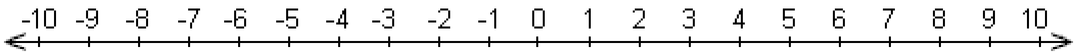


SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases

Svetlana Kiritchenko, Saif M. Mohammad

National Research Council Canada

and

Mohammad Salameh

University of Alberta

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

Sentiment Lexicons

- Sentiment lexicon: a list of terms (usually single words) with association to positive (negative) sentiment

happy	0.9
awful	-0.9
award	0.6

- Applications:
  - sentence-, tweet-, message-level sentiment classification
  - stance detection
  - literary analysis
  - detecting personality traits

Sentiment Composition

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

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

bad luck	-0.75
bad	-0.41
luck	0.58

These lexicons are especially useful for studying sentiment composition.

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

Existing Manually Created Data

- most include only single words (lemmas)
- most have only coarse levels of sentiment (positive vs. negative)
- no fine-grained sentiment lexicons for phrases, other languages

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

Annotation Method

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

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

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

- comparative in nature
- helps with consistency issues

Crowdsourcing:

- Each 4-tuple is annotated by at least eight respondents

Best–Worst Scaling:

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

Data

Three subtasks/domains:

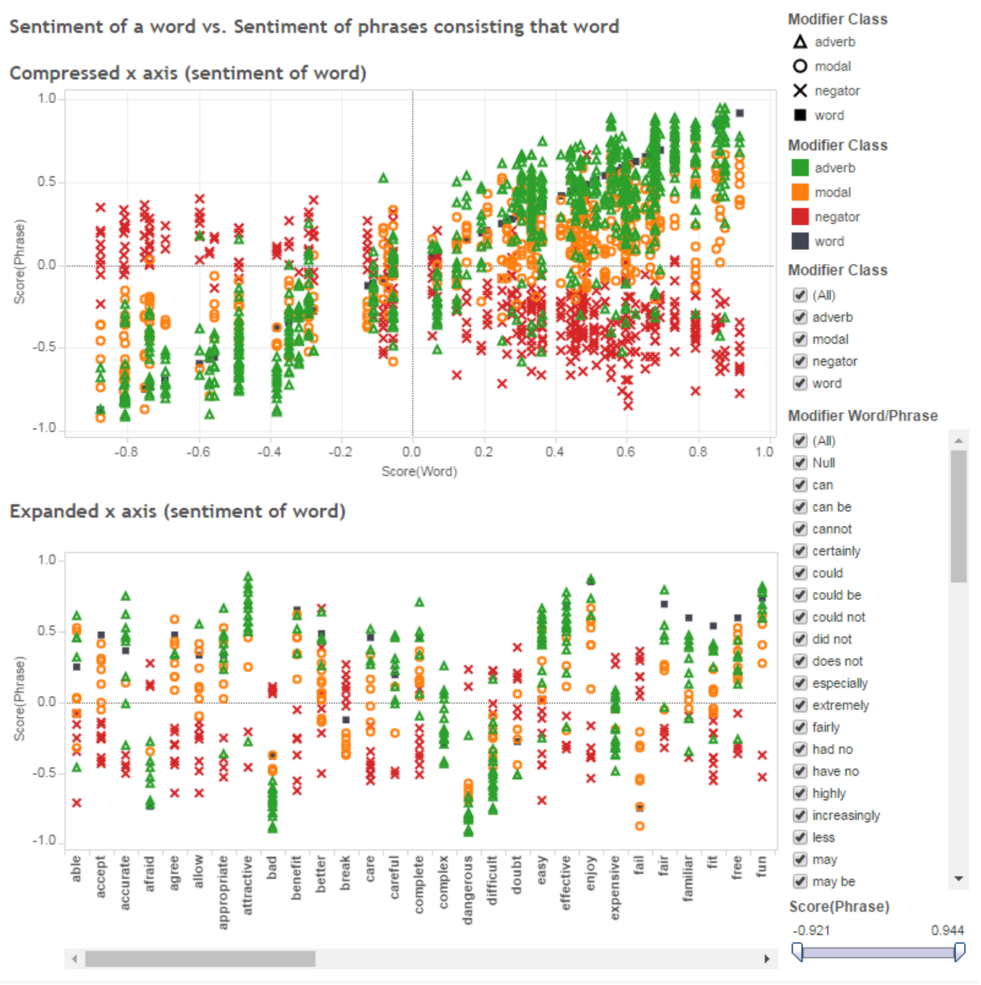
- General English Sentiment Modifiers:
  - 2,999 single words and phrases with negators, modals, and degree adverbs (e.g., *delightful*, *rather dangerous*, *may not know*)
- English Twitter Mixed Polarity:
  - 1,269 single words and phrases with at least one positive and at least one negative word (e.g., *lazy sundays*, *best winter break*, *happy accident*)
- Arabic Twitter:
  - 1,366 single words and simple negated phrases (e.g., كارث, # عشق, مش هيتحقق, صدااااع)

Quality of Annotations

- Annotations are reliable
  - re-doing the annotations with different sets of annotators produces a very similar order of terms (an average Spearman rank correlation of 0.98)

Svetlana Kiritchenko and Saif M. Mohammad. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing. *NAACL-2016*.

Interactive Visualization for General English Sentiment Modifiers (SCL-NMA)



Sentiment of a word vs. Sentiment of phrases consisting that word

Compressed x axis (sentiment of word)

Expanded x axis (sentiment of word)

Modifier Class

- △ adverb
- ○ modal
- ✕ negator
- ■ word

Modifier Class

- 🟩 adverb
- 🟧 modal
- 🟥 negator
- ■ word

Modifier Class

- ☑ (All)
- ☑ adverb
- ☑ modal
- ☑ negator
- ☑ word

Modifier Word/Phrase

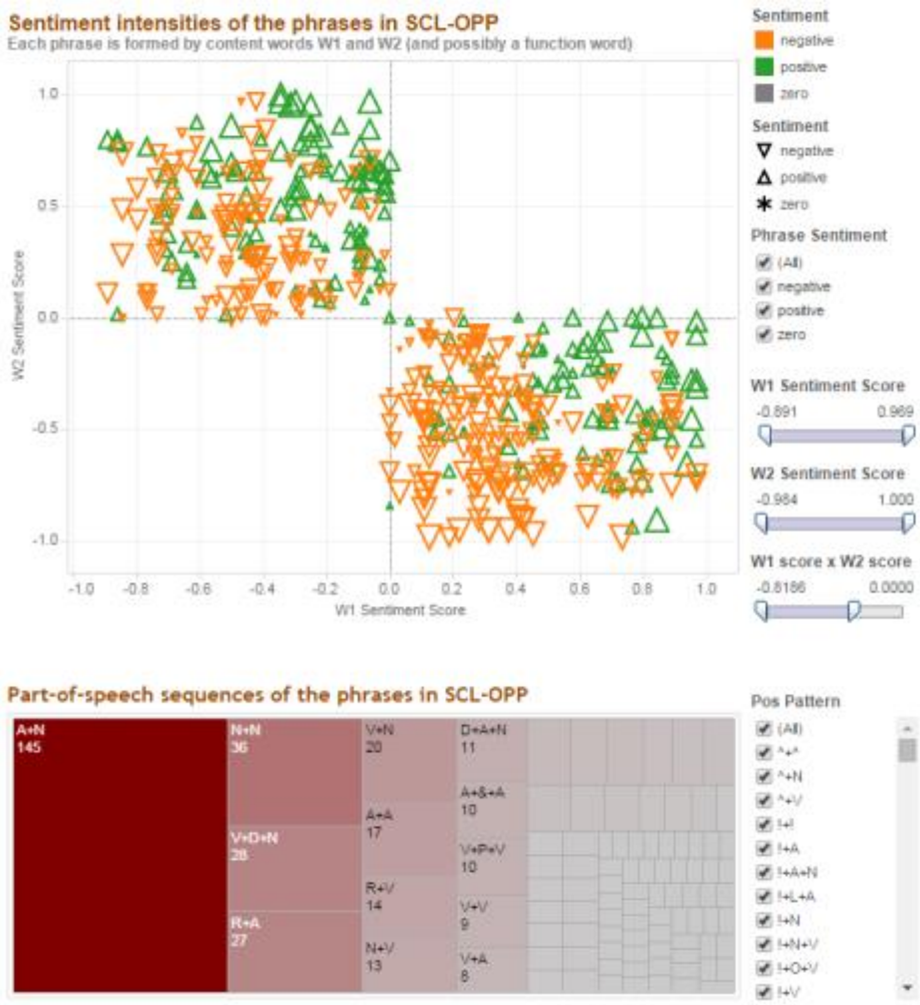
- ☑ (All)
- ☑ Null
- ☑ can
- ☑ can be
- ☑ cannot
- ☑ certainly
- ☑ could
- ☑ could be
- ☑ could not
- ☑ did not
- ☑ does not
- ☑ especially
- ☑ extremely
- ☑ fairly
- ☑ had no
- ☑ have no
- ☑ highly
- ☑ increasingly
- ☑ less
- ☑ may
- ☑ may be

Score(Phrase)

-0.921 0.944

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

Interactive Visualization for English Twitter Mixed Polarity (SCL-OPP)



Sentiment intensities of the phrases in SCL-OPP

Each phrase is formed by content words W1 and W2 (and possibly a function word)

Phrase Sentiment

☑ (All)
☑ negative
☑ positive
☑ zero

W1 Sentiment Score: -0.891 – 0.969

W2 Sentiment Score: -0.984 – 1.000

W1 score x W2 score: -0.8186 – 0.0000

Part-of-speech sequences of the phrases in SCL-OPP

Pos Pattern

☑ (All)
☑ ^+^
☑ ^+N
☑ ^+V
☑ !+!
☑ !+A
☑ !+A+N
☑ !+L+A
☑ !+N
☑ !+N+V
☑ !+O+V
☑ !+V

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

Previous Edition of the Task

SemEval-2015 Task 10 Subtask E

- Domain:
  - high-frequency terms from English tweets
- Phrase length:
  - single words (e.g., *fake*)
  - two-word negated phrases (e.g., *can't wait*)
- Term categories:
  - regular English words (e.g., *happy*)
  - hashtagged words (e.g., *#loveumom*)
  - misspelled or creatively spelled words (e.g., *happeeee*)
  - abbreviations (e.g., *lmao*)
  - slang (e.g., *smexy*)
  - emoticons (e.g., *<33*)
  - etc.

Evaluation

Data distribution: for each subtask,

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

Evaluation measures:

- Kendall’s rank correlation (primary)
- Spearman’s rank correlation (secondary)

Participants

5 teams, 3 submissions per subtask

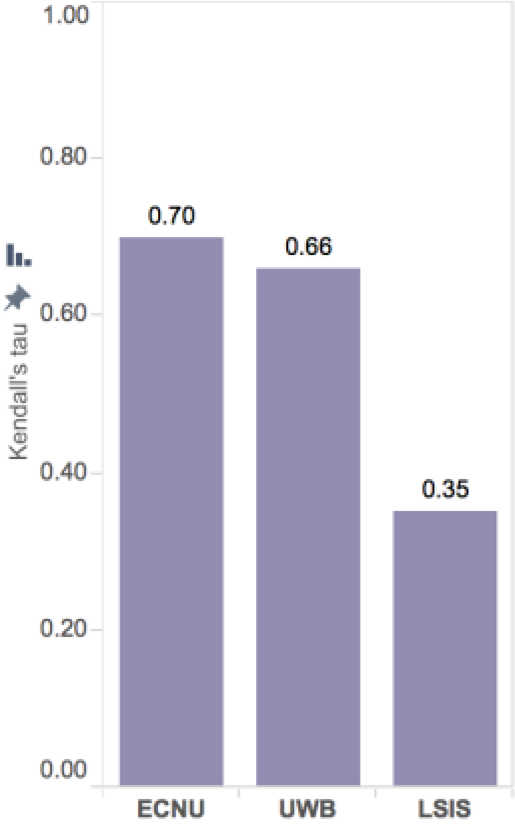
- *ECNU:* East China Normal University, China
- *iLab-Edinburgh:* Heriot-Watt University, UK
- *LSIS:* Aix-Marseille University, France
- *NileTMRG:* Nile University, Egypt
- *UWB:* University of West Bohemia, Czech Republic

Participated Systems

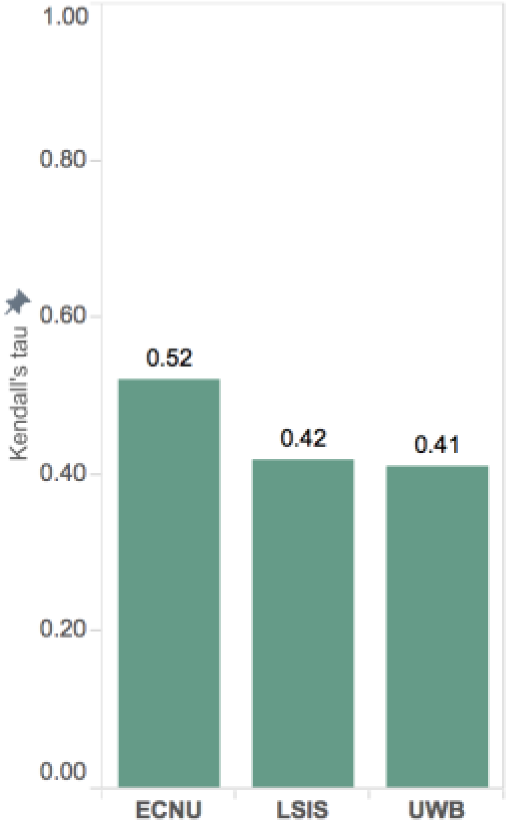
- Supervised vs. unsupervised:
  - most systems trained regression models on dev. set and available sentiment lexicons and corpora;
  - the winning team *ECNU* treated the task as rank prediction;
  - one system *LSIS* was unsupervised leveraging information from sentiment lexicons, corpora, and Google search.
- 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

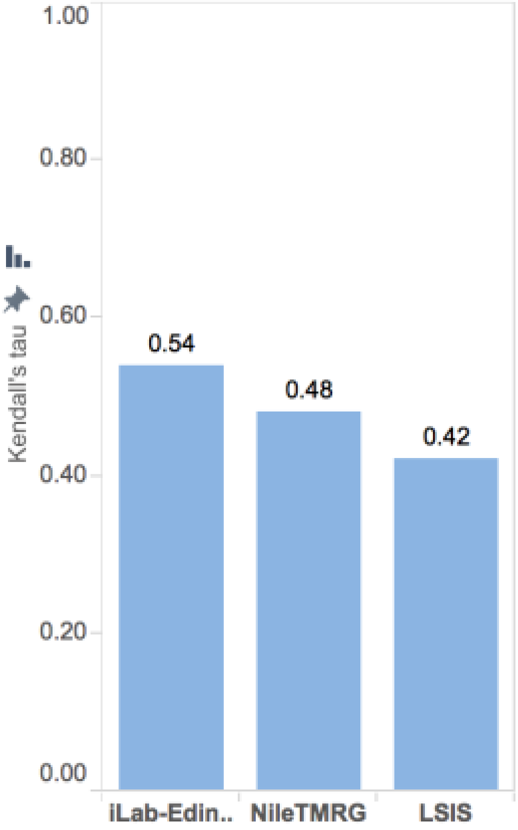
General English
Sentiment Modifiers



English Twitter
Mixed Polarity



Arabic Twitter



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.

Conclusions

- Strong correlations between predicted and gold rankings:
  - for general English domain,
  - for single words in the other two domains.
- Correlations are markedly weaker:
  - for multi-word phrases in the English Mixed Polarity set,
  - for Arabic Twitter set.

We hope that the availability of these datasets will foster further research towards automatic methods for sentiment composition in English and other languages.

Task website: http://alt.qcri.org/semeval2016/task7/