Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End Cold-Start KBP

Heng Ji[1] **and Joel Nothman**[2]
[1] Computer Science Department, Rensselaer Polytechnic Institute
jih@rpi.edu
[2] Sydney Informatics Hub, University of Sydney
joel.nothman@gmail.com

Abstract

In this paper we give an overview of the Tri-lingual Entity Discovery and Linking (EDL) task at the Knowledge Base Population (KBP) track at TAC2016. We will summarize several new and effective research directions including cross-lingual knowledge transfer and exploiting language-specific resources. In this year we make the EDL task as part of end-to-end Tri-lingual cold-start knowledge base construction on a very large document collection (90,003 documents). So we will also measure and investigate the impact of EDL on cold-start slot filling.

1 Introduction

The success of Tri-lingual (English, Chinese and Spanish) Entity Discovery and Linking (EDL) techniques at the NIST TAC Knowledge Base Population (KBP) track has been mainly demonstrated through their *portability* (the performance of Chinese and Spanish EDL is comparable to that of English). Compared to the KBP2015 EDL task (Ji et al., 2015), we made the following changes and improvement in KBP2016.

We intend to investigate whether the *quality* of state-of-the-art Tri-lingual EDL methods is sufficient for end-to-end KBP. We combined EDL with Tri-lingual slot filling to form up a new end-to-end tri-lingual KBP task. We also targeted at a larger scale data processing to measure the *scalability* of EDL, by significantly increasing the size of source collections from 500 documents to 90,003 documents for the end-to-end KBP task.

Automatically discovering and clustering nominal mentions, and grounding them to KB or other name mentions is crucial to many real-world applications such as intelligence analysis and product profiling. Therefore we extended the task of detecting specific, individual nominal mentions to all five entity types and all three languages.

The rest of this paper is structured as follows. Section 2 describes the definition of the full Tri-lingual EDL task. Section 3 briefly summarizes the participants. Section 4 highlights some annotation efforts. Section 5 summarize evaluation results, comparison across languages 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

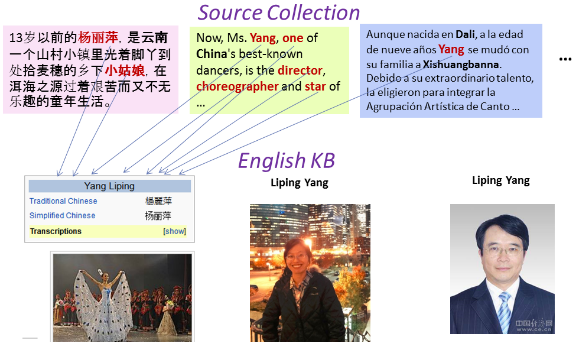
This section will summarize the Tri-lingual Entity Discovery and Linking tasks conducted at KBP 2016. More details regarding data format and scoring software can be found in the task website[1].

2.1 Task Definition

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 three languages (English, Chinese and Spanish), classify them into one of the following pre-defined

[1] http://nlp.cs.rpi.edu/kbp/2016/

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.



13岁以前的杨丽萍，是云南一个山村小镇里光着脚丫到处拾麦穗的乡下小姑娘，在洱海之源过着艰苦而又不无乐趣的童年生活。

Now, Ms. Yang, one of China's best-known dancers, is the director, choreographer and star of ...

Aunque nacida en Dali, a la edad de nueve años Yang se mudó con su familia a Xishuangbanna. Debido a su extraordinario talento, la eligieron para integrar la Agrupación Artística de Canto ...

Figure 1: Tri-lingual EDL Input/Output Example

Besides name mentions, nominal mentions referring to specific, real-world individual entities should also be extracted. 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.

We set two evaluation windows, the first in August and the second in September to check the progress. The slot filling teams were encouraged to use EDL results to detect entities, and the EDL systems in the second evaluation window were also encouraged to use the results of EDL systems from the first window.

2.2 Scoring Metrics

As detailed in Table 1, TAC 2016 reports measures for detection of mentions and their types, identification of KB links, and clustering of mentions with or without links. 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. `NERLC` evaluates the intersection of `NERC` and `NERL`.

Results below also refer to other diagnostic measures, including `NELC` which reports linking, mention detection and classification performance, discarding NIL annotations; `NENC` 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

We also evaluate EDL as a cross-document coreference task, in which the set of tuples is

Short name	Name in scoring software	Scope	Key	Evaluates
Mention evaluation				
NER	strong_mention_match	all	*span*	Identification
NERC	strong_typed_mention_match	all	*span,type*	+ classification
Linking evaluation				
NERLC	strong_typed_all_match	all	*span,type,kbid*	+ linking
NELC	strong_typed_link_match	KB-linked mentions	*span,type,kbid*	Link recognition and class
NENC	strong_typed_nil_match	NIL mentions	*span,type*	NIL recognition and class
Tagging evaluation				
KBIDs	entity_match	KB-linked mentions	*docid,kbid*	Document tagging
Clustering evaluation				
CEAFm	mention_ceaf	all	*span*	Identification and clustering
CEAFmC	typed_mention_ceaf	all	*span,type*	+ classification
CEAFmC+	typed_mention_ceaf_plus	all	*span,type,kbid*	+ linking
Clustering diagnostics				
CEAFm-doc	mention_ceaf;docid=<micro>	micro-average across docs	*span*	Within-document clustering
CEAFm-1st	mention_ceaf:is_first:span	doc's 1st mention of entity	*span*	Cross-document clustering

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.

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. We further constrain clustering evaluation to require correct mention type classification (CEAFmC) and correct KB link targets (CEAFmC+, which includes type).

2.2.3 Cross-document clustering diagnostics

The overall clustering measures do not distinguish between the task of clustering mentions within a document and clustering across documents. Because clustering within a document is able to exploit local discourse features, including a “one referent per document” assumption, cross-document and within-document coreference resolution should ideally be evaluated as separate tasks. We report CEAFm-doc as a summary of within-document CEAFm coreference performance, micro-averaging across all documents. This score bounds overall CEAFm from above, as cross-document coreference errors reduce the number of true positives in the maximum-score bijection.

We may also attempt to separately evaluate cross-document clustering, in order to disregard within-document clustering errors, and remove the bias of CEAFm and CEAFm-doc to long within-document coreference chains. This is, however, non-trivial to do, as we need to identify the correspondence of a gold and predicted entity in each document without requiring that all mentions be matched. We approximate cross-document performance by limiting evaluation to the first mention per document of each predicted and gold entity, in CEAFm-1st.[2] This biases evaluation to documents and genres where the first mention of each gold entity is easily resolved, e.g. by use of a canonical name, but should provide an estimate of cross-document clustering performance.

2.2.4 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 a system annotates each document independently; and intervals are not reliable where a system uses global clustering information in its mention detection, classification and KB linking. For similar reasons, we do not calculate confidence intervals for clustering metrics.

3 Participants Overview

Table 2 summarizes the participants for KBP2016 Trilingual EDL task. In total 11 teams submitted for the first evaluation window and 11 teams submitted runs for the second evaluation window, and 7 teams submitted to both windows. 6 teams (BUPT, IBM, OSU, RPI, CMU and USTC) developed systems for all three languages, among them BUPT, IBM, OSU and RPI participated in both windows.

4 Data Annotation and Resources

The details of the data annotation for KBP2016 are presented in a separate paper by the Linguistic Data Consortium (Ellis et al., 2016). In this section we only highlight a few aspects that affect the effectiveness of some new methods.

4.1 Knowledge Base

We use the same reference knowledge base as in 2015, namely BaseKB [3].

4.2 Source Document Collection

The source corpus consists of 90,003 documents, equally distributed across three languages and two genres (news articles and discussion forum (DF) posts published in recent years). Table 3 presents the number of documents for each genre and language.

4.3 Evaluation Set

500 documents from the source collection were selected for evaluation. LDC prepared manual annotations for those 500 documents as ground truth. Tables 4, 5 and 6 summarize the statistics of this evaluation set. We can see that a majority (78.2%) of mentions are linkable, and linkable entities are more often referred to by name mentions (77.4%) rather than nominal mentions. 900 entities (15.8 %) appear across two languages, while 325 entities (5.7%) appear across three languages.

Finally, we also devoted a lot of time at collecting related publications and tutorials [4], resources and softwares [5] to lower down the entry cost for EDL.

5 Evaluation Results

5.1 Overall Performance

Table 7 and Table 8 summarize the results of EDL1 and EDL2 respectively. For public release we have anonymized the team names: each team is numbered with the rank of its best submission (across EDL1 and EDL2) according to CEAFm. We then report results upon the best (by CEAFm) submission per team in each of EDL1 and EDL2. Systems not providing output for Chinese or Spanish are elided from those sections.

5.2 Progress from August to September

Four teams participated the full tri-lingual task for both evaluation windows. Table 9 shows their progress from August to September. Team 4 and 9 achieved significant improvement from one month focused research and development efforts. Team 4 developed new name taggers for the second window, instead of using open source taggers as in the first window. They also had time to exploit language-specific resources and cross-lingual knowledge transfer methods for the second evaluation window. Compared to the first window, Team 9 added full functionality of embeddings into the submissions for the second window, and also exploited team 5's entity mention extraction results. These results may inspire us to avoid packing multiple tasks within

[2] This corresponds to Pure-CDEC evaluation in Upadhyay et al. (2016) (personal correspondence).

[3] http://basekb.com/

[4] http://nlp.cs.rpi.edu/kbp/2016/elreading.html

[5] http://nlp.cs.rpi.edu/kbp/2016/tools.html

Team	Affiliation	Language			Evaluation	
		CMN	ENG	SPA	1st	2nd
BUPT	Beijing University of Post and Telecommunication	✓	✓	✓	✓	✓
IBM	International Business Machines Corporation	✓	✓	✓	✓	✓
OSU	Oregon State University	✓	✓	✓	✓	✓
RPI_BLENDER	Rensselaer Polytechnic Institute	✓	✓	✓	✓	✓
CMU_CS	Carnegie Mellon University	✓	✓	✓	✓	
USTC	University of Science and Technology of China	✓	✓	✓	✓	
HITS	Heidelberg Institute for Theoretical Studies		✓		✓	✓
WednesdayGo	Zhejiang University		✓		✓	✓
summa	University College London		✓		✓	✓
EastLand	National University of Defense Technology	✓			✓	
REDES	Universidad de Jaen and Universidad de Alicante		✓	✓	✓	
hltcoe	Human Language Technology Center of Excellence	✓	✓	✓		✓
UI_CCG	UIUC Cognitive Computation Group	✓		✓		✓
UTAustin	University of Texas at Austin	✓	✓	✓		✓
YorkNRM	York University	✓	✓	✓		✓

Table 2: Runs Submitted by KBP2016 Trilingual Entity Discovery and Linking Participants

	Chinese	Spanish	English	All
News	15,000	15,001	15,001	45,002
DF	15,001	14,999	15,001	45,001
All	30,001	30,000	30,002	90,003

Table 3: Total # of Documents in Source Collection

	Chinese	Spanish	English	All
News	85	83	84	252
DF	82	85	84	251
All	167	168	168	503

Table 4: Total # of Documents in Evaluation Data

	Chinese	Spanish	English	All
Linkale Name	6,060	4,054	5,618	15,732
Linkale Nominal	1,055	1,086	1,589	3,730
NIL Name	1,263	1,331	1,391	3,985
NIL Nominal	467	493	633	1,593
All	8,845	6,964	9,231	25,130

Table 5: Total # of Mentions in Evaluation Data

	Chinese	Spanish	English	All
Linkable	762	748	919	2,429
NIL	1,060	963	1,246	3,269
All	1,822	1,711	2,165	5,698

Table 6: Total # of Entities in Evaluation Data

a short evaluation window, so teams will have more time to attempt more advanced methods and conduct thorough analysis.

5.3 Comparison on Languages

Table 10 compares the break down scores for various languages. Compared to English, Chinese and Spanish entity mention extraction scores are slightly lower. However, for the combined score of entity extraction, clustering and linking, Chinese scores are higher than English. Compared to the other Information Extraction and KBP tasks (RPI team's performance shown in Figure 2), we are very glad to see the gap between foreign languages and English is being filled for EDL.

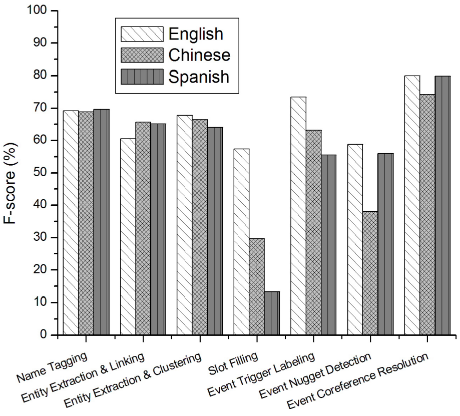


Figure 2: Information Extraction and KBP Performance Comparison Across Languages

System	NER			NERC			NERLC			KBIDs			CEAFm			CEAFmC			CEAFm-doc			CEAFm-1st		
	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	P	R	F_1
TRILINGUAL																								
1	86.9	**74.5**	**80.2**	82.2	**70.4**	**75.9**	70.1	**60.0**	**64.7**	82.3	**67.8**	**74.3**	74.0	**63.3**	**68.2**	71.6	**61.3**	**66.0**	79.6	**68.1**	**73.4**	63.5	**62.0**	**62.7**
3	85.9	60.9	71.3	80.7	57.2	66.9	71.2	50.5	59.0	69.4	57.9	63.1	76.7	54.4	63.7	73.4	52.0	60.9	82.7	58.6	68.6	59.8	56.1	57.9
5	86.9	62.8	72.9	83.0	59.9	69.6	66.4	48.0	55.7	77.6	56.2	65.2	73.3	52.9	61.5	70.8	51.1	59.3	82.4	59.5	69.1	64.9	56.4	60.4
4	87.7	51.1	64.6	**83.2**	48.4	61.2	**76.4**	44.5	56.3	71.0	56.3	62.8	80.7	47.0	59.4	**77.3**	45.0	56.9	**84.9**	49.5	62.5	61.5	50.3	55.3
7	81.3	53.6	64.6	74.6	49.2	59.3	66.2	43.6	52.6	66.4	48.1	55.8	70.4	46.4	55.9	67.7	44.6	53.8	78.2	51.5	62.1	55.1	45.8	50.0
9	64.2	41.6	50.5	55.2	35.8	43.4	46.2	30.0	36.4	46.8	52.8	49.6	55.7	36.1	43.8	50.4	32.7	39.7	60.5	39.2	47.6	33.2	29.9	31.5
11	87.9	26.6	40.9	**83.2**	25.2	38.7	75.6	22.9	35.2	79.7	25.8	39.0	78.8	23.9	36.6	75.9	23.0	35.3	82.0	24.9	38.2	**68.1**	23.9	35.3
12	**89.8**	24.3	38.3	71.1	19.3	30.3	63.4	17.2	27.0	**85.5**	21.6	34.5	73.9	20.0	31.5	66.8	18.1	28.5	79.1	21.4	33.7	57.4	22.4	32.2
13	87.3	19.8	32.3	79.8	18.1	29.5	71.0	16.1	26.2	73.8	21.5	33.3	**81.1**	18.4	30.0	75.1	17.0	27.7	83.5	18.9	30.8	67.6	20.7	31.7
14	72.8	24.9	37.2	49.6	17.0	25.3	23.0	7.9	11.8	32.6	23.7	27.4	53.0	18.2	27.0	36.3	12.4	18.5	64.4	22.1	32.9	33.7	17.4	23.0
15	52.3	3.2	6.0	36.9	2.2	4.2	34.1	2.1	3.9	39.1	3.6	6.5	44.3	2.7	5.1	34.2	2.1	3.9	47.2	2.9	5.4	33.5	2.1	3.9
2	85.9	23.2	36.6	81.8	22.1	34.8	19.7	5.3	8.4	0.0	0.0	0.0	3.9	1.1	1.7	3.9	1.0	1.6	22.3	6.0	9.5	9.5	0.2	0.3
CHINESE																								
1	83.7	**78.1**	**80.8**	78.9	**73.7**	**76.2**	67.3	**62.8**	**65.0**	**82.5**	**67.1**	**74.0**	75.2	**70.1**	**72.6**	72.7	**67.8**	**70.2**	77.8	**72.6**	**75.1**	**67.8**	**67.2**	**67.5**
3	85.4	61.7	71.7	81.3	58.8	68.2	75.9	54.9	63.7	75.9	60.2	67.1	**82.7**	59.7	69.4	79.0	57.1	66.3	**84.4**	61.0	70.8	66.8	60.0	63.2
5	79.6	56.6	66.2	76.1	54.1	63.3	66.6	47.4	55.4	76.7	55.6	64.4	75.3	53.6	62.6	72.5	51.6	60.3	78.2	55.7	65.0	61.1	56.3	58.6
4	**85.5**	45.4	59.3	**82.0**	43.5	56.8	**76.1**	40.4	52.7	68.3	49.9	57.7	82.2	43.6	57.0	**79.3**	42.0	54.9	83.9	44.5	58.2	64.8	47.6	54.9
7	72.1	50.8	59.6	67.1	47.3	55.5	58.1	40.9	48.0	57.1	41.4	48.0	65.6	46.2	54.3	63.5	44.7	52.5	70.3	49.6	58.2	48.5	43.1	45.6
9	71.5	45.5	55.6	64.5	41.0	50.2	57.8	36.8	45.0	52.4	56.7	54.4	67.6	43.1	52.6	62.3	39.7	48.5	69.6	44.3	54.2	42.1	32.0	36.3
15	52.3	9.0	15.4	36.9	6.4	10.9	34.1	5.9	10.0	39.1	11.3	17.5	44.3	7.6	13.0	34.2	5.9	10.0	47.2	8.1	13.9	33.5	6.5	10.9
ENGLISH																								
11	87.9	72.2	79.3	83.2	68.4	75.0	75.6	**62.2**	**68.2**	79.7	69.0	74.0	78.8	**64.7**	**71.1**	75.9	**62.4**	**68.5**	82.0	**67.4**	**74.0**	68.1	**63.4**	65.7
1	89.5	**75.1**	**81.7**	84.6	**71.0**	**77.2**	73.1	61.3	66.6	83.6	**69.5**	**75.9**	76.4	64.0	69.7	74.1	62.2	67.6	80.1	67.2	73.1	67.4	62.6	64.9
4	**93.5**	57.4	71.1	**88.4**	54.3	67.3	**81.6**	50.1	62.1	76.1	65.1	70.2	**87.6**	53.7	66.6	**83.3**	51.1	63.3	**90.7**	55.7	69.0	**71.4**	58.2	64.2
5	92.0	68.7	78.7	87.9	65.7	75.2	65.8	49.2	56.3	72.8	57.4	64.2	75.8	56.6	64.8	73.3	54.8	62.7	85.7	64.0	73.3	70.6	61.7	**65.8**
3	92.2	58.8	71.8	87.5	55.9	68.2	74.0	47.2	57.7	72.9	58.5	64.9	81.7	52.1	63.6	78.4	50.1	61.1	87.5	55.9	68.2	67.3	57.3	61.9
13	87.3	53.7	66.5	79.8	49.1	60.8	71.0	43.6	54.0	73.8	57.6	64.7	81.1	49.9	61.8	75.1	46.2	57.2	83.5	51.3	63.5	67.6	55.1	60.7
12	89.8	66.0	76.1	71.1	52.2	60.2	63.4	46.6	53.7	**85.5**	57.9	69.0	73.9	54.3	62.6	66.8	49.1	56.6	79.1	58.1	66.9	57.4	59.5	58.4
7	88.8	53.6	66.9	81.4	49.1	61.3	72.4	43.7	54.5	74.6	50.5	60.3	77.5	46.7	58.3	73.8	44.5	55.5	84.5	51.0	63.6	65.4	49.6	56.4
9	55.3	37.0	44.3	44.5	29.8	35.7	37.9	25.4	30.4	41.0	48.1	44.2	50.4	33.7	40.4	42.0	28.1	33.7	52.3	34.9	41.9	29.9	28.7	29.2
14	81.3	41.7	55.2	62.1	31.9	42.1	28.7	14.7	19.5	35.4	35.2	35.3	65.0	33.3	44.1	49.2	25.2	33.3	71.4	36.6	48.4	44.3	29.6	35.5
2	85.9	63.0	72.7	81.8	60.0	69.3	19.7	14.4	16.6	0.0	0.0	0.0	3.9	2.9	3.3	3.9	2.8	3.3	22.3	16.3	18.8	9.5	0.4	0.8
SPANISH																								
1	88.1	**69.0**	**77.4**	83.9	**65.6**	**73.6**	**70.1**	**54.9**	**61.6**	80.4	**66.4**	**72.8**	75.3	**59.0**	**66.2**	72.3	**56.6**	**63.5**	81.5	**63.8**	**71.5**	59.2	**60.3**	59.8
5	**89.3**	62.8	73.7	**84.7**	59.5	69.9	67.1	47.2	55.4	**85.8**	55.5	67.4	**76.1**	53.5	62.8	**73.1**	51.4	60.3	**83.1**	58.4	68.6	**67.9**	54.7	**60.6**
3	79.8	62.7	70.2	72.5	56.9	63.8	62.5	49.1	55.0	59.9	54.9	57.3	69.4	54.5	61.1	65.4	51.3	57.5	75.5	59.3	66.4	50.8	55.1	52.8
7	84.5	57.1	68.1	76.5	51.6	61.6	69.6	47.0	56.1	66.8	51.8	58.3	74.6	50.4	60.1	69.8	47.1	56.3	81.1	54.7	65.4	55.7	47.8	51.5
4	82.4	50.0	62.3	77.4	47.0	58.5	69.9	42.4	52.8	66.8	52.3	58.7	72.5	44.0	54.7	69.4	42.1	52.4	78.3	47.5	59.1	49.4	44.7	46.9
9	67.2	42.8	52.3	58.4	37.2	45.4	43.0	27.4	33.5	48.8	54.5	51.5	54.7	34.9	42.6	49.9	31.8	38.8	60.3	38.4	46.9	32.9	32.1	32.5
14	62.2	34.4	44.3	34.3	19.0	24.4	16.0	8.8	11.4	29.6	33.9	31.6	51.8	28.6	36.9	29.0	16.0	20.7	55.7	30.8	39.7	31.5	27.4	29.3

Table 7: Overall Entity Discovery and Linking Performance (%) during the First Evaluation Window

System	NER			NERC			NERLC			KBIDs			CEAFm			CEAFmC			CEAFm-doc			CEAFm-1st		
	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	P	R	F_1
Trilingual																								
2	85.6	**66.8**	**75.1**	81.9	**63.9**	**71.8**	70.6	**55.1**	**61.9**	82.0	**61.6**	**70.3**	73.4	**57.2**	**64.3**	71.7	**55.9**	**62.8**	78.2	**61.0**	68.5	61.6	54.2	57.7
3	85.8	62.6	72.4	80.6	58.9	68.0	70.8	51.7	59.7	69.5	59.0	63.8	75.7	55.3	63.9	72.4	52.9	61.2	81.6	59.6	68.9	59.2	**57.2**	58.2
4	86.7	57.7	69.3	81.2	54.1	64.9	74.3	49.5	59.4	72.5	61.1	66.3	**79.8**	53.1	63.8	75.9	50.6	60.7	84.2	56.1	67.3	64.4	55.7	59.7
5	86.9	62.8	72.9	83.0	59.9	69.6	66.4	48.0	55.7	77.6	56.2	65.2	73.3	52.9	61.5	70.8	51.1	59.3	82.4	59.5	**69.1**	64.9	56.4	**60.4**
6	87.7	53.4	66.4	**84.0**	51.2	63.6	74.6	45.4	56.5	77.6	51.9	62.2	78.9	48.1	59.7	**76.0**	46.3	57.6	82.1	50.0	62.2	63.6	52.9	57.8
8	70.1	61.2	65.3	65.6	57.3	61.2	48.0	41.9	44.7	73.6	40.9	52.6	56.9	49.7	53.1	55.3	48.3	51.6	66.6	58.2	62.1	47.2	49.2	48.2
10	83.8	40.8	54.9	79.7	38.8	52.2	69.8	34.0	45.7	**83.5**	38.5	52.7	75.1	36.6	49.2	72.3	35.2	47.3	80.4	39.2	52.7	**69.3**	35.5	47.0
9	70.4	50.3	58.7	62.9	45.0	52.5	51.4	36.7	42.9	50.4	58.5	54.1	59.6	42.6	49.7	55.8	39.9	46.6	64.4	46.1	53.7	39.0	35.9	37.4
11	87.9	26.6	40.9	83.2	25.2	38.7	**75.6**	22.9	35.2	79.7	25.8	39.0	78.8	23.9	36.6	75.9	23.0	35.3	82.1	24.9	38.2	68.2	23.9	35.3
12	**89.0**	24.9	38.9	73.3	20.5	32.0	62.3	17.4	27.2	80.8	21.5	33.9	75.6	21.1	33.0	68.6	19.2	30.0	79.5	22.2	34.7	59.0	23.0	33.1
13	88.8	22.0	35.2	81.1	20.1	32.2	70.7	17.5	28.1	75.7	22.2	34.3	79.4	19.7	31.5	74.0	18.3	29.4	**84.6**	20.9	33.6	68.0	21.0	32.0
Chinese																								
3	**85.0**	63.9	**73.0**	81.0	60.9	69.5	76.2	**57.3**	**65.4**	76.1	**63.1**	**69.0**	**82.1**	**61.7**	**70.4**	**78.4**	**58.9**	**67.3**	**83.9**	**63.0**	**72.0**	66.7	**61.7**	**64.1**
2	82.2	**65.1**	72.7	78.9	**62.5**	**69.8**	69.4	55.0	61.3	**84.0**	58.3	68.8	75.2	59.6	66.5	73.5	58.3	65.0	77.0	61.0	68.1	65.2	54.9	59.6
4	81.4	56.7	66.8	77.5	54.0	63.7	70.8	49.3	58.1	74.2	58.0	65.1	78.8	54.9	64.7	75.7	52.7	62.1	80.6	56.1	66.1	**66.9**	56.9	61.5
10	84.0	64.0	72.6	79.3	60.4	68.5	68.8	52.4	59.5	81.0	59.3	68.5	74.7	56.9	64.6	71.3	54.3	61.6	81.9	62.3	70.8	66.8	56.0	60.9
6	84.3	53.2	65.2	**81.8**	51.6	63.3	**77.1**	48.6	59.6	82.5	52.3	64.0	80.5	50.8	62.3	78.3	49.4	60.6	82.6	52.1	63.9	65.0	51.7	57.6
5	79.6	56.6	66.2	76.1	54.1	63.3	66.6	47.4	55.4	76.7	55.6	64.4	75.3	53.6	62.6	72.5	51.6	60.3	78.2	55.7	65.0	61.1	56.3	58.6
8	69.3	63.6	66.3	64.9	59.6	62.2	43.3	39.7	41.4	75.8	33.9	46.8	60.7	55.7	58.1	58.5	53.7	56.0	66.5	61.0	63.6	46.7	57.6	51.6
9	71.5	45.5	55.6	64.5	41.0	50.2	58.0	36.9	45.1	55.6	55.8	55.7	66.3	42.2	51.6	61.1	38.9	47.5	69.2	44.0	53.8	44.0	30.7	36.1
English																								
11	87.9	**72.2**	**79.3**	83.2	**68.3**	75.0	75.6	**62.1**	**68.2**	79.7	**69.0**	**74.0**	78.8	**64.7**	**71.1**	75.9	**62.3**	**68.4**	82.1	**67.4**	**74.0**	68.2	**63.4**	65.7
2	87.8	71.5	78.8	83.6	68.0	75.0	71.5	58.2	64.2	**81.3**	63.3	71.2	74.0	60.2	66.4	72.4	58.9	65.0	77.4	63.0	69.4	64.8	56.2	60.2
4	**93.4**	57.5	71.1	**88.3**	54.3	67.3	**81.5**	50.1	62.1	76.0	65.0	70.1	**87.4**	53.7	66.5	**83.1**	51.1	63.3	**90.5**	55.7	69.0	**71.2**	58.3	64.1
5	92.0	68.7	78.7	87.9	65.7	**75.2**	65.8	49.2	56.3	72.8	57.4	64.2	75.8	56.6	64.8	73.3	54.8	62.7	85.7	64.0	73.3	70.6	61.7	**65.8**
3	91.8	61.4	73.6	87.3	58.4	70.0	72.2	48.3	57.8	72.7	58.9	65.1	79.0	52.8	63.3	75.9	50.7	60.8	84.7	56.6	67.9	64.8	58.8	61.7
13	88.8	59.6	71.4	81.1	54.4	65.1	70.7	47.5	56.8	75.7	59.4	66.6	79.4	53.3	63.8	74.0	49.7	59.4	84.6	56.8	68.0	68.0	55.7	61.2
12	89.0	67.5	76.8	73.3	55.6	63.2	62.3	47.3	53.8	80.8	57.4	67.1	75.6	57.3	65.2	68.6	52.0	59.2	79.5	60.3	68.6	59.0	61.1	60.1
6	91.6	57.2	70.4	87.2	54.4	67.0	73.2	45.7	56.3	72.5	54.7	62.4	79.2	49.5	60.9	75.7	47.2	58.2	81.8	51.1	62.9	65.5	58.1	61.6
8	68.0	66.8	67.4	64.1	62.9	63.5	49.1	48.2	48.6	76.0	51.4	61.3	59.9	58.8	59.3	57.9	56.9	57.4	63.9	62.7	63.3	53.5	55.9	54.6
9	68.9	59.2	63.7	63.1	54.2	58.3	49.2	42.3	45.5	45.5	63.9	53.2	57.2	49.2	52.9	54.4	46.7	50.2	60.2	51.7	55.6	38.1	43.4	40.6
Spanish																								
10	83.6	**65.5**	73.4	80.3	**62.9**	**70.5**	71.0	**55.6**	**62.4**	**85.9**	**63.7**	**73.1**	76.8	**60.2**	**67.5**	74.4	**58.3**	**65.4**	78.6	**61.6**	**69.1**	**72.8**	**58.6**	**64.9**
2	87.1	62.8	73.0	83.5	60.2	70.0	71.1	51.2	59.6	81.0	62.8	70.8	75.6	54.5	63.3	73.4	52.9	61.5	81.0	58.4	67.8	58.2	54.4	56.2
5	**89.3**	62.8	**73.7**	**84.7**	59.5	69.9	67.1	47.2	55.4	85.8	55.5	67.4	76.1	53.5	62.8	73.1	51.4	60.3	**83.1**	58.4	68.6	67.9	54.7	60.6
4	85.4	59.4	70.1	77.5	53.9	63.5	70.3	48.9	57.7	66.9	59.6	63.1	76.0	52.8	62.3	70.8	49.3	58.1	81.4	56.6	66.8	57.0	53.2	55.0
3	79.8	62.7	70.2	72.5	56.9	63.8	62.5	49.1	55.0	59.9	54.9	57.3	69.4	54.5	61.1	65.4	51.3	57.5	75.5	59.3	66.4	50.8	55.1	52.8
6	86.6	48.7	62.3	82.4	46.3	59.2	**73.0**	41.0	52.5	80.1	48.0	60.1	**79.0**	44.3	56.8	**75.9**	42.6	54.6	82.0	46.0	58.9	61.6	49.3	54.8
8	75.4	50.9	60.8	69.8	47.1	56.2	54.0	36.4	43.5	68.0	35.2	46.4	64.0	43.1	51.5	60.9	41.0	49.0	72.2	48.7	58.1	50.9	43.1	46.7
9	71.5	44.6	54.9	60.5	37.7	46.5	46.8	29.2	36.0	53.3	54.6	53.9	62.1	38.7	47.7	54.7	34.1	42.0	66.1	41.2	50.7	39.4	34.9	37.0

Table 8: Overall Entity Discovery and Linking Performance (%) during the Second Evaluation Window

Team	3	4	5	9
August	60.9	56.9	59.3	39.7
September	61.2	60.7	59.3	46.6

Table 9: Progress from August to September: CEAFmC scores (%) to indicate Entity Extraction+Clustering+Linking performance

Languages	ENG	CMN	SPA
Extraction	77.2	76.2	73.6
Extraction+Linking	68.2	65.4	62.4
Extraction+Clustering	71.1	72.6	67.5
Extraction+Clustering+Linking	68.5	70.2	65.4

Table 10: Top F-scores (%) Comparison Across Languages

5.4 Entity Types and Textual Genres

This year's evaluation extended nominal mention detection to all five entity types. Figures 3, 4, 5 and 6 show the break-down scores for various entity types and genres. As we have observed in previous years, the bottlenecks still lie on facilities and nominals. Interestingly we don't observe any significant performance gap between newswire and discussion forums any more, which indicate that EDL techniques are now well adapted to informal genres.

5.5 NIL and Non-NIL Comparison

Figures 7, 9, 10, 8, 11, 13, 14 and 12 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 all languages, all systems except team 4 achieved comparable performance on NIL and Non-NILs. Team 4 encoded specific heuristic rules to detect whether an entity mention is specific. Last year we observed that for Spanish, NIL mentions are more challenging than Non-NIL mentions. But this year systems performed very well on Spanish NIL mentions. Comparing NERL and NERLC scores, we can also see that mention typing accuracy keeps very high for all three languages.

5.6 EDL Assembling

Team 6 tried assembling EDL results from the first evaluation window by both supervised and unsupervised methods. Their assembling approaches assume no prior knowledge about the history performance of any individual system, which is a challenging setting. The assembled results did out beat top 1-5 teams, but outperform teams with lower ranks.

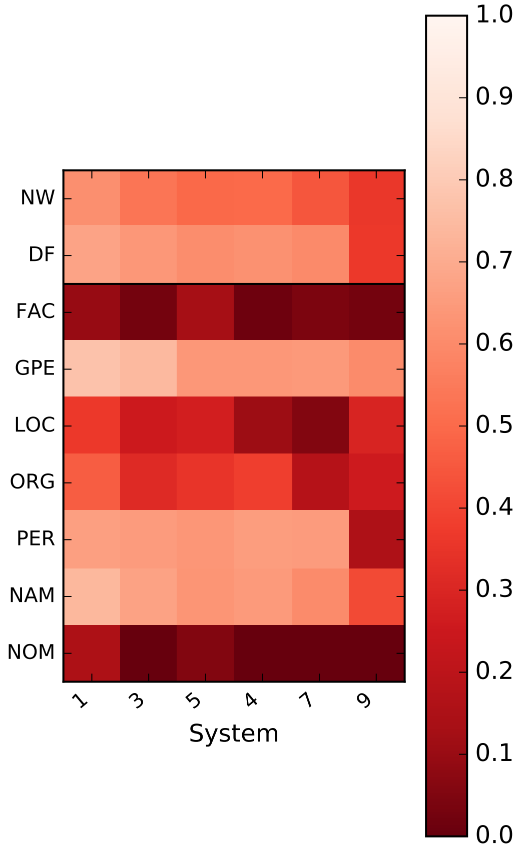


Figure 3: Evaluation Window 1 Breakdown Entity Mention Extration and Linking Performance for Entity Types and Genres

5.7 Impact on Cold-Start Slot Filling

Due to the short development time, only a few cold-start slot filling teams (RPI, Stanford) exploited EDL results in superficial ways. For example, RPI used the DBPedia properties of linkable entities to expand query mentions (e.g., expanding "*Morsi*" to "*Mohamed Morsi*"; expanding "*Cuomo*" to "*Mario Cuomo*") and enrich slot fillers. The mean F-score of RPI cold-start slot filling is improved from 13.27% to 14.57% after using EDL results.

In the future EDL techniques should be used to replace simple document retrieval or document clustering components that are currently used by most slot filling systems. We believe they will be very useful, especially for highly ambiguous queries. For example, some organization names like "*Ministry of Environmental Protection*" might be shared by many countries. Usually a single query document may not mention the country

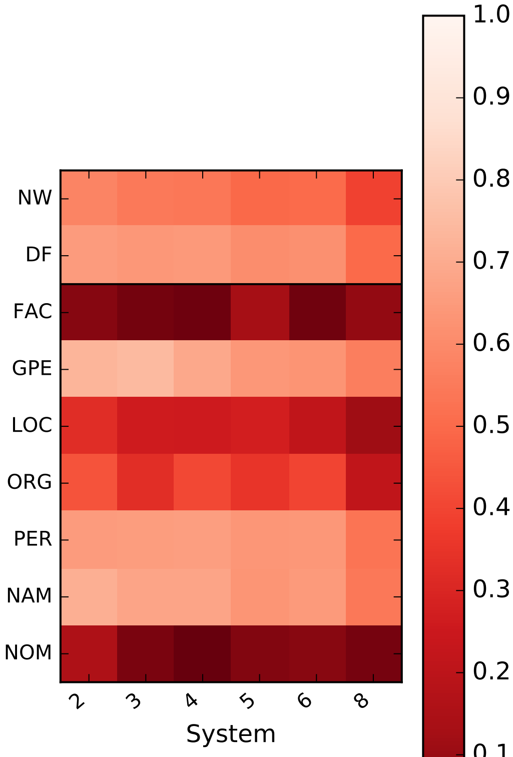


Figure 4: Evaluation Window 2 Breakdown Entity Mention Extration and Linking Performance for Entity Types and Genres

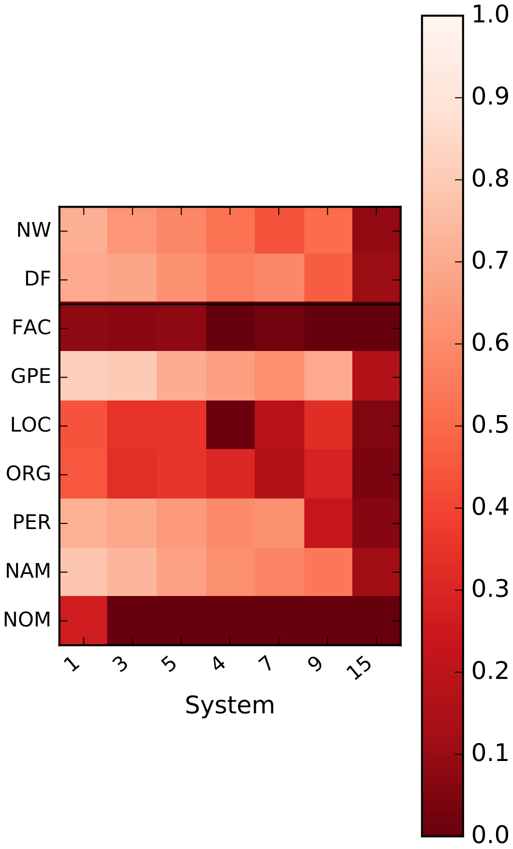


Figure 5: Evaluation Window 1 Breakdown Entity Mention Extration and Clustering Performance for Entity Types and Genres

name. But if an EDL system can cluster multiple documents that involve the same entity (e.g., "*Ministry of Environmental Protection of the People's Republic of China*"), we can obtain much richer contexts and attributes to fill in slots.

5.8 Agreement across submissions

Last year's overview discussed the extent to which correct links detected by each system were common. By plotting the performance of systems with respect to each other's output (i.e. treating another team's submission as the gold standard), we can also see that the outputs of the best systems are much more similar to each other than they are to the gold standard. The heatmaps in Figures 15 and 16 suggest that they are making the same errors and omissions with each other, consistent with the general difficulty of obtaining recall in information extraction's long tail.

6 What's New and What Works

6.1 Keep Name Taggers Young

Most teams have used supervised models for extracting entity mentions. We have found it's crucial to keep name taggers up-to-date, which is not surprising as observed by (Mota and Grishman, 2008). We found for all three languages (English, Chinese and Spanish), training a name tagger from new data from similar time periods as the evaluation data, performs much better than a tagger trained from old data from a decade ago even though with a ten times size. Tables 11 and 12 show the detailed performance of a bi-directional LSTMs (Long Short Term Memory) networks based name tagger trained from various conditions.

6.2 Language-specific Resources

Regardless of the success of many teams at applying deep neural networks to name tagging which can save efforts at human

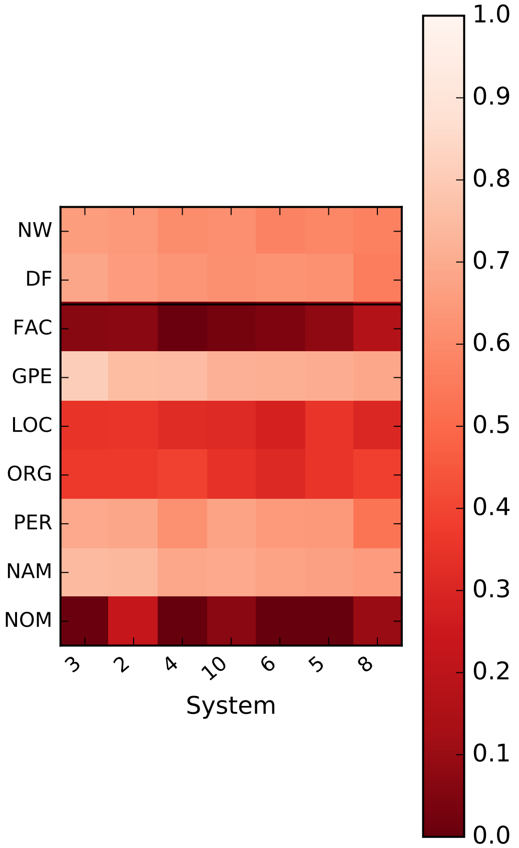


Figure 6: Evaluation Window 2 Breakdown Entity Mention Extration and Clustering Performance for Entity Types and Genres

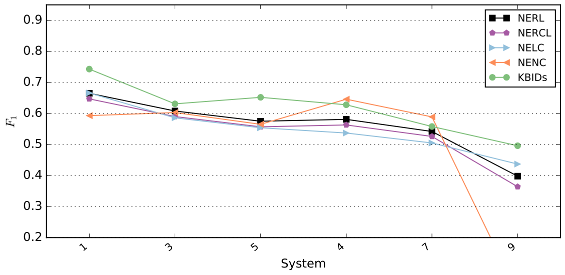


Figure 7: Evaluation Window 1 Tri-lingual EDL NIL and Non-NIL Performance Comparison

Training Data	F-score
Ontonotes (1,911 docs from 2005-2006)	56.7
EDL2015 (147 docs from 2014-2015)	49.9
Ontonotes+EDL	70.2

Table 11: Chinese Name Tagging Performance (%)

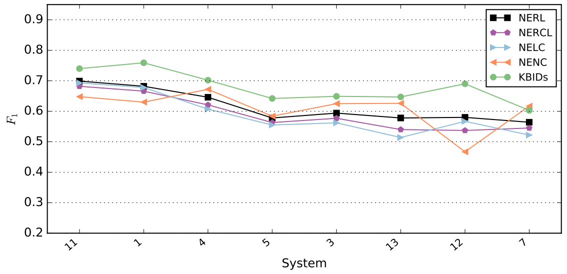


Figure 8: Evaluation Window 1 English EDL NIL and Non-NIL Performance Comparison

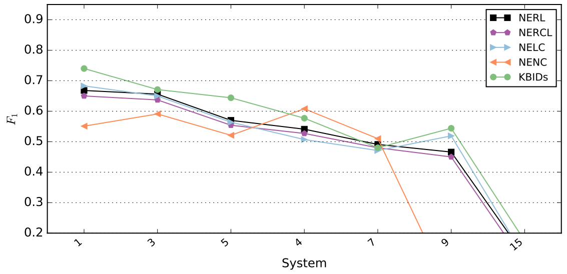


Figure 9: Evaluation Window 1 Chinese EDL NIL and Non-NIL Performance Comparison

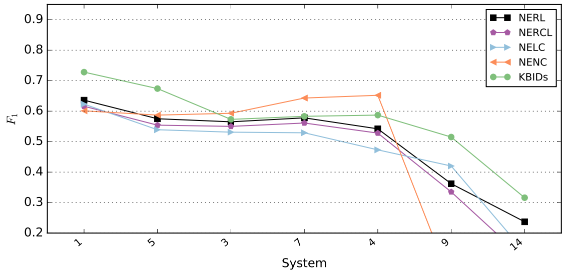


Figure 10: Evaluation Window 1 Spanish EDL NIL and Non-NIL Performance Comparison

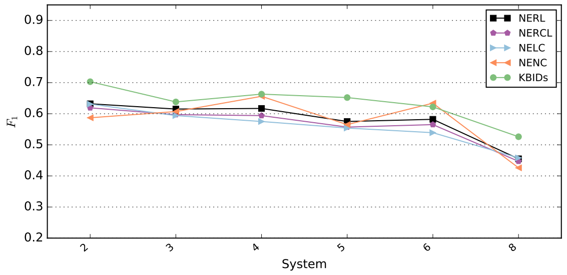


Figure 11: Evaluation Window 2 Tri-lingual EDL NIL and Non-NIL Performance Comparison

feature engineering, we have found explicit language-specific resources still provide significant gains. RPI team exploited the following Chinese language-specific resources and achieved significant improvement on Chinese name tagging as shown in Figure 17.

- Chinese first name and last name character

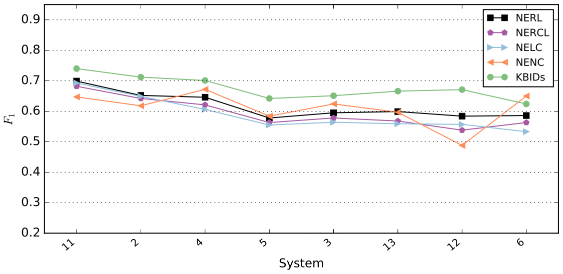


Figure 12: Evaluation Window 2 English EDL NIL and Non-NIL Performance Comparison

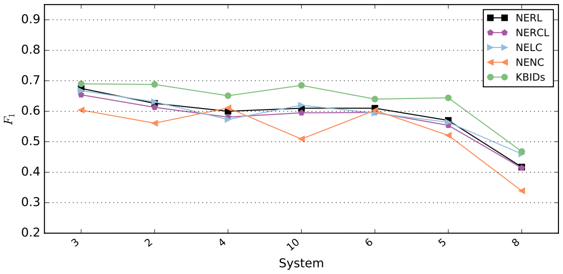


Figure 13: Evaluation Window 2 Chinese EDL NIL and Non-NIL Performance Comparison

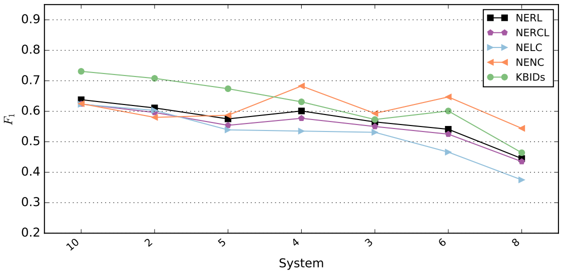


Figure 14: Evaluation Window 2 Spanish EDL NIL and Non-NIL Performance Comparison

Training Data	F-score
CONLL (273, 037 tokens from 2000)	56.7
EDL2015 (129 docs from 2014-2015)	58.2
Ontonotes+EDL	69.6

Table 12: Spanish Name Tagging Performance (%)

lists

- Foreign transliterated name initial/middle/end character lists
- Name gazetteers
- Title words
- Action verbs that are likely to appear before or after animate entities

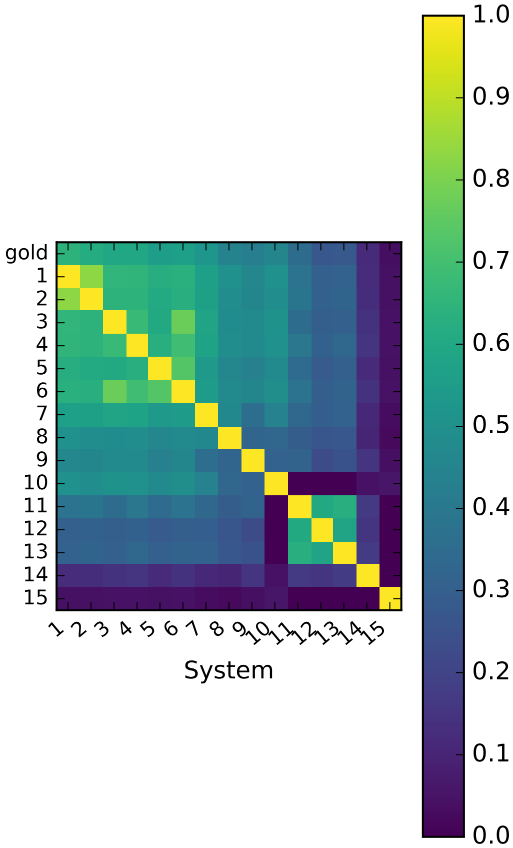


Figure 15: NERLC F_1 when treating true gold standard and each system's output as gold standard.

- Triggers for events involving each type of entity
- Nominal lists
- Conjunction words
- Time expressions

6.3 Cross-lingual Knowledge Transfer

After two years development of EDL resources, Chinese and Spanish are not low-resource languages any more in terms of available labeled data. However, we are still lack of knowledge resources for Chinese and Spanish. For example, there are advanced knowledge representation parsing tools available (e.g., Abstract Meaning Representation (AMR) (Banarescu et al., 2013)) and large-scale knowledge bases for English EDL, but not for Chinese and Spanish. RPI team (Lu et al., 2016) developed an approach to discover comparable documents from the source collection based on cross-lingual topic

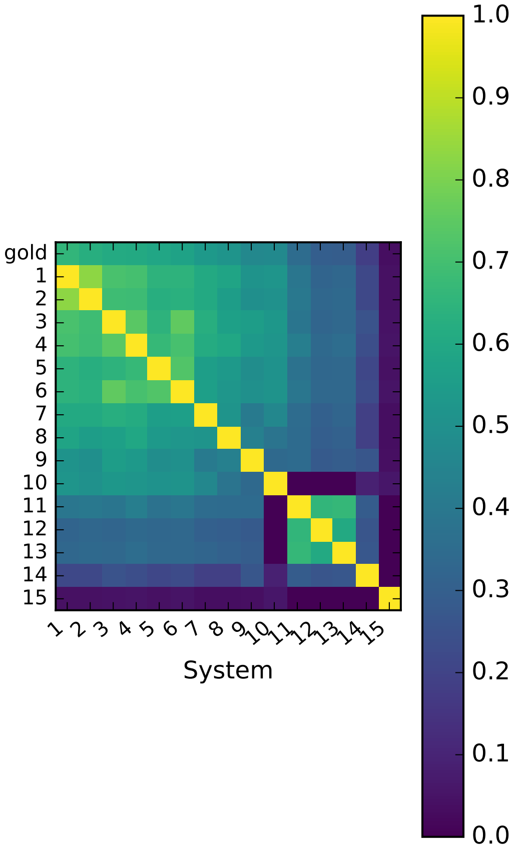


Figure 16: CEAFm F_1 when treating true gold standard and each system's output as gold standard.

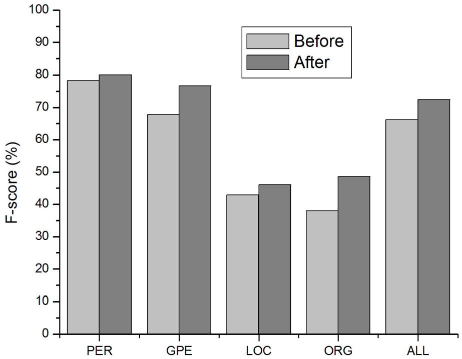


Figure 17: Impact of Language-specific Resources on Chinese Name Tagging

modeling using LDA and bi-lingual lexicons, and project Entity Discovery and Linking results from English documents (Pan et al., 2015) based on Abstract Meaning Representation back to Chinese and Spanish. On the other hand, some local entity mentions may have more concrete evidence in foreign documents for linking. Therefore knowledge from foreign languages can also be leveraged to enhance the performance of English EDL. Figure 18 shows that this approach provides significant improvement for all three languages on all EDL metrics.

6.4 Foreign Language EDL or MT+English EDL

There are two basic approaches to cross-lingual EDL: (1) Foreign Language EDL + Entity Translation; and (2) Full Document Machine Translation (MT) + English EDL. This year all systems except team 10 adopted approach (1). We can see that team 10 achieved competitive results for Spanish but 5.7% lower CEAFmC F-score on Chinese compared to the top performance. This is mainly because Spanish-to-English MT is more mature than Chinese-to-English MT. Team 10 used Google online MT service. In contrast we found that if we use an academic MT system for Spanish EDL, pipeline (1) can still outperform (2), as shown in Table 13.

Training Data	F-score
Spanish Name Tagger	69.6
MT + English name tagging + projection	65.7

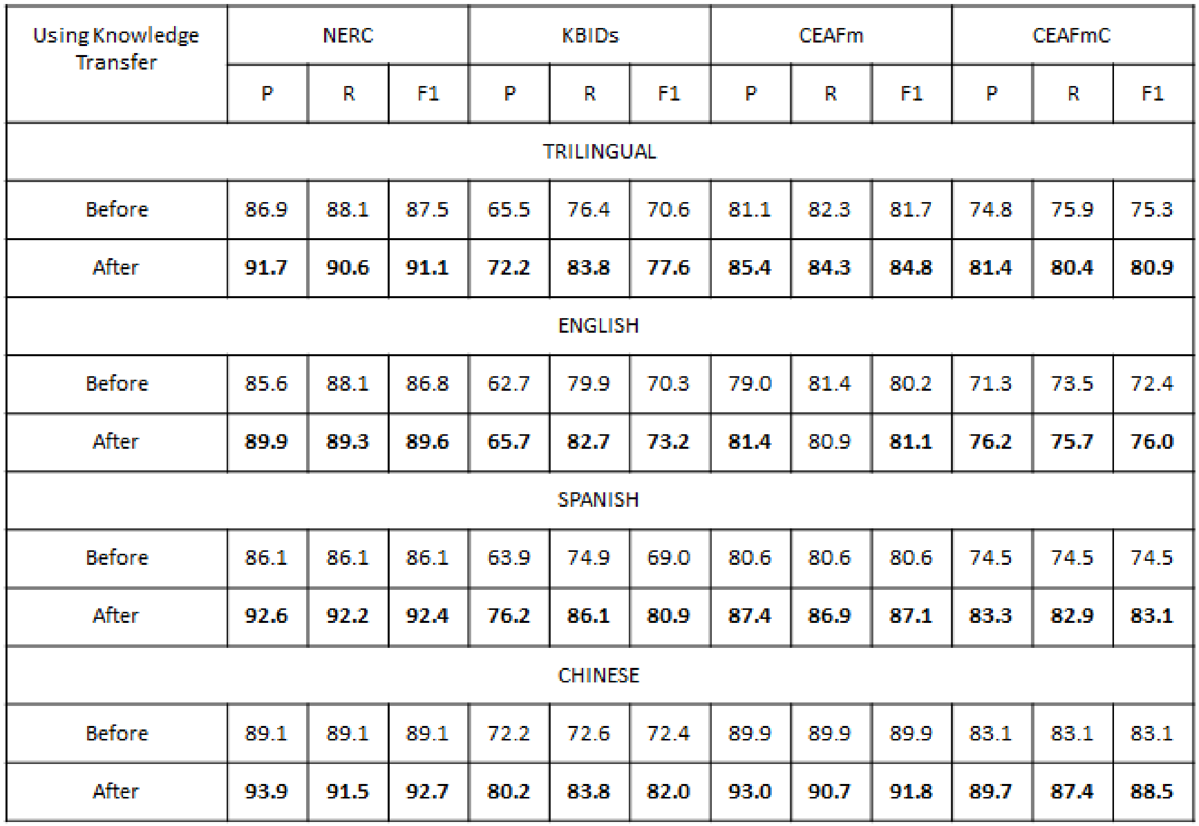
Table 13: Spanish Name Tagging Performance (%)

7 Remaining Challenges

7.1 Entity Mention Extraction

We are making good progress on Entity Mention Extraction (the best F-score is improved from last year's 72.4% to this year's 75.9%), but it still remains the most challenging step in EDL, for all three languages. Morphs remain the most challenging mentions in Chinese data. Partially due to the rapid evolution of language in a social media under censorship, more and more common entities are referred by morph mentions in recent data. For example, "*Polar*" refers to Russia, "*White Head Eagle*" refers to the US, and "*rabbit*" and "*big green fruit*" refers to China.

We need to develop EDL techniques specially for this kind of coded language in social media. The EDL corpora we have annotated so far are still not sufficient to generalize morph mentions that were created by a variety of categories. Coded language defeats most contextual features used by state-of-the-art EDL techniques. Less-mature



Using Knowledge Transfer	NERC			KBIDs			CEAFm			CEAFmC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TRILINGUAL												
Before	86.9	88.1	87.5	65.5	76.4	70.6	81.1	82.3	81.7	74.8	75.9	75.3
After	**91.7**	**90.6**	**91.1**	**72.2**	**83.8**	**77.6**	**85.4**	**84.3**	**84.8**	**81.4**	**80.4**	**80.9**
ENGLISH												
Before	85.6	88.1	86.8	62.7	79.9	70.3	79.0	81.4	80.2	71.3	73.5	72.4
After	**89.9**	**89.3**	**89.6**	**65.7**	**82.7**	**73.2**	**81.4**	80.9	**81.1**	**76.2**	**75.7**	**76.0**
SPANISH												
Before	86.1	86.1	86.1	63.9	74.9	69.0	80.6	80.6	80.6	74.5	74.5	74.5
After	**92.6**	**92.2**	**92.4**	**76.2**	**86.1**	**80.9**	**87.4**	**86.9**	**87.1**	**83.3**	**82.9**	**83.1**
CHINESE												
Before	89.1	89.1	89.1	72.2	72.6	72.4	89.9	89.9	89.9	83.1	83.1	83.1
After	**93.9**	**91.5**	**92.7**	**80.2**	**83.8**	**82.0**	**93.0**	**90.7**	**91.8**	**89.7**	**87.4**	**88.5**

Figure 18: Impact of Cross-lingual Knowledge Transfer for a Cluster of Topically-Related Documents

metaphor detection (Wang et al., 2006; Tsvetkov, 2013; Heintz et al., 2013) techniques also have trouble with abstract coded language. Word-sense disambiguation techniques (Yarowsky, 1995; Mihalcea, 2007; Navigli, 2009) deal in sense inventories that are fairly static over time, whereas code language evolves rapidly. Coded language effectively hides in plain sight because anything is a candidate mention. Moreover, common words and phrases are often used as code words to avoid censorship. For example, "姐夫 *(jie fu)*" can be used to refer to both its literal meaning "*sisterinlaw*" or the Russian polician Medvedev whose name's Chinese transliteration ends with syllables that have similar pronunciations as "*jie fu*". Likewise it's difficult to link or cluster them to the real targets after they are identified.

7.2 Within-document and Cross-document Coreference Resolution

Cross-document entity clustering is a cascaded effect of both within-document and cross-document coreference resolution. This year we introduced two new diagnostic metric, `CEAFm-doc` as a summary of within-document `CEAFm` coreference performance, micro-averaging across all documents, and `CEAFm-1st` to approximate cross-document performance by limiting evaluation to the first mention per document of each predicted and gold entity. From the results we can see that both of them generally correlate with the final CEAFmC scores. Very limited efforts are currently being made on Chinese and Spanish within-document coreference resolution. In the future we should consider preparing more knowledge resources to improve within-document coreference resolution, which is also currently a major bottleneck for slot filling and event coreference resolution.

In fact, within-document coreference resolution has been one of the major reasons for the low precision for knowledge graph construction in many industrial companies such as Google and Microsoft. For example, in the following sentences: "*Jutting into the Long Island Sound with rocky outcroppings, marshy inlets and lush forest, Pelham Bay Park looks more like Maine than the Bronx. The city's largest park at 2,766 acres - three times the size of Central Park - it* *takes hours to explore.*", Google fails to resolve "*the city's largest park*" to "*Pelham Bay Park*" and thus returns a wrong result "*Central Park*" for a query "*the largest park in New York City*".

Within-document coreference resolution errors are also propagated to event coreference resolution. Let's look at the following Chinese sentence: "笔者*(writer)* 到*(went to)*阳城县*(Yangcheng County)*新阳东街*(Xinyang East Street)*营业厅*(business hall)*询问 *(consulted)*，工作人员 *(personnel)* 称*(claimed)* 他们*(they)*会*(will)*在24小时内*(within 24 hours)*抢修通*(fix the tunnel)*.". Without knowing that "*writer*" and "*personnel*" refer to two different entities, an event coreference resolver may mistakenly cluster "*consulted*" and "*claimed*" as the same event.

7.3 Is DNN Working for EDL?

Many teams including IBM, IBM, YorkNRM and ZJU used Deep Neural Networks (DNN) for name tagging. Most systems find DNN provides a general, powerful underlying model for name tagging. RPI team found that compared to Conditional Random Fields (CRFs), DNN usually provides up to 6% higher F-score for various languages, if a massive amount of clean annotated data is available. IBM team had opposite observations - their DNN model got 2%-3% lower performance than CRFs. In any case, such data annotations are expensive to prepare. In order to compensate this data requirement, various automatic annotation generation methods have been attempted in previous work, including knowledge base driven distant supervision, cross-lingual projection and naturally existing noisy annotations such as Wikipedia markups. Annotations produced from these methods are usually very noisy, while DNN is sensitive to noise just like many other machine learning methods. For example, RPI found that the F-score of the same DNN model learned from noisy training data is 45% lower than that trained from clean data for some low-resource languages. Furthermore, DNN based name taggers are not easily portable to a new data set from a different epoch or domain. For Entity Linking, the systems based on unsupervised learning (e.g., team 4) achieved comparable (only 2% lower) performance to DNN based supervised methods (e.g., team 2) trained from a large amount of labeled data.

8 Looking Ahead

The new Tri-lingual EDL task has created many interesting research problems and new directions. In KBP2017 we will consider the following possible extensions and improvement:

- combine with tri-lingual slot filling, event extraction and sentiment/belief anaysis, allow more development time to achieve end-to-end KBP performance.

- Many teams continued to benefit from leveraging the graph structures in the KB. They represent Wikipedia as a directed weighted graph, and then compute the discriminative power of each link type in the graph based on entropy measure or the number of triangles that the link belongs to. What about if the KB is just a list of names without any rich structures (e.g., geonames, product names)?

- Multi-lingual EDL for 300+ languages. One possible way to push the development of language-universal techniques is not to evaluate for 1-3 languages, but for all the languages in the world, at least for those we have Wikipedia entries. Some teams including RPI and Sydney have been cleaning and enriching Wikipedia markups so we can get silver standard annotations for EDL for all the 300+ languages in Wikipedia.

- Multi-media EDL and KBP. Extend the input data from text to multiple data modalities (videos, images, comments, social media, news articles, structured data bases, etc.).

- Encourage teams to produce multiple or n-best hypotheses at all levels, from individual assertions to subgraphs, and all kb assertions to be tagged with meaningful confidence/certainty measures.

- 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 streaming data into the source collection;

Acknowledgments

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*.

J. Ellis, J. Getman, N. Kuster, Z. Song, A. Bies, and S. Strassel. 2016. Overview of linguistic resources for the tac kbp 2016 evaluations: Methodologies and results. In *Proc. Text Analysis Conference (TAC2016)*.

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.

I. Heintz, R. Gabbard, M. Srivastava, D. Barner, D. Black, M. Friedman, and R. Weischedel. 2013. Automatic extraction of linguistic metaphors with LDA topic modeling. In *Proc. ACL2013 Workshop on Metaphor in NLP*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proc. Text Analysis Conference (TAC2015)*.

D. Lu, X. Pan, N. Pourdamghani, S.-F. Chang, H. Ji, and K. Knight. 2016. A multi-media approach to cross-lingual entity knowledge transfer. In *Proc. the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016)*.

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

R. Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2007)*.

R. Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(10):1–69.

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)*.

Y. Tsvetkov. 2013. Cross-lingual metaphor detection using common semantic features. In *Proc. ACL2013 Workshop on Metaphor in NLP*.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In *COLING*.

Z. Wang, H. Wang, H. Duan, S. Han, and S. Yu. 2006. Chinese noun phrase metaphor recognition with maximum entropy approach. In *Proc. of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2006)*.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting on Association for Computational Linguistics (ACL1995)*.