Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking

Heng Ji[1], **Joel Nothman**[2], **Ben Hachey**[3], **Radu Florian**[4]

[1] Computer Science Department, Rensselaer Polytechnic Institute
jih@rpi.edu
[2] Department of Computing and Information Systems, University of Melbourne
joel.nothman@gmail.com
[3] School of Information Technologies, University of Sydney
ben.hachey@gmail.com
[4] IBM TJ Waston Research Center
raduf@us.ibm.com

Abstract

In this paper we give an overview of the Tri-lingual Entity Discovery and Linking task at the Knowledge Base Population (KBP) track at TAC2015. In this year we introduced a new end-to-end Tri-lingual entity discovery and linking task which requires a system to take raw texts from three languages (English, Chinese and Spanish) as input, automatically extract entity mentions, link them to an English knowledge base, and cluster NIL mentions across languages. More entity types and mention types were also added into some languages. In this paper we provide an overview of the task definition, annotation issues, successful methods and research challenges associated with this new task. This new task has attracted a lot of participants and has intrigued many interesting research problems and potential approaches. We believe it's a promising task to be combined with Tri-lingual slot filling to form a new Tri-lingual cool-start KBP track in TAC2016.

1 Introduction

We have achieved some promising successes in English Entity Discovery and Linking (EDL) in the previous years. However, for certain entities, a lot of new and detailed information is only available in documents written in a foreign language for which there may be very few linguistic resources (annotated data, tools, etc.) available. For example, when the Ebola outbreak started in 2014, news articles in Yoruba tend to report the newest updates with many details such as individual hospitals, researchers and local town names. In contrast, news articles in English tend to only focus on general statistics such as the number of deaths, or non-local information such as a foreign government's reaction to the outbreak. Therefore, we believe it will be highly valuable to automatically link and fuse the knowledge across languages so we can construct a more complete profile in order to gain comprehensive understanding of an entity or event. On the other hand, the new incident language (IL) often has very low linguistic resources and data for rapid development of an EDL system. But once the cross-lingual links are built, we can take advantage of the high-resources in English, including annotated data, gazetteers and rich knowledge representations, and transfer them to develop and enhance EDL in IL. Therefore, this year we extend EDL from mono-lingual to cross-lingual. An EDL system is required to discover entity mentions from all three languages instead of one language, and link them to an English Knowledge Base (KB) or cluster them into NIL entities across languages.

Previous Entity Linking and EDL tasks mainly focused on three main types: Person, Organization and Geo-political entities. Recent work (Xiao and Weld, 2012; Lee et al., 2006) suggests that using a larger set of fine-grained types can lead to substantial improvement in downstream NLP applications. We aim to gradually add new entity types into the KBP program. This year we add locations and facilities for all three languages. This addition triggered some new research interests in fine-grained entity typing and universal entity schema discovery.

Finally, it's valuable to cluster nominal mentions for creating new entries in KB. We also explored person nominal mention extraction from English and identified some new challenges.

To summarize, compared to the KBP2014 EDL task (Ji et al., 2014), the main changes and improvement in KBP2015 include:

- Extended English EDL task from mono-lingual to tri-lingual;
- Added two new entity types - natural locations (LOC) and facilities (FAC), for all three languages;
- Added person nominal mentions for English;
- Prepared and used a new KB based on Freebase snapshot;
- Defined a new diagnostic task of English Entity Discovery within the Cold-start KBP track.

The rest of this paper is structured as follows. Section 2 describes the definition of the full Tri-lingual EDL task and various diagnostic tasks. Section 3 briefly summarizes the participants. Section 4 highlights some annotation efforts. Section 5 summarize evaluation results and some general progress report over years. Section 6 summarizes new and effective methods, while Section 7 provides some detailed analysis and discussion about remaining challenges. Section 8 sketches our future directions.

2 Task Definition and Evaluation Metrics

This section will summarize the Tri-lingual Entity Discovery and Linking tasks conducted at KBP 2015. More details regarding data format and scoring software can be found in the task website[1].

2.1 Full Task

Given a document collection in three languages (English, Chinese and Spanish) as input, a tri-lingual EDL system is required to automatically identify entity mentions from a source collection of textual documents in multiple languages (English, Chinese and Spanish), classify them into one of the following pre-defined five types: Person (PER), Geo-political Entity (GPE), Organization (ORG), Location (LOC) and Facility (FAC), and link them to an existing English Knowledge Base (KB), and cluster mentions for those NIL entities that don't have corresponding KB entries. Figure 1 illustrates an example for the full task.

Figure 1: Tri-lingual EDL Input/Output Example

Besides name mentions, person nominal mentions referring to specific, real-world individual entities should also be extracted from English. The system output includes the following fields:

- system run ID;
- mention ID: unique for each entity mention;
- mention head string: the full head string of the entity mention;
- document ID: mention head start offset mention head end offset: an ID for a document in the source corpus from which the mention head was extracted, the starting offset of the mention head, and the ending offset of the mention head;
- reference KB link entity ID, or NIL cluster ID: A unique NIL ID or an entity node ID, correspondent to entity linking annotation and NIL-coreference (clustering) annotation respectively;
- entity type: GPE, ORG, PER, LOC, FAC type indicator for the entity;
- mention type: NAM (name), NOM (nominal) type indicator for the entity mention;
- confidence value.

[1] http://nlp.cs.rpi.edu/kbp/2015/

Short name	Name in scoring software	Filter	Key	Evaluates
Mention evaluation				
`NER`	`strong_mention_match`	NA	*span*	Identification
`NERC`	`strong_typed_mention_match`	NA	*span,type*	+ classification
Linking evaluation				
`NERLC`	`strong_typed_all_match`	NA	*span,type,kbid*	+ linking
`NELC`	`strong_typed_link_match`	*is linked*	*span,type,kbid*	Link recognition and classification
`NENC`	`strong_typed_nil_match`	*is nil*	*span,type*	NIL recognition and classification
Tagging evaluation				
`KBIDs`	`entity_match`	*is linked*	*docid,kbid*	Document tagging
Clustering evaluation				
`CEAFm`	`mention_ceaf`	NA	*span*	Identification and clustering
`CEAFmC`	`typed_mention_ceaf`	NA	*span,type*	+ classification
`CEAFmC+`	`typed_mention_ceaf_plus`	NA	*span,type,kbid*	+ linking

Table 1: Evaluation measures for entity discovery and linking, each reported as P, R, and F_1. *Span* is shorthand for (*document identifier, begin offset, end offset*). *Type* is PER, ORG or GPE. *Kbid* is the KB identifier or NIL.

2.2 Scoring Metrics

TAC 2015 continues the 2014 measures for entity detection and linking (EDL) and its diagnostic (EL) variant, listed in Table 1. EL provides gold standard mentions to systems, isolating linking and clustering performance. The scorer is available at `https://github.com/wikilinks/neleval`.

2.2.1 Set-based metrics

Recognizing and linking entity mentions can be seen as a tagging task. Here evaluation treats an annotation as a set of distinct tuples, and calculates precision and recall between gold (G) and system (S) annotations:

$$P = \frac{|G \cap S|}{|S|} \qquad R = \frac{|G \cap S|}{|G|}$$

For all measures P and R are combined as their balanced harmonic mean, $F_1 = \frac{2PR}{P+R}$.

By selecting only a subset of annotated fields to include in a tuple, and by including only those tuples that match some criteria, this metric can be varied to evaluate different aspects of systems (cf. Hachey et al. (2014) which also relates such metric variants to the entity disambiguation literature). As shown in Table 1, `NER` and `NERC` metrics evaluate mention detection and classification, while `NERL` measures linking performance but disregards entity type and NIL clustering. In the EL task where mentions are given, `NERL` is equivalent to the linking accuracy score reported in previous KBP evaluations.

Results below also refer to other diagnostic measures, including `NEL` which reports linking (and mention detection) performance, discarding NIL annotations; `NEN` reports the performance of NIL annotations alone. `KBIDs` considers the set of KB entities extracted per document, disregarding mention spans and discarding NILs. This measure, elsewhere called *bag-of-titles evaluation*, does not penalize boundary errors in mention detection, while also being a meaningful task metric for document indexing applications of named entity disambiguation.

2.2.2 Clustering metrics

Alternatively, entity linking is understood as a cross-document coreference task, in which the set of tuples is partitioned by the assigned entity ID (for KB and NIL entities), and a coreference evaluation metric is applied. To evaluate clustering, we apply Mention CEAF (Luo, 2005), which finds the optimal alignment between system and gold standard clusters, and then evaluates precision and recall micro-averaged over mentions, as in a multiclass classification evaluation. While other metrics reward systems for correctly identifying coreference within clusters, a system which splits an entity into multiple clusters will only be rewarded for the largest and purest of those clusters. `CEAFm` performance is bounded from above by `NER`, `CEAFmC` by `NERC` and so on.

Mention CEAF (`CEAFm`) is calculated as follows. Let $G_i \in \mathcal{G}$ describe the gold partitioning, and $S_i \in \mathcal{S}$ the system, we calculate the maximum score bijection m:

$$m = \arg\max_m \sum_{i=1}^{|\mathcal{G}|} \left|G_i \cap S_{m(i)}\right|$$
$$\text{s.t. } m(i) = m(j) \iff i = j$$

Then CEAFm is calculated by:

$$P_{\texttt{CEAFm}} = \frac{\sum_{i=1}^{|\mathcal{G}|} \left|G_i \cap S_{m(i)}\right|}{\sum_{i=1}^{|\mathcal{S}|} |S_i|}$$
$$R_{\texttt{CEAFm}} = \frac{\sum_{i=1}^{|\mathcal{G}|} \left|G_i \cap S_{m(i)}\right|}{\sum_{i=1}^{|\mathcal{G}|} |G_i|}$$

As with set-based metrics, selecting a subset of fields or filtering tuples introduces variants that only award score when, for example, the system matches the gold standard KB link or entity type. Compared to KBP2014 CEAFm metrics, we added two new enhanced variants:

- CEAFmC: adding type match into CEAFm
- CEAFmC+: combining CEAFmC and KB ID matching, which can serve as an end-to-end metric for EDL to measure the overall performance of extraction, linking and clustering.

2.2.3 Confidence intervals

We calculate $c\%$ confidence intervals for set-based metrics by bootstrap resampling documents from the corpus, calculating these pseudo-systems' scores, and determining their values at the $\frac{100-c}{2}$th and $\frac{100+c}{2}$th percentiles of 2500 bootstrap resamples. This procedure assumes that and system annotates documents independently, and intervals are not reliable where systems use global clustering information in their set-based output (i.e. beyond NIL cluster assignment). For similar reasons, we do not calculate confidence intervals for clustering metrics.

2.2.4 Weak boundary matching

Since the introduction of mention detection to TAC EDL in 2014, boundary errors in the detection of mentions for linking are common. For some applications, it is appropriate to award systems for near matches on mention boundaries, which can be due to spurious ambiguity. In other cases systems' performance could be improved by fixing boundaries as a post-process, such as merging adjacent mentions with the same entity link, so we diagnostically consider these more leniently.

To evaluate mention extraction performance, we also added a variant for partial (weak) mention boundary matching. It checks the number of overlapped characters between a system generated mention and a ground-truth mention. The impact of this strategy will be reported in section 5.6.

In addition, we have developed a method to align unmatched system mentions to unmatched gold mentions, for which there can be ambiguity in a coreference task such as NIL clustering. We outline the approach to finding a maximal alignment for evaluating under Mention CEAF, which will be described in full at a later venue:

1. Candidate mention pairs are grouped into pairs of gold and predicted entities.
2. The number of potential matches is calculated for each gold-predicted entity pair. Since one gold mention may have multiple unmatched mentions from the same predicted entity, and vice-versa, this is non-trivial.
3. These potential counts are added onto the contingency matrix.
4. The maximum-scoring assignment of predicted to gold entities is found, as for CEAF calculation.
5. This can be used directly to calculate Mention CEAF where potential alignments for assigned entities are fixed.
6. Further boundary errors may be fixed by repeating this process while disregarding entity pairs involved in the maximum-scoring assignment.

Note that duplicated system mentions for a single gold mention (or vice-versa) will still result in precision (or recall) errors, as at most one system is fixed to each gold mention. Thus we also report Mention CEAF under this optimal fixed span condition (CEAFm-weak).

2.3 Diagnostic Tasks

In order to investigate compare EDL's performance on various steps, languages, entity types, mention types, and check the progress over years, we also allow systems to submit results on their favored combinations. For example, a team can choose to focus on 1-2 languages, or name mentions, or linking task using perfect mentions. A perfect mention ('query' as in Entity Linking tasks in previous years) includes the following five fields:

- query id: A query ID, unique for each entity mention;
- mention: The full head string of the query entity mention;
- docid: An ID for a document in the source corpus from which the mention head was extracted;
- The starting offset for the mention head;
- The ending offset for the mention head;

For example:

```
⟨query id=``EDL15_ENG_0001"⟩
⟨name⟩cairo⟨/name⟩
⟨docid⟩bolt-eng-DF-200-192451-5799099⟨/docid⟩
⟨beg⟩2450⟨/beg⟩
⟨end⟩2454⟨/end⟩
⟨/query⟩
```

The output format is the same as the full EDL task.

3 Participants Overview

Table 2 summarizes the participants for the Trilingual EDL task. In total 10 teams submitted 35 runs for the full task and 10 teams submitted 25 runs for the diagnostic task (Entity Linking with perfect mentions as input).

Six teams performed linking across three languages in the full task, with other teams – including all those exclusively in the diagnostic task – linking one or two languages. Only four teams identified person nominal mentions, though most distinguished other entity types.

4 Data Annotation and Resources

The details of the data annotation for KBP2015 are presented in a separate paper by the Linguistic Data Consortium (Ellis et al., 2015). This year we used a new reference knowledge base derived from BaseKB, a cleaned version of English Freebase. The detailed statistics of the training and evaluation source collections are summarized in Table 3 and Table 4 respectively.

	Chinese	Spanish	English	All
News	84	82	85	251
DF	63	47	83	193
All	147	129	168	444

Table 3: Total # of Documents in Training Data

	Chinese	Spanish	English	All
News	84	84	82	250
DF	82	83	85	250
All	166	167	167	500

Table 4: Total # of Documents in Evaluation Data

The corpus consists of topic-focused news articles and discussion forum posts published in recent years, topically related comparable (but non-parallel) across languages. LDC human annotators selected documents so that a substantial amount of entities appear across two or three languages. Figure 2 and Figure 3 show that 8.3% and 7.5% coreferential entities are across languages for training data and evaluation data respectively, which provide great opportunities for cross-lingual knowledge transfer (section 6.1). In the future systems may attempt working on streaming news or social media data so they can discover cross-lingual comparable documents automatically.

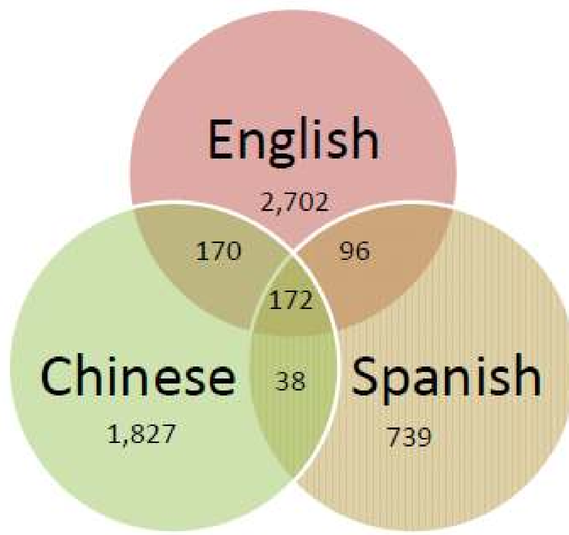

Figure 2: # of Coreferential Entities in Training Data

IBM team reported some annotation errors on overlapping mentions (Sil and Florian, 2015). There are also a few linking annotation errors. For example, in the following post: "*"I am my own man !" This is the phrase you will start hearing today from Jeb Bush as he tries to distance himself from his brother. But is he really his own man?*", the human annotator mistakenly linked the second mention "*man*" to "*George W. Bush*", while top systems correctly linked it to "*Jeb Bush*". LDC will fix these errors and release updated annotation

			Language			Entity and mention type						Genre	
Task	Team	Affiliation	CMN	ENG	SPA	FAC	GPE	LOC	PER	PER/NOM	ORG	NW	DF
EDL	IBM	International Business Machines Corporation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EDL	OSU	Oregon State University	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EDL	RPI_BLENDER	Rensselaer Polytechnic Institute	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EDL	hltcoe	Human Language Technology Center of Excellence	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EDL	BUPTTeam	Beijing University of Post and Telecommunication	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
EDL	CMU_Edvisees	Carnegie Mellon University	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
EDL	HITS	Heidelberg Institute for Theoretical Studies		✓			✓		✓		✓	✓	✓
EDL	UI_CCG	UIUC Cognitive Computation Group			✓	✓	✓	✓	✓		✓	✓	✓
EDL	ZJU_DCD_EDL	Zhejiang University		✓		✓	✓	✓	✓		✓	✓	✓
EDL	sdu	Shandong University	✓						✓			✓	✓
EL	BUPTTeam	Beijing University of Post and Telecommunication	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EL	CMU_Edvisees	Carnegie Mellon University	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EL	IBM	International Business Machines Corporation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EL	RPI_BLENDER	Rensselaer Polytechnic Institute	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EL	GWU	George Washington University		✓		✓	✓	✓	✓		✓	✓	✓
EL	HITS	Heidelberg Institute for Theoretical Studies		✓			✓		✓		✓	✓	✓
EL	SYDNEY	Sydney University		✓					✓				✓
EL	ZJU_DCD_EDL	Zhejiang University		✓					✓			✓	✓
EL	lvic_edl	Vision and Content Engineering Laboratory		✓		✓	✓	✓	✓		✓	✓	✓
EL	sinica	Academia Sinica	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓

Table 2: Runs Submitted by KBP2015 Trilingual Entity Discovery and Linking Participants

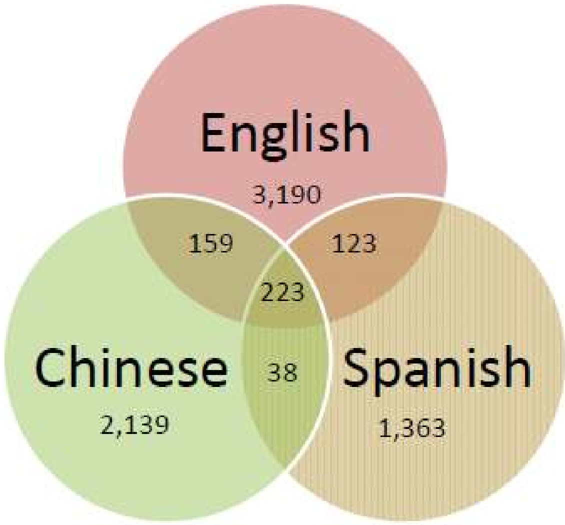

Figure 3: # of Coreferential Entities in Evaluation Data

packages.

Finally, we also devoted a lot of time at collecting related publications and tutorials [2], resources and softwares [3] to lower down the entry cost for EDL.

5 Evaluation Results

5.1 Overall Performance

Table 5 and Table 6 summarize the results of the full EDL and the diagnostic EL tracks respectively. We selected the best run from each system for comparison. For public release purpose we anonymized the team names. English Extraction and Linking is much more difficult than the same task in KBP2014, based on the comparison of the same top systems acros years. Nevertheless, as a new cross-lingual KBP task, the overall results of both EDL and EL are very encouraging. Compared to last year's top mono-lingual English tracks (Ji et al., 2014), Tri-lingual EDL performance (CEAFm) is only 11% lower than mono-lingual English EDL. The best Spanish-to-English Entity Linking CEAFm score is improved from 82.9% last year to 91.2% this year. In fact, cross-lingual setting has provided unique opportunities for cross-lingual inference. Some systems have applied cross-lingual coreference and inference methods to transfer knowledge from one language to another and significantly improved the performance of various components: entity typing, clustering and linking. With effective cross-lingual knowledge transfer RPI system (Hong et al., 2015) was able to achieve top CEAFmC score in Tri-lingual EL, even though it is based on an unsupervised linking algorithm without using any labeled data. More details will be presented in section 6.1.

Comparing the full EDL performance (Figure 4) with the diagnostic EL performance (Figure 5), we can see the best CEAFmC score dropped from 75.3% to 55.1% (system 1), and the best CEAFmC+ score dropped from 72.4% to 59.4% (system 2). We can see that the best Tri-lingual mention extraction F-score is not great: 72.4%; and the highest mention identification recall is 72.7%. This indicates that entity mention extraction is a major challenge for EDL (detailed analysis will be presented in section 7.1). And there is no single system that achieved the best performance at both mention extraction and linking/clustering.

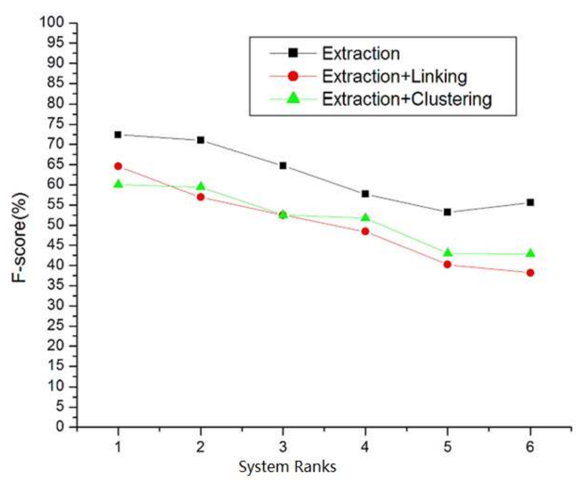

Figure 4: Tri-lingual EDL Performance

5.2 Progress from ACE to KBP Mention Extraction

Mention extraction (Florian et al., 2006; Li et al., 2014; Li and Ji, 2014; Lu and Roth, 2015) has been an unsolved challenge since the Automatic Content Extraction (ACE) program [4]. But we have been making great progress since then. IBM applied their English ACE mention extraction system (Florian et al., 2006) to this year's EDL evaluation data and only obtained 84.4%

[2] http://nlp.cs.rpi.edu/kbp/2015/elreading.html

[3] http://nlp.cs.rpi.edu/kbp/2015/tools.html

[4] http://www.itl.nist.gov/iad/mig/tests/ace/

System	NER			NERC			NERLC			KBIDs			CEAFm			CEAFm-weak			CEAFmC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
TRILINGUAL																					
3	79.7	**72.7**	**76.1**	75.9	**69.3**	**72.4**	**67.7**	**61.8**	**64.6**	63.2	**65.1**	**64.1**	64.6	**58.9**	**61.6**	66.5	**60.7**	63.5	62.9	**57.4**	**60.0**
2	**83.8**	67.0	74.5	**79.8**	63.9	71.0	64.0	51.2	56.9	72.8	54.2	62.1	**69.3**	55.4	**61.6**	**72.2**	57.8	**64.2**	**66.9**	53.5	59.4
11	72.0	69.3	70.6	66.0	63.5	64.7	53.6	51.5	52.5	**74.3**	54.9	63.1	55.9	53.8	54.8	58.5	56.2	57.3	53.5	51.5	52.5
1	79.1	53.6	63.9	71.4	48.4	57.7	59.9	40.6	48.4	58.9	47.8	52.8	68.3	46.2	55.1	71.6	48.5	57.8	64.1	43.4	51.7
4	56.2	63.5	59.7	50.2	56.7	53.2	37.9	42.8	40.2	27.0	41.5	32.7	43.5	49.1	46.1	47.0	53.1	49.9	40.5	45.8	43.0
12	72.7	55.3	62.8	64.2	48.9	55.6	44.2	33.6	38.2	45.5	48.3	46.9	54.7	41.6	47.3	56.9	43.4	49.2	49.5	37.7	42.8
CHINESE																					
3	82.2	**77.7**	**79.9**	79.2	**74.8**	**76.9**	**74.4**	**70.2**	**72.2**	68.1	**67.5**	**67.8**	**78.2**	**73.9**	**76.0**	80.2	**75.8**	**77.9**	**76.1**	**71.9**	**74.0**
11	78.8	68.7	73.4	74.7	65.1	69.6	61.2	53.3	57.0	**83.7**	47.7	60.8	71.0	61.8	66.1	75.2	65.5	70.0	68.2	59.4	63.5
2	**85.1**	57.8	68.8	**81.9**	55.7	66.3	65.5	44.5	53.0	75.1	39.7	52.0	76.8	52.2	62.2	**81.7**	55.5	66.1	74.7	50.8	60.5
4	54.9	67.5	60.6	50.0	61.4	55.1	42.9	52.7	47.3	25.7	46.1	33.0	48.9	60.1	53.9	52.6	64.7	58.0	45.6	56.0	50.3
1	79.4	38.9	52.2	70.1	34.3	46.0	57.9	28.3	38.0	57.6	28.5	38.1	68.0	33.3	44.7	73.9	36.2	48.6	64.8	31.7	42.6
12	76.2	36.9	49.7	72.0	34.9	47.0	45.8	22.2	29.9	65.2	28.4	39.6	52.8	25.6	34.4	55.4	26.8	36.2	50.6	24.5	33.0
13	12.3	10.2	11.2	2.7	2.3	2.5	1.0	0.8	0.9	66.5	44.0	53.0	9.1	7.5	8.2	24.2	20.2	22.0	1.5	1.3	1.4
ENGLISH																					
3	**83.4**	67.5	74.6	**80.0**	64.7	71.5	**69.8**	**56.5**	**62.5**	65.3	65.6	65.4	**76.5**	61.9	**68.4**	**79.2**	64.1	**70.8**	**74.3**	60.1	**66.4**
2	83.2	70.1	**76.1**	79.2	**66.7**	**72.4**	66.8	56.2	61.1	70.2	**65.7**	67.9	72.3	60.9	66.1	75.6	63.7	69.1	69.5	58.6	63.6
11	74.8	**74.0**	74.4	67.1	66.4	66.7	53.1	52.5	52.8	**74.8**	58.8	65.8	65.2	**64.5**	64.8	67.8	**67.1**	67.5	61.3	**60.6**	60.9
1	81.8	64.9	72.4	74.6	59.1	65.9	62.9	49.8	55.6	62.6	64.7	63.6	73.7	58.4	65.2	76.7	60.8	67.8	68.2	54.1	60.3
9	80.5	60.1	68.8	71.8	53.6	61.4	63.4	47.4	54.2	63.8	65.1	64.5	73.9	55.2	63.2	76.2	57.0	65.2	67.1	50.1	57.4
12	77.1	69.1	72.9	67.4	60.5	63.7	50.1	45.0	47.4	48.2	62.9	54.6	63.6	57.1	60.2	66.1	59.3	62.5	56.9	51.0	53.8
5	82.4	68.9	75.0	62.7	52.5	57.1	53.9	45.1	49.1	73.5	64.0	**68.4**	74.8	62.6	68.2	76.8	64.3	70.0	57.6	48.2	52.5
4	53.4	58.9	56.0	46.8	51.7	49.1	31.1	34.3	32.6	24.1	40.2	30.2	42.0	46.3	44.1	46.6	51.4	48.9	38.2	42.2	40.1
SPANISH																					
14	80.1	**77.2**	78.7	76.1	**73.4**	74.7	**70.5**	**67.9**	**69.2**	73.0	**74.0**	**73.5**	75.0	**72.3**	**73.7**	77.2	**74.4**	75.8	72.2	**69.6**	**70.9**
2	**83.4**	76.7	**79.9**	**78.4**	72.2	**75.2**	59.7	54.9	57.2	**76.9**	53.2	62.9	**76.9**	70.7	**73.7**	**79.4**	73.1	**76.1**	**72.9**	67.1	69.9
3	69.4	77.1	73.1	63.7	70.7	67.0	53.9	59.8	56.7	54.7	61.1	57.7	58.4	64.8	61.4	61.4	68.2	64.6	54.9	60.9	57.7
4	67.5	68.1	67.8	60.2	60.8	60.5	45.6	46.0	45.8	39.6	37.9	38.8	54.1	54.5	54.3	56.8	57.3	57.0	49.7	50.1	49.9
1	71.0	52.1	60.1	64.2	47.1	54.3	54.2	39.7	45.9	51.2	41.0	45.5	63.0	46.2	53.3	66.0	48.4	55.9	58.5	42.9	49.5
11	54.7	58.2	56.4	49.8	53.0	51.4	42.7	45.4	44.0	65.6	56.0	60.4	45.2	48.1	46.6	47.3	50.3	48.7	42.8	45.5	44.1
12	58.2	54.3	56.2	48.8	45.6	47.1	27.6	25.8	26.6	32.9	46.7	38.6	47.0	43.9	45.4	51.7	48.2	49.9	40.2	37.5	38.8

Table 5: Overall Entity Discovery and Linking Performance

System	NERC			NERLC			NENC			KBIDs			CEAFm			CEAFmC			CEAFmC+		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
TRILINGUAL																					
1	86.9	**88.1**	**87.5**	71.1	72.2	71.6	76.0	71.3	73.6	65.5	76.4	70.6	81.1	82.3	81.7	74.8	**75.9**	**75.3**	68.2	69.2	68.7
2	85.8	85.9	85.9	**74.1**	**74.2**	**74.2**	69.5	**78.4**	**73.7**	**77.1**	**78.9**	**78.0**	**82.9**	**83.0**	**82.9**	**75.0**	75.1	75.1	**72.3**	**72.4**	**72.4**
3	**89.7**	84.0	86.8	73.5	68.9	71.1	**77.8**	69.6	73.4	62.9	71.0	66.7	73.5	68.9	71.1	70.0	65.6	67.7	65.1	61.0	63.0
4	68.4	66.1	67.2	47.5	45.9	46.7	32.3	71.5	44.5	68.2	40.1	50.5	62.8	60.7	61.7	50.9	49.2	50.0	44.7	43.2	43.9
CHINESE																					
3	**94.4**	**89.4**	**91.8**	**85.4**	**80.9**	**83.1**	**89.6**	81.4	**85.3**	79.7	**76.7**	78.2	**92.9**	88.0	**90.4**	**89.3**	**84.6**	**86.9**	**83.8**	**79.4**	**81.5**
1	89.1	89.1	89.1	78.1	78.1	78.1	73.8	79.7	76.7	72.2	72.6	72.4	89.9	**89.9**	89.9	83.1	83.1	83.1	76.4	76.4	76.4
2	82.8	82.8	82.8	71.6	71.6	71.6	57.9	78.2	66.5	**80.6**	76.2	**78.3**	87.5	87.5	87.5	74.8	74.8	74.8	71.3	71.3	71.3
4	70.6	70.6	70.6	54.9	54.9	54.9	33.6	**82.4**	47.7	77.5	43.7	55.9	76.8	76.8	76.8	60.5	60.5	60.5	53.2	53.2	53.2
6	31.5	31.5	31.5	4.8	4.8	4.8	8.2	13.8	10.3	10.1	12.9	11.3	66.0	65.9	66.0	22.3	22.3	22.3	4.2	4.2	4.2
ENGLISH																					
2	86.8	87.1	**87.0**	**73.6**	**73.8**	**73.7**	73.6	**75.0**	**74.3**	71.5	78.4	74.8	79.9	80.1	80.0	74.2	**74.4**	**74.3**	**70.5**	**70.7**	**70.6**
3	**88.6**	81.3	84.8	69.2	63.5	66.2	**76.9**	58.0	66.1	57.3	72.8	64.2	80.1	73.5	76.6	**75.7**	69.4	72.4	67.7	62.1	64.8
1	85.6	**88.1**	86.8	66.1	68.1	67.1	76.4	64.8	70.1	62.7	79.9	70.3	79.0	**81.4**	**80.2**	71.3	73.5	72.4	63.1	65.0	64.0
5	82.3	76.0	79.0	67.9	62.7	65.2	58.2	64.1	61.0	**73.4**	72.5	72.9	79.3	73.3	76.2	69.7	64.4	67.0	65.1	60.2	62.5
4	66.8	62.0	64.3	40.5	37.5	38.9	29.1	61.2	39.4	61.4	36.2	45.6	65.9	61.1	63.4	49.9	46.3	48.1	38.2	35.4	36.7
7	61.6	56.9	59.1	46.2	42.7	44.4	32.9	66.6	44.1	60.2	28.5	38.7	60.0	55.5	57.7	46.3	42.8	44.5	39.9	36.9	38.3
8	61.6	56.9	59.2	48.6	44.9	46.7	42.4	53.6	47.3	62.6	57.6	60.0	52.0	48.0	49.9	40.3	37.3	38.7	34.1	31.6	32.8
9	38.6	35.6	37.0	27.2	25.2	26.2	75.8	51.5	61.4	68.9	**83.9**	**75.7**	**83.1**	76.9	79.9	30.5	28.2	29.3	26.2	24.2	25.2
10	38.7	17.4	24.0	27.1	12.2	16.9	63.0	29.3	39.9	70.9	33.4	45.4	80.9	36.4	50.2	31.7	14.3	19.7	25.9	11.7	16.1
6	30.5	28.2	29.3	3.2	3.0	3.1	3.9	7.5	5.1	15.0	12.7	13.8	36.3	33.6	34.9	15.6	14.5	15.0	2.1	2.0	2.1
SPANISH																					
2	**88.9**	**88.9**	**88.9**	**80.4**	**80.4**	**80.4**	**80.5**	**87.5**	**83.8**	**84.2**	**83.5**	**83.8**	**91.2**	**91.2**	**91.2**	**83.3**	**83.3**	**83.3**	**79.6**	**79.6**	**79.6**
1	86.1	86.1	86.1	71.5	71.5	71.5	78.5	76.2	77.3	63.9	74.9	69.0	80.6	80.6	80.6	74.5	74.5	74.5	66.6	66.6	66.6
3	83.8	81.0	82.4	62.2	60.1	61.2	66.6	82.7	73.8	55.9	60.1	57.9	72.6	70.1	71.3	66.1	63.8	64.9	57.0	55.0	56.0
4	70.5	70.5	70.5	48.2	48.2	48.2	35.3	85.9	50.1	76.4	36.2	49.2	69.9	69.9	69.9	54.2	54.2	54.2	42.2	42.2	42.2

Table 6: Overall Entity Linking Performance

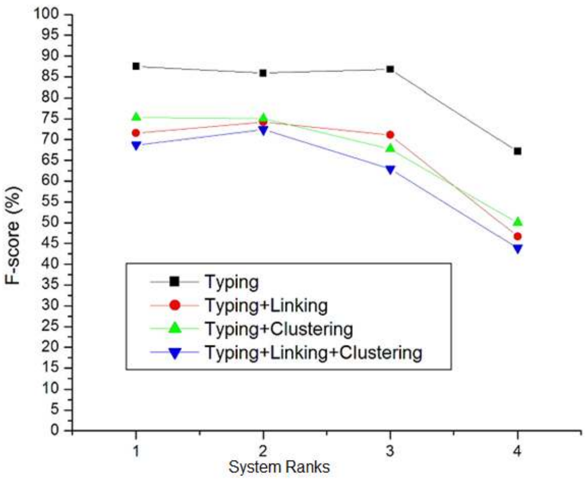

Figure 5: Tri-lingual EL Performance

Precision, 49.9% Recall and 62.7% F-score, significantly lower than the F-score (71.5%) of their KBP2015 EDL system on the same data set. The low recall indicates the traditional mention extractor might be "getting too old" for new names due to the fast language evolution and informal style of discussion forum posts.

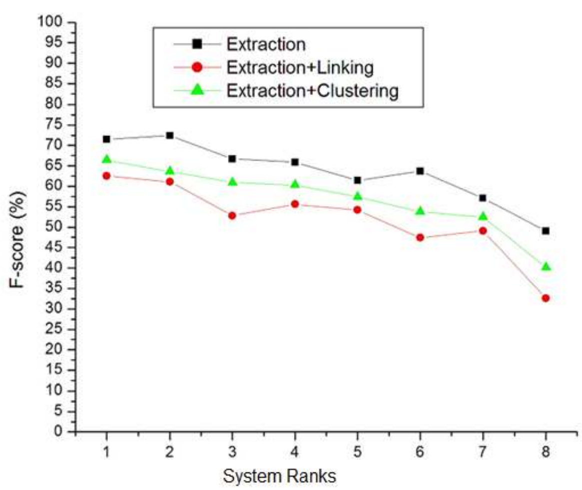

Figure 6: English EDL Performance

5.3 Comparison on Languages

Figures 7, 8, 9, 10 and 11 compare the break down scores for various languages. Overall we don't observe any language is particularly more challenging than the other. There are two basic approaches to cross-lingual EDL: (1) Foreign Language EDL + Entity Translation; and (2) Full Document Machine Translation (MT) + English EDL. Most Chinese-to-English systems adopted approach (1) while most Spanish-to-English systems adopted (2), mainly because Spanish-to-English MT is more mature than Chinese-to-English MT. UI_CCG system (Sammons et al., 2015) reported that the F-score of English name tagging on Spanish-to-English MT output is 10% higher than Spanish name tagging. In contrast, Chinese-to-English MT tends to miss and incorrectly translate many names. We will present more detailed error analysis in section 7.2.

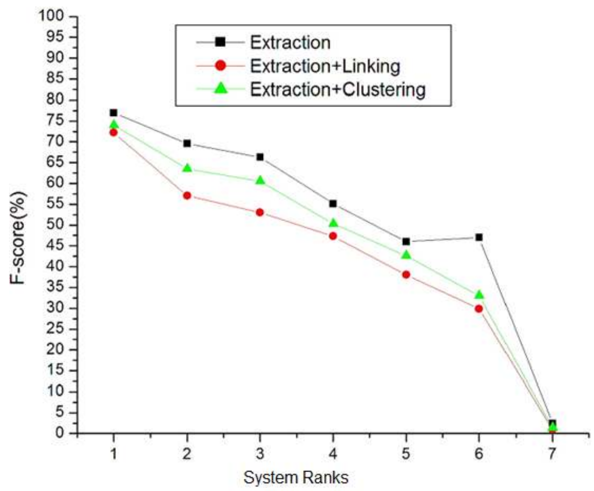

Figure 7: Chinese EDL Performance

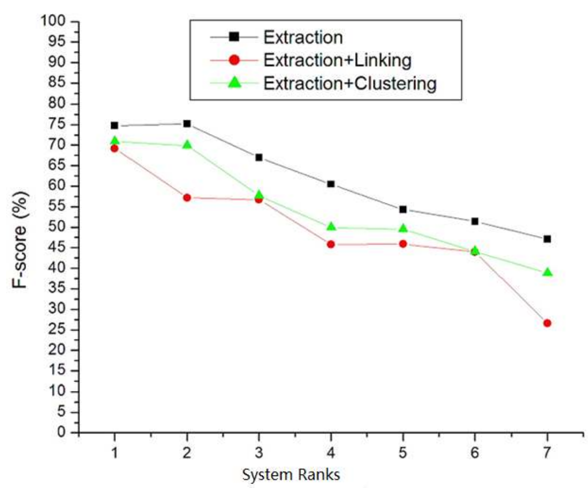

Figure 8: Spanish EDL Performance

5.4 Entity Types and Textual Genres

This year's evaluation introduced new entity types for TAC KBP: FAC, GPE and LOC, as well as nominal mentions of PER entities. Figure 12 and Figure 13 show that overall `NERL` performance closely follows those for the frequent PER and GPE categories. Nominal mention detection is

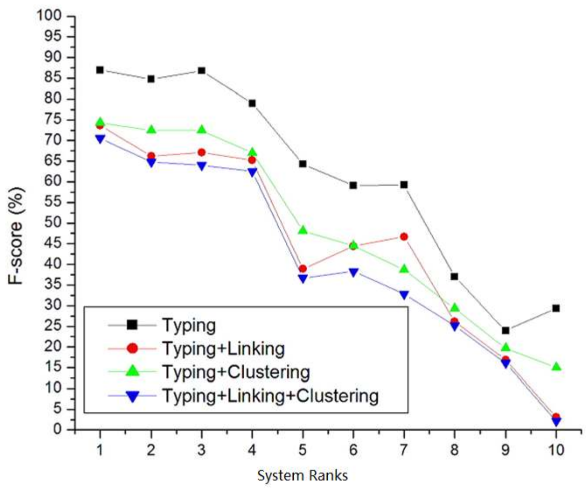

Figure 9: English EL Performance

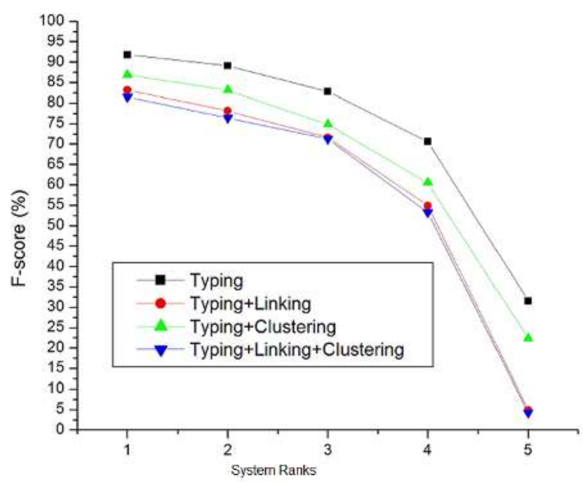

Figure 10: Chinese EL Performance

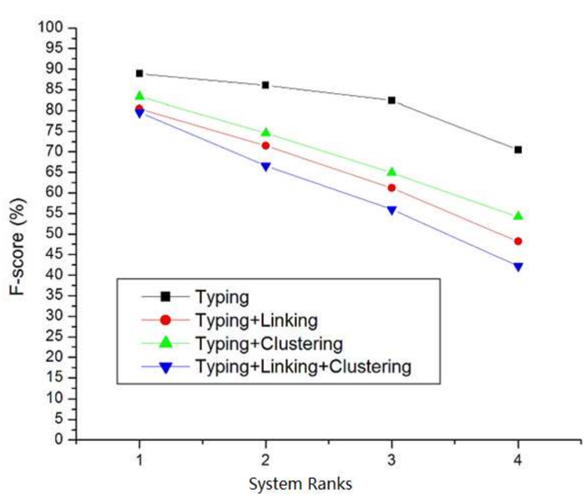

Figure 11: Spanish EL Performance

more challenging than name mention detection, and good nominal detection is not necessary for strong end-to-end performance due to its relatively low popularity in the source collection. RPI System (Hong et al., 2015) designed special heuristic rules and constraints to detect whether a person nominal mention is generic or specific, and link them to KB via within-document coreference resolution (details in section 6.2) and achieved the best score on both nominal mention extraction (Figure 14) and linking (Figure 15). Of the named entities, FAC performance is lowest by far.

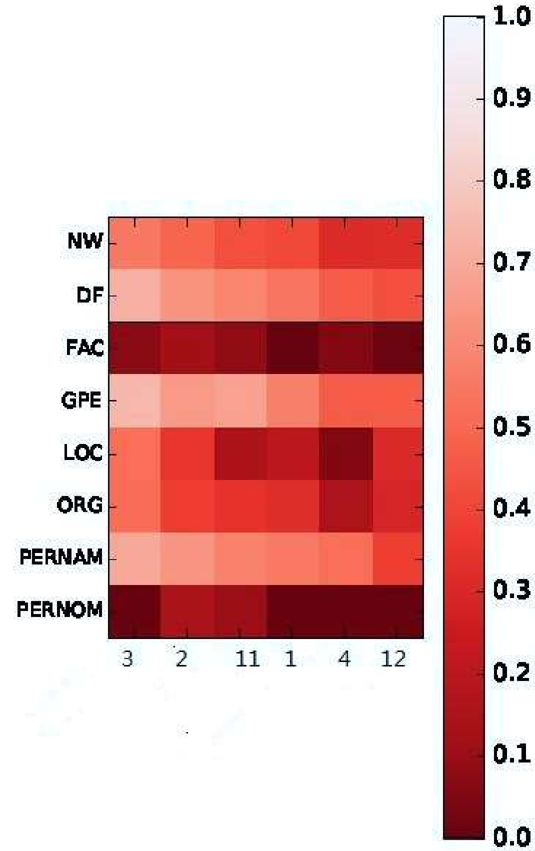

Figure 12: EDL NERL F_1 when selecting a subset of annotations by text genre and entity type. NW = newswire, DF = discussion forum; FAC = facility, GPE = geopolitical entity, LOC = natural location, ORG = organization, PERNAM = person name, PERNOM = person nominal mention.

5.5 NIL and Non-NIL Comparison

Figures 16, 18, 19 and 17 compare the EDL performance of NIL mentions and Non-NIL mentions. Comparing NELC (Link recognition and classification) and NENC (NIL recognition and classification), we can see for English and Chinese, NENC scores are significantly higher, which indicates that NIL mentions are not more

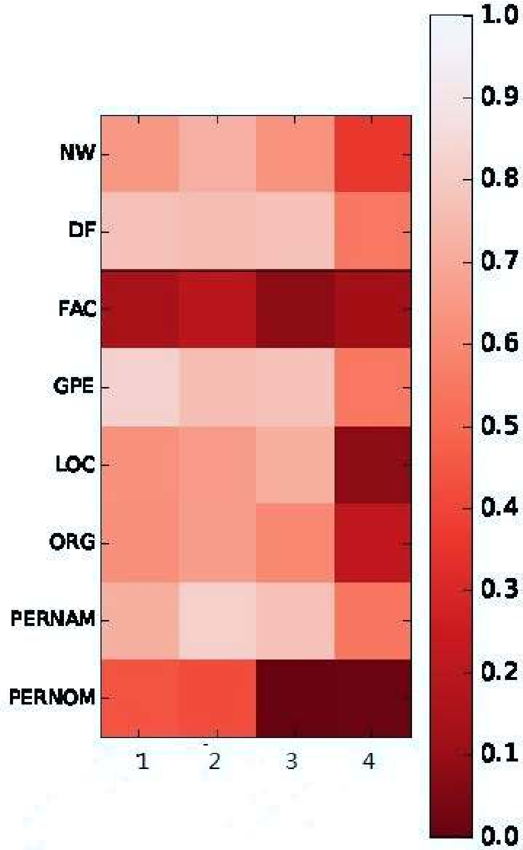

Figure 13: EDL NERL F_1 when selecting a subset of annotations by text genre and entity type. NW = newswire, DF = discussion forum; FAC = facility, GPE = geopolitical entity, LOC = natural location, ORG = organization, PERNAM = person name, PERNOM = person nominal mention.

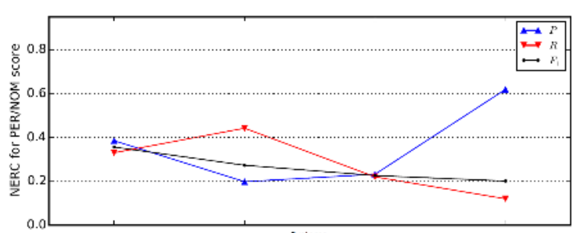

Figure 14: Person Nominal Mention Extraction Performance

difficult to extract than Non-NIL mentions, and the linking requirement brings extra challenges to Non-NIL mentions. In contrast, for Spanish, NIL mentions are more challenging than Non-NIL mentions. Comparing NERL and NERLC scores, we can also see that mention typing accuracy is very high for all three languages.

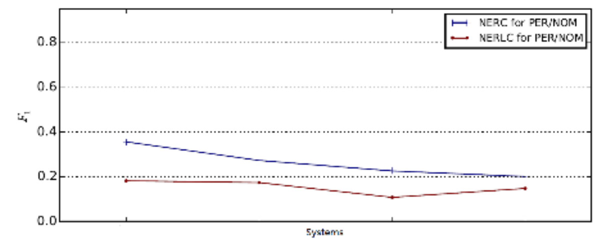

Figure 15: Person Nominal Mention Linking Performance

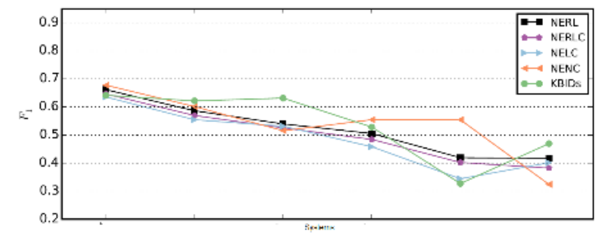

Figure 16: Tri-lingual EDL NIL and Non-NIL Performance Comparison

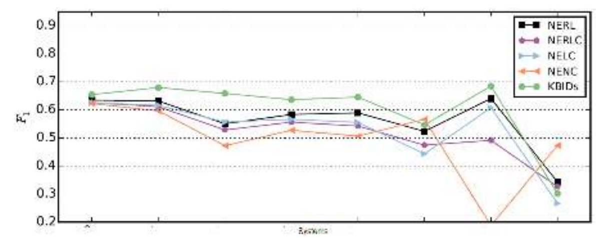

Figure 17: English EDL NIL and Non-NIL Performance Comparison

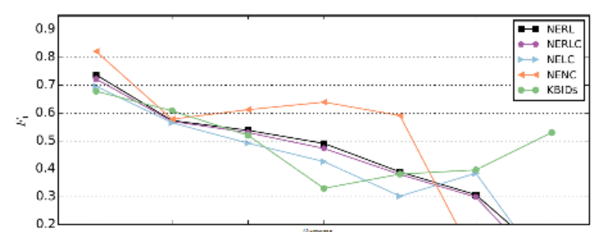

Figure 18: Chinese EDL NIL and Non-NIL Performance Comparison

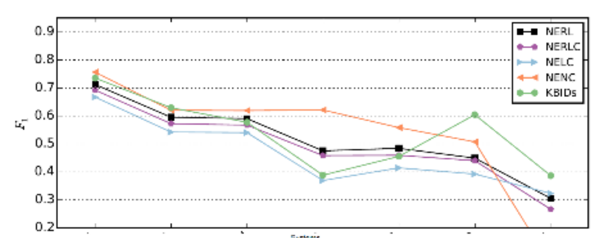

Figure 19: Spanish EDL NIL and Non-NIL Performance Comparison

5.6 Impact of Boundary Match

We evaluated the impact of the partial (weak) mention boundary matching strategy as described in section 2.2.4. The score changes of Tri-lingual EDL are presented in Figure 20. We can see that partial matching can boost up to 13% F-score gain for mention identification, and up to 9% F-score gain for overall mention extraction (identification and classification). The overall mention extraction and linking scores of two systems got a minor decrease (less than 1%) due to the propagation of mention boundary detection errors to linking.

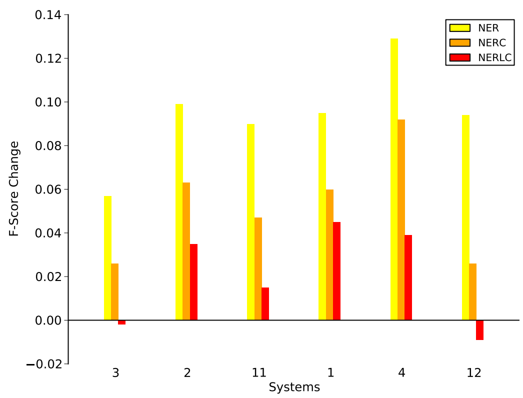

Figure 20: Difference between Strong and Weak Boundary Match Scores

5.7 Linking: Are We Picking the Same Low-Hanging Fruits?

Figures 21, 22, 23 and 24 show the entity linking performance on perfect mentions without considering typing. We can see that the performance of top systems is very encouraging, generally around or above 80% for all three languages.

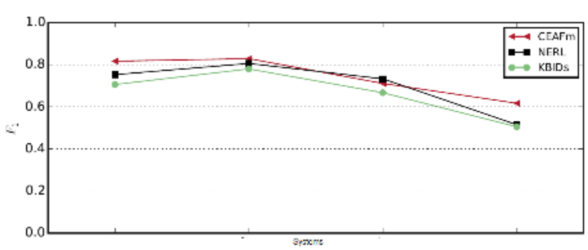

Figure 21: Tri-lingual Entity Linking on Perfect Mentions

Two further questions naturally come up though: (1) Are we solving the same easy problems? and (2) Are we still facing the same

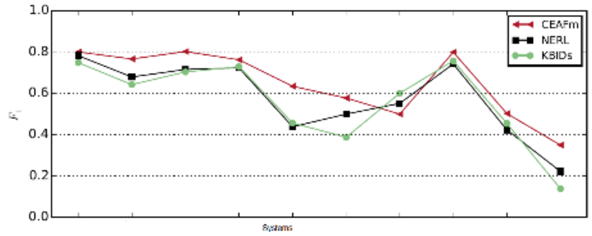

Figure 22: English Entity Linking on Perfect Mentions

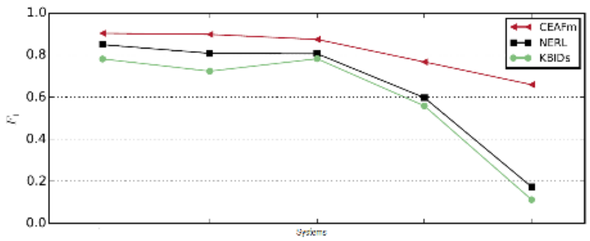

Figure 23: Chinese Entity Linking on Perfect Mentions

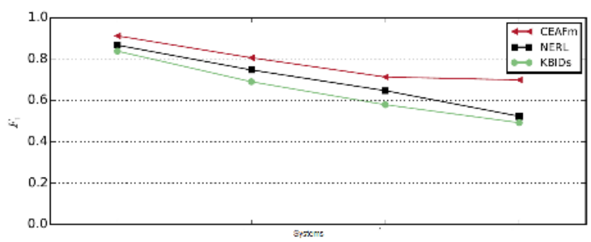

Figure 24: Spanish Entity Linking on Perfect Mentions

challenges? In order to answer these questions, we compare the overlapped instances from top three entity linking systems: RPI, IBM and HITS. Figure 25 and Figure 26 depict the number of overlapped correctly linked mentions and errors among three systems respectively. We can draw the following conclusions from these two figures: "*3/4 happy families resemble each other, while unhappy ones are 9/10 different.*".

Three systems faced the same challenges on the following two cases:

- **Rare entities**. For example, in the following sentence "*The 3 High passes routes - I have seen little information about Kongma La, Cho La and renjo La trails or the lodges either side of Cho La but lodges in the Thame Valley have suffered - Lungden is closed and **Thame** itself has a lot of damage.*", all three systems mistakenly linked "*Thame*" to the town in Oxfordshire instead of the

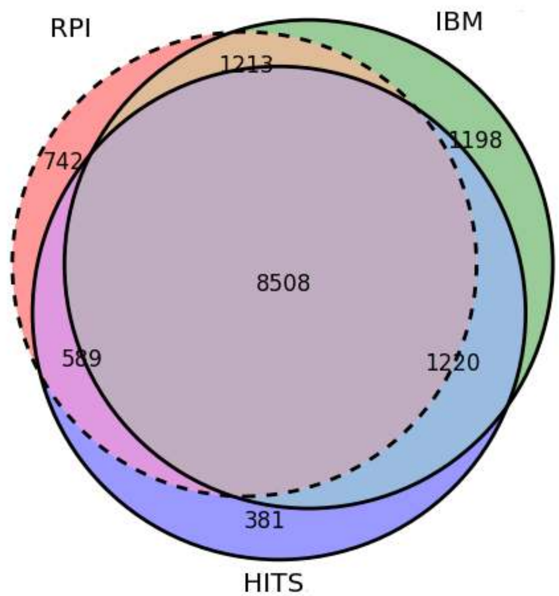

Figure 25: Overlapped Correctly Linked Mentions among RPI, IBM and HITS

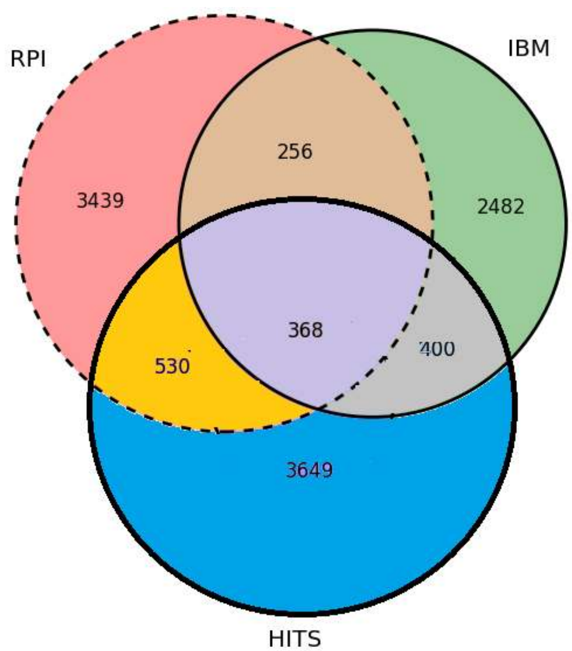

Figure 26: Overlapped Linking Errors among RPI, IBM and HITS

correct village in Nepal, because most of other mentions in the context are NILs.

- **Popularity Bias**. Without knowing the global topic knowledge about a mention, all three systems tend to mistakenly link it to a more popular entity. For example, in the following sentence "*Between the historicity of Clinton's campaign, the Dems' structural advantage in the Electoral College, and Clinton's advantage with low-information voters, I think a lot of things out of his control would have to break right for **Bush**.*", all three systems mistakenly linked "*Bush*" to "*George W. Bush*" instead of "*Jeb Bush*".

- **World Knowledge**. In the following sentence "*The whole corruption scandal has been ongoing for decades, its just conspicuous that "suddenly" now the FBI and IRS decided to clamp down on the irregularities now that FIFA did not abide to **Washington**'s pressures.*", all three systems mistakenly linked "*Washington*" to the city instead of the correct entity "*US government*", due to the lack of knowledge that a country's capital is often used to refer to its government.

Table 7 compares the overlapped instances between RPI and IBM system outputs for three languages. We can see that two systems share the most similar strength and weakness for Chinese because of the unique challenges on morph resolution and name translation. In contrast two systems differ the most for English because RPI's linker is unsupervised and IBM's linker is supervised. RPI's entity linker tries to avoid using labeled training data and relies only on co-occurred mentions for collective inference. This method has some advantage when indicative entity mentions appear in local contexts. For example, In the following sentence "*This is more than the emergency needs of the UN ' s World Fund Program (**WFP**) in the country, but we ' re just talking about food here.*", RPI linker successfully linked "*WFP*" to its correct KB entry "*World Fund Program*" which appears right before the mention. In contrast, IBM linker mistakenly linked it to "*World Food Programme*" due to the noise introduced from other contextual words such as "*food*". On the other hand, IBM linker benefited from its richer contextual feature extraction and supervised model for other cases. For example, in the following sentence "*It's being reported that of the 21 people reportedly advising Jeb Bush, 19 are veterans of the first **Bush** administration, the second **Bush** administration, or in a few cases, both.*", IBM linker successfully linked "*Bush*" in "*the first Bush administration*" to "*George H. Bush*", and "*WFP*" in "*the second Bush administration*" to "*George W. Bush*", while

RPI linker mistakenly linked both of them to "*Jeb Bush*" which was the only contextual mention used for collective inference. Most systems including HITS often make mistakes on mentions with ambiguous entity types. For example, in the following sentence "*A statement from the **White House** also condemned the attack "in the strongest possible terms"*.", HITS system mistakenly linked "*White House*" to the facility entry in the KB. Due to the joint linking and typing model, both RPI and IBM systems successfully linked it to the organization entity "*Executive Office of the President of the United States*".

To summarize, top systems are facing different challenges, and have developed complementary techniques. In the future we may consider assembling multiple linkers through effective system combination and validation, similar to the slot filling validation task (Ji et al., 2010; Ji et al., 2011b).

	Chinese	Spanish	English	All
Correct	82.0	77.1	72.2	76.2
Error	18.2	9.9	8.5	11.3

Table 7: % of Overlapped Instances (Overlap/Union) between RPI and IBM EL system outputs

6 What's New and What Works

6.1 Cross-lingual Knowledge Linking, Inference and Transfer

State-of-the-art EDL methods rely on entity profiling and collective inference (Ji et al., 2014). In high-resource languages like English, we can use some advanced knowledge representations such as Abstract Meaning Representation (AMR) (Banarescu et al., 2013) to effectively select semantic neighbors for entity profiling and collaborators for collective inference (Pan et al., 2015). However, such representations are not available for low-resource languages. Moreover, generally other Natural Language Processing (NLP) tools such as dependency parsers in foreign languages also perform worse than their counterparts in English. As a result, recent work on foreign language entity linking attempted to shifted the focus on avoid excessive linguistic analysis on the source documents and fully leverage KB structure (Wang et al., 2015). Fortunately, this new Tri-lingual EDL task provides new and unique opportunities for cross-lingual knowledge transfer via entity translation, linking and inference. We could build cross-lingual links via name translation in the source collection, and utilize the existing cross-lingual links between KBs, and then transfer knowledge from a high-resource language to a low-resource language.

Figures 27 and 28 illustrate a motivating example. From the Chinese source collection we might not have enough resources to conduct deep understanding and thus can only extract co-occurrence based knowledge graphs, which would not be sufficient to correctly link two mentions "罗姆尼" and "保罗" to their referent entities "*Mitt Romney*" and "*Ron Paul*" respectively in the English KB. But if we can align these two mentions with two other mentions "*Romney*" and "*Paul*" in English source documents by name translation, then we can use the rich knowledge representation from English documents (generated by AMR parser in this example) to infer the links to the KB, because "*Romney*" and "*Paul*" are connected by a conjunction relation in the source so they can be used as collaborators for collective inference; "*Romney*" and "*Mitt Romney*" share many neighbor nodes, and "*Paul*" and "*Ron Paul*" also share many neighbor nodes. Furthermore, the rich hyperlinks and cross-lingual links in multi-lingual KBs can help jointly confirm the linking decisions. RPI system (Hong et al., 2015) developed a joint mention extraction, translation and linking model, which enhanced the quality of mention boundary identification, typing, translation and linking simultaneously. HLT-COE system (Finin et al., 2015) also found this cross-lingual coreference and inference approach can greatly decrease the number of candidate clusters and thus reduce ambiguity.

Let's look at another challenging example below. From the Chinese discussion forum post itself, it would be almost impossible to resolve the morphed mention "刀锋战士*(Knife Warrior)*". However, if we can link the post to its topically related Chinese news article first, then we can easily infer that "刀锋战士*(Knife Warrior)*" refers to "皮斯托瑞斯*(Pistorius)*" and should be linked to the South African sprint runner "*Oscar Pistorius*".

- **Chinese discussion forum post: 刀锋战士**

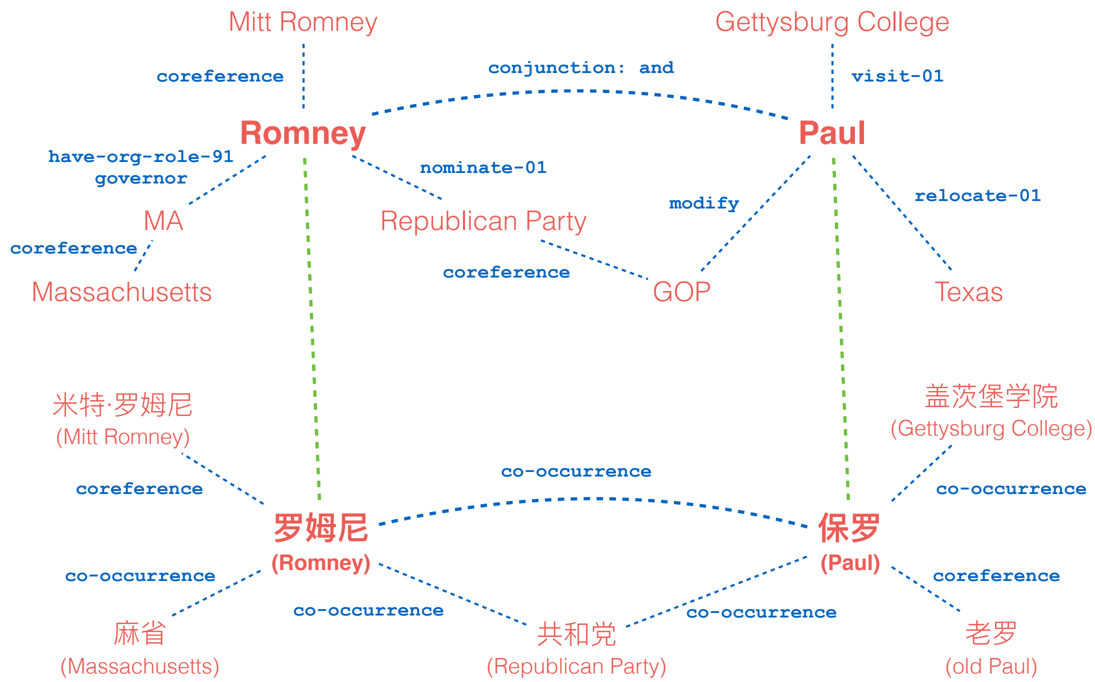

Figure 27: Multi-lingual Knowledge Graphs from the Source Collection

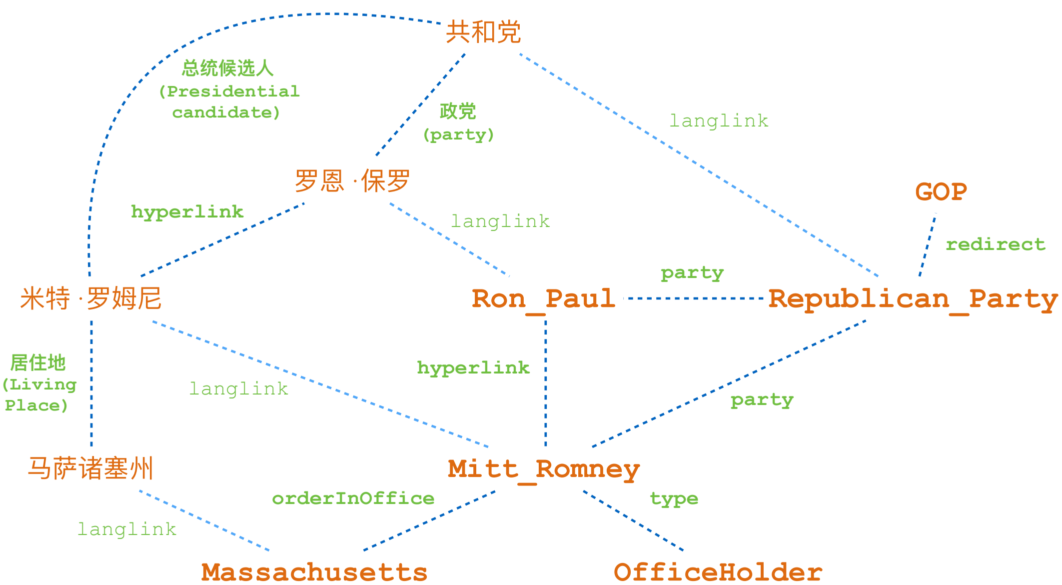

Figure 28: Multi-lingual Knowledge Graphs from multi-lingual KBs

醒来发现女友不在身旁...**刀锋战士** 以过失杀人罪被判处5年监禁 *(Knife Warrior found his girl friend was not beside him when he woke up...Knife Warrior received a five year prison sentence.)*

- **Chinese news document:** 南非比勒陀利亚高等法院*(The Pretoria High Court)21*日宣布判处南非残疾运动员**皮斯托瑞斯***(**Oscar Pistorius**)5*年有期徒刑。皮斯托瑞斯是南非著名的残疾人田径选手，有”**刀锋战士**”之称。*(On the 21th, the Pretoria High Court of South Africa sentenced the disabled sportsman Oscar Pistorius 5 year prison. Pistorius is a famous runner in South Africa, also named as ”Knife Warrior”.)*

Figure 29 verifies this hypothesis. We can see that the linking recall scores for entities across three languages are much higher (up to 38% difference) than entities that only appear in one language, because the linkers benefit from the enriched contexts and KB properties across languages and cross-lingual inference and propagation for linking decisions. For cross-document NIL mention clustering, the recall scores of cross-lingual entities are lower than mono-lingual entities due to the extra errors from name translation.

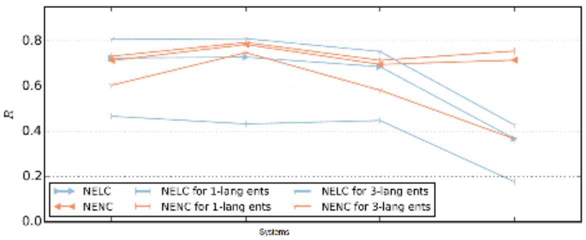

Figure 29: Recall Comparison on Mono-lingual and Cross-lingual Entities

6.2 Tackling Nominal Mentions

This year, a EDL system is also required to extract person nominal mentions referring to specific, real-world individual entities. There are two major challenges as follows.

(1) How to distinguish specific and generic nominal mentions. State-of-the-art three-class realis classification accuracy for events is still below 50% (Hong et al., 2015). Similarly, we often need to analyze the entire context sentence to determine whether a nominal entity mention is generic or specific. RPI system (Hong et al., 2015) encoded heuristic rules to detect the following categories:

(a) Hypothetical: can be detected using keywords such as indefinite articles (e.g., “a/an”) and conditional conjunctions (e.g., “if”). For example, the nominal mentions in the following sentences are all generic:

- Apparently, Bibi *assumes* that the next US **President** *will* see things much more his way than does Barack Obama. *If* H is the nominee, he may be right.

- because it is so widely *presumed* that H is the inevitable **candidate**.

- And *if* something happened to Hillary he *could* become **president**.

(b) Subjective Mood: can be detected by discourse structures. For example, “Vice President” in the following sentence is generic:

- while President Hillary Clinton lived in the White House ... **Vice President** Bill C *would* reside at N.

(c) Generic Referent: many cases can be detected by keywords such as “should” and “a”:

- instead it ’ s the US **President** who *should* take the abuse and apologize for having *a* backbone.

However, some cases need background knowledge to infer a nominal mention is generic. For example, only if we know the following sentence is quoted from a legal document, we can infer that “President” is a generic mention:

- ”which lists they shall sign and certify, and transmit sealed to the seat of the government of the United States, directed to the **President** of the Senate.”

This approach achieved about 46% F-score for person nominal mention extraction from English training data.

(2) How to link or cluster these nominal mentions. In order to link a nominal mention to KB, or assign it to a NIL cluster), the most effective approach is to apply within-document coreference resolution to resolve it to a name mention. Although state-of-the-art within-document coreference resolution performance for nominal mentions is still quite low, linking each identified person nominal mention to its closest person name mention can yield 67% accuracy (Hong et al., 2015).

6.3 Global Knowledge Derived from KB

Due to the lack of resources in foreign languages, Tri-lingual EDL systems have been trying to be more creative at utilizing global knowledge in the English KB. For example, RPI system (Hong et al., 2015) leveraged knowledge graph based embeddings, developed an entropy based measure to quantify the discriminative power of each link type in the KB. IBM system (Sil and Florian, 2015) proposed a global feature based on computing pointwise mutual information for the Wikipedia categories of consecutive pairs of entities, and significantly enhanced linking accuracy.

In addition, many systems including HITS (Heinzerling et al., 2015), LVIC (Besancon et al., 2015), IBM (Sil and Florian, 2015) and RPI (Hong et al., 2015) utilized entity linking results as feedback to map the entity properties in KB to one of the five types defined in KBP. RPI system utilized Abstract Meaning Representation (AMR) corpus (Banarescu et al., 2013) which contains over 100 fine-grained entity types and human annotated KB titles. DBPedia[5] also provides rich types for each page. Therefore, they generated a mapping table between AMR type and DBPedia `rdf:type` (e.g., *university* - `TechnicalUniversitiesAndColleges`). The typing F1 score is 93.2% for perfect mentions in English training data. HITS system (Heinzerling et al., 2015) removed all sense annotations whose entity types from name taggers don't agree with the rdf:type in KB.

6.4 Name Translation Mining

In the pilot Tri-lingual EDL study conducted in the DEFT program earlier this year, we found that a pipeline of foreign language ED + Name Translation + Linking is more effective than full document Machine Translation + English EDL. Therefore, Name Translation becomes a critical component. Despite that many challenges still remain for name translation (section 6), teams have been developing creative approaches to automatically mine name translation pairs. For example, RPI system has developed a novel unsupervised approach to construct and align bursty knowledge graphs from multi-lingual streaming data, incorporating novel criteria based on multi-dimensional clues from pronunciation, translation, burst, neighbor and graph topological structure. This approach was able to mine many high-quality name translation pairs.

Many Chinese news documents include name translation pairs in the parenthesis pattern. For example, we can mine from many pairs from the news document below: "鲍里斯·*Y*·涅姆佐夫*(Boris Y. Nemtsov)*", "俄罗斯内政部*(Interior Ministry)*", "鲍里斯·*N*·叶利钦*(Boris N. Yeltsin)*", "弗拉基米尔·*V*·普京*(Vladimir V. Putin)*", "国际文传电讯社*(Interfax)*", "对话*(Sobesednik)*", "福布斯*(Forbes)*", "保罗·克列布尼科夫*(Paul Khlebnikov)*", "安娜·波利特科夫斯卡娅*(Anna Politkovskaya)*" and "纳塔利娅·埃斯蒂米洛娃*(Natalya Estemirova)*" using the methods described in previous work (Lin et al., 2008; Ji et al., 2009).

- 周五深夜，俄罗斯知名反对派领袖**鲍里斯·*Y*·涅姆佐夫*(Boris Y. Nemtsov)***在莫斯科市中心遭枪击身亡，遇难地距离克里姆林宫的围墙几步之遥。周六凌晨*1*点刚过不久，**俄罗斯内政部*(Interior Ministry)***证实了涅姆佐夫遇刺一事。享年*55*岁的涅姆佐夫曾是**鲍里斯·*N*·叶利钦*(Boris N. Yeltsin)***的第一副总理，后参与组织反对**弗拉基米尔·*V*·普京*(Vladimir V. Putin)***的示威活动。**国际文传电讯社*(Interfax)***援引一名警方知情人士称，这似乎是一次买凶杀人。本月早些时候，涅姆佐夫曾告诉《**对话**》***(Sobesednik)***杂志，他想过俄罗斯总统**弗拉基米尔·*V*·普京*(Vladimir V. Putin)***可能会杀掉自己，不过，他似乎并没有把这个想法当回事。他说他的母亲比他忧虑的多。*2004*年，《**福布斯**》***(Forbes)***杂志的**保罗·克列布尼科夫*(Paul Khlebnikov)***遭枪击身亡；*2006*年，以反对车臣 战争的激烈论战而闻名的**安娜·波利特科夫斯卡娅*(Anna Politkovskaya)***遭到枪杀。*2009*年，**纳塔利娅·埃斯蒂米洛娃*(Natalya Estemirova)***在北高加索遭遇绑架，并被枪击身亡。

6.5 Fine-grained Entity Typing

Adding new entity types into KBP2015 has triggered some new research efforts on fine-grained entity typing. RPI system took one step further to discover new entity types automatically. They start from learning general embeddings for each entity mention, compose the embeddings of specific contexts using linguistic structures, link the mention to knowledge bases and learn its related knowledge representations. Then they developed a novel joint hierarchical clustering and linking algorithm to type all mentions using these representations. The types of all entity mentions are automatically discovered based on a set of clusters, which can capture fine-grained types customized for any input corpus. This framework doesn't rely on any annotated data, predefined typing schema, or hand-crafted features, therefore it can be quickly adapted to a new domain, genre and language. For example, RPI Tri-lingual EDL system can be easily adapted to the biomedical domain, by replacing Freebase with biomedical

[5]http://dbpedia.org

ontologies. For example, given a sentence "The effects of the **MEK** inhibitor on total **HER2** , **HER3** and on phosphorylated **pHER3** were dose dependent.", it can link "**HER2**" to "`ERBB2`" in BioPortal and extract the type '*Proto-Oncogenes→Oncogenes→Genes→Genome Components→Genome→Phenomena and Processes*' as the type for this entity mention.

6.6 Portability for a New Language

Adding foreign languages into the EDL task has shifted some research focus from quality to portability. Supervised learning usually produces better Entity Linking results than unsupervised learning (Ji et al., 2011a). However, they suffered from the high cost of large-scale manual annotation - 90% linking accuracy requires about 20,000 query mentions for training. This year more systems have been seeking new methods for rapid low-cost development or adaption of EDL techniques to a new language. For example, RPI's unsupervised entity typing and linking framework (Hong et al., 2015) was able to apply a new surprise language overnight. The mention-level typing accuracy with perfect boundary is very promising: 85.42% for Hausa and 72.26% for Yoruba. IBM system (Sil and Florian, 2015) trained their linking component from English Wikipedia and thus it can be adapted to new languages without re-training.

7 Remaining Challenges

7.1 Mention Extraction

Regardless of the progress since ACE, the performance of mention extraction is still not satisfactory. The best KBP2015 Tri-lingual mention extraction system's F-score is 72.4%, and the best English mention extraction F-score is 76.1%. In the following we will highlight some remaining major challenges across top systems.

7.1.1 OOV Mention Identification

The highest mention identification recall is only 72.7%. Some English error types include:

- **Informal Nominals**: Most missing errors are caused by the mentions that rarely appeared in traditional newswire training data. For example, in a discussion forum post "*OK, what in cheney's background makes you think he's the* ***puppeteer*** *and bush is the puppet?*", the person nominal mention "puppeteer" doesn't appear in English training data so it's very difficult for a mention detector to identify it unless a good coreference resolver can be applied to link it to "he". Likewise "mastermind" in the following sentence is a person nominal mention and can be identified by resolving it to "Mohamed Mohamud": "*Kenyan enyan authorities also put a $220,000 (200,000 euro) bounty on the head of the alleged* ***mastermind*** *of the attack, Mohamed Mohamud, - also known as Dulyadin Gamadhere - who is believed to be in Somilia.*". Probably it's worth re-visiting some previous work about joint inference between coreference and name tagging (Ji et al., 2005).

- **Abbreviations**: Abbreviations, especially single-letter ones, are difficult to identify. For example, in a discussion forum post "*Even the Bush family knew W needed help moving his lips.*", "*W*" should be identified as a person name that refers to George W. Bush.

- **Names embedded in URLs**: It's also difficult to identify uncommon names from URLs, such as "netanyahus" in "*http://news.yahoo.com/liberal-israelis-**netanyahus**-win-reality-check -115401998.html*". Perhaps next year we should remove this requirement

Typical Chinese error types include:

- **Code-switch**: Chinese documents often include English names, which need to be identified by specific patterns. For example, many systems failed to identify the organization name "*Dahabshiil*" from the following sentence: "*Dahabshiil*就是名单上一家汇款公司，它的业务涉及整个非洲之角地区。".

- **Ancient entities**: Chinese discussion forum posts often quote ancient person names mentioned in a document written in classical Chinese. For example, the following discussion forum post tries to explain "revolution" by quoting a sentence from a book *I ching* written during 1000750 BC: "详细解释：出处及演变"革命"一词的古义是变革天命，最早见于《周易·革卦·象传》："天地革而四时成，**汤**

武革命，顺乎天而应乎人。” *(Detailed explanation: the origin and evolution of the word “revolution” are from its ancient meaning, change destiny, which appeared earliest in ”I ching. Gegua. Tuanzhuan”: ”The timing is right for the revolution of the Heaven and the Earth, Tang against Wu's revolutions obey the will of Heaven and be in harmony with men.”)*”, from this quoted sentence a name tagger needs to identify “汤 (Tang)” and “武 (Wu)” as person names, referring to two historical figures “商汤 (Shang Tang)” and “周武王(Zhou Wu King)” respectively.

- **Morphs**: In the KBP2014 EDL overview paper (Ji et al., 2014) we pointed out a unique challenge from Chinese discussion forum - Entity ”Morphs”, a special type of fake alternative names, to achieve certain communication goals such as expressing strong sentiment or evading censors. This year we extended the Chinese track from Entity Linking to Entity Discovery and Linking. Compared to linking, it's even more difficult to identify these morph mentions from texts (Zhang et al., 2015). In KBP2015 EDL training data, about 16% mentions are morphs. For example, in a discussion forum post “所以，我觉得，乡港人还是太没有气魄了，只占领*中*环算什么？ *(So I think Hong Kong people are still not brave, what's the big deal to occupy the Central?)*”, “乡港 (Xiang Gang)” is a morph referring to “香港 (Hong Kong)”, and “*中*环 (Central)” is another morph referring to “*中*环 (Central)”. It requires an EDL system to incorporate new techniques such as new name discovery or a morph identified trained specifically from a large amount of Chinese social media data.

7.1.2 Boundary Errors due to Informal Contexts

Compared to newswire, the top EDL systems made a lot more mention boundary errors on discussion forum due to its informal nature. For example, one of the top systems mistakenly identified “*Pres Obama*” as a person name from “*while Pres Obama and Holder are willing to give them a pass.*” Some effective text normalization techniques might be worth adding before mention extraction, e.g., to expand “*Pres*” to “*President*”. In another post: “*Let's give Obama a THIRD TERM as President Have Pelosi and Reid introduce a Bill to bomb some place that republicans hate.*”, many systems mistakenly identified the informally capitalized word “*Have*” as part of an incorrect name mention “*Have Pelosi*”.

7.1.3 Joint Linking and Typing: When to Trust Whom More?

Despite the great success achieved by joint modeling of entity typing and linking, there is still no clear and elegant solution to automatically decide when to trust typing and when to trust linking. Most EDL systems chose to use linking feedback to override typing results from name tagging, with the assumption that linking accuracy is higher than typing. However, this strategy may introduce errors for some highly ambiguous mentions, especially in discussion forum posts. For example, in the following sentence “*His stands on fracking, the TPP, indefinite detention, the Patriot Act, torture, war crimes, drone killings, persecution of whistle-blowers and journalists, etc., preclude him from being a ”Populist President”.*”, “*TPP*” refers to “*Trans-Pacific Partnership*”. However, most top EDL systems mistakenly linked it to “*Comandante FAP Guillermo del Castillo Paredes Airport (IATA: TPP)*” and thus labeled it as a facility. Therefore the current joint inference approaches still need to be improved by integrating more reliable confidence estimation.

7.2 Machine Translation and Entity Translation

Most Spanish-to-English EDL systems used Google Translation service or Bing Translation service to translate Spanish documents into English and then apply English EDL to the MT output. Compared to Spanish-to-English Machine Translation (MT), state-of-the-art Chinese-to-English MT performance is still not satisfactory. For example, the following shows the translation results from various MT systems:

- **Chinese**: 经过一个展厅，从条幅上看，写的是敬贺**本焕**老和尚百年寿辰书画展，这才知道老和尚生于1907年，已经年满百岁。

- **Reference Translation**: After passing through an exhibition hall, from the banner, we could see that it was a painting and calligraphy exhibition to celebrate the old monk **Ben Huan**'s centennial birthday. Not until then we knew that this old monk was born in 1907 and so he is already 100 years old.

- **Research MT System 1**: After a Gallery, from banner, write the Worship Celebration, the old monk 100 years birthday dynasties that old monk who was born in 1907, has over 100 years old.

- **Research MT System 2**: after an exhibition hall, from the point of view of scrolls, writing is **hwan** old monk centennial birthday of painting and calligraphy, he did not know it until this old monk was born in 1907 years, is already 100 years old.

- **Google MT**: After an exhibition hall, from banners point of view, this is written, **He Jing Huan** birthday centuries old monk painting and calligraphy exhibition, to realize that the old monk born in 1907 , has over a hundred .

Both research MT systems are state-of-the-art statistical models trained from 2 million sentence pairs. We can see that even though the general topics of this sentence are correctly translated, none of these three MT systems translated the person name "本焕(Ben Huan)" correctly. Our previous analysis (Ji et al., 2009) showed that a state-of-the-art Chinese-to-English MT system mis-translated about 40% person names.

Therefore, most Chinese-to-English EDL systems have chosen an alternative pipeline: Chinese EDL + Entity Translation. Unfortunately, Chinese-to-English Entity Translation is not a trivial task either. Our KBP2011 overview paper (Ji et al., 2011a) presented the detailed error distributions for Entity Translation. Many of these challenges still remain, especially when the Chinese name is composed based on meanings while the English translation is based on pronunciation. For example, "上海国际化自贸区" should be translated and linked to "*Causeway Bay*" in the English KB, but its literal translation "*Shanghai International Free Trade Zone*" doesn't share any words with the official translation. We need to improve name structure parsing, and develop some automatic mechanism to decide when to translate or transliterate a name or a name component. Some systems such as RPI (Hong et al., 2015) have developed better name transliteration models. But these models usually rely on a large amount of name pairs, which might not be adaptable for other low-resource foreign languages. It would be more promising to focus on automatic name translation mining techniques.

7.3 Entity Linking

7.3.1 Abbreviations and Nicknames

In Chinese, GPE names are often abbreviated as single characters, such as "美" for "美国*(United States)*" and "中" for "中国*(China)*". These single characters are highly ambiguous in various contexts. For example, "美" can also be a common adjective which means "*beautiful*".

For many abbreviations and nicknames, an EDL system usually should assign more weight to context similarity than popularity. For example, in the following post "香港人都支持占中*?(Do all of Hong Kong people support Occupy the Central?)*", the GPE mention "中" should be linked to "*Central*", the central business district of Hong Kong, instead of the more popular entry "*China*". And in "*First, as much as Bibi says Israel will go it alone if it has to, I doubt that it would/could.*", we should link "*Bibi*" should be linked to the Israeli politician "*Benjamin Netanyahu*", instead of the more popular Swedish actress "*Bibi Andersson*".

7.3.2 Re-visit "Collaborators" in Collective Inference

State-of-the-art Entity Linking techniques rely on the idea of "*Collective Inference*", where a set of *mention collaborators* are linked simultaneously by choosing an "optimal" or maximally "coherent" set of named entity targets - one target entity for each mention in the coherent set. Many existing methods extract"collaborators" based on co-occurrence (Ratinov et al., 2011; Wang et al., 2015), topic relevance (Cassidy et al., 2012), social distance (Cassidy et al., 2012; Huang et al., 2014), dependency relation (Ling et al., 2014), or a combination of these through meta-paths (Huang et al., 2014). However, two mentions are qualified as collaborators for collective inference not because they are involved in a syntactic structure. Rather it's because they are often involved in some specific types of relations and events. Some other work tried to restrict collaborators that bear a specific type of relation (Cheng and Roth, 2013). But high-quality relation/event extraction (e.g., ACE) is limited to a fixed set of pre-defined types. One potential solution is to construct a background knowledge base in a never-ending way to gather relations and events for each entity in real time, then we can infer collaborators from this background knowledge base.

7.3.3 Knowledge Representation

Recent advances in rich knowledge representations such as AMR have greatly promoted unsupervised entity linking (Pan et al., 2015). However, in some cases even AMR cannot capture implicit relations among entities and concepts. For example, in the following sentence "*The Stockholm Institute stated that 23 of 25 major armed conflicts in the world in 2000 occurred in impoverished nations.*", the concept "*armed*" is crucial to determine that "*The Stockholm Institute*" should be linked to "*Stockholm International Peace Research Institute*" instead of "*Stockholm Institute of Education*". However, the AMR graph for this sentence (as depicted in Figure 30) cannot capture the semantic connections between "*The Stockholm Institute*" and "*armed*". Entity Linking is likely to benefit from adding even richer types of nodes and edges, and Cross-sentence nominal and pronoun coreference resolution into AMR.

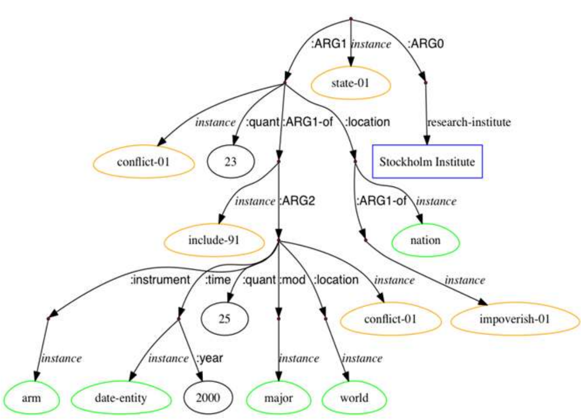

Figure 30: AMR graph for the example sentence.

7.3.4 Background Knowledge

Most discussion forum posters assume the readers already know the on-topic entities and events, and thus they don't bother to elaborate the background for these target entities. Also they tend to use short and informal mentions for efficient discussions. As a result, directly comparing the contexts of a mention in a post with a candidate KB entity's text description is often not sufficient. For example, in order to figure out that "*Gulf*" in the following post "*I went to youtube and checked out the **Gulf** oil crisis: all of the posts are one month old, or older*···" refers to "*Gulf of Mexico*", we need to know it suffered a catastrophic explosion. For the following post "*What words about Bush, Cruz, Romney, Carson, Walker?*", we first need to know it's about presidential election, then link "*Cruz*" to "*Ted Cruz*" instead of the more popular Spanish actress "*Pen é lope Cruz*". Similarly, we need to know the following post "*Whitewater, Monica, B, infidelity, Foster, health care reform, sniper fire, Sir Edmund Hillary, head trauma, I war vote, support for the Iraq war, email, servers, lesbians........*" is talking about Clinton in order to link "*Whitewater*" to "*Whitewater Development Corporation*", and "*Monica*" to "*Monica Lewinsky*". An entity linker needs to automatically construct a background knowledge base as a bridge between the source collection and the KB.

7.3.5 Commonsense Knowledge

Entity linkers are still lack of exploiting commonsense knowledge. For example, top systems mistakenly link "*Kenyatta*" in the following sentence "*In his first televised address since the attack ended on Thursday, Kenyatta condemned the "barbaric slaughter" and asked help from the Muslim community in rooting out radical elements.* " to "*Jomo Kenyatta (1891-1978)*", because it share the same last name as the correct entry "*Uhuru Kenyatta*", and both served as the President of Kenya. But from the post we can clearly see the target entity is still alive because he made an announcement. Other types of comprehensive commonsense knowledge is required to disambiguate some difficult cases. For example, we need to know that a capital city can be used to refer to a country's government, in order to link "*Washington*" to "*Federal government of the United States*" in the following post "*Millions of Americans went to war for America, and came back broken or otherwise gave up a lot, and now we look to take a huge chunk of their hide because Washington no longer works.*".

7.3.6 Morph Decoding

Morphs remain the most challenging mentions in Chinese Entity Linking. There are many different techniques that have been used to encode these morphs (Zhang et al., 2014), therefore surface features and context similarity are far from enough to decode them successfully. For example, "天朝*(Heaven Dynasty)*" is a morph created for "*China*" due to its long history. "帖木儿*(Post Wood Er)*" is created to refer to "*Genghis Khan*" because its pronunciation "*Tie Mu Er*" is close to his born name "*Temüjin*". In English, 2% mentions are morphs. For example, in "*They passed a bill, and Christie the Hutt decides he's stull sucking up to be RomBot's running mate. I think the Good Doctor is too crazy to hang it up.*", "*Christie the Hutt*", "*RomBot*" and "*Good Doctor*" refer to "*Chris Christie*", "*Mitt Romney*" and "*Ron Paul*" respectively.

8 Looking Ahead

The new Tri-lingual EDL task has created many interesting research problems and new directions. In KBP2016 we will consider the following possible extensions and improvement:

- Combine with tri-lingual slot filling and tri-lingual event extraction to form up an end-to-end cool-start tri-lingual KBP task;
- Target at a larger scale data processing, by increasing the size of source collections from 500 documents to 10,000 documents;
- Add EDL for individual specific nominals for PER, GPE, ORG, LOC and FAC entities into all three languages; This addition is likely to promote two research directions: (1) It introduces a new definition of mentions from the end usage of KB construction; (2) It may promote within-document coreference resolution research which is currently a bottleneck for all KBP tracks (EDL, Slot Filling, Event KBP);
- Add more fine-grained entity types, or allow EDL systems to automatically discover new entity types; We may start by adding Weapon, Vehicle, Commodity and other Product subtypes as defined in AMR (Banarescu et al., 2013) such as work-of-art, picture, music, show, broadcast-program, publication, book, newspaper, magazine and journal;
- Add a new task of EDL assembling;
- Add streaming data into the source collection;
- Perhaps replace Spanish with a new low-resource language for which full-document MT techniques are less mature.

Acknowledgments

Special thanks to Hoa Trang Dang (NIST) for coordinating and overseeing this track. Thanks Joe Ellis (LDC) for providing data statistics. Thanks Xiaoman Pan (RPI) for his assistance on result analysis. Heng Ji's work has been supported by U.S. DARPA DEFT Program No. FA8750-13-2-0041, ARL NS-CTA No. W911NF-09-2-0053, and NSF IIS-1523198. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proc. Linguistic Annotation Workshop*.

R. Besancon, H. Daher, H. L. Borgne, O. Ferret, A. Daquo, and A. Popescu. 2015. Cea list participation at tac edl english diagnostic task. In *Proc. Text Analytics Conference (TAC2015)*.

T. Cassidy, H. Ji, L. Ratinov, A. Zubiaga, and H. Huang. 2012. Analysis and enhancement of wikification for microblogs with context expansion. In *Proceedings of COLING 2012*.

X. Cheng and D. Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

J. Ellis, J. Getman, D. Fore, N. Kuster, Z. Song, A. Bies, and S. Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *Proc. Text Analysis Conference (TAC2015)*.

T. Finin, D. Lawrie, P. McNamee, J. Mayfield, D. Oard, N. Gao, N. Peng, Y. Lin, J. MacKin, and T. Dowd. 2015. HLTCOE participation in TAC KBP 2015: Cold start and TEDL. In *Proceedings of the Eighth Text Analysis Conference*, Gaithersburg, Maryland, November. NIST.

R. Florian, H. Jing, N. Kambhatla, and I. Zitouni. 2006. Factorizing complex models: A case study in mention detection. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL'06)*, Sydney, Australia, July.

B. Hachey, J. Nothman, and W. Radford. 2014. Cheap and easy entity evaluation. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 2)*, pages 464–469.

B. Heinzerling, A. Judea, and M. Strube. 2015. Hits at tac kbp 2015: Entity discovery and linking, and event nugget detection. In *Proc. Text Analysis Conference (TAC2015)*.

Y. Hong, D. Lu, D. Yu, X. Pan, X. Wang, Y. Chen, L. Huang, and H. Ji. 2015. Rpi_blender tac-kbp2015 system description. In *Proc. Text Analysis Conference (TAC2015)*.

H. Huang, Y. Cao, X. Huang, H. Ji, and C. Lin. 2014. Collective tweet wikification based on semi-supervised graph regularization. In *Proc. the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*.

H. Ji, D. Westbrook, and R. Grishman. 2005. Using semantic relations to refine coreference decisions. In *Proc. Human Language Technology Conf./ Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, Vancouver, B.C., Canada, Oct.

H. Ji, R. Grishman, D. Freitag, M. Blume, J. Wang, S. Khadivi, R. Zens, and H. Ney. 2009. Name extraction and translation for distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Text Analysis Conf. (TAC) 2010*.

H. Ji, R. Grishman, and H. T. Dang. 2011a. An overview of the tac2011 knowledge base population track. In *Proc. Text Analysis Conf. (TAC'11)*, Gaithersburg, Maryland, Nov.

H. Ji, R. Grishman, and H.T. Dang. 2011b. Overview of the tac 2011 knowledge base population track. In *Text Analysis Conf. (TAC) 2011*.

H. Ji, H. T. Dang, J. Nothman, and B. Hachey. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*.

C. Lee, Y. Hwang, H. Oh, S. Lim, J. Heo, C. Lee, H. Kim, J. Wang, and M. Jang. 2006. Fine-grained named entity recognition using conditional random fields for question answering. In *Information Retrieval Technology*, pages 581–587. Springer.

Q. Li and H. Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proc. the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*.

Q. Li, H. Ji, Y. Hong, and S. Li. 2014. Constructing information networks using one single model. In *Proc. the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP2014)*.

D. Lin, S. Zhao, B. V. Durme, and M. Paca. 2008. Mining parenthetical translations from the web by word alignment. In *Proc. Association for Computational Linguistics Conference (ACL2008)*.

X. Ling, S. Singh, and D. S. Weld. 2014. Context representation for named entity linking. In *Proc. Pacific Northwest Regional NLP Workshop (NW-NLP 2014)*.

W. Lu and D. Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*.

X. Luo. 2005. On coreference resolution performance metrics. In *Proc. HLT/EMNLP*, pages 25–32.

X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proc. the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL-HLT 2015)*.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

M. Sammons, H. Peng, Y. Song, S. Upadhyay, C.-T. Tsai, P. Reddy, S. Roy, and D. Roth. 2015. Illinois ccg tac 2015 event nugget, entity discovery and linking, and slot filler validation systems. In *Proc. Text Analytics Conference (TAC2015)*.

A. Sil and R. Florian. 2015. The ibm systems for trilingual entity discovery and linking at tac 2015. In *Proc. Text Analysis Conference (TAC2015)*.

H. Wang, J. G. Zheng, X. Ma, P. Fox, and H. Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*.

L. Xiao and D. S. Weld. 2012. Fine-grained entity recognition. In *AAAI*.

B. Zhang, H. Huang, X. Pan, H. Ji, K. Knight, Z. Wen, Y. Sun, J. Han, and B. Yener. 2014. Be appropriate and funny: Automatic entity morph encoding. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL2014)*.

B. Zhang, H. Huang, X. Pan, S. Li, C.-Y. Lin, H. Ji, K. Knight, Z. Wen, Y. Sun, J. Han, and B. Yener. 2015. Context-aware entity morph decoding. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL2015)*.