

# Cold Start 2016

Hoa Dang    Shahzad Rajput

National Institute of Standards and Technology

TAC 2016

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 KB Entity Discovery Evaluation
  - Participants
  - Results
- 3 SF/KB Evaluation
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 KB Entity Discovery Evaluation
  - Participants
  - Results
- 3 SF/KB Evaluation
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion

- Knowledge Base Construction - Queries not provided
  - ED Evaluation
  - SF Evaluation
- Slot Filling - Queries provided
- Slot Filler Validation

# Task Variants

## Queries

### LDC Query

```
<query id='CS16_9999'>
  <mentions>
    <mention>
      <name>June McCarthy</name>
      <docid>ENG_142</docid>
      <beg>16931</beg>
      <end>16943</end>
    </mention>
    <mention>
      <name>Junio McCarthy</name>
      <docid>SPA_142</docid>
      <beg>2863</beg>
      <end>2869</end>
    </mention>
  </mentions>
  <enttype>per</enttype>
  <nodeid>per_049</nodeid>
  <slot0>per:children</slot0>
  <slot1>per:age</slot1>
</query>
```

### SF Query

```
<query id='CSSF16_ENG_abcabdefde'>
  <name>June McCarthy</name>
  <docid>ENG_142</docid>
  <beg>16931</beg>
  <end>16943</end>
  <enttype>PER</enttype>
  <slot>per:children</slot>
  <slot0>per:children</slot0>
  <slot1>per:age</slot1>
</query>
<query id='CSSF16_SPA_defdeabcab'>
  <name>Junio McCarthy</name>
  <docid>SPA_142</docid>
  <beg>2863</beg>
  <end>2869</end>
  <enttype>PER</enttype>
  <slot>per:children</slot>
  <slot0>per:children</slot0>
  <slot1>per:age</slot1>
</query>
```

# Task Variants

## Examples

### Knowledge Base

```
:e4 type PER
:e4 mention ‘‘Bart Simpson’’ Doc726:37-48
:e4 nominal_mention ‘‘brother’’ Doc726:15-21

:e4 per:siblings :e7 Doc124:283-288,Doc885:173-179,Doc885:274-281
:e4 per:age ‘‘10’’ Doc124:180-181,Doc885:173-179 0.9
```

---

### Slot Filling

```
Q4 org:city_of_headquarters myrun1 Doc42:3-8,Doc8:3-11 Baltimore GPE Doc8:3-11 1.0
Q5 per:siblings myrun1 Doc124:283-288,Doc885:173-179 Lisa PER Doc124:283-286 0.7
Q6 per:age myrun1 Doc124:180-181,Doc885:173-179 10 STRING Doc124:180-181 0.9
```

---

# Changes in 2016

- Tasks were cross-lingual
  - entity mentions, slot fillers and provenance from any document
- Three diagnostic monolingual versions
  - entity mentions, slot fillers and provenance from only the single language.
- KB: PER, ORG, GPE + LOC, FAC — slot inventory was not modified
- SF/KB: Nominal mention

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 **KB Entity Discovery Evaluation**
  - **Participants**
  - **Results**
- 3 SF/KB Evaluation
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion



# KB Entity Detection

## Participants

<b>Teams</b>	ENG	SPA	CMN	XLING	Total
BBN	5	-	-	-	5
ICTCAS	4	-	-	-	4
Stanford	3	-	2	4	9
UMass	5	5	-	5	15
hltcoe	5	4	4	4	17
lilian	3	-	-	-	3
summa	3	-	-	-	3
<b>Total</b>	<b>28</b>	<b>9</b>	<b>6</b>	<b>13</b>	<b>56</b>

# Results

## KB Entity Detection Scores

Lang.	Team	typed_mention_ceaf			mention_ceaf			b_cubed		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
CMN	hltcoe	<b>0.661</b>	<b>0.519</b>	<b>0.582</b>	0.682	0.536	0.600	0.673	0.413	0.512
	Stanford	0.661	0.368	0.473	0.734	0.408	0.525	0.729	0.273	0.397
ENG	BBN	<b>0.764</b>	<b>0.598</b>	<b>0.671</b>	0.785	0.614	0.689	0.779	0.515	0.620
	ICTCAS_OKN	0.749	0.531	0.621	0.782	0.554	0.648	0.854	0.443	0.584
	hltcoe	0.656	0.557	0.603	0.677	0.575	0.622	0.636	0.465	0.537
	lilian	0.666	0.435	0.526	0.718	0.469	0.567	0.803	0.347	0.484
	Stanford	0.600	0.441	0.508	0.647	0.475	0.548	0.632	0.344	0.445
	UMass.IESL	0.752	0.352	0.479	0.787	0.368	0.501	0.845	0.233	0.366
	summa	0.553	0.268	0.361	0.577	0.280	0.377	0.697	0.169	0.272
SPA	hltcoe	<b>0.632</b>	<b>0.383</b>	<b>0.477</b>	0.662	0.401	0.499	0.653	0.289	0.401
	UMass.IESL	0.612	0.261	0.366	0.698	0.297	0.417	0.800	0.176	0.288
XLING	hltcoe	<b>0.595</b>	<b>0.465</b>	<b>0.522</b>	0.610	0.476	0.535	0.635	0.351	0.452
	Stanford	0.607	0.284	0.387	0.663	0.310	0.422	0.667	0.173	0.275
	UMass.IESL	0.671	0.195	0.302	0.714	0.208	0.322	0.824	0.098	0.175

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 KB Entity Discovery Evaluation
  - Participants
  - Results
- 3 SF/KB Evaluation**
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion

# SF/KB Definitions

## Scoring

- Wrong or ineXact is *Spurious*
- Hop 1 filler whose Hop 0 parent filler is Wrong or ineXact, is *Spurious*
- Correct responses are grouped into equivalence classes (EC). At most one response is *Right*; all other *Spurious*
- NAM mention in EC, or NOM mentions and the EC is NOM, then one is *Right*; otherwise, if only NOM mentions in a NAM EC, then one is *Ignored*
- **Reference** = number of single-valued pseudo-slots with a correct response + number of equivalence classes for all list-valued pseudo-slots
- Recall =  $\#Right / Reference$
- Precision =  $\#Right / (\#Right + \#Spurious)$
- F1 =  $2 * Precision * Recall / (Precision + Recall)$
- **Applied only to queries with a known correct answer**

# SF/KB Definitions

## Metrics

Score Variants	Aggregates Reported	
	Micro-average	Macro-average
SF	Yes	Yes
LDC-MAX	Yes	Yes
LDC-MEAN	<b>No</b>	Yes

- SF: consider all entrypoints as a separate query.
- LDC-MAX: Considering the run's best entrypoint per LDC query on the basis of F1 score across both hops.
- LDC-MEAN: Precision, Recall, and F1 for each LDC query is the mean Precision, mean Recall, and mean F1 for all entrypoints for that LDC query.

**Micro-averages** are computed as:

$$Total\_Precision = \frac{Total\_Right}{Total\_Right + Total\_Wrong}$$

$$Total\_Recall = \frac{Total\_Right}{Total\_GT}$$

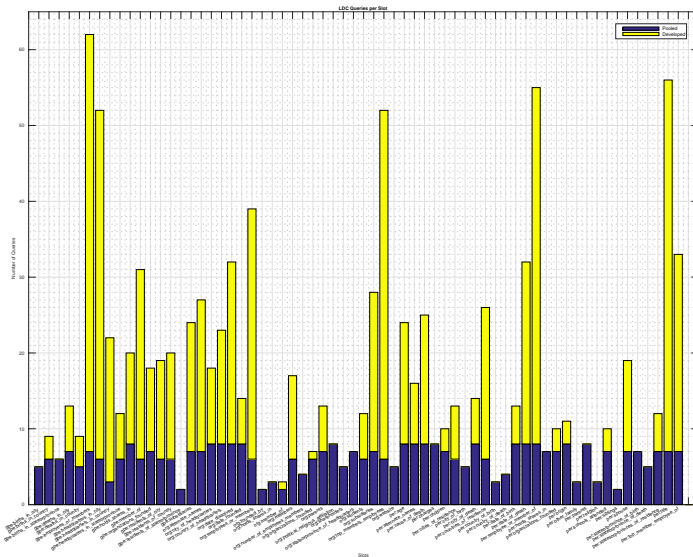
$$Total\_F1 = \frac{2 \times Total\_Precision \times Total\_Recall}{Total\_Precision + Total\_Recall}$$

**Macro-averages** are computed as the mean Precision, mean Recall, and mean F1.

	Language	SF Queries		LDC Queries		
		Developed	Pooled	Developed	Pooled	Nil
ALL	English	1,350	487	1,001	355	123
	Spanish	1,156	402	893	298	101
	Chinese	1,170	371	901	302	100
	<b>Total</b>	<b>3,676</b>	<b>1,260</b>	<b>1,077</b>	<b>392</b>	<b>123</b>
Ambiguous	English	464	187	268	111	35
	Spanish	457	157	252	108	35
	Chinese	343	138	254	105	32
	<b>Total</b>	<b>1,164</b>	<b>482</b>	<b>290</b>	<b>124</b>	<b>35</b>

# Queries

## LDC Queries





# Participants

## SF/KB Teams

Teams	KB runs					SF runs					Tot.
	ENG	SPA	CMN	XLING	Tot.	ENG	SPA	CMN	XLING	Tot.	
BBN	5	-	-	-	5	-	-	-	-	-	5
CMUML	-	-	-	-	-	5	-	-	-	5	5
DCD	-	-	-	-	-	2	-	-	-	2	2
ICTCAS	4	-	-	-	4	-	-	-	-	-	4
IRTSX	-	-	-	-	-	3	-	-	-	3	3
LDC	-	-	-	-	-	1	1	1	1	4	4
MSR	-	-	-	-	-	5	-	-	-	5	5
NAIST	-	-	-	-	-	3	-	-	-	3	3
RPI	-	-	-	-	-	3	1	2	-	6	6
SoochowNLP	-	-	-	-	-	5	-	-	-	5	5
Stanford	3	-	2	4	9	5	-	4	4	13	22
UMass	5	5	-	5	15	5	5	-	5	15	30
UNIST	-	-	-	-	-	4	-	-	-	4	4
doughnutPRIS	-	-	-	-	-	2	-	-	-	2	2
hltcoe	5	4	4	4	17	-	-	-	-	-	17
lilian	3	-	-	-	3	-	-	-	-	-	3
summa	3	-	-	-	3	-	-	-	-	-	3
<b>Total</b>	<b>28</b>	<b>9</b>	<b>6</b>	<b>13</b>	<b>56</b>	<b>43</b>	<b>7</b>	<b>7</b>	<b>10</b>	<b>67</b>	<b>123</b>

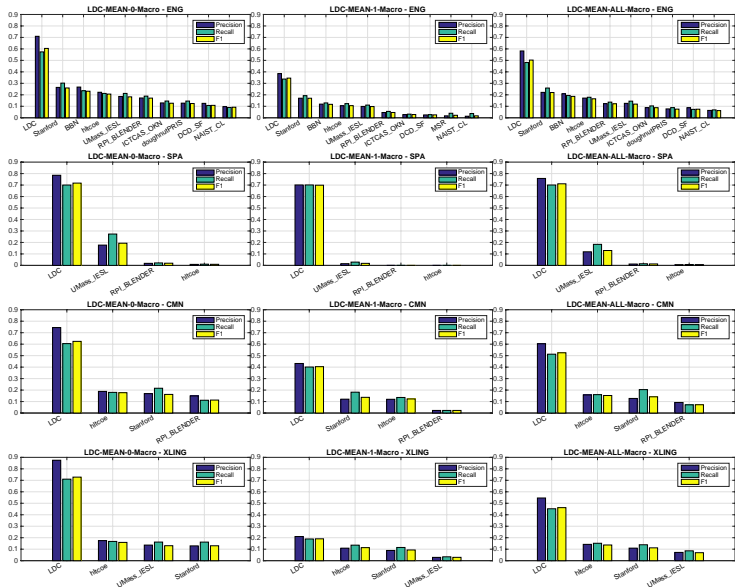
# SF/KB Participants

## Summary of Approaches Used

- Search Engines (Lucene, Bing)
- NLP Tools (NLTK, Stanford's CoreNLP, BBN's SERIF)
- Background knowledge: NELL KB, WordNet, FrameNet, Freebase
- Rule-based clustering, Graphs / Neural networks
- Training Data: Internal manual data, Past SF Data, CoNLL 2012, Angelis dataset
- Validation/Verification: RelationFactory, Social-network based filler verification, Wikipedia anchor text statistics

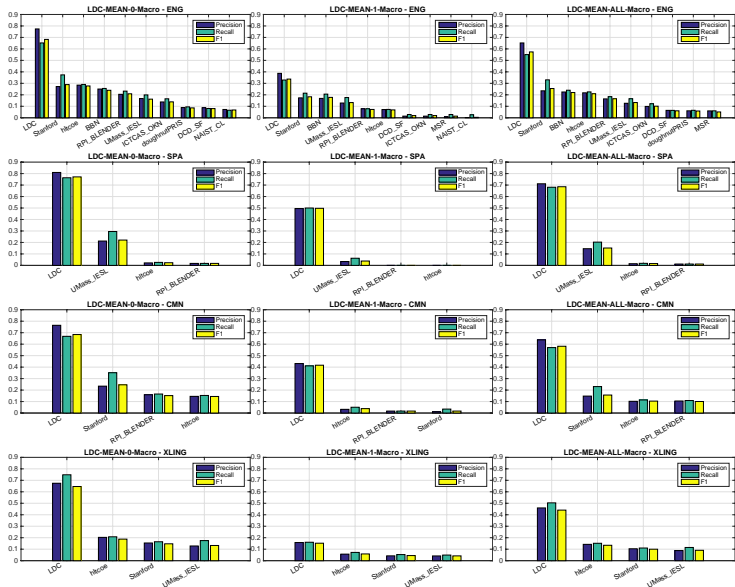
# SF/KB Results

## CS LDC-MEAN-Macro - All Queries



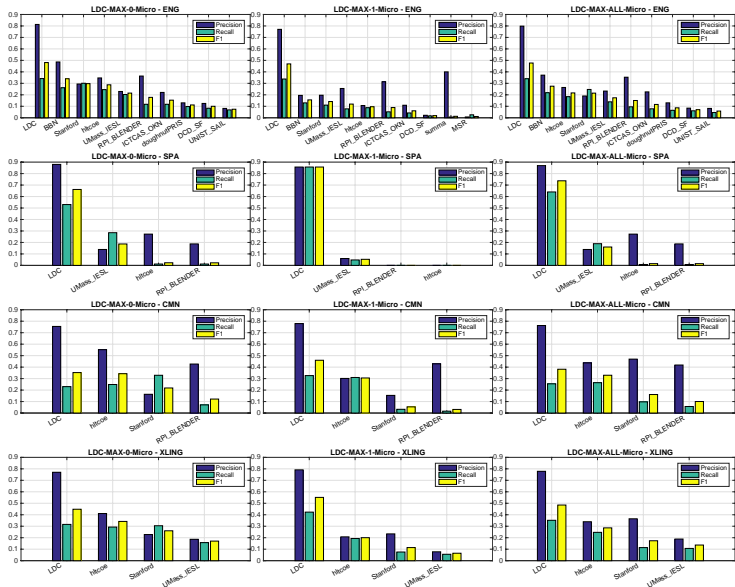
# SF/KB Results

## CS LDC-MEAN-Macro - Ambiguous Queries



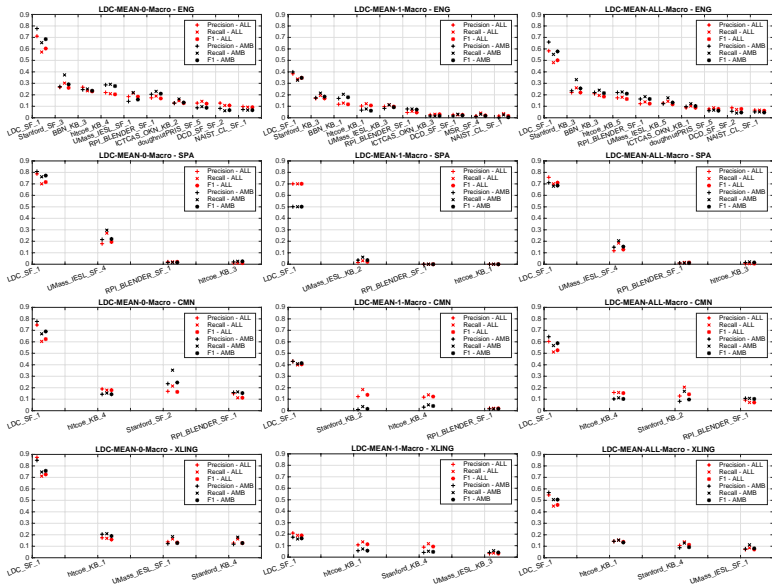
# SF/KB Results

## CS LDC-MAX-Micro - All Queries



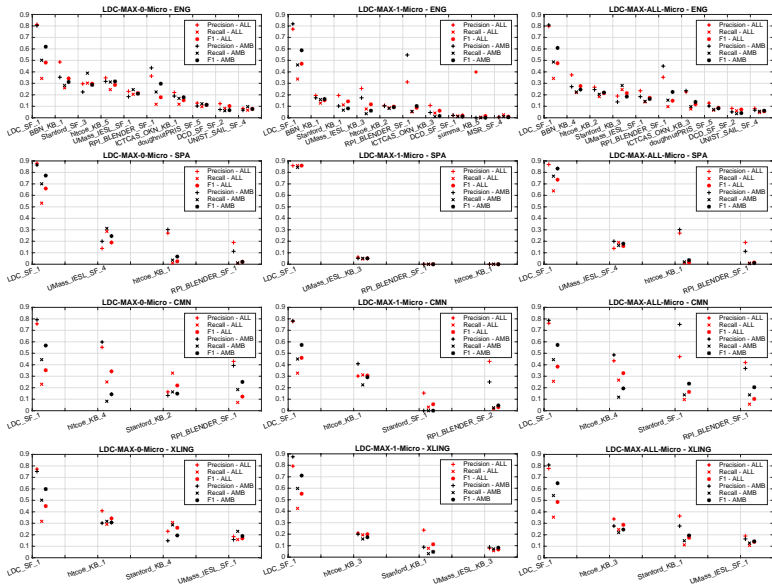
# SF/KB Results

## CS LDC-MEAN-Macro - ALL vs AMBIGUOUS



# SF/KB Results

## CS LDC-MAX-Micro - ALL vs AMBIGUOUS



# SF/KB Results

Confidence Interval - LDC-MEAN-MACRO ENG

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.4358	0.5027	0.5695	0.4381	0.5027	0.5717
2	Stanford	0.1774	0.2201	0.2649	0.1771	0.2201	0.2668
3	BBN	0.1512	0.1864	0.2249	0.1483	0.1864	0.2248
4	hltcoe	0.1285	0.1655	0.2163	0.1235	0.1655	0.2057
5	RPI_BLENDER	0.0924	0.1221	0.1558	0.0908	0.1221	0.1577
6	UMass_IESL	0.0913	0.1192	0.1550	0.0876	0.1192	0.1509
7	ICTCAS_OKN	0.0633	0.0878	0.1168	0.0631	0.0878	0.1145
8	doughnutPRIS	0.0546	0.0754	0.1015	0.0544	0.0754	0.1003
9	DCD_SF	0.0517	0.0747	0.1037	0.0492	0.0747	0.1034
10	NAIST_CL	0.0416	0.0625	0.0947	0.0382	0.0625	0.0906



# SF/KB Results

Confidence Interval - LDC-MEAN-MACRO SPA, CMN, XLING

## Spanish

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.6428	0.7111	0.7710	0.6448	0.7111	0.7769
2	UMass_IESL	0.0965	0.1298	0.1682	0.0936	0.1298	0.1669
3	RPI_BLENDER	0.0031	0.0129	0.0363	0.0000	0.0129	0.0303
4	hltcoe	0.0000	0.0064	0.0240	0.0000	0.0064	0.0178

## Chinese

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.4155	0.5250	0.6076	0.4355	0.5250	0.6283
2	hltcoe	0.0982	0.1527	0.2836	0.0848	0.1527	0.2364
3	Stanford	0.0918	0.1415	0.2098	0.0863	0.1415	0.2037
4	RPI_BLENDER	0.0461	0.0723	0.1083	0.0447	0.0723	0.1028

## Cross-lingual

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.3917	0.4619	0.5104	0.4075	0.4619	0.5261
2	hltcoe	0.1057	0.1366	0.2129	0.0956	0.1366	0.1861
3	Stanford	0.0897	0.1118	0.1362	0.0898	0.1118	0.1373
4	UMass_IESL	0.0533	0.0695	0.0870	0.0535	0.0695	0.0897

# SF/KB Results

Confidence Interval - LDC-MAX-MICRO ENG

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.4122	0.4771	0.5543	0.4069	0.4771	0.5516
2	BBN	0.2189	0.2752	0.3270	0.2205	0.2752	0.3245
3	hltcoe	0.1453	0.2167	0.2569	0.1533	0.2167	0.2630
4	Stanford	0.1744	0.2147	0.2674	0.1677	0.2147	0.2585
5	UMass_IESL	0.1441	0.1741	0.2079	0.1443	0.1741	0.2056
6	RPI_BLENDER	0.1128	0.1506	0.1972	0.1110	0.1506	0.1933
7	ICTCAS_OKN	0.0860	0.1162	0.1501	0.0888	0.1162	0.1486
8	doughnutPRIS	0.0668	0.0869	0.1123	0.0654	0.0869	0.1101
9	DCD_SF	0.0506	0.0706	0.0971	0.0483	0.0706	0.0951
10	UNIST_SAIL	0.0417	0.0585	0.0849	0.0390	0.0585	0.0805

# SF/KB Results

Confidence Interval - LDC-MAX-MICRO SPA

## Spanish

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.6514	0.7373	0.8440	0.6364	0.7373	0.8230
2	UMass_IESL	0.1067	0.1603	0.2095	0.1161	0.1603	0.2241
3	hltcoe	0.0000	0.0155	0.0572	0.0000	0.0155	0.0412
4	RPI_BLENDER	0.0046	0.0153	0.0474	0.0000	0.0153	0.0374

## Chinese

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.2980	0.3816	0.4745	0.3081	0.3816	0.4817
2	hltcoe	0.2407	0.3293	0.4292	0.2332	0.3293	0.4148
3	Stanford	0.0969	0.1613	0.2649	0.0856	0.1613	0.2517
4	RPI_BLENDER	0.0664	0.1004	0.1437	0.0638	0.1004	0.1372

## Cross-lingual

Rank	Run	BCA			PER		
		95%(	F1	)95%	95%(	F1	)95%
1	LDC	0.4168	0.4847	0.5655	0.4163	0.4847	0.5628
2	hltcoe	0.2329	0.2858	0.3385	0.2357	0.2858	0.3328
3	Stanford	0.1352	0.1735	0.2097	0.1358	0.1735	0.2147
4	UMass_IESL	0.1123	0.1362	0.1598	0.1126	0.1362	0.1618

# SF/KB Results

Scores per slot - LDC-MEAN-MACRO ENG

Slot	#LDC/SF Queries	TopTeam (S)	Precision LDC/S	Recall LDC/S	F1 LDC/S	Diff
org_number_of_employees_members	3/4	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_age	7/12	RPI_BLENDER	1.00/0.21	1.00/0.21	1.00/0.21	0.7900
org_organizations_founded	6/9	DCD_SF	1.00/0.33	0.83/0.33	0.89/0.33	0.5600
per_cause_of_death	7/9	DCD_SF	1.00/0.50	1.00/0.50	1.00/0.50	0.5000
per_holds_shares_in	6/12	BBN	1.00/0.25	0.67/0.25	0.75/0.25	0.5000
...	...	...	...	...	...	...
gpe_member_of	6/7	UMass_IESL	0.75/0.75	0.62/0.62	0.67/0.67	0.0000
per_city_of_birth	4/4	ICTCAS_OKN	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
per_stateorprovince_of_birth	7/8	RPI_BLENDER	0.50/0.50	0.50/0.50	0.50/0.50	0.0000
org_holds_shares_in	3/5	BBN	1.00/1.00	0.50/0.50	0.67/0.67	0.0000
per_country_of_death	4/6	DCD_SF	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
org_shareholders	5/8	UMass_IESL	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
...	...	...	...	...	...	...
gpe_employees_or_members	7/9	hlcoe	0.36/0.83	0.24/0.56	0.25/0.63	-0.3800
gpe_residents_of_stateorprovince	2/3	Stanford	0.50/0.44	0.07/0.69	0.12/0.54	-0.4200
per_schools_attended	7/10	RPI_BLENDER	0.60/1.00	0.45/1.00	0.48/1.00	-0.5200
per_country_of_birth	3/3	BBN	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000
per_other_family	3/4	NAIST_CL	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000

# SF/KB Results

Scores per slot - LDC-MEAN-MACRO SPA

Slot	#LDC/SF Queries	TopTeam (S)	Precision LDC/S	Recall LDC/S	F1 LDC/S	Diff
org_number_of_employees_members	2/3	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
gpe_births_in_city	3/5	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_date_of_death	8/9	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_website	3/4	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_city_of_headquarters	4/6	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
...	...	...	...	...	...	...
gpe_headquarters_in_stateorprovince	5/6	UMass.IESL	0.50/0.14	0.10/0.84	0.17/0.24	-0.0700
gpe_headquarters_in_country	3/3	UMass.IESL	0.50/0.09	0.04/0.89	0.07/0.16	-0.0900
per_children	7/11	UMass.IESL	0.75/0.75	0.50/0.75	0.58/0.71	-0.1300
per_spouse	5/9	UMass.IESL	0.50/0.50	0.50/1.00	0.50/0.67	-0.1700
gpe_births_in_stateorprovince	5/5	UMass.IESL	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000

# SF/KB Results

Scores per slot - LDC-MEAN-MACRO CMN

Slot	#LDC/SF Queries	TopTeam (S)	Precision	Recall	F1	Diff
			LDC/S	LDC/S	LDC/S	
org_website	2/2	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_other_family	3/5	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_cause_of_death	7/11	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
gpe_member_of	6/9	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_date_founded	7/10	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
...	...	...	...	...	...	...
per_date_of_death	6/8	Stanford	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
gpe_births_in_city	4/4	Stanford	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
per_top_member_employee_of	3/3	RPI_BLENDER	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
per_title	5/7	Stanford	0.75/0.68	0.62/0.70	0.67/0.67	0.0000
per_country_of_death	3/4	RPI_BLENDER	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
...	...	...	...	...	...	...
gpe_headquarters_in_country	3/3	hltscoe	0.50/0.56	0.03/0.40	0.05/0.46	-0.4100
gpe_births_in_country	5/5	Stanford	0.00/0.31	0.00/0.82	0.00/0.42	-0.4200
gpe_residents_of_country	6/9	hltscoe	0.33/0.47	0.04/0.62	0.07/0.53	-0.4600
org_top_members_employees	5/6	Stanford	0.50/1.00	0.12/0.69	0.20/0.77	-0.5700
org_city_of_headquarters	4/6	Stanford	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000

# SF/KB Results

Scores per slot - LDC-MEAN-MACRO XLING

Slot	#LDC/SF Queries	TopTeam (S)	Precision <i>LDC/S</i>	Recall <i>LDC/S</i>	F1 <i>LDC/S</i>	Diff
per_stateorprovince_of_death	5/16	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_founded_by	2/6	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_religion	3/12	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_number_of_employees_members	4/9	-	0.83/0.00	1.00/0.00	0.89/0.00	0.8900
org_organizations_founded	6/21	UMass.IESL	1.00/0.08	0.83/0.04	0.89/0.06	0.8300
...	...	...	...	...	...	...
org_shareholders	5/17	UMass.IESL	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
gpe_employees_or_members	7/25	hltcoe	0.74/0.46	0.34/0.46	0.41/0.44	-0.0300
gpe_headquarters_in_country	3/9	hltcoe	0.83/0.35	0.07/0.15	0.12/0.20	-0.0800
gpe_headquarters_in_stateorprovince	6/21	hltcoe	0.97/0.49	0.08/0.20	0.14/0.27	-0.1300
gpe_residents_of_country	6/23	hltcoe	0.94/0.45	0.10/0.41	0.17/0.43	-0.2600

# SF/KB Results

Scores per slot - LDC-MAX-MICRO ENG

Slot	#LDC/SF Queries	TopTeam (S)	Precision	Recall	F1	Diff
			LDC/S	LDC/S	LDC/S	
org_number_of_employees_members	3/4	-	0.67/0.00	1.00/0.00	0.80/0.00	0.8000
org_date_dissolved	8/11	DCD_SF	1.00/0.14	1.00/0.50	1.00/0.22	0.7800
per_age	7/12	Stanford	1.00/1.00	1.00/0.29	1.00/0.44	0.5600
org_date_founded	7/7	RPI_BLENDER	1.00/0.40	1.00/0.50	1.00/0.44	0.5600
per_cause_of_death	7/9	Stanford	1.00/0.50	1.00/0.50	1.00/0.50	0.5000
...	...	...	...	...	...	...
gpe_member_of	6/7	UMass.IESL	0.75/0.75	0.60/0.60	0.67/0.67	0.0000
per_city_of_birth	4/4	hltcoe	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
per_cities_of_residence	5/7	UMass.IESL	1.00/1.00	0.50/0.50	0.67/0.67	0.0000
per_spouse	5/7	BBN	1.00/1.00	0.75/0.75	0.86/0.86	0.0000
per_country_of_death	4/6	ICTCAS_OKN	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
...	...	...	...	...	...	...
gpe_residents_of_country	6/7	BBN	1.00/0.67	0.08/0.37	0.15/0.48	-0.3300
gpe_residents_of_stateorprovince	2/3	Stanford	1.00/0.44	0.08/0.67	0.15/0.53	-0.3800
per_schools_attended	7/10	RPI_BLENDER	1.00/1.00	0.44/1.00	0.62/1.00	-0.3800
per_other_family	3/4	NAIST_CL	0.00/0.33	0.00/1.00	0.00/0.50	-0.5000
per_country_of_birth	3/3	ICTCAS_OKN	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000



# SF/KB Results

Scores per slot - LDC-MAX-MICRO SPA

Slot	#LDC/SF Queries	TopTeam (S)	Precision <i>LDC/S</i>	Recall <i>LDC/S</i>	F1 <i>LDC/S</i>	Diff
per_stateorprovince_of_birth	6/8	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_city_of_death	8/11	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_stateorprovince_of_headquarters	3/3	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_stateorprovince_of_death	5/6	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_statesorprovinces_of_residence	3/3	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
...	...	...	...	...	...	...
gpe_births_in_country	6/8	UMass.IESL	1.00/0.31	0.27/0.67	0.42/0.43	-0.0100
gpe_headquarters_in_country	3/3	UMass.IESL	1.00/0.08	0.05/0.84	0.10/0.15	-0.0500
gpe_headquarters_in_stateorprovince	5/6	UMass.IESL	1.00/0.12	0.09/0.83	0.16/0.22	-0.0600
org_subsidiaries	4/5	UMass.IESL	0.50/1.00	0.50/0.50	0.50/0.67	-0.1700
gpe_births_in_stateorprovince	5/5	UMass.IESL	0.00/1.00	0.00/1.00	0.00/1.00	-1.0000

# SF/KB Results

Scores per slot - LDC-MAX-MICRO CMN

Slot	#LDC/SF Queries	TopTeam (S)	Precision	Recall	F1	Diff
			LDC/S	LDC/S	LDC/S	
org_parents	4/4	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_children	6/8	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_statesorprovinces_of_residence	5/6	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_stateorprovince_of_death	5/5	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_stateorprovince_of_headquarters	2/2	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
...	...	...	...	...	...	...
per_origin	6/8	hltcoe	0.80/1.00	0.80/0.60	0.80/0.75	0.0500
per_employee_or_member_of	3/3	hltcoe	1.00/1.00	0.75/0.75	0.86/0.86	0.0000
gpe_births_in_city	4/4	Stanford	1.00/1.00	1.00/1.00	1.00/1.00	0.0000
per_title	5/7	Stanford	0.75/0.67	0.50/0.67	0.60/0.67	-0.0700
org_alternate_names	5/6	Stanford	0.50/0.50	0.20/0.40	0.29/0.44	-0.1500
...	...	...	...	...	...	...
org_top_members_employees	5/6	Stanford	1.00/1.00	0.22/0.67	0.36/0.80	-0.4400
gpe_headquarters_in_country	3/3	hltcoe	0.50/0.61	0.05/0.50	0.08/0.55	-0.4700
per_top_member_employee_of	3/3	RPI-BLENDER	0.33/1.00	1.00/1.00	0.50/1.00	-0.5000
gpe_residents_of_country	6/9	hltcoe	0.67/0.55	0.03/0.71	0.06/0.62	-0.5600
org_city_of_headquarters	4/6	Stanford	0.00/0.50	0.00/1.00	0.00/0.67	-0.6700

# SF/KB Results

Scores per slot - LDC-MAX-MICRO XLING

Slot	#LDC/SF Queries	TopTeam (S)	Precision LDC/S	Recall LDC/S	F1 LDC/S	Diff
per_stateorprovince_of_death	5/16	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_founded_by	2/6	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
per_religion	3/12	-	1.00/0.00	1.00/0.00	1.00/0.00	1.0000
org_date_dissolved	8/29	-	0.67/0.00	1.00/0.00	0.80/0.00	0.8000
per_city_of_death	8/29	-	0.50/0.00	1.00/0.00	0.67/0.00	0.6700
...	...	...	...	...	...	...
org_stateorprovince_of_headquarters	7/13	Stanford	0.75/0.83	1.00/0.83	0.86/0.83	0.0300
gpe_subsidiaries	7/23	hltcoe	0.75/0.46	0.26/0.32	0.39/0.38	0.0100
gpe_births_in_country	6/21	Stanford	0.80/0.21	0.16/0.33	0.27/0.26	0.0100
per_holds_shares_in	7/28	hltcoe	1.00/1.00	0.50/0.50	0.67/0.67	0.0000
org_political_religious_affiliation	8/28	hltcoe	0.44/1.00	0.80/0.40	0.57/0.57	0.0000
...	...	...	...	...	...	...
gpe_headquarters_in_country	3/9	hltcoe	0.80/0.50	0.05/0.28	0.10/0.36	-0.2600
gpe_headquarters_in_city	6/18	hltcoe	1.00/0.50	0.12/0.45	0.21/0.48	-0.2700
gpe_headquarters_in_stateorprovince	6/21	Stanford	0.93/0.37	0.06/0.47	0.12/0.42	-0.3000
gpe_residents_of_country	6/23	hltcoe	0.94/0.50	0.09/0.46	0.17/0.48	-0.3100
org_shareholders	5/17	UMass.IESL	0.50/1.00	1.00/1.00	0.67/1.00	-0.3300

# SF/KB Results

## NIL Queries

Team	LDC-MEAN-0-Macro			NIL Queries		
	Precision	Recall	F1	Precision	Recall	F1
LDC	0.7098	0.5741	0.6043	0.2921	0.9024	0.4413
Stanford	0.2647	0.3019	0.2597	0.2902	0.8943	0.4382
BBN	0.2676	0.2357	0.2314	0.3085	0.9756	0.4688
hltcoe	0.2223	0.2122	0.2064	0.3013	0.9431	0.4567
UMass_IESL	0.1857	0.2123	0.1824	0.2548	0.7480	0.3801
RPI_BLENDER	0.1726	0.1887	0.1713	0.2846	0.8699	0.4289
ICTCAS_OKN	0.1295	0.1458	0.1268	0.2995	0.9350	0.4537
doughnutPRIS	0.1275	0.1456	0.1242	0.1898	0.5122	0.2770
DCD_SF	0.1257	0.1073	0.1082	0.0528	0.1220	0.0737
NAIST_CL	0.0970	0.0898	0.0920	0.1063	0.2602	0.1509
MSR	0.0945	0.0910	0.0778	0.0074	0.0163	0.0102
UNIST_SAIL	0.0441	0.0581	0.0424	0.2507	0.7317	0.3734
CMUML	0.0388	0.0257	0.0282	0.3138	1.0000	0.4777
lilian	0.0350	0.0227	0.0249	0.3067	0.9675	0.4658
SoochowNLP	0.0174	0.0587	0.0230	0.2788	0.8455	0.4193
summa	0.0293	0.0138	0.0171	0.3138	1.0000	0.4777
IRTSX	0.0031	0.0072	0.0042	0.3013	0.9431	0.4567

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 KB Entity Discovery Evaluation
  - Participants
  - Results
- 3 SF/KB Evaluation
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion

- Refine the output from systems
- Input: SF/KB submissions, preliminary assessments and scores

Team	Filtering (ENG)	Ensembling (ENG)	Ensembling (CMN)	Ensembling (SPA)	Ensembling (XLING)	Total
gator_dsr	-	3	-	-	1	4
IRTSX	4	4	-	-	-	8
SAFT_ISI	-	3	2	2	-	7
UI_CCG	3	-	-	-	-	3
UTAustin	-	5	5	5	-	15
<b>Total</b>	<b>7</b>	<b>15</b>	<b>7</b>	<b>7</b>	<b>1</b>	<b>37</b>

# Results

## SFV Confidence Interval - SF-MICRO ENG

Rank	Team	95%(	F1	)95%
1	LDC_SF_ENG_1	0.4110	0.4873	0.5742
2	gator_dsr4.ENG.ensemble	0.2639	0.3242	0.3714
3	SAFT_ISI1.ENG.ensemble	0.2375	0.2864	0.3386
4	SAFT_ISI2.ENG.ensemble	0.2358	0.2864	0.3358
5	gator_dsr1.ENG.ensemble	0.2088	0.2818	0.3157
6	UTAustin4.ENG.ensemble	0.2461	0.2734	0.3312
7	gator_dsr2.ENG.ensemble	0.1974	0.2732	0.3026
8	BBN_KB_ENG_4	0.2133	0.2703	0.3393
9	SAFT_ISI3.ENG.ensemble	0.2351	0.2702	0.3129
10	BBN_KB_ENG_1	0.2052	0.2693	0.3253
11	UI_CCG3.BBN_KB_ENG_4.filtered	0.1828	0.2657	0.3021
12	BBN_KB_ENG_3	0.2142	0.2651	0.3327
13	UI_CCG3.BBN_KB_ENG_1.filtered	0.2112	0.2648	0.3305
14	UTAustin5.ENG.ensemble	0.2153	0.2614	0.2994
15	UI_CCG2.BBN_KB_ENG_4.filtered	0.1922	0.2495	0.3136
16	UI_CCG2.BBN_KB_ENG_1.filtered	0.1678	0.2485	0.2948
17	UI_CCG2.BBN_KB_ENG_3.filtered	0.1805	0.2447	0.3037



# Results

## SFV Confidence Interval - SF-MACRO ENG

Rank	Team	95%(	F1	)95%
1	LDC_SF_ENG_1	0.4397	0.5287	0.6013
2	gator_dsr1.ENG.ensemble	0.2501	0.3249	0.3571
3	gator_dsr2.ENG.ensemble	0.2601	0.3126	0.3500
4	gator_dsr4.ENG.ensemble	0.2162	0.2781	0.3190
5	SAFT_ISI3.ENG.ensemble	0.1915	0.2335	0.2677
6	UTAustin4.ENG.ensemble	0.1815	0.2226	0.2577
7	SAFT_ISI2.ENG.ensemble	0.1719	0.2195	0.2560
8	SAFT_ISI1.ENG.ensemble	0.1747	0.2195	0.2610
9	UTAustin5.ENG.ensemble	0.1499	0.1935	0.2219
10	BBN_KB_ENG_3	0.1464	0.1896	0.2320
11	BBN_KB_ENG_4	0.1465	0.1870	0.2381
12	BBN_KB_ENG_1	0.1368	0.1857	0.2214
13	UI_CCG3.BBN_KB_ENG_4.filtered	0.1426	0.1821	0.2255
14	UI_CCG3.BBN_KB_ENG_1.filtered	0.1380	0.1808	0.2243
15	UI_CCG2.BBN_KB_ENG_3.filtered	0.1102	0.1632	0.1895
16	UI_CCG2.BBN_KB_ENG_4.filtered	0.1222	0.1607	0.2067
17	UI_CCG2.BBN_KB_ENG_1.filtered	0.1247	0.1594	0.2071

# Results

## SFV Confidence Interval - LDC-MEAN-MACRO ENG

Rank	Team	95%(	F1	)95%
1	LDC_SF_ENG_1	0.4374	0.5207	0.5894
2	gator_dsr1.ENG.ensemble	0.2643	0.3122	0.3610
3	gator_dsr2.ENG.ensemble	0.2490	0.3001	0.3476
4	gator_dsr4.ENG.ensemble	0.2199	0.2710	0.3209
5	SAFT_ISI3.ENG.ensemble	0.1915	0.2261	0.2615
6	UTAustin4.ENG.ensemble	0.1866	0.2198	0.2567
7	SAFT_ISI2.ENG.ensemble	0.1717	0.2071	0.2561
8	SAFT_ISI1.ENG.ensemble	0.1677	0.2071	0.2496
9	UTAustin5.ENG.ensemble	0.1542	0.1907	0.2275
10	BBN_KB_ENG_3	0.1450	0.1849	0.2278
11	BBN_KB_ENG_4	0.1459	0.1832	0.2304
12	BBN_KB_ENG_1	0.1459	0.1822	0.2273
13	UI_CCG3.BBN_KB_ENG_4.filtered	0.1435	0.1778	0.2229
14	UI_CCG3.BBN_KB_ENG_1.filtered	0.1381	0.1769	0.2184
15	UI_CCG2.BBN_KB_ENG_3.filtered	0.1263	0.1622	0.2028
16	UI_CCG2.BBN_KB_ENG_4.filtered	0.1220	0.1607	0.1994
17	UI_CCG2.BBN_KB_ENG_1.filtered	0.1262	0.1598	0.2039

# Results

## SFV Confidence Interval - LDC-MAX-MICRO ENG

Rank	Team	95%(	F1	)95%
1	LDC_SF_ENG_1	0.4167	0.4969	0.5748
2	gator_dsr4.ENG.ensemble	0.2742	0.3301	0.3835
3	SAFT_ISI2.ENG.ensemble	0.2384	0.2832	0.3304
4	SAFT_ISI1.ENG.ensemble	0.2343	0.2830	0.3336
5	BBN_KB_ENG_4	0.2166	0.2786	0.3316
6	BBN_KB_ENG_1	0.2208	0.2786	0.3366
7	UTAustin4.ENG.ensemble	0.2341	0.2769	0.3153
8	SAFT_ISI3.ENG.ensemble	0.2332	0.2735	0.3091
9	UI_CCG3.BBN_KB_ENG_1.filtered	0.2169	0.2723	0.3264
10	UI_CCG3.BBN_KB_ENG_4.filtered	0.2190	0.2722	0.3316
11	gator_dsr1.ENG.ensemble	0.2134	0.2721	0.3210
12	BBN_KB_ENG_3	0.2187	0.2718	0.3273
13	UTAustin5.ENG.ensemble	0.2209	0.2676	0.3110
14	gator_dsr2.ENG.ensemble	0.2109	0.2647	0.3154
15	UI_CCG2.BBN_KB_ENG_1.filtered	0.1980	0.2577	0.3130
16	UI_CCG2.BBN_KB_ENG_4.filtered	0.2046	0.2572	0.3190
17	UI_CCG2.BBN_KB_ENG_3.filtered	0.1946	0.2512	0.3143

# Outline

- 1 Introduction
  - Task Variants
  - Changes in 2016
- 2 KB Entity Discovery Evaluation
  - Participants
  - Results
- 3 SF/KB Evaluation
  - Definitions
  - Queries
  - Participants
  - Results
- 4 SFV Evaluation
  - Setup
  - Participants
  - Results
- 5 Conclusion

# Conclusion

- Encouraging number of participants - given first year of cross-lingual
- Automatic systems appear to get about 50% or less performance as compared to manual systems.
- Systems getting better at dealing with name variants
- Easy slots may perhaps be removed
- Focus more on specific and difficult slots