The IBM Systems for Trilingual Entity Discovery and Linking at TAC 2015

Avirup Sil and **Georgiana Dinu** and **Radu Florian**
IBM T.J. Watson Research Center
1101 Kitchawan Rd
Yorktown Heights, NY 10598
{avi,gdinu,raduf}@us.ibm.com

Abstract

This paper describes the IBM systems for the Trilingual Entity Discovery and Linking (EDL) for the TAC 2015 Knowledge-Base Population track. The entity discovery or mention detection (MD) system is based on system combination of deep neural networks and conditional random fields. The entity linking (EL) system is based on a language independent probabilistic disambiguation model. The same EL model was applied across all 3 languages: English , Spanish and Chinese. We submitted 4 runs for the EDL track and 3 runs for the diagnostic EL track. The system obtains the best score of 0.661 in the end-to-end mention ceaf metric. It also obtains the second best score in the entity discovery and linking components in terms of the "strong typed mention match" and "strong all match" scores proving its robustness across different languages and genres.

1 System Description

1.1 Introduction

This year (2015), the EDL task has been extended to the Trilingual Entity Discovery and Linking (EDL) task. Systems need to extract mentions in documents from 3 languages: English , Spanish and Chinese and link the mentions extracted to the English version of a snapshot of Freebase. If the mentions refer to a NIL entity, then they need to get clustered with a unique identifier resolving them. There were several challenges involved in this new EDL task: mentions have to be extracted from multiple languages and genre, the mentions need to be linked to Freebase and then clustered across the languages with a single unique identifier if they are referring to the same canonical entity.

1.2 Mention Detection

The IBM mention detection system was a combination of two mention detection systems - one being a Neural Net-based (NN) system and one being a Conditional Random Fields (CRF) system, both trained to predict the standard IOB mention detection encoding (for English, the tag also has a bit specifying whether the mention is named or nominal). The Chinese model was a character-based model, while the English and Spanish models are more standard token-based models. All models were trained and applied using the IBM Statistical Information Relation and Extraction toolkit (SIRE).

The CRF model is a linear-chain CRF model of size 1 (the previous tag is used in features), using a multitude of features including words in context, capitalization flags, various entity dictionaries, both supervised (lists extracted from the ACE'05 data, the CoNLL'03 data, etc) and unsupervised (the system output on Gigaword), word clustering (Brown clusters), cache features, word length and IDF. In addition, the output of a KLUE model (an information extraction system with 50 mention types and relation types) was used as an additional input (for a minor improvement in performance).

All parameters of the model were estimated by 5-fold cross-validation on the training data. Additionally, the Spanish CRF model was trained on *both* the English and the Spanish data, effectively creating a bi-lingual system. This improved performance by 0.8F on the training data (from 0.836 to 0.844) - the reason for trying such a model was that Spanish and English share many lexical tokens and some of the features (such as word capitalization flags).

The NN system [1] uses a feed-forward neural net

[1]The NN system was used only for Spanish and English; the Chinese system was built as a combination of CRF models

to predict entity labels. The network architecture (Figure 1) is similar to that proposed in (Collobert et al., 2011) and uses as input the concatenation of the target and context words (symmetric window of size 4) to which we add vectors for three of the features used in the CRF model: dictionaries, capitalization flags and suffix/prefix features. For these additional features, when multiple values fire, their vectors are averaged (e.g. the capitalization vector for CEO is the mean of *allcap*, *initcap* and *3upper* vectors. The vectors for the suffix/prefix feature are initialized randomly (size 50) while the other features are initialized with one-hot representations. The word vectors are initialized with 300-dimensional pre-trained embeddings build on a concatenation of Gigaword, Bolt and Wikipedia, (totaling $\approx$ 6 billion tokens). Embeddings are built using a variant of the word2vec CBOW architecture, which predicts a target word from the concatenation of its context words, rather than the average. This variant outperforms CBOW both on standard word similarity benchmarks as well as in mention detection experiments. Both the additional feature vectors as well as word vectors are fine-tuned during training (i.e. error is back-propagated to the input representation).

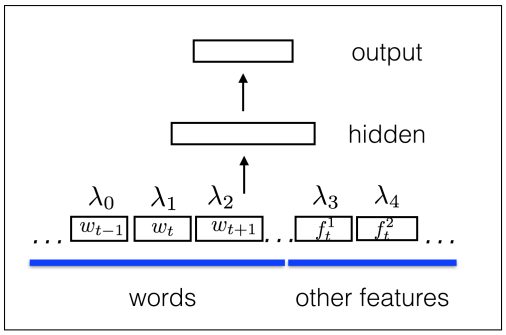

Figure 1: Architecture of the neural network used for mention detection

Additionally, we attach scalar weights to each of the features (λ_i), allowing the model to more easily learn the relative importance of each word/feature used in the input representations. (Learning for example that the target word has the highest weight and context word weights decay with increasing distance to the target.) We use one hidden layer of size 100 and sigmoid as its activation function. The cost function is the sentence-level log-likelihood described in (Collobert et al., 2011). In a standard word-level model the probability of the correct label is normalized w.r.t. the other labels using a softmax function. The sentence-level log-likelihood models the dependencies between different labels by introducing a transition matrix and maximizing the log-probability of the label path of an entire sentence, normalized w.r.t. all the possible label paths. The score of a path y_i^T is given by:

$$s(x_1^T, y_1^T) = \sum_{t=1}^{T} (A_{y_{t-1},y_t} + f_{y_t,t}) \qquad (1)$$

where A is a parameter label transition matrix, and $f_{y_t,t}$ is the neural network assigned score for tag y_t at time t. The score is normalized over al possible paths:

$$\log p(\hat{y}_1^T | x_1^T) = s(x_1^T, \hat{y}_1^T) - \underset{\forall y_1^T}{\text{logadd}}\, s(x_1^T, y_1^T) \qquad (2)$$

where $\text{logadd}_i z_i = \log(\sum_i e^{z_i})$. The two systems were combined in a simple scheme described below. We noticed that all models were slanted towards precision (meaning, precision was 5-6 points higher than recall), and we combined them as follows:

- The initial system output is the best performing system (NNs for English and CRF for Spanish)
- Considering the remaining systems in the order of performance, add any mentions that do not overlap with the combined system

The combination resulted in improvements of 0.5-1F on the cross-validated data.

For coreference, we used the SIRE system as such (the TAC types are a subset of the KLUE

	CRF	NN	Combination
English	0.715	0.718	0.727
Spanish	0.703	0.698	0.752

Table 1: Classifier combination results for English and Spanish There is a surprising gain in Spanish, which most likely comes from the fact that both systems have much higher precision than recall (0.863P/0.594R for CRF and 0.753P/0.651R for NN). The CRF/Spanish number here is obtained by training on both English and Spanish data as submitted; when trained on Spanish alone, the resulting CRF as a performance of 0.717, yet the combination yields the same F-measure, 0.752.

types, so this was directly possible). The KLUE model, however, identifies more entity types, and also pronouns, and the absence of those mentions in the system output negatively affected the coreference output. To account for this issue, we aligned the KLUE output with the TAC output and propagated the KLUE coreference to the TAC document, resulting in better entities - this was our main submission, while the KLUE coreference applied directly to the TAC output was our second one.

1.3 Entity Linking

The fundamental structure of the IBM EL system for 2015 is based on the 2014 system of (Sil and Florian, 2014) which obtained the top score in the official Spanish evaluation in 2014. The full document global entity disambiguation approach partitions the full set of mentions m of an input document d into smaller sets of mentions which appear near one another. We refer to these sets as the *connected components* of d, or $CC(d)$. We perform classification over the set of entity-mention tuples $E(cc)$ that are formed using candidate entities within the same connected component $cc \in CC(d)$. Consider this small snippet of text:

"... Home Depot CEO Nardelli quits ..."

In this example text, the phrase "Home Depot CEO Nardelli" would constitute a connected component, since the mentions "Home Depot" and "Nardelli" are separated by three or fewer tokens. Two of the entity-mention tuples for this connected component would be:

1. ([Home Depot], `Home_Depot`, [Nardelli], `Robert_Nardelli`)
2. ([Home Depot], `Home_Depot`, [Nardelli], `Steve_Nardelli`).

We use a maximum-entropy model to estimate $P(b|d, cc)$, the probability of an entity-mention tuple b for a given connected component $cc \in CC(d)$. Here $b_i = (m_1, e_1, \ldots, m_{n_i}, e_{n_i})$, where each e_j is taken from the Wikipedia dump of April 2014 (for English, Spanish and Chinese) for mention m_j detected by the mention detection component. The model involves a vector of real-valued feature functions $\mathbf{f}(b, d, cc)$ and a vector of real weights $\mathbf{w}$, one weight per feature function. The probability is given by

$$P(b|d, cc, \mathbf{w}) = \frac{\exp\left(\mathbf{w} \cdot \mathbf{f}(b, d, cc)\right)}{\sum_{b' \in B(cc)} \exp\left(\mathbf{w} \cdot \mathbf{f}(b', d, cc)\right)} \quad (3)$$

We use L2-regularized conditional log likelihood (CLL) as the objective function for training:

$$CLL(T, \mathbf{w}) = \sum_{(b,d,cc) \in T} \log P(b|d, cc, \mathbf{w}) + \sigma \|\mathbf{w}\|_2^2$$

where $(b, d, cc) \in T$ indicates that b is the correct tuple of entities and mentions for connected component cc in document d in training set T, and σ is a regularization parameter. LBFG-S can be used to solve this gradient-based convex optimization.

Some of the feature functions used in the IBM EDL system is as follows:

Local Features. The most basic versions of these features include: **COUNT-EXACT-MATCH**, which counts the number of mentions whose surface form matches exactly with one of the names for the linked entity stored in Wikipedia; **ALL-EXACT-MATCH**, which is true if all mentions in b match a Wikipedia title exactly; and **ACRONYM-MATCH**, if the mention's surface form is an acronym for a name of the linked entity in Wikipedia. The system also uses features based on redirect counts, cosine similarity of source and target texts, as well as counts of Wikipedia inlinks, outlinks etc. Besides computing the cosine similarity of texts mentioned in source and target documents, the system also computes **COSINE-SIM-LEMMA** which converts the text into its lemmatized format and then computes the cosine. The system also uses information from word embeddings and uses features based on cosine and nearest neighbors.

Global Features. Some of the global features include the **ENTITY-CATEGORY-PMI** and **ENTITY-CATEGORY-PRODUCT-PMI**. These make use of Wikipedia's category information system to find patterns of entities that commonly appear next to one another. Let $T(e)$ be the set of Wikipedia categories for entity e. We remove common Wikipedia categories which are associated with almost every entity in text, like `Living People` etc., since they have lower discriminating power. From the training data, the system first computes pointwise mutual information (PMI) (Turney, 2002) scores for the Wikipedia categories of consecutive pairs of

entities, (e_1, e_2):

$$PMI(T(e_1), T(e_2)) = \frac{\sum_{(e,e') \in T} \mathbf{1}[T(e_1) = T(e) \wedge T(e_2) = T(e')]}{\sum_{e \in T} \mathbf{1}[T(e_1) = T(e)] \times \sum_{e \in T} \mathbf{1}[T(e_2) = T(e)]}$$

where the sum in the numerator is taken over consecutive pairs of entities (e, e') in training data. The feature **ENTITY-CATEGORY-PMI** adds these scores up for every consecutive (e_1, e_2) pair in b. We also include another feature **ENTITY-CATEGORY-PRODUCT-PMI** which does the same, but uses an alternative product variant of the PMI score. Other features include categorical overlap of entities in the document and features similar to the Normalized Google Distance (NGD).

2 NIL Clustering and Entity Typing

The IBM Entity Linking system links the mentions extracted from the text to the Wikipedia dump of the respective language that the document is in: e.g. mentions in Chinese documents will be linked to the Chinese Wikipedia. In the next step, we attempt to link back these non-English links to the English Wikipedia title using Wikipedia's inter-language links and whatever does not match gets a NIL label. Finally, once all mentions either have a English Wikipedia title or a NIL label, we assign a TAC KB (Freebase) id using the "Freebase to Wikipedia" mapping.

Entity Typing which is predicting either PER, ORG, FAC or LOC is first done at the mention detection step. The IBM EDL system also uses feedback from the Wikipedia links produced: for every title in Wikipedia we train a maximum entropy classifier based on n-grams from the first paragraph of a Wikipedia title. The aim of the classifier is to attach a IBM KLUE entity type to every Wikipedia title. Finally, this classifier is run on every Wikipedia page (4.5 million titles) to generate a dictionary of Wikipedia titles to its entity type. For the TAC task, we only looked at the once where the classifier was at a confidence level of more than 90%. Hence, for every links produced by the EL system, we update the entity types by the MD system to the once predicted by this classifier if it exists in our dictionary. This strategy was used for the IBM run2 which brought a slight improvement (0.716) over IBM run1 (0.710).

Systems	MD	EL	End-to-End
Rank 1	0.724	0.661	**0.616**
Rank 2	**0.716**	**0.586**	0.616
Rank 3	0.647	0.539	0.551

Table 2: Comparing our system with the others in terms of F1 scores: MD indicates mention detection, EL indicates linking the detected mentions to the KB or NIL and finally, End-to-End indicates the final performance in terms clustering the linked entities together across languages and the KB. Bold numbers indicate the IBM system scores.

Since the TAC guidelines prohibit fictional entities we also train a rule-based binary classifier which looks at cosine-similarity based features trained from n-grams of fictional entities from Wikipedia. This classifier discards mentions like Bruce Wayne or Mickey Mouse (since these are fictional characters).

3 Experiments and some results

The IBM system for MD was trained on the TAC 2015 training data and the language independent EL system was trained on a sample of the 2011 English Wikipedia dump made publicly available by (Ratinov et al., 2011) and also on the CoNLL 2003 train component of the NER task made available by Hoffart et al (2011). EL model has been ported to the Spanish and Chinese EL task without the need for re-training. Most of the system development for the English data has been performed on the TAC 2009 data and particular attention was provided on the entities of the type PER. For the Spanish system development, we performed most experiments on the TAC 2013 evaluation data.

3.1 Results

Table 2 lists our official results for the TEDL task. Our system obtains the best score in terms of the end-to-end metric (mention CEAF). The system also comes second in terms of MD and EL. The MD component also obtains the best score in English and Spanish in terms of language as shown in Table 3.

Comparison in terms of the performance of our EL system compared with the full EDL system is shown in Figure 2. We observe that there are significant gains when brought to the EL system when full gold mentions are given. The EL score

Systems	MD	EL	End-to-End
English	0.727 (1)	0.631 (3)	0.661 (3)
Spanish	0.752 (1)	0.595 (2)	0.737 (2*)
Chinese	0.68 (1)	0.538 (3)	0.622 (3)

Table 3: The IBM system runs are shown above. Numbers in the braces show the rank obtained. The MD systems for English and Spanish obtained the top scores while the EL and end-to-end metrics display a robust performance. * indicates that the Spanish end-to-end metric is tied with rank 1.

	P	R	F1
Strong All Match	0.805	0.807	0.806
Mention Ceaf	0.829	0.83	0.829

Table 4: The results for the Trilingual Entity Linking evaluation is displayed above.

(strong all match) for the TEL task is 0.806 compared to 0.586, which obtained Rank 2 in the evaluation. Finally, the in terms of the end-to-end metric (mention ceaf) which considers both linking and clustering we observe a similar trend: an F1 score of 0.829 is obtained in comparison to 0.616 which is the best score in the TEDL evaluation in 2015.

4 Conclusion

We described the IBM E(D)L systems for English, Spanish and Chinese. Our language-independent EL strategy allowed us to train a single entity disambiguation system on one language and port it to another without the need for re-training. The system displays a robust performance across the genres and languages in the task and obtains the best end-to-end score.

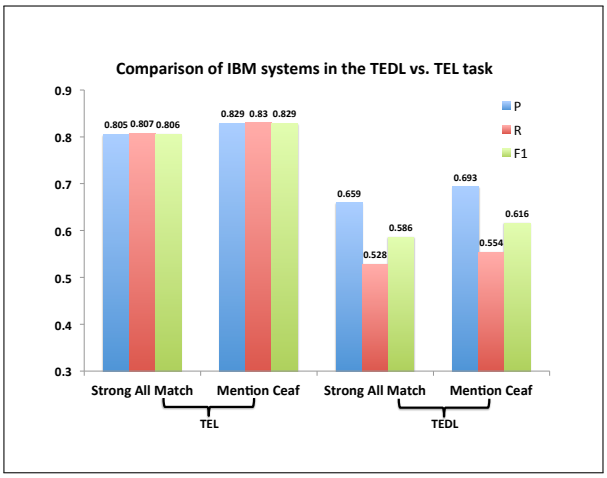

Figure 2: A comparison of our EL performance on the 2 tasks in 2015: Full trilingual EDL vs. EL. The bars show big improvements in terms of both mention CEAF and strong all match over the EDL scores which obtained the top score in the task.

References

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Furstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum1. 2011. Robust Disambiguation of Named Entities in Text. In *EMNLP*, pages 782–792.

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

Avirup Sil and Radu Florian. 2014. IThe IBM Systems for English Entity Discovery and Linking and Spanish Entity Linking at TAC 2014. In *TAC 2014*.

P. D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Procs. of ACL*, pages 417–424.