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
† Information Sciences Institute, University of Southern California

University of Illinois at Urbana-Champaign
{xiaoman6,hengji}@illinois.edu
{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 We make all software and resources publicly available for research purpose at \( \text{http://panx27.github.io/wikiann} \).

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 Wikipedia2, 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 inter-language links3 between 302 languages. We propose a novel cross-lingual joint entity and word (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 https://www.wikipedia.org
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.

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).](image)

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).

![figs/pca_zh.pdf](figs/pca_zh.pdf)

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

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 “computer” (computer) and “fruit” (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, “phone” (phone) and “company” (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 $Z^S$ and the target language embedding $Z^T$, and a set of pre-aligned word pairs, a linear mapping $W$ is learned to transform $Z^S$ 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 $X$ and $Y$ to denote two aligned matrices which contain the embedding of the pre-aligned words from $Z^S$ and $Z^T$ respectively. A linear mapping $W$ can be learned such that:

$$\operatorname{argmin}_W ||WX - Y||_F$$

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

$$\operatorname{argmin}_W ||WX - Y||_F = UV^T$$

Then $W$ can be obtained from the singular value decomposition (SVD) of $YX^T$ such that:

$$U\Sigma V^T = \text{SVD}(YX^T)$$

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

$$Z^T = \{z_{e_1}, \ldots, z_{e_{|E_T|}}, z_{w_1}, \ldots, z_{w_{|W_T|}}\}$$

Similarly, we use $E_S$ and $W_S$ to denote the sets of entities and words in the source language associated with the source entity and word embedding $Z^S$:

$$Z^S = \{z_{e_1}^s, \ldots, z_{e_{|E_S|}}^s, z_{w_1}^s, \ldots, z_{w_{|W_S|}}^s\}$$

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

$$Z^S = \{z_{e_1}^{s'}, \ldots, z_{e_{|E_T'|}}^{s'}, z_{e_1}, \ldots, z_{e_{|E_S|-|E_T'|}}^s, z_{w_1}^{s'}, \ldots, z_{w_{|W_S|}}^{s'}\}$$

Note that $z_{e_i}^{s'}$ and $z_{e_i}^s$ are the embedding of $e_i$ in $Z^T$ and $Z^S$ respectively. Therefore, using entities in $E_T'$ as anchors, we can learn a linear mapping $W$.\]
that maps $Z^S$ into the vector space of $Z^T$, and obtain the cross-lingual joint entity and word embedding $Z$.

We adopt the refinement procedure proposed by Conneau et al. (2017) to improve the quality of $W$. A set of new high-quality anchors is generated to refine $W$ learned from $Z^T$. 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 $W$. 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 $Wx$ and a target embedding $y$, the mean cosine similarity of $Wx$ and $y$ for their $K$ nearest neighbors in the other language, $r_T(Wx)$ and $r_S(y)$ are computed respectively. The comparison metric is defined as follows:

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

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 $Z$, the vectors of context words are averaged to obtain the context vector representation of $m$:

$$v_m = \frac{1}{|W_m|} \sum_{w \in W_m} z_w$$

where $W_m$ is the set of context words of $m$, and $z_w \in Z$ 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 $v_m$ and entity embedding $z_e \in Z$ such that:

$$F_{\text{my}}(e) = \cos(v_m, z_e) = \frac{v_m \cdot z_e}{|v_m| |z_e|}$$
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 $Z$, 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 $C_e$ 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:

$$F_{coh}(e) = \frac{1}{|C_e|} \sum_{c_e \in C_e} \cos(z_e, z_{c_e})$$

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 $Z$.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{prior}(e)$</td>
<td>Entity Prior: $\frac{</td>
</tr>
<tr>
<td>$F_{prob}(e</td>
<td>m)$</td>
</tr>
<tr>
<td>$F_{type}(e</td>
<td>m,t)$</td>
</tr>
</tbody>
</table>

Table 1: Mention disambiguation features.

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:

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

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

$$v_s = \frac{1}{|T_s|} \sum_{t \in T_s} IDF(t, S) \cdot z_t$$

where $T_s$ is the set of terms of $s$ and $z_t \in Z$ 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.

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.

We further observe that Context Similarity and Coherence features derived from \( Z \) 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

<table>
<thead>
<tr>
<th>Source-Target</th>
<th>( P@1 )</th>
<th>( P@5 )</th>
<th>( P@10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>es-en</td>
<td>79.1</td>
<td>89.2</td>
<td>92.3</td>
</tr>
<tr>
<td>it-en</td>
<td>74.5</td>
<td>86.9</td>
<td>90.5</td>
</tr>
<tr>
<td>ru-en</td>
<td>68.4</td>
<td>82.8</td>
<td>86.7</td>
</tr>
<tr>
<td>tr-en</td>
<td>59.0</td>
<td>79.9</td>
<td>86.3</td>
</tr>
<tr>
<td>uk-en</td>
<td>63.0</td>
<td>79.7</td>
<td>85.9</td>
</tr>
<tr>
<td>zh-en</td>
<td>63.1</td>
<td>83.8</td>
<td>89.2</td>
</tr>
<tr>
<td>uk-ru</td>
<td>78.1</td>
<td>90.3</td>
<td>92.8</td>
</tr>
<tr>
<td>ru-uk</td>
<td>75.8</td>
<td>90.2</td>
<td>93.7</td>
</tr>
</tbody>
</table>

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).

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

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).
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.

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

<table>
<thead>
<tr>
<th>Language Pairs</th>
<th>Prefect</th>
<th>Partial</th>
<th>Word</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese-English</td>
<td>81%</td>
<td>10%</td>
<td>92.3%</td>
<td>95.5%</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>75%</td>
<td>13%</td>
<td>89.7%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Russian-Ukrainian</td>
<td>70%</td>
<td>16%</td>
<td>82.4%</td>
<td>90.3%</td>
</tr>
</tbody>
</table>

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).

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.

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
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 downstream 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 unlinked 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.

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


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.*


