Overview of the TAC2013 Knowledge Base Population Evaluation: **English Slot Filling**

Mihai Surdeanu

with a lot of help from: Hoa Dang, Joe Ellis, Heng Ji, and Ralph Grishman

Introduction

- **Slot filling (SF)**: extract values of specified attributes for a given entity from a large collection of natural language texts.

- This was the 5th year for the KBP SF evaluation
- A few new things this year

New: Annotation Guidelines

- per:title
  - Titles at different organizations are different
  - Mitt Romney
    - CEO at **Bain Capital**
    - CEO at **Bain & Company**
    - CEO at **2002 Winter Olympics**

different fillers!

- per:employee_of + per:member_of = per:employee_or_member_of

- Entities in meta data can be used as query input or output
  - Consider post authors as filler candidates

New: Provenance and Justification

- Exact provenance and justification required
  - Up to two mentions for slot/filler provenance
  - Up to two sentences for justification

New: Provenance and Justification

query
entity: Michelle Obama
slot: per:spouse

Michelle Obama started her career as a corporate lawyer specializing in marketing and intellectual property. Michelle met Barack Obama when she was employed as a corporate attorney with the law firm Sidley Austin. She married him in 1992.

output
Entity provenance: "She", "Michelle Obama"
Filler provenance: "him", "Barack Obama"
Justification: "She married him in 1992."

New: Source Corpus

- One million documents from Gigaword
- One million web documents (similar to 2012)
- ~100,000 documents from web discussion fora

- Released as a single corpus for convenience

Scoring Metric

- Each non-NIL response is assessed as: **C**orrect, ine**X**act, **R**edundant, or **W**rong
  - Justification contains >2 sentences ➔ W
  - Provenance and/or justification incomplete ➔ W
  - Filler string incomplete or include extraneous material ➔ X
  - Text spans justify the extraction and filler is exact
    - Filler exists in the KB ➔ R
    - Filler does not exist in KB ➔ C
- Credit given for C and R

Scoring Metric

- Precision, recall, and F1 score computed considering C and R fillers as correct

- Recall is tricky
  - Gold keys constructed from
    - System outputs judged as correct
    - A manual key prepared by LDC annotators independently

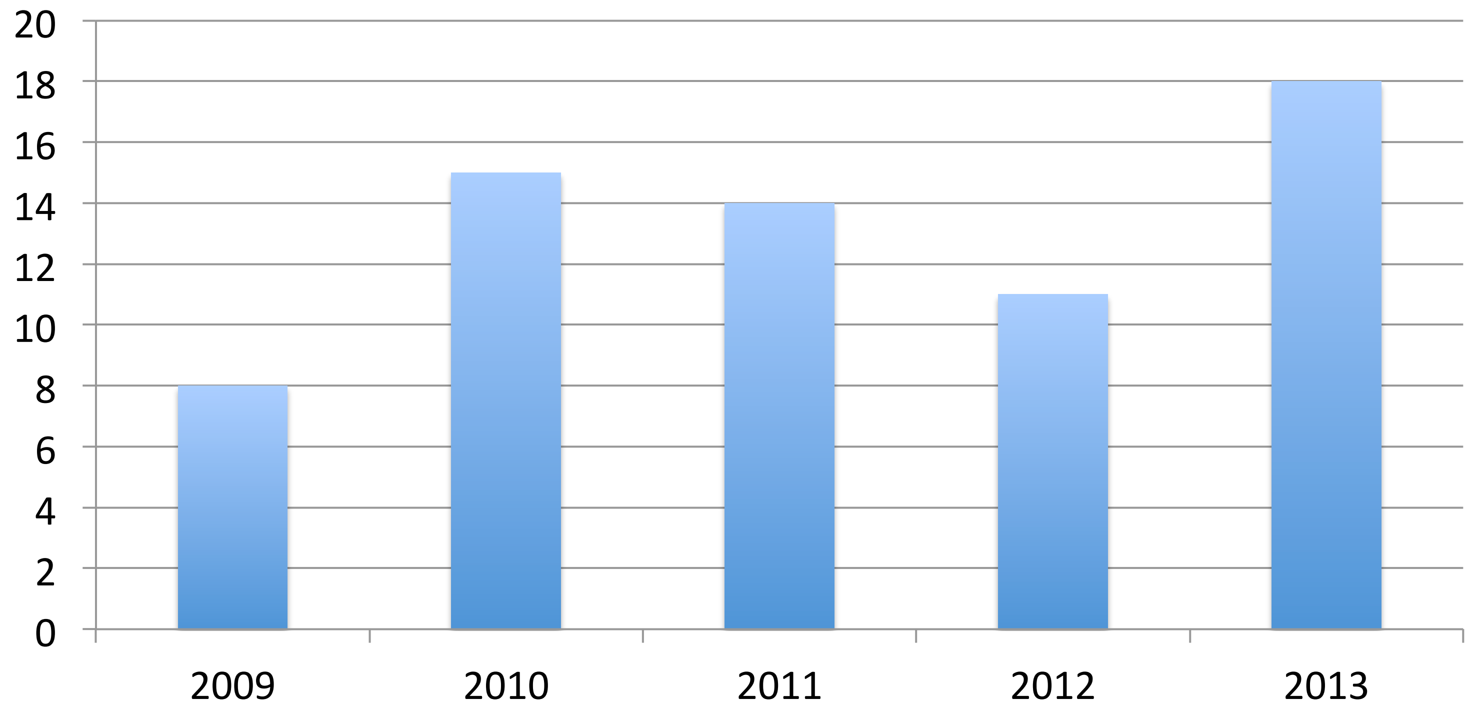
Also serves as a fair performance ceiling

PARTICIPANTS

Participants

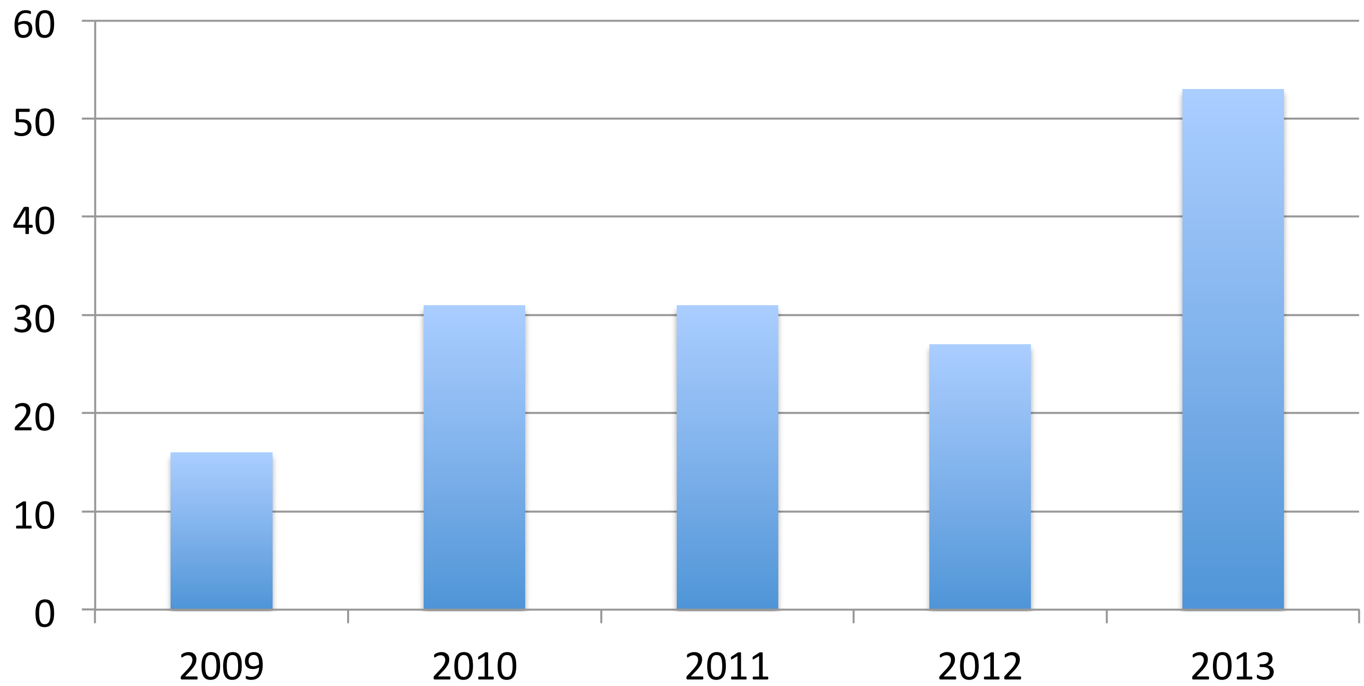
Team Id	Organization(s)	SF?	TSF?
ARPANI	Bhilai Institute of Technology	√	
CMUML	Carnegie Mellon University	√	√
PRIS2013	Beijing University of Posts and Telecommunications	√	
TALP_UPC	TALP Research Center of Technical University of Catalonia (UPC)	√	
UWashington	Department of Computer Science and Engineering, University of Washington	√	
utaustin	University of Texas at Austin – AI Lab	√	
SINDI	Korea Institute of Science and Technology Information	√	
CohenCMU	Carnegie Mellon University	√	
UMass_IESL	University of Massachusetts Amherst, Information Extraction and Synthesis Lab	√	
BIT	Beijing Institute of Technology	√	
SAFT_KRes	University of Southern California Information Sciences Institute	√	
UNED	Universidad Nacional de Educación a Distancia	√	√
IIRG	University College Dublin	√	
NYU	New York University	√	
Stanford	Stanford University	√	
lsv	Saarland University	√	
Compreno	ABBYY	√	√
RPI-BLENDER	Rensselaer Polytechnic Institute	√	√
MS_MLI	Microsoft Research		√

Participation Trends



Number of teams who submitted at least one SF run

Participation Trends



Number of SF submissions

RESULTS

The Task Was Harder This Year

- Harder
  - Stricter scoring
  - More complex queries, with a more uniform slot distribution
- Easier
  - Extracting redundant fillers is somewhat easier

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	47.83	**36.18**	27.45	50.38	35.54
RPI-BLENDER	30.62	38.19	33.98	29.02	40.73	33.89
PRIS2013	27.82	35.33	31.13	27.59	38.87	32.27
BIT	22.06	57.86	31.94	21.73	61.35	32.09
Stanford	28.46	32.30	30.26	28.41	35.86	31.70
NYU	17.35	50.70	25.85	16.76	53.83	25.56
UWashington	10.31	**59.72**	17.59	10.29	**63.45**	17.70
CMUML	10.63	28.79	15.53	10.69	32.30	16.07
SAFT_KRes	13.43	12.43	12.91	14.99	15.67	15.32
UMass_IESL	18.47	9.43	12.48	18.46	10.88	13.69
utaustin	7.91	21.85	11.62	8.11	25.16	12.26
UNED	9.11	15.08	11.36	9.33	17.59	12.19
Compreno	13.19	8.69	10.48	12.74	9.74	11.04
TALP_UPC	9.67	6.54	7.81	9.81	7.69	8.62
IIRG	3.20	7.38	4.46	2.86	7.72	4.17
SINDI	2.80	7.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	Precision	F1	Recall	Precision	F1
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPA	[illegible]	.83	**36.18**	27.45	50.38	35.54
RPI-I	[illegible]	.19	33.98	29.02	40.73	33.89
PRIS	[illegible]	.33	31.13	27.59	38.87	32.27
BIT	[illegible]	.86	31.94	21.73	61.35	32.09
Stanf	[illegible]	.30	30.26	28.41	35.86	31.70
NYU	[illegible]	.70	25.85	16.76	53.83	25.56
UWa	[illegible]	**.72**	17.59	10.29	**63.45**	17.70
CMU	[illegible]	.79	15.53	10.69	32.30	16.07
SAFT	[illegible]	.43	12.91	14.99	15.67	15.32
UMa	[illegible]	[illegible]	12.48	18.46	10.88	13.69
utaus	[illegible]	[illegible]	11.62	8.11	25.16	12.26
UNE	[illegible]	[illegible]	11.36	9.33	17.59	12.19
Comp	[illegible]	[illegible]	10.48	12.74	9.74	11.04
TAL	[illegible]	54	7.81	9.81	7.69	8.62
IIRG	[illegible]	38	4.46	2.86	7.72	4.17
SIND	[illegible]	.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Official scores are generally higher than diagnostic scores

Redundant fillers are easier to extract. That's why they are already in the KB?

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	47.83	**36.18**	27.45	50.38	35.54
RPI-BLENDER	30.62	38.19	33.98	29.02	40.73	33.89
PRIS2013	27.82	35.33	31.13	27.59	38.87	32.27
BIT	22.06	57.86	31.94	21.73	61.35	32.09
Stanford	28.46	32.30	30.26	28.41	35.86	31.70
NYU	17.35	50.70	25.85	16.76	53.83	25.56
UWashington	10.31	**59.72**	17.59	10.29	**63.45**	17.70
CMUML	10.63	28.79	15.53	10.69	32.30	16.07
SAFT_KRes	13.43	12.43	12.91	14.99	[illegible]	[illegible]
UMass_IESL	18.47	9.43	12.48	[illegible]	[illegible]	[illegible]
utaustin	7.91	21.85	11.62	8.11	[illegible]	[illegible]
UNED	9.11	15.08	11.36	9.33	[illegible]	[illegible]
Compreno	13.19	8.69	10.48	[illegible]	[illegible]	[illegible]
TALP_UPC	9.67	6.54	7.81	9.81	[illegible]	[illegible]
IIRG	3.20	7.38	4.46	2.86	[illegible]	[illegible]
SINDI	2.80	7.26	4.04	2.59	7.84	[illegible]
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	$\boxed{68.49}$

Harder task:
this score was
81.4 in 2012.

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	[illegible]	[illegible]	[illegible]	[illegible]	35.54
RPI-BLENDER	30.62	[illegible]	[illegible]	[illegible]	[illegible]	33.89
PRIS2013	27.82	[illegible]	[illegible]	[illegible]	[illegible]	32.27
BIT	22.06	[illegible]	[illegible]	[illegible]	[illegible]	32.09
Stanford	28.46	[illegible]	[illegible]	[illegible]	[illegible]	31.70
NYU	17.35	[illegible]	[illegible]	[illegible]	[illegible]	25.56
UWashington	10.31	[illegible]	[illegible]	[illegible]	[illegible]	17.70
CMUML	10.63	[illegible]	[illegible]	[illegible]	[illegible]	16.07
SAFT_KRes	13.43	12.43	12.91	14.99	15.67	15.32
UMass_IESL	18.47	9.43	12.48	18.46	10.88	13.69
utaustin	7.91	21.85	11.62	8.11	25.16	12.26
UNED	9.11	15.08	11.36	9.33	17.59	12.19
Compreno	13.19	8.69	10.48	12.74	9.74	11.04
TALP_UPC	9.67	6.54	7.81	9.81	7.69	8.62
IIRG	3.20	7.38	4.46	2.86	7.72	4.17
SINDI	2.80	7.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Increased performance:
6 systems over 30 F1. Last year, there were only 2.

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	47.83	**36.18**	27.45	50.38	35.54
RPI-BLENDER	30.62	38.19	33.98	29.02	40.73	33.89
PRIS2013	27.82	35.33	31.13	27.59	38.87	32.27
BIT	22.06	57.86	31.94	21.73	61.35	32.09
Stanford	28.46	32.30	30.26	28.41	35.86	31.70
NYU	17.35	[illegible]	[illegible]	[illegible]	[illegible]	25.56
UWashington	10.31	[illegible]	[illegible]	[illegible]	[illegible]	17.70
CMUML	10.63	[illegible]	[illegible]	[illegible]	[illegible]	16.07
SAFT_KRes	13.43	[illegible]	[illegible]	[illegible]	[illegible]	15.32
UMass_IESL	18.47	[illegible]	[illegible]	[illegible]	[illegible]	13.69
utaustin	7.91	[illegible]	[illegible]	[illegible]	[illegible]	12.26
UNED	9.11	[illegible]	[illegible]	[illegible]	[illegible]	12.19
Compreno	13.19	[illegible]	[illegible]	[illegible]	[illegible]	11.04
TALP_UPC	9.67	[illegible]	[illegible]	[illegible]	[illegible]	8.62
IIRG	3.20	7.38	4.46	2.86	7.72	4.17
SINDI	2.80	7.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Increased performance:
Median: 15.7 F1.
Last year: 9.9 F1.

Overall Results

	Diagnostic Scores			Official Scores		
	Recall	Precision	F1	Recall	Precision	F1
lsv	**32.93**	38.50	35.50	**33.17**	42.53	**37.28**
ARPANI*	29.10	47.83	**36.18**	27.45	50.38	35.54
RPI-BLENDER	30.62	38.19	33.98	29.02	[illegible]	[illegible]
PRIS2013	27.82	35.33	31.13	27.59	[illegible]	[illegible]
BIT	22.06	57.86	31.94	21.73	[illegible]	[illegible]
Stanford	28.46	32.30	30.26	28.41	[illegible]	[illegible]
NYU	17.35	50.70	25.85	16.76	[illegible]	[illegible]
UWashington	10.31	**59.72**	17.59	10.29	[illegible]	[illegible]
CMUML	10.63	28.79	15.53	10.69	[illegible]	[illegible]
SAFT_KRes	13.43	12.43	12.91	14.99	15.67	15.32
UMass_IESL	18.47	9.43	12.48	18.46	10.88	13.69
utaustin	7.91	21.85	11.62	8.11	25.16	12.26
UNED	9.11	15.08	11.36	9.33	17.59	12.19
Compreno	13.19	8.69	10.48	12.74	9.74	11.04
TALP_UPC	9.67	6.54	7.81	9.81	7.69	8.62
IIRG	3.20	7.38	4.46	2.86	7.72	4.17
SINDI	2.80	7.26	4.04	2.59	7.84	3.89
CohenCMU	3.68	1.69	2.32	3.68	1.98	2.57
LDC	58.35	83.81	68.80	57.08	85.60	68.49

Perspective: We are at 54% of human performance

Distribution of Slots in Evaluation Queries

	Entity Count	Value Count (Pct)
per:title	33	142 (10.8%)
org:top_members_employees	41	116 (8.8%)
org:alternate_names	45	82 (6.2%)
per:employee_or_member_of	28	72 (5.5%)
per:children	23	52 (3.9%)
per:cities_of_residence	30	51 (3.9%)
per:age	31	51 (3.9%)
per:date_of_death	36	48 (3.6%)
per:cause_of_death	33	47 (3.5%)
per:charges	13	45 (3.4%)
per:alternate_names	24	45 (3.4%)
per:countries_of_residence	25	36 (2.7%)
per:city_of_death	32	35 (2.6%)
org:country_of_headquarters	34	34 (2.6%)
org:website	32	32 (2.4%)
per:origin	28	32 (2.4%)
per:spouse	23	28 (2.1%)
per:statesorprovinces_of_residence	23	28 (2.1%)
per:schools_attended	16	27 (2.0%)
...		

Distribution of Slots in Evaluation Queries

	Entity Count	Value Count (Pct)
per:title	33	142 (10.8%)
org:top_members_		116 (8.8%)
org:alternate_nan		82 (6.2%)
per:employee_or		72 (5.5%)
per:children		52 (3.9%)
per:cities_of_resi		51 (3.9%)
per:age		51 (3.9%)
per:date_of_death		48 (3.6%)
per:cause_of_dea		47 (3.5%)
per:charges		45 (3.4%)
per:alternate_nan		45 (3.4%)
per:countries_of_		36 (2.7%)
per:city_of_death		35 (2.6%)
org:country_of_head		34 (2.6%)
org:website	32	32 (2.4%)
per:origin	28	32 (2.4%)
per:spouse	23	28 (2.1%)
per:statesorprovinces_of_residence	23	28 (2.1%)
per:schools_attended	16	27 (2.0%)
...		

Harder data:

- 13 slots needed to cover 60% of data
- Some of these are hard.
- In 2011, only 7 slots needed to cover 60% of data.

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	44.46	37.13	+3.24
ARPANI*	27.49	50.36	35.57	28.72	52.38	37.10	+1.56
Stanford	29.20	36.80	32.56	32.49	40.76	36.16	+4.46
PRIS2013	28.03	39.44	32.78	29.34	41.07	34.23	+1.86
BIT	21.90	61.73	32.33	22.55	63.27	33.25	+1.16
NYU	16.98	54.49	25.90	18.16	57.99	27.66	+2.10
IIRG	10.50	28.31	15.32	14.39	38.60	20.97	**+16.80**
UWashington	10.44	**64.29**	17.96	11.38	**69.75**	19.56	+1.86
CMUML	10.71	32.30	16.09	11.72	35.19	17.58	+1.51
SAFT_KRes	15.55	16.24	15.89	17.20	17.88	17.53	+2.21
utaustin	8.46	26.22	12.79	10.76	33.19	16.25	+3.99
Compreno	13.48	10.26	11.64	17.82	13.54	15.39	+4.35
UNED	9.69	18.23	12.65	11.65	21.82	15.19	+3.00
UMass_IESL	18.49	10.88	13.70	20.49	12.01	15.14	+1.45
TALP_UPC	10.16	7.96	8.93	13.02	10.15	11.41	+2.79
SINDI	2.66	8.04	4.00	3.43	10.31	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	[illegible]	[illegible]	3.24
ARPANI*	27.49	50.36	35.57	28.72	[illegible]	[illegible]	.56
Stanford	29.20	36.80	32.56	32.49	[illegible]	[illegible]	.46
PRIS2013	28.03	39.44	32.78	29.34	[illegible]	[illegible]	.86
BIT	21.90	61.73	32.33	22.55	[illegible]	[illegible]	.16
NYU	16.98	54.49	25.90	18.16	[illegible]	[illegible]	.10
IIRG	10.50	28.31	15.32	14.39	[illegible]	[illegible]	**6.80**
UWashington	10.44	**64.29**	17.96	11.38	[illegible]	[illegible]	.86
CMUML	10.71	32.30	16.09	11.72	[illegible]	[illegible]	.51
SAFT_KRes	15.55	16.24	15.89	17.20	[illegible]	[illegible]	.21
utaustin	8.46	26.22	12.79	10.76	[illegible]	[illegible]	.99
Compreno	13.48	10.26	11.64	17.82	[illegible]	[illegible]	.35
UNED	9.69	18.23	12.65	11.65	[illegible]	[illegible]	.00
UMass_IESL	18.49	10.88	13.70	20.49	[illegible]	[illegible]	+1.45
TALP_UPC	10.16	7.96	8.93	13.02	10.15	11.4	+2.79
SINDI	2.66	8.04	4.00	3.43	10.31	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	$\boxed{+2.14}$

Not directly comparable with the official scores due to collapsing of per:title

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	[illegible]	[illegible]	[illegible]
ARPANI*	27.49	50.36	35.57	28.72	[illegible]	[illegible]	[illegible]
Stanford	29.20	36.80	32.56	32.49	[illegible]	[illegible]	[illegible]
PRIS2013	28.03	39.44	32.78	29.34	41.07	[illegible]	+1.86
BIT	21.90	61.73	32.33	22.55	63.27	33.25	+1.16
NYU	16.98	54.49	25.90	18.16	57.99	27.66	+2.10
IIRG	10.50	28.31	15.32	14.39	38.60	20.97	**+16.80**
UWashington	10.44	**64.29**	17.96	11.38	**69.75**	19.56	+1.86
CMUML	10.71	32.30	16.09	11.72	35.19	17.58	+1.51
SAFT_KRes	15.55	16.24	15.89	17.20	17.88	17.53	+2.21
utaustin	8.46	26.22	12.79	10.76	33.19	16.25	+3.99
Compreno	13.48	10.26	11.64	17.82	13.54	15.39	+4.35
UNED	9.69	18.23	12.65	11.65	21.82	15.19	+3.00
UMass_IESL	18.49	10.88	13.70	20.49	12.01	15.14	+1.45
TALP_UPC	10.16	7.96	8.93	13.02	10.15	11.41	+2.79
SINDI	2.66	8.04	4.00	3.43	10.31	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

System bug?

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	[illegible]	[illegible]	[illegible]
ARPANI*	27.49	50.36	35.57	28.72	[illegible]	[illegible]	[illegible]
Stanford	29.20	36.80	32.56	[illegible]	[illegible]	[illegible]	[illegible]
PRIS2013	28.03	39.44	32.78	[illegible]	[illegible]	[illegible]	[illegible]
BIT	21.90	61.73	32.33	[illegible]	[illegible]	[illegible]	[illegible]
NYU	16.98	54.49	25.90	[illegible]	[illegible]	[illegible]	[illegible]
IIRG	10.50	28.31	15.32	14.39	[illegible]	[illegible]	[illegible]
UWashington	10.44	**64.29**	17.96	11.38	[illegible]	[illegible]	[illegible]
CMUML	10.71	32.30	16.09	11.72	[illegible]	[illegible]	[illegible]
SAFT_KRes	15.55	16.24	15.89	17.20	[illegible]	[illegible]	[illegible]
utaustin	8.46	26.22	12.79	10.76	[illegible]	[illegible]	[illegible]
Compreno	13.48	10.26	11.64	17.82	[illegible]	[illegible]	[illegible]
UNED	9.69	18.23	12.65	11.65	[illegible]	[illegible]	[illegible]
UMass_IESL	18.49	10.88	13.70	20.49	[illegible]	[illegible]	[illegible]
TALP_UPC	10.16	7.96	8.93	13.02	[illegible]	[illegible]	[illegible]
SINDI	2.66	8.04	4.00	3.43	[illegible]	[illegible]	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

About the same as the official scores.
If systems identify the correct docs, they can extract correct offsets.

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	44.46	37.13	+3.24
ARPANI*	27.49	50.36	35.57	28.72	52.38	37.10	+1.56
Stanford	29.20	36.80	32.56	32.49	40.76	36.16	+4.46
PRIS2013	28.03	39.44	32.78	29.34	41.07	34.23	+1.86
BIT	21.90	61.73	32.33	22.55	63.27	33.25	+1.16
NYU	16.98	54				27.66	+2.10
IIRG	10.50	28				20.97	**+16.80**
UWashington	10.44	**64**				19.56	+1.86
CMUML	10.71	32				17.58	+1.51
SAFT_KRes	15.55	16				17.53	+2.21
utaustin	8.46	26.22	12.79	10.76	33.19	16.25	+3.99
Compreno	13.48	10.26	11.64	17.82	13.54	15.39	+4.35
UNED	9.69	18.23	12.65	11.65	21.82	15.19	+3.00
UMass_IESL	18.49	10.88	13.70	20.49	12.01	15.14	+1.45
TALP_UPC	10.16	7.96	8.93	13.02	10.15	11.41	+2.79
SINDI	2.66	8.04	4.00	3.43	10.31	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

These are much higher for some systems.

Results with Lenient Scoring

	Official Score with `ignoreoffsets`			Official Score with `anydoc`			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	44.46	37.13	+3.24
ARPANI*	27.49	[illegible]	[illegible]	[illegible]	[illegible]	37.10	+1.56
Stanford	29.20	[illegible]	[illegible]	[illegible]	[illegible]	36.16	+4.46
PRIS2013	28.03	[illegible]	[illegible]	[illegible]	[illegible]	34.23	+1.86
BIT	21.90	[illegible]	[illegible]	[illegible]	[illegible]	33.25	+1.16
NYU	16.98	[illegible]	[illegible]	[illegible]	[illegible]	27.66	+2.10
IIRG	10.50	[illegible]	[illegible]	[illegible]	[illegible]	20.97	**+16.80**
UWashington	10.44	[illegible]	[illegible]	[illegible]	[illegible]	19.56	+1.86
CMUML	10.71	32.30	16.09	11.72	35.19	17.58	+1.51
SAFT_KRes	15.55	16.24	15.89	17.20	17.88	17.53	+2.21
utaustin	8.46	26.22	12.79	10.76	33.19	16.25	+3.99
Compreno	13.48	10.26	11.64	17.82	13.54	15.39	+4.35
UNED	9.69	18.23	12.65	11.65	21.82	15.19	+3.00
UMass_IESL	18.49	10.88	13.70	20.49	12.01	15.14	+1.45
TALP_UPC	10.16	7.96	8.93	13.02	10.15	11.41	+2.79
SINDI	2.66	8.04	4.00	3.43	10.31	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

Extracted fillers from documents not in the source corpus.

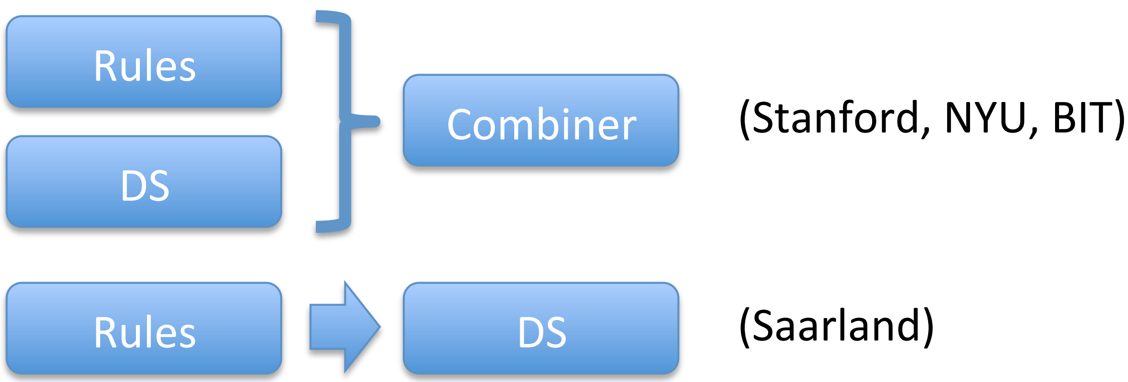
Results with Lenient Scoring

	Official Score with ignoreoffsets			Official Score with anydoc			
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**F1 Increase**
lsv	**33.56**	42.97	**37.69**	**35.84**	45.67	**40.17**	+2.89
RPI-BLENDER	29.13	40.82	34.00	31.87	44.46	37.13	+3.24
ARPANI*	27.49	50.36	35.57	28.72	52.38	37.10	+1.56
Stanford	29.20	36.80	32.56	32.49	40.76	36.16	+4.46
PRIS2013	28.03	39.44	32.78	29.34	41.07	34.23	+1.86
BIT	21.90	61.73	32.33	22.55	63.27	33.25	+1.16
NYU	16.98	54.49	25.90	18.16	57.99	27.66	+2.10
IIRG	10.50	28.31	15.32	14.39	38.60	20.97	**+16.80**
UWashington	10.44	**64.29**	17.96	11.38	**69.75**	19.56	+1.86
CMUML	10.71	[illegible]	[illegible]	[illegible]	35.19	17.58	+1.51
SAFT_KRes	15.55	[illegible]	[illegible]	[illegible]	[illegible]	17.53	+2.21
utaustin	8.46	[illegible]	[illegible]	[illegible]	[illegible]	16.25	+3.99
Compreno	13.48	[illegible]	[illegible]	[illegible]	[illegible]	15.39	+4.35
UNED	9.69	[illegible]	[illegible]	[illegible]	[illegible]	15.19	+3.00
UMass_IESL	18.49	[illegible]	[illegible]	[illegible]	[illegible]	15.14	+1.45
TALP_UPC	10.16	[illegible]	[illegible]	[illegible]	[illegible]	11.41	+2.79
SINDI	2.66	[illegible]	[illegible]	[illegible]	[illegible]	5.14	+1.25
CohenCMU	3.89	2.09	2.72	5.55	2.97	3.87	+1.30
LDC	57.36	85.90	68.79	59.01	87.95	70.63	+2.14

Inferred relations not explicitly stated in text.

Technology

- Most successful approaches combine distant supervision (DS) with rules



- DS models with built-in noise reduction (Stanford)

Technology

- KBP system based on OpenIE (Washington)
  - Extracted tuples (Arg1, Rel, Arg2) from the KBP corpus
  - Manual written rules to map these tuples to KBP relations
- Bootstrapping dependency-based patterns (Beijing University of Posts and Telecommunications)

Technology

- Unsupervised clustering of patterns (UPC)
- Combining observed and unlabeled data through matrix factorization (UMass)
- Inferring new relations from the stated ones using first-order logic rules learned BLP (UT)

Conclusions

- Positive trends
  - Most popular SF evaluation to date
  - Best performance on average
- Things that need more work
  - Still at ~50% of human performance
  - Participant retention rate at lower than 50%
    - Reduce barrier of entry: offer preprocessed data?
      - Sentences containing <entity, filler> in training
      - Sentences containing entity in testing
      - NLP annotations