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

Svetlana Kiritchenko and Saif M. Mohammad

National Research Council Canada

National Research Council Canada | Conseil national de recherches Canada

Word–Sentiment Associations

- Adjectives
  - reliable and stunning are typically associated with positive sentiment
  - rude and broken are typically associated with negative sentiment

- Nouns and verbs
  - holiday and smiling are typically associated with positive sentiment
  - death and crying are typically associated with negative sentiment

Note: to be associated with sentiment, the word does not need to express sentiment. Words associated with positive (negative) sentiment tend to occur in positive (negative) sentences.

Sentiment Lexicons

- Sentiment lexicon: a list of terms (usually single words) with
  - binary association to positive or negative sentiment
  - numerical score indicating the degree of association

happy	0.9
awful	-0.9
award	0.6

Sentiment Composition

Sentiment composition: determining sentiment of a phrase (or a sentence) from its constituents

Sentiment composition lexicon (SCL): a list of phrases and their constituent words with association to positive (negative) sentiment

would not be happy	-0.6
happy	0.9

These lexicons are especially useful for studying sentiment composition.

Our goal: Manually capture fine-grained (real-valued) sentiment associations for single words and multi-word phrases.

Uses of Sentiment Lexicons

- Sentence-, tweet-, message-level sentiment classification
- Tracking the distribution of sentiment words in text
  - tracking sentiment towards products, services, people, issues, etc.
  - literary analysis
  - detecting personality traits
  - detecting bullying, depression, online trolls, etc.
- Linguistic studies
  - how words are used to convey affect
  - sentiment composition
- Information visualization
  - digital humanities

Manually Created Sentiment Lexicons

- Features:
  - more accurate than automatically generated lexicons
  - less coverage than automatic lexicons
- Uses (that cannot be fulfilled by automatic lexicons):
  - to create automatic lexicons
  - to directly evaluate automatic lexicons
  - linguistic analysis
    - help understand how sentiment is conveyed by words and phrases
    - how sentiment is perceived by native speakers

Manually Created Sentiment Lexicons

- **Features**
  - more accurate than automatically generated lexicons
  - less coverage than automatic lexicons
- **Uses (that cannot be fulfilled by automatic lexicons):**
  - to create automatic lexicons
  - to directly evaluate automatic lexicons
  - linguistic analysis
    - help understand how sentiment is conveyed by words and phrases
    - how sentiment is perceived by native speakers

Existing Manually Created Lexicons

- most include only single words (lemmas)
- most have only coarse levels of sentiment (positive vs. negative)

Obtaining real-valued sentiment annotations is challenging:

- higher cognitive load than simply marking positive, negative, neutral
- hard to be consistent across multiple annotations
- difficult to maintain consistency across annotators
  - 0.8 for one annotator may be 0.7 for another

Comparative Annotations

Paired Comparisons (Thurstone, 1927; David, 1963)**:**

If X is the property of interest (positive, useful, etc.),

give two terms and ask which is more X

- less cognitive load
- helps with consistency issues
- requires a large number of annotations
  - order N^2, where N is number of terms to be annotated

Comparative Annotations

Paired Comparisons (Thurstone, 1927; David, 1963):

If X is the property of interest (positive, useful, etc.),

give two terms and ask which is more X

Best–Worst Scaling (Louviere & Woodworth, 1990):
(a.k.a. Maximum Difference Scaling or MaxDiff)

Give k terms and ask which is most X, and which is least X
(*k is usually 4 or 5*)

- preserves the comparative nature
- keeps the number of annotations down to about 2N
- leads to more reliable annotations
  - less biased and more discriminating (Cohen, 2003)

Outline

I. Capturing fine-grained sentiment associations
   - Best–Worst Scaling annotation method
   - new fine-grained sentiment composition lexicons
   - robustness of the annotations

II. Using the created lexicons
   - to gain new understandings of human perception of sentiment
   - to study the effect of modifiers (negators, modals, adverbs) on sentiment
   - to study sentiment composition in opposing polarity phrases

Outline

I. Capturing fine-grained sentiment associations
   - Best–Worst Scaling annotation method
   - new fine-grained sentiment composition lexicons
   - robustness of the annotations

II. Using the created lexicons
   - to gain new understandings of human perception of sentiment
   - to study the effect of modifiers (negators, modals, adverbs) on sentiment
   - to study sentiment composition in opposing polarity phrases

Best–Worst Scaling

- The annotator is presented with four terms (a 4-tuple) and asked:
  - which term is the most positive
  - which term is the most negative

- By answering just these two questions, five out of the six inequalities are known
  - For example, given the terms A, B, C, and D:
    - if A is most positive and D is most negative,

      then we know:

$$A > B, A > C, A > D, B > D, C > D$$

Example Annotation Instance

Focus words:
1. worse 2. was not sufficient 3. more afraid 4. banish

Q1. Identify the word that is associated with the MOST amount of POSITIVE sentiment (or, least amount of negative sentiment) -- the most positive term.

- ○ worse
- ○ was not sufficient
- ○ more afraid
- ○ banish

Q2. Identify the word that is associated with the MOST amount of NEGATIVE sentiment (or, least amount of positive sentiment) -- the most negative term.

- ○ worse
- ○ was not sufficient
- ○ more afraid
- ○ banish

Best–Worst 4-tuples

We generate 4-tuples such that:

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

This is to maximize the chance that each term is seen in a sufficient number, and a diverse set of 4-tuples.

Converting Responses to Real-Valued Scores

- Responses converted into real-valued scores for all the terms:
  - a simple counting procedure (Orme, 2009):

$$score(t) = \frac{\#most\ positive(t) - \#most\ negative(t)}{\#annotations(t)}$$

The scores range from:

-1 (least association with positive sentiment)

to 1 (most association with positive sentiment)

  - terms can then be ranked by sentiment

Annotation by Crowdsourcing

- Manual annotation through crowdsourcing
- Crowdsourcing platform: CrowdFlower
- Each question was answered by at least eight respondents
- Quality control through a small set of gold answers

Outline

I. Capturing fine-grained sentiment associations
   - Best–Worst Scaling annotation method
   - new fine-grained sentiment composition lexicons
   - robustness of the annotations

II. Using the created lexicons
   - to gain new understandings of human perception of sentiment
   - to study the effect of modifiers (negators, modals, adverbs) on sentiment
   - to study sentiment composition in opposing polarity phrases

Datasets and Domains

- Languages:
  - English
  - Arabic
- Domains:
  - general
  - Twitter
- Types of composition:
  - common modifiers (negators, modals, degree adverbs)
  - words with opposing polarities

SemEval-2015 English Twitter Lexicon

- Includes 1,515 terms from tweets:
  - regular English words: peace, jumpy
  - tweet-specific terms
    - hashtags and conjoined words: #inspiring, #needsleep
    - misspellings: appriciate
    - creative spellings: goooood, cant w8
    - abbreviations: smfh, lol
    - emoticons: :'(, <33
  - negated terms: not nice, nothing better, can't wait

SemEval-2016 Arabic Twitter Lexicon

- Includes 1,367 terms from Arabic tweets:
  - Modern Standard Arabic (MSA) and Levantine dialect
  - regular Arabic words
  - tweet-specific terms
    - hashtags and conjoined words
    - creative spellings
  - negated terms

SemEval-2016 General English Sentiment Modifiers Lexicon

a.k.a. Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA)

- Includes 3,207 general English terms:
  - 1,621 single words (Osgood's positive and negative lists)
  - 1,586 multi-word phrases 'modifier w', where w is an Osgood word and modifier is one of the following:
    - a negator: did not harm, will not be interested
    - a modal verb: should be better
    - a degree adverb: certainly agree, much trouble
    - a combination of the above: would be very easy

SemEval-2016 English Twitter Mixed Polarity Lexicon

a.k.a. Sentiment Composition Lexicon for Opposing Polarity Phrases

- Opposing Polarity Phrase (OPP): includes at least one positive word and at least one negative word

- Lexicon includes 1,661 English terms:
  - 851 OPP bigrams and trigrams: happy accident, guilty pleasures, best winter break
  - 810 unigrams that are part of the selected ngrams: happy, accident, winter

Outline

I. Capturing fine-grained sentiment associations
- Best–Worst Scaling annotation method
- new fine-grained sentiment composition lexicons
- robustness of the annotations

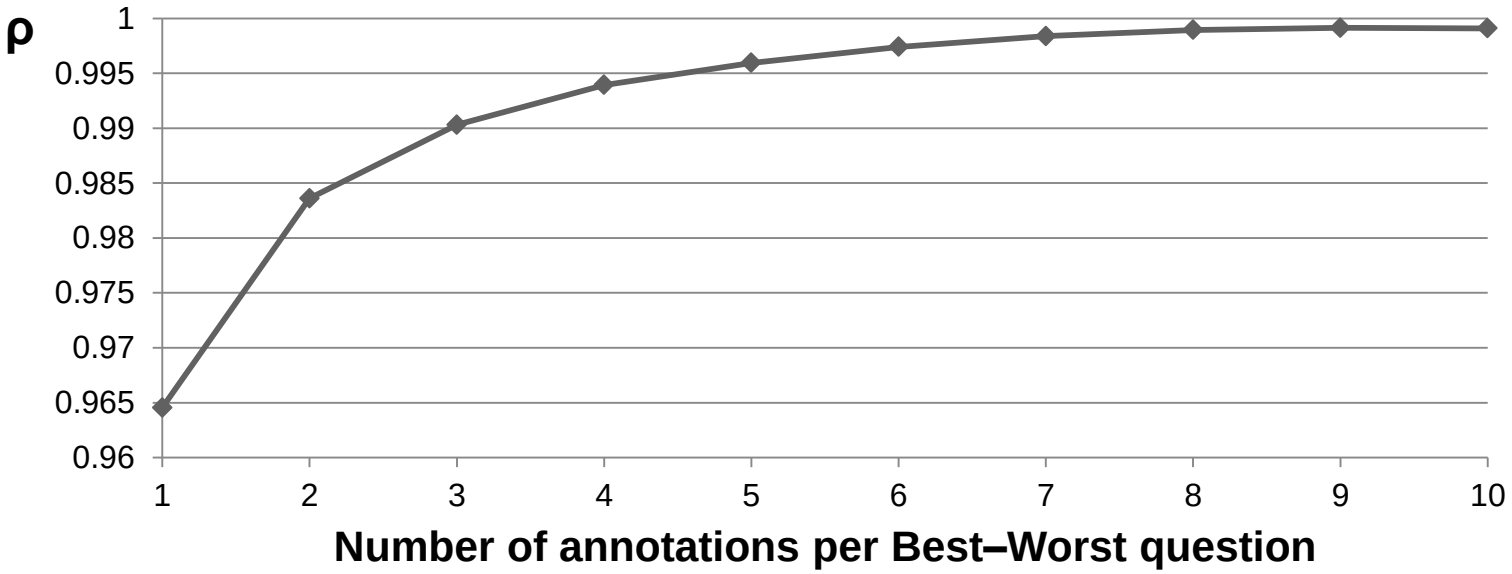
II. Using the created lexicons
- to gain new understandings of human perception of sentiment
- to study the effect of modifiers (negators, modals, adverbs) on sentiment
- to study sentiment composition in opposing polarity phrases

Robustness of the Annotations

- Divided the Best–Worst responses for each question into two halves
- Generated scores and rankings based on each set individually
- The two sets produced very similar results:
  - Spearman Rank Correlation coefficient between the two rankings was 0.98 for all four lexicons
  - Pearson Correlation coefficient between the two sets of scores was 0.98 for all four lexicons

Least Number of Annotations

- For each question (4-tuple), randomly choose n annotations (n=1..10)
- Calculate sentiment scores based on the selected subset of annotations
- Spearman rank correlation (ρ) between scores obtained with a full set and a subset of n annotations per question:



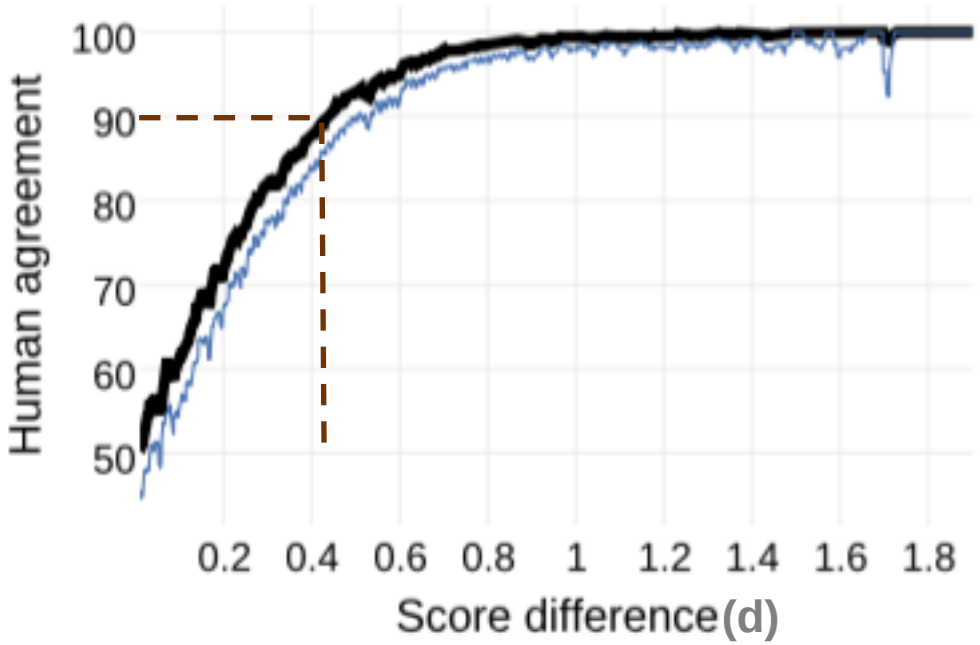
Outline

I. Capturing fine-grained sentiment associations
- Best–Worst Scaling annotation method
- new fine-grained sentiment composition lexicons
- robustness of the annotations

II. Using the created lexicons
- to gain new understandings of human perception of sentiment
- to study the effect of modifiers (negators, modals, adverbs) on sentiment
- to study sentiment composition in opposing polarity phrases

Human Agreement vs. Sentiment Difference

- For word pair w_1 and w_2 such that score(w_1) > score(w_2), we calculate human agreement for score(w_1) > score(w_2)
- We plot average human agreement as a function of d = score(w_1) – score(w_2)



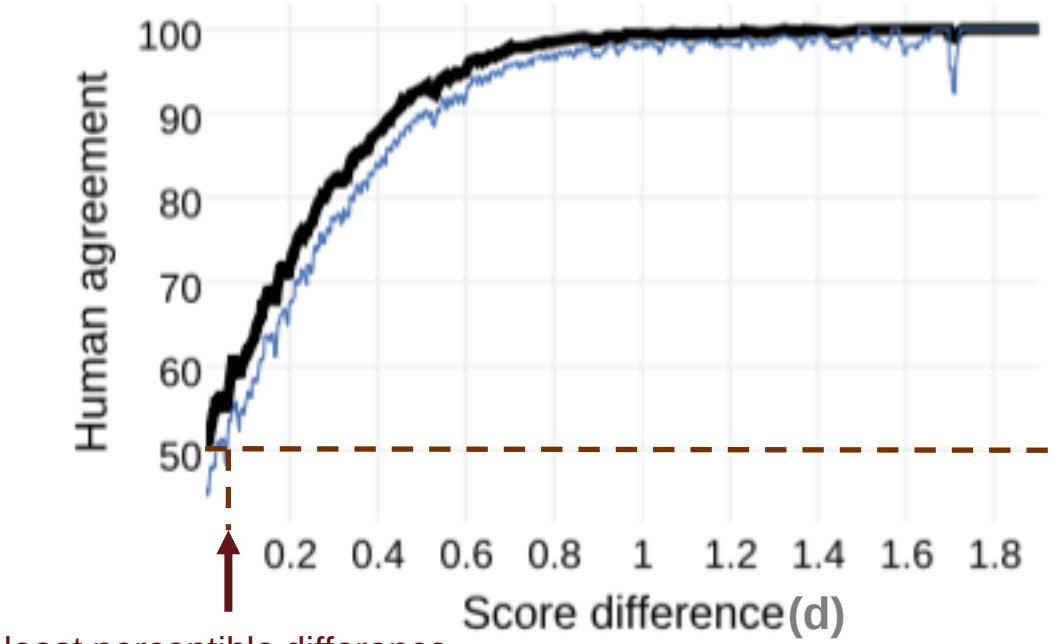
Least Perceptible Difference

- Least perceptible difference aka just-noticeable difference
  - a concept from psychophysics
  - 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)

- With our fine-grained sentiment scores, we can measure the least perceptible difference in sentiment
  - useful in studying sentiment composition (e.g., to determine whether a modifier significantly impacts the sentiment of the word it modifies)

Measuring the Least Perceptible Difference

- Least perceptible difference in sentiment scores is a point *d* at which we can say with high confidence that the two terms do not have the same sentiment associations



Least Perceptible Differences in lexicons:

General English	: 0.069
English Twitter	: 0.080
Arabic Twitter	: 0.087

(~4% in range -1..1)

Outline

I. Capturing fine-grained sentiment associations
- Best–Worst Scaling annotation method
- new fine-grained sentiment composition lexicons
- robustness of the annotations

II. Using the created lexicons
- to gain new understandings of human perception of sentiment
- to study the effect of modifiers (negators, modals, adverbs) on sentiment
- to study sentiment composition in opposing polarity phrases

Sentiment Composition Lexicon for Negators, Modals, and Adverbs (SCL-NMA)

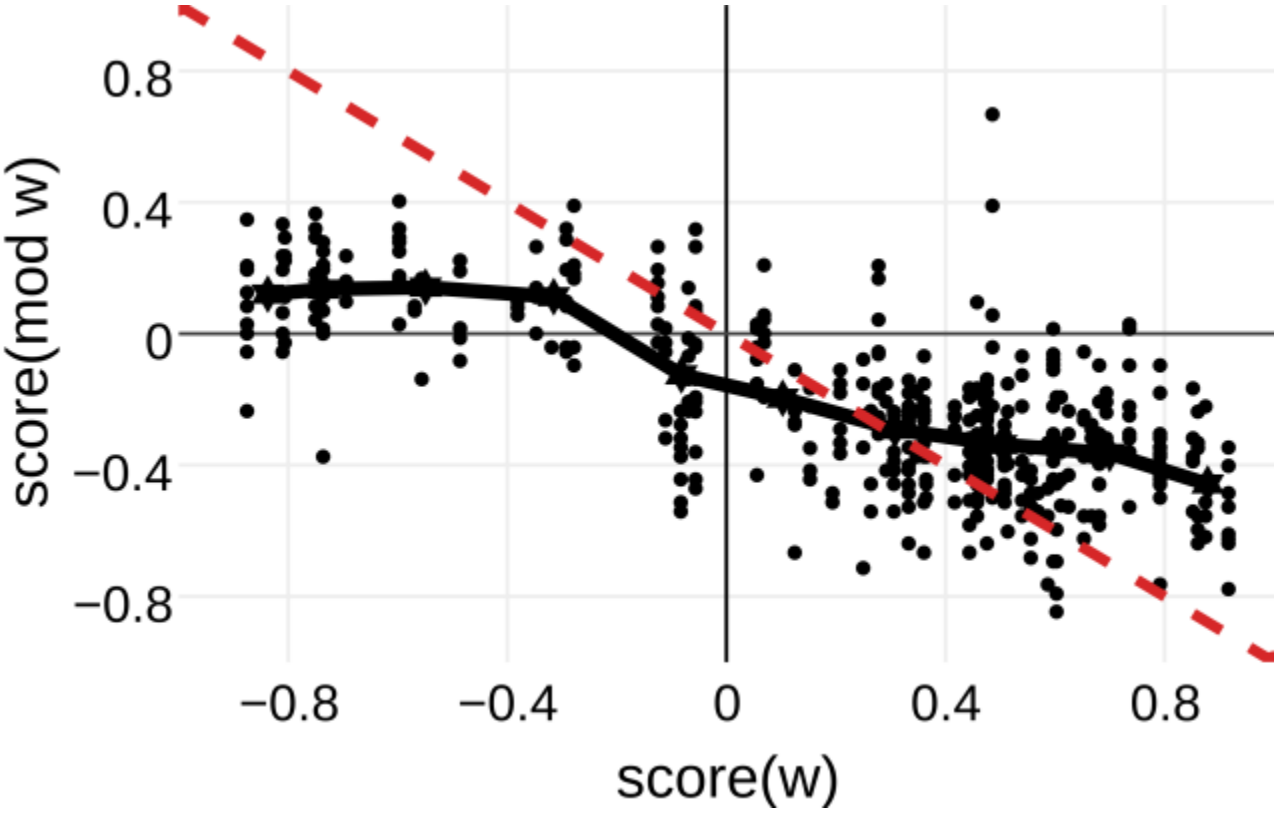
- SCL-NMA provides fine-grained sentiment associations for
  - phrases involving
    - negators (e.g., did not harm)
    - modal verbs (e.g., should be better)
    - degree adverbs (e.g., certainly agree)
    - combinations of the above (e.g., would be very easy)
  - their constituent content words (e.g., harm, better, agree, easy)

- Use SCL-NMA to help understand how modifiers (negators, modal verbs, degree adverbs) affect sentiment in phrases

Overall Impact of Sentiment Modifiers

	On positive words		On negative words	
Modifier group	Avg. diff	# of pairs	Avg. diff.	# of pairs
negators	-0.93	265	0.79	71
modals	-0.32	258	0.24	72
degree adverbs	0.20	435	0.17	163

Impact of Negation on Sentiment

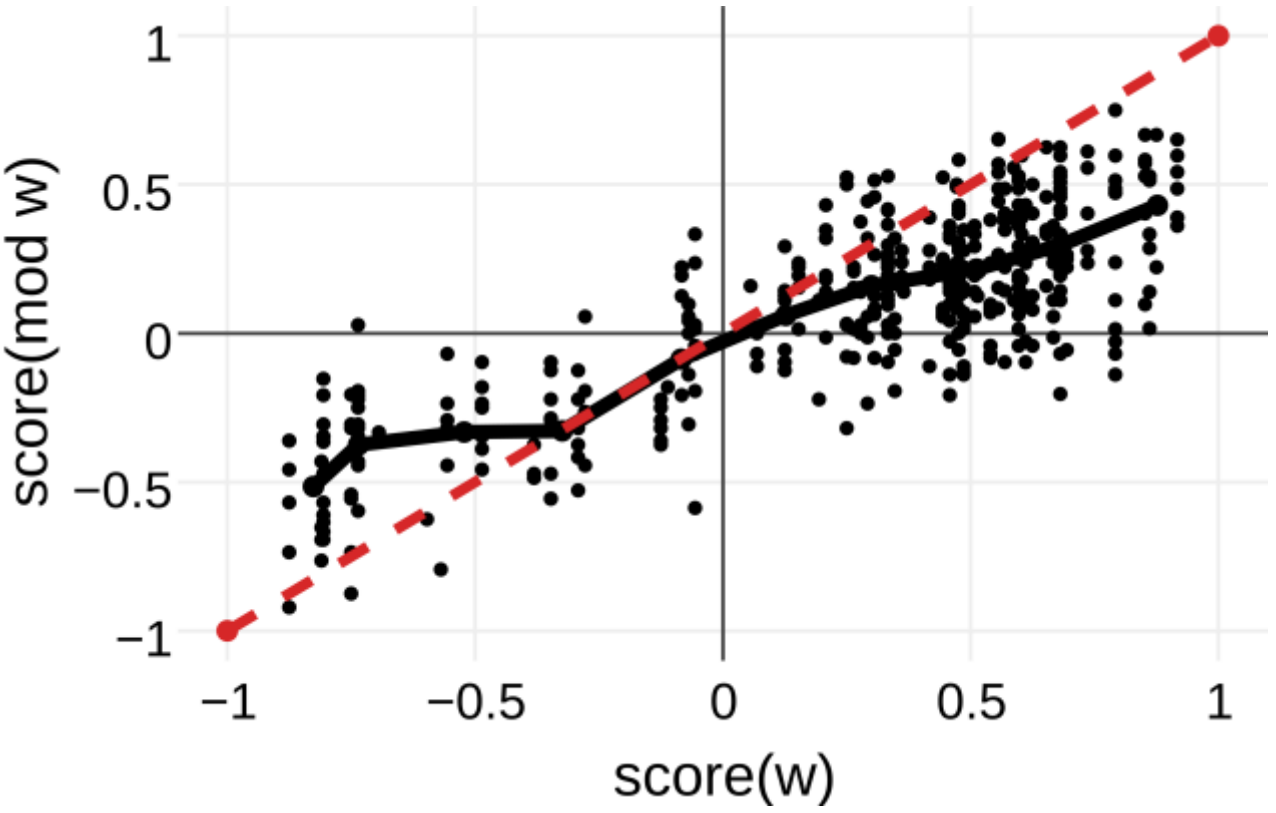


The black line shows an average effect of the negators group.
The red line shows the reversing hypothesis: score(mod w) = –score(w).

Impact of Negation on Sentiment

- Most negators
  - decrease sentiment of positive words by 0.8-1.0 points
  - increase sentiment of negative words by 0.7-0.9 points
- The greatest shift is caused by will not be and will not
- The weakest effect is by may not, nothing, and never
- Verb tense seems not to affect the behavior of negators significantly
- Modals in combination with negators slightly influence the behavior of the modifier:
  - stronger negators: will not, will not be, and cannot
  - weaker negators: could not, would not, and may not

Impact of Modal Verbs on Sentiment



The black line shows an average effect of the modals group.
The red line shows the function: score(mod w) = score(w).

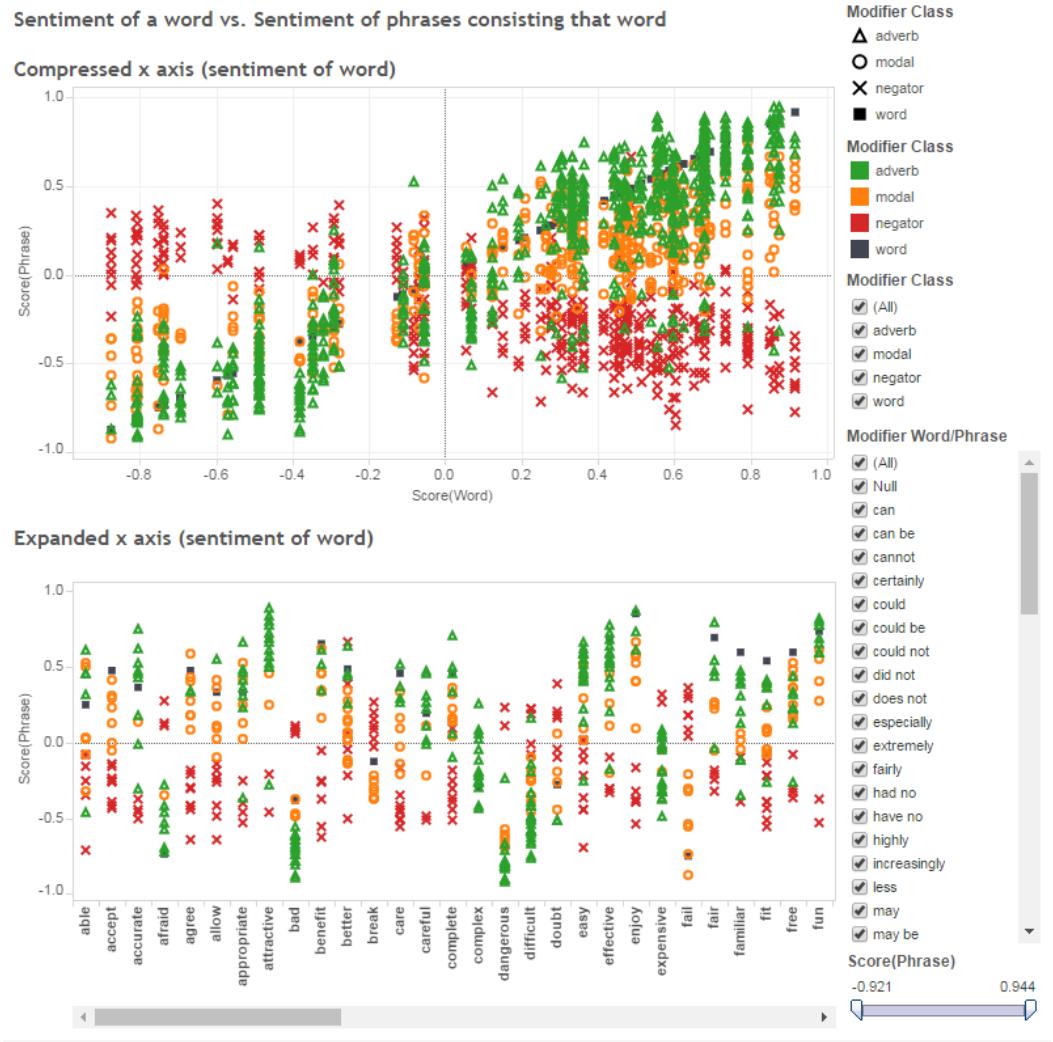
Impact of Modal Verbs on Sentiment

- Most modal verbs
  - decrease sentiment of positive words by 0.2-0.4 points
  - increase sentiment of negative words by 0.2-0.3 points
- The greatest shift (about 0.4 points) is observed for words with high absolute sentiment values
- The most influential modal modifier is would have been
- Consistent and relatively strong modifiers are formed by modals could and might
- Smallest effect on sentiment is caused by can, can be, would, and would be

Impact of Degree Adverbs on Sentiment

- Many degree adverbs have a small and rather inconsistent effect on sentiment
- The only degree adverb that affects sentiment to a large extent (0.835 points) is less
  - acts as negator
- Modifiers that consistently reduce the intensity of positive words are was too, too, probably, fairly, and relatively
- One modifier, highly, consistently and significantly increases the sentiment of positive words
- The sentiment of negative words is noticeably lowered by modifiers extremely and very very

Interactive Visualization for SCL-NMA



http://www.saifmohammad.com/WebPages/SCL.html#NMA

Outline

I. Capturing fine-grained sentiment associations
   - Best–Worst Scaling annotation method
   - new fine-grained sentiment composition lexicons
   - robustness of the annotations

II. Using the created lexicons
   - to gain new understandings of human perception of sentiment
   - to study the effect of modifiers (negators, modals, adverbs) on sentiment
   - to study sentiment composition in opposing polarity phrases

Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP)

- Opposing Polarity Phrase (OPP) consists of:
  - at least one positive word AND
  - at least one negative word

- Sentiment Composition Lexicon for Opposing Polarity Phrases provides real-valued sentiment associations for:
  - OPPs (311 bigrams and 265 trigrams)
  - their constituent single words (602 words)

Goals

- Analyze the linguistic patterns in OPPs:
  - Are there common patterns?
  - Are some POS more influential in determining the sentiment of a phrase than others?
- Apply unsupervised and supervised techniques of sentiment composition to determine their efficacy on OPPs:
  - How accurate are simple (intuitive) rules?
  - Can accurate models of sentiment composition for OPPs be learned?

Linguistic Patterns in OPPs

SCP	Occ.	# phrases
▽adj. + △adj. → △phrase	0.76	17
▽adj. + △noun → ▽phrase	0.59	68
△adj. + ▽noun → ▽phrase	0.53	73
△adverb + ▽adj. → ▽phrase	0.89	18
△adverb + ▽verb → ▽phrase	0.91	11
▽noun + △noun → △phrase	0.60	10
△noun + ▽noun → ▽phrase	0.52	25
▽verb + det. + △noun → ▽phrase	0.65	17
▽verb + △noun → ▽phrase	0.82	17

Table 2: Sentiment composition patterns (SCPs) in SCL-OPP. △denotes a positive word or phrase, ▽denotes a negative word or phrase. 'Occ.' stands for occurrence rate of an SCP.

Unsupervised Baselines

- Majority label
- First/Last unigram
- Most polar unigram
- POS rules involving adjectives and verbs

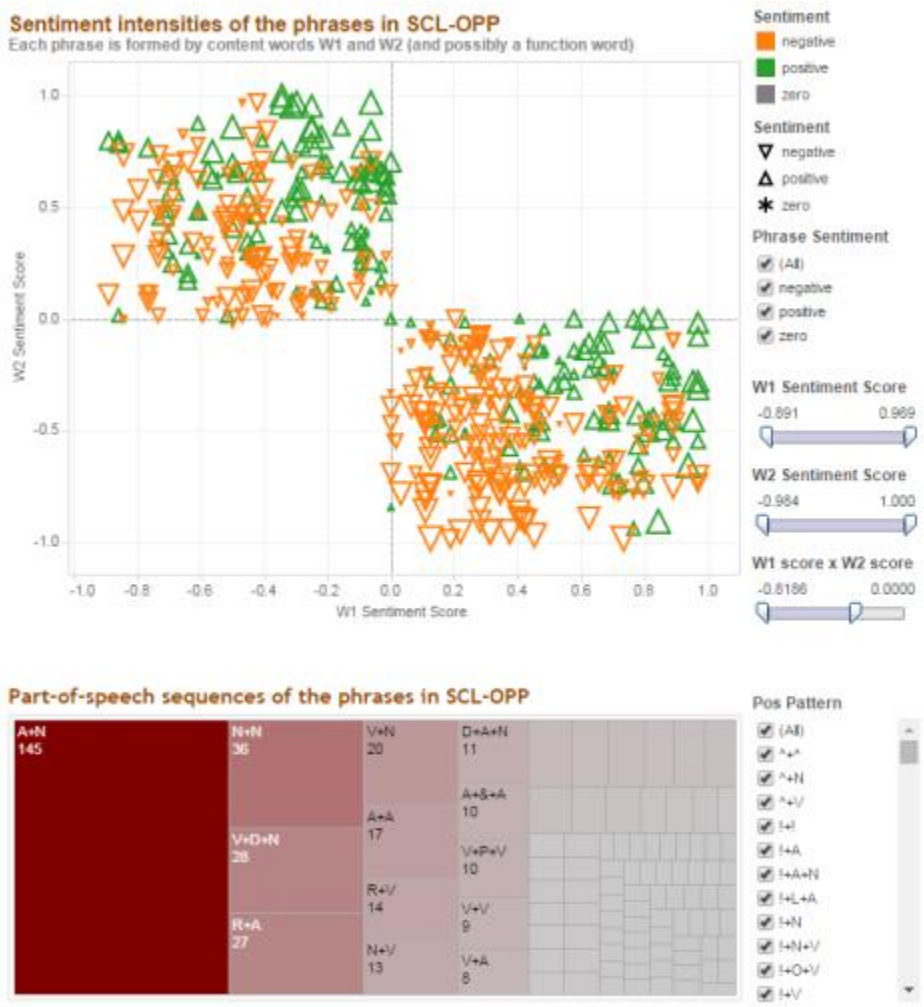
Supervised Learning

- Sentiment composition tasks:
  - binary classification (the phrase is positive or negative)
  - regression (real-valued sentiment score of the phrase)
- ML method:
  - Support Vector Machines (SVM) with RBF kernel
  - 10-fold cross-validation
- Features:
  - unigrams
  - POS tags
  - sentiment labels (binary)
  - sentiment scores (real-valued)
  - word embeddings

Results

- Sentiment of the first or last unigram is not predictive of the phrase's sentiment.
- Adjectives and verbs do not always dominate the sentiment in a phrase.
- Real-valued sentiment scores of unigrams are substantially more beneficial than binary labels.
- Sentiment of a phrase depends on its constituents and not only on their sentiment.
- Best results (accuracy over 80%) are achieved with all the features.

Interactive Visualization for SCL-OPP



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

Conclusions

- Created four real-valued sentiment composition lexicons through manual annotation and Best–Worst Scaling:
  - English Twitter
  - Arabic Twitter
  - English sentiment modifiers
  - English opposing polarity phrases
- Demonstrated the robustness of the Best–Worst Scaling annotation method
- Used the created lexicons to intrinsically evaluate automatic sentiment lexicons (SemEval-2015, SemEval-2016)

Conclusions (cont.)

- Used the English Sentiment Modifiers lexicon to analyze the impact of negators, modals, and adverbs on sentiment:
  - cannot be easily modeled with simple heuristics
  - the type of the modifier as well as the modifier word and the content word themselves affect the behavior

- Used the English Opposing Polarity Phrases lexicon to study sentiment composition in OPPs:
  - OPP sentiment cannot be predicted from POS and sentiment of the constituents
  - words, POS, sentiment scores, and embeddings are all beneficial in sentiment prediction in OPPs

Lexicons Availability

All lexicons, their interactive visualizations, and the corresponding papers are available at:

http://www.saifmohammad.com/WebPages/SCL.html

Code for Best–Worst Scaling is available at:
http://www.saifmohammad.com/WebPages/BestWorst.html

All lexicons were used as test sets in:

- SemEval-2015 Task 10: http://alt.qcri.org/semeval2015/task10/
- SemEval-2016 Task 7: http://alt.qcri.org/semeval2016/task7/

SemEval-2016 Task 7

Determining Sentiment Intensity of English and Arabic Phrases

Task: Determining Sentiment Intensity of English and Arabic Phrases

Task Description:

- Input: a list of terms
  - single words
  - multiword phrases
- Output: score indicative of the term's strength of association with positive sentiment
  - a more positive term should have a higher score than a less positive term.

Motivation:

- intrinsic evaluation of automatically created sentiment lexicons for:
  - single words
  - phrases (sentiment composition)

Task: Example

Input:

certainly agree
did not harm
favor
much trouble
severe
should be better
was so difficult
would be very easy

Output:

favor	0.83
would be very easy	0.72
certainly agree	0.67
did not harm	0.60
should be better	0.54
was so difficult	0.24
much trouble	0.17
severe	0.08

Evaluation

Data:

- SemEval-2015 Task 10 Subtask E
  - English Twitter
- SemEval-2016 Task 7
  - English Sentiment Modifiers
  - English Opposing Polarity Phrases
  - Arabic Twitter

Data distribution: for each subtask,

- no training data;
- development set: 200 terms with scores;
- unseen test set with no scores.

Evaluation measure: Kendall's rank correlation

Participated Systems

- Supervised vs. unsupervised:
  - In SemEval-2015, most systems were unsupervised leveraging information from sentiment lexicons, corpora, and Google search
  - In SemEval-2016, most systems trained regression models on dev. set and available sentiment lexicons and corpora
- Features:
  - information from sentiment lexicons
  - general and sentiment-specific word embeddings
  - pointwise mutual information (PMI) between terms and sentiment classes in labeled corpora
  - lists of negators, intensifiers, and diminishers

Results

- Results on the General English Sentiment Modifiers set are markedly higher than the results on the other datasets.
- Results on the Arabic Twitter test set are substantially lower than the results on the similar English Twitter data used in the 2015 competition.
- Results on single words are noticeably higher than the corresponding results on multi-word phrases:
  - especially apparent on the Arabic Twitter data.