Joint Learning of Local and Global Features for Entity Linking via Neural Networks

Thien Huu Nguyen[†], **Nicolas Fauceglia**[#], **Mariano Rodriguez Muro**[§],
Oktie Hassanzadeh[§], **Alfio Massimiliano Gliozzo**[§] and **Mohammad Sadoghi**[‡]
[†] New York University, [#] Carnegie Mellon University, [‡] Purdue University
[§] IBM T.J. Watson Research Center, Yorktown Heights, New York, USA
thien@cs.nyu.edu,fauceglia@cs.cmu.edu
{mrodrig,hassanzadeh,gliozzo}.us.ibm.com,msadoghi@purdue.edu

Abstract

Previous studies have highlighted the necessity for entity linking systems to capture the local entity-mention similarities and the global topical coherence. We introduce a novel framework based on convolutional neural networks and recurrent neural networks to simultaneously model the local and global features for entity linking. The proposed model benefits from the capacity of convolutional neural networks to induce the underlying representations for local contexts and the advantage of recurrent neural networks to adaptively compress variable length sequences of predictions for global constraints. Our evaluation on multiple datasets demonstrates the effectiveness of the model and yields the state-of-the-art performance on such datasets. In addition, we examine the entity linking systems on the domain adaptation setting that further demonstrates the cross-domain robustness of the proposed model.

1 Introduction

We address the problem of entity linking (EL): mapping entity mentions in documents to their correct entries (called target entities) in some existing knowledge bases (KB) like Wikipedia. For instance, in the sentence "*Liverpool suffered an upset first home league defeat of the season.*", an entity linking system should be able to identify the entity mention "*Liverpool*" as a football club rather than a city in England in the knowledge bases. This is a challenging problem of natural language processing, as the same entity might be presented in various names, and the same entity mention string might refer to different entities in different contexts. Entity linking is a fundamental task for other applications such as information extraction, knowledge base construction etc.

In order to tackle the ambiguity in EL, previous studies have first generated a set of target entities in the knowledge bases as the referent candidates for each entity mention in the documents, and then solved a ranking problem to disambiguate the entity mention. The key challenge in this paradigm is the ranking model that computes the relevance of each target entity candidate to the corresponding entity mention using the available context information in both the documents and the knowledge bases.

The early approach for the ranking problem in EL has resolved the entity mentions in documents *independently* (*the local approach*), utilizing various *discrete* and *hand-designed* features/heuristics to measure the local mention-to-entity relatedness for ranking. These features are often specific to each entity mention and candidate entity, covering a wide range of linguistic and/or structured representations such as lexical and part-of-speech tags of context words, dependency paths, topical features, KB infoboxes (Bunescu and Pasca, 2006; Mendes et al., 2011; Cassidy et al., 2011; Ji and Grishman, 2011; Shen et al., 2014) etc. Although the local approach can exploit a rich set of discrete structures for EL, its limitation is twofold:

(i) The independent ranking mechanism in the local approach overlooks the topical coherence among the target entities referred by the entity mentions within the same document. This is undesirable as the topical coherence has been shown to be effective for EL in the previous work (Han et al.,

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2310–2320, Osaka, Japan, December 11-17 2016.

2011; Hoffart et al., 2011; Ratinov et al., 2011; He et al., 2013b; Alhelbawy and Gaizauskas, 2014; Pershina et al., 2015).

(ii) The local approach might suffer from the data sparseness issue of unseen words/features, the difficulty of calibrating, and the failure to induce the underlying similarity structures at high levels of abstraction for EL (due to the extensive reliance on the hand-designed coarse features) (Sun et al., 2015; Francis-Landau et al., 2016).

The first drawback of the local approach has been overcome by *the global models* in which all entity mentions (or a group of entity mentions) within a document are disambiguated *simultaneously* to obtain a *coherent* set of target entities. The central idea is that the referent entities of some mentions in a document might in turn introduce useful information to link other mentions in that document due to the semantic relatedness among them. For example, the appearances of "*Manchester*" and "*Chelsea*" as the football clubs in a document would make it more likely that the entity mention "*Liverpool*" in the same document is also a football club. Unfortunately, the coherence assumption of the global approach does not hold in some situations, necessitating the discrete/coarse features in the local approach as a mechanism to compensate for the potential exceptions of the coherence assumption (Ratinov et al., 2011; Hoffart et al., 2011; Sil et al., 2012; Durrett and Klein, 2014; Pershina et al., 2015). Consequently, the global approach is still subject to the second limitation of data sparseness of the local approach due to their use of discrete features.

Recently, the surge of neural network (NN) models has presented an effective mechanism to mitigate the second limitation of the local approach. In such models, words are represented by *continuous* representations (Bengio et al., 2003; Turian et al., 2010; Mikolov et al., 2013) and features for the entity mentions and candidate entities are automatically learnt from data. This essentially alleviates the data sparseness problem of unseen words/features and helps to extract more effective features for EL in a given dataset (Kalchbrenner et al., 2014; Nguyen et al., 2016a). In practice, the features automatically induced by NN are combined with the discrete features in the local approach to extend their coverage for EL (Sun et al., 2015; Francis-Landau et al., 2016). However, as the previous NN models for EL are local, they cannot capture the global interdependence among the target entities in the same document (the first limitation of the local approach).

Guided by these analyses, in this paper, we propose to use neural networks to model both the local mention-to-entity similarities and the global relatedness among target entities in an unified architecture. This allows us to inherit all the benefits from the previous systems as well as overcome their inherent issues. Our work is an extension of (Francis-Landau et al., 2016) which only considers the local similarities.

Given a document, we simultaneously perform linking for every entity mention from the beginning to the end of the document. For each entity mention, we utilize convolutional neural networks (CNN) to obtain the distributed representations for the entity mention as well as its target candidates. These distributed representations are then used for two purposes: (i) computing the local similarities for the entity mention and target candidates, and (ii) functioning as the input for the recurrent neural networks (RNN) that runs over the entity mentions in the documents. The role of the RNNs is to accumulate information about the previous entity mentions and target entities, and provide them as the global constraints for the linking process of the current entity mention. We systematically evaluate the proposed model on multiple datasets in both the general setting and the domain adaptation setting. The experiment results show that the proposed model outperforms the current state-of-the-art models on the evaluated datasets. To our knowledge, this is also the first work investigating the EL problem in the domain adaptation setting.

2 Model

The entity linking problem in this work can be formalized as follows. Let D be the input document and $M = \{m_1, m_2, \ldots, m_k\}$ be the entity mentions in D. The goal is to map each entity mention m_i to its corresponding Wikipedia page (entity) or return "*NIL*" if m_i is not present in Wikipedia. For each entity mention $m_i \in D$, let $P_i = \{p_{i1}, p_{i2}, \dots p_{in_i}\}$ be its set of Wikipedia candidate pages (entities)[1] where n_i is the number of page candidates for m_i. Also, let $p_i^* \in P_i$ be the correct target entity for m_i.

Following Francis-Landau et al. (2016), we represent each entity mention m_i by the triple $m_i = (s_i, c_i, d_i)$, where s_i is the *surface string* of m_i, c_i is the *immediate context* (within some predefined window) of m_i and d_i is the *entire document* containing m_i. Essentially, s_i, c_i and d_i are the sequences of words to capture the contexts or topics of m_i at multiple granularities. For the target candidate pages p_{ij}, we use the *title* t_{ij} and *body content* b_{ij} to represent them ($p_{ij} = (t_{ij}, b_{ij})$). For convenience, we also denote $p_i^* = (t_i^*, b_i^*)$ for the correct entity pages. Again, t_{ij}, b_{ij}, t_i^* and b_i^* are also sequences of words.

In order to link the entity mentions, the strategy is to assign a relevance score $\phi(m_i, p_{ij})$ for each target candidate p_{ij} of m_i, and then use these scores to rank the candidates for each mention. In this work, we decompose $\phi(m_i, p_{ij})$ as the sum of the two following factors:

$$\phi(m_i, p_{ij}) = \phi_{local}(m_i, p_{ij}) + \phi_{global}(m_1, m_2, \dots, m_i, P_1, P_2, \dots, P_i)$$

In this formula, $\phi_{local}(m_i, p_{ij})$ represents the local similarities between m_i and p_{ij}, i.e, only using the information related to m_i and p_{ij}. $\phi_{global}(m_1, m_2, \dots, m_i, P_1, P_2, \dots, P_i)$, on the other hand, additionally considers the other mentions and candidates in the document, attempting to model the interdependence among these objects. The denotation $\phi_{global}(m_1, m_2, \dots, m_i, P_1, P_2, \dots, P_i)$ implies that we are computing the ranking scores for all the target candidates of all the entity mentions in each document simultaneously, preserving the order of the entity mentions from the beginning to the end of the document.

The model in this work consists of three main components: (i) the encoding component that applies convolutional neural networks to induce the distributed representations for the input sequences s_i, c_i, d_i, t_{ij}, and b_{ij}, (ii) the local component that computes the local similarities $\phi_{local}(m_i, p_{ij})$ for each entity mention m_i, and (iii) the global component that runs recurrent neural networks on the entity mentions $\{m_1, m_2, \dots, m_k\}$ to generate the global features $\phi_{global}(m_1, m_2, \dots, m_i, P_1, P_2, \dots, P_i)$.

2.1 Encoding

Let x be some context word sequence of the entity mentions or target candidates (i.e, $x \in \{s_i, c_i, d_i\}_i \cup \{t_{ij}, p_{ij}\}_{i,j} \cup \{t_i^*, b_i^*\}_i$). In order to obtain the distributed representation for x, we first transform each word $x_i \in x$ into a real-valued, h-dimensional vector w_i using the word embedding table E (Mikolov et al., 2013): $w_i = E[x_i]$. This essentially converts the word sequence x into a sequence of vectors that is padded with zero vectors to form a fixed-length sequence of vectors $w = (w_1, w_2, \dots, w_n)$ of length n.

In the next step, we apply the convolution operation over w to generate the hidden vector sequence, that is then transformed by a non-linear function G and pooled by the sum function (Francis-Landau et al., 2016). Following the previous work on CNN (Nguyen and Grishman, (2015a; 2015b)), we utilize the set L of multiple window sizes to parameterize the convolution operation. Each window size $l \in L$ corresponds to a convolution matrix $M_l \in \mathbb{R}^{v \times lh}$ of dimensionality v. Eventually, the concatenation vector $\bar{x}$ of the resulting vectors for each window size in L would be used as the distributed representation for x:

$$\bar{x} = \bigoplus_{l \in L} \sum_{i=1}^{n-l+1} G(M_l w_{i:(i+l-1)})$$

where $\bigoplus$ is the concatenation operation over the window set L and $w_{i:(i+l-1)}$ is the concatenation vector of the given word vectors.

For convenience, let $\bar{s}_i$, $\bar{c}_i$, $\bar{d}_i$, $\bar{t}_{ij}$, $\bar{b}_{ij}$, $\bar{t}_i^*$ and $\bar{b}_i^*$ be the distributed representations of s_i, c_i, d_i, t_{ij}, p_{ij}, t_i^* and b_i^* obtained by the convolution procedure above, respectively. Note that we apply the same set of convolution parameters for each type of text granularity in the source document D as well as in the

[1] For comparison purpose, we use the target candidates provided by Francis-Landau et al. (2016). Essentially, a query generation is executed for each entity mention, whose outputs are combined with link counts to retrieve the potential entities (including "*NIL*"). The query generation itself involves removing stop words, plural suffixes, punctuation, and leading or tailing words.

target entity side. The vector representations of the context would then be fed into the next components to compute the features for EL.

2.2 Local Similarities

We employ the local similarities $\phi_{local}(m_i, p_{ij})$ from (Francis-Landau et al., 2016), the state-of-the-art neural network model for EL. In particular:

$$\phi_{local}(m_i, p_{ij}) = \phi_{sparse}(m_i, p_{ij}) + \phi_{CNN}(m_i, p_{ij}) = W_{sparse}F_{sparse}(m_i, p_{ij}) + W_{CNN}F_{CNN}(m_i, p_{ij})$$

In this formula, W_{sparse} and W_{CNN} are the weights for the feature vectors F_{sparse} and W_{CNN} respectively. $F_{sparse}(m_i, p_{ij})$ is the sparse feature vector obtained from (Durrett and Klein, 2014). This vector captures various linguistic properties and statistics that have been discovered in the previous studies for EL. The representative features include the anchor text counts from Wikipedia, the string match indications with the title of the Wikipedia candidate pages, or the information about the shape of the queries for candidate generations (Francis-Landau et al., 2016).

$F_{CNN}(m_i, p_{ij})$, on the other hand, involves the cosine similarities between the representation vectors at multiple granularities of m_i and p_{ij}. In particular:

$$F_{CNN}(m_i, p_{ij}) = [cos(\bar{s}_i, \bar{t}_{ij}), cos(\bar{c}_i, \bar{t}_{ij}), cos(\bar{d}_i, \bar{t}_{ij}), cos(\bar{s}_i, \bar{b}_{ij}), cos(\bar{c}_i, \bar{b}_{ij}), cos(\bar{d}_i, \bar{b}_{ij})] \quad (1)$$

The intuition for this computation is that the similarities at different levels of contexts might help to enforce the potential topic compatibility between the contexts of the entity mentions and target candidates for EL (Francis-Landau et al., 2016).

2.3 Global Similarities

In order to encapsulate the coherence among the entity mentions and their target entities, we run recurrent neural networks over the sequences of the representation vectors for the entity mentions (i.e, the vector sequences for the surface strings $(\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_k)$ and for the immediate contexts $(\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k)$) and the target entities (i.e, the vector sequences for the page titles $(\bar{t}^*_1, \bar{t}^*_2, \ldots, \bar{t}^*_k)$ and for the body contents $(\bar{b}^*_1, \bar{b}^*_2, \ldots, \bar{b}^*_k)$)[2].

Let us take the representation vector sequence of the body contents of the target pages $(\bar{b}^*_1, \bar{b}^*_2, \ldots, \bar{b}^*_k)$[3] as an example. The recurrent neural network with the recurrent function Φ for this sequence will generate the hidden vector sequence $(h^b_1, h^b_2, \ldots, h^b_k)$ where: $h^b_i = \Phi(h^b_{i-1}, \bar{b}^*_i)$.

Each vector h^b_i in this sequence encodes or summarizes the information about the content of the previous target entities (i.e, before i) in the document due to the property of RNN.

Given the hidden vector sequence, when predicting the target entity for the entity mention m_i, we ensure that the target entity is consistent with the global information stored in h^b_{i-1}. This is achieved by using the cosine similarities between h^b_{i-1} and the representation vectors of each target candidate p_{ij} of m_i, (i.e, $cos(h^b_{i-1}, \bar{t}_{ij})$ and $cos(h^b_{i-1}, \bar{b}_{ij})$) as the global features for the ranking score.

We can repeat this process for the other representation vector sequences in both the entity mention side and the target entity side. The resulting global features would then be grouped into a single feature vector to compute the global similarity score $\phi_{global}(m_1, m_2, \ldots, m_i, P_1, P_2, \ldots, P_i)$ as in the local similarity section. An overview of the whole model is presented in Figure 1.

Regarding the reccurent function Φ, we employ the gated recurrent units (GRU) (Cho et al., 2014) to alleviate the "*vanishing gradient problem*" of RNN. GRU is a simplified version of long-short term memory units (LSTM) that has been shown to achieve comparable performance (Józefowicz et al., 2015).

Finally, for training, we jointly optimize the parameters for the CNNs, RNNs and weight vectors by maximizing the log-likelihood of a labeled training corpus. We utilize the stochastic gradient descent algorithm and the AdaDelta update rule (Zeiler, 2012). The gradients are computed via back-propagation. Following (Francis-Landau et al., 2016), we do not update the word embedding table during training.

[2] Note that we have different recurrent neural networks for different context vector sequences.

[3] In the training process, $(\bar{b}^*_1, \bar{b}^*_2, \ldots, \bar{b}^*_k)$ are obtained from the golden target entities while in the test time, they are retrieved from the predicted target entities.

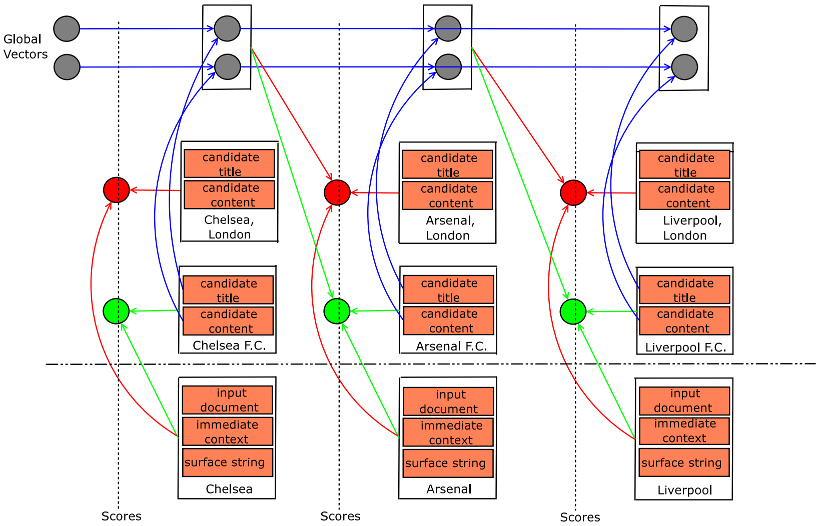

Figure 1: Joint model for learning local and global features for a document with 3 entity mentions: *Chelsea*, *Arsenal* and *Liverpool*. Each of the entity mentions has two entity candidate pages (either a football club or a city).The orange rectangles denote the CNN-induced representation vectors $\bar{s}_i$, $\bar{c}_i$, $\bar{d}_i$, $\bar{t}_{ij}$ and $\bar{b}_{ij}$. The circles in red and green are the ranking scores for the target candidates, in which the green circles correspond to the correct target entities. Finally, the circles in grey are the hidden vectors (i.e, the global vectors) of the RNNs running over the entity mentions. We only show the global entity vectors in this figure to improve the visualization.

3 Experiments

3.1 Datasets

Following (Francis-Landau et al., 2016), we evaluate the models on 4 different entity linking datasets:

i) ACE (Bentivogli et al., 2010): This corpus is from the 2005 evaluation of NIST. It is also used in (Fahrni and Strube, 2014) and (Durrett and Klein, 2014).

ii) CoNLL-YAGO (Hoffart et al., 2011): This corpus is originally from the CoNLL 2003 shared task of named entity recognition for English.

iii) WP (Heath and Bizer, 2011): This dataset consists of short snippets from Wikipedia.

iv) WIKI (Ratinov et al., 2011): This dataset contains 10,000 randomly sampled Wikipedia articles. The task is to disambiguate the links in each article[4].

For all the datasets, we use the standard data splits (for training data, test data and development data) as the previous works for comparable comparison (Francis-Landau et al., 2016).

3.2 Parameters and Resources

For all the experiments below, in the CNN models to learn the distributed representations for the inputs, we use window sizes in the set $L = \{2, 3, 4, 5\}$ for the convolution operation with the dimensionality $v = 200$ for each window size[5]. The non-linear function for transformation is $G = \tanh$.

We employ the English Wikipedia dump from June 2016 as our reference knowledge base.

Regarding the input contexts for the entity mentions and the target candidates, we utilize the window size of 10 for the immediate context c_i, and only extract the first 100 words in the documents for d_i and b_{ij}.

Finally, we pre-train the word embedings on the whole English Wikipedia dump using the word2vec toolkit (Mikolov et al., 2013). The training parameters are set to the default values in this toolkit. The dimensionality of the word embeddings is 300.

Note that every parameter and resource in this work is either taken from the previous work (Nguyen and Grishman, 2016b; Francis-Landau et al., 2016) or selected by the development data.

[4] As noted by Francis-Landau et al. (2016) and Nguyen et al. (2014b), the original Wikipedia dump in Ratinov et al. (2011) is no longer accessible, so we cannot duplicate the results or conduct comparable experiments with (Ratinov et al., 2011). We instead compare our performance with (Francis-Landau et al., 2016) that provides the access to their Wikipedia dump.

[5] As we need to compute the cosine similarities between the hidden vectors of the RNN models and the representation vectors of the target candidates, the number of hidden units for the RNN is set to $200|L| = 800$ naturally.

3.3 Evaluating the Global Features

In this section, we evaluate the effectiveness of the global features for EL. In particular, we differentiate two types of global features based on the side of information we expect to enforce the coherence. The first type of global features (***global-mention***) concerns the entity mention side and involves applying the global RNN models on the CNN-induced representation vectors of the entity mentions (i.e, the surface vectors $(\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_k)$ and the immediate context vectors $(\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k)$). The second type of global features (***global-entity***), on the other hand, focuses on the target entity side and models the coherence with the representation vectors of the target entities (i.e, the page title vectors $(\bar{t}^*_1, \bar{t}^*_2, \ldots, \bar{t}^*_k)$ and the body content vectors $(\bar{b}^*_1, \bar{b}^*_2, \ldots, \bar{b}^*_k)$). Table 1 reports the development performance (F1 scores) of the proposed model on different cases where the *global-mention* and *global-entity* features are included or excluded from the model.

Global Features	Dataset		
	ACE	CoNLL	WP
No	86.1	89.3	84.0
global-mention	86.8	90.2	84.2
global-entity	**86.9**	**90.7**	**84.2**
global-mention + *global-entity*	86.2	90.6	84.0

Table 1: Performance of the global features on the development set. *No* means not using the global features.

The most important observation from the table is that the global features, in general, help to improve the performance of the model on different datasets. This is substantial on the ACE and CoNLL datasets when only one type of the global features (either *global-mention* or *global-entity*) is integrated into the model. The combination of *global-mention* and *global-entity* is not very effective as it is actually worse than the performance of the individual global feature types. This suggests that *global-mention* and *global-entity* might cover overlapping information and their combination would inject redundancy into the model. The best performance is achieved by the *global-entity* features that would be used in all the evaluations below.

3.4 Comparing to the Previous Work

This section compares the proposed system (called *Global-RNN*) with the state-of-the-art models on our four datasets. These systems include the neural network model in (Francis-Landau et al., 2016), the joint model for entity analysis in (Durrett and Klein, 2014) and the AIDA-light system with two-stage mapping in (Nguyen et al., 2014b)[6]. Table 2 shows the performance of the systems on the test sets with the reference knowledge base of the June 2016 Wikipedia dump. We also include the performance of the systems on the December 2014 Wikipedia dump that was used and provided by (Francis-Landau et al., 2016) for further and compatible comparison.

Systems	Wikipedia 2014				Wikipedia 2016			
	ACE	CoNLL	WP	WIKI	ACE	CoNLL	WP	WIKI
DK2014 (Durrett and Klein, 2014)	79.6	-	-	-	-	-	-	-
AIDA-LIGHT (Nguyen et al., 2014b)	-	84.8	-	-	-	-	-	-
Local CNN (Francis-Landau et al., 2016)	**89.9**	85.5	90.7	82.2	86.1	84.5	90.4	81.4
Global-RNN	89.7	**87.2**†	**91.2**†	**83.7**†	**87.8**†	**86.5**†	**91.2**†	**81.7**

Table 2: Performance of the systems. Cells marked with †designate the *Global-RNN* models that significantly outperform the *Local CNN* model ($\rho < 0.05$).

First, we see that the performance of the systems drop significantly when we switch from Wikipedia 2014 to Wikipedia 2016 (especially for the datasets ACE and CoNLL). This is can be partly explained by the inclusion of new entities (pages) into Wikipedia from 2014 to 2016 that has made the entity mentions in the datasets more ambiguous[7]. Second and more importantly, *Global-RNN* significantly outperforms

[6] We note that (Alhelbawy and Gaizauskas, 2014) and (Pershina et al., 2015) also use the CoNLL-YAGO dataset for their experiments. However, since they evaluate the models on the whole dataset rather than the test set as the other works do, they are not comparable to the performance we report in this paper.

[7] The number of Wikipedia pages in 2014 is about 4.5 million while this number is 5 million in June 2016.

the all the compared models (except for the ACE dataset on Wikipedia 2014 and the WIKI dataset on Wikipedia 2016), thereby demonstrating the benefits of the joint modeling for local and global features via neural networks for EL in this work.

3.5 Domain Adaptation Experiments

The purpose of this section is to further evaluate the models in the domain adaptation setting to investigate their cross-domain robustness for EL.

It is often observed in many natural language processing tasks that the performance of a model trained on a source domain would degrade significantly when it is applied to a different target domain (Blitzer et al., 2006; Daume, 2007; McClosky et al., 2010; Plank and Moschitti, 2013; Nguyen and Grishman, 2014a). Such a performance loss originates from a variety of mismatches between the source and the target domains, including the differences in vocabulary, data distributions, styles etc. This has motivated the domain adaptation research that aims to improve the cross-domain performance of the models by adaptation techniques.

One of the key strategies of the domain adaptation techniques is the search for the domain-independent features that are discriminative across different domains (Blitzer et al., 2006; Jiang, 2009; Plank and Moschitti, 2013; Nguyen and Grishman, 2014a). These invariants serve as the connectors between different domains and help to transfer the knowledge from one domain to the others. For EL, we hypothesize that the global coherence is an effective domain-independent feature that would help to improve the cross-domain performance of the models. The intuition is that the entities mentioned in a document of any domains should be related to each other. Eventually, we expect that the proposed model with global coherence features would be more robust to domain shifts than the local approach (Francis-Landau et al., 2016).

3.5.1 Dataset

We use the ACE dataset to evaluate the cross-domain performance of the models. ACE involves documents in 6 different domains: broadcast conversation (bc), broadcast news (bn), telephone conversation (cts), newswire (nw), usenet (un) and webblogs (wl). Following the common practice of domain adaptation research on this dataset (Plank and Moschitti, 2013; Nguyen et al., 2015c; Gormley et al., 2015), we use **news** (the union of **bn** and **nw**) as the source domain and **bc**, **cts**, **wl**, **un** as four different target domains. We take half of **bc** as the development set and use the remaining data for testing. We note that **news** consists of formally written documents while a majority of the other domains is informal text, making the source and target domains very divergent in terms of vocabulary and styles (Plank and Moschitti, 2013).

3.5.2 Evaluation

Table 3 compares *Global-RNN* with the neural network EL model in (Francis-Landau et al., 2016), the best reported model on the ACE dataset in the literature[8]. In this table, the models are trained on the source domain **news**, and evaluated on **news** itself (in-domain performance) (via 5-fold cross validation) as well as on the 4 target domains **bc**, **cts**, **wl**, **un** (out-of-domain performance). The experiments in this section are done with the 2016 Wikipedia dump.

Models	Domain				
	in-domain	bc	cts	wl	un
Local CNN (Francis-Landau et al., 2016)	90.6	87.8	88.7	80.2	82.1
Global-RNN	**91.0**	**88.7**†	**88.9**	**81.3**†	**83.1**†

Table 3: Cross-domain performance. Cells marked with †designate the *Glob-RNN* models that significantly outperform the *Local CNN* model ($\rho < 0.05$).

The first observation from the table is that the performance of all the compared systems on the target domains is much worse than the corresponding in-domain performance. In particular, the performance gap between the in-domain performance and the the worst out-of-domain performance (on the domain

[8] The performance of the model from (Francis-Landau et al., 2016) reported in this work is obtained by running their actual released system.

wl) is up to 10%, thus indicating the mismatches between the source and the target domains for EL. Second and most importantly, *Global-RNN* is consistently better than the model with only local features in (Francis-Landau et al., 2016) over all the target domains (although it is less pronounced in the **cts** domain). This demonstrates the cross-domain robustness of the proposed model and confirms our hypothesis about the domain-independence of the global coherence features for EL.

3.5.3 Analysis

In order to better understand the performance gap in the domain adaptation experiments for EL, we visualize the representation vectors of the entity mentions in different domains. In particular, after *Global-RNN* is trained, we retrieve the representation vectors $\bar{c}_i$ for the immediate contexts of the entity mentions in the source and target domains, project them into the 2-dimension space via the t-SNE algorithm and plot them. Figure 2 shows the plot.

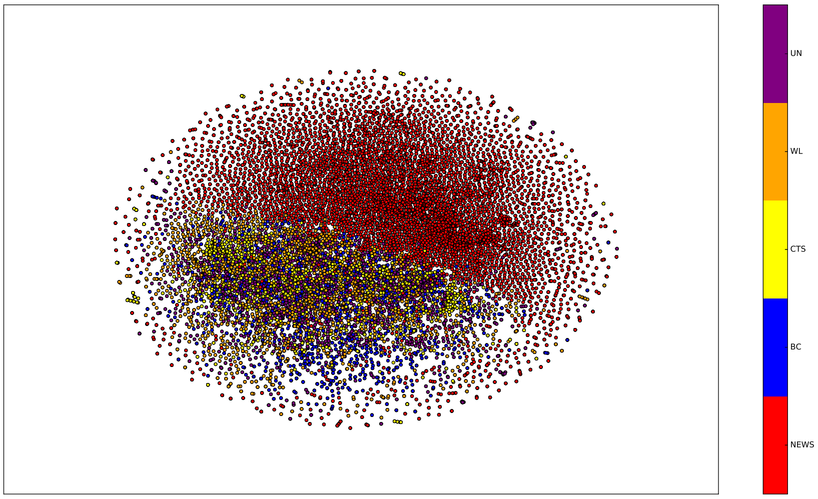

Figure 2: t-SNE visualization on the representation vectors c_i of different domains.

As we can see from the figure, the entity mentions in the target domains **bc**, **cts**, **wl** and **un** are quite separated from those of the source domain **news**, thereby explaining the performance loss in the domain adaption experiments.

It is not clear in Figure 2 why the models perform much worse on the target domains **wl** and **un** than the other domains (i.e, **bc** and **cts**). We further investigate this problem by computing the similarities between the target domains and the source domain. While there are several methods to estimate domain similarities (Plank and van Noord, 2011), in this work, we employ the mean of the cosine similarities of every mention pairs in the two domains of interest. Specifically, let E and F be the two domains of interest, and $E = \{e_1, e_2, \ldots, e_g\}$ and $F = \{f_1, f_2, \ldots, f_w\}$ be the sets of the representation vectors for the entity mentions in E and F respectively ($g = |E|, w = |F|$). The similarity between E and F is then given by:

$$Sim(E, F) = 100 \times \frac{\sum_{i=1}^{g} \sum_{j=1}^{w} cos(e_i, f_j)}{gw}$$

Table 4 shows the similarities between the source domain **news** and each target domains **bc**, **cts**, **wl** and **un** with respect to the representation vectors of the immediate context $\bar{c}_i$ (*context*) and the target entity titles $\bar{t}_i^*$ (*title*) for the entity mentions m_i. We also include the similarities in which the representation vectors are the local feature vectors $F_{CNN}(m_i, t_i^*)$ in Equation 1 (*interaction*). The goal of the local feature similarities is to characterize how the entity mentions in different domains interact with their target entities.

It is clear from the table that **wl** is the most dissimilar domain from the source domain. This is followed by **un** and partly explains the performance in Table 3.

Domain	*context*	*title*	*interaction*
bc	10.7	2.0	34.4
cts	11.4	2.0	32.6
wl	9.2	0.8	30.3
un	9.5	1.4	31.1

Table 4: Similarities to the source domain **news**.

4 Related Work

Entity linking or disambiguation has been studied extensively in NLP research, falling broadly into two major approaches: local and global disambiguation. Both approaches share the goal of measuring the similarities between the entity mentions and the target candidates in the reference KB. The local paradigm focuses on the internal structures of each separate mention-entity pair, covering the name string comparisons between the surfaces of the entity mentions and target candidates, entity popularity or entity type and so on (Bunescu and Pasca, 2006; Milne and Witten, 2008; Zheng et al., 2010; Ji and Grishman, 2011; Mendes et al., 2011; Cassidy et al., 2011; Shen et al., 2014). In contrast, the global approach jointly maps all the entity mentions within documents to model the topical coherence. Various techniques have been exploited for capturing such semantic consistency, including Wikipedia category agreement (Cucerzan, 2007), Wikipedia link-based measures (Kulkarni et al., 2009; Hoffart et al., 2011; Shen et al., 2012), Point-wise Mutual Information measures (Ratinov et al., 2011), integer linear programming (Cheng and Roth, 2013), PageRank (Alhelbawy and Gaizauskas, 2014; Pershina et al., 2015), stacked generalization (He et al., 2013a), to name a few. The entity linking techniques and systems have been actively evaluated at the NIST-organized Text Analysis Conference (Ji et al., 2014).

Neural networks are applied to entity linking very recently. He et al. (2013b) learn enttiy representation via Stacked Denoising Auto-encoders. Sun et al. (2015) employ convolutional neural networks and neural tensor networks to model mentions, entities and contexts while Francis-Landau et al. (2016) combine CNN-based representations with sparse features to improve the performance. However, none of these work utilize recurrent neural networks to capture the coherence features as we do in this work.

5 Conclusion

We present a joint model to learn the local context similarities and the global topical relatedness features for entity linking. CNNs are employed to capture the local similarities while RNNs are utilized to introduce the coherence. The model achieves the state-of-the-art performance on multiple datasets for entity linking. It is also shown to be more robust to domain shifts. Our future work is threefold: (i) integrating entity embedding models into the current work, (ii) exploring new neural models to jointly perform entity linking and entity extraction (Nguyen et al., 2016c), and (iii) further evaluating the models in the cross-lingual settings.

References

Ayman Alhelbawy and Robert Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *ACL*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. In *Journal of Machine Learning Research 3*.

Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. 2010. Extending english ace 2005 corpus annotation with ground-truth links to wikipedia. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Taylor Cassidy, Zheng Chen, Javier Artiles, Heng Ji, Hongbo Deng, Lev-Arie Ratinov, Jing Zheng, Jiawei Han, and Dan Roth. 2011. Cuny-uiuc-sri tac-kbp2011 entity linking system description. In *TAC*.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *EMNLP*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP*.

Hal Daume. 2007. Frustratingly easy domain adaptation. In *ACL*.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *TACL*.

Angela Fahrni and Michael Strube. 2014. A latent variable model for discourse-aware concept and entity disambiguation. In *EACL*.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *NAACL*.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *EMNLP*.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: A graph-based method. In *SIGIR*.

Zhengyan He, Shujie Liu, Yang Song, Mu Li, Ming Zhou, and Houfeng Wang. 2013a. Efficient collective entity linking with stacking. In *EMNLP*.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013b. Learning entity representation for entity disambiguation. In *ACL*.

Tom Heath and Christian Bizer. 2011. Linked data: Evolving the web into a global data space. In *Morgan and Claypool, 1st edition*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *EMNLP*.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *ACL*.

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

Jing Jiang. 2009. Multi-task transfer learning for weakly-supervised relation extraction. In *ACL-IJCNLP*.

Rafal Józefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *ICML*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL*.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *SIGKDD*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *NAACL*.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *CIKM*.

Thien Huu Nguyen and Ralph Grishman. 2014a. Employing word representations and regularization for domain adaptation of relation extraction. In *ACL*.

Thien Huu Nguyen and Ralph Grishman. 2015a. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st NAACL Workshop on Vector Space Modeling for NLP (VSM)*.

Thien Huu Nguyen and Ralph Grishman. 2015b. Event detection and domain adaptation with convolutional neural networks. In *ACL-IJCNLP*.

Thien Huu Nguyen and Ralph Grishman. 2016b. Combining neural networks and log-linear models to improve relation extraction. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence (DLAI)*.

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014b. Aida-light: High-throughput named-entity disambiguation. In *WWW*.

Thien Huu Nguyen, Barbara Plank, and Ralph Grishman. 2015c. Semantic representations for domain adaptation: A case study on the tree kernel-based method for relation extraction. In *ACL-IJCNLP*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016a. Joint event extraction via recurrent neural networks. In *NAACL*.

Thien Huu Nguyen, Avirup Sil, Georgiana Dinu, and Radu Florian. 2016c. Toward mention detection robustness with recurrent neural networks. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence (DLAI)*.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized page rank for named entity disambiguation. In *NAACL*.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *ACL*.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *ACL*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *ACL*.

Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2012. Linden: Linking named entities with knowledge base via semantic knowledge. In *WWW*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. In *TKDE*.

Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking named entities to any database. In *EMNLP*.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL*.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. In *CoRR, abs/1212.5701*.

Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Learning to link entities with knowledge base. In *NAACL*.