sentiment composition in opposing polarity phrases. Finally, we list the current and possible future applications of the lexicon.
Keywords: sentiment composition, sentiment lexicon, opposing polarity phrases, Best–Worst Scaling, crowdsourcing

1. Introduction

Words have associations with sentiment. For example, *honest* and *competent* are associated with positive sentiment, whereas *dishonest* and *dull* are associated with negative sentiment. Further, the degree of positivity (or negativity), also referred to as sentiment intensity, can vary. For example, most people will agree that *succeed* is more positive (or less negative) than *improve*, and *failure* is more negative (or less positive) than *decline*.

Sentiment associations are commonly captured in *sentiment lexicons*—lists of associated word–sentiment pairs (optionally with a score indicating the degree of association). They are mostly used in sentiment analysis, but are also valuable in stance detection (Mohammad et al., 2016a; Mohammad et al., 2016b), literary analysis (Hartner, 2013; Kleres, 2011), and other applications.

Manually created sentiment lexicons usually include only single words. However, 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, based on its constituents. Lexicons that include sentiment associations for phrases as well as their constituent words can be very useful in studying sentiment composition. We will refer to them as *sentiment composition lexicons (SCLs)*.

In this work, we focus on sentiment composition in phrases that include at least one positive and at least one negative word—for example, phrases such as *happy accident*, *best winter break*, *couldn't stop smiling*, and *lazy sundays*.[1] We refer to them as *opposing polarity phrases*. We describe how we created a sentiment composition lexicon for opposing polarity phrases and their constituent words.

Most existing manually created sentiment lexicons provide only lists of positive and negative words with very coarse levels of sentiment (Stone et al., 1966; 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 (Taboada et al., 2011). However, obtaining real-valued sentiment annotations is challenging for several reasons. 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. Further, it is difficult for an annotator to remain consistent with his/her annotations. 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 of N^2, where N is the number of terms to be annotated).

Here, in contrast to most previous work on sentiment annotation, we create a lexicon that provides real-valued scores of association of a phrase with positive sentiment. For this, we employ the *Best–Worst Scaling* method of annotation, which is commonly used in marketing research (Louviere and Woodworth, 1990). It exploits the comparative approach to annotation while keeping the number of annotations small. When applied on the task of sentiment annotation, Best–Worst Scaling has been shown to produce remarkably consistent annotations of terms (Kiritchenko and Mohammad, 2016a).

In this paper, we describe how we compiled real-valued sentiment association scores for opposing polarity phrases and their constituents through Best-Worst Scaling. We refer to this resource as the *Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP)*. The lexicon includes entries for 265 trigrams, 311 bigrams, and 602 unigrams. We show that re-doing the annotation with different sets of annotators produces consistent rankings of terms by sentiment, proving that the obtained sentiment scores are reliable.

We explore the entries in SCL-OPP in search for linguistic regularities that govern the sentiment composition in opposing polarity phrases. We also show the frequency distribution of opposing polarity phrases pertaining to various part-of-speech sequences. The most common sequence in our dataset is 'Adjective + Noun'. We observe that adjec-

[1] Observe that *lazy* is associated with negative sentiment whereas *sundays* is associated with positive sentiment.

tives and verbs are frequently the primary source of sentiment in the phrase; however, some nouns can override their sentiment, as in *new crisis* or *leave a #smile*. Overall, sentiment composition for many part-of-speech sequences is not straightforward, and SCL-OPP has phrases corresponding to many different cases. In related work (not described here), we also created a sentiment composition lexicon for negators, modals, and adverbs (SCL-NMA). We refer the readers to (Kiritchenko and Mohammad, 2016b) for more details on that lexicon. Lexicons such as SCL-OPP and SCL-NMA, which include entries for phrases as well as their constituents, are useful in understanding how meaning (especially sentiment) is composed. They are made freely available to the research community.[2]

2. Related Work

Sentiment Lexicons: There exist a number of manually created lexicons that provide lists of positive and negative words, for example, General Inquirer (Stone et al., 1966), Hu and Liu Lexicon (Hu and Liu, 2004), MPQA Subjectivity Lexicon (Wilson et al., 2005), and NRC Emotion Lexicon (Mohammad and Turney, 2013). Only a few manually created lexicons provide real-valued scores for sentiment. ANEW is a collection of 1,034 English words manually rated for valence, arousal, and dominance using a 9-point rating scale (Bradley and Lang, 1999). Warriner et al. (2013) extended this list to include 13,915 English lemmas, again manually rated for valence, arousal, and dominance on a 9-point scale. In a similar way, the LabMT lexicon was created (Dodds et al., 2011). It provides real-valued estimates of association of English single words with happiness. Later, the lexicon was extended to include frequently used words from ten languages (Dodds et al., 2015). None of these lexicons, however, contain multi-word phrases.

Manually created sentiment lexicons can be used to automatically generate larger sentiment lexicons using semi-supervised techniques (Esuli and Sebastiani, 2006; Turney and Littman, 2003; Mohammad et al., 2013). Automatically collected lexicons often have real-valued sentiment association scores, are larger in scale, and can easily be collected for a specific domain; therefore, they are often more beneficial in downstream applications, such as sentence-level sentiment prediction (Kiritchenko et al., 2014). Yet, their intrinsic evaluation has been limited due to the lack of manually created real-valued sentiment lexicons. Further, any analysis of the relationship between the sentiment of a phrase and its constituents is less reliable when made from an automatically generated resource as opposed to when made from a manually created resource (as automatically generated resources are less accurate). In this work, we create a fine-grained sentiment composition lexicon for opposing polarity phrases through manual annotation.

Annotation techniques: A widely used method of annotation for obtaining numerical scores is the rating scale method—where an annotator is asked to rate an item on a five-, ten-, or hundred-point scale. While easy to understand, rating items on a scale is not natural for people. Different people may assign different scores to the same target item, and it is hard for annotators to remain consistent when annotating a large number of items. Also, respondents can mark many terms as equally positive making the annotations less useful. Furthermore, respondents often use just a limited part of the scale providing a bias and reducing the discrimination among items. To obtain reliable annotations, the rating scale methods require a high number of responses, typically 15 to 20 (Warriner et al., 2013; Graham et al., 2015).

A more natural annotation task for humans is to compare items (e.g., whether one word is more positive than the other). Most commonly, the items are compared in pairs (Thurstone, 1927; David, 1963). In this work, we use Best–Worst Scaling (described in Section 3.2.), which is another annotation technique that exploits the comparative approach to annotation. Best–Worst Scaling forces the respondent to indicate a choice (Best and Worst), while still producing real-valued scores reflective of relative importance. In a small marketing study that compared three scaling techniques (9-point rating scale, paired comparisons, and Best–Worst Scaling), Cohen (2003) observed that Best–Worst Scaling produces more reliable, unbiased, and more discriminating results than the other rating annotation methods do. Kiritchenko and Mohammad (2016a) applied Best–Worst Scaling to annotate terms (words and phrases) for sentiment intensity and showed that it produces remarkably consistent annotations.

3. Creating a Sentiment Lexicon for Opposing Polarity Phrases

In order to create a real-valued sentiment composition lexicon for opposing polarity phrases, we first selected a diverse set of opposing polarity ngrams and their constituents (Section 3.1.), and then annotated them for sentiment through Best–Worst Scaling and crowdsourcing (Section 3.2.). Re-doing the annotation with different sets of annotators produced consistent rankings of terms by sentiment, proving that the obtained annotations are reliable (Section 3.3.).

3.1. Term Selection

Opposing polarity phrases frequently occur in many domains. For this work, we chose English tweets as our source of phrases. We polled the Twitter API (from 2013 to 2015) to collect about 11 million tweets that contain emoticons: ':)' or ':('. We will refer to this corpus as the *Emoticon Tweets Corpus*. From this corpus, we selected bigrams and trigrams that had at least one positive word and at least one negative word. The polarity labels (positive or negative) of the words were determined by simple look-up in existing sentiment lexicons: Hu and Liu lexicon (Hu and Liu, 2004), NRC Emotion lexicon (Mohammad and Turney, 2013), MPQA lexicon (Wilson et al., 2005), and NRC's automatically generated Twitter-specific lexicon (Kiritchenko et al., 2014).[3] Apart from the requirement of having at least one positive and at least one negative word, an ngram must satisfy the following criteria:

[2] http://www.saifmohammad.com/WebPages/SCL.html

[3] If a word was marked with conflicting polarity in two lexicons, then that word was not considered as positive or negative.

- the ngram must have a clear meaning on its own, (for example, the ngram should not start or end with a conjunction like '*or*' or '*and*');
- the ngram should not include a named entity;
- the ngram should not include obscene language.

In addition, we ensured that there was a good variety of phrases—for example, even though there were a large number of ngrams of the form *super w*, where *w* is a negative adjective, only a small number of such ngrams were included. Finally, we aimed to achieve a good spread in terms of degree of sentiment association (from very negative terms to very positive terms, and all the degrees of polarity in between). For this, we estimated the sentiment score of each phrase using an automatic PMI-based method described in (Kiritchenko et al., 2014).[4] Then, the full range of sentiment values was divided into 5 bins, and about a hundred bigrams and a hundred trigrams were selected from each bin, except for the middle bin from which only 50 bigrams and 50 trigrams were selected.[5]

In total, 851 ngrams (bigrams and trigrams) were selected. We also chose for annotation all unigrams that appeared in the selected set of bigrams and trigrams.[6] There were 810 such unigrams. The master list consisted of 1,661 terms. Note that since the multi-word phrases and single-word terms were drawn from a corpus of tweets, they include a small number of hashtag words (e.g., *#wantit*) and creatively spelled words (e.g., *plssss*). However, a majority of the terms are those that one would use in everyday English.

3.2. Term Annotation with Best–Worst Scaling

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.

As the first step, the master list of 1,661 terms was randomly sampled (with replacement) to create 3,322 (2 x 1,661) 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.

Next, the set of 4-tuples was annotated through a crowdsourcing platform, CrowdFlower. The annotators were presented with four terms (single words and multi-word phrases) 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.[7]

Focus terms:
1. shameless self promotion 2. happy tears 3. hug
4. major pain

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. shameless self promotion
2. happy tears
3. hug
4. major pain

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. shameless self promotion
2. happy tears
3. hug
4. major pain

Each 4-tuple was annotated by eight respondents. Only native speakers of English residing in the United States were asked to annotate the terms. To ensure data quality, we followed best practices of crowdsourcing, such as providing clear and easy to follow instructions, using check questions for which the correct answer was already determined by internal annotations, and discarding annotations provided by annotators who obtained low accuracy (less than 70%) on the check questions.

The responses were then translated into real-valued scores 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). These real-valued scores also produce a ranking of terms by sentiment.

When selecting the terms, we used sentiment associations obtained from both manual and automatic lexicons. As a result, some unigrams had erroneous sentiment associations. After manually annotating the full set of 1,661 terms (that include unigrams, bigrams, and trigrams), we found that 114 bigrams and 161 trigrams have all their comprising unigrams of the same polarity. These 275 ngrams were discarded from the further analysis. We will refer to the remaining set of 1,178 opposing polarity bigrams, trigrams and their constituent unigrams with the corresponding manually obtained sentiment association scores as the *Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP)*. For analysis in the rest of the paper, we consider an ngram positive if its manually annotated sentiment

[4] Our goal is to create a sentiment lexicon with reliable annotations of terms; therefore, the core sentiment annotations are done manually (as described in Section 3.2.). The automatically estimated sentiment scores obtained with the PMI-based method, which are less reliable than manual annotations, help only to select terms with different degrees of sentiment association.

[5] We wanted to include fewer neutral terms.

[6] Stopword unigrams, namely numbers, auxiliary verbs, pronouns, prepositions, exclamations, articles, and conjunctions, were not selected for annotation.

[7] The full set of instructions to annotators is available at http://www.saifmohammad.com/WebPages/BestWorst.html

Ngrams	Number of terms		
	positive	negative	total
unigrams	292	310	602
bigrams	135	176	311
trigrams	104	161	265
all	531	647	1,178

Table 1: The number of unigrams, bigrams, and trigrams in SCL-OPP.

Term	Sentiment score
Unigrams	
friends	0.703
great	0.625
long	-0.094
breaking	-0.500
loss	-0.672
Bigrams	
bloody great	0.609
long holiday	0.484
great capture	0.250
breaking free	0.172
great loss	-0.734
Trigrams	
best winter break	0.844
long time friends	0.734
nice long walk	0.469
isn't long enough	-0.188
heart breaking moment	-0.797

Table 2: Example entries with real-valued sentiment scores from SCL-OPP.

score is greater than or equal to zero, and negative if its sentiment score is less than zero. Table 1 shows the total number of unigrams, bigrams, and trigrams in SCL-OPP. In addition, the table displays the distribution of positive and negative ngrams in the set. Table 2 shows a few example entries from the lexicon.

3.3. Quality of Annotations

Let *majority answer* refer to the option chosen most often for a question. 81% of the responses to the Best–Worst questions matched the majority answer.

We also tested the consistency of the aggregated scores by randomly dividing the sets of eight responses to each question into two halves and comparing the rankings obtained from these two groups of responses. The Spearman rank correlation coefficient between the two sets of rankings was found to be 0.98. (The Pearson correlation coefficient between the two sets of sentiment scores was also 0.98.) Thus, even though annotators might disagree about answers to individual questions, the aggregated scores produced by applying the counting procedure on the Best–Worst annotations are remarkably reliable at ranking terms by sentiment.

4. Sentiment Composition Patterns

SCL-OPP allows us to explore sentiment composition patterns in opposing polarity phrases. We define a *Sentiment Composition Pattern (SCP)* as a rule that includes on the left-hand side the parts of speech and the sentiment associations of the constituent unigrams (in the order they appear in the phrase), and on the right-hand side the sentiment association of the phrase. The part-of-speech (POS) sequence of a phrase is determined by looking up the most common part-of-speech sequence for that phrase in the Emoticon Tweets Corpus.[8] Table 3 shows all POS sequences for which our lexicon contains more than ten phrases. For each POS sequence, the table lists all SCPs that have at least two examples in the lexicon.[9] The parts of speech are encoded as follows: 'A' stands for adjective, 'N' for noun, 'V' for verb, 'R' for adverb, 'P' for preposition or subordinating conjunction, 'D' for determiner, and '&' for coordinating conjunction. The polarity of the terms is shown with the following symbols: a green '△' denotes a positive word or phrase, and an orange '▽' denotes a negative word or phrase. The table also shows an example phrase for each SCP.

The most frequent POS sequence in the lexicon is 'A+N': there are 141 phrases that have an adjective as the first word and a noun as the second. The entries under 'A+N' in Table 3 show that there are four different SCPs formed by this POS sequence differing in the polarity of the phrase and the polarity of the constituent words. For example, there are 34 positive and 39 negative phrases where the first word is a positive adjective and the second word is a negative noun. In other words, a positive adjective and a negative noun can form either a positive phrase (e.g., *happy accident*) or a negative phrase (e.g., *great loss*). In general, there are many POS sequences (e.g., 'A+N', 'N+N', 'V+D+N', etc.) for which SCPs with the same left-hand side can form either a positive or a negative phrase, and our lexicon contains representatives of both cases.

Some parts of speech impact the sentiment of the phrase in a predictable manner. For example, adverbs often play a role of an intensifier—a word that increases or decreases the sentiment intensity of the following word (e.g., *incredibly slow*, *dearly missed*). Only the intensity of the next word is changed while its polarity (positive or negative) is often preserved. Some adjectives can also play the role of an intensifier when combined with another adjective (e.g., *crazy talented*) or a noun (e.g., *epic fail*). For example, the adjective *great*, often considered highly positive, becomes an intensifier when combined with a noun (e.g., *great loss*, *great capture*). Yet, other adjectives determine the polarity of the entire phrase (e.g., *happy accident*, *bad luck*). Overall, even though adjectives and verbs are frequently the primary source of sentiment in the phrase, some nouns can override their sentiment as in *new crisis* or *leave a #smile*. Since SCL-OPP has phrases corresponding to many different kinds of Sentiment Composition Patterns, it is a useful resource for linguistic studies of sentiment composition as well as for testing automatic techniques that estimate sentiment intensity of opposing polarity phrases.

[8]The corpus was automatically POS tagged using the CMU Tweet NLP tool (Gimpel et al., 2011). A small percentage of phrases were POS tagged incorrectly.

[9]The complete list of all SCPs is available at http://www.saifmohammad.com/WebPages/SCL.html#OPP.

POS sequence SCP	# of phrases	Example
A+N	141	
▽A + △N → ▽phrase	40	bad luck
▽A + △N → △phrase	28	late nap
△A + ▽N → ▽phrase	39	great loss
△A + ▽N → △phrase	34	happy accident
N+N	35	
▽N + △N → ▽phrase	4	split personality
▽N + △N → △phrase	6	sneak peek
△N + ▽N → ▽phrase	13	heart attack
△N + ▽N → △phrase	12	coffee break
R+A	27	
▽R + △A → △phrase	9	ridiculously happy
△R + ▽A → ▽phrase	16	incredibly slow
△R + ▽A → △phrase	2	fashionably late
V+D+N	26	
▽V + D + △N → ▽phrase	11	lost a child
▽V + D + △N → △phrase	6	leave a #smile
△V + D + ▽N → ▽phrase	3	found the debris
△V + D + ▽N → △phrase	6	solved the problem
A+A	20	
▽A + △A → ▽phrase	4	stinking hot
▽A + △A → △phrase	13	crazy talented
△A + ▽A → ▽phrase	2	plain sad
V+N	20	
▽V + △N → ▽phrase	14	losing hope
▽V + △N → △phrase	3	break dance
△V + ▽N → ▽phrase	2	committed suicide
R+V	14	
▽R + △V → ▽phrase	2	barely afford
△R + ▽V → ▽phrase	10	truly missed
N+V	13	
▽N + △V → ▽phrase	2	hospital visit
▽N + △V → △phrase	3	problem solved
△N + ▽V → ▽phrase	7	love hurts
V+P+V	11	
▽V + P + △V → ▽phrase	3	sucks to live
△V + P + ▽V → ▽phrase	5	like to torture
D+A+N	11	
D + ▽A + △N → ▽phrase	4	a bad deal
D + ▽A + △N → △phrase	2	a long nap
D + △A + ▽N → △phrase	4	a good problem
A+N+N	11	
▽A + △N + △N → ▽phrase	3	shameless self promotion
▽A + △N + △N → △phrase	3	long time friends
△A + △N + ▽N → △phrase	2	new hair cut

Table 3: Sentiment composition patterns (SCPs) in SCL-OPP. 'A' stands for adjective, 'N' stands for noun, 'V' stands for verb, 'R' stands for adverb, 'P' stands for preposition or subordinating conjunction, 'D' stands for determiner, and '&' stands for coordinating conjunction. The polarity of the terms is shown with the following symbols: a green '△' denotes a positive word or phrase, and an orange '▽' denotes a negative word or phrase.

5. An Interactive Visualization of SCL-OPP

For ease of exploration of the Sentiment Composition Lexicon for Opposing Polarity Phrases, we created an online interactive visualization.[10] Figure 1 shows a screenshot of default view. It has two components. The main (top) component shows a scatter plot of sentiment scores for opposing polarity phrases. Each point in the scatter plot corresponds to an ngram (a bigram or a trigram). The x-axis is the sentiment score of the first content word in the ngram; the y-axis is the sentiment score of the second content word in the ngram. All bigrams and trigrams that have two content words (and possibly a stop word) are shown. The polarity of the phrase is represented by the color and direction of the triangle: a green '△' corresponds to a positive phrase, and an orange '▽' corresponds to a negative phrase. The size of a triangle is proportional to the absolute value of the phrase's sentiment score. The exact sentiment score of the phrase, as well as the sentiment scores of its constituent words, can be viewed by hovering over the point in the graph with the mouse. Notice that all points lie in the top left and bottom right quadrants of the plot since in each phrase one of the words is positive and the other one is negative.

The second (bottom) visualization component, known as a treemap, shows tiles corresponding to each part-of-speech (POS) sequence present in the dataset. The size (area) of a tile is proportional to the number of instances corresponding to that POS sequence. One can see that the most frequent sequence in the dataset is the bigram where the first word is an adjective and the second word is a noun (A+N).

Interactivity: The two components are linked so as to allow filtering of the information in one component by making a selection in the other component. The main visualization component can be filtered by clicking on the POS sequence of interest in the treemap. For example, clicking on the 'A+N' updates the scatter plot to show only phrases with an adjective as the first word and a noun as the second (see Figure 2). By clicking on a point in the main component, the treemap is updated to show only the tile corresponding to the POS sequence of the selected phrase. The filtering in both components can also be done by checking the boxes corresponding to the POS sequences or polarity of phrases on the right. The selection can further be narrowed down by adjusting the ranges of sentiment scores for the phrase and the constituents using the sliders on the right. Thus, the viewers can easily explore the subsets of the data they are interested in.

6. Applications of SCL-OPP

The Sentiment Composition Lexicon for Opposing Polarity Phrases can be used in various ways. Here we describe a few of its current and possible future applications.

6.1. Linguistic Analysis of Sentiment Composition in Opposing Polarity Phrases

Kiritchenko and Mohammad (2016c) further analyze regularities present in different kinds of phrases in SCL-OPP

[10] http://www.saifmohammad.com/WebPages/SCL.html#OPP

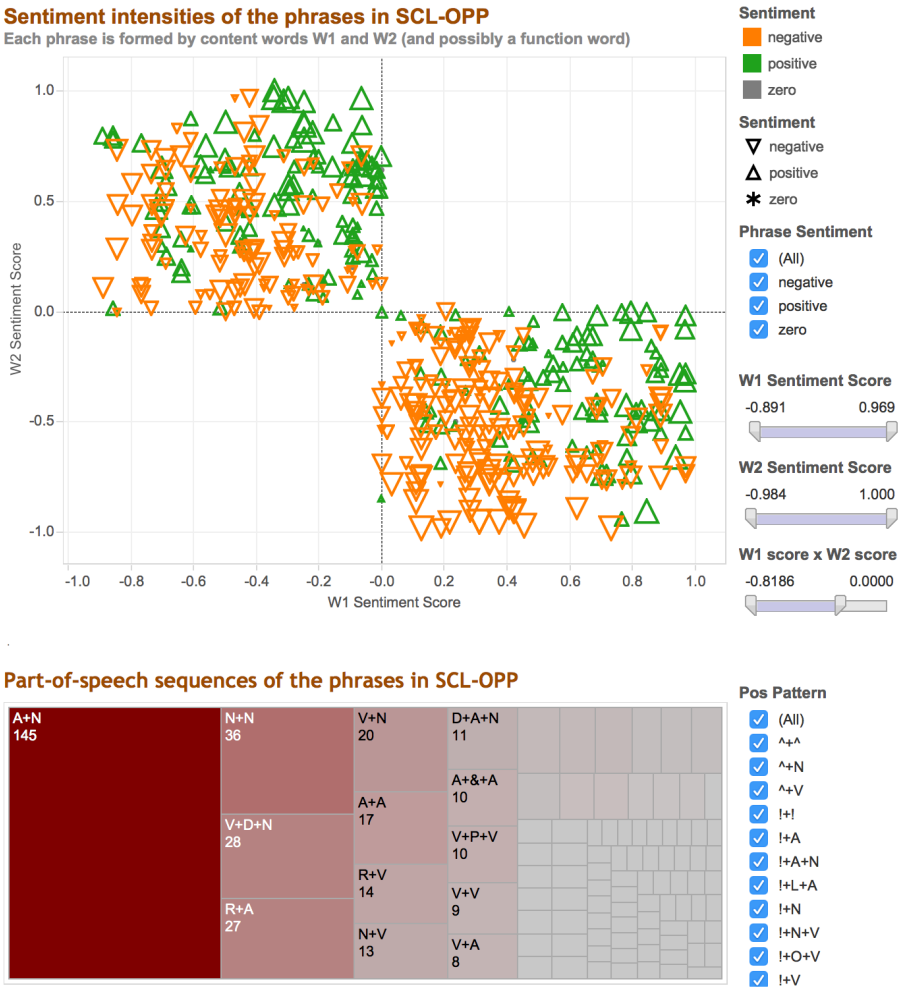

Figure 1: A screenshot of the default view in the interactive visualization for SCL-OPP. The top component shows a scatter plot of sentiment scores for opposing polarity phrases. Each point in the scatter plot corresponds to an ngram (a bigram or a trigram). The x-axis is the sentiment score of the first content word in the ngram; the y-axis is the sentiment score of the second content word in the ngram. The polarity of the phrase is represented by the color and direction of the triangle. The size of a triangle is proportional to the absolute value of the phrase's sentiment score. The bottom component (treemap) shows tiles corresponding to each POS sequence present in the dataset. The size of a tile is proportional to the number of instances corresponding to that POS sequence.

to get insights into how sentiment is composed. They conclude that for most phrases the sentiment of the phrase cannot be reliably predicted only from the parts of speech and polarities of their constituent words. They also propose several unsupervised and supervised techniques for determining sentiment of opposing polarity phrases (sentiment composition). Furthermore, they show that the constituent words, their parts of speech, sentiment scores, and embeddings are all useful features in supervised sentiment prediction on this dataset.

6.2. Evaluating Automatic Methods that Predict Sentiment Intensity of Phrases

Portions of SCL-OPP were used as development and test sets in SemEval-2016 shared task (Task 7) 'Determining Sentiment Intensity of English and Arabic Phrases' (Kiritchenko et al., 2016).[11] The objective of this task was to automatically predict sentiment intensity scores for multiword phrases. The task consisted of three subtasks, one for each of the three domains: general English, English Twitter, and Arabic Twitter. The development and test datasets for the English Twitter domain, *English Twitter Mixed Polarity Sets*, were constructed from the master list described in Section 3.1. They include a large number of opposing polarity phrases, some same polarity phrases, and their single-word constituents. Some terms that appeared in a previous iteration of the shared task were removed. Two hundred terms with the corresponding manual sentiment annotations were released to participants as a development set. A total of 1,069 terms were used as the test set. Five teams submitted nine system outputs for the task. The best result on the English Twitter Mixed Polarity test set was achieved with a supervised method by exploiting a variety of available sentiment resources; the highest Kendall's rank correlation between the predicted and gold term rankings was 0.523.

[11] http://alt.qcri.org/semeval2016/task7/

Figure 2: A screenshot of the interactive visualization when one of POS sequences (A+N) is selected. One can also see the tool tip box showing information about the phrase (*lucky devil*) over which the mouse is hovering (the mouse pointer is not shown).

6.3. Generating a Large Sentiment Lexicon of Opposing Polarity Phrases

A possible future application of SCL-OPP is the automatic creation of a high-coverage sentiment lexicon of opposing polarity phrases. One can use the manually created lexicon entries as training data and learn a regression model over a set of features that captures the co-occurrence information (e.g., word and phrase embeddings). Such a high-coverage lexicon will be useful in downstream applications such as sentence-level sentiment classification and stance detection (Mohammad et al., 2016b).

6.4. Determining How Sentiment is Composed in the Human Brain

One of our long-term goals is to study how humans process sentiment in words and phrases. It is known that the activity in the human brain differs when the person is shown a positive or a negative word. We can gather brain activity of participants when they are presented with opposing polarity phrases, as well as when they are presented with the constituent words alone. One of the hypotheses that can be tested is whether an opposing polarity phrase triggers a valenced response pertaining to each of its constituents or a valenced response pertaining only to the meaning of the whole phrase (or both).

7. Summary

We have created a real-valued sentiment lexicon of opposing polarity phrases (such as *happy accident*) and their constituent words (such as *happy* and *accident*) through manual annotation. We used an annotation technique known as Best–Worst Scaling, which has been shown to provide unbiased, reliable, and highly discriminating results. The resulting lexicon, the Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP), includes a variety of phrases pertaining to many different part-of-speech sequences. Lexicons such as SCL-OPP, which include entries for phrases as well as their constituents, are useful in understanding how meaning (especially sentiment) is composed.

Since opposing polarity phrases are particularly challenging for automatic sentiment analysis systems, we used entries from SCL-OPP in an official test set of SemEval-2016 Task 7 (a shared task on automatically determining sentiment intensity of phrases). Additionally, we envision the following ways in which the lexicon can be used: (1) to automatically create a large coverage sentiment lexicon of multi-word phrases and apply it in downstream applications such as sentence-level sentiment classification, and (2) to investigate how the human brain processes sentiment composition. The lexicon is made freely available to the research community.

8. Bibliographic References

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

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

David, H. A. (1963). *The method of paired comparisons*. Hafner Publishing Company, New York.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS One*, 6(12):e26752.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., et al. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Esuli, A. and Sebastiani, F. (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.

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

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Graham, Y., Mathur, N., and Baldwin, T. (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.

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

Hu, M. and Liu, B. (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.

Kiritchenko, S. and Mohammad, S. M. (2016a). Capturing reliable fine-grained sentiment associations by crowdsourcing. 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.

Kiritchenko, S. and Mohammad, S. M. (2016b). The effect of negators, modals, and degree adverbs on sentiment composition. *Submitted*.

Kiritchenko, S. and Mohammad, S. M. (2016c). Sentiment composition of words with opposing polarities. 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.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Kiritchenko, S., Mohammad, S. M., and Salameh, M. (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.

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

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

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

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

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, Atlanta, Georgia.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016a). A dataset for detecting stance in tweets. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2016b). Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, Submitted.

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

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., and associates. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273.

Turney, P. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4).

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Wilson, T., Wiebe, J., and Hoffmann, P. (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.