Overview of the TAC2012 Knowledge Base Population Entity Linking Tasks

JAMES MAYFIELD
(JOHNS HOPKINS UNIVERSITY)

JAVIER ARTILES
(RAKUTEN INSTITUTE OF TECHNOLOGY)

HOA TRANG DANG
(NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY)

JOE ELLIS, XUANSONG LI, KIRA GRIFFITT, STEPHANIE M. STRASSEL, JONATHAN WRIGHT
(LINGUISTIC DATA CONSORTIUM)

Molingual Entity Linking Task

English source document

BBC News Sport Weather Travel

Page last updated at 08:31 GMT, Monday, 15 October 2012 09:31 UK

Moore: "Why put barriers up between Scotland and UK?"

The terms of a referendum, giving Scots the chance to choose independence from the United Kingdom, will be formally approved in Edinburgh today.

Scotland's first minister Alex Salmond will sign the agreement with Prime Minister David Cameron, for a vote to be held in two years' time.

The BBC's political editor Nick Robinson and BBC Scotland's political editor Brian Taylor explained that the terms under which the referendum will be held, are crucially important to its outcome.

And talking to the Today programme's John Humphrys, Scottish MP Michael Moore argued that being part of the UK provides Scotland with security. "When two of our biggest banks - RBS and Bank of Scotland - collapsed, it was the strength and size of the UK economy that helped us to cope."

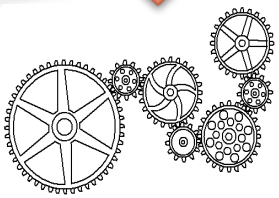
"If you look at our businesses", he added, "we export twice as much to England as we do to any other part of the world.

"Why would you put barriers up between us?"

Query

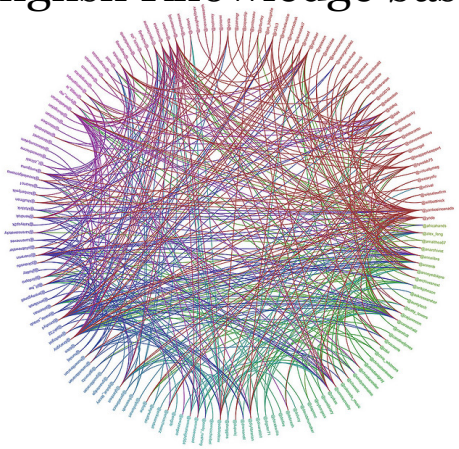
And talking to the Today programme's John Humphrys, Scottish MP **Michael Moore** argued that being part of the UK provides Scotland with security. "When two of our biggest banks - RBS and Bank of Scotland - collapsed, it was th[illegible] he UK economy that helped us to [illegible]

Query name offsets



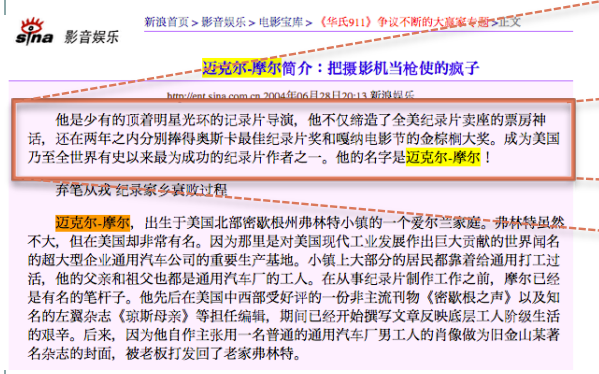
Entity Linking System

English Knowledge base



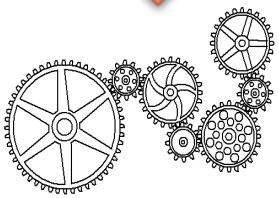
Cross-lingual Entity Linking Task

Chinese source document



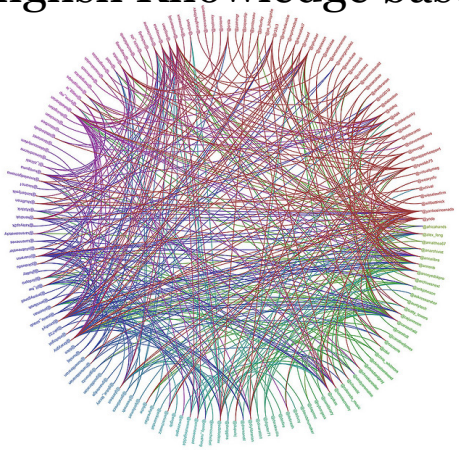
Query

他是少有的顶着明星光环的记录片导演，他不仅缔造了全美纪录片卖座的票房神话，还在两年之内分别捧得奥斯卡最佳纪录片奖和嘎纳电影节的金棕榈大奖。成为美国乃至全世界有史以来最为成功的纪录片作者之一。他的名字是**迈克尔-摩尔**！



Entity Linking System

English Knowledge base



New Spanish subtask

Entity Linking Task

Spanish source document

EL PAÍS PORTADA

INTERNACIONAL

Escocia votará en 2014 sobre su independencia

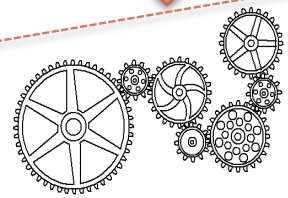
- Londres y Edimburgo ultiman un acuerdo para que el referéndum sea vinculante y se ciña exclusivamente al sí o el no a la secesión

WALTER OPPENHEIMER | Londres | 10 OCT 2012 - 20:50 CET 618

El preacuerdo, alcanzado tras varios meses de negociaciones entre el ministro británico para Escocia, Michael Moore, y la número dos del Gobierno escocés, Nicola Sturgeon, fue cerrado el lunes por la noche en una conversación telefónica entre ambos, que se reunirán para el visto bueno final el viernes que viene.

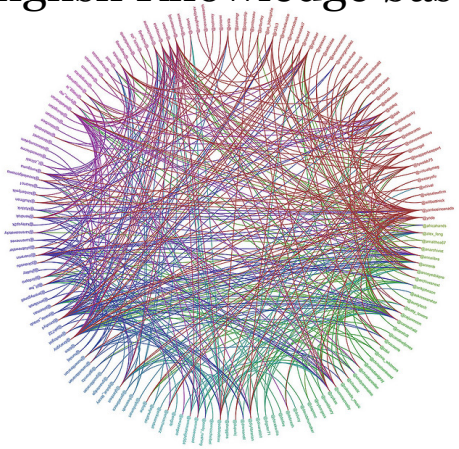
Query

El preacuerdo, alcanzado tras varios meses de negociaciones entre el ministro británico para Escocia, **Michael Moore**, y la número dos del Gobierno escocés, Nicola Sturgeon, fue cerrado el lunes por la noche en una conversación telefónica entre ambos, que se reunirán para el visto bueno final el viernes que viene



Entity Linking System

English Knowledge base



Scoring metric: B-Cubed+

L(e) and *C(e)* the category and the cluster of an item *e*,

SI(e) and *GI(e)* represent, respectively, the system (*i.e.*, participant-submitted) and gold-standard (ground truth) KB identifiers for an item *e*.

We define the correctness of the relation between *e* and *e*' in the distribution as:

$$\mathrm{G}(e, e') = \begin{cases} 1 \text{ iff } L(e) = L(e') \wedge C(e) = C(e') \wedge \\ GI(e) = SI(e) = GI(e') = SI(e') \\ \\ 0 \text{ otherwise} \end{cases}$$

Two items are correctly related when they share a category if and only if they appear in the same cluster and share the same KB identifier in the system and the gold standard.

$$\text{Precision} = \text{Avg}_e[\text{Avg}_{e'.C(e)=C(e')}[\mathrm{G}(e, e')]]$$

$$\text{Recall} = \text{Avg}_e[\text{Avg}_{e'.L(e)=L(e')}[\mathrm{G}(e, e')]]$$

English Entity Linking Results (B^3+ F1 score)

Run ID	All (2226)	in-KB (1177)	NIL (1049)	NW (1471)	WB (755)	PER (918)	ORG (706)	GPE (602)
MS_MLI3	0.730	0.685	0.781	0.782	0.630	0.809	0.715	0.627
hltcoe1	0.699	0.653	0.749	0.759	0.580	0.840	0.615	0.579
lcc20123	0.689	0.620	0.765	0.710	0.646	0.714	0.717	0.614
Blender_CUNY3	0.688	0.595	0.789	0.741	0.583	0.833	0.666	0.488
SYDNEY_CMCRC1	0.665	0.656	0.675	0.700	0.598	0.739	0.598	0.631
HITS2	0.621	0.561	0.683	0.644	0.572	0.719	0.600	0.492
JVN_TDT1	0.586	0.491	0.688	0.596	0.564	0.719	0.597	0.365
UGENT_IBCN1	0.585	0.548	0.626	0.632	0.494	0.723	0.490	0.484
basistech3	0.569	0.528	0.614	0.610	0.488	0.783	0.396	0.438
UMass2	0.563	0.539	0.589	0.579	0.531	0.605	0.585	0.471
nlprir1	0.562	0.477	0.657	0.597	0.495	0.691	0.517	0.414
ualberta2	0.547	0.501	0.598	0.574	0.494	0.700	0.468	0.403

- 25 teams submitted a total of 98 runs.
- A high rate of singleton NIL clusters gives a strong clustering baseline. One-in-one clustering achieves a higher score than any of the top systems (see Blender_CUNY team presentation for more details).
  - High coverage of entities in the KB
  - Entities not in the KB are more likely to be rare (singletons).

Cross-lingual EL results

Chinese Entity Linking Results (B^3+ F1 score) 4 teams / 12 runs

run id	All 2122	in-KB 1240	NIL 882	NW 1362	WB 760	PER 699	ORG 718	GPE 705
HITS1	0.740	0.736	0.737	0.717	0.781	0.604	0.777	0.834
lcc20122	0.667	0.613	0.741	0.662	0.675	0.628	0.659	0.705
basistech3	0.566	0.476	0.693	0.564	0.570	0.559	0.489	0.645
ualberta1	0.434	0.329	0.577	0.434	0.433	0.474	0.325	0.485

Spanish Entity Linking Results (B^3+ F1 score) 4 teams / 17 runs

run id	All 2066	in-KB 923	NIL 1143	PER 669	ORG 539	GPE 858
lcc20121	0.643	0.623	0.657	0.663	0.659	0.608
hltcoe4	0.632	0.491	0.745	0.784	0.602	0.530
basistech3	0.599	0.544	0.643	0.805	0.522	0.479
HITS3	0.538	0.468	0.579	0.582	0.645	0.412

Name ambiguity/variety

Monolingual EL

measure	year	all	NIL	non-NIL
ambiguity	2010	12.9	9.3	5.7
	2011	13.1	7.1	12.1
	2012	46.4	35.2	40.9
variety	2010	2.1	1.7	2.5
	2011	1.6	0.9	2.4
	2012	6.4	8.3	4.7

Cross-lingual EL

EL task	measure	year	all	NIL	non-NIL
Chinese	ambiguity	2011	20.9	5.7	9.3
		2012	8.4	14.8	3.4
	variety	2011	1.6	2.4	0.9
		2012	52.9	32.1	84.2
Spanish	ambiguity	2012	41.2	35.7	34.7
	variety	2012	4.0	3.9	4.0

- *Ambiguity*: the percentage of name strings referring to more than one entity cluster.
- *Variety:* the percentage of entity clusters expressed by more than one name string.

- A drop in overall system scores compared to previous year can be attributed in part to a higher name ambiguity/variety.

External resources and tools

- Cross-Lingual Dictionary for English Wikipedia Concepts (Spitkovsky and Chang, 2012).
  - higher recall at the KB candidate generation step!
- Newer versions of Wikipedia (linked back to the reference KB).
  - Time to upgrade the reference KB?
  - Other Knowledge Bases: Freebase, DBPedia.
- KBP specific and publicly available: *CUNY-Blender KBP toolkit*.
  - *Reusability!*
- Most popular tool among participants: *Stanford NLP toolkit* (NER, POS, parsing, etc.).
  - Other tools include: NER C&C Tools (Curran et al., 2007). LCC's CiceroLite NER system (Lehmann et al., 2007), etc.

Some highlights from the EL systems

- Trend towards global document EL vs local EL.
  - Should we have a full document EL task in the way WSD had during Senseval ?

Clustering approaches:

- System with the best NIL score in English EL (0.78) does not use a sophisticated clustering approach [MSR].
  - *"acronym expansion matching in the text, and the identity of surface forms identified in target documents by the entity extraction and disambiguation component "*
- MCMC clustering showed similar results to Agglomerative clustering but is suitable for large datasets [LCC2012].

Join inference:

- Joint infrence through Markov Random Field for entity mention resolution on small contexts. [HLTCOE]
  - Structured prediction cascades for reducing the number of KB candidates per mention. [HLTCOE]
- Recognition of unknown entities and clustering jointly using Markov Logic Networks [HITS] .

Some highlights from the EL systems

KB Candidate selection:

- Delay the entity mention boundary detection until the disambiguation stage [MSR] .
- Entity names transformation rule extraction for KB candidate retrieval [SYDNEY CMCRC].

- Geo-location based features.
- Densiometric zoning approach of Kohlschtter and Nejdl (2008) to select the appropriate zones which contain unstructured natural language text [LCC2012] .

- Two main strategies in cross-lingual EL:
  - Translate source document to the target KB language (allows re-using existing English EL systems).
  - Map articles in the source collection language Wikipedia to entries in the English Wikipedia (e.g. using interlanguage links).

LDC EL presentation

Some possibilities for KBP 2013

- Micro-blogging source documents.
  - New challenges, including informal text and extremely short context.
  - New features derived of the social network structure.
- Confidence values in EL system's output.
- A more strict definition of query name boundaries.

- Further discussion on this at the end of the workshop.