Neural Cross-Lingual Coreference Resolution And Its Application To Entity Linking

Gourab Kundu and **Avirup Sil** and **Radu Florian** and **Wael Hamza**
IBM Research
1101 Kitchawan Road
Yorktown Heights, NY 10598
{gkundu, avi, raduf, whamza}@us.ibm.com

Abstract

We propose an entity-centric neural cross-lingual coreference model that builds on multi-lingual embeddings and language-independent features. We perform both intrinsic and extrinsic evaluations of our model. In the intrinsic evaluation, we show that our model, when trained on English and tested on Chinese and Spanish, achieves competitive results to the models trained directly on Chinese and Spanish respectively. In the extrinsic evaluation, we show that our English model helps achieve superior entity linking accuracy on Chinese and Spanish test sets than the top 2015 TAC system without using any annotated data from Chinese or Spanish.

1 Introduction

Cross-lingual models for NLP tasks are important since they can be used on data from a new language without requiring annotation from the new language (Ji et al., 2014, 2015). This paper investigates the use of multi-lingual embeddings (Faruqui and Dyer, 2014; Upadhyay et al., 2016) for building cross-lingual models for the task of coreference resolution (Ng and Cardie, 2002; Pradhan et al., 2012). Consider the following text from a Spanish news article:

"Tormenta de nieve afecta a 100 millones de personas en EEUU. Unos 100 millones de personas enfrentaban el sábado nuevas dificultades tras la enorme tormenta de nieve de hace días en la costa este de Estados Unidos."

The mentions "EEUU" ("US" in English) and "Estados Unidos" ("United States" in English) are coreferent. A coreference model trained on English data is unlikely to coreference these two mentions in Spanish since these mentions did not appear in English data and a regular English style abbreviation of "Estados Unidos" will be "EU" instead of "EEUU". But in the bilingual English-Spanish word embedding space, the word embedding of "EEUU" sits close to the word embedding of "US" and the sum of word embeddings of "Estados Unidos" sit close to the sum of word embeddings of "United States". Therefore, a coreference model trained using English-Spanish bilingual word embeddings on English data has the potential to make the correct coreference decision between "EEUU" and "Estados Unidos" without ever encountering these mentions in training data.

The contributions of this paper are two-fold. Firstly, we propose an entity-centric neural cross-lingual coreference model. This model, when trained on English and tested on Chinese and Spanish from the TAC 2015 Trilingual Entity Discovery and Linking (EDL) Task (Ji et al., 2015), achieves competitive results to models trained directly on Chinese and Spanish respectively. Secondly, a pipeline consisting of this coreference model and an Entity Linking (henceforth EL) model can achieve superior linking accuracy than the official top ranking system in 2015 on Chinese and Spanish test sets, without using any supervision in Chinese or Spanish.

Although most of the active coreference research is on solving the problem of noun phrase coreference resolution in the Ontonotes data set, invigorated by the 2011 and 2012 CoNLL shared task (Pradhan et al., 2011, 2012), there are many important applications/end tasks where the mentions of interest are not noun phrases. Consider the sentence,

"(U.S. president Barack Obama who started ((his) political career) in (Illinois)), was born in (Hawaii)."

The bracketing represents the Ontonotes style

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 395–400
Melbourne, Australia, July 15 - 20, 2018.

noun phrases and underlines represent the phrases that should be linked to Wikipedia by an EL system. Note that mentions like "U.S." and "Barack Obama" do not align with any noun phrase. Therefore, in this work, we focus on coreference on mentions that arise in our end task of entity linking and conduct experiments on TAC TriLingual 2015 data sets consisting of English, Chinese and Spanish.

2 Coreference Model

Each mention has a *mention* type (m_type) of either name or nominal and an *entity* type (e_type) of Person (PER) / Location (LOC) / GPE / facility (FAC) / organization (ORG) (following standard TAC (Ji et al., 2015) notations).

The objective of our model is to compute a function that can decide whether two partially constructed entities should be coreferenced or not. We gradually merge the mentions in the given document to form entities. Mentions are considered in the order of names and then nominals and within each group, mentions are arranged in the order they appear in the document. Suppose, the sorted order of mentions are $m_1, \ldots, m_{N_1}, m_{N_1+1}, \ldots, m_{N_1+N_2}$ where N_1 and N_2 are respectively the number of the named and nominal mentions. A singleton entity is created from each mention. Let the order of entities be $e_1, \ldots, e_{N_1}, e_{N_1+1}, \ldots, e_{N_1+N_2}$.

We merge the named entities with other named entities, then nominal entities with named entities in the same sentence and finally we merge nominal entities across sentences as follows:

Step 1: For each *named* entity e_i ($1 \leq i \leq N_1$), antecedents are all entities e_j ($1 \leq j \leq i-1$) such that e_j and e_i have same e_type. Training examples are triplets of the form (e_i, e_j, y_{ij}). If e_i and e_j are coreferent (meaning, y_{ij}=1), they are merged.

Step 2: For each nominal entity e_i ($N_1 + 1 \leq i \leq N_1 + N_2$), we consider antecedents e_j such that e_i and e_j have the same e_type and e_j has some mention that appears in the *same sentence* as some mention in e_i. Training examples are generated and entities are merged as in the previous step.

Step 3: This is similar to previous step, except e_i and e_j have *no* sentence restriction.

Features: For each training triplet (e_1, e_2, y_{12}), the network takes the entity pair (e_1, e_2) as input and tries to predict y_{12} as output. Since each entity represents a set of mentions, the entity-pair embedding is obtained from the embeddings of mention pairs generated from the cross product of the entity pair. Let $M(e_1, e_2)$ be the set $\{(m_i,m_j) \mid (m_i,m_j) \in e_1 \times e_2\}$. For each $(m_i, m_j) \in M(e_1, e_2)$, a feature vector ϕ_{m_i,m_j} is computed. Then, every feature in ϕ_{m_i,m_j} is embedded as a vector in the real space. Let v_{m_i,m_j} dentote the concatenation of embeddings of all features in ϕ_{m_i,m_j}. Embeddings of all features except the words are learned in the training process. Word embeddings are pre-trained. v_{m_i,m_j} includes the following language independent features:

String match: whether m_i is a substring or exact match of m_j and vice versa (e.g. m_i = "Barack Obama" and m_j = "Obama")

Distance: word distance and sentence distance between m_i and m_j discretized into bins

m_type: concatenation of m_types for m_i and m_j

e_type: concatenation of e_types for m_i and m_j

Acronym: whether m_i is an acronym of m_j or vice versa (e.g. m_i = "United States" and m_j = "US")

First name mismatch: whether m_i and m_j belong to e_type of PERSON with the same last name but different first name (e.g. m_i="Barack Obama" and m_j = "Michelle Obama")

Speaker detection: whether m_i and m_j both occur in the context of words indicating speech e.g. "say", "said"

In addition, v_{m_i,m_j} includes the average of the word embeddings of m_i and average of the word embeddings of m_j.

2.1 Network Architecture

The network architecture from the input to the output is shown in figure 1.

Embedding Layer: For each training triplet (e_1, e_2, y), a sequence of vectors v_{m_i,m_j} (for each $((m_i, m_j) \in M(e_1, e_2))$) is given as input to the network.

Relu Layer: $v^r_{m_i,m_j} = max(0, W^{(1)} v_{m_i,m_j})$

Attention Layer: To generate the entity-pair embedding, we need to combine the embeddings of mention pairs generated from the entity-pair. Consider two entities e_1 = (President[1], Obama)} and e_2 = {(President[2], Clinton)}. Here the superscripts are used to indicate two different mentions with the same surface form. Since the named mention pair (Obama, Clinton) has no string overlap, e_1 and e_2 should not be coreferenced even though the

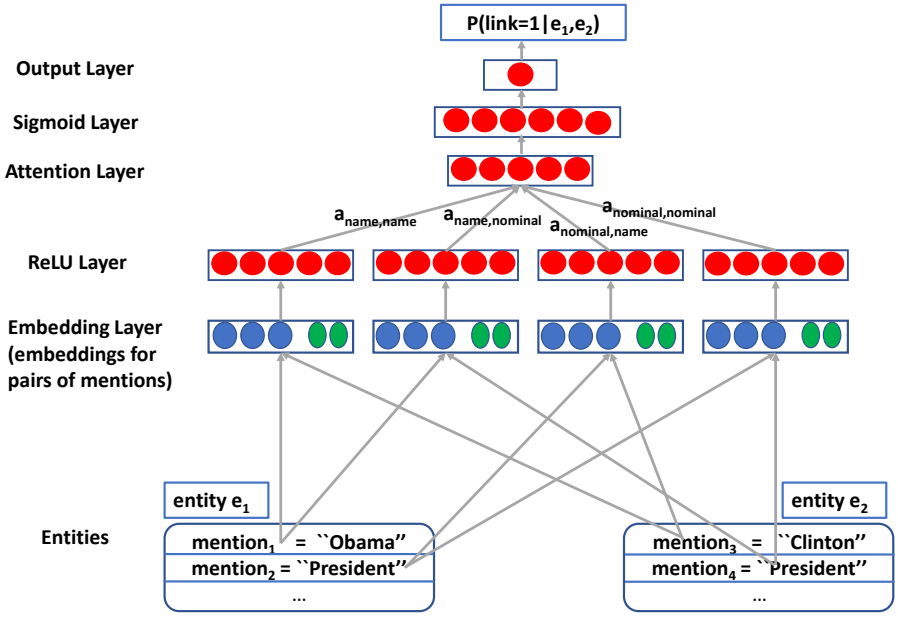

Figure 1: Network architecture for our coreference system. Blue circles in mention-pair embeddings layer represent embeddings of features. Green circles represent word embeddings.

nominal mention pair (President[1], President[2]) has full string overlap. So, while combining the embeddings for the mention pairs, mention pairs with m_type (name, name) should get higher weight than mention pairs with m_type (nominal, nominal). The entity pair embedding is the weighted sum of the mention-pair embeddings. We introduce 4 parameters $a_{name,name}$, $a_{name,nominal}$, $a_{nominal,nominal}$ and $a_{nominal,name}$ as weights for mention pair embeddings with m_types of (name, name), (name, nominal), (nominal, nominal) and (nominal, name) respectively. The entity pair embedding is computed as follows:

$$v^a_{e_1,e_2} = \sum_{(m_i,m_j)\in M(e_1,e_2)} \frac{a_{m\_type(m_i),m\_type(m_j)}}{N} v^r_{m_i,m_j}$$

Here N is a normalizing constant given by:

$$N = \sqrt{\sum_{(m_i,m_j)\in M(e_1,e_2)} a^2_{m\_type(m_i),m\_type(m_j)}}$$

This layer represents attention over the mention pair embeddings where attention weights are based on the m_types of the mention pairs.

Sigmoid Layer: $v^s_{e_1,e_2} = \sigma(W^{(2)} v^a_{e_1,e_2})$

Output Layer:

$$P(y_{12} = 1|e_1, e_2) = \frac{1}{1 + e^{-w^s . v^s_{e_1,e_2}}}$$

The training objective is to maximize L.

$$L = \prod_{d\in D} \prod_{(e_1,e_2,y_{12})\in S_d} P(y_{12}|e_1, e_2; W^{(1)}, W^{(2)}, a, w^s) \quad (1)$$

Here D is the corpus and S_d is the training triplets generated from document d.

Decoding proceeds similarly to training algorithm, except at each of the three steps, for each entity e_i, the highest scoring antecedent e_j is selected and if the score is above a threshold, e_i and e_j are merged.

3 A Zero-shot Entity Linking model

We use our recently proposed cross-lingual EL model, described in (Sil et al., 2018), where our target is to perform "zero shot learning" (Socher et al., 2013; Palatucci et al., 2009). We train an EL model on English and use it to decode on any other language, provided that we have access to multi-lingual embeddings from English and the target language. We briefly describe our techniques here and direct the interested readers to the paper. The EL model computes several similarity/coherence *scores* S in a "feature abstraction layer" which computes several measures of similarity between the context of the mention m in the query document and the context of the candidate link's Wikipedia page which are fed to a

feed-forward neural layer which acts as a binary classifier to predict the correct link for m. Specifically, the feature abstraction layer computes cosine similarities (Sil and Florian, 2016) between the representations of the source query document and the target Wikipedia pages over various granularities. These representations are computed by performing CNNs and LSTMs over the context of the entities. Then these similarities are fed into a Multi-perspective Binning layer which maps each similarity into a higher dimensional vector. We also train fine-grained similarities and dissimilarities between the query and candidate document from multiple perspectives, combined with convolution and tensor networks.

The model achieves state-of-the-art (SOTA) results on English benchmark EL datasets and also performs surprisingly well on Spanish and Chinese. However, although the EL model is "zero-shot", the within-document coreference resolution in the system is a language-dependent SOTA coreference system that has won multiple TAC-KBP (Ji et al., 2015; Sil et al., 2015) evaluations but is trained on the target language. Hence, our aim is to apply our proposed coreference model to the EL system to perform an extrinsic evaluation of our proposed algorithm.

4 Experiments

We evaluate cross-lingual transfer of coreference models on the TAC 2015 Tri-Lingual EL datasets. It contains mentions annotated with their grounded Freebase [1] links (if such links exist) or corpus-wide clustering information for 3 languages: English (henceforth, En), Chinese (henceforth, Zh) and Spanish (henceforth, Es). Table 1 shows the size of the training and test sets for the three languages. The documents come from two genres of newswire and discussion forums. The mentions in this dataset are either named entities or nominals that belong to five types: PER, ORG, GPE, LOC and FAC.

Hyperparameters: Every feature is embedded in a 50 dimensional space except the words which reside in a 300 dimensional space. The Relu and Sigmoid layers have 100 and 500 neurons respectively. We use SGD for optimization with an initial learning rate of 0.05 which is linearly reduced to 0.0001. Our mini batch size is 32 and we train for 50 epochs and keep the best model based on dev set.

[1]TAC uses BaseKB, which is a snapshot of Freebase. SIL18 links entities to Wikipedia and in-turn links them to BaseKB.

	En	Es	Zh
Train	168	129	147
Test	167	167	166

Table 1: No of documents for the TAC 2015 Tri-Lingual EL Dataset

	MUC	B^3	CEAF	CoNLL
This work	87.8	86.8	80.9	**85.2**
C&M16	83.6	78.7	69.2	77.2

Table 2: Coreference results on the En test set of TAC 15 competition. Our model significantly outperforms C&M16.

Coreference Results: For each language, we follow the official train-test splits made in the TAC 2015 competition. Except, a small portion of the training set is held out as development set for tuning the models. All experimental results on all languages reported in this paper were obtained on the official test sets. We used the official CoNLL 2012 evaluation script and report MUC, B^3 and CEAF scores and their average (CONLL score). See Pradhan et al. (2011, 2012).

To test the competitiveness of our model with other SOTA models, we train the publicly available system of Clark and Manning (2016) (henceforth, C&M16) on the TAC 15 En training set and test on the TAC 15 En test set. The C&M16 system normally outputs both noun phrase mentions and their coreference and is trained on Ontonotes. To ensure a fair comparison, we changed the configuration of the system to accept gold mention boundaries both during training and testing. Since the system was unable to deal with partially overlapping mentions, we excluded such mentions in the evaluation. Table 2 shows that our model outperforms C&M16 by 8 points.

For cross-lingual experiments, we build monolingual embeddings for En, Zh and Es using the widely used CBOW word2vec model (Mikolov et al., 2013a). Recently Canonical Correlation Analysis (CCA) (Faruqui and Dyer, 2014), Multi-CCA (Ammar et al., 2016) and Weighted Regression (Mikolov et al., 2013b) have been proposed for building the multi-lingual embedding space from monolingual embedding. In our prelimi-

	MUC	B^3	CEAF	CoNLL
Es Test Set				
En model	89.5	91.2	87.2	89.3
Es Model	90	91.4	88	**89.8**
Zh Test Set				
En model	95.5	93.3	88.7	92.5
Zh Model	96	92.8	89.6	**92.8**

Table 3: Coreference results on the Es and Zh test sets of TAC 15. En model performs competitively to the models trained on target language data.

Systems	Train on Target Lang	Acc. on Es	Acc. on Zh
EL - Coref	No	78.1	81.3
EL + En Coref	**No**	**81.1**	**83.9**
TAC Rank 1	Yes	80.4	83.1
SIL18	Yes	82.3	84.4

Table 4: Performance comparison on the TAC 2015 Es and Zh datasets. EL + En Coref outperforms the best 2015 TAC system (Rank 1) without requiring any Es or Zh coreference data.

nary experiments, the technique of Mikolov et al. (2013b) performed the best and so we used it to project the embeddings of Zh and Es onto En.

In Table 3, "En Model" refers to the model that was trained on the En training set of TAC 15 using *multi-lingual* embeddings and tested on the Es and Zh testing set of TAC 15. "Es Model" refers to the model trained on Es training set of TAC 15 using Es embeddings. "Zh Model" refers to the model trained on the Zh training set of TAC 15 using Zh embeddings. The En model performs 0.5 point below the Es model on the Es test set. On the Zh test set, the En model performs only 0.3 point below the Zh model. Hence, we show that without using any target language training data, the En model with multi-lingual embeddings gives comparable results to models trained on the target language.

EL Results: We replace the in-document coreference system (trained on the target language) of SIL18 with our En model to investigate the performance of our proposed algorithm on an extrinsic task. Table 4 shows the EL results on Es and Zh test sets respectively. "EL - Coref" refers to the case where the first step of coreference is not used and EL is used to link the mentions directly to Freebase. "EL + En Coref" refers to the case where the neural english coreference model is first used on Zh or Es data followed by the EL model. The former is 3 points below the latter on Es and 2.6 points below Zh, implying coreference is a vital task for EL. Our "EL + En Coref" outperforms the 2015 TAC best system by 0.7 points on Es and 0.8 points on Zh, without requiring any training data for coreference on Es and Zh respectively. Finally, we show the SOTA results on these two data sets recently reported by SIL18. Although their EL model does not use any supervision from Es or Zh, their coreference resolution model is trained on a large internal data set on the same language as the test set .Without using any in-language training data, our results are competitive to their results (1.2% below on Es and 0.5% below on Zh).

5 Related Work

Rule based (Raghunathan et al., 2010) and statistical coreference models (Bengtson and Roth, 2008; Rahman and Ng, 2009; Fernandes et al., 2012; Durrett et al., 2013; Clark and Manning, 2015; Martschat and Strube, 2015; Björkelund and Kuhn, 2014) are hard to transfer across languages due to their use of lexical features or patterns in the rules. Neural coreference is promising since it allows cross-lingual transfer using multilingual embedding. However, most of the recent neural coreference models (Wiseman et al., 2015, 2016; Clark and Manning, 2015, 2016; Lee et al., 2017) have focused on training and testing on the same language. In contrast, our model performs cross-lingual coreference. There have been some recent promising results regarding such cross-lingual models for other tasks, most notably mention detection(Ni et al., 2017) and EL (Tsai and Roth, 2016; Sil and Florian, 2016). In this work, we show that such promise exists for coreference also.

The tasks of EL and coreference are intrinsically related, prompting joint models (Durrett and Klein, 2014; Hajishirzi et al., 2013). However, the recent SOTA was obtained using pipeline models of coreference and EL (Sil et al., 2018). Compared to a joint model, pipeline models are easier to implement, improve and adapt to a new domain.

6 Conclusion

The proposed cross-lingual coreference model was found to be empirically strong in both intrinsic and extrinsic evaluations in the context of an entity linking task.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *ACL*.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *ACL*.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *ACL*.

Greg Durrett, David Leo Wright Hall, and Dan Klein. 2013. Decentralized entity-level modeling for coreference resolution. In *ACL*.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*.

Eraldo Rezende Fernandes, Cícero Nogueira Dos Santos, and Ruy Luiz Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *EMNLP-CoNLL*.

Hannaneh Hajishirzi, Leila Zilles, Daniel S Weld, and Luke Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *EMNLP*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *TAC*.

Heng Ji, Joel Nothman, Ben Hachey, et al. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *TAC*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *ACL*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *ACL*.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *NIPS*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *EMNLP-CoNLL*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *CoNLL*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *EMNLP*.

Avirup Sil, Georgiana Dinu, and Radu Florian. 2015. The ibm systems for trilingual entity discovery and linking at tac 2015. In *TAC*.

Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. In *ACL*.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *AAAI*.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *HLT-NAACL*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *ACL*.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning global features for coreference resolution. In *NAACL*.

Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *ACL*.