Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking

Heng Ji[1], **Xiaoman Pan**[1], **Boliang Zhang**[1], **Joel Nothman**[2],
James Mayfield[3], **Paul McNamee**[3] **and Cash Costello**[3]
[1] Computer Science Department, Rensselaer Polytechnic Institute
{jih, panx2, zhangb8}@rpi.edu
[2] Sydney Informatics Hub, University of Sydney
joel.nothman@gmail.com
[3] Johns Hopkins University
{mayfield, mcnamee}@jhu.edu
cash.costello@jhuapl.edu

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 TAC2017, and of the Ten Low Resource Language EDL Pilot. We will summarize several new and effective research directions including multi-lingual common space construction for cross-lingual knowledge transfer, rapid approaches for silver-standard training data generation and joint entity and word representation. We will also sketch out remaining challenges and future research directions.

1 Introduction

The Entity Discovery and Linking (EDL) track at TAC-KBP has experienced nine years of joy and prosperity, thanks to the successful community efforts and DARPA and NIST's support at creating valuable resources and shared tasks. Table 1 summarizes the overall progress of EDL research in the last decade. In addition to improved quality at each subtask (mention extraction, linking and NIL clustering), the major recent accomplishment lies in the dramatically enhanced portability. State-of-the-art EDL techniques today can take an arbitrary large-size corpus as input, and extract fine-grained types (16,000+) of entities from hundreds of languages (Pan et al., 2017), and link them to English knowledge bases with either rich properties (e.g., DBPedia) or scarce properties (e.g., World Fact Book or simply a product name list). These techniques are performed in a "*Cold-start*" fashion, without using human-defined schema or manual annotations, and thus they can be easily adapted to a new domain, genre or language.

The secret weapon behind these successes is embracing symbolic semantics and distributional semantics into a unified framework. The key idea is to bottom-up discovery instead of top-down classification by clustering semantically similar entities, and then learning a universal grounding function to assign a type to each cluster. In this way, multiple sources can share a common semantic space and transfer knowledge and resources across related words and entities (Cao et al., 2017), and across thousands of languages (Zhang et al., 2017b).

In TAC-KBP EDL2016, the top Chinese and Spanish systems achieved comparable performance as the top English systems (Ji and Nothman, 2016). However, most of the success was due to clean manual annotation efforts made by LDC or participants. Clean data annotation is often not available for low-resource languages and difficult to obtain during emergent settings. In order to compensate this data requirement, various automatic annotation generation methods have been proposed to create "*Silver Standard*", including knowledge base driven distant supervision, cross-lingual projection, and leveraging naturally existing noisy annotations such as Wikipedia markups (Pan et al., 2017). Compared to the KBP2016 EDL task (Ji and Nothman, 2016), we added 10 low-resource languages as a pilot study, and aim to answer the following research questions:

- How to fill in the performance gap between silver standard and gold standard?
- Can we advance the field by exploring non-traditional linguistic resources which are beyond human data annotation?
- Is there any performance ceiling for cross-lingual EDL? To what extent is it due to the lack of language-specific knowledge?
- Silver-standard annotations are usually very noisy, while many machine learning methods are sensitive to noise. How to make these learning models more robust to noise?

The rest of this paper is structured as follows. Section 2 describes the definition of the Tri-lingual EDL task and the ten languages EDL Pilot. Section 3 briefly summarizes the participants. Section 4 highlights some annotation efforts and elaborate details at preparing data sets for the ten languages EDL task. Section 5 summarize evaluation results for both tasks. Section 6 summarizes new and effective methods, while Section 7 provides some detailed analysis and discussion about remaining challenges. Section 9 sketches our future directions.

2 Task Definition and Evaluation Metrics

This section will summarize the Entity Discovery and Linking tasks conducted at KBP 2017. More details regarding data format and scoring software can be found in the task website[1].

2.1 Tri-lingual EDL 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 three 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. We use the same reference knowledge base as in 2016, namely BaseKB [2]. 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 July as part of the Cold-Start KB construction task and the second in September to check the progress.

2.2 Ten Low Resource Language EDL Pilot

NIST and DARPA chose the following ten low-resource languages for the pilot study, by considering multiple factors including the amount of available resources, language diversity and end user needs: Polish, Chechen, Albanian, Swahili, Kannada, Yoruba, Northern Sotho, Nepali, Kikuyu and Somali. The pilot study generally follows the tri-lingual EDL task specification, but it does not require NIL clustering, and it does not include facility entity type or nominal mentions. Also we link the entity mentions to a Wikipedia dump of March 5, 2016[3], instead of the BaseKB.

2.3 Scoring Metrics

As detailed in Table 2, we report measures for detection of mentions and their types,

[1] http://nlp.cs.rpi.edu/kbp/2017/

[2] http://basekb.com/

[3] http://nlp.cs.rpi.edu/wikiann/

Grow with DARPA DEFT and LORELEI	**2006-2011**	**2012-2017**
Mention Extraction	Human (most)	Automatic
NIL Clustering	None	64 methods
Foreign Languages	Chinese (5%-10% lower than English)	System for 282 languages (Chinese/Spanish comparable to or outperform English); Research toward 3,000 languages
Document Size	-	Extended from 500 to 90,000 documents
Genre	News, web blog	News, Discussion Forum, Web blog, Tweets, Scientific Literature
Entity Types	PER, GPE, ORG	PER, GPE, ORG, LOC, FAC, hundreds/thousands of fine-grained types for typing
Mention Types	Name or all concepts (most)	Name, Nominal, Pronoun
KB	Wikipedia	Freebase, Scarce KB (e.g., Geoname, World Factbook, name list
Training Data	20,000 queries (entity mentions)	from 500 to 0 documents; unsupervised linking comparable to supervised linking
# Good Papers	62	120 (new KBP track at ACL conferences); 6 tutorials at top conferences

Table 1: A Decade of Progress on EDL

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 2: 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, GPE, LOC or FAC. *Kbid* is the KB identifier or NIL.

identification of KB links, and clustering of mentions with or without links. The scorer is available at `https://github.com/wikilinks/neleval`.

2.3.1 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 2, `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.

2.3.2 Clustering metrics

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

2.3.3 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.[4] 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.

2.3.4 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.

[4] This corresponds to Pure-CDEC evaluation in **?**) (personal correspondence).

3 Participants Overview

Table 3 summarizes the participants for KBP2017 EDL tasks. In total 8 teams submitted for the first Tri-lingual EDL evaluation window as part of the cold-start KB construction task, 16 teams submitted runs for the second evaluation window, and 3 teams submitted to the ten languages EDL pilot.

4 Data Annotation and Resources

The details of the data annotation for KBP2017 Tri-lingual EDL are presented in a separate paper by the Linguistic Data Consortium (Getman et al., 2017). In this section we only elaborate how we prepare the ground truth for the ten languages EDL pilot.

For Chechen, Somali and Yoruba, we use LDC released LORELEI LRLPs and REFLEX corpus. The new challenge is that the remaining languages don't have any gold-standard training data annotated by native speakers. For five of them (Albanian, Kannada, Nepali, Polish and Swahili), fortunately we can crawl news data from the Voice of America news website[5], the Kannada Prabha news website[6], the British Broadcasting Corporation news website[7], and the Wiadomoci news website[8]. Then RPI and JHU made a joint effort at developing "*Chinese Room*" interfaces and annotated 50 documents for each language by five non-native speakers. We adjudicated our name tagging annotations and then created silver-standard entity linking through the RPI Chinese Room interface. We were not able to obtain news data for Kikuyu and Northern Sotho, and so we use the Wikipedia derived silver-standard data (Pan et al., 2017). Table 4 summarizes the resources prepared for each language.

Now we elaborate some implementation details about the RPI Chinese Room. More detailed results and analysis about the interface can be found in (Cheung et al., 2017). We applied cross-lingual topic modeling based on lexicons to clustered all news documents, then we selected incident related documents based on the keywords related to the situation frame types defined in the DARPA LORELEI program. We built a "Chinese Room" EDL interface where a foreign language document is displayed, and words and candidate names are translated based on lexicons and gazetteers. A non-native user can also collect and provide his/her knowledge about an IL in the interface, such as name designators. If a language is not written in roman alphabet, we also apply a universal romanizer[9] to display the romanized results. This interface allows a user to identify, classify and translate names in each IL sentence. The interface also allows a user to delete a sentence with low annotation confidence.

The JHUAPL Dragonfly annotation tool, which JHU used to perform its Chinese Room annotations, is similar in spirit to the RPI tool. For each word of the sentence, Dragonfly displays the word itself, a romanization of the word (if necessary) using the uroman tool, any translations of the word from available dictionaries, and any translations of other words in the Brown cluster for that word. Machine translation output (in this case from Google Translate) is also presented. The annotator has the ability to add translations to a local dictionary as they are discovered; these translations are then automatically displayed when new documents are annotated. More information on the Dragonfly annotation tool and the JHU ground truth annotation effort is available in Finin et. al (2017).

Finally, we also devoted a lot of time at collecting related publications and tutorials,[10] resources and software[11] to lower the entry cost for EDL.

5 Evaluation Results

5.1 Overall Performance

Table 5, Table 6 and Table 7 summarize the results. For public release we have anonymized the team names: each team is numbered with the rank of its best submission. Overall the EDL track is a great success again this year, especially that given three years of annotations and resources, the performance of foreign languages (English and Chinese) is comparable to or even better than that of English for various measures. The best end-to-end extraction, linking and clustering performance of Chinese is 4% higher than that of English.

[5] https://www.voanews.com/

[6] http://www.kannadaprabha.com/

[7] http://www.bbc.com/news

[8] http://wiadomosci.gazeta.pl/

[9] https://www.isi.edu/ ulf/uroman.html

[10] http://nlp.cs.rpi.edu/kbp/2017/elreading.html

[11] http://nlp.cs.rpi.edu/kbp/2017/tools.html

Team	Affiliation	Tri-lingual CMN	 ENG	 SPA	10 Languages
		1st Evaluation Window			
A2KD_Adept	Raytheon BBN Technologies	✓	✓		
ICTCAS_OKN	Institute of Computing Technology, Chinese Academy of Sciences		✓		
ISCAS_Sogou	Institute of Software, Chinese Academy of Sciences & Sogou, Inc.	✓			
SAFT_ISI	USC Information Sciences Institute	✓	✓	✓	
STANFORD	Stanford University	✓	✓	✓	
TinkerBell	RPI, UIUC, Stanford, Columbia, Cornell, JHU, UPenn	✓	✓	✓	
hltcoe	Human Language Technology Center of Excellence	✓	✓		
newbie_mr	Machine Reading Co		✓		
		2nd Evaluation Window			
2089Pacific	Individual		✓		
BUPTTeam	Beijing University of Posts and Telecommunications	✓	✓	✓	
Boun	Boğaziči University University		✓		
CMUCS	Language Technologies Institute, Carnegie Mellon University	✓	✓	✓	
CRIM	Computer Research Institute of Montreal		✓		
hltcoe	Human Language Technology Center of Excellence				✓
IBM	IBM Research	✓	✓	✓	✓
IRIS	Paul Sabatier University		✓		
NUDT	College of Computer, National University of Defense Technology	✓	✓	✓	
RPI_BLENDER	Rensselaer Polytechnic Institute	✓	✓	✓	✓
SUMMA	University College London	✓	✓	✓	
TAI	AI platform department of Tencent	✓	✓	✓	
UI_CCG	University of Illinois at Urbana Champaign	✓	✓	✓	
Ugglan	Lund University	✓	✓	✓	
YorkNRM	York University	✓	✓	✓	
rise_dcd_zju	College of Computer Science and Technology, Zhejiang University	✓	✓	✓	
srcb	Ricoh Software Research Center (Beijing) Co.,Ltd.	✓	✓		

Table 3: Runs Submitted by KBP2017 13 Languages Entity Discovery and Linking Participants

Languages	Training	Test	Data Source
Albanian	40 documents	10 documents	Silver+
Chechen	83 documents	30 documents	Gold
Kannada	40 documents	10 documents	Silver+
Kikuyu	1,404 sentences	1,055 sentences	Silver
Nepali	40 documents	10 documents	Silver+
Northern Sotho	1,356 sentences	1,125 sentences	Silver
Polish	40 documents	10 documents	Silver+
Somali	605 documents	50 documents	Gold
Swahili	40 documents	10 documents	Silver+
Yoruba	197 documents	50 documents	Gold

Table 4: 10 Language EDL Resources (Silver: Wikipedia derived annotation; Silver+: Chinese Room; Gold: LDC released annotation)

5.2 Performance Comparison across Types and Genres

Figures 1 and 2 show the performance comparison across different entity types, mention types and genres. It's clear that facility entities and nominal mentions remain the most challenging across systems

5.3 Performance Comparison across Languages

Figure 3 compares the performance across three languages. For the first time, the top Chinese end-to-end EDL performance is 4% higher than English.

6 What's New and What Works

6.1 Joint Name Tagging and Entity Linking

Similar to previous years, joint modeling of name tagging and entity linking continues to show improvement. The MSRA team (Luo et al., 2017) achieved 1.3% name tagging F-score gain by designing one single joint conditional random fields (CRFs) model for joint name tagging and entity linking.

6.2 Joint Word and Entity Embeddings

Similar to the above joint modeling idea, mention extraction and linking, especially typing mentions would benefit tremendously from knowing both of the common words in source context of the mention and the candidate entity's properties and connected entities in the KB. The CMU

Team	NER			NERC			NERLC			KBIDs			CEAFmC+		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Tri-lingual															
5	**83.2**	**67.3**	**74.4**	**76.8**	**62.2**	**68.8**	**62.6**	**50.7**	**56.0**	**73.1**	**64.9**	**68.8**	**60.7**	**49.1**	**54.3**
18	52.8	54.8	53.8	29.8	30.9	30.3	22.6	23.4	23.0	64.1	46.9	54.2	19.7	20.5	20.1
16	81.7	53.0	64.3	71.7	46.5	56.4	5.5	3.5	4.3	0.0	0.0	0.0	4.8	3.1	3.7
Chinese															
5	84.8	62.9	**72.2**	79.6	59.1	**67.8**	**65.1**	**48.3**	**55.4**	79.9	**64.9**	**71.7**	**64.0**	**47.5**	**54.5**
14	75.0	60.5	67.0	70.0	56.5	62.6	47.8	38.5	42.7	**84.4**	38.7	53.1	46.3	37.4	41.4
18	68.2	47.4	55.9	38.8	26.9	31.8	31.5	21.9	25.8	62.3	44.4	51.8	30.6	21.3	25.1
15	79.8	56.2	66.0	73.9	52.0	61.1	14.7	10.3	12.1	0.0	0.0	0.0	13.9	9.8	11.5
20	56.2	**71.5**	63.0	51.7	**65.9**	57.9	9.9	12.7	11.1	0.0	0.0	0.0	8.9	11.4	10.0
16	**85.4**	50.8	63.7	**81.1**	48.3	60.5	5.0	3.0	3.7	0.0	0.0	0.0	4.6	2.8	3.5
English															
5	77.5	66.7	71.7	71.5	61.5	66.1	**57.9**	49.8	**53.5**	63.6	**68.2**	**65.8**	**54.1**	46.5	**50.1**
14	78.6	79.1	**78.8**	72.6	**73.0**	**72.8**	52.9	**53.2**	53.0	**70.4**	49.8	58.4	48.8	**49.1**	49.0
15	73.0	**79.5**	76.1	66.1	71.9	68.9	23.2	25.3	24.2	0.0	0.0	0.0	21.1	22.9	22.0
21	**90.8**	62.5	74.1	**83.3**	57.3	67.9	26.9	18.5	21.9	0.0	0.0	0.0	23.5	16.2	19.2
18	55.9	70.5	62.4	31.7	39.9	35.3	19.5	24.6	21.8	66.9	50.5	57.6	16.0	20.2	17.9
16	78.5	48.9	60.3	71.3	44.5	54.8	7.8	4.9	6.0	0.0	0.0	0.0	7.0	4.4	5.4
24	51.5	32.9	40.1	29.7	19.0	23.2	5.2	3.3	4.0	0.0	0.0	0.0	4.9	3.1	3.8
Spanish															
5	**86.6**	**74.3**	**80.0**	**78.5**	**67.4**	**72.5**	**64.1**	**55.0**	**59.2**	**76.4**	**62.1**	**68.5**	**62.8**	**53.9**	**58.0**
18	40.9	50.4	45.1	22.7	28.0	25.1	19.9	24.6	22.0	64.0	46.6	53.9	16.2	20.0	17.9
16	84.9	58.7	69.4	63.5	43.9	51.9	5.2	3.6	4.2	0.0	0.0	0.0	4.5	3.1	3.7

Table 5: Overall Tri-lingual Entity Discovery and Linking Performance (%) during the First Evaluation Window.

team (Ma et al., 2017) and RPI team (Zhang et al., 2017b) leaned joint word and embeddings, significantly improved both mention extraction and entity linking. The RPI system followed a Multi-Prototype Mention Embedding model proposed by (Cao et al., 2017).

6.3 Return of Supervised Models

From 2009 to 2017 TAC-KBP has provided the community substantial amount of annotations for both mention extraction (1,500+ documents) and entity linking (5,000+ query entities). Along with resources developed by other programs such as ACE, CONLL, OntoNotes and ERE, supervised models have become popular again this year for each step of EDL (Sil et al., 2017).

For name tagging, generally distributional semantic features are more effective than symbolic semantic features (Celebi and Ozgur, 2017), while combining them significantly enhanced both of the quality and robustness to noise for low-resource languages (Zhang et al., 2017b; Zhang et al., 2017a).

More teams (Sil et al., 2017; Moreno and Grau, 2017; Yang et al., 2017) have returned to supervised models to rank candidate entities for entity linking. The new neural entity linker designed by IBM (Sil et al., 2017) achieved higher entity linking accuracy than state-of-the-art on the KBP2010 data set.

6.4 Corpus-level Coherence for NIL Clustering

The traditional way of measuring coherence is applied to document-level. Namely that multiple mentions which connected in the source document should be linked to entities in the KB which are also strongly connected to each other. The SUMMA team (Mendes et al., 2017) designed a new method to measure coherence based on corpus-level and achieved 1.7% absolute gain on CEAFmC F-score.

6.5 Chinese Room

For five low-resource languages, the IBM team (Sil et al., 2017) mainly used the silver-standard annotations derived from Wikipedia markups (Pan et al., 2017) for training. In contrast the RPI team (Zhang et al., 2017b) and the JHU HLT-COE team used Chinese Room interfaces to annotate silver-standard for training data that has the same genre as the evaluation data. From the final results we can see that the in-domain Chinese Room annotations are more effective than Wikipedia derived annotations, achieving 26% higher mention extraction F-score and 8% higher extraction and linking F-score.

Team	NER			NERC			NERLC			KBIDs			CEAFmC+		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Tri-lingual															
1	88.5	71.4	79.0	85.0	68.6	75.9	76.0	**61.3**	**67.8**	78.7	**73.7**	**76.1**	75.4	**60.9**	**67.4**
2	**91.9**	65.0	76.1	**88.2**	62.4	73.1	**81.2**	57.5	67.3	80.5	70.1	75.0	**79.0**	55.9	65.5
3	88.8	64.5	74.7	85.0	61.7	71.5	69.6	50.6	58.6	**81.4**	61.4	70.0	68.7	49.9	57.8
7	83.8	**75.7**	**79.6**	80.8	**72.9**	**76.7**	64.4	58.2	61.1	72.2	64.8	68.3	59.4	53.6	56.4
6	89.4	58.4	70.6	83.0	54.3	65.6	74.9	48.9	59.2	80.1	61.7	69.7	70.9	46.4	56.1
4	87.7	61.5	72.3	81.2	57.0	67.0	71.5	50.1	58.9	68.2	61.2	64.5	67.8	47.5	55.9
8	89.4	60.2	71.9	85.8	57.8	69.1	75.3	50.8	60.7	74.5	65.2	69.5	67.4	45.4	54.3
9	79.5	62.7	70.1	74.3	58.6	65.5	59.8	47.2	52.8	74.1	60.3	66.5	58.3	45.9	51.4
10	88.5	67.7	76.7	85.2	65.3	73.9	68.4	52.4	59.3	76.2	63.1	69.0	58.3	44.6	50.5
11	83.4	51.3	63.5	76.0	46.7	57.9	66.8	41.1	50.9	64.3	47.4	54.5	62.6	38.5	47.6
12	80.2	64.5	71.5	72.5	58.3	64.6	38.9	31.2	34.6	32.0	39.2	35.2	38.3	30.8	34.1
Chinese															
1	89.4	**71.3**	**79.3**	87.1	**69.5**	**77.3**	80.0	**63.8**	**71.0**	85.3	**76.8**	**80.8**	79.3	**63.3**	**70.4**
2	90.6	68.1	77.8	87.4	65.7	75.0	**81.3**	61.1	69.8	82.1	74.8	78.3	**79.8**	60.0	68.5
4	**92.2**	62.5	74.5	**88.0**	59.6	71.1	80.3	54.4	64.9	80.3	68.0	73.7	79.1	53.6	63.9
8	87.6	58.1	69.9	85.3	56.6	68.0	77.3	51.3	61.7	82.5	68.1	74.6	76.3	50.6	60.9
7	82.4	69.6	75.4	79.6	67.3	72.9	67.4	57.0	61.7	75.5	66.5	70.7	65.7	55.5	60.2
10	86.6	59.8	70.8	84.1	58.1	68.7	71.2	49.2	58.2	78.6	63.6	70.3	68.8	47.5	56.2
6	89.1	57.4	69.8	82.4	53.1	64.6	72.3	46.5	56.6	**85.4**	64.3	73.4	71.2	45.8	55.8
3	85.2	60.2	70.6	81.5	57.6	67.5	62.4	44.1	51.7	80.4	57.7	67.2	61.8	43.7	51.2
13	90.4	54.8	68.3	86.9	52.7	65.6	67.6	41.0	51.0	55.3	57.2	56.2	66.7	40.5	50.4
11	78.4	47.3	59.0	72.2	43.6	54.4	62.4	37.7	47.0	62.3	44.4	51.8	61.4	37.1	46.2
9	68.4	42.4	52.4	62.0	38.4	47.5	49.7	30.8	38.0	70.0	47.9	56.8	49.4	30.6	37.8
12	80.1	53.5	64.2	75.2	50.3	60.3	41.3	27.6	33.1	32.4	36.5	34.3	40.9	27.3	32.8
English															
3	**92.3**	68.0	78.3	**88.6**	65.3	75.2	**78.7**	58.1	66.8	82.8	70.7	76.3	**78.1**	57.6	**66.3**
1	85.0	**84.8**	**84.9**	80.3	**80.1**	80.2	68.5	**68.3**	**68.4**	76.0	**78.5**	**77.2**	66.2	**66.0**	66.1
2	90.0	69.4	78.4	85.6	66.0	74.6	78.1	60.3	68.0	72.9	76.2	74.5	73.9	57.0	64.4
9	85.4	78.7	81.9	81.5	75.2	78.2	68.1	62.8	65.3	75.9	73.6	74.8	65.7	60.5	63.0
7	89.9	72.1	80.0	87.6	70.3	78.0	74.2	59.5	66.1	79.1	68.5	73.4	67.1	53.8	59.8
17	87.0	74.9	80.5	82.6	71.1	76.4	66.0	56.8	61.0	64.9	60.7	62.7	64.3	55.4	59.5
6	90.6	65.0	75.7	83.8	60.1	70.0	74.3	53.3	62.1	82.3	62.1	70.7	69.4	49.8	58.0
8	89.0	63.4	74.0	85.2	60.7	70.9	73.2	52.1	60.9	69.2	68.2	68.7	68.6	48.9	57.1
10	89.6	75.7	82.0	86.1	72.7	78.9	67.7	57.2	62.0	74.6	68.0	71.1	60.8	51.4	55.7
4	89.4	63.6	74.3	83.4	59.4	69.4	71.1	50.6	59.1	67.0	62.6	64.7	66.6	47.4	55.4
11	88.0	58.0	69.9	79.3	52.3	63.0	69.0	45.5	54.8	66.9	52.1	58.6	63.7	42.0	50.7
13	87.6	73.3	79.8	81.7	68.3	74.4	58.9	49.3	53.6	43.7	66.6	52.8	54.9	45.9	50.0
19	87.9	80.3	84.0	85.0	77.6	**81.1**	60.7	55.4	57.9	**85.4**	48.0	61.5	50.7	46.3	48.4
22	83.3	48.2	61.1	71.6	41.4	52.5	67.1	38.8	49.1	68.3	44.6	53.9	57.9	33.5	42.5
12	78.4	72.1	75.1	69.4	63.8	66.5	34.2	31.4	32.7	27.1	37.5	31.5	33.4	30.7	32.0
23	67.0	38.0	48.5	47.9	27.2	34.7	33.7	19.1	24.4	45.2	50.5	47.7	33.2	18.8	24.0
Spanish															
1	89.1	69.5	78.1	85.5	66.7	74.9	74.2	**57.8**	**65.0**	75.3	**67.5**	71.2	73.9	**57.6**	**64.8**
2	92.3	60.1	72.8	88.4	57.6	69.7	**80.2**	52.3	63.3	81.6	63.3	**71.3**	**79.6**	51.9	62.8
3	90.4	67.2	77.1	86.2	64.1	73.5	70.7	52.5	60.3	81.2	57.4	67.3	69.1	51.4	59.0
9	84.5	**76.1**	**80.1**	79.1	**71.2**	75.0	62.7	56.4	59.4	78.4	64.8	70.9	60.7	54.7	57.5
10	**92.7**	58.7	71.8	88.7	56.1	68.8	77.3	48.9	59.9	73.6	60.1	66.1	72.9	46.1	56.5
6	88.5	59.1	70.8	84.8	56.6	67.9	76.6	51.1	61.3	**83.9**	59.5	69.6	70.0	46.7	56.0
8	92.2	60.2	72.9	87.2	56.9	68.9	74.7	48.8	59.0	71.2	59.7	64.9	68.0	44.4	53.7
7	91.9	69.6	79.2	**89.0**	67.4	**76.7**	66.6	50.4	57.4	69.6	57.9	63.2	59.5	45.0	51.3
11	85.7	50.4	63.5	78.0	45.9	57.8	70.8	41.7	52.5	64.0	46.6	53.9	63.1	37.1	46.7
4	80.0	58.1	67.3	70.0	50.9	58.9	59.9	43.5	50.4	57.1	52.5	54.7	53.7	39.0	45.2
12	82.1	72.8	77.2	72.5	64.3	68.2	40.9	36.2	38.4	36.1	43.5	39.5	40.4	35.7	37.9

Table 6: Overall Tri-lingual Entity Discovery and Linking Performance (%) during the Second Evaluation Window.

6.6 Common Semantic Space

The RPI team (Zhang et al., 2017b) developed a common semantic space to allow multiple languages to share distributed representations. They designed a multi-level, multi-encoder, multi-decoder framework. They extended the auto-encoder from monolingual semantic space projection to multilingual common semantic space construction by incorporating rich syntactic and grammatic knowledge from available linguistic resources. This common

Language	Team	NER			NERC			NERLC			KBIDs		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
10 Languages	8	**84.2**	**73.9**	**78.7**	**80.1**	**70.2**	**74.8**	**70.5**	**61.9**	**65.9**	**62.9**	**53.9**	**58.1**
	3	57.5	49.0	52.9	53.1	45.3	48.9	43.6	37.2	40.1	52.1	48.4	50.2
	14	82.0	17.9	29.4	73.9	16.2	26.5	58.2	12.7	20.9	61.5	25.1	35.6
Polish	8	**75.8**	51.2	61.1	**66.0**	44.6	53.2	**56.2**	38.0	**45.3**	**58.4**	58.4	**58.4**
	14	70.7	63.8	67.0	58.7	53.0	55.7	42.1	38.0	39.9	51.9	**61.1**	56.1
	3	71.9	**70.4**	**71.1**	64.1	**62.7**	**63.4**	39.5	**38.7**	39.1	53.2	58.4	55.7
Somali	8	**81.5**	**78.2**	**79.8**	**80.2**	**76.9**	**78.5**	**57.3**	**54.9**	**56.0**	65.5	**59.9**	**62.6**
	3	50.1	35.1	41.3	46.4	32.5	38.2	32.8	23.0	27.1	**67.0**	41.0	50.8
Northern Sotho	8	**90.6**	**91.3**	**91.0**	**90.4**	**91.2**	**90.8**	**85.1**	**85.8**	**85.5**	81.3	55.4	65.9
	3	46.4	44.7	45.5	42.9	41.3	42.1	38.7	37.3	38.0	**81.9**	**68.6**	**74.7**
Albanian	14	**89.2**	**80.5**	**84.6**	**80.0**	**72.2**	**75.9**	**60.1**	**54.2**	**57.0**	63.4	**66.7**	65.0
	8	84.3	77.0	80.5	78.1	71.4	74.6	59.0	53.9	56.3	**68.0**	64.5	**66.2**
	3	78.9	56.5	65.9	65.7	47.1	54.9	37.7	27.0	31.5	56.3	40.7	47.2
Kannada	8	**79.1**	56.3	**65.8**	67.3	47.9	56.0	**52.9**	**37.7**	**44.0**	69.1	**48.7**	**57.1**
	14	75.3	**56.7**	64.7	**67.9**	**51.2**	**58.4**	46.3	34.9	39.8	**70.2**	42.3	52.8
	3	57.6	35.3	43.8	42.4	26.0	32.3	23.5	14.4	17.9	61.3	24.4	34.9
Chechen	8	**63.8**	**51.4**	**57.0**	**62.1**	**50.1**	**55.4**	**58.9**	**47.5**	**52.6**	**0.0**	**0.0**	**0.0**
	3	0.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	**0.0**	**0.0**
Nepali	14	**75.2**	**59.0**	**66.1**	**73.9**	**58.0**	**65.0**	**57.8**	**45.4**	**50.8**	61.0	**57.1**	**59.0**
	8	59.6	42.4	49.6	52.7	37.6	43.9	41.8	29.8	34.8	75.0	38.1	50.5
	3	55.2	25.9	35.2	50.0	23.4	31.9	49.0	22.9	31.2	**84.4**	42.9	56.8
Yoruba	8	**76.9**	**54.0**	**63.4**	**60.0**	**42.1**	**49.5**	**43.2**	**30.3**	**35.6**	**50.9**	**48.8**	**49.8**
	3	53.1	43.7	48.0	47.5	39.1	42.9	32.0	26.4	28.9	44.3	42.8	43.5
Kikuyu	8	**93.9**	84.3	**88.8**	**93.8**	84.2	**88.7**	**93.8**	84.2	**88.7**	**75.9**	35.2	**48.1**
	3	72.7	**90.1**	80.5	72.6	**89.9**	80.3	72.0	**89.2**	79.7	24.6	**56.8**	34.3
Swahili	14	83.1	**80.0**	**81.5**	75.6	**72.8**	**74.2**	66.5	**64.1**	**65.3**	63.2	**67.2**	**65.2**
	8	**84.2**	68.2	75.4	**75.7**	61.3	67.8	**67.1**	54.3	60.0	**65.8**	56.6	60.9
	3	72.8	70.4	71.6	67.5	65.3	66.4	51.8	50.2	51.0	55.8	57.0	56.4

Table 7: Overall 10 Languages Pilot Entity Discovery and Linking Performance (%).

semantic space significantly improved the name tagging performance for languages like Chechen by borrowing resources and knowledge from Russian.

For Tri-lingual EDL, many mention extraction systems used character embeddings. The common semantic space can further expand the positive impact of character embeddings from high-resource languages to low-resource languages. Mentions referring to the same entity across languages may share a set of similar characters, e.g., Semsettin Gunaltay (English) = emsettin Günaltay (Turkish) = Semsetin Ganoltey (Somali). The RPI system (Zhang et al., 2017b) composed word embeddings from shared character embeddings using Convolutional Neural Networks (CNN). Word embeddings learned in this way achieved significant improvement compared to learning word embeddings directly from text using word as a basic unit.

6.7 Impact of Name Translation

Some low-resource languages such as Tagalog and Swahili tend to include many English words in code-switch form or borrowed words which look similar to English. So it raises a natural question - if name translation is still helping cross-lingual EDL for these languages? We replaced the name translation component with a string match method between foreign language name and English Wikipedia title in the RPI system. The results are shown in Table 8.

took out name translation component from ELISA cross-lingual EDL system, replaced it with direct string match against English Wikipedia titles, and kept all other components (salience, similarity, coherence for disambiguation etc.). Here are some numbers (extraction+linking F-scores) to assure you that name translation is super crucial for cross-lingual EDL:-)

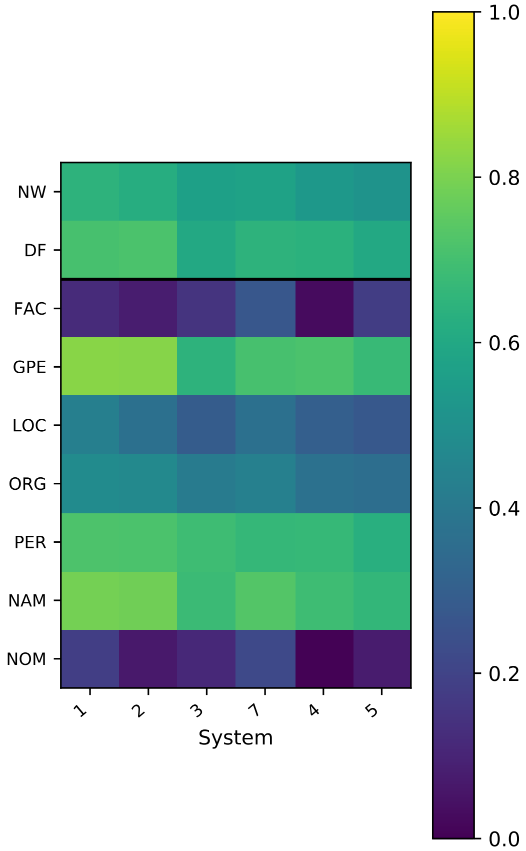

Figure 1: Breakdown Entity Mention Extraction and Linking Performance for Entity Types and Genres.

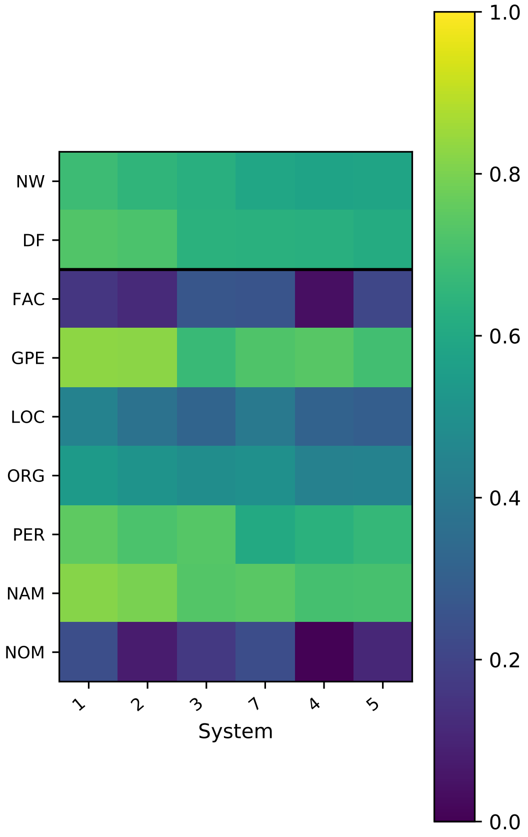

Figure 2: Breakdown Entity Mention Extraction and Clustering Performance for Entity Types and Genres.

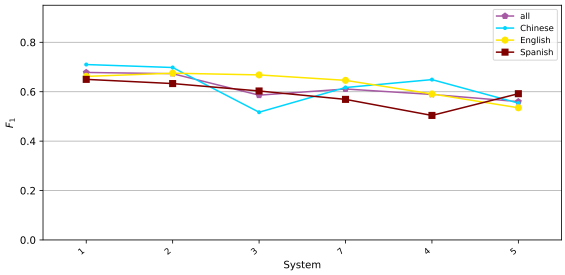

Figure 3: Tri-lingual EDL Performance Comparison.

7 Remaining Challenges

7.1 Duplicability Problem about DNN

Among all of the supervised learning frameworks for mention extraction this year, the most popular one is a combined Deep Neural Networks architecture consisted of Bi-directional Long Short-Term Memory networks (Bi-LSTM) (Graves et al., 2013) and CRFs (Lample et al., 2016). This framework fits the problem of name tagging because: (1). Predicting the tag for each token needs evidence from both of its previous context and future context in the entire sentence. Bi-LSTM networks meet this need by processing each sequence in both directions with two separate hidden layers, which are then fed into the same output layer. (2). There are strong classification dependencies among name tags in a sequence. For example, “I-LOC” cannot follow “B-ORG”. CRFs model, which is particularly good at jointly modeling tagging decisions, can be built on top of the Bi-LSTM networks.

Many teams (Zhao et al., 2017; Bernier-Colborne et al., 2017; Zhang et al., 2017b; Li et al., 2017; Mendes et al., 2017; Yang et al., 2017) trained this framework from the same training data (KBP2015 and KBP2016 EDL corpora) and the same set of features (word and entity embeddings), but got very different results. They are ranked at the 1st, 2nd, 4th, 11th, 15th, 16th, 21st respectively. The mention extraction

Languages	w/o NT	w/t NT
Albanian	18.8%	58.8%
Kannada	7.1%	46.7%
Nepali	3.4%	38.7%
Northern Sotho	20.4%	85.6%
Polish	13.7%	49.5%
Somali	37.8%	56.7%
Swahili	50.3%	63.7%
Yoruba	39.8%	42.7%

Table 8: Impact of Name Translation (NT) (Extraction+Linking F-scores)

F-score gap between the best system and the worst system is about 24%. The reasons that cause these different scores still need to be figured out by detailed analysis. However, DNN requires engineering efforts at tuning hyper-parameters. We should require each team who has adopted DNN framework to report detailed model configurations, any additional training data and dictionary resources (and share them with the community), and details at learning embeddings, and present detailed qualitative analysis on results instead of just reporting performance numbers.

7.2 Challenges for Low-resource Languages

The pilot study on ten low-resource languages showed promising results. However, overall the end-to-end EDL score for these ten languages is about 15% lower than the best Tri-lingual EDL score for three high-resource languages.

Both of the RPI team (Zhang et al., 2017b) and JHU HLT-COE team developed *Chinese Room* interfaces to allow non-native speakers to annotate name tagging and translation for low-resource languages. However, it was difficult to further boost the performance by adding more annotations by non-native speakers through the Chinese Room, possibly because all low-hanging fruits were already picked, while the limited coverage of lexicon and automatic romanized form did not provide non-native speakers enough support to achieve high recall.

7.3 Entity Linking Still Lacks of Background Knowledge

Addressing most of the remaining entity linking errors still requires deep background knowledge discovery from English Wikipedia and large English corpora. Some examples as follows.

- *Before 2000, the regional capital of Oromia was Addis Ababa, also known as "Finfinne".* It's from the text description of "*Oromia Region*" entry in Wikipedia, which teaches us "*Finfinne*" can be linked to "*Addis Ababa*" in the KB.

- *The armed Oromo units in the Chercher Mountains were adopted as the military wing of the organization, the Oromo Liberation Army or OLA.* It's from the text description of "*Oromo Liberation Front*" entry in Wikipedia, which teaches us "*WBO (Oromo Liberation Army)*" is part of "*ABO (Oromo Liberation Front)*" and thus they refer to two different entities.

- The names of the same region may have got frequently changed in the history. The same name mention may refer to different entities at different time points. For example, the Wikipedia entry for "*Jimma Horo*" teaches us that *Jimma Horo may refer to the following: Jimma Horo, East Welega, former woreda (district) in East Welega Zone, Oromia Region, Ethiopia; Jimma Horo, Kelem Welega, current woreda (district) in Kelem Welega Zone, Oromia Region, Ethiopia.* So we would really need to figure out what kind of events and situations these mentions were involved, at what time, in order to be able to correctly linking and clustering them. This is even the major challenge that the state-of-the-art English entity linking is still facing.

- *EPRDF* = *OPDO* + *ANDM* + *SEPDM* + *TPLF* because of the following facts described in Wikipedia articles:
  - *EPRDF*: Ethiopian People's Revolutionary Democratic Front, also called *Ehadig*.
  - *OPDO*: Oromo Peoples' Democratic Organization.
  - *ANDM*: Amhara National Democratic Movement.
  - *SEPDM*: Southern Ethiopian People's Democratic Movement
  - *TPLF*: Tigrayan People's Liberation Front, also called *Weyane* or *Second Weyane*, perhaps because there was a rebellion group called Woyane/Weyane in the Tigray province in 1943.

- *Qeerroo* is not an organization although it has its own website, based on what's described in news articles:
  - *The overwhelming belief is that its leaders are handpicked by the TPLF puppet-masters, and the new generation of Oromo youth known as the ' Qeerroo ' have seen that it is business as usual after the latest reform.*
  - *The Qeerroo, also called the Qubee generation, first emerged in 1991 with the participation of the Oromo Liberation Front (OLF) in the transitional government of Ethiopia. In 1992 the Tigrayan-led minority regime pushed the OLF out of government and the activist networks of Qeerroo gradually blossomed as a form of Oromummaa or Oromo nationalism.*
  - *Today the Qeerroo are made up of Oromo youth. These are predominantly students from elementary school to university, organising collective action through social media. It is not clear what kind of relationship exists between the group and the OLF. But the Qeerroo clearly articulate that the OLF should replace the Tigrayan-led regime and recognise the Front as the origin of Oromo nationalism.*

- "*Somali (Somali region)*", "*Somalia*" and "*Somaliland*" refer to three different entities:
  - *The Ethiopian Somali Regional State (Somali: Dawlada Deegaanka Soomaalida Itoobiya) is the easternmost of the nine ethnic divisions (kililoch) of Ethiopia.*
  - *Somalia, officially the Federal Republic of Somalia(Somali: Jamhuuriyadda Federaalka Soomaaliya), is a country located in the Horn of Africa.*
  - *Somaliland (Somali: Somaliland), officially the Republic of Somaliland (Somali: Jamhuuriyadda Somaliland), is a self-declared state internationally recognised as an autonomous region of Somalia.*

8 Resources

The EDL community has been sharing many valuable systems, resources and data sets. It has become difficult to keep track of them, but many of them will have pointers in the EDL resource page [12]. RPI's cross-lingual EDL system for 282 languages, including the models developed under KBP2017 are all publicly available for research purpose: system APIs [13], trained models, data sets and resources [14], online live EDL demo [15], and heatmap demo [16].

9 Looking Ahead

In KBP2018 and beyond, the following research directions will be worth exploring.

- **Bridge the performance gap between high-resource languages and low-resource languages**: the great progress of Chinese and Spanish EDL mainly benefits from several years of resource development under KBP. In emergent situations we may need to rapidly develop a cross-lingual EDL system within a couple of days or even hours. More research needs to be done to relieve the reliance on training data and resources. Chinese Room is a creative and effective idea, but it reaches performance ceiling quickly, and moves the human labor from data annotation to interface development to some extent. The most promising direction seems to build a large-scale common semantic space for knowledge and resource transfer. In the meanwhile we do need gold-standard data to validate and measure our research progress. Perhaps we could ask LDC and other resource providers to prepare lots of development and test sets in lots of languages.

- **Multi-media EDL**: It's an exciting extension from text-only to multiple data modalities (text, speech, image and video). And text EDL techniques seem mature enough for this extension. But we need to plan out carefully: how to build a common cross-media schema? What type of entity

[12] http://nlp.cs.rpi.edu/kbp/2017/tools.html

[13] http://blender02.cs.rpi.edu:3300/elisa_ie/api

[14] http://nlp.cs.rpi.edu/wikiann/

[15] http://blender02.cs.rpi.edu:3300/elisa_ie

[16] http://blender02.cs.rpi.edu:3300/elisa_ie/heatmap

mentions should we focus on (named entities like linking Obama's picture to his KB entry, or all concepts including nominals such as 'riot police' which appear more frequently in images and videos, or a mixture of both, such as 'Korean ferry')? How much inference is needed and should be required (e.g., should we link a banner of 'Occupy Wall Street' to 'New York City'? should we link NIST building to NIST?)?

- **Extended entity type**: let's extend the number of entity types from five to thousands, so EDL can be utilized to enhance other NLP tasks such as Machine Translation. The English tokens in Wikipedia with YAGO entity types occupy 10% vocabulary.

- **Streaming Data**: We have been talking about how important and timely to move from batch mode to streaming data for years, let's just start to do that. It also brings several new and exciting research problems: How to perform extraction, linking and clustering at real-time? how to dynamically adjust measures and construct/update KB? In addition, clustering must be more efficient than agglomerative clustering techniques that require O(n2) space and time; and smarter collective inference strategy is required for taking advantage of evidence in both local context and global context.

- **Submit systems instead of results**: So our techniques are more duplicable and more resources can be shared with a wider community.

Acknowledgments

Heng Ji's work has been supported by U.S. DARPA DEFT Program No. FA8750-13-2-0041, LORELEI Program No. HR0011-15-C-0115, and ARL NS-CTA No. W911NF-09-2-0053. 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.

This project uses the universal romanizer software 'uroman' written by Ulf Hermjakob, USC Information Sciences Institute (2015-2016).

References

Gabriel Bernier-Colborne, Caroline Barriere, and Pierre Andre Menard. 2017. Crim's systems for the tri-lingual entity detection and linking task. In *Proc. TAC2017*.

Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proc. ACL2017*.

Arda Celebi and Arzucan Ozgur. 2017. Description of the boun system for the trilingual entity detection and linking tasks at tac kbp 2017. In *Proc. TAC2017*.

Leon Cheung, Thamme Gowda, Ulf Hermjakob, Nelson Liu, Jonathan May, Alexandra Mayn, Nima Pourdamghani, Michael Pust, Kevin Knight, Nikolaos Malandrakis, Pavlos Papadopoulos, Anil Ramakrishna, Karan Singla, Victor Martinez, Colin Vaz, Dogan Can, Shrikanth Narayanan, Kenton Murray, Toan Nguyen, David Chiang, Xiaoman Pan, Boliang Zhang, Ying Lin, Di Lu, Lifu Huang, Kevin Blissett, Tongtao Zhang, Heng Ji, Ondrej Glembek, Murali Karthick Baskar, Santosh Kesiraju, Lukas Burget, Karel Benes, Igor Szoke, Karel Vesely, Jan "Honza" Cernocky, Camille Goudeseune, Mark Hasegawa Johnson, Leda Sari, Wenda Chen, and Angli Liu. 2017. Elisa system description for lorehlt 2017. In *Proc. LoReHLT2017*.

Tim Finin, Dawn Lawrie, James Mayfield, Paul McNamee, and Cash Costello. 2017. HLTCOE participation in TAC KBP 2017: Cold Start TEDL and low-resource EDL. In *Proc. Text Analysis Conference (TAC2017)*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2017. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *Proc. Text Analysis Conference (TAC2017)*.

Alan Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding, 2013 IEEE Workshop on*.

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.

Heng Ji and Joel Nothman. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end kbp. In *Proc. Text Analysis Conference (TAC2016)*.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceeddings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*.

Zhenzhen Li, Qun Zhang, Ting Li, Jun Xu, and Dawei Feng. 2017. A hybrid model for trilingual entity discovery and linking tasks at tac kbp 2017. In *Proc. TAC2017*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2017. Joint named entity recognition and disambiguation. In *Proc. TAC2017*.

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

Xuezhe Ma, Nicolas Fauceglia, Yiu chang Lin, and Eduard Hovy. 2017. Cmu system for entity discovery and linking at tac-kbp 2017. In *Proc. TAC2017*.

Afonso Mendes, David Nogueira, Samuel Broscheit, Filipe Aleixo, Pedro Balage, Rui Martins, Sebastiao Miranda, and Mariana S. C. Almeida. 2017. Summa at tac knowledge base population task 2017. In *Proc. TAC2017*.

Jose G. Moreno and Brigitte Grau. 2017. Iris&limsi at tac kbp trilingual entity discovery and linking 2017. In *Proc. TAC2017*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*.

Avirup Sil, Georgiana Dinu, Gourab Kundu, and Radu Florian. 2017. The ibm systems for entity discovery and linking at tac 2017. In *Proc. TAC2017*.

Tao Yang, Dong Du, and Feng Zhang. 2017. The tai system for trilingual entity discovery and linking track in tac kbp 2017. In *Proc. TAC2017*.

Boliang Zhang, Di Lu, Xiaoman Pan, Ying Lin, Halidanmu Abudukelimu, Heng Ji, and Kevin Knight. 2017a. Embracing non-traditional linguistic resources for low-resource language name tagging. In *Proc. IJCNLP2017*.

Boliang Zhang, Xiaoman Pan, Ying Lin, Tongtao Zhang, and Heng Ji. 2017b. Rpi_blender tac-kbp2017 multi-lingual edl system description. In *Proc. TAC2017*.

Huasha Zhao, Yi Yang, Qiong Zhang, and Luo Si. 2017. Improve neural mention detection and classification via enforced training and inference consistency. In *Proc. TAC2017*.