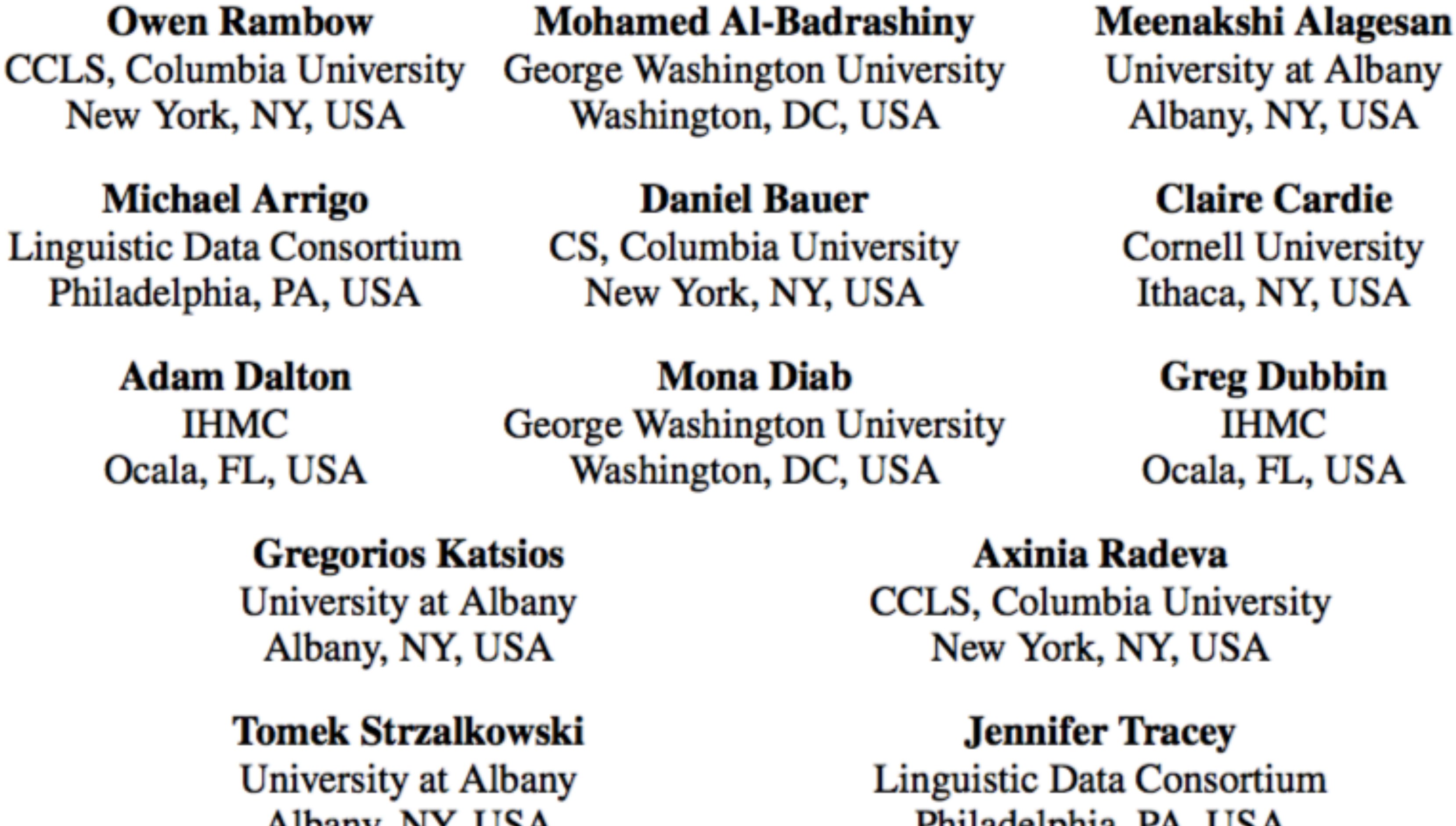
The BeSt Eval at the 2017 NIST TAC KBP



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

Mohamed Al-Badrashiny
George Washington University
Washington, DC, 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

Mona Diab
George Washington University
Washington, DC, USA

Greg Dubbin
IHMC
Ocala, FL, 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

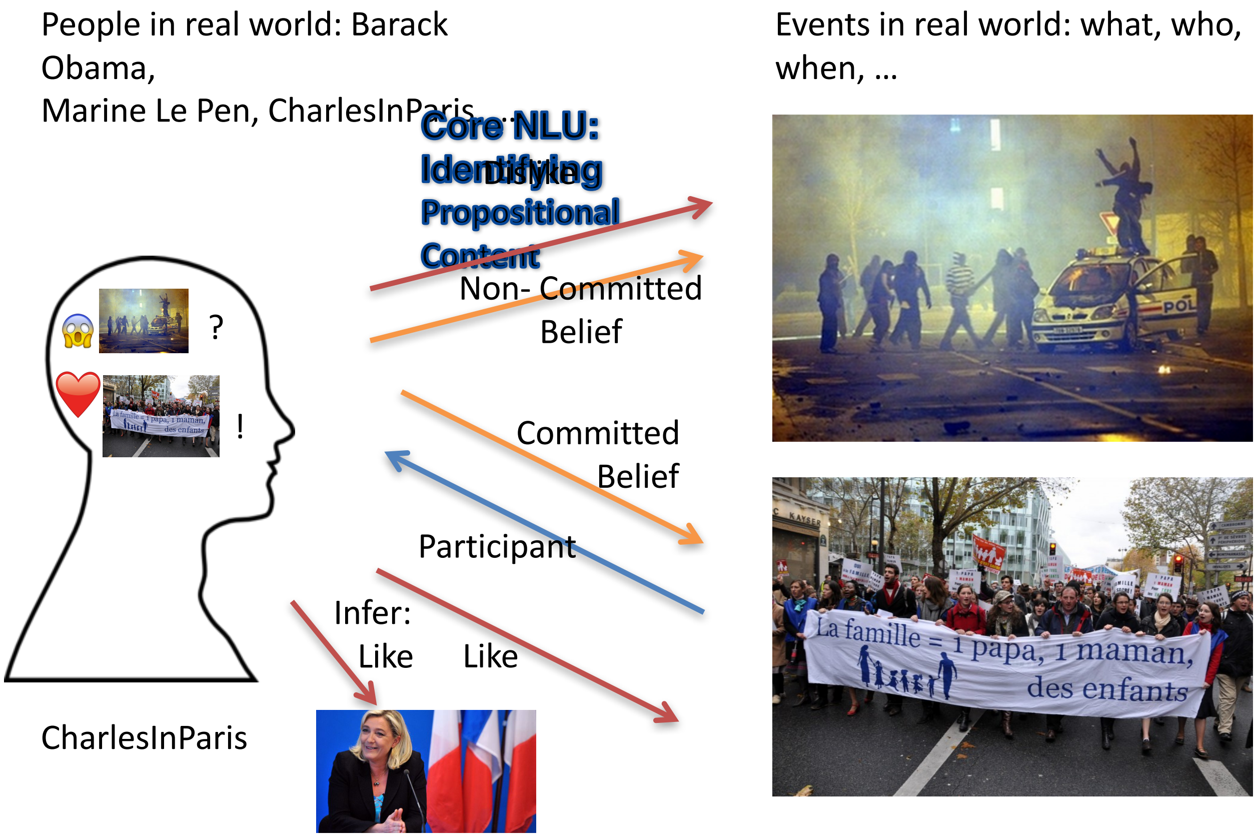
Jennifer Tracey
Linguistic Data Consortium
Philadelphia, PA, USA

BeSt: Evaluating Mind Reading

There was a demonstration against gay marriage in Paris yesterday.

CharlesInParis: I was at the pro-family demo yesterday!

It was such a good show of support for French values!
So different from the alleged riots…



People in real world: Barack Obama,
Marine Le Pen, CharlesInParis, …

Events in real world: what, who, when, …

Core NLU: Identifying Propositional Content

Dislike

Non- Committed Belief

Committed Belief

Participant

Infer: Like

Like

😱 ?

❤️ !

CharlesInParis

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)
  - Sources: Entities,
  - Targets: Entites (sentiment only), relations and events (EREs)
  - 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

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

Training Data: Same as for BeSt Eval 2016

English	All data	Discussion Forums (%)	Newswire (%)
Train	157K words	89%	11%
Evaluation 2016	88K words	52%	48%
Evaluation 2017	95K words	65%	35%

Spanish	All data	Discussion Forums (%)	Newswire (%)
Train	79K words	100%	0%
Evaluation 2016	67K words	61%	39%
Evaluation 2017	89K words	66%	34%

Chinese	All data	Discussion Forums (%)	Newswire (%)
Train	133K words	100%	0%
Evaluation 2016	122K words	65%	35%

English Training 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	13.7%	9.3%	26.6%

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

Participation

- Participation increased over 2016, but still low
  - Hard and new problem
  - Decided not to advertise

BeSt Eval Participants
Belief

	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												
Albany	---	---	---	---	---	---	---	---	---	---	---	---
Chinese Ac. Sci.	X	X	X	X	X	X	X	X	X	X	X	X
Columbia/GWU	X	X	X	X	X	X	X	X	X	X	X	X
Cornell/Mich	---	---	---	---	---	---	---	---	X	X	X	X
Jaén Sinai	X	X	---	---	---	---	---	---	---	---	---	---
IBM Dublin	---	---	---	---	---	---	---	---	---	---	---	---
Best Perform.												

BeSt Eval Participants
Belief: 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			X	X								
Albany	---	---	---	---	---	---	---	---	---	---	---	---
Chinese Ac. Sci.	X	X	X	X	X	X	X	X	X	X	X	X
Columbia/GWU	X	X	X	X	X	X	X	X	X	X	X	X
Cornell/Mich	---	---	---	---	---	---	---	---	X	X	X	X
Jaén Sinai	X	X	---	---	---	---	---	---	---	---	---	---
IBM Dublin	---	---	---	---	---	---	---	---	---	---	---	---
Best Perform.	78	63	1	1	78	68	0	1	82	64	0	0

BeSt Eval Participants
Sentiment

	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												
Albany	X	X	X	X	---	---	---	---	---	---	---	---
Chinese Ac. Sci.	X	X	X	X	X	X	X	X	X	X	X	X
Columbia/GWU	X	X	X	X	X	X	X	X	X	X	X	X
Cornell/Mich	---	---	---	---	---	---	---	---	X	X	X	X
Jaén Sinai	---	---	---	---	---	---	---	---	---	---	---	---
IBM Dublin	X	X	X	X	---	---	---	---	---	---	---	---
Best perform.												

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		X		X			X	X			X	
Albany	X	X	X	X	---	---	---	---	---	---	---	---
Chinese Ac. Sci.	X	X	X	X	X	X	X	X	X	X	X	X
Columbia/GWU	X	X	X	X	X	X	X	X	X	X	X	X
Cornell/Mich	---	---	---	---	---	---	---	---	X	X	X	X
Jaén Sinai	---	---	---	---	---	---	---	---	---	---	---	---
IBM Dublin	X	X	X	X	---	---	---	---	---	---	---	---
Best perform.	25	12	8	3	21	11	7	3	28	14	4	2

Observations

- Predicted ERE is hard
  - Same in 2016
- Will continue to emerge as a central topic of research in NLP as we move towards deep understanding of language
  - Cognitive science
  - Pragmatics
  - Discourse & dialog

Many Thanks

- NIST (Hoa Dang) for organizing the evaluation
- LDC (Jennifer Tracey and Michael Arrigo) for annotations
  - The corpora will continue to be used
- DARPA (Boyan Onyshkevych) for funding the research
- All teams that participated in planning the evaluation
- All teams that participated in the evaluation