Cross-lingual Wikification Using Multilingual Embeddings

Chen-Tse Tsai and **Dan Roth**
University of Illinois at Urbana-Champaign
201 N. Goodwin, Urbana, Illinois, 61801
{ctsai12, danr}@illinois.edu

Abstract

Cross-lingual Wikification is the task of grounding mentions written in non-English documents to entries in the English Wikipedia. This task involves the problem of comparing textual clues across languages, which requires developing a notion of similarity between text snippets across languages. In this paper, we address this problem by jointly training multilingual embeddings for words and Wikipedia titles. The proposed method can be applied to all languages represented in Wikipedia, including those for which no machine translation technology is available. We create a challenging dataset in 12 languages and show that our proposed approach outperforms various baselines. Moreover, our model compares favorably with the best systems on the TAC KBP2015 Entity Linking task including those that relied on the availability of translation from the target language to English.

1 Introduction

Wikipedia has become an indispensable resource in knowledge acquisition and text understanding for both human beings and computers. The task of Wikification or Entity Linking aims at disambiguating mentions (sub-strings) in text to the corresponding titles (entries) in Wikipedia or other Knowledge Bases, such as FreeBase. For English text, this problem has been studied extensively (Bunescu and Pasca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007; Ratinov et al., 2011; Cheng and Roth, 2013). It also has been shown to be a valuable component of several natural language processing and information extraction tasks across different domains.

Recently, there has also been interest in the cross-lingual setting of Wikification: given a mention from a document written in a foreign language, the goal is to find the corresponding title in the English Wikipedia. This task is driven partly by the fact that a lot of information around the world may be written in a foreign language for which there are limited linguistic resources and, specifically, no English translation technology. Instead of translating the whole document to English, *grounding* the important entity mentions in the English Wikipedia may be a good solution that could better capture the key message of the text, especially if it can be reliably achieved with fewer resources than those needed to develop a translation system. This task is mainly driven by the Text Analysis Conference (TAC) Knowledge Base Population (KBP) Entity Linking Tracks (Ji et al., 2012; Ji et al., 2015; Ji et al., 2016), where the target languages are Spanish and Chinese.

In this paper, we develop a general technique which can be applied to all languages in Wikipedia even when no machine translation technology is available for them.

The challenges in Wikification are due both to ambiguity and variability in expressing entities and concepts: a given mention in text, e.g., Chicago, may refer to different titles in Wikipedia (Chicago Bulls, the City, Chicago Bears, the band, etc.), and a title can be expressed in the text in multiple ways, such as synonyms and nicknames. These challenges are usually resolved by calculating some similarity between the representation of the mention and candidate titles. For instance, the mention could be represented using its neighboring words, whereas a ti-

tle is usually represented by the words and entities in the document which introduces the title. In the cross-lingual setting, an additional challenge arises from the need to match words in a foreign language to an English title.

In this paper, we address this problem by using multilingual title and word embeddings. We represent words and Wikipedia titles in both the foreign language and in English in the same continuous vector space, which allows us to compute meaningful similarity between mentions in the foreign language and titles in English. We show that learning these embeddings only requires Wikipedia documents and language links between the titles across different languages, which are quite common in Wikipedia. Therefore, we can learn embeddings for all languages in Wikipedia without any additional annotation or supervision.

Another notable challenge for the cross-lingual setting that we do not address in this paper is that of *generating* English candidate titles given a foreign mention when there is no corresponding title in the foreign language Wikipedia. If a title exists in both the English and the foreign language Wikipedia, there could be examples of using this title in the foreign language Wikipedia text, and this information could help us determine the possible English titles. For example, Vladimir N. Vapnik exists in both the English Wikipedia (`en/Vladimir_Vapnik`)[1] and the Chinese Wikipedia (`zh/`弗拉基米·万普尼克). In the Chinese Wikipedia, we may see the use of the mention 萬普尼克 as a reference, that is, 萬普尼克 is linked to the title `zh/`弗拉基米·万普尼克. Following the inter-language links in Wikipedia, we can reach the English title `en/Vladimir_Vapnik`. On the other hand, Dan Roth does not have a page in the Chinese Wikipedia, it would have been harder to get to `en/Dan_Roth` from the Chinese mention. In this case, a transliteration model may be needed. Note that the difference between these two cases is only in generating English title candidates from the given foreign mention. The disambiguation method which identifies the most probable title is conceptually the same, so our method could generalize as is to this case.

[1]We use `en/Vladimir_Vapnik` to refer to the title of `en.wikipedia.org/wiki/Vladimir_Vapnik`

For evaluation purposes, we focus in this paper on mentions that have corresponding titles in both the English and the foreign language Wikipedia, and concentrate on disambiguating titles across languages. This allows us to evaluate on a large number of Wikipedia documents. Note that under this setting, a natural approach is to do wikification on the foreign language and then follow the language links to obtain the corresponding English titles. However, this approach requires developing a separate wikifier for each foreign language if it uses language-specific features, while our approach is generic and only requires using the appropriate embeddings. Importantly, the aforementioned approach will also not generalize to the cases where the target titles only exist in the English Wikipedia while ours does.

We create a challenging Wikipedia dataset for 12 foreign languages and show that the proposed approach, WikiME (**Wiki**fication using **M**ultilingual **E**mbeddings), consistently outperforms various baselines. Moreover, the results on the TAC KBP2015 Entity Linking dataset show that our approach compares favorably with the best Spanish system and the best Chinese system despite using significantly weaker resources (no need for translation). We note that the need for translation would have prevented the wikification of 12 languages used in this paper.

2 Task Definition and Model Overview

We formalize the problem as follows. We are given a document d in a foreign language, a set of mentions $M_d = \{m_1, \cdots, m_n\}$ in d, and the English Wikipedia. For each mention in the document, the goal is to retrieve the English Wikipedia title that the mention refers to. If the corresponding entity or concept does not exist in the English Wikipedia, "NIL" should be the answer.

Given a mention $m \in M_d$, the first step is to generate a set of title candidates C_m. The goal of this step is to quickly produce a short list of titles which includes the correct answer. We only look at the surface form of the mention in this step, that is, no contextual information is used.

The second and the key is the ranking step where we calculate a score for each title candidate $c \in C_m$, which indicates how relevant it is to the given men-

tion. We represent the mention using various contextual clues and compute several similarity scores between the mention and the English title candidates based on multilingual word and title embeddings. A ranking model learnt from Wikipedia documents is used to combine these similarity scores and output the final score for each title candidate. We then select the candidate with the highest score as the answer, or output NIL if there is no appropriate candidate.

The rest of paper is structured as follows. Section 3 introduces our approach of generating multilingual word and title embeddings for all languages in Wikipedia. Section 4 presents the proposed cross-lingual wikification model which is based on multilingual embeddings. Evaluations and analyses are presented in Section 5. Section 6 discusses related work. Finally, Section 7 concludes the paper.

3 Multilingual Entity and Word Embeddings

In this section, we describe how we generate a vector representation for each word and Wikipedia title in any language.

3.1 Monolingual Embeddings

The first step is to train monolingual embeddings for each language separately. We adopt the "Alignment by Wikipedia Anchors" model proposed in Wang et al. (2014). For each language, we take all documents in Wikipedia and replace the hyperlinked text with the corresponding Wikipedia title. For example, consider the following Wikipedia sentence: "It is led by and mainly composed of **Sunni** Arabs from **Iraq** and **Syria**.", where the three bold faced mentions are linked to some Wikipedia titles. We replace those mentions and the sentence becomes "It is led by and mainly composed of `en/Sunni_Islam` Arabs from `en/Iraq` and `en/Syria`." We then learn the skip-gram model (Mikolov et al., 2013a; Mikolov et al., 2013b) on this newly generated text. Since a title appears as a token in the transformed text, we will obtain an embedding for each word and title from the model.

The skip-gram model maximizes the following objective:

$$\sum_{(w,c)\in D} \log \frac{1}{1+e^{-v'_c \cdot v_w}} + \sum_{(w,c)\in D'} \log \frac{1}{1-e^{-v'_c \cdot v_w}},$$

where w is the target token (word or title), c is a context token within a window of w , v_w is the target embedding represents w, v'_c is the embedding of c in context, D is the set of training documents, and D' contains the sampled token pairs which serve as negative examples. This objective is maximized with respect to variables v_w's and v'_w's. In this model, tokens in the context are used to predict the target token. The token pairs in the training documents are positive examples, and the randomly sampled pairs are negative examples.

3.2 Multilingual Embeddings

After getting monolingual embeddings, we adopt the model proposed in Faruqui and Dyer (2014) to project the embeddings of a foreign language and English to the same space. The requirement of this model is a dictionary which maps the words in English to the words in the foreign language. Note that there is no need to have this mapping for every word. The aligned words are used to learn the projection matrices, and the matrices can later be applied to the embeddings of each word to obtain the enhanced new embeddings. Faruqui and Dyer (2014) obtain this dictionary by picking the most frequent translated word from a parallel corpus. However, there is a limited or no parallel corpus for many languages. Since our monolingual embedding model consists also of title embeddings, we can use the Wikipedia title alignments between two languages as the dictionary.

Let $A_{en} \in R^{a\times k_1}$ and $A_{fo} \in R^{a\times k_2}$ be the matrices containing the embeddings of the aligned English and foreign language titles, where a is the number of aligned titles and k_1 and k_2 are the dimensionality of English embeddings and foreign language embeddings respectively (i.e., each row is a title embedding). Canonical correlation analysis (CCA) (Hotelling, 1936) is applied to these two matrices:

$$P_{en}, P_{fo} = CCA(A_{en}, A_{fo}),$$

where $P_{en} \in R^{k_1\times d}$ and $P_{fo} \in R^{k_2\times d}$ are the projection matrices for English and foreign language

FEATURE TYPE	DESCRIPTIONS
Basic	$Pr(c\|m)$ and $Pr(m\|c)$, the fraction of times the title candidate c is the target page given the mention m, and the fraction of times c is referred by m
Other Mentions	Cosine similarity of $e(c)$ and the average of vectors in $other\text{-}mentions(m)$ The maximum and minimum cosine similarity of the vectors in $other\text{-}mentions(m)$ and $e(c)$
Local Context	Cosine similarity of $e(c)$ and $context_j(m)$, for $j = 30, 100$, and 200
Previous Titles	Cosine similarity of $e(c)$ and the average of vectors in $previous\text{-}titles(m)$ The maximum and minimum cosine similarity of the vectors in $previous\text{-}titles(m)$ and $e(c)$

Table 1: Features for measuring similarity of an English title candidate c and a mention m in the foreign language, where $e(c)$ is the English title embedding of c. $other\text{-}mentions(m)$, $previous\text{-}titles(m)$, and $context_j(m)$ are defined in Section 4.2.

embeddings, and d is the dimensionality of the projected vectors, which is a parameter in CCA.

Let $E_{en} \in R^{n_1 \times k_1}$ be the matrix containing the monolingual embeddings for all words and titles in English, where the number of words and titles is n_1, We obtain the multilingual embeddings of English words and titles by

$$E'_{en} = E_{en}P_{en} \in R^{n_1 \times d}.$$

Similarly, the multilingual embeddings of the foreign words and titles are stored in the rows of

$$E'_{fo} = E_{fo}P_{fo} \in R^{n_2 \times d},$$

where there are n_2 words and titles in the foreign language. The rows of E'_{en} and E'_{fo} are the representations of words and titles that we use to create the similarity features in the ranker.

Faruqui and Dyer (2014) show that the multilingual embeddings perform better than monolingual embeddings on various English word similarity datasets. Since synonyms in English may be translated into the same word in a foreign language, the CCA model could bring the synonyms in English closer in the embedding space. In this paper, we further show that projecting the embeddings of the two languages into the same space helps us computing better similarity between the words and titles across languages and that a bilingual dictionary consisting of pairs of Wikipedia titles is sufficient to induce these embeddings.

4 Cross-lingual Wikification

We now describe the algorithm for finding the English title given a foreign mention.

4.1 Candidate Generation

Given a mention m, the first step is to select a set of English title candidates C_m, a subset of all titles in the English Wikipedia. Ideally the correct title is included in this set. The goal is to produce a manageable number of candidates so that a more sophisticated algorithm can be applied to disambiguate them.

Since we focus on the titles in the intersection of English and the foreign language Wikipedia, we can build indices from the anchor texts in the foreign language Wikipedia. More specifically, we create two dictionaries and apply a two-step approach. The first dictionary maps each hyperlinked mention string in the text to the corresponding English titles. We simply lookup this dictionary by using the query mention m to retrieve all possible titles. The title candidates are initially sorted by $Pr(title|mention)$, the fraction of times the title is the target page of the given mention. This probability is estimated from all Wikipedia documents. The top k title candidates are then returned.

If the first high-precision dictionary fails to generate any candidate, we then lookup the second dictionary. We break each hyperlinked mention string into tokens, and create a dictionary which maps tokens to English titles. The tokens of m are used to query this dictionary. Similarly, the candidates are sorted by $Pr(title|token)$ and the top k candidates are returned.

4.2 Candidate Ranking

Given a mention m and a set of title candidates C_m, we compute a score for each title in C_m which indi-

cates how relevant the title is to m. For a candidate $c \in C_m$, we define the relevance as:

$$s(m, c) = \sum_i w_i \phi_i(m, c), \quad (1)$$

a weighted sum of the features, ϕ_i, which are based on multilingual title and word embeddings. We represent the mention m by the following contextual clues and use these representation to compute feature values:

- $context_j(m)$: use the tokens within j characters of m to compute the TF-IDF weighted average of their embeddings in the foreign language.
- $other\text{-}mentions(m)$: a set of vectors that represent other mentions. For each mention in the document other than m, we represent it by averaging the embeddings of the tokens in the mention surface string.
- $previous\text{-}titles(m)$: a set of vectors that represent previous entities. For each mention before m, we represent it by the English embedding of the disambiguated title.

Let $e(c)$ be the English embedding of the title candidate c. The features used in Eq. (1) are shown in Table 1. We train a linear ranking SVM model with the proposed features to obtain the weights, w_i, in Eq. (1). Finally, the title which has the highest relevant score is chosen as the answer to m.

5 Experiments

We evaluate the proposed method on the Wikipedia dataset of 12 langugaes and the TAC'15 Entity Linking dataset.

For all experiments, we use the Word2Vec implementation in Gensim[2] to learn the skip-gram model with dimensionality 500 for each language. The CCA code for projecting mono-lingual embeddings is from Faruqui and Dyer (2014)[3] in which the ratio parameter is set to 0.5 (i.e., the resulting multilingual embeddings have dimensionality 250).

We use Stanford Word Segmenter (Chang et al., 2008) for tokenizing Chinese, and use the Java built-in BreakIterator for Thai. For all other languages, tokenization is based on whitespaces. The number of tokens we use to learn the skip-gram model and the number of title alignments used by the CCA are given in Table 2. For learning the weights in Eq. (1), we use the implementation of linear ranking SVM in Lee and Lin (2014). Parameter selection and feature engineering are done by conducting cross-validation on the training data of Spanish Wikipedia dataset.

LANGUAGE	#TOKENS	#ALIGN. TITLES
German	616,347,668	960,624
Spanish	460,984,251	754,740
French	357,553,957	1,088,660
Italian	342,038,537	836,154
Chinese	179,637,674	469,982
Hebrew	75,076,391	137,821
Thai	68,991,911	72,072
Arabic	67,954,771	255,935
Turkish	47,712,534	162,677
Tamil	12,665,312	50,570
Tagalog	4,925,785	48,725
Urdu	3,802,679	83,665

Table 2: The number of tokens used in training the skip-gram model and the number of titles which can be aligned to the corresponding English titles via the language links in Wikipedia.

5.1 Wikipedia Dataset

We create this dataset from the documents in Wikipedia by taking the anchors (hyperlinked texts) as the query mentions and the corresponding English Wikipedia titles as the answers. Note that we only keep the mentions for which we can get the corresponding English Wikipedia titles by the language links. As observed in previous work (Ratinov et al., 2011), most of the mentions in Wikipedia documents are easy, that is, the baseline of simply choosing the title that maximizes $Pr(title|mention)$, the most frequent title given the mention surface string, performs quite well. In order to create a more challenging dataset, we randomly select mentions such that the number of easy mentions is about twice the number of hard mentions (those mentions for which the most common title is not the correct title). This generation process is inspired by (and close to) the distribution generated in the TAC KBP2015 Entity Linking Track. Another problem that occurs when creating a dataset from Wikipedia documents is that even though training documents are different from

[2] https://radimrehurek.com/gensim/

[3] https://github.com/mfaruqui/crosslingual-cca

LANGUAGE	#TRAINING	#TEST (#HARD)
German	23,124	9,798 (3,266)
Spanish	30,471	12,153 (4,051)
French	37,860	14,358 (4,786)
Italian	34,185	12,775 (4,254)
Chinese	44,246	11,394 (3,798)
Hebrew	20,223	16,146 (5,382)
Thai	16,819	11,381 (3,792)
Arabic	22,711	10,646 (3,549)
Turkish	12,942	13,798 (4,598)
Tamil	21,373	11,346 (3,776)
Tagalog	4,835	1,074 (358)
Urdu	1,413	1,389 (463)

Table 3: The number of training and test mentions of the Wikipedia dataset. The mentions are from the hyperlinked text in randomly selected Wikipedia documents. We ensure that there are at least one-third of test mentions are hard (cannot be solved by the most common title given the mention).

test documents, many mentions and titles actually overlap. To test that the algorithms really generalize from training examples, we ensure that no (mention, title) pair in the test set appear in the training set. Table 3 shows the number of training mentions, test mentions, and hard mentions in the test set of each language. This dataset is publicly available at `http://bilbo.cs.illinois.edu/~ctsai12/xlwikifier-wikidata.zip`.

The performance of the proposed method (WikiME) is shown in Table 4 along with the following approaches:

MonoEmb: In this method, we use the monolingual embeddings before applying CCA while all the other settings are the same as in WikiME. Since the monolingual embeddings are learnt separately for each language, calculating the cosine similarity of the word embedding in the foreign language and an English title embedding does not produce a good similarity function. The ranker, though, learns that the most important feature is $Pr(title|mention)$, and, consequently, performs well on easy mentions but has poor performance on hard mentions.

WordAlign: Instead of using the aligned Wikipedia titles in generating multilingual embeddings, the CCA model operates on the word alignments as originally proposed in Faruqui and Dyer (2014). We use the word alignments provided by Faruqui and Dyer (2014), which are obtained

LANGUAGE	METHOD	HARD	EASY	TOTAL
German	MonoEmb	35.18	**96.92**	76.34
	WordAlign	52.39	95.32	81.01
	WikiME	**53.28**	95.53	**81.45**
	Ceiling	90.20	100	96.73
Spanish	EsWikifier	40.11	**99.28**	79.56
	MonoEmb	38.46	96.12	76.90
	WordAlign	48.75	95.78	80.10
	WikiME	**54.46**	94.83	**81.37**
	Ceiling	93.46	100	97.69
French	MonoEmb	23.17	**97.16**	72.50
	WordAlign	41.70	96.08	77.96
	WikiME	**47.51**	95.72	**79.65**
	Ceiling	89.41	100	96.47
Italian	MonoEmb	32.68	**97.48**	75.90
	WikiME	**48.28**	95.52	**79.79**
	Ceiling	87.99	100	96.00
Chinese	MonoEmb	43.73	97.85	79.81
	WikiME	**57.61**	**98.03**	**84.55**
	Ceiling	94.29	100	98.10
Hebrew	MonoEmb	42.59	**98.16**	79.64
	WikiME	**56.67**	97.71	**84.03**
	Ceiling	96.84	100	98.95
Thai	MonoEmb	53.43	99.08	83.87
	WikiME	**70.02**	**99.17**	**89.46**
	Ceiling	94.49	100	98.16
Arabic	MonoEmb	39.81	**98.99**	79.26
	WikiME	**62.05**	98.17	**86.13**
	Ceiling	93.27	100	97.76
Turkish	MonoEmb	40.47	**98.15**	78.93
	WikiME	**60.18**	97.55	**85.10**
	Ceiling	94.08	100	98.03
Tamil	MonoEmb	34.51	98.65	77.30
	WikiME	**54.13**	**99.13**	**84.15**
	Ceiling	95.60	100	98.54
Tagalog	MonoEmb	35.47	**99.44**	78.12
	WikiME	**56.70**	98.46	**84.54**
	Ceiling	90.78	100	96.93
Urdu	MonoEmb	63.71	98.81	87.11
	WikiME	**74.51**	**99.35**	**91.07**
	Ceiling	90.06	100	96.69

Table 4: Ranking performance (Precision@1) of different approaches on various languages. Since about one-third of the test mentions are non-trivial, a baseline is 66.67 for all languages, if we pick the most common title given the mention. **Bold** signifies highest score for each column.

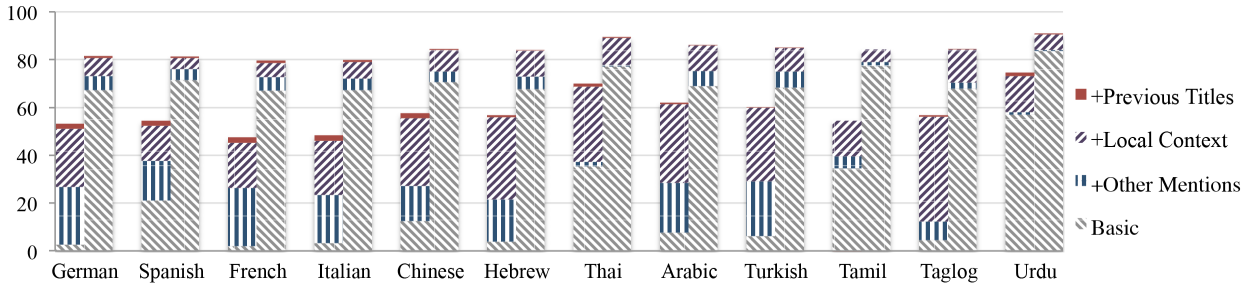

Figure 1: Feature ablation study of WikiME. The left bar of each language shows the performance on hard mentions, whereas the right bar corresponds to the performance of all mentions. The descriptions of feature types are listed in Table 1.

from the parallel news commentary corpora combined with the Europarl corpus for English to German, France, and Spanish. The number of aligned words for German, France, and Spanish are 37,484, 37,582, and 37,554 respectively. WikiME performs statistically significantly better than WordAlign on all three languages.

EsWikifier: We use Illinois Wikifier (Ratinov et al., 2011; Cheng and Roth, 2013) on a Spanish Wikipedia dump and train its ranker on the same set of documents that are used in WikiME.

Ceiling: These rows show the performance of title candidate generation. That is, the numbers indicate the percentage of mentions that have the gold title in its candidate set, therefore upper-bounds the ranking performance.

In sum, WikiME can disambiguate the hard mentions much better than other methods without sacrificing the performance on the easy mentions much. Comparing across different languages, it is important to note that languages which have a smaller size Wikipedia tend to have better performance, despite the degradation in the quality of the embeddings (see below). This is due to the difficulty of the datasets. That is, there is less ambiguity because the number of articles in the corresponding Wikipedia is small.

Figure 1 shows the feature ablation study of WikiME. For each language, we show results on hard mentions (the left bar) and all mentions (the right bar). We do not show the performance on easy mentions since it always stays high and does not change much. We can see that *Local Context* and *Other Mentions* are very effective for most of the languages. In particular, on hard mentions, the performance gain of the three feature groups is from almost 0 to around 50. For the easier dataset such as Urdu, *Basic features* alone work quite well.

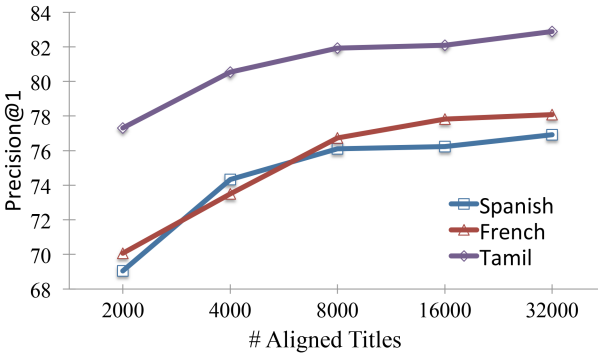

Figure 2: The number of aligned titles used in generating multilingual embeddings versus the performance of WikiME.

Figure 2 shows the performance of WikiME when we vary the number of aligned titles in generating multilingual embeddings. The performance drops a lot when there are only few aligned titles, especially for Spanish and French, where the results are even worse than MonoEmb when only 2000 titles are aligned. This indicates that the CCA method needs enough aligned pairs in order to produce good embeddings. The performance does not change much when there are more than 16,000 aligned titles.

5.2 TAC KBP2015 Entity Linking

To evaluate our system on documents outside Wikipedia, we conduct an experiment on the evaluation documents in TAC KBP2015 Tri-Lingual Entity Linking Track. In this dataset, there are 166 Chinese documents (84 news and 82 discussion forum articles) and 167 Spanish documents (84 news and 83 discussion forum articles). The mentions in this dataset are all named entities of five types: Person, Geo-political Entity, Organization, Location,

and Facility.

Table 5 shows the results. Besides the Spanish Wikifier (EsWikifier) that we used in the previous experiment, we implemented another baseline for Spanish Wikification. In this method, we use Google Translate to translate the whole documents from Spanish to English, and then the English Illinois Wikifier is applied to disambiguate the English gold mentions. Note that the target Knowledge Base of this dataset is FreeBase, therefore we use the FreeBase API to map the resulting English or Spanish Wikipedia titles to the corresponding FreeBase ID. If this conversion fails to find the corresponding FreeBase ID, "NIL" is returned instead.

The ranker models used in all three systems are trained on Wikipedia documents. We can see that WikiME outperforms both baselines significantly on Spanish. It is interesting to see that the translation-based baseline performs slightly better than the Spanish Wikifier, which indicates that the machine translation between Spanish and English is quite reliable. Note that this translation-based baseline got the highest score in this shared task when the mention boundaries were not given.

The row "Top TAC'15 System" lists the best scores of the diagnostic setting in which mention boundaries are given (Ji et al., 2016). Since the official evaluation metric considers not only the linked FreeBase IDs but also the entity types, namely, an answer is counted as correct only if the FreeBase ID and the entity type are both correct, we built two simple 5-class classifiers to classify each mention into the five entity types so that we can compare with the state of the art. One classifier uses FreeBase types of the linked FreeBase ID as features, and this classifier is only applied to mentions that are linked to some entry in FreeBase. For NIL mentions, another classifier which uses word form features (words in the mention, previous word, and next word) is applied. Both classifiers are trained on the training data of this task. From the last two rows of Table 5, we can see that WikiME achieves better results than the best TAC participants.

APPROACH	SPANISH	CHINESE
Translation + EnWikifier	79.35	N/A
EsWikifier	79.04	N/A
WikiME	**82.43**	**85.07**
+Typing		
Top TAC'15 System	80.4	83.1
WikiME	**80.93**	**83.63**

Table 5: TAC KBP2015 Entity Linking dataset. All results use gold mentions and the metric is precision@1. The top section only evaluates the linked FreeBase ID. To compare with the best systems in TAC, we also classify each mention into the five entity types. The results which evaluate both FreeBase IDs and entity types are shown in the bottom section.

6 Related Work

Wikification on English documents has been studied extensively. Earlier works (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007) focus on local features which compare context words with the content of candidate Wikipedia pages. Later, several works (Cucerzan, 2007; Milne and Witten, 2008; Han and Zhao, 2009; Ferragina and Scaiella, 2010; Ratinov et al., 2011) proposed to explore global features, trying to capture coherence among titles that appear in the text. In our method, we compute local and global features based on multilingual embeddings, which allow us to capture better similarity between words and Wikipedia titles across languages.

The annual TAC KBP Entity Linking Track has used the cross-lingual setting since 2011 (Ji et al., 2012; Ji et al., 2015; Ji et al., 2016), where the target foreign languages are Spanish and Chinese. To our best knowledge, most of the participants use one of the following two approaches: (1) Do entity linking in the foreign language, and then find the corresponding English titles from the resulting foreign language titles; and (2) Translate the query documents to English and do English entity linking. The first approach relies on a large enough Knowledge Base in the foreign language, whereas the second depends on a good machine translation system. The approach developed in this paper makes significantly simpler assumptions on the availability of such resources, and therefore can scale also to lower-resource languages, while doing very well also on high-resource languages.

Wang et al. (2015) proposed an unsupervised method which matches a knowledge graph with a graph constructed from mentions and the corre-

sponding candidates of the query document. This approach performs well on the Chinese dataset of TAC'13, but falls into the category (1). Moro et al. (2014) proposed another graph-based approach which uses Wikipedia and WordNet in multiple languages as lexical resources. However, they only focus on English Wikification.

McNamee et al. (2011) aims at the same cross-lingual Wikification setting as we do, where the challenge is in comparing foreign language words with English titles. They treat this problem as a cross-lingual information retrieval problem. That is, given the context words of the target mention in the foreign language, retrieve the most relevant English Wikipedia page. However, their approach requires parallel text to estimate word translation probabilities. In contrast, our method only needs Wikipedia documents and the inter-language links.

Besides the CCA-based multilingual word embeddings (Faruqui and Dyer, 2014) that we extend in Section 3, several other methods also try to embed words in different languages into the same space. Hermann and Blunsom (2014) use a sentence aligned corpus to learn bilingual word vectors. The intuition behind the model is that representations of aligned sentences should be similar. Unlike the CCA-based method which learns monolingual word embeddings first, this model directly learns the cross-lingual embeddings. Luong et al. (2015) propose Bilingual Skip-Gram which extends the monolingual skip-gram model and learns bilingual embeddings using a parallel copora and word alignments. The model jointly considers within language co-occurrence and meaning equivalence across languages. That is, the monolingual objective for each language is also included in their learning objective. Several recent approaches (Gouws et al., 2014; Coulmance et al., 2015; Shi et al., 2015; Soyer et al., 2015) also require a sentence aligned parallel corpus to learn multilingual embeddings. Unlike other approaches, Vulić and Moens (2015) propose a model that only requires comparable corpora in two languages to induce cross-lingual vectors. Similar to our proposed approach, this model can also be applied to all languages in Wikipedia if we treat documents across two Wikipedia languages as a comparable corpus. However, the quality and quantity of this comparable corpus for low-resource languages will be low, we believe.

We choose the CCA-based model because we can obtain multilingual word and title embeddings for all languages in Wikipedia without any additional data beyond Wikipedia. In addition, by decoupling the training of the monolingual embeddings from the cross-lingual alignment we make it easier to improve the quality of the embeddings by getting more text in the target language or a better dictionary between English and the target language. Nevertheless, as cross-lingul wikification provides another testbed for multilingual embeddings, it would be very interesting to compare these recent models on Wikipedia languages.

7 Conclusion

We propose a new, low-resource, approach to Wikification across multiple languages. Our first step is to train multilingual word and title embeddings jointly using title alignments across Wikipedia collections in different languages. We then show that using features based on these multilingual embeddings, our wikification ranking model performs very well on a newly constructed dataset in 12 languages, and achieves state of the art also on the TAC'15 Entity Linking dataset.

An immediate future direction following our work is to improve the title candidate generation process so that it can handle the case where the corresponding titles only exist in the English Wikipedia. This only requires augmenting our method with a transliteration tool and, together with the proposed disambiguation approach across languages, this will be a very useful tool for low-resource languages which have a small number of articles in Wikipedia.

Acknowledgments

This research is supported by NIH grant U54-GM114838, a grant from the Allen Institute for Artificial Intelligence (allenai.org), and Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.)

References

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Chapter of the ACL (EACL)*.

P.-C. Chang, M. Galley, and C. D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation, Association for Computational Linguistics*.

X. Cheng and D. Roth. 2013. Relational inference for wikification. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

J. Coulmance, J.-M. Marty, G. Wenzek, and A. Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of EMNLP*.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL*, pages 708–716.

M. Faruqui and C. Dyer. 2014. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics.

P. Ferragina and U. Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

S. Gouws, Y. Bengio, and G. Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. In *Deep Learning Workshop, NIPS*.

X. Han and J. Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 215–224. ACM.

K. M. Hermann and P. Blunsom. 2014. Multilingual distributed representations without word alignment. In *Proceedings of ICLR*.

H. Hotelling. 1936. Relations between two sets of variates. *Biometrika*, pages 321–377.

H. Ji, R. Grishman, and H. T. Dang. 2012. Overview of the tac2011 knowledge base population track. In *Text Analysis Conference (TAC2011)*.

H. Ji, J. Nothman, and B. Hachey. 2015. Overview of tac-kbp2014 entity discovery and linking tasks. In *Text Analysis Conference (TAC2014)*.

H. Ji, J. Nothman, B. Hachey, and R. Florian. 2016. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *Text Analysis Conference (TAC2015)*.

C.-P. Lee and C.-J. Lin. 2014. Large-scale linear RankSVM. *Neural computation*, 26(4):781–817.

T. Luong, H. Pham, and C. 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*.

P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard, and D. S. Doermann. 2011. Cross-language entity linking. In *Proceedings of IJCNLP*, pages 255–263.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 233–242.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

D. Milne and I. H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 509–518.

A. Moro, A. Raganato, and R. Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 231–244.

L. Ratinov, D. Downey, M. Anderson, and D. Roth. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

T. Shi, Z. Liu, Y. Liu, and M. Sun. 2015. Learning crosslingual word embeddings via matrix co-factorization. In *Proceedings of ACL*.

H. Soyer, P. Stenetorp, and A. Aizawa. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of ICLR*.

I. Vulić and M.-F. Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of ACL*.

Z. Wang, J. Zhang, J. Feng, and Z. Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of EMNLP*.

H. Wang, J. G. Zheng, X. Ma, P. Fox, and H. Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proceedings of EMNLP*.