RPI_BLENDER TAC-KBP2017 13 Languages EDL System

Boliang Zhang, Xiaoman Pan, Ying Lin, Tongtao Zhang, Kevin Blissett,
Samia Kazemi, Spencer Whitehead, Lifu Huang and Heng Ji
Computer Science Department, Rensselaer Polytechnic Institute
jih@rpi.edu

1 Introduction

This year RPI participated in both of the Trilingual Entity Discovery and Linking (EDL) task and 10 languages EDL task. We will present system development details in the following sections.

2 Trilingual EDL

2.1 Named Mention Extraction:

We consider name tagging as a sequence labeling problem, where each token in a sentence is tagged as the Beginning (B), Inside (I) or Outside (O) of a name mention with one of five types: Person (PER), Organization (ORG), Geopolitical Entity (GPE), Location (LOC) and Facility (FAC). Prediction of the tags requires evidence from context in the entire sentence and Bi-LSTM networks (Graves et al., 2013; Lample et al., 2016) meet such requirement by processing each sequence in both directions with two separate hidden layers, which are then fed into the same output layer. Moreover, classification dependencies constrain the tags in a sequence, *e.g.*, "I-LOC" should never follow "B-ORG". Therefore, we adopt CRFs model, which is particularly good at jointly modeling sequential tagging decisions, on tp of the Bi-LSTM networks. External information (*e.g.*, gazetteers, Brown Clustering, etc.) is proved to be beneficial for name tagging as well. Hence, we use an additional Bi-LSTM to consume the external feature embeddings of each token and concatenate both Bi-LSTM encodings of feature embeddings and word embeddings before the output layer. Figure 1 depicts the framework of our model.

We set the word input dimension to 100, word LSTM hidden layer dimension to 100, character input dimension to 50, character LSTM hidden layer dimension to 25, input dropout rate to 0.5, and use stochastic gradient descent with learning rate 0.01 for optimization.

Nominal and Pronominal Mention Extraction: we utilize a deep neural networks based entity coreference resolution system (Clark and Manning, 2016) in Stanford CoreNLP toolkit (Manning et al., 2014) to extract nominal and pronominal mentions.

2.2 Name Translation:

We translate Chinese mentions into English based on name translation dictionaries mined with various approaches described in (Ji et al., 2009; Pan et al., 2017). If a Chinese entity mention cannot be translated, we use Pinyin to transliterate it. In addition, we create a corpus which contains Chinese words and English entities from Chinese Wikipedia, by replacing Chinese anchor links with English entity IDs according to cross-lingual links. To this extent, we are able to learn distributed representations of multi-lingual words and English entities to match Chinese mentions and English candidate entities in the KB.

2.3 Entity Linking:

Given a set of entity mentions $M = \{m_1, m_2, ..., m_n\}$, we first generate an initial list of candidate entities $E_m = \{e_1, e_2, ..., e_n\}$ for each entity mention m, then we rank them and select the candidate entity with the highest score as the appropriate entity for linking.

We adopt a dictionary-based candidate generation approach (Medelyan and Legg, 2008). In oder to expand the coverage of the dictionary, we also generate a secondary dictionary by normalizing all keys in the primary dictionary using a phonetic algorithm NYSIIS (Taft, 1970). If an entity mention m is not in the primary dictionary, we will use the secondary dictionary to generate candidates.

Then we rank these entity candidates based on

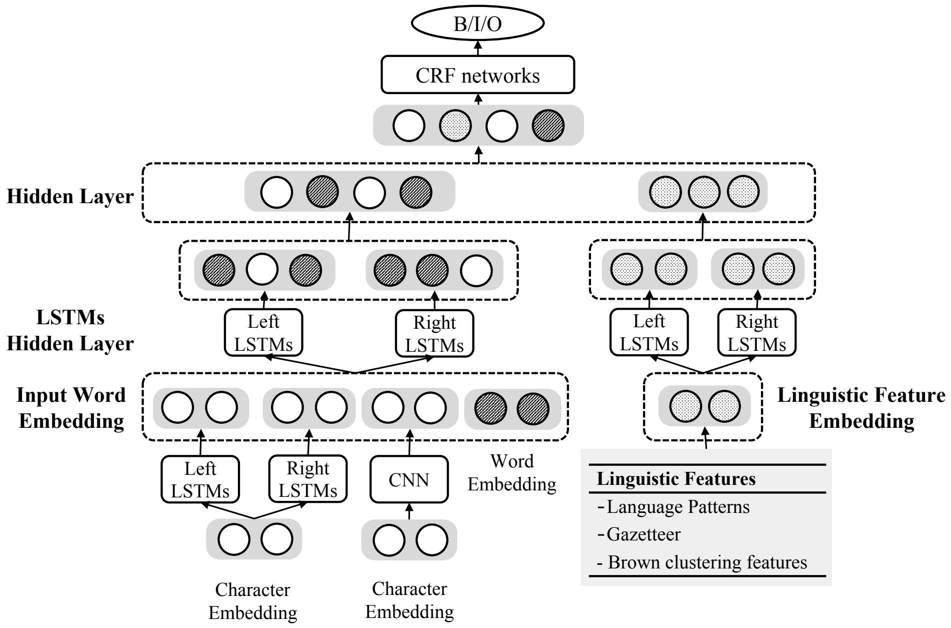

Figure 1: Name Tagging Model with Explicit Linguistic Features.

three measures: *salience*, *similarity* and *coherence* (Pan et al., 2015).

We utilize Wikipedia anchor links to compute *salience* based on entity prior:

$$p_{prior}(e) = \frac{A_{*,e}}{A_{*,*}} \quad (1)$$

where $A_{*,e}$ is a set of anchor links that point to entity e, and $A_{*,*}$ is a set of all anchor links in Wikipedia. We define mention-to-entity probability as

$$p_{mention}(e|m) = \frac{A_{m,e}}{A_{m,*}} \quad (2)$$

where $A_{m,*}$ is a set of anchor links with the same anchor text m, and $A_{m,e}$ is a subset of $A_{m,*}$ which points to entity e.

Then we compute the *similarities* between an individual mention and its entity candidates. We first utilize entity type of the mention which are extracted from name tagging. For each entity e in the KB, we assign a coarse-grained entity type t (PER, ORG, GPE, LOC, Miscellaneous (MISC)) using a Maximum Entropy based entity classifier (Pan et al., 2017). We incorporate entity type by combining it with mention-to-entity probability $p_{mention}(e|m)$ (Ling et al., 2015):

$$p_{type}(e|m,t) = \frac{p(e|m)}{\sum\limits_{e \mapsto t} p(e|m)} \quad (3)$$

where $e \mapsto t$ indicates that t is the entity type of e. We also adopt a neural network model that jointly learns distributed representations of words and entities from Wikipedia (Yamada et al., 2017; Cao et al., 2017). Considering all Wikipedia anchor links as entity annotations, a training corpus can be created by replacing anchor links with unique entity IDs. Such training corpus can be used to train the distributed representations of words and entities simultaneously. For each entity mention m, we build the vector representation of its context v_t using the vector representation of each word (exclude entity mention itself and stop words) in the context. Then we compute cosine similarity between the vector representation of each candidate entity v_e and v_t, which can be used to measure similarity between mention and entity $p_{sim}(m, e)$.

To compute *coherence*, we construct a weighted undirected graph $G = (E, D)$ from DBpedia, where E is a set of all entities in DBpedia and $d_{ij} \in D$ indicates that two entities e_i and e_j share some DBpedia properties as described in (Huang et al., 2017). The weight of d_{ij}, denoted as w_{ij}, is computed with

$$w_{ij} = \frac{|p_i \cap p_j|}{\max(|p_i|, |p_j|)} \quad (4)$$

where p_i, p_j are the sets of DBpedia properties of e_i and e_j respectively. After constructing the knowledge graph, we apply the graph embedding framework proposed by (Tang et al., 2015) to generate knowledge representations for all entities in the KB. We compute cosine similarity between

Rule	Description
Exact match	Create initial clusters based on mention surface form.
Normalization	Normalize surface forms (e.g., remove designators and stop words) and group mentions with the same normalized surface form.
NYSIIS (Taft, 1970)	Obtain soundex NYSIIS representation of each mention and group mentions with the same representation longer than 4 letters.
Edit distance	Cluster two mentions if the edit distance between their normalized surface forms is not greater than D, where $D = length(mention_1)/8 + 1$.
Translation	Merge two clusters if they include mentions with the same translation.

Table 1: Heuristic Rules for NIL Clustering.

the vector representations of two entities to model coherence between these two entities $coh(e_i, e_j)$. Given a entity mention m and its candidate entity e, we defined coherence score as:

$$p_{coh}(e) = \frac{1}{|C_m|} \sum_{c \in C_m} coh(e, c) \qquad (5)$$

where C_m is the union of entities for coherent mentions of m.

Finally, we combine these measures and compute final score for each candidate entity e.

2.4 NIL Clustering:

For entity mentions that cannot be linked to the KB, we apply heuristic rules described in Table 1 to cluster these NIL entity mentions. For each cluster, we assign the most frequent entity mention as the document-level canonical mention.

2.5 10 Languages EDL

RPI also organized and participated in the EDL pilot evaluation for ten languages: Polish, Chechen, Albanian, Swahili, Kannada, Yoruba, Northern Sotho, Nepali, Kikuyu and Somali. The underlying system framework is the same as RPI's English and Chinese EDL system. The major challenge lies in the lack of data annotation and resources for most of these languages. We attempted various creative ways to create silver-standard training data. We derived some entity annotations from Wikipedia markups as described in (Pan et al., 2017). In addition, we developed a "Chinese Room" EDL interface where a foreign language document is displayed, and some words and candidate names are translated based on lexicons and gazetteers. A user can also collect and provide their knowledge about a language in the interface, such as name designators. The romanized transliteration of non-latin script languages is also displayed. This interface enables a user to identify, classify and translate names in each sentence and it also allows a user to delete a sentence with low annotation confidence. We, system developers who are non-native speakers of these languages, use this interface to generate noisy name annotations.

Moreover, we developed a common semantic space to allow multiple languages to share distributed representations. In this work, we extend 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. We design a multi-level, multi-encoder, multi-decoder framework. For each language, we adopt a character-aware neural language model to learn word embeddings. We apply a Convolutional Neural Network (CNN) over the sequence of characters of each word, and a max-over-time pooling function to obtain word representations. Then we further optimize all word representations by a multi-layer Long Short-term Memory (LSTM) and a softmax function, minimizing the loss between the predicted distribution over next word and the actual next word. Then we project mono-media mono-lingual semantic representations for each word into a common semantic space based on multi-level alignment: (1) Word Alignment: based on multi-lingual dictionaries including Wikitionary, Panlex and Open Multilingual WordNet. (2) Structure Alignment: based on multi-lingual structural knowledge resources such as World Atlas of Linguistic Structure (WALS). The goal of this multi-lingual multi-level autoencoder is to automatically learn a reduced dimensional vector representation for each mono-lingual embedding input and reconstruct the input from a new vector from the shared common space, and minimize the reconstruction error. We minimize the following three loss functions: (1) Mono-lingual reconstruction errors: project from mono-lingual embedding space to common semantic space, then reconstruct this mono-lingual embedding space; (2) Cross-lingual reconstruction errors: project from common semantic space to mono-lingual embedding space, using aligned words and knowledge elements from other lan-

Languages	F1 (%)	# of Docs	# of Words	Data Source
Albanian	75.81	40	34,827	Silver+
Chechen	66.01	113	110,518	Gold
Kannada	63.47	40	6,717	Silver+
Kikuyu	91.80	104	1,249	Silver
Nepali	67.97	40	14,067	Silver+
Northern Sotho	93.20	71	1,435	Silver
Polish	55.19	40	16,583	Silver+
Somali	83.13	605	72,114	Gold
Swahili	77.39	40	35,205	Silver+
Yoruba	66.76	197	36.291	Gold

Table 2: 10 Languages Name Tagging Resources and Performance (Silver: Wikipedia derived annotation; Silver+: Chinese Room; Gold: LDC released annotation)

guages; and (3) Cross-lingual alignment errors: minimize the distance of the aligned word representations in the common semantic space. This common semantic space significantly improved the name tagging performance for languages like Chechen by borrowing resources and knowledge from Russian.

Table 2 summarizes the resources used and name tagging performance for each language.

References

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. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*. https://nlp.stanford.edu/pubs/clark2016deep.pdf.

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*.

Lifu Huang, Jonathan May, Xiaoman Pan, Heng Ji, Xiang Ren, Jiawei Han, Lin Zhao, and James A. Hendler. 2017. Liberal entity extraction: Rapid construction of fine-grained entity typing systems. *Big Data* 5. https://doi.org/10.1089/big.2017.0012.

Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens, and Hermann Ney. 2009. Name extraction and translation for distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation* .

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*.

Xiao Ling, Sameer Singh, and Daniel Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics* 3.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*.

O. Medelyan and C. Legg. 2008. Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense. In *Proc. AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*.

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*.

Robert L Taft. 1970. *Name Search Techniques*. New York State Identification and Intelligence System, Albany, New York, US.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*. International World Wide Web Conferences Steering Committee, pages 1067–1077.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *CoRR* abs/1705.02494.