# The 2016 TAC KBP BeSt Evaluation

**Owen Rambow**
CCLS, Columbia University
New York, NY, USA

**Meenakshi Alagesan**
University at Albany
Albany, NY, USA

**Michael Arrigo**
Linguistic Data Consortium
Philadelphia, PA, USA

**Daniel Bauer**
CS, Columbia University
New York, NY, USA

**Claire Cardie**
Cornell University
Ithaca, NY, USA

**Adam Dalton**
IHMC
Ocala, FL, USA

**Hoa Dang**
NIST
Gaithersburg, MD, USA

**Mona Diab**
George Washington University
Washington, DC, USA

**Greg Dubbin**
IHMC
Ocala, FL, USA

**Jason Duncan**
The MITRE Corporation
McLean, VA, USA

**Gregorios Katsios**
University at Albany
Albany, NY, USA

**Axinia Radeva**
CCLS, Columbia University
New York, NY, USA

**Tomek Strzalkowski**
University at Albany
Albany, NY, USA

**Jennifer Tracey**
Linguistic Data Consortium
Philadelphia, PA, USA

rambow@ccls.columbia.edu

## Abstract

This paper provides a summary of the Belief and Sentiment (BeSt) evaluation that was part of the 2016 NIST TAC KBP evaluation. The evaluation is based on the accuracy of adding belief or sentiment links in a knowledge base between existing knowledge base objects. This is the first evaluation to cover both belief and sentiment.

## 1 Introduction

This document summarizes the 2016 NIST TAC KBP BeSt (Belief and Sentiment) Evaluation. It is an evaluation of sentiment and belief detection with source and target, where sources are named entities and targets are named entities or events or relations. The underlying assumption is that people (and perhaps other entities) *have* attitudes (beliefs and sentiments) towards various targets, and that texts *express* these attitudes. The goal of the evaluation, unlike other sentiment-oriented evaluations, is to interpret text in order to determine the attitudes of discourse participants and others (i.e., part of their cognitive states).

The evaluation has the following characteristics:

- It is interested in sources, attitudes, and targets: who has what mental attitude (belief or sentiment) towards what?
- The evaluation is not interested in trigger words or linguistic markers of the detected attitude, only in detection of the attitude itself.
- The evaluation includes belief and sentiment.
- The source is an entity of type Person, Geo-Political Entity (GPE), or Organization. The target can be any relation, or any event. In addition, for sentiment only, the target can also be any entity.
- There are two conditions for the evaluation:
  - We provide gold entities, relations, and events (EREs).
  - We provide predicted EREs. The predictions will come from a combination of systems.

For both conditions, participants will have access to files specifying EREs of interest; this includes in-document co-reference of entity mentions and event mentions. The tasks of finding entities, relations, and/or events, and related tasks such as co-reference, are not part of this evaluation.

- The evaluation is on English texts, but will also contain smaller Chinese and Spanish tracks.

This evaluation builds on the previous Sentiment Slot Filling evaluations that were part of the NIST TAC evaluation campaigns in 2013 and 2014 (Mitchell, 2013). Our evaluation has in common with the Sentiment Slot Filling evaluation the identification of the sentiment from sources to targets. Our evaluation extends this orientation in the following ways:

- We incorporate belief as well as sentiment.
- We allow entities, relations and events as targets, not only entities.
- We annotate on top of an ERE annotation (gold or predicted), and do not require the performers to identify possible sources and targets, or perform tasks such as co-reference resolution. This allows the performers to concentrate on the sentiment and belief tasks, and it allows us to identify the difficulty of the source-and-target belief and sentiment tasks on their own.
- Our evaluation is an automatic evaluation against a gold standard which is annotated like the training data, rather than by human assessment. This allows for more interpretable results, and allows the performers and other groups to reuse this data set for future development.

## 2 Data and Baseline

The LDC made the following data sets available for training for this evaluation:

- LDC2016E27_DEFT_English_Belief _and_Sentiment_Annotation_V2

  Total: 157k words

- LDC2016E61_DEFT_Chinese_Belief _and_Sentiment_Annotation

  Total: 133k words

- LDC2016E62_DEFT_Spanish_Belief_ and_Sentiment_Annotation

  Total: 79k words

Participating teams were free to use other data sets, such as MPQA (Wiebe et al., 2005) or FactBank (Saurí and Pustejovsky, 2009) or any other relevant resource.

The evaluation sets comprised about 200 documents in each language. While the training data was predominantly discussion forum data, the evaluation data was balanced.

ADD INFO ON DISTRIBUTION OF BELIEF AND SENTIMENT AMONG DATA SETS HERE.

The annotation in the BeSt files is on top of the annotation of entities, relations and events (ERE). In general, we distinguish between objects and object mentions. For example, the person Barack Obama is an entity, and he may be mentioned several times in a text. The mentions are textual occurrences and are represented using a pointer to a specific text file and character offsets in that file. The entity himself is not a textual occurrence and does not have an offset. Instead, Obama is an actual person. The ERE files contains lists of all mentions of annotated entities, relations, and events. While all person, GPE (geo-political entities), and organization mentions are annotated, not all relations or events are annotated, only those of certain types. Please refer to the ERE annotation manual for details. Note that by grouping mentions within an object, the ERE file provides co-reference annotation. Currently, this is only within-document.

There is a terminological wrinkle: while we have events and event mentions, and relations and relation mentions, the terminology is different for events: we have hoppers and event mentions. A hopper is a (metaphorical) container in which several related events are grouped; this is done because the notion of ‘event is more complex than that of an entity, with less of a clear single referent in the real world, and thus with less clarity about which event mentions actually refer to the same event. For the sake of this evaluation, we think of the term hopper as actually meaning event, as the ontological malaise caused by the problem of event coreference is not directly relevant to the belief and sentiment task. We will use the hopper and "event" interchangeably in this paper.

For the baseline, we determine on the training set what the majority value is; for sentiment it is always (across languages and genres) "neg", for belief always (across languages and genres) "CB". We then create an attitude for each possible target mention found in the text: for beliefs, towards each relation and each event; for sentiment, towards each entity, relation, and event. (We take gold or predicted ERE files, as the case may be, as the source of the EREs.) We use the triggers to identify target mentions; if a target has no trigger (as is the case with many relations), we identify the target mention through the mentions of its argument(s). For the source, we always assume it is the author, so we determine the author of the target mention, and then choose the appropriate mention of the author as the source mention. Some newswire files have no author mention, so we fill in None for the source (which is what the gold expects). This extended baseline is the official baseline for the evaluation.

## 3 Conceptual Description of Task

This section provides a conceptual description of the task. The actual implementation of the task in terms of input and output files (including file formats) is detailed in the task description available prior to the evaluation.

The following questions illustrate what the evaluation is getting at. We use the term private state to refer to either belief or sentiment.

- Does JohnFromTulsa like Obama?
- Who has (or is claimed to have) negative sentiment towards Obama?
- Who is self-reporting a belief about Obama, and what is it?
- What private states does BigGuyAtlanta express (or do others report he expresses) about the the annexation of Crimea?
- What private states of others is BigGuyAtlanta reporting?
- What is Hillary Clintons sentiment towards the Benghazi hearings?
- Does BigGuyAtlanta have a belief about Obama?
- Does BigGuyAtlanta believe that Obama was born in Kenya?

The systems determine the sentiment and/or belief from a holder (source) towards a target, which is an entity, a relation, or an event.

The basis of the evaluation are private state tuples (PSTs), which are 4-tuples of the following form:

> (source-entity, target-object, value, provenance-list)

The 4-tuples express the belief or sentiment of the source-entity towards the target-object (which can be an entity, a relation, or an event). The value is one from the following two sets:

- A sentiment value (positive, negative).
- A belief value (CB, NCB, ROB) where: CB = committed belief, meaning that the source is convinced the target is true. Note that this does not mean it happened in the past, a source can hold a committed belief about an event in the future. NCB = non-committed belief, meaning that the source thinks it is possible or probable that the target is true, but is not certain. ROB = reported belief. Sometimes, a writer reports on a different sources belief, without letting the reader know what his or her belief state is.

The provenance-list is a list of pointers to the text passages which support the identified claim about belief or sentiment. The provenance-list contains an entry for every single piece of textual evidence that supports the specific private state claim expressed by the PST. We consider an instance of provenance to be the target mention ID, along with the file name.

All the private states expressed in a document collection can be expressed as a collection of PSTs. The same (source-entity, target-object) pair can occur several times with different values. There are two reasons for this:

A source can have several different private states with respect to the same target. For example, the writer can have positive sentiment towards the election of Clinton, and also have a non-committed belief towards it. A source can even have conflicting private states, for example both positive and negative sentiment. This happens when someone changes his or her mind, or when they react to different aspects of the target. In this evaluation, all private states should be found; there is no aggregation or temporal analysis of conflicting private states. (In future work, the PSTs can easily be extended to record temporal information.) Because the provided ERE files only record in-document coreference, it is possible that what is in fact the same source and target and the same private state get recorded multiple times (if they are expressed in multiple documents).

The task for the evaluation is as follows:

- Input: a source text file and an ere.xml file which lists entity mentions, relation mentions, and event mentions, as well as intra-document coreference among them.

- Output: a best.xml file which refers to the input ere.xml file and which lists the belief and sentiment relations from entity mentions to entity mentions, relation mentions, and event mentions. All mentions will be mentions introduced in the ere.xml file.

Note that performers participating in the evaluation do not do entity, relation, event recognition, or coreference resolution. These tasks will already have been performed.

For the numerical evaluation results, the parameters of the evaluation are as follows:

1. There are two attitudes: belief and sentiment.

2. There are three languages: Chinese, English, and Spanish.

3. There are two genres: discussion forums (DF) and newswire (NW).

4. There are two annotation conditions:

    - The entity mentions, relation mentions, and event mentions in the ere.xml file are gold annotations.

    - The entity mentions, relation mentions, and event mentions in the ere.xml file are the output of an automatic system (provided by RPI).

    Training data of both types (gold and predicted) will be available, so that performers can choose to have two systems optimized for the two annotation conditions.

This gives us 24 separate tasks. We use recall, precision, and F-measure as measures.

## 4 Evaluation Metric

The scoring is based on the PST 4-tuples (see Section 3). We perform a recall-precision analysis on the predicted 4-tuples against the gold 4-tuples. In this way, we are evaluation how well we would our systems would be able to populate a knowledge base. However, as the 4-tuples contain lots of information, we assign partial credit. When assigning partial credit, we always require that the target is correct.

Partial credit is given if the target is correct, but not the source. Partial credit is given if the type of attitude is correct (i.e., belief or sentiment), but not the value (pos or neg for sentiment, CB, NCB, ROB for belief). No partial credit is given if belief is predicted when there is a sentiment and vice versa. Partial credit is given for the provenance list (i.e., pointers to documents and specific text passages that support the claimed attitude from source to target). There are two conditions. In the full-provenance condition, partial credit is given based on recall-precision analysis of the provenance list. In the single-provenance condition, full credit is given if at least one correct provenance is detected.

Here is a detailed description.

1. Given a predicted best.xml and a gold best.xml file, both are first converted to the 4-tuple notation of Section 2. This happens as follows:

    For each (source-mention, target-mention) pair in the best.xml file, the corresponding source and target objects are retrieved from the rich-ere.xml file. The value (a belief or sentiment value) is retrieved from the best.xml file. If there is no 4-tuple with the source, target, and value in the first three positions, a new 4-tuple is created. The provenance list of this new tuple is set to be a list containing the target-mention . If there already is a 4-tuple with the source, target, and value in the first three positions, the target-mention is added to the provenance list. This gives us two sets of 4-tuples that express the same content as the gold and predicted best.xml files (in light of the shared rich-ere.xml file), respectively.

2. We then perform an initial analysis on the first three fields of the predicted 4-tuples against the gold 4-tuples. We sort all predicted 4-tuples by the type of match against the gold 4-tuples. We then process all tuples of this match type before moving to tuples of the next match type. Whenever a gold tuple is part of a successful match, it is removed from the pool of possible matches for subsequent predicted tuples.

   The match types are as follows. They are processed in the order given.

   (a) If a predicted 4-tuple does not match any gold tuple on target and attitude type (belief or sentiment), it is a false positive.

   (b) If the source, target, and value of a candidate tuple match a gold 4-tuple, then the tuple counts as a true positive with a true positive matching score of 1.

   (c) If the source and target and attitude type match a gold 4-tuple, then the tuple counts as a true positive with a true positive matching score of $\frac{2}{3}$.

   (d) If the value and target of a candidate tuple match a gold 4-tuple, then the tuple counts as a true positive with a true positive matching score of $\frac{2}{3}$.

   (e) If only the target and attitude type of a candidate tuple match a gold 4-tuple, then the tuple counts as a true positive with a true positive matching score of $\frac{1}{3}$.

   (f) Any gold 4-tuple that is not matched at least partially by a predicted 4-tuple under rules (b), (c), (d), or (e) counts as a false negative.

3. If a predicted tuple counts as a true positive under (2), we check the provenance list. Recall that an instance of provenance is the target mention, so two instances of a provenance match if they are the same target mention. For the provenance list, there are two conditions:

   (a) In the full-provenance condition, a recall-precision matching of the predicted provenance list against the gold provenance list is performed. The resulting f-measure is used to scale the true positive matching score obtained in step (2). Note that this can be 0, if no correct instances of the provenance are identified.

   (b) In the single-provenance condition, we check if any predicted provenance is correct; if yes, the tuple remains a true positive and the matching score from (2) is retained; if no, the tuple is counted as a false positive.

   Experiments on the submitted systems showed that the one-is-enough approach resulted in scores that are fairly consistently around 2 percentage points better, and thus the two modes of evaluation do not provide very different information. (Presumably this is the case because for many gold instances of belief and sentiment, there is a single provenance, in which case the two approaches give the same result.) We report the stricter full-provenance condition in this paper for all evaluations.

The resulting sums of true positives matching scores, the count of false positives, and the count of false negatives are used in a standard recall-precision calculation, with the f-measure as the final result.

The following information which is found in the gold best.xml files was not used in the evaluation this year:

- Polarity for belief
- Sarcasm
- Belief towards event arguments
- Instances of the NA value for beliefs

Validation and scoring scripts were distributed in early June for use in development. Note that the micro-average reported by the script is relevant and will be used in the evaluation. For the macro-average, the evaluation script calculates the recall, precision, and f-measure for each file, averages the recall and precision across files, and then calculates the f-measure. The macro-averaged results are affected by some files being outliers with no data points to be found, which results in a recall of 1. In contrast, for micro-average, the calculation script merges all files into one large data set and then calculates recall and precision on this merged data set. The fact that some files have no data points does not affect the overall evaluation results.

## 5 Participants

Four teams participated in the evaluation. We summarize their approaches here.

- The REDESB group focuses on belief classification rather than sentiment identification and polarity classification. Tehy assume that the source is always the author. They use a Bayesian approach, using features such as event/relation sub-type, the type of the entities, the arguments of the relations and events, and the POS of the trigger word related to the belief.

- The CUBISM team has separate sentiment and belief systems. For sentiment, they extend the previously developed affect calculus algorithm (ACA). ACA combines information about the syntactic and semantic structure of a sentence with base polarity values of words and phrases in order to estimate polarity and intensity of sentiment from the holder to- wards the target. The belief classifier operates over a graph constructed from the entity, relation, and event data provided for the task, and also uses dialogue acts. Nodes in the graph are also assigned membership to a community on the assumption that authors who interact with have the same type of beliefs on similar event and relation types. Beliefs are then created for each event and relation and labeled using a Naive Bayes Classifier.

- The Columbia_GWU team is the only one with submissions for all 24 conditions. For sentiment, they propose two approaches. In the first approach, they assume that the source is the author, and then classifies possible targets. They use random forests, and lexical features, the types of the targets, and also experiment with syntax to find relevant features within a longer sentence. The second approach treats source-and-target sentiment like relation extraction. For belief, they assume the source is the target and use a word-tagger to identify the type of belief, and combine that system with a default system.

- The Cornell-Michigan-Pittsburgh team (aka CORNPITTMICH) submitted systems for English and Chinese. The systems employ a combination of simple, hand-crafted rules and machine learning-based approaches.

  All three English systems have two stages: the first stage performs link prediction; the second, belief and sentiment classification. The systems differ in the implementation of the two stages. For all three systems, the rule-based link predictor bases its decisions on information drawn from the text span between the source and target entity/relation/event pairs. All stage 2 systems are trained on gold positive links and spurious NONE/NA links predicted by the rule-based component.

  Our Chinese systems use separate components for handling sentiment vs. belief. A hybrid approach to **sentiment classification** is used for both discussion forum and newswire. The model consists of: (a) An LSTM-based neural network for sentence-level sentiment analysis trained with about 4k sentences from Weibo with polarity annotated (positive/negative/none); (b) a rule-based model for finding the source of a sentiment in the discussion forum vs. newswire; (c) a rule-based model that outputs the final results based on the output of model (a), the source output by model (b), as well as a number of high level features such as indicators of entity/relation/event, text length, number of entities in the sentence, etc.. The main function of this model is to set different thresholds of accepting the positive/negative predictions from the neural network for different scenarios. The parameters of this model are automatically tuned on the BeSt training data. The only difference between our submissions is the metric for tuning the parameters of model (c). In particular, the metrics for submission 1,2,3 are focused on good F-score, recall, precision respectively.

The **belief classification** component for all three Chinese submissions is identical. For discussion forums, the output was obtained by a combination of a rule-based model and a linear model trained on the BeSt training data. For newswire, the output was obtained by a rule-based model. Specifically, the rule-based model is a simple model that always outputs type="cb" for each relation and event; it it uses the same model as sentiment (b) to find the source of a belief. The linear model takes the text around the relation/event mention and decides whether or not there is a belief. If the answer is no, it removes the corresponding belief output (produced by the rule-based model) from the final output.

## 6 Results

We provide the F-Measure results for Belief in Table 2 and for Sentiment in Table 4. The baseline is nontrivial and for belief it is very high, since most possible targets are in fact committed beliefs of the author, resulting in a high precision and recall. In addition, we summarize the results on Belief for the four participating teams in Table 1 and the results on Sentiment in Table 3, where the best performer is marked with $\odot$ and systems which beat the base line is marked with $\diamondsuit$ . In all tables we simply give the results for the best performinig system from a specific performer.

## 7 Discussion

This was the first evaluation to focus on source-and-target sentiment, the first to concentrate on source-and-target belief, and the first evaluation to combine belief and sentiment. However, the evaluation revealed several areas in which improvement will be possible.

- **Large Number of Conditions** We ended up with 24 rather distinct conditions. This may have been too many conditions, since every single condition would have (probably) profited from individual tuning. However, it is also important to understand what factors affect performance.

- **Predicted ERE** There are no separate gold belief and sentiment annotations on predicted ERE, and therefore we needed to map predicted ERE to gold ERE as part of our evaluation. This in effect amounts to an evaluation of ERE, which is of course not our goal. We first establish a mapping between entity mentions based on an exact match of the annotated text span and entity type. Event mentions are identified by their trigger words. We then use the set of mapped mentions to compute the mapping for entities and events by maximizing overlap of their mention sets. Relations often do not have triggers. We therefore match relation mentions based on the annotated relation type and an exact match between mentions of relation argument. Because we require exact matches of annotated spans and types, and missed mentions are not taken into account to compute entity, event, and relation mappings, our technique strongly prefers precision over recall. This results in very low scores for all submitted system in the predicted ERE condition. In addition, the predicted ERE annotations for Chinese do not contain any relation annotations to begin with, so the submitted systems were not able to produce annotations for relations.

- **Data Issues** While the amount of data available for training was significant in terms of words, the amount of sentiment differed. In particular, for Chinese, there was little sentiment in the training data and therefore Chinese systems that relied only on the provided data did not perform well.

## 8 Conclusion

We believe that the approach taken in this evaluation is an important step towards modeling the meaning

| | English | | | | Spanish | | | | Chinese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | |
| | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW |
| Columbia/GWU | ⊙ | ⊙ | ○ | ○ | ⊙ | ⊙ | ○ | ○ | ○ | ⊙ | ○ | ○ |
| cornpittmich | ○ | ○ | ⊙ | ⊙ | — | — | — | — | ⊙ | ○ | ○ | ○ |
| CUBISM | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| REDES | ○ | ○ | — | — | — | — | — | — | — | — | — | — |

**Table 1** Results on belief for the four participating teams; ○ = participated (no result beat the baseline); ⊙ = best performer (a column without best performer means that all performers achieved a result of zero)

| | English | | | | Spanish | | | | Chinese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | |
| | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW |
| Baseline | 0.783 | 0.677 | 0.097 | 0.089 | 0.782 | 0.655 | 0 | 0 | 0.841 | 0.694 | 0 | 0 |
| Columbia/GWU | 0.779 | 0.664 | 0.042 | 0.039 | 0.678 | 0.591 | 0 | 0 | 0.797 | 0.670 | 0 | 0 |
| cornpittmich | 0.764 | 0.657 | 0.055 | 0.084 | — | — | — | — | 0.841 | 0.596 | 0 | 0 |
| CUBISM | 0.633 | 0.654 | 0 | 0 | 0.532 | 0.486 | 0 | 0 | 0.679 | 0.610 | 0 | 0 |
| REDES | 0.523 | 0.603 | — | — | — | — | — | — | — | — | — | — |

**Table 2** Results on belief for the four participating teams (f-measure)

| | English | | | | Spanish | | | | Chinese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | |
| | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW |
| Columbia/GWU | ⊙ ♢ | ⊙ ♢ | ⊙ ♢ | ○ ♢ | ⊙ ♢ | ⊙ ♢ | ⊙ ♢ | ⊙ | ○ ♢ | ○ ♢ | ○ | ○ |
| cornpittmich | ○ ♢ | ○ | ○ ♢ | ○ | — | — | — | — | ⊙ ♢ | ⊙ ♢ | ⊙ | ⊙ ♢ |
| CUBISM | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ ♢ |
| REDES | ○ | ○ | — | — | — | — | — | — | — | — | — | — |

**Table 3** Results on Sentiment for the four participating teams; ○ = participated; ♢ = beat the baseline; ⊙ = best performer (a column without best performer means that all performers achieved a result of zero)

| | English | | | | Spanish | | | | Chinese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | | Gold ERE | | Predicted ERE | |
| | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW | DF | NW |
| Baseline | 0.145 | 0.072 | 0.066 | 0.040 | 0.161 | 0.091 | 0.026 | 0.026 | 0.107 | 0.021 | 0.035 | 0.011 |
| Columbia/GWU | 0.206 | 0.094 | 0.095 | 0.048 | 0.226 | 0.085 | 0.032 | 0.004 | 0.170 | 0.040 | 0.010 | 0.006 |
| cornpittmich | 0.195 | 0.007 | 0.084 | 0.001 | — | — | — | — | 0.399 | 0.096 | 0.025 | 0.028 |
| CUBISM | 0.151 | 0.029 | 0 | 0 | 0.068 | 0.024 | 0.007 | 0.002 | 0.078 | 0.028 | 0.016 | 0.029 |
| REDES | 0 | 0 | — | — | — | — | — | — | — | — | — | — |

**Table 4** Results on Sentiment for the four participating teams (f-measure)

of natural language communication: the goal of understanding communication is not simply to identify propositional content, but to understand the cognitive states (beliefs, sentiments, intentions) of the discourse participants (Austin, 1962). This will become increasingly crucial for the information extraction and knowledge base population, and this evaluation is a first step towards understanding how well we are performing at understanding discourse.

## Acknowledgments

The work on this paper by the following co-authors was supported by the DARPA DEFT Program: Rambow, Arrigo, Bauer, Diab, Duncan, Radeva, Tracey. This material is based in part on research sponsored by Air Force Research Laboratory and Defense Advance Research Projects Agency under agreement number FA8750-13-2-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government.

## References

J. L. Austin. 1962. *How to do things with words*. Oxford University Press.

Margaret Mitchell. 2013. Overview of the tac2013 knowledge base population evaluation: English sentiment slot filling. In *TAC 2013 Proceedings*. NIST. http://tac.nist.gov/publications/2013/papers.html.

Roser Saurí and James Pustejovsky. 2009. Fact-Bank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.