The IBM Systems for Trilingual Entity Discovery and Linking at TAC 2016

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 2016 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 described in (Sil and Florian, 2016). However, the system is different than TAC 2015 as it is trained using more training data from previous TAC evaluations and a significant portion of the Wikipedia. The same EL model was applied across all 3 languages: English, Spanish and Chinese. We submitted 3 runs for the first EDL evaluation window and 5 for the next one.

1 System Description

1.1 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.

We also used part of the test data from 2015 for training - we separated about 20 documents in each language for development, and added the rest to the training set. On this small training set, the performance improved about 5F for English by using this additional data.

The NN system uses a feed-forward neural net 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 two of the features used in the CRF model: dictionaries and capitalization flags. 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.) 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 1000 and sigmoid as its activation function. The cost function is the word-level log-likelihood described in (Collobert et al., 2011) in which the probability of the correct label is normalized w.r.t. the other labels using a softmax function.

We additionally incorporate character-level rep-

resentations by concatenating the output of a forward and a backward LSTM on the sequence of characters of the current token (Lample et al., 2016), where the character embeddings are randomly initialized for English and Spanish and pre-trained for Chinese. For English, we utilize an additional feature consisting of the label assigned by a mention detection model assigning one of 50 predefined labels and trained on additional data.

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).

For Chinese we also use positional character embeddings (Peng and Dredze, 2015) with each character being concatenated with its position in the word, leading from a vocabulary of 8K characters to 18K positional characters. The target word is represented as the mean of the positional character embeddings. For Chinese, both word and character embeddings are 300-dimensional and learned with word2vec on GigaWord.

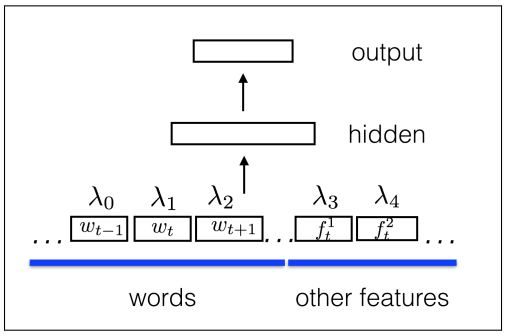

Figure 1: Architecture of the neural network used for mention detection

The two systems were combined in a simple scheme described below (same combination we have performed in 2015). 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 small development data, as can be seen in Table 1.

We have also built a coreference model directly from the training data. To be able to produce nominal mentions for the types for which such information was not provided (Spanish and Chinese data, and all types besides PER for English), we have done the following:

- Ran KLUE model described above on the test data
- Aligned the coreference chains from the TAC output and the KLUE output, by the maximum mention overlap
- Added any nominal mentions found in a chain of a proper type (i.e. PER, ORG, etc) to the corresponding TAC chain (if one existed); the mentions that belong to a chain that does not align are thrown out.

1.2 Entity Linking

The fundamental structure of the IBM EL system for 2016 is based on (Sil and Florian, 2016), a version of which was used in (Sil and Florian, 2014), which obtained the top score in the official Spanish evaluation in 2014 and top score in the diagnostic Tri-lingual evaluation in 2016. 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.

	NN	Best/NN	Vote/NN	CRF	Combo
English	74.0($\pm$0.4)	74.7	74.8	76.0	77.1
Spanish	75.2($\pm$0.9)	76.6	75.0	78.5	80.0
Chinese	73.4($\pm$0.6)	74.3	74.4	-	-

Table 1: NN - mean and standard deviation over 10 runs - Results on the small dev set of 20 documents for each language, from the evaluation data in 2015 (about 30K tokens for EN, 15k for Spa and Cmn)

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

$$P(b|d, cc, \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot \mathbf{f}(b, d, cc))}{\sum_{b' \in B(cc)} \exp(\mathbf{w} \cdot \mathbf{f}(b', d, cc))} \quad (1)$$

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).

	Cheng&Roth	LIEL	LIEL+ more Data
ACE	0.853	0.862	0.868
MSNBC	0.812	0.850	0.860

Table 2: Comparison of systems with more training data. Both LIEL and (Cheng and Roth, 2013) are trained on the training data provided by (Ratinov et al., 2011). Adding in more Wikipedia data actually helps the system as is evident in the last column.

2 More Training Data

The previous year's submissions included a system trained on the Wikipedia data provided by (Ratinov et al., 2011). However, this year the IBM EL system also included more freely available data from a Wikipedia dump of 2014. The effects of adding in more training data are shown in Table 2.

3 NIL Clustering and Discarding Fictional Entities

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 interlanguage 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.

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).

References

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. *Urbana*, 51:61801.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal, September. Association for Computational Linguistics.

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*.

Avirup Sil and Radu Florian. 2016. One for All: Towards Language Independent Named Entity Linking. In *Association of Computational Linguistics, ACL 2016. Berlin, Germany.*

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