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

Mihai Surdeanu

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

Introduction

- **Temporal Slot filling (TSF)**: grounds fillers extracted by SF by finding the start and end dates when they were valid.

- This was the 2nd year for a KBP TSF evaluation
  - There was a pilot evaluation in 2011
- A few new things this year

~ New: Seven Slots Considered

- per:spouse
- per:title
- per:employee_or_member_of
- per:cities_of_residence
- per:statesorprovinces_of_residence
- per:countries_of_residence
- org:top_employees/members

New: Input Queries

```
Column 1: TEMP70711
Column 2: per:spouse
Column 3: Barack Obama
Column 4: AFP_ENG_20081208.0592.LDC2009T13
Column 5: Michelle Obama
Column 6: XXX-YYY
Column 7: ZZZ-WWW
Column 8: SSS-TTT
Column 9: 1.0
Column 10: E0566375
Column 11: E0082980
```

New: Input Queries

```
Column 1: TEMP70711
Column 2: per:spouse
Column 3: Barack Obama
Column 4: AFP_ENG_20081208.0592.LDC2009T13
Column 5: Michelle Obama
Column 6: XXX-YYY
Column 7: ZZZ-WWW
Column 8: SSS-TTT
Column 9: 1.0
Column 10: E0566375
Column 11: E0082980
```

Both entity and filler given!

New: Input Queries

```
Column 1: TEMP70711
Column 2: per:
Column 3: Bara
Column 4: AFP_ENG_ 208.0392.LDC2009T13
Column 5: Michelle Obama
Column 6: XXX-YYY
Column 7: ZZZ-WWW
Column 8: SSS-TTT
Column 9: 1.0
Column 10: E0566375
Column 11: E0082980
```

Provenances and justification given!

New: Provenance of Dates

```
<DOC>
<DOCID> AFP_ENG_20081231.0121.LDC2009T13 </DOCID>
<DOCTYPE SOURCE="newswire"> NEWS STORY </DOCTYPE>
<DATETIME> 2008-12-31 </DATETIME>
<BODY>
<HEADLINE>
Thousands protest in Brussels against
Israeli action in Gaza
</HEADLINE>
<TEXT>
<P>
Thousands took the streets in Brussels on Wednesday
calling for an end to Israeli bombing of the
Palestinian Gaza Strip ...
</DOC>
```

New: Provenance of Dates

```
<DOC>
<DOCID> AFP_ENG_20081231.0121.LDC2009T13 </DOCID>
<DOCTYPE SOURCE="newswire"> NEWS STORY </DOCTYPE>
<DATETIME> 2008-12-31 </DATETIME>
<BODY>
<HEADLINE>
Thousands prot        Brussels against
Israeli
</HEADLI
<TEXT>
<P>
Thousands took the streets in Brussels on Wednesday
calling for an end to Israeli bombing of the
Palestinian Gaza Strip ...
</DOC>
```

Provenance of date mentions used for normalization must be reported!

Scoring Metric

- Same four-tuple used to represent dates: [T1 T2 T3 T4]
  - Relation is true for period beginning between T1 and T2 and ending between T3 and T4
- Has limitations
  - Recurring events

Scoring Metric

- For each query:
  - System output S = <t1, t2, t3, t4>
  - Gold tuple S_g = <g1, g2, g3, g4>
  - Individual query score: $Q(S) = \frac{1}{4} \sum_i \frac{1}{1 + |t_i - g_i|}$
- Overall: $Accuracy = \frac{\sum_{S^i \in S} Q(S^i)}{N}$

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 Summary

	Teams	Submissions
2011	4	7
2013	5	16

RESULTS

Data

- 273 queries
- Only 201 were actually scored
  - 5 dropped because neither LDC nor systems found correct fillers
  - 67 dropped because gold annotations had an invalid temporal interval
    - Valid interval: $T1 \leq T2$, $T3 \leq T4$, and $T1 \leq T4$

Scoring and Baseline

- Justification ignored (for now) in scoring
- DCT-WITHIN baseline of Ji et al. (2011)
  - Assumption: the relation is valid at the doc date
  - Tuple: <-∞, doc date, doc date, +∞>

Results

	org:top_members_employees	per:cities_of_residence	per:countries_of_residence	per:employee_or_member_of	per:spouse	per:stateorprovinces_of_residence	per:title	**All**
Baseline	24.70	17.40	15.18	17.83	14.75	21.08	23.20	19.10
MS_MLI	**31.94**	**36.06**	**32.85**	**40.12**	**33.04**	**31.85**	**27.35**	**33.15**
RPI-BLENDER	31.19	13.07	14.93	26.71	29.04	17.24	34.68	23.42
UNED	26.20	6.88	8.16	15.24	14.47	14.41	19.34	14.79
CMUML	19.95	7.46	8.47	16.52	13.43	5.65	11.95	11.53
Compreno	0.0	2.42	8.56	0.0	13.50	7.91	0.0	5.14
LDC	69.87	60.22	58.26	72.27	81.10	54.07	91.18	68.84

Results

	org:top_members_employees	per:cities_of_residence	per:countries_of_residence	per:employee_or_member_of	per:spouse	per:stateorprovinces_of_residence	per:title	All
Baseline	24.70	17.40	15.18	[illegible]	[illegible]	[illegible]	[illegible]	19.10
MS_MLI	**31.94**	**36.06**	**32.85**	[illegible]	[illegible]	[illegible]	[illegible]	**33.15**
RPI-BLENDER	31.19	13.07	14.93	[illegible]	[illegible]	[illegible]	[illegible]	23.42
UNED	26.20	6.88	8.16	[illegible]	[illegible]	[illegible]	[illegible]	14.79
CMUML	19.95	7.46	8.47	[illegible]	[illegible]	[illegible]	[illegible]	11.53
Compreno	0.0	2.42	8.56	[illegible]	[illegible]	[illegible]	[illegible]	5.14
LDC	69.87	60.22	58.26	[illegible]	[illegible]	[illegible]	[illegible]	68.84

- 2/5 systems outperformed the baseline
- 3/4 did in 2011

Results

	org:top_members_employees	per:cities_of_residence	per:countries_of_residence	per:employee_or_member_of	per:spouse	per:stateorprovinces_of_residence	per:title	**All**
Baseline	24.70	17.40	15.18	17.83	[illegible]	[illegible]	[illegible]	19.10
MS_MLI	**31.94**	**36.06**	**32.85**	**40.12**	[illegible]	[illegible]	[illegible]	**33.15**
RPI-BLENDER	31.19	13.07	14.93	26.71	[illegible]	[illegible]	[illegible]	23.42
UNED	26.20	6.88	8.16	15.24	[illegible]	[illegible]	[illegible]	14.79
CMUML	19.95	7.46	8.47	16.52	[illegible]	[illegible]	[illegible]	11.53
Compreno	0.0	2.42	8.56	0.0	[illegible]	[illegible]	[illegible]	5.14
LDC	69.87	60.22	58.26	72.27	[illegible]	[illegible]	[illegible]	68.84

Perspective:
Top system is at
48% of human
performance

Results

	org:top_members_employees	per:cities_of_residence	per:countries_of_residence	per:employee_or_member_of	per:spouse	per:stateorprovinces_of_residence	per:title	**All**
Baseline	24.70	17.40	15.18	17.83	14.75	21.08	23.20	19.10
MS_MLI	**31.94**	**36.06**	**32.85**	**40.12**	**33.04**	**31.85**	**27.35**	**33.15**
RPI-BLENDER	31.19	13.07	14.93	26.71	29.04	17.24	34.68	23.42
UNED	26.20	6.88	8.16	15.24	14.47	14.41	19.34	14.79
CMUML	19.95	7.46	8.47	16.52	13.43	5.65	11.95	11.53
Compreno	0.0	2.42	8.56	0.0	13.50	7.91	0.0	5.14
LDC	69.87	60.22	58.26	72.27	81.10	54.07	91.18	68.84

Locations of residence tend to perform worse than average

Results

	org:top_members_employees	per:cities_of_residence	per:countries_of_residence	per:employee_or_member_of	per:spouse	per:stateorprovinces_of_residence	per:title	**All**
Baseline	24.70	17.40	15.18	17.83	14.75	21.08	23.20	19.10
MS_MLI	**31.94**	**36.06**	**32.85**	**40.12**	**33.04**	**31.85**	**27.35**	**33.15**
RPI-BLENDER	31.19	13.07	14.93	26.71	29.04	17.24	34.68	23.42
UNED	26.20	6.88	8.16	15.24	14.47	14.41	19.34	14.79
CMUML	19.95	7.46	8.47	16.52	13.43	5.65	11.95	11.53
Compreno	0.0	2.42	8.56	0.0	13.50	7.91	0.0	5.14
LDC	69.87	60.22	58.26	72.27	81.10	54.07	91.18	68.84

Employment relations tend to perform better than average

Results:

Identification of Filler Mentions: A Hidden Problem!

	TSF Accuracy	SF F1	SF Precision	SF Recall
LDC	68.8	83.1	97.3	72.5
MS_MLI	33.1	77.3	96.8	64.4
RPI-BLENDER	23.4	51.8	69.2	41.4
UNED	14.8	46.6	69.9	35.0
CMUML	11.5	32.2	38.5	27.6
Compreno	5.1	18.5	53.6	11.2

Slide by Hoa Dang

Results:
Identification of Slots: A Hidden Problem

	TSF Accuracy	SF F1	SF Precision	SF Recall
LDC	68.8	83.1	97.3	72.5
MS_MLI	33.1	77.3	96.8	64.4
RPI-BLENDER	23.4	51.8	69.2	41.4
UNED	14.8	46.6	69.9	35.0
CMUML	11.5	32.2	38.5	27.6
Compreno	5.1	18.5	53.6	11.2

Identification of correct mentions of fillers is a challenge!

Slide by Hoa Dang

Technology

- Most groups used distant supervision (DS) to assign labels to <entity, filler, date> tuples
  - Training data:
    - Freebase (structured) – RPI, UNED
    - Wikipedia infoboxes (semi-structured) – Microsoft
  - Labels: Start, End, In, Start-And-End
- Ensemble models for DS (RPI)
  - Explicit features + tree kernels

Technology

- Language model to clean up DS noise (Microsoft)
  - Learns that *n*-grams such as "FILLER and ENTITY were married" are indicative of per:spouse
  - These *n*-grams then used in a boosted decision tree classifier, which identifies noisy tuples

Conclusions

- Slight increase in participation
- On average, performance worse than in 2011
  - 2/5 systems outperformed the baseline vs. 3/4
  - New and complex task!
- Notable contributions
  - Noise reduction for TSF
  - Ensemble models for TSF