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

Owen Rambow, Tao Yu, Axinia Radeva, Sardar Hamidian,
Alexander Fabbri, Debanjan Ghosh
Christopher Hidey (PRESENTER)
Tianrui Peng, Mona Diab
Kathleen McKeown, Smaranda Muresan

Columbia, George Washington, Rutgers
rambow@ccls.columbia.edu

November 15, 2016

Contents

Data We Used

- LDC2016E27_DEFT_English_Belief_and_Sentiment_Annotation_V2
- LDC2016E61_DEFT_Chinese_Belief_and_Sentiment_Annotation
- LDC2016E62_DEFT_Spanish_Belief_and_Sentiment_Annotation

No other data sources

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

Assumption: The source is the author; in vast majority of sentiment cases for both discussion forum and newswire data sets are from the author.

We pursue two approaches.

- **Target-oriented approach**: target-specific features.
  - Long complex sentences
  - Many possible targets per sentence
  - We isolate potential targets in "small sentences" using a parser
- **Context-oriented method**: consider larger context.
  - Do not use "small sentences"
  - Instead model larger context (post, all posts by author, file) using word embeddings

We use the context-oriented method as it performs better

Features

We employ widely used text classification features and task-specific features:

- Word embeddings
- Sentiment word counts
- Mention types of the target

The features are extracted on the target, sentence, post and file levels.
We use Support Vector Machines (SVM) with linear kernels and Random Forest classifiers.

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

Test on ⟶	Disc. Forums			Newswire		
Train on ↓	Prec.	Rec.	F-ms.	Prec.	Rec.	F-ms.
Disc. For.	37.2%	74.4%	49.7%	15.5%	22.8%	18.5%
Disc. For. + Newswire	35.6%	75.3%	48.4%	19.6%	22.8%	21.1%

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

We treat source-and-target sentiment as a relation extraction from source to target; reuse SINNET for social event extraction (Agarwal & Rambow 2010)

- Replace potential source and target by marker
- Use many linguistic representations (linear, phrase structure syntax, dependency syntax, FrameNet parse)
- Use sequence and tree kernels

Caveat: we did not introduce sentiment-specific features (lack of time)

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

Test on ⟶	Disc. Forums			Newswire		
Train on ↓	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Disc. For.	35.5%	59.2%	44.4%	7.0%	13.0%	9.9%
Disc. For. + Newswire	34.5%	57.0%	43.0%	4.0%	4.0%	4.0%
Best Sys-1	37.2%	74.4%	49.7%	19.6%	22.8%	21.1%

Not bad on DF, given that we are using no sentiment-specific features!

Results for our English Sentiment Systems on Eval Data

Boldface = top F-measure in eval

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

Basic Approach

- Same approach as for English sentiment 1 (context-oriented method)
- Word segmentation, POS tagging, Polyglot word embeddings
- HowNet Chinese Sentiment Lexicon

Results for our Chinese Sentiment System on “SuperDev” Data

Low performance due to:

- Few sentiment cases
- Annotation errors

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

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

- Same approach as for English sentiment 1 (context-oriented method)
- Stanford CoreNLP Spanish tokenizer, POS tagger, and parser
- Word embeddings from Spanish Billion-Word Corpus
- Spanish Sentiment Lexicon (Pérez-Rosas et al., 2012)
- System 2 uses the same features as System 1, but uses a 2-layer MLP and allows the embeddings to vary during training

Results for our Spanish Sentiment Systems on Eval Data

Boldface = top F-measure in eval

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	DF	9.2%	61.8%	16.1%	1.8%	5.1%	2.6%
	NW	5.3%	33.1%	9.1%	1.9%	3.9%	2.6%
*Sent*1	DF	16.5%	35.8%	**22.6%**	7.4%	2.0%	**3.2%**
	NW	16.1%	2.3%	4.0%	8%	0.2%	**0.4%**
*Sent*2	DF	18.0%	18.0%	18.0%	1.8%	0.4%	0.6 %
	NW	19.1%	5.5%	**8.5%**	0%	0%	0%

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

Three systems:

- System 3: A default system (every target is CB)
- System 2: A word-based tagger, based on 2014 evaluation (Werner et al. 2015); high-precision, low recall
- System 1: Combination system: If System 2 makes a prediction for a target, use its prediction; otherwise, use System 3

English Belief Results

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%

On the "superdev" set (more newswire), promise of system combination does not pay off

Results for our English Belief Systems on Eval Data

Boldface = top F-measure in eval

Sys.		Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Bl.	DF	69.67%	89.42%	78.32%	14.06%	7.34%	9.65%
	NW	82.65%	57.37%	67.73%	23.64%	5.47%	8.88%
S1	DF	74.92%	81.03%	**77.85%**	8.88%	2.26%	3.60%
	NW	83.79%	53.75%	65.49%	20.56%	2.08%	3.78%
S2	DF	77.42%	24.45%	37.16%	14.30%	1.41%	2.56%
	NW	85.93%	15.60%	26.40%	32.25%	1.30%	2.51%
S3	DF	68.26%	85.86%	76.06%	8.33%	2.77%	4.16%
	NW	82.41%	55.65%	**66.43%**	19.33%	2.19%	3.93%

Here, for DF, our system combination System 1 pays off

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

- Follow English approach
- System 3 = majority baseline system
- System 2 = high-precision, low-recall, uses Chinese word tagger (Colomer at al. 2016)
- System 1 = combination of System 3 + System 2 when it makes a prediction
- Vary parameters to get high-recall and high-precision systems

Results for our Chinese Belief Systems on Eval Data

Boldface = top F-measure in eval; no results by any team on predicted ERE

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

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Basic Approach

- Used simple approach based on probability of different belief categories given target type
- Random choice with hand-tuned probabilities based on observed probabilities

Adding choice based on target type boosts performance considerably (= System 2)

Results of Spanish Belief System

Boldface = top F-measure in eval; no results by any team on predicted ERE

System	Genre	Gold ERE		
		Prec.	Rec.	F-meas.
Baseline	DF	76.77%	77.39%	77.08%
	NW	74.78%	54.21%	62.86%
System 2	DF	63.86%	69.65%	**66.63%**
	NW	64.90%	48.92%	**55.79%**

Data

English Sentiment 1

English Sentiment 2

Chinese Sentiment

Spanish Sentiment

English Belief

Chinese Belief

Spanish Belief

Ongoing and Future Work

- Sentiment ratio across different files and genres differs drastically; develop system to probe amount of sentiment first before making specific predictions?
- Sentiment: use of relation extraction approach promising; will add more features and investigate how we can combine it with target-focused approach
- Belief: will use relation extraction approach on belief to capture non-author beliefs
- Belief: will use better “official” baseline in all languages

Thanks!

Questions?