Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining

Xiaoman Pan[*], **Thamme Gowda**[‡], **Heng Ji**[*†], **Jonathan May**[‡], **Scott Miller**[‡]
[*] Department of Computer Science [†] Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{xiaoman6,hengji}@illinois.edu
[‡] Information Sciences Institute, University of Southern California
{tg,jonmay,smiller}@isi.edu

Abstract

Entities, which refer to distinct objects in the real world, can be viewed as language universals and used as effective signals to generate less ambiguous semantic representations and align multiple languages. We propose a novel method, *CLEW*, to generate cross-lingual data that is a mix of entities and contextual words based on Wikipedia. We replace each anchor link in the source language with its corresponding entity title in the target language if it exists, or in the source language otherwise. A cross-lingual joint entity and word embedding learned from this kind of data not only can disambiguate linkable entities but can also effectively represent unlinkable entities. Because this multilingual common space directly relates the semantics of contextual words in the source language to that of entities in the target language, we leverage it for unsupervised cross-lingual entity linking. Experimental results show that *CLEW* significantly advances the state-of-the-art: up to 3.1% absolute F-score gain for unsupervised cross-lingual entity linking. Moreover, it provides reliable alignment on both the word/entity level and the sentence level, and thus we use it to mine parallel sentences for all $\binom{302}{2}$ language pairs in Wikipedia.[1]

1 Introduction

The sheer amount of natural language data provides a great opportunity to represent named entity mentions by their probability distributions, so that they can be exploited for many Natural Language Processing (NLP) applications. However, named entity mentions are fundamentally different from common words or phrases in three aspects. First, the semantic meaning of a named entity mention (*e.g.,* a person name "*Bill Gates*") is not a simple summation of the meanings of the words it contains ("*Bill*" + "*Gates*"). Second, entity mentions are often highly ambiguous in various local contexts. For example, "*Michael Jordan*" may refer to the basketball player or the computer science professor. Third, representing entity mentions as mere phrases fails when names are rendered quite differently, especially when they appear across multiple languages. For example, "*Ang Lee*" in English is "*Li An*" in Chinese.

Fortunately, entities, the objects which mentions refer to, are unique and equivalent across languages. Many manually constructed entity-centric knowledge base resources such as Wikipedia[2], DBPedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007) are widely available. Even better, they are massively multilingual. For example, up to August 2018, Wikipedia contains 21 million interlanguage links[3] between 302 languages. We propose a novel **c**ross-**l**ingual joint **e**ntity and **w**ord (*CLEW*) embedding learning framework based on multilingual Wikipedia and evaluate its effectiveness on two practical NLP applications: Cross-lingual Entity Linking and Parallel Sentence Mining.

Wikipedia contains rich entity anchor links. As shown in Figure 2, many mentions (*e.g.,* "小米" (*Xiaomi*)) in a source language are linked to the entities in the same language that they refer to (*e.g.,* zh/小米科技 (*Xiaomi Technology*)), and some mentions are further linked to their corresponding English entities (*e.g.,* Chinese mention "苹果" (*Apple*) is linked to entity en/Apple_Inc. in English). We replace each mention (anchor link) in the source language with its corresponding entity title in the target language if it exists, or in

[1] We make all software and resources publicly available for research purpose at http://panx27.github.io/wikiann.

[2] https://www.wikipedia.org

[3] https://en.wikipedia.org/wiki/Help:Interlanguage_links

the source language otherwise. After this replacement, each entity mention is treated as a unique disambiguated entity, then we can learn joint entity and word embedding representations for the source language and target language respectively.

Furthermore, we leverage these shared target language entities as pivots to learn a rotation matrix and seamlessly align two embedding spaces into one by linear mapping. In this unified common space, multiple mentions are reliably disambiguated and grounded, which enables us to directly compute the semantic similarity between a mention in a source language and an entity in a target language (*e.g.,* English), and thus we can perform Cross-lingual Entity Linking in an unsupervised way, without using any training data. In addition, considering each pair of Wikipedia articles connected by an inter-language link as comparable documents, we use this multilingual common space to represent sentences and extract many parallel sentence pairs.

The novel contributions of this paper are:

- We develop a novel approach based on rich anchor links in Wikipedia to learn cross-lingual joint entity and word embedding, so that entity mentions across multiple languages are disambiguated and grounded into one unified common space.
- Using this joint entity and word embedding space, entity mentions in any language can be linked to an English knowledge base without any annotation cost. We achieve state-of-the-art performance on unsupervised cross-lingual entity linking.
- We construct a rich resource of parallel sentences for $\binom{302}{2}$ language pairs along with accurate entity alignment and word alignment.

2 Approach

2.1 Training Data Generation

Wikipedia contains rich entity anchor links. For example, in the following sentence from English Wikipedia markup: "**[[Apple Inc.|apple]]** is a technology company.", where **[[Apple Inc.|apple]]** is an anchor link that links the anchor text *"apple"* to the entity `en/Apple_Inc.`[4]

[4]In this paper, we use `langcode/entity_title` to represent entities in Wikipedia in each individual language. For example, `en/*` refers to an entity in English Wikipedia `en.wikipedia.org/wiki/*`.

Traditional approaches to derive training data from Wikipedia usually replace each anchor link with its anchor text, for example, "**apple** is a technology company.". These methods have two limitations: (1) **Information loss**: For example, the anchor text *"apple"* itself does not convey information such as the entity is a company; (2) **Ambiguity** (Faruqui et al., 2016): For example, the fruit sense and the company sense of *"apple"* mistakenly share one surface form. Similar to previous work (Wang et al., 2014; Tsai and Roth, 2016; Yamada et al., 2016), we replace each anchor link with its corresponding entity title, and thus treat each entity title as a unique word. For example, "`en/Apple_Inc.` is a technology company.". Using this kind of data mix of entity titles and contextual words, we can learn joint embedding of entities and words.

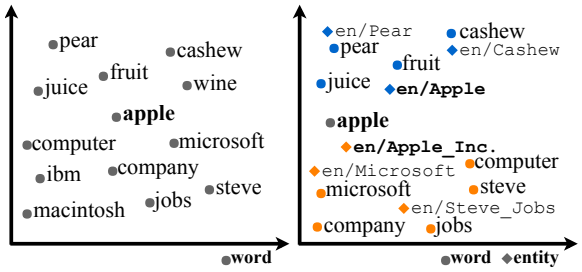

Figure 1: Traditional word embedding (left), and joint entity and word embedding (right).

The results from traditional word embedding and joint entity and word embedding for "*apple*" are visualized through Principal Component Analysis (PCA) in Figure 1. Using the joint embedding we can successfully separate those words referring to fruit and others referring to companies in the vector space. Moreover, the similarity can be computed based on entity-level instead of word-level. For example, `en/Apple_Inc` and `en/Steve_Jobs` are close in the vector space because they share many context words and entities.

Moreover, the above approach can be easily extended to the cross-lingual setting by using Wikipedia inter-language links. We replace each anchor link in a source language with its corresponding entity title in a target language if it exists, and otherwise replace each anchor link with its corresponding entity title in the source language. An example is illustrated in Figure 2.

Using this approach, the entities in a target language can be embedded along with words and the entities in a source language, as illustrated in Fig-

Figure 2: Using Wikipedia inter-language links to generate sentences which contain words and entities in a source language (*e.g.,* Chinese) and entities in a target language (*e.g.,* English).

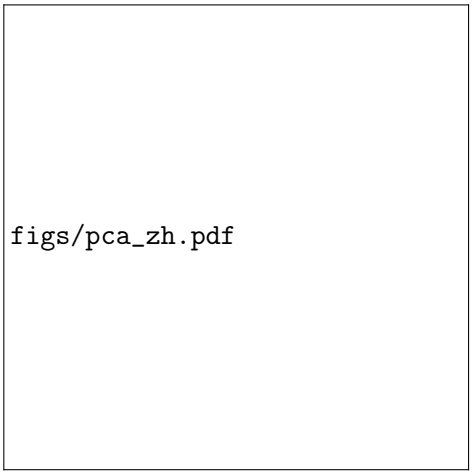

Figure 3: Embedding which includes entities in English, and words and entities in Chinese (English words in brackets are human translations of Chinese words).

ure 3.

This joint representation has two advantages: (1) **Disambiguation**: For example, two entities `en/Apple_Inc.` and `en/Apple` can be differentiated by their distinct neighbors "电脑" (*computer*) and "水果" (*fruit*) respectively. (2) **Effective representation of unknown entities**: For example, the new entity `zh/`小米科技 (*Xiaomi Technology*), a Chinese mobile phone manufacturer, may not have an English Wikipedia page yet. But because it's close to neighbors such as `en/Microsoft`, "手机" (*phone*) and "公司" (*company*), we can infer it's likely to be a technology company.

2.2 Linear Mapping across Languages

Word embedding spaces have similar geometric arrangements across languages (Mikolov et al., 2013b). Given two sets of independently trained word embedding, the source language embedding $\mathcal{Z}^S$ and the target language embedding $\mathcal{Z}^T$, and a set of pre-aligned word pairs, a linear mapping $\mathbf{W}$ is learned to transform $\mathcal{Z}^S$ into a shared space where the distance between the embedding of the source language word and the embedding of its pre-aligned target language word is minimized. For example, given a set of pre-aligned word pairs, we use $\mathbf{X}$ and $\mathbf{Y}$ to denote two aligned matrices which contain the embedding of the pre-aligned words from $\mathcal{Z}^S$ and $\mathcal{Z}^T$ respectively. A linear mapping $\mathbf{W}$ can be learned such that:

$$\underset{\mathbf{W}}{\operatorname{argmin}}||\mathbf{WX} - \mathbf{Y}||_{\mathrm{F}}$$

Previous work (Xing et al., 2015; Smith et al., 2017) shows that enforcing an orthogonal constraint $\mathbf{W}$ yields better performance. Consequently, the above equation can be transferred to Orthogonal Procrustes problem (Conneau et al., 2017):

$$\underset{\mathbf{W}}{\operatorname{argmin}}||\mathbf{WX} - \mathbf{Y}||_{\mathrm{F}} = \mathbf{UV}^\top$$

Then $\mathbf{W}$ can be obtained from the singular value decomposition (SVD) of $\mathbf{YX}^\top$ such that:

$$\mathbf{U\Sigma V}^\top = \mathrm{SVD}(\mathbf{YX}^\top)$$

In this paper, we propose using entities instead of pre-aligned words as anchors to learn such a linear mapping $\mathbf{W}$. The basic idea is illustrated in Figure 4. We use $\mathcal{E}_T$ and $\mathcal{W}_T$ to denote the sets of entities and words in the target language associated with the target entity and word embedding $\mathcal{Z}^T$:

$$\mathcal{Z}^T = \{\mathbf{z}^t_{e_1}, .., \mathbf{z}^t_{e_{|\mathcal{E}_T|}}, \mathbf{z}^t_{w_1}, .., \mathbf{z}^t_{w_{|\mathcal{W}_T|}}\}$$

Similarly, we use $\mathcal{E}_S$ and $\mathcal{W}_S$ to denote the sets of entities and words in the source language associated with the source entity and word embedding $\mathcal{Z}^S$:

$$\mathcal{Z}^S = \{\mathbf{z}^s_{e_1}, .., \mathbf{z}^s_{e_{|\mathcal{E}_S|}}, \mathbf{z}^s_{w_1}, .., \mathbf{z}^s_{w_{|\mathcal{W}_S|}}\}$$

and use $\mathcal{E'}_T$ to denote the set of entities in the source language which are replaced with the corresponding entities in the target language, where $\mathcal{E'}_T \in \mathcal{E}_T$. Then $\mathcal{Z}^S$ can be represented as

$$\mathcal{Z}^S = \{\mathbf{z}^{t'}_{e_1}, .., \mathbf{z}^{t'}_{e_{|\mathcal{E'}_T|}}, \mathbf{z}^s_{e_1}, .., \mathbf{z}^s_{e_{|\mathcal{E}_S|-|\mathcal{E'}_T|}}, \mathbf{z}^s_{w_1}, .., \mathbf{z}^s_{w_{|\mathcal{W}_S|}}\}$$

Note that $\mathbf{z}^t_{e_i}$ and $\mathbf{z}^{t'}_{e_i}$ are the embedding of e_i in $\mathcal{Z}^T$ and $\mathcal{Z}^S$ respectively. Therefore, using entities in $\mathcal{E'}_T$ as anchors, we can learn a linear mapping

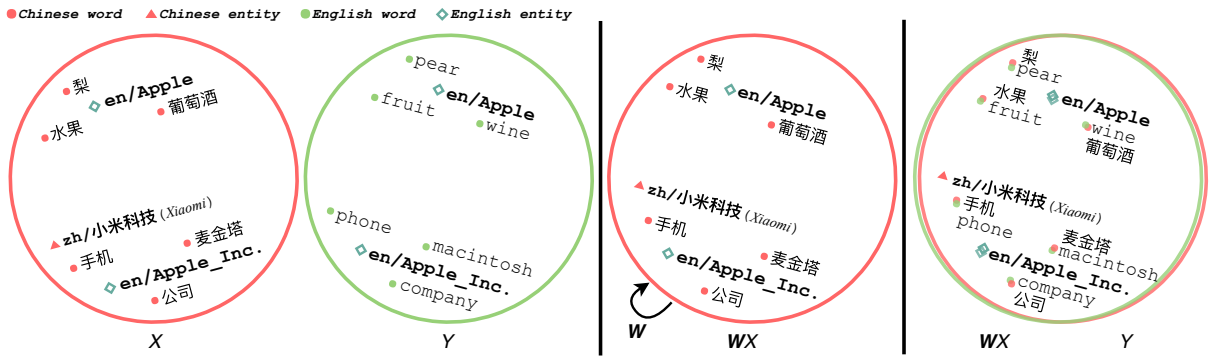

Figure 4: Using the aligned entities as anchors to learn a linear mapping (rotation matrix) which maps a source language embedding space to a target language embedding space.

$\mathbf{W}$ that maps $\mathcal{Z}^S$ into the vector space of $\mathcal{Z}^T$, and obtain the cross-lingual joint entity and word embedding $\mathcal{Z}$.

We adopt the refinement procedure proposed by Conneau et al. (2017) to improve the quality of $\mathbf{W}$. A set of new high-quality anchors is generated to refine $\mathbf{W}$ learned from $\mathcal{E}'_T$. High-quality anchors refer to entities that have high frequency (*e.g.,* top 5,000) and entities that are mutual nearest neighbors. We iteratively apply this procedure to optimize $\mathbf{W}$. At each iteration, the new high-quality anchors are exploited to learn a new mapping.

Conneau et al. (2017) also propose a novel comparison metric, Cross-domain Similarity Local Scaling (CSLS), to relieve the hubness phenomenon, where some vectors (*hubs*) are the nearest neighbors of many others. For example, entity `en/United_States` is a *hub* in the vector space. By employing this metric, the similarity of isolated vectors is increased, while the similarity of vectors in dense areas is decreased. Specifically, given a mapped source embedding $\mathbf{Wx}$ and a target embedding $\mathbf{y}$, the mean cosine similarity of $\mathbf{Wx}$ and $\mathbf{y}$ for their K nearest neighbors in the other language, $r_T(\mathbf{Wx})$ and $r_S(\mathbf{y})$ are computed respectively. The comparison metric is defined as follows:

$$\text{CSLS}(\mathbf{Wx}, \mathbf{y}) = \cos(\mathbf{Wx}, \mathbf{y}) - r_T(\mathbf{Wx}) - r_S(\mathbf{y})$$

Conneau et al. (2017) show that the performance is essentially the same when $K = 5, 10, 50$. Following this work, we set $K = 10$.

3 Downstream Applications

We apply *CLEW* to enhance two important downstream tasks: Cross-lingual Entity Linking and Parallel Sentence Mining.

3.1 Unsupervised Cross-lingual Entity Linking

Cross-lingual Entity Linking aims to link an entity mention in a source language text to its referent entity in a knowledge base (KB) in a target language (*e.g.,* English Wikipedia). A typical Cross-lingual Entity Linking framework includes three steps: mention translation, entity candidate generation, and mention disambiguation. We use translation dictionaries collected from Wikipedia (Ji et al., 2009) to translate each mention into English. If a mention has multiple translations, we merge the linking results of all translations at the end. We adopt a dictionary-based approach (Medelyan and Legg, 2008) to generate entity candidates for each mention. Then we use *CLEW* to implement the following two widely used mention disambiguation features: Context Similarity and Coherence.

Context Similarity refers to the context similarity between a mention and a candidate entity. Given a mention m, we consider the entire sentence containing m as its local context. Using *CLEW* embedding $\mathcal{Z}$, the vectors of context words are averaged to obtain the context vector representation of m:

$$\mathbf{v}_m = \frac{1}{|\mathcal{W}_m|} \sum_{w \in \mathcal{W}_m} \mathbf{z}_w$$

where $\mathcal{W}_m$ is the set of context words of m, and $\mathbf{z}_w \in \mathcal{Z}$ is the embedding of the context word w. We measure context similarity between m and each of its entity candidates by using the cosine similarity between $\mathbf{v}_m$ and entity embedding $\mathbf{z}_e \in \mathcal{Z}$ such that:

$$\mathcal{F}_{\text{txt}}(e) = \cos(\mathbf{v}_m, \mathbf{z}_e) = \frac{\mathbf{v}_m \cdot \mathbf{z}_e}{\|\mathbf{v}_m\| \, \|\mathbf{z}_e\|}$$

Feature	Description
$\mathcal{F}_{\text{prior}}(e)$	Entity Prior: $\frac{\lvert A_{e,*}\rvert}{\lvert A_{*,*}\rvert}$, where $A_{e,*}$ is a set of anchor links that link to entity e and $A_{*,*}$ is all anchor links in the KB
$\mathcal{F}_{\text{prob}}(e\lvert m)$	Mention to Entity Probability: $\frac{\lvert A_{e,m}\rvert}{\lvert A_{*,m}\rvert}$, where $A_{*,m}$ is a set of anchor links with anchor text m and $A_{e,m}$ is a subset that links to entity e.
$\mathcal{F}_{\text{type}}(e\lvert m,t)$	Entity Type (Ling et al., 2015): $\frac{p(e\lvert m)}{\sum_{e\mapsto t} p(e\lvert m)}$, where $e \mapsto t$ indicates that t is one of e's entity types. Conditional probability $p(e\lvert m)$ can be estimated by $\mathcal{F}_{\text{prob}}(e\lvert m)$.

Table 1: Mention disambiguation features.

Coherence is driven by the assumption that if multiple mentions appear together within a context window, their referent entities are more likely to be strongly connected to each other in the KB. Previous work (Cucerzan, 2007; Milne and Witten, 2008; Hoffart et al., 2011; Ratinov et al., 2011; Cheng and Roth, 2013; Ceccarelli et al., 2013; Ling et al., 2015) considers the KB as a knowledge graph and models coherence based on the overlapped neighbors of two entities in the knowledge graph. These approaches heavily rely on explicit connections among entities in the knowledge graph and thus cannot capture the coherence between two entities that are implicitly connected. For example, two entities `en/Mosquito` and `en/Cockroach` only have very few overlapped neighbors in the knowledge graph, but they usually appear together and have similar contexts in text. Using *CLEW* embedding $\mathcal{Z}$, the coherence score can be estimated by cosine similarity between the embedding of two entities. This coherence metric pays more attention to semantics.

We consider mentions that appear in the same sentence as coherent. Let m be a mention, and $\mathcal{C}_e$ be the set of corresponding entity candidates of m's coherent mentions. The coherence score for each of m's entity candidates is the average:

$$\mathcal{F}_{\text{coh}}(e) = \frac{1}{|\mathcal{C}_e|} \sum_{c_e \in \mathcal{C}_e} \cos(\mathbf{z}_e, \mathbf{z}_{c_e})$$

Finally, we linearly combine these two features with several other common mention disambiguation features as shown in Table 1.

3.2 Parallel Sentence Mining

One major bottleneck of low-resource language machine translation is the lack of parallel sentences. This inspires us to mine parallel sentences from Wikipedia automatically using *CLEW* embedding $\mathcal{Z}$.

Wikipedia contributors tend to translate some content from existing articles in other languages while editing an article. Therefore, if there exists an inter-language link between two Wikipedia articles in different languages, these two articles can be considered comparable and thus they are very likely to contain parallel sentences. We represent a Wikipedia sentence in any of the 302 languages by aggregating the embedding of entities and words it contains. In order to penalize high frequent words and entities, we apply a weighted metric:

$$\text{IDF}(t, \mathcal{S}) = \log\left(\frac{|\mathcal{S}|}{|\{s \in \mathcal{S} : t \in s\}|}\right)$$

where t is a term (entity or word), $\mathcal{S}$ is an article containing $|\mathcal{S}|$ sentences, and $|\{s \in \mathcal{S} : t \in s\}|$ is the total number of sentences containing t. The embedding of a sentence $\mathbf{v}_s$ can be computed as:

$$\mathbf{v}_s = \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \text{IDF}(t, \mathcal{S}) \cdot \mathbf{z}_t$$

where $\mathcal{T}_s$ is the set of terms of s and $\mathbf{z}_t \in \mathcal{Z}$ is the embedding of t.

Given two comparable Wikipedia articles connected by an inter-language link, we compute the similarity of all possible sentence pairs using the CSLS metric described in Section 2.2 and rank them. If the CSLS score of a sentence pair is greater than a threshold (in this paper, we empirically set the threshold to 0.1 based on a separate small development set), then the sentence pair is considered as parallel. An advantage of our approach is that it provides a similarity score for every term pair, which can be used for improving word alignment and entity alignment.

4 Experiments

4.1 Training Data

We use an April 1, 2018 Wikipedia XML dump to generate data to train the joint entity and word embedding. We only select and analyze those main Wikipedia pages (`ns` tag is 0) which are not redirected (`redirect` tag is None) using the approach described in Section 2.1. We use the Skip-gram model in Word2Vec (Mikolov et al., 2013a,c) to learn the unaligned embeddings. The number of dimensions of the embedding is set to 300, and the minimal number of occurrences, the size of the context window, and the learning rate are set to 5, 5, and 0.025 respectively.

4.2 Linear Mapping

A large number of aligned entities can be obtained using the approach described in Section 2.1. For example, there are about 400,000 aligned entities between English and Spanish. However, the mapping algorithm does not perform well if we try to align all anchors, because the embedding of rare entities is updated less often, and thus their contexts are very different across languages. Therefore, we learn the global mapping using only high-quality anchors, and select high-frequency entities only as anchors using the salience metric described in Table 1. We use 5,000 anchors for training and 1,500 anchors for testing for each language pair. Our proposed method is applied to 9 language pairs in our experiments. Table 2 shows the statistics and the performance. We can see that mapping a language to its related language (*e.g.,* Ukrainian to Russian) usually achieves better performance.

Source-Target	$P@1$	$P@5$	$P@10$
es-en	79.1	89.2	92.3
it-en	74.5	86.9	90.5
ru-en	68.4	82.8	86.7
tr-en	59.0	79.9	86.3
uk-en	63.0	79.7	85.9
zh-en	63.1	83.8	89.2
uk-ru	78.1	90.3	92.8
ru-uk	75.8	90.2	93.7

Table 2: Linear entity mapping statistics and performance (Precision (%) at K) (en: English, es: Spanish, it: Italian, ru: Russian, so: Somali, tr: Turkish, uk: Ukrainian, zh: Chinese).

4.3 Cross-lingual Entity Linking

We use the training set and evaluation set (LDC2015E75 and LDC2015E103) in TAC Knowledge Base Population (TAC-KBP) 2015 Tri-lingual Entity Linking Track (Ji et al., 2015) for the cross-lingual entity linking experiments, because these data sets include the most recent and comprehensive gold-standard annotations on this task and we can compare our model with previously reported state-of-the-art approaches on the same benchmark.

We first compare our unsupervised approach to the top TAC2015 unsupervised system reported by Ji et al. (2015). In order to have a fair comparison with the state-of-the-art supervised methods, we also combine the features as described in Section 3.1 in a point-wised learning to rank algorithm based on Gradient Boosted Regression Trees (Friedman, 2000). The learning rate and the maximum depth of the decision trees are set to 0.01 and 4 respectively. The results are shown in Table 3. We can see that our unsupervised and supervised approaches significantly outperform the best TAC15 systems.

Method	ENG	CMN	SPA
Best TAC15 Unsupervised	67.1	78.1	71.5
Our Unsupervised	**70.0**	**81.2**	**73.4**
w/o Context Similarity	66.9	79.0	70.6
w/o Coherence	68.5	78.6	71.4
Best TAC15 Supervised	73.7	83.1	80.4
(Tsai and Roth, 2016)	-	83.6	80.9
(Sil et al., 2017)	-	84.4	82.3
Our Supervised	**74.8**	84.2	82.1
w/o Context Similarity	72.2	80.4	79.5
w/o Coherence	73.3	82.1	77.8

Table 3: F1 (%) of the evaluation set in TAC KBP 2015 Tri-lingual Entity Linking Track (Ji et al., 2015) (ENG: English, CMN: Chinese, SPA: Spanish).

We further observe that Context Similarity and Coherence features derived from $\mathcal{Z}$ play significant roles. Without such features, the performance drops significantly, as shown in Table 3. For example, in the following sentence: "欧盟委员会副主席**雷丁**就此表示... (European Commission vice president ***Redding*** said that...)", without Context Similarity feature, mention "**雷丁(*Redding*)**" is likely to be linked to the football club `en/Reading_F.C.` or the city `en/Redding,_California`. Using contextual words such as "委员会(*commission*)" and "主

席(*president*)", we can successfully link this mention to the target entity `en/Viviane_Reding`.

4.4 Parallel Sentence Mining

The proposed parallel sentence mining approach can be applied to any two languages in Wikipedia. Therefore, we have mined parallel sentences from a total number of $\binom{302}{2}$ language pairs and made this data set publicly available for research purpose. Table 4 shows some examples of mined parallel sentences from Wikipedia, with word and entity alignment highlighted.

Amharic - English
* ዓርብ የሳምንቱ ስድስተኛ ቀን ሲሆን ሐሙስ በኋላ ቅዳሜ በፊት ይገኛል ።
* Friday is the day after Thursday and the day before Saturday .
Yoruba - English
* Glasgow ni ilu totobijulo ni orile-ede Skotlandi ati eyi totobijulo keta ni Britani .
* Glasgow is the largest city in Scotland , and third largest in the United Kingdom .
Uyghur - English
* جۈمە ، پەيشەنبە بىلەن شەنبە ئوتتۇرسىدىكى ، ھەپتىنىڭ بەشىنچى كۈندۇر .
* Friday is the day after Thursday and the day before Saturday .
Vietnamese - English
* Bardolph là một làng thuộc quận McDonough , tiểu bang Illinois , HoaKỳ .
* Bardolph is a village in McDonough County , Illinois , United States .
Russian - Ukrainian
* Статья 2 - я Конституции СССР 1977 года провозглашала : « Вся власть в СССР принадлежит народу .
* Стаття 2 - га Конституції СРСР 1977 року проголошувала : " Вся влада в СРСР належить народові . (*Article 2 of the Constitution of the USSR in 1977 proclaimed: "All power in the USSR belongs to the people."*)
Classical Chinese - Modern Chinese
* 至二战之时，南斯拉夫屡败，终为德意志、义大利所分。
* 在二次世界大战期间，南斯拉夫多次战败，分别被德国、意大利占领。 (*During the World War II, Yugoslavia was defeated several times and was occupied by Germany and Italy.*)

Table 4: Examples of mined parallel sentences from Wikipedia. A portion of alignments are highlighted using the same colors.

We randomly select 100 mined parallel sentence pairs for each of 3 language pairs, and ask linguistic experts to judge the quality of these sentence pairs (perfect, partial, or not parallel). The results are shown in Table 5. We can see that the quality of mined parallel sentence is promising and the quality of word and entity alignment is decent.

Furthermore, we evaluate the quality of mined parallel sentences extrinsically using a neural machine translation (NMT) model. We use the Transformer model (Vaswani et al., 2017) implemented by Tensor2Tensor[5]. Our Transformer model has 6 encoder and decoder layers, 8 attention heads, 512-dimension hidden states, 2048-dimension feed-forward layers, dropout of 0.1 and label smoothing of 0.1. The model is trained up to 128,000 optimizer steps.

Language Pairs	**Prefect**	**Partial**	**Word**	**Entity**
Chinese-English	81%	10%	92.3%	95.5%
Spanish-English	75%	13%	89.7%	91.1%
Russian-Ukrainian	70%	16%	82.4%	90.3%

Table 5: Quality of the mined parallel sentences (Perfect and Partial stand for the percentage of perfect and partial respectively; Word and Entity stand for the Accuracy of word and entity alignments respectively).

Using the NMT model as a black box, we perform two experiments using the following training and tuning settings:

- *Baseline:* 44,000 training and 1,000 tuning sentences randomly sampled from the WMT17 News Commentary v12 Russian-English Corpus (Bojar et al., 2016).
- *Our approach:* Adding 44,000 training and 1,000 tuning sentences mined from Wikipedia using *CLEW*.

Using 1,000 randomly selected sentences from WMT 17 corpus for testing, the baseline achieves 19.0% BLEU score while our approach achieves 20.8% BLEU score.

5 Related Work

Cross-lingual Word Embedding Learning. Mikolov et al. (2013b) first notice that word embedding spaces have similar geometric arrangements across languages. They use this property to learn a linear mapping between two spaces. After that, several methods attempt to improve the mapping (Faruqui and Dyer, 2014; Xing et al., 2015; Lazaridou et al., 2015; Ammar et al., 2016; Artetxe et al., 2017; Smith et al., 2017). The measures used to compute similarity between a foreign word and an English word often include distributed monolingual representations on character-level (Costa-jussà and Fonollosa, 2016; Luong and Manning, 2016), subword-level (Anwarus Salam et al., 2012; Rei et al.,

[5] https://github.com/tensorflow/tensor2tensor

2016; Sennrich et al., 2016; Yang et al., 2017), and bi-lingual word embedding (Madhyastha and España-Bonet, 2017). Recent attempts have shown that it is possible to derive cross-lingual word embedding from unaligned corpora in an unsupervised fashion (Zhang et al., 2017; Conneau et al., 2017; Artetxe et al., 2018).

Another strategy for cross-lingual word embedding learning is to combine monolingual and cross-lingual training objectives (Zou et al., 2013; Klementiev et al., 2012; Luong et al., 2015; Ammar et al., 2016; Vulić et al., 2017). Compared to our direct mapping approach, these methods generally require large size of parallel data.

Our work is largely inspired from (Conneau et al., 2017). However, our work focuses on better representing entities, which are fundamentally different from common words or phrases in many aspects as described in Section 1. Previous multilingual word embedding efforts including (Conneau et al., 2017) do not explicitly handle entity representations. Moreover, we perform comprehensive extrinsic evaluations based on down-stream NLP applications including cross-lingual entity linking and machine translation, while previous work on cross-lingual embedding only focused on intrinsic evaluations.

Cross-lingual Joint Entity and Word Embedding Learning. Previous work on cross-lingual joint entity and word embedding methods largely neglect unlinkable entities (Tsai and Roth, 2016) and heavily rely on parallel or comparable sentences (Cao et al., 2018). Tsai and Roth (2016) apply a similar approach to generate code-switched data from Wikipedia, but their framework does not keep entities in the source language. Using all aligned entities as a dictionary, they adopt canonical correlation analysis to project two embedding spaces into one. In contrast, we only choose salient entities as anchors to learn a linear mapping. Cao et al. (2018) generate comparable data via distant supervision over multilingual knowledge bases, and use an entity regularizer and a sentence regularizer to align cross-lingual words and entities. Further, they design knowledge attention and cross-lingual attention to refine the alignment. Essentially, they train cross-lingual embedding jointly, while we align two embedding spaces that trained independently. Moreover, compared to their approach that relies on comparable data, aligned entities are easier to acquire.

Parallel Sentence Mining. Automatic mining parallel sentences from comparable documents is an important and useful task to improve Statistical Machine Translation. Early efforts mainly exploited bilingual word dictionaries for bootstrapping (Fung and Cheung, 2004). Recent approaches are mainly based on bilingual word embeddings (Marie and Fujita, 2017) and sentence embeddings (Schwenk, 2018) to detect sentence pairs or continuous parallel segments (Hangya and Fraser, 2019). To the best of our knowledge, this is the first work to incorporate joint entity and word embedding into parallel sentence mining. As a result the sentence pairs we include reliable alignment between entity mentions which are often out-of-vocabulary and ambiguous and thus receive poor alignment quality from previous methods.

6 Conclusions and Future Work

We developed a simple yet effective framework to learn cross-lingual joint entity and word embedding based on rich anchor links in Wikipedia. The learned embedding strongly enhances two down-stream applications: cross-lingual entity linking and parallel sentence mining. The results demonstrate that our proposed method advances the state-of-the-art for unsupervised cross-lingual entity linking task. We have also constructed a valuable repository of parallel sentences for all language pairs in Wikipedia to share with the community. In the future, we will extend the framework to capture better representation of other types of knowledge elements such as relations and events.

Acknowledgments

This research is based upon work supported in part by U.S. DARPA LORELEI Program HR0011-15-C-0115, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9116, and ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Khan Md. Anwarus Salam, Setsuo Yamada, and Tetsuro Nishino. 2012. Sublexical translations for low-resource language. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 39–52, Mumbai, India. The COLING 2012 Organizing Committee.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198. Association for Computational Linguistics.

Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Chengjiang Li, Xu Chen, and Tiansi Dong. 2018. Joint representation learning of cross-lingual words and entities via attentive distant supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 227–237. Association for Computational Linguistics.

Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Learning relatedness measures for entity linking. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 139–148, New York, NY, USA. ACM.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Jerome H. Friedman. 2000. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain. Association for Computational Linguistics.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

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*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *Proc. Text Analysis Conference (TAC2015)*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474. The COLING 2012 Organizing Committee.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280. Association for Computational Linguistics.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Minh-Thang Luong and Christopher Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of ACL2016*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159. Association for Computational Linguistics.

Pranava Swaroop Madhyastha and Cristina España-Bonet. 2017. Learning bilingual projections of embeddings for vocabulary expansion in machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 139–145, Vancouver, Canada. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 392–398, Vancouver, Canada. 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*.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*.

D. Milne and I.H. Witten. 2008. Learning to link with wikipedia. In *Proc. ACM international conference on Information and knowledge management (CIKM 2008)*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan. The COLING 2016 Organizing Committee.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL2016*.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2017. Neural cross-lingual entity linking. *CoRR*, abs/1712.01813.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2546–2558. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Baosong Yang, Derek F. Wong, Tong Xiao, Lidia S. Chao, and Jingbo Zhu. 2017. Towards bidirectional hierarchical representations for attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1432–1441, Copenhagen, Denmark. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398. Association for Computational Linguistics.