

The BeSt Eval at the 2016 NIST TAC KBP

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

Overview

- BeSt Eval
 - Task
 - The Role of ERE Annotation
- Data
 - Basic Annotation
 - Differences in Belief vs. Sentiment
 - Differences by Genre
 - Differences in Gold vs. Predicted ERE
- Evaluation Script
- Submitted Systems and Results
- Conclusions

BeSt Eval


- BeSt Eval organized by the DEFT BeSt group
 - Albany, Columbia, Cornell, GWU, IHMC, LDC, MITRE, NIST, Pittsburgh
- Task: Evaluate addition of belief and sentiment to existing KB objects (EREs)
 - EREs are the sources and targets
 - Want to evaluate KB population, not text tagging
 - Want to exclude ERE KBP tasks from belief and sentiment tasks
 - Allows component-level research improvements and system development
- First evaluation to cover both belief and sentiment

BeSt Eval:

The Role of ERE Annotation

- Assume ERE annotation as input
 - ERE annotation (LDC): straightforward representation of entities, relations and events in KB with pointers to mentions in text
 - Distinction between object vs. object mention
- Currently no cross-document co-reference in LDC gold or predicted ERE data, so analysis is one document at a time
 - If cross-document co-reference is available, nothing changes for evaluation framework
 - Most systems would not change given cross-document co-reference

Two Conditions for EREs

- Use gold ERE annotation from LDC
 - Use predicted annotation
 - From RPI, co-reference by Stanford, much support from UIUC – many thanks!
 - Transformed at Columbia into ERE format
 - Task of creating predicted ERE file is not straightforward, since we need to link it to gold BeSt file so we can perform evaluation
 - Basically same problem as evaluating ERE!
-  Mapping from predicted EREs required *exact* match on mention/trigger or argument mentions

Data: Basic Annotation

English	All data	Discussion Forums (%)	Newswire (%)
Train	157K words	89%	11%
Evaluation	88K words	52%	48%

Spanish	All data	Discussion Forums (%)	Newswire (%)
Train	79K words	100%	0%
Evaluation	67K words	61%	39%

Chinese	All data	Discussion Forums (%)	Newswire (%)
Train	133K words	100%	0%
Evaluation	122K words	65%	35%

Data: Belief vs. Sentiment Disc. Forums vs. Newswire

Percentage of targets that have:

	All data	Discussion Forums	Newswire
Sentiment from any source	18.9%		
Sentiment from author	16.3%		
Sentiment from other source	2.6%		
Belief from any source			
Belief from author			
Belief from other source			

Data: Belief vs. Sentiment Disc. Forums vs. Newswire

Percentage of targets that have:

	All data	Discussion Forums	Newswire
Sentiment from any source	18.9%	21.2%	6.8%
Sentiment from author	16.3%		
Sentiment from other source	2.6%		
Belief from any source			
Belief from author			
Belief from other source			

Data: Belief vs. Sentiment Disc. Forums vs. Newswire

Percentage of targets that have:

	All data	Discussion Forums	Newswire
Sentiment from any source	18.9%	21.2%	6.8%
Sentiment from author	16.3%	19.0%	1.8%
Sentiment from other source	2.6%	2.2%	5.0%
Belief from any source			
Belief from author			
Belief from other source			

Data: Belief vs. Sentiment Disc. Forums vs. Newswire

Percentage of targets that have:

	All data	Discussion Forums	Newswire
Sentiment from any source	18.9%	21.2%	6.8%
Sentiment from author	16.3%	19.0%	1.8%
Sentiment from other source	2.6%	2.2%	5.0%
Belief from any source	100%	100%	100%
Belief from author	94.3%	99.3%	79.2%
Belief from other source	5.7%	0.7%	20.8%

Note: Belief includes “NA” tag which was not included in evaluation

Evaluation Script

- Eval script written at Columbia based on community consensus
- Goal: evaluate accuracy of links added to KB
 - Not focused on text annotation (except for Provenance)
- Target must be correct
- Partial credit
 - For incorrect source
 - If value of sentiment (pos, neg) or of belief (CB, NCB, ROB) is wrong
 - For target “provenance”, two conditions:
 - At least one span in list must be correct (WHAT WE USED)
 - Score weighted by the F-measure of predicted mentions against correct mentions
 - “At-least-one” condition gets pretty consistently 2% better scores than the weighted approach, with no change in order of system results

BeSt Eval Tasks

24 conditions:

- 2 cognitive attitudes (belief and sentiment)
- 3 languages
- 2 conditions (gold ERE and predicted ERE)
- 2 genres

Because of important differences in data, each condition is very different

BeSt Eval Participants

Belief: Beat the Baseline

	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
compittmich	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)

BeSt Eval Participants

Sentiment: Top Performers

	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
compittmich	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)

Conclusions/Outlook

- Participation low: hard and new problem
- Need to review matching of predicted ERE to gold ERE
 - No predicted relations/events at all in Chinese!
 - Be more lenient?
- Set of conditions very complex, maybe need to simplify