The Columbia-GWU System at the 2016 TAC KBP BeSt Evaluation

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

Tao Yu
DSI, Columbia University
New York, NY, USA

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

Sardar Hamidian
CS, George Wahington Uni.
Washington, DC, USA

Alexander Fabbri
CC, Columbia University
New York, NY, USA

Debanjan Ghosh
SCI, Rutgers University
New Brunswick, NJ, USA

Christopher Hidey
CS, Columbia University
New York, NY, USA

Tianrui Peng
CS, Columbia University
New York, NY, USA

Mona Diab
CS, George Wahington Uni.
Washington, DC, USA

Kathleen McKeown
CS, Columbia University
New York, NY, USA

Smaranda Muresan
CCLS, Columbia University
New York, NY, USA

rambow@ccls.columbia.edu

Abstract

We present the components of the Columbia-GWU contribution to the 2016 TAC KBP BeSt Evaluation.

1 Introduction

The 2016 TAC KBP BeSt evaluation of source-and-target belief and sentiment covers 24 different scenarios: two phenomena to detect, three languages, two genres, and two conditions. As a result, we did not build a unified system; instead, we use several different base approaches which we tailor to different situations. We list them here:

- A sentiment system based on identifying the target only, with the source implicitly the author. This is developed for English (Section 3), and adapted for Chinese (Section 5) and Spanish (Section 6).

- A sentiment system based on a relation extraction system, with the notion that sentiment from source to target is a relation from source to target. This is used only for English (Section 4).

- A belief system that combines high-precision word tagging with a high-recall default system. This is used for English (Section 7) and adapted for Chinese (Section 8).

- A belief system that is based on a weighted random choice of tags, which we use for Spanish (Section 9).

While we perform experiments on the two genres (discussion forums and newswire) to find optimized training data for each test data, we do not perform special adaptation for the two conditions of gold and predicted ERE (entities, relations, events). Instead, all results are obtained by training on gold ERE.

2 Data

We use the following data sets from the LDC.

- LDC2016E27_DEFT_English_Belief_and_Sentiment_Annotation_V2

- LDC2016E61_DEFT_Chinese_Belief_and_Sentiment_Annotation

- LDC2016E62_DEFT_Spanish_Belief_and_Sentiment_Annotation

We did not use any other sentiment-annotated resources (other than sentiment lexicons). For English and Chinese belief (Section 7 and 8), we used belief word taggers which are trained on data using a different annotation, in which only targets are identified (the source being always the author):

- LDC2014E55_DEFT_Committed_Belief_Annotation_R1_V1.1 and LDC2014E106_DEFT_Committed_Belief_Annotation_R2
- LDC2015E99_DEFT_Chinese_Committed_Belief_Annotation

For the sentiment systems, on the "Belief and Sentiment" data sets, we first created a division into train-dev-text of approximately 80%-10%-10% for each of three languages, in which we attempted to keep the percentage of occurrences of sentiment and belief about equal. We subsequently realized that the eval data would include much more newswire than the training data. For English, we therefore created a new development set which we call "superdev", in which we combined all newswire files from our previous dev and test sets. (We abandoned our own test set since the eval data would serve as test set.) In this paper, we report results on this superdev set for English, as well as for the eval set.

We did not train separately on predicted ERE.

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 a sentiment for each possible target: for beliefs, for each relation and each event; for sentiment, for each entity, relation, and event. (We take gold or predicted ERE files, as the case may be, as the source of the EREs.) For the source, we always assume it is the author, so we determine the author and choose the appropriate mention 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). In our belief systems, we use this baseline as one of our systems and as a component in another system; see Section 7, Section 8, and Section 9. We subsequently improved the baseline by extending the set of identified targets by identifying relations and events not only through the trigger, but also through the arguments. This extended baseline is the official baseline for the evaluation, and it outperforms the baseline used in our belief experiments.

All results are given using the "one-is-enough" option for provenance scoring, which requires at least one overlap between the predicted provenance list and the gold provenance list for each predicted sentiment or belief.

3 English Sentiment 1

3.1 Basic Approach

We employ widely used text classification features such as word embeddings and sentiment words count. Also, we implement some task-specific features such as the mention types of the target. The features are extracted on the target, sentence, post and file levels. As classifiers, we use Support Vector Machines (SVM) with linear kernels and Random Forest classifiers are used to train our models. We apply two different approaches, target-oriented and context-oriented, with similar features on this task.

3.2 Data Pre-Processing

Pre-processing steps include sentence and word tokenization and other common NLP tasks such as part-of-speech tagging and parsing. For the tokenization, we make use of the the NLTK package (Bird et al., 2009). Sentences are also syntactically parsed and marked with part-of-speech tags with the Stanford CoreNLP tool (Manning et al., 2014).

We then collect all possible targets from the ERE file (gold or predicted). We exclude all entity mentions which refer to the author of that particular passage (since the author is unlikely to express sentiment towards herself); these are typically signaled by texts that are simply *I* or *my* in our data set. For each of the other possible targets, we extract the sentence where it is located and save the indicators of which file, post and author it is from.

3.3 Approach

The task we address in this evaluation is source-target dependent sentiment analysis, which means we are interested not only the sentiment towards a target, but also in the person who has this sentiment. Nonetheless, we still assume for the approach discussed in this section that the source is the author.

This is because the vast majority of sentiment cases for both discussion forum and newswire data sets are from the author. We pursue two approaches.

First, we apply a target-oriented approach. The main idea of this method is to have target-specific features. One of most challenging problems of this task is that the sentences in our dataset are complex, and they are relatively long. Also, it is common to have more than 5 potential targets in each sentence. The most important step behind our approach is to get a "small sentence" which is most related to the target. In order to get this small sentence, we use the S labels from the parsing output to divide the sentence into chunks of clauses and non-clausal words. We then scan the two chunks both before and after the chunk where the target is located, and discard the chunks that contain other mentions and that have the S label. Finally, we combine the remaining clauses into a small sentence that includes the target.

We generate features based on our small sentence. Sentiment lexicons including the NRC Emotion Lexicon (Mohammad and Turney, 2010), Bing Liu's Lexicon (Hu and Liu, 2004), and MPQA Subjectivity Lexicon (Wilson et al., 2005) help count the sentiment words in the small sentence. We added the pre-trained word embedding of each word in the small sentence, weighted by POS (as determined by the tagger). The weights are chosen through experiments on the training set. We also experimented with word2vec (Mikolov et al., 2013) and different Glove (Pennington et al., 2014) word embedding resources. Glove 840B300d gives us the best performance. Other features include the types of possible target entities, relations and events such as GPE, LOC, conflict, etc., as well as the relative position of the targets in the sentence.

Second, we apply a context-oriented method. We use the same features described for the first approach. However, we no longer cut the sentence into clauses to get the small sentence. Instead, we have the features of original sentence, post and file.

We apply linear SVMs and Random Forest classifiers in this task for both approaches. Hyperparameters were tuned via cross-validation.

We experimented with many other features such as different sentiment lexicons, negation, punctuation and POS tagger counts, but they did not prove helpful for this task.

3.4 Results

The model trained with features of the small sentence gets 6% higher than of the original large sentence on F score if we only work on the sentence level. However, the context-oriented approach outperforms the target-oriented method. We therefore use the context-oriented approach without small sentences.

We then experiment with the different genres, DF and NW. For developing our system we conducted a set of experiments on "SuperDev" to determine which model to use. We train separate models on DF and on DF and NW combined. The size of the training set from the NW corpus was very small and the model trained only on newswire data did not perform well. The results from the experiments are reported in Table 1. As we can see, for DF, the best system to use is trained on DF only, while for NW (and for the combined dev set of DF+NW) training on a combination of DF and NW is best.

Results on the evaluation set are shown in Table 3. For the English Sentiment result on the evaluation set, each of three teams submitted 3 systems for all genres and conditions. So there are 36 submissions in total. Our system using the context-oriented approach performed the best among all of them on all genres and conditions.

3.5 Ongoing and Future Work

We find that the sentiment ratio across different files and genres differs drastically. It would be useful to determine the sentiment level in each file first and apply different approaches to detect sentiment for files with high or low sentiment level.

4 English Sentiment 2

4.1 Basic Approach

Our general approach is based on relation extraction and directional sentiment as a relation between source and target. Relation extraction typically identifies relations between two entities in raw text. We extend the notion of relation to cover sentiment. The sources of the sentiment relation are always entities, while the target can be entities, events or relations. We use a supervised approach to detect positive or negative sentiment. The English Sentiment System-2 is build on top of SINNET system, which was de-

Test on Train on	Disc. Forums			Newswire			Disc. Forums + Newswire		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Disc. Forums	37.2%	74.4%	49.7%	15.5%	22.8%	18.5%	37.4%	59.4%	45.9%
Disc. Forums + Newswire	35.6%	75.3%	48.4%	19.6%	22.8%	21.1%	35.9%	70.4%	47.6%

Table 1 Results for our English Sentiment System-1 on "SuperDev" Data

veloped under DEFT funding and is a particular relation extraction system for social event extraction. Social events are the building blocks of social networks (Agarwal and Rambow, 2010).

4.2 Data Pre-Processing

Some significant pre-processing work was done to adapt the SINNET system to the task of source-and-target sentiment extraction (Figure 1).

We model source-and-target sentiment as a relation between the source and target. However the original system requires both parties to the relation to be mentioned explicitly in the text, and thus does not capture the implicit sentiment of the author. It only considers the sentiment relations between entities within the sentence. In order to capture those cases, we add at the beginning of each sentence the words *The author says*, with *the author* a mention of the author entity (which is already defined in discussion forums).

4.3 Approach

The English Sentiment System-2 is based on the SINNET system and relies on it for all machine learning aspects. The SINNET system uses tree kernels and Support Vector Machines (SVMs) to extract social events from natural language text. The SINNET system uses a combination of structures derived from the linear order, syntactic parses (phrase structure trees and dependency trees), and semantic parses. The English Sentiment System-2 takes as training data sentences with pairs of (entity, ERE) marked, and with an annotation for that sentence and (source, target) pair showing what type of sentiment the source has towards the target. The input to the system consists of two types of files: a raw text as well as text with entity and sentiment annotations. There are two main modules. The first is the linguistic pre-processing module which generates all linguistic information that is used by the machine learning (ML) module, and the ML module itself which uses SVMs with tree kernels.

4.4 Results

For developing our system we conducted a set of experiments on "SuperDev" to determine which model to use on each of the genres, as we did for our other sentiment system (see Section 3.4). The results from the experiments are reported in Table 2. As we can see, for both DF and NW, the best system to use is trained on DF only.

Results on the evaluation set are shown in Table 3. We see that our system 2 performs slightly worse than our system 1, but above baseline, on both genres.

4.5 Ongoing and Future Work

We plan to run experiments with different linguistic features, in particular semantic. We also plan to add sentiment-relevant features such as lexical features derived from the sentiment lexicons (which we have not yet used in our system 2). We plan to also work on those cases in which the source is not the author, since our system 2 based on SINNET can find those cases, while our system 1 cannot.

The training data is very skewed. There are many more pairs of entities without sentiment than pairs of entities that have sentiment. We can use random under-sampling and random over-sampling techniques to achieve better results. Also, the current system needs to be boosted with sentiment features in order to improve its accuracy.

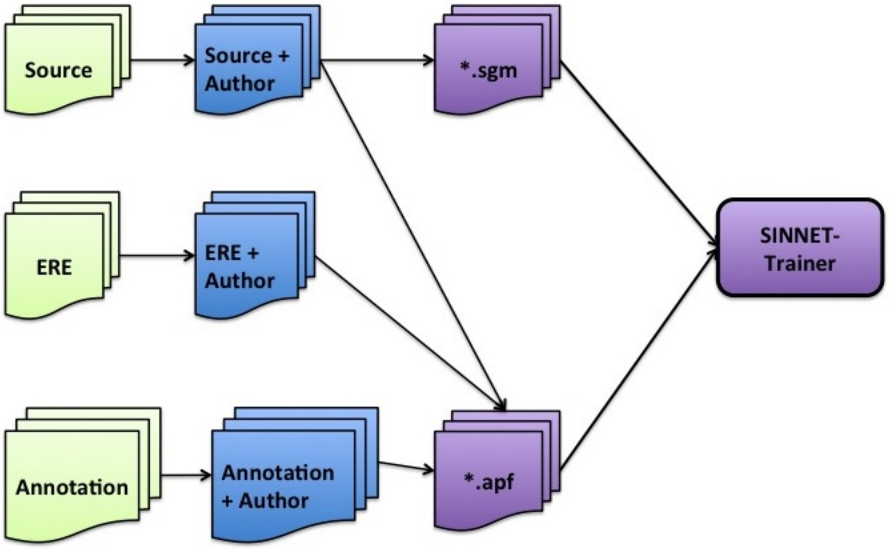

Figure 1 A pipeline to add The author says; convert ERE-annotated files into the proper APF (Ace Program Format) files and convert Source files to *.sgm format which is the format expected by the SINNET system.

Test on / Train on	Disc. Forums			Newswire		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Disc. Forums	35.5%	59.2%	44.4%	7.0%	13.0%	9.9%
Disc. Forums + Newswire	34.5%	57.0%	43.0%	4.0%	4.0%	4.0%

Table 2 Results for our English Sentiment System-2 on "SuperDev" Data

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	8.1%	70.6%	14.5%	3.7%	29.7%	6.5%
	Newswire	4.0%	35.5%	7.2%	2.3%	16.3%	4.0%
System 1	Disc. Forums	14.1%	38.5%	20.7%	6.2%	20.6%	9.5%
	Newswire	7.3%	16.5%	10.1%	2.7%	9.0%	4.2%
System 2	Disc. Forums	12.0%	38.3%	18.3%	5.5%	18.4%	8.4%
	Newswire	4.2%	5.6%	4.8%	2.4%	3.0%	2.7%

Table 3 Results for our two English Sentiment systems on KBP Eval Data, along with the baseline

5 Chinese Sentiment

5.1 Basic Approach

We follow the same approach as for English sentiment 1 (Section 3).

5.2 Data Pre-Processing

We segment sentences based on a list of some specific punctuations. For Chinese, it is critical to group words into meaningful phrases and to do POS tagging. This is performed by the Jiaba text segmentation package[1]. In addition, we treat all mentions except the author mentions as possible targets.

5.3 Approach

We apply the context-oriented method described in the English Sentiment 1 system (Section 3) to the Chinese sentiment task. We used Polyglot Chinese (Al-Rfou et al., 2013) to obtain word embeddings, and the HowNet Chinese Sentiment Lexicon[2].

5.4 Results

Because there are many annotation errors and few sentiment cases in the Chinese dataset, the model trained on this data does not performs well.

For developing our system we conducted a set of experiments on our dev set to determine which model to use. The results from the experiments are reported in Table 4. Also, our results on the real evaluation set are shown in Table 5.

There were a total of 7 systems submitted for all genres and conditions of Chinese sentiment. Our systems didn't do well on the Chinese Sentiment task (though we do beat the majority baseline for discussion forums).

5.5 Ongoing and Future Work

Because there are many annotation errors and few sentiment cases in the Chinese dataset, the model trained on this data performs badly. It would be helpful to have a new and similarly labeled dataset, or to investigate the use of alternate annotations.

6 Spanish Sentiment

6.1 Basic Approach

The main contributions of this approach involve generalizations of English capabilities for Spanish sentiment between a source and a target. We use a pre-processing step that limits the target range to only a relevant span of words. We also trained classifiers using features based on both Spanish lexicons and word embeddings. The pre-processing and training steps require Spanish dictionaries, corpora, tokenizers, and parsers. Reuse of these methods for other languages and domains would require similar resources.

6.2 Data Pre-Processing

¡ Similar to English pre-processing, we tokenize and parse the data to determine the relevant text. For Spanish, we used the Stanford CoreNLP tokenizer, POS tagger, and parser (Manning et al., 2014). The relevant Spanish phrasal categories to determine the small sentences are similar to English- 'S' also marks a clause.

6.3 Approach

The overall approach for Spanish is to use word embeddings combined with lexical features and other document features. We first identify the words in the relevant range for the target. Then we identify the word embedding for each word in this range and average these word embeddings to obtain a representation for the entire sequence. The word embeddings are the 300-dimensional embeddings created from the Spanish Billion-Word Corpus (Cardellino, 2016). We use a Spanish lexicon to count the number of polar words in each data instance (Perez-Rosas et al., 2012). We also included features to account for the overall polarity in a file, in an individual post, or from a specific author. These features were represented as an average of the target range embeddings for all data in a file, post, or from the same author. Finally, we included a feature based on the mention type of the target. The idea is that certain categories of mentions (i.e., businesses, life events) are more or less inherently likely to include sentiment towards the respective target. This feature was represented as a categorical, one-hot encoded feature.

[1] https://github.com/fxsjy/jieba

[2] http://www.keenage.com/html/e_index.html

Test on Train on	Disc. Forums		
	Prec.	Rec.	F-meas.
Disc. Forums	14.9%	25.0%	18.7%

Table 4 Results for our Chinese Sentiment System on the development dataset

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	5.8%	77.1%	10.8%	2.2%	8.3%	3.5%
	Newswire	1.1%	34.0%	2.1%	0.6%	3.7%	1.1%
System 1	Disc. Forums	12.6%	26.0%	17.0%	4.1%	0.6%	1.0%
	Newswire	2.5%	9.7%	4.0%	7.4%	0.3%	0.6%

Table 5 Results for our Chinese Sentiment system on KBP Eval Data, along with the baseline

We experimented with different training variations. For instance, in one variation we allowed the word embeddings to vary rather than be fixed for the entirety of training. Another variation is to learn a weighted average of the embeddings for each target range using an attention mechanism over words. We also experimented with averaging embeddings only over the content words (i.e., after removal of Spanish stop-words.). In addition, rather than a categorical representation of the mention type feature, we created an embedding for the mention type (varying between 10 and 15 dimensions). We also experimented with linear SVMs and up to 2 hidden layers of a Multi-Layer Perceptron. Hyperparameters were tuned via cross-validation.

Apart from the above experiments, we experimented with additional lexicon features from the Spanish Dictionary of Affect and Language (Rıos and Gravano, 2013) and Spanish SentiWordNet, but they did not help in improving the accuracy.

6.4 Results

In Table 6, we report the results on TAC KBP evaluation data for two systems, one using a linear SVM ($Sent1$) and one using a Multi-Layer Perceptron ($Sent2$) where in the latter system, the embeddings and weights were allowed to vary during training.

The final systems were the best-performing variations on development data for an SVM with a linear kernel and a Multi-Layer Perceptron (MLP).

- The first system ($Sent1$) was the linear SVM where the features included the average of the words in the target range, post, author, and file as well as a polarity score and the mention type as a categorical feature.

- The second system ($Sent2$) was the MLP with 2 hidden layers and an attention mechanism over all words in the target range. The embeddings and weights were allowed to vary during training.

Two teams submitted a total of 4 systems. $Sent1$ was the best performing system on all genres and conditions, except for newswire with gold ERE, where $Sent2$ performed the best. $Sent1$ beat the baseline on discussion forums with predicted and gold ERE and $Sent2$ beat the baseline on discussion forums with gold ERE.

6.5 Ongoing and Future Work

Our ongoing and future work include utilizing novel kernels for word embeddings, such as the model introduced in (Ghosh et al., 2015) where the authors utilized a greedy alignment technique to measure the similarity between two data instances as kernel similarity. Also, although we have utilized some lexicons in our current experiments, we are interested in investigating more lexicons, for instance the Spanish LIWC (Pennebaker et al., 2001).

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	9.2%	61.8%	16.1%	1.8%	5.1%	2.6%
	Newswire	5.3%	33.1%	9.1%	1.9%	3.9%	2.6%
$Sent1$	Disc. Forums	16.5%	35.8%	22.6%	1.8%	0.4%	0.6%
	Newswire	16.1%	2.3%	4.0%	8%	0.2%	0.4%
$Sent2$	Disc. Forums	18.0%	18.0%	18.0%	5.5%	18.4%	8.4%
	Newswire	19.1%	5.5%	8.5%	0%	0%	0%

Table 6 Results for our two Spanish Sentiment systems on TAC KBP Evaluation Data

7 English Belief

7.1 Approach

Our approach is to use an existing English word-level belief tagger (Prabhakaran et al., 2010; Werner et al., 2015). It tags each word which is the syntactic head of a proposition with one of the belief labels:

- CB for committed belief where the writer strongly believes in the proposition.

- NCB for non-committed belief, where the writer has a weak belief in the proposition.

- ROB for reported belief, which is the case when the writers intention is to report on someone elses stated belief, whether or not they themselves believe it or not.

- NA for propositions which are not beliefs. NA is discarded as it is not used in this evaluation.

All other words are tagged as “Other”. We then check whether the trigger for a relation mention or an event mention contains a word that has been tagged with a belief tag, and then apply that tag to the entire relation or event mention. We combine this approach with a majority baseline for our best results.

We have three systems (note that we do not represent them in order of their submission number, but in an order that allows an easier exposition):

- The word-tagger based system (System 2). In this system three sets of files have been applied in order to generate the final output, including input text files, output of the English belief tagger, and the ERE files (gold or predicted, as the case may be). By extracting target entities, relation from the ERE files we look for the associated committed belief tags within the output of English belief tagger. Moreover, using the word offset we find the corresponding authors for each case. This is a high-precision, low-recall system.

- The majority-baseline system (System 3). The majority Baseline labels all the extracted target entities, and relations as the majority belief tag, which is CB. The source is again the author of the passage which contains the target mention. This is a high-recall, lower-precision system.

- The combination system (System 1). System 1 employs both systems within the single pipeline. We applied System-3 to label all the untagged entities and relations, which were not tagged by the committed belief tagger and tag them as CB. Applying the combination system could lead to significant improvements in our experiments.

7.2 Results

Table 7 shows the performance of each system on Superdev dataset. As expected, the word-tagger based system (System 2) has the highest precision, but on this data set our combination does not actually beat the majority baseline. We hypothesize that this is related to the amount of newswire data in this dataset, for which the baseline System 2 performs quite well (as there is much committed belief).

The results on the evaluation set are shown in Table 8. Our System 3 is a majority baseline, but the official majority baseline (at the top of the table) includes a better matching algorithm for targets, and thus obtains better scores than our System 3. On discussion forums (which have more varied types of beliefs compared to newswire), the precision of both majority baselines is lower than in the superdev set, so that for the DF genre our System 2 (the word tagger-based system) can improve on precision. As a result, the combination System 1 outperforms our baseline System 3 (though not the improved official baseline). In contrast, for newswire, the baselines have good precision (and of course recall), so that our combined System 1 cannot improve on that performance.

In comparison to other submission, for discussion forums with gold ERE, we are the best performing system by a small margin, while other systems beat us by a small margin in the other conditions. For both genres and both ERE conditions, no system beats the official baseline.

7.3 Ongoing and Future Work

We will start out by incorporating the improved baseline system into our belief prediction system. We are in the progress of adding new set of features such as word embeddings, sentiment, negation and hedge words. In addition to novel features, various machine learning approaches will be applied in order to improve the overall performance of the Belief tagger.

We will also experiment using the approach based on relation extraction that we used for sentiment in our second system (Section 4). Thi system can detect sources which are not the author, and we expect it to contribute to newswire, where cases of reported belief are more common.

8 Chinese Belief

8.1 Basic Approach

Our basic approach is smilar to the approach we use for English (Section 7). We use an existing Chinese word-level belief tagger. Like the English word-level belief tagger, it tags each word which is the syntactic head of a proposition with one of the belief labels. All other words are tagged as “Other”. We then check whether the trigger for a relation mention or an event mention contains a word that has been tagged with a belief tag, and then apply that tag to the relation or event mention.

8.2 Data Pre-Processing

Because Chinese does not have space between words, we first use the Stanford Chinese word segmenter (Manning et al., 2014) to split sentences into sequences of words.

8.3 Approach

As in English, we have three systems:

- The word-tagger based system (System 2). We use the belief word-tagger presented in (Colomer et al., 2016). We part-of-speech tag the text. Then for each word, the system extracts the word, three words before this word, one word after this word, and their pos tags as a features. We predict the belief tag.
- The majority-baseline system (System 3).
- The combination system (System 1).

8.4 Results

The results on the evaluation set are shown in Table 9.

Summary of the results

- For predicted ERE, no system, including the baseline, found any beliefs, since no relations and few events were predicted; therefore, we do not show the predicted ERE condition. Since all the scores for Predicted ERE are 0, we are only going to compare the result of Gold ERE with DF and NW.
- We first look at discussion forums with gold ERE. As expected, our word-tagger System 2 has very low recall but a high precision. Combining our baseline system (System 3) with System 2 yields the system with the highest precision of all our systems, and also higher than the baseline, but at the cost of much lower recall. Thus, in terms of f-measure, our best system is, disappointingly, our baseline system, System 3.
- We now consider newswire with gold ERE. Among our three systems, our word-tagger System 2 fails to find any instances of belief, and thus combination of System 2 with out

System	Superdev		
	Prec.	Rec.	F-meas.
System 1 (Combination)	77.78%	85.57%	81.49%
System 2 (Word tagger)	83.10%	24.87%	38.28%
System 3 (Majority)	78.15%	85.50%	81.66%

Table 7 Results for our three English Belief systems on the superdev set

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	69.67%	89.42%	78.32%	14.06%	7.34%	9.65%
	Newswire	82.65%	57.37%	67.73%	23.64%	5.47%	8.88%
System 1	Disc. Forums	74.92%	81.03%	77.85%	8.88%	2.26%	3.60%
	Newswire	83.79%	53.75%	65.49%	2.05%	2.08%	3.78%
System 2	Disc. Forums	77.42%	24.45%	37.16%	14.30%	14.08%	2.56%
	Newswire	85.86%	55.64%	66.43%	32.25%	1.30%	2.51%
System 3	Disc. Forums	68.26%	85.86%	76.06%	8.33%	2.77%	4.16%
	Newswire	86.21%	53.13%	65.74%	1.93%	2.19%	3.93%

Table 8 Results for our three English Belief systems on KBP Eval Data, along with the official baseline

baseline System 3 does not improve on System 3 (neither on precision, nor on recall). Our baseline System 3 got the highest score among all teams, but it is worse than the official baseline system.

- For the predicted ERE, no system (including the official baseline) predicts any belief, and we omit the numbers.

9 Spanish Belief

For Spanish belief, we adopted a simpler approach, which we tuned on the entire data set.

- System 1 is a baseline system, with a strict matching against targets.
- System 2 is the same baseline system, but we model a different probability for each label (CB, NCB, ROB) for each type of target (relation type or event type).
- System 3 is a slight variant of System 1 (an attempt at finding more targets), which did not produce the desired results.

The results on the evaluation set are shown in Table 10. For predicted ERE, no system, including the baseline, found any beliefs, since no relations and few events were predicted; therefore, we do not show the predicted ERE condition. We see that modeling the target types separately (system 2) provides a large boost. Our System 2 beat the only other system submitted on both genres. The offical baseline beats all submitted systems.

10 Conclusion

There are several clear areas for us to improve our existing systems.

1. We need to train systems on predicted ERE data.
2. We need to experiment with our relation-extraction system (Section 4) in order to add sentiment-specific features.
3. We need to incorporate the improved baselines into our belief systems. For all three languages, we start with a baseline system and improve on the one we used. We will use the updated baseline system which is the official baseline and investigate if our methods improve on it.

System	Genre	Gold ERE		
		Prec.	Rec.	F-meas.
Baseline	Disc. Forums	80.77%	87.70%	84.09%
	Newswire	81.95%	60.23%	69.43%
System 1	Disc. Forums	82.66%	67.67%	74.42%
	Newswire	79.72%	53.02%	63.68%
System 2	Disc. Forums	74.37%	11.12%	19.34%
	Newswire	100.00%	0.00%	0.00%
System 3	Disc. Forums	79.38%	79.98%	79.68%
	Newswire	80.83%	57.15%	66.96%

Table 9 Results for our three Chinese Belief systems on KBP Eval Data, along with the official baseline. Note that no belief was predicted on the predicted ERE, neither by the official baseline, nor by our systems; we therefore omit this condition from the table.

System	Genre	Gold ERE		
		Prec.	Rec.	F-meas.
Baseline	Disc. Forums	76.77%	77.39%	77.08%
	Newswire	74.78%	54.21%	62.86%
System 1	Disc. Forums	53.64%	45.19%	49.06%
	Newswire	59.96%	34.68%	43.94%
System 2	Disc. Forums	63.86%	69.65%	66.63%
	Newswire	64.90%	48.92%	55.79%
System 3	Disc. Forums	53.68%	45.20%	49.08%
	Newswire	59.95%	34.68%	43.94%

Table 10 Results for our three Spanish Belief systems on KBP Eval Data, along with the official baseline. Note that no belief was predicted on the predicted ERE, neither by the official baseline, nor by our systems; we therefore omit this condition from the table.

4. We need to make use of all available annotated resources: currently, we do not actually train on the belief portions of the source-and-target belief and sentiment corpora, we only use them for tuning.

5. We need to investigate in more detail how to train systems for different conditions. Specifically, even within a genre such as discussion forums, the amount of sentiment expressed can differ from what was encountered in training, which can decrease results.

We expect the performance of automatic source-and-target belief-and-sentiment systems to greatly increase over the next few years as the data is better understood and appropriate features and machine learning techniques are discovered.

Acknowledgments

The work of the following co-authors of this paper was supported by the DARPA DEFT Program: Rambow, Yu (in part), Radeva, Hamidian, Ghosh, Hidey, Diab, McKeown, Muresan. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1034. Association for Computational Linguistics.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings, March.

Juan Pablo Colomer, Keyu Lai, and Owen Rambow. 2016. Detecting level of belief in chinese and spanish. In *Proceedigs of the ExProM Workshop at Coling*, Osaka. To appear.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal, September. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *LREC*, volume 12, page 73.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August. Coling 2010 Organizing Committee.

Matıas G DellAmerlina Rıos and Agustın Gravano. 2013. Spanish dal: A spanish dictionary of affect in language. *WASSA 2013*, page 21.

Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado, June. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.