The Cool Cucumber System at the 2017 TAC KBP BeSt Evaluation

Thanh-Son Nguyen
Singapore Management University
School of Information Systems
tsnguyen.2013@smu.edu.sg

Yufang Hou
IBM Research - Ireland
Dublin, Ireland
yhou@ie.ibm.com

Charles Jochim
IBM Research - Ireland
Dublin, Ireland
charlesj@ie.ibm.com

Elizabeth M. Daly
IBM Research - Ireland
Dublin, Ireland
elizabeth.daly@ie.ibm.com

Léa A. Deleris
IBM Research - Ireland
Dublin, Ireland
lea.deleris@ie.ibm.com

ABSTRACT

The 2017 BeSt Evaluation is an evaluation of targeted belief and sentiment detection. It seeks to answer the question "*who has what mental attitude towards whom/what?*". Our team participated in the sentiment evaluation for English, starting from known mentions of entities, relations and events. In this situation, the difficulty lies in part in determining the pairs that carry sentiment from the multitude of potential pairs that can be created by combining all the known mentions. To deal with this issue, we propose a two-stage classifier to (1) eliminate objective pairs (i.e., pairs that do not have sentiment) and (2) predict sentiment polarity values for the predicted-subjective pairs.

1 INTRODUCTION

The 2017 BeSt Evaluation is an evaluation of targeted belief and sentiment detection. It seeks to answer the question "*who has what mental attitude towards whom/what?*".

The datasets provided include documents from *discussion forums* and *newswire text*. For each document, we have information about:

> *Source*: the original content of the document (including tags).
> *ERE*: containing information about *Entities*, *Relations*, *Events*, as well as the information of the corresponding *mentions* (e.g., offset, length, id) in the source file.
> *Annotation*: gold-pair annotations (.best.xml), including *belief* and *sentiment* annotations.

The task has different evaluation scenarios including two different types of prediction labels (i.e., *belief* and *sentiment*) and three different languages (i.e., English, Chinese, and Spanish).

Our team participated in the evaluation of *sentiment* for English. The task is defined as follows:

Problem 1 (BeSt Sentiment Evaluation Task). *Given the text content and the ERE annotations of a document, extract pairs between entity and entity, relation, or event that have sentiment and the sentiment polarity values of the pairs.*

Given the *source* (text content) and the *ERE* file, we need to determine the pairs of mentions that have sentiment, and the value of the associated sentiment. The source is always an entity mention. The target for sentiment can be an entity mention, a relation mention or an event mention. Possible values for sentiment are *pos* (positive) and *neg* (negative).

The evaluation focuses not only on the sentiment values but also on determining which pairs contain positive or negative sentiment. This constitute an essential step in the task as one can form a large collection of potential pairs from the combinations of entities and entities, relations, or event. If we train a highly effective classifier to predict sentiment for all the possible pairs, we might be able to predict correct sentiments for the gold pairs (i.e., the pairs that are annotated in the gold-pair annotations) but the performance for the BeSt evaluation would remain low as we include too many false positive pairs (i.e., pairs that do not appear in the gold-pair annotations).

To deal with this issue, we propose to use a two-stage classifier. In the first stage, we build a subjective/objective classifier, which identifies which pairs contain sentiment (i.e., subjective, or sentiment pairs) and which do not (i.e., objective or no-sentiment pairs). The pairs predicted as subjective are passed to the second stage, a sentiment classifier, to be assigned with positive or negative label.

The approach is similar to Wilson et al.'s [6], even though the objectives are different. While our method may seem similar to an one-vs.-all strategy in a multi-label classifier, it is indeed different. Specifically, if we apply one-vs.-all, we need to build three models, one for classifying "objective vs. (negative or positive)", one for "negative vs. (objective or positive)", and one for "positive vs. (objective or negative)". The latter two have to differentiate pairs having sentiment with a group of pairs having sentiment and no sentiment. Our hypothesis is that it makes more sense to differentiate pairs having sentiment with pairs that do not. Therefore, our two-stage classifier is proposed to avoid grouping sentiment pairs and no-sentiment pairs into one class.

2 DATASET

We use the datasets provided by the organizer:

Table 1: Information of the Number of Documents, Number of Positive and Negative Gold Pairs for Discussion Forum and Newswire Documents in E27 and E114 Datasets

Dataset	DocSource	#Docs	#PosPairs	#NegPairs
E27	Disc. Forum	209	1079	2792
	Newswire	37	109	292
E114	Disc. Forum	84	511	539
	Newswire	81	606	644

LDC2016E27_DEFT_English_Belief_and_Sentiment_Annotation_V2: we call *E27* for short.
LDC2016E114_TAC_KBP_2016_Belief_and_Sentiment_Evaluation_Gold_Standard_Annotation_V2: we call *E114* for short.

We use *E27* for training, and keep *E114* for evaluating the models.

Gold pairs. We extract gold annotated pairs[1], i.e., pairs that are annotated in the gold-pair annotations, based on the *annotations* files. For each *target*, if there is a *source* having a valid sentiment (i.e., positive or negative), we form a gold pair of the *source* and *target* with the corresponding sentiment label. We only consider the targets appearing in the valid content of the document (i.e., not in the *quotes* content if the document is from *discussion forum*).

As we can see in Table 1, for the training dataset *E27*, most of the documents are from discussion forums. Regarding to sentiment labels, the distribution is clearly skewed towards *negative*, with more than 70% of the gold pairs being negative for both document sources. *E114* is more balanced in both the number of documents in each source and the sentiment labels.

3 TWO-STAGE CLASSIFICATION

Figure 1 displays the two-stage classifier framework. Given a document, from the input text content and the ERE annotation, we assemble a set of potential pairs by defining rules to combine mentions together. Then, the first stage will classify those potential pairs into subjective or objective. All the subjective pairs are then used as input for the second stage, sentiment classification. For both stages, we train the models based on *Support Vector Machine*. Training and classification are performed using SVMLight[3].

We now describe this process in more details. We outline our potential pairs extraction strategy in Section 3.1, Section 3.2 discusses the features that we use. Finally, the two classifiers are described in Section 3.3 and Section 3.4.

3.1 Potential Pairs Extraction

To form a potential pair, we need to identify the source and the target of that pair. For both types of document, we require the target of a pair to belong to a "*valid sentence*". A

[1] We call \gold pairs" for short.

Table 2: Dataset *E27*: Number of Positive or Negative Gold Pairs in the Generated Potential Pairs and Coverage Over All the Gold Pairs

DataSource	#Positive (%Cov)	#Negative (%Cov)
Disc. Forum	589 (55%)	1564 (56%)
Newswire	30 (28%)	56 (19%)

valid sentence is the sentence that does not appear in *quotes*[2] (for discussion forum documents) and contains at least one sentiment word[3] (for both discussion forums and newswire documents).

We chose to rely on different heuristic rules to generate potential pairs for documents from discussion forums and newswire.

Discussion forum. Discussion forum documents are in form of posts where each post has an author. Authors usually write posts to express their own opinions about a topic (e.g., people, event). Therefore, for this type of document, we use the author as the source, and the entity, relation or event mentions in the post as the targets. We only consider the targets that appear in valid contents.

Source: the post's author
Target: entity mention, relation mention or event mention.

Newswire. Unlike discussion forum documents, newswire documents are in the form of paragraphs. Some newswire documents do not have any author, some have at most one author. For this type of document, we form the pairs as follows:

Source: entity mention (only PER (person) or ORG (organization)) or the author (if applicable)
Target: entity mention (not the same as source), relation mention or event mention appearing in the same sentence with the source (if the source is not the author).

To prevent over generating pairs, we limit the pairs that each target can form to at most three, using closest entity mentions (excluding the author).

To have a sense of how good the proposed strategy is in terms of covering the gold pairs, Table 2 shows the number of gold positive and negative pairs appearing in the generated potential pairs and also the percentage of the coverage for the dataset *E27*. Following the rules described above, for discussion forum documents, our method covers 55% of the positive gold pairs and 56% of the negative gold pairs. It is lower for newswire documents with 28% and 19% for positive pairs and negative pairs, respectively.

As we mentioned before, a significant challenge for this task is to identify the subjective pairs. Table 3 shows the number of subjective and objective pairs generated from our rules for potential pair extraction. In all the generated potential pairs for discussion forums, only about 24% are

[2] As instructed by the task description
[3] Based on Bing Liu's sentiment lexicon[2]

Figure 1: Two-Stage Classifier

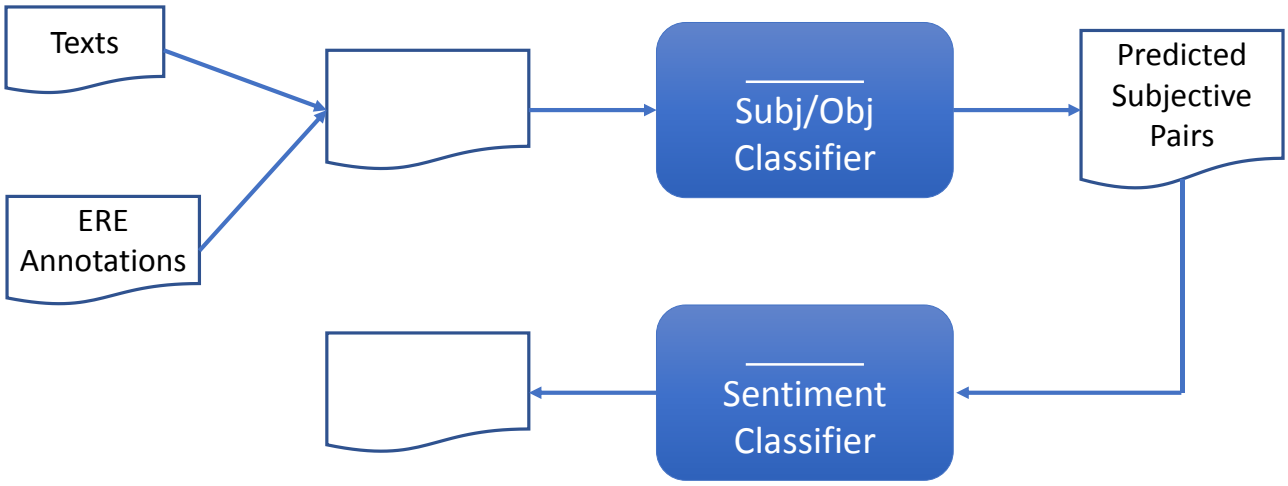

Table 3: Dataset *E27*: Number and Percentage of Subjective Pairs and Objective Pairs within the Generated Potential Pairs

DataSource	#Subjective (%)	#Objective (%)
Disc. Forum	2153 (24%)	6933 (76%)
Newswire	86 (1%)	9175 (99%)

annotated as having sentiment and more than 70% of the pairs are objective. It is even worse for newswire as there are only 1% of the generated potential pairs are subjective and the rest, 99%, is objective.

We investigated modifications to our rules so as to reduce the proportion of missed potential pairs. However, from our explorations, adding a new rule to increase the recall of gold pairs led to the generation of many more objective pairs. Altogether, the percentage of subjective pairs was decreasing. Therefore, we decided to keep the rules as described.

3.2 Feature Extraction

We have two main types of feature. The first type is based on weighted word embeddings and the second one is based on sentiment lexicon.

Weighted word embeddings-based features. We use the pre-trained Glove word embeddings 840B.300d from [4] containing 300-dimension vectors for 840 billion tokens. As in [5], the word embeddings are weighted based on the part-of-speech (POS) tag of the word. The POS tags weights are manually assigned. For example, the weight for *JJ* (*adjective*, e.g., nice) is 1.0 because adjective words usually indicate sentiment polarity, while the weight for *DT* (*determiner*, e.g., the) is 0.05 as stop words are not important for the task. In addition to the POS tags weights, word embeddings are also weighted based on the distance between the word and the target. The distance weight is computed as $d^{-0.5}$, following [1], where d is the number of tokens between the word and the corresponding target.

Sentiment lexicon-based features. For sentiment lexicon-based features, we count the number of positive and negative words in a text content (e.g., a sentence, a window of text around the target) based on Bing Liu's sentiment lexicon [2]. It is then normalized by a pre-counted max sentiment counts.

We also use some other features based on source/target, such as "is the source the author?", "Is source in the sentence?", or "whether the source appears before the target in the sentence".

3.3 Stage 1: Subj/Obj Classifier

The main goal of this stage is to identify the subjective pairs among the potential pairs.

The features used in this stage are:

> Feature source: a text window of tokens around the target will be used to generate the features. To extract the text window, we take N tokens before the target, the target and N tokens after the target. In our experiments, we use $N = 5$.
> Features: weighted word embeddings and normalized sentiment words count, as described in Section 3.2.

The classifier in this stage is trained using the following setting:

> Training data: generated potential pairs from the training set.
> Labels: A potential pair that is found in the gold-pair annotations will be labelled as 1 (i.e., subjective), otherwise -1 (i.e., objective).
> Learning: as we have many more objective data instances compared to subjective instances, we control the proportion of objective pairs in the training set

α	*Prec.*	*Recall*	*F*-Score
30%	11.40%	49.50%	18.60%
40%	13.20%	42.10%	20.10%
50%	**17.10%**	**31.80%**	**22.20%**
60%	23.40%	16.60%	19.40%
70%	42.70%	1.20%	2.40%
80%	100.00%	0.00%	0.00%

(a) Train: DF; Test: DF

α	*Prec.*	*Recall*	*F*-Score
30%	2.30%	42.20%	4.30%
40%	3.40%	27.40%	6.10%
50%	**4.80%**	**17.40%**	**7.50%**
60%	6.20%	8.10%	7.00%
70%	22.20%	0.30%	0.60%
80%	100.00%	0.00%	0.00%

(b) Train: DF; Test: NW

α	*Prec.*	*Recall*	*F*-Score
30%	12.20%	46.10%	19.30%
40%	12.90%	43.70%	20.00%
50%	14.00%	40.00%	20.70%
60%	**15.60%**	**32.10%**	**21.00%**
70%	21.30%	18.90%	20.10%
80%	47.40%	2.90%	5.40%

(c) Train: DF+NW; Test: DF

α	*Prec.*	*Recall*	*F*-Score
30%	3.50%	24.30%	6.10%
40%	3.90%	16.00%	6.20%
50%	**4.40%**	**10.70%**	**6.30%**
60%	6.20%	6.30%	6.20%
70%	9.20%	3.00%	4.60%
80%	22.20%	0.30%	0.60%

(d) Train: DF+NW; Test: NW

Table 4: Stage 1 Training: Varying Objective Proportion (α). Evaluations on Different Training and Testing Datasets. DF: Discussion Forum; NW: Newswire.

such that only α% of the training data instances are objective. We determine the optimal value for α in Section 4.1.

3.4 Stage 2: Sentiment Classi er

The goal of the second stage is to assign predicted-subjective pairs from stage 1 a *positive* or *negative* sentiment label.

The features used in this stage are as follows:

> Feature source: the sentence containing the target will be used to generate the features for a pair.
> Features: weighted word embeddings, normalized sentiment words count, and source/target-based features, as described in Section 3.2.

For training, all the gold pairs extracted from the training set will be used .

4 EXPERIMENTS

To evaluate the proposed two-stage classifier, we use dataset *E27* as the training data and keep dataset *E114* as the testing data. We use the evaluator script provided for the task to evaluate the predicted results[4]. As we have too few newswire documents in *E27*, we have two different settings for the training documents:

> **DF**: the model is trained on discussion forum documents only
> **DF+NW**: the model is trained on discussion forum documents and newswire documents

[4] The evaluation is based on the generated 4-tuples. For more details, please refer to the task's description, available at http://www.cs.columbia.edu/~rambow/best-eval-2017/task-spec-v2.8.pdf

4.1 Stage 1 Training: Varying Objective Proportion

As mentioned before, we control the proportion of the objective pairs in training for stage 1, α. To see how the proportion affects the performance and to choose a trade-off value, we vary the percentage of objective pairs when training stage 1. Table 4 shows the results for different values of the objective proportion α as we train and test on different dataset (as noted in the subtables). For any setting, as α is decreased, recall increases and precision decreases. The reason is that when we decrease the proportion of objective pairs in the training set, stage 1 tends to predict more subjective pairs, thus increasing the recall (as gold subjective pairs have more chance to be predicted correctly by stage 1). However, it will also increase *false positive* as many pairs predicted subjective are actually objective. Based on the *F*-score shown in subtables, we choose $\alpha = 50\%$ for later experiments.

4.2 Comparing with the Baseline

We compare our results with the results generated by the baseline system described in [5]. In the baseline, the source is always the author (*null* if there is no author), the target is each mention of entity, relation or event, and the sentiment is always *negative*.

We train our models using dataset *E27*, training on discussion forum documents only, and use $\alpha = 50\%$ for the proportion of objective pairs in stage 1's training. We generate sentiment pairs (i.e., subjective pairs with sentiment polarity values) using the trained models for dataset *E114*. The sentiment pairs for the baselines are also generated for dataset *E114*.

Table 5 shows the comparison results between the baseline and our two-stage classifier (i.e., *Cool-Cucumber*). *Cool-Cucumber* underperforms compared to the baseline in terms

Table 6: Cool-Cucumber's Performance on the Final Evaluation Data

	# Files	*Cool-Cucumber*	*Prec.*	*Recall*	*F*-Score
Disc. Forum	84	System 1	28.4%	21.9%	24.7%
		System 2	17.2%	41.2%	24.3%
		System 3	27.6%	17.5%	21.4%
Newswire	83	System 1	7.2%	11.4%	8.8%
		System 2	7.2%	11.4%	8.8%
		System 3	7.5%	8.6%	8.0%

Table 5: Comparing with the Baseline. Training: *E27*, Testing: *E114*

		Prec.	*Recall*	*F*-Score
Disc. Forum	*Baseline*	8.1%	70.6%	14.5%
	Cool-Cucumber	17.1%	31.8%	22.2%
Newswire	*Baseline*	4.0%	35.5%	7.2%