A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification

Anonymous ACL submission

Abstract

We present a multilingual bag-of-entities model that effectively boosts the performance of zero-shot cross-lingual text classification by extending a multilingual pre-trained language model (e.g., M-BERT). It leverages the multilingual nature of Wikidata: entities in multiple languages representing the same concept are defined with a unique identifier. This enables entities described in multiple languages to be represented using shared embeddings. A model trained on entity features in a resource-rich language can thus be directly applied to other languages. Our experimental results on cross-lingual topic classification (using the MLDoc and TED-CLDC datasets) and entity typing (using the SHINRA2020-ML dataset) show that the proposed model consistently outperforms state-of-the-art models.

1 Introduction

In the zero-shot approach to cross-lingual transfer learning, models are trained on annotated data in a resource-rich language (the source language) and then applied to another language (the target language) without any training. Substantial progress in cross-lingual transfer learning has been made using multilingual pre-trained language models (PLMs), such as multilingual BERT (M-BERT), jointly trained on massive corpora in multiple languages (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a). However, recent empirical studies have found that cross-lingual transfer learning with PLMs does not work well for languages with insufficient pre-training data or between distant languages (Conneau et al., 2020b; Lauscher et al., 2020), which suggests the difficulty of cross-lingual transfer based solely on textual information.

We propose a multilingual bag-of-entities (M-BoE) model that boosts the performance of zero-shot cross-lingual text classification by injecting features of language-agnostic knowledge base (KB) entities into PLMs. KB entities, unlike words, can capture unambiguous semantics in documents and be effectively used to address text classification tasks (Gabrilovich and Markovitch, 2006; Chang et al., 2008; Negi and Rosner, 2013; Song et al., 2016; Yamada and Shindo, 2019). In particular, our model extends PLMs by using Wikidata entities as input features (see Figure 1). A key idea behind our model is to leverage the multilingual nature of Wikidata: entities in multiple languages representing the same concept (e.g., *Apple Inc.*, 애플, アップル) are assigned a unique identifier across languages (e.g., Q312). Given a document to be classified, our model extracts Wikipedia entities from the document, converts them into the corresponding Wikidata entities, and computes the entity-based document representation as the weighted average of the embeddings of the extracted entities. Inspired by previous work (Yamada and Shindo, 2019; Peters et al., 2019), we compute the weights using an attention mechanism that selects the entities relevant to the given document. We then compute the sum of the entity-based document representation and the text-based document representation computed using the PLM and feed it into a linear classifier. Since the entity vocabulary and entity embedding are shared across languages, a model trained on entity features in the source language can be directly transferred to multiple target languages.

We evaluated the performance of the M-BoE model on three cross-lingual text classification tasks: topic classification on the MLDoc (Schwenk and Li, 2018) and TED-CLDC (Hermann and Blunsom, 2014) datasets and entity typing on the SHINRA2020-ML (Sekine et al., 2020) dataset. We trained the model using training data in the source language (English) and then evaluated it on the target languages. It outperformed our base PLMs (i.e., M-BERT (Devlin et al., 2019) and the XLM-R model (Conneau et al., 2020a)) for all tar-

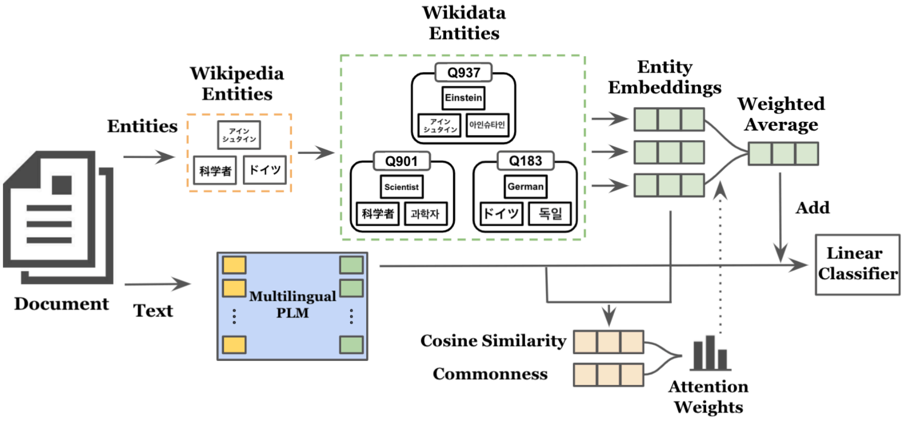


Figure 1: Architecture of M-BoE. Given a document, the model extracts Wikipedia entities, converts them into corresponding Wikidata entities, and calculates the entity-based document representation by using the weighted average of the embeddings of the entities selected by an attention mechanism. The sum of the entity-based representation and the representation computed using a multilingual PLM is used to perform linear classification for the task.

get languages on all three tasks, thereby demonstrating the effectiveness of the entity-based representation. Furthermore, our model performed better than state-of-the-art models on the MLDoc dataset.

Our contributions are as follows:

- We present a method for boosting the performance of cross-lingual text classification by extending multilingual PLMs to leverage the multilingual nature of Wikidata entities. Our method successfully improves the performance on multiple target languages simultaneously without expensive pre-training or additional text data in the target languages.
- Inspired by previous work (Yamada and Shindo, 2019; Peters et al., 2019), we introduce an attention mechanism that enables entity-based representations to be effectively transferred from the source language to the target languages. The mechanism selects entities that are relevant to address the task.
- We present experimental results for three cross-lingual text classification tasks demonstrating that our method outperformed our base PLMs (i.e., M-BERT and XLM-R) for all languages on the three tasks and outperformed state-of-the-art methods on the MLDoc dataset.

2 Related Work

Cross-lingual PLMs Zero-shot cross-lingual transfer learning approaches have relied on parallel corpora (Xu and Wan, 2017) or multilingual word representation (Duong et al., 2017). Considerable progress has been made on PLMs for various cross-lingual transfer tasks. The representative models are M-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), which are multilingual extensions of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), respectively. Both models are pre-trained on massive corpora of approximately 100 languages. LASER (Artetxe and Schwenk, 2019) is a PLM trained on a parallel corpus of 93 languages by using a sequence-to-sequence architecture.

Improving cross-lingual transfer learning Several studies have attempted to improve cross-lingual transfer learning by using additional text data in the target language. Lai et al. (2019) proposed using an unlabeled corpus in the target language to bridge the gap between the language and the domain. Dong et al. (2020) and Keung et al. (2019) incorporated adversarial training using unlabeled target language examples. Dong and de Melo (2019) and Eisenschlos et al. (2019) presented methods for data augmentation in which pseudo-labels are assigned to an unlabeled corpus in the target language. Conneau and Lample (2019) additionally pre-trained BERT-based models using a parallel corpus. However, these methods require extra training on additional text data for each target language, and their resulting models work well only on a single target language. Unlike these methods, our method does not require extra training and improves performance simultaneously for all target languages with only a single PLM. Furthermore, our method can be easily applied to these models since it is a simple extension of a PLM and does not modify its internal architecture.

Enhancing monolingual PLMs using entities Several methods have been proposed for improving the performance of PLMs through pre-training using entities. ERNIE (Zhang et al., 2019) and KnowBert (Peters et al., 2019) enrich PLMs by using pre-trained entity embeddings. LUKE (Yamada et al., 2020b) and EaE (Févry et al., 2020) train entity embeddings from scratch during pre-training. However, all of these methods are aimed at improving the performance of monolingual tasks and require pre-training with a large corpus, which is computationally expensive. Our method dynamically injects entity information into PLMs during fine-tuning without expensive pre-training.

Several studies have attempted to incorporate entity information into PLMs after pre-training to enhance the performance of monolingual tasks. Ostendorff et al. (2019) concatenated contextualized representations with knowledge graph embeddings to represent author entities and used them as features for the book classification task. E-BERT (Poerner et al., 2020) inserts KB entities next to the entity names in the input sequence to improve BERT's performance for entity-centric tasks. Verlinden et al. (2021) introduced a mechanism for combining span representations and KB entity representations within a BiLSTM-based end-to-end information extraction model. Unlike these methods, our method aims to improve the cross-lingual text classification by combining PLMs with language-agnostic entity embeddings.

Text classification models using entities Several methods have been commonly used to address text classification using entities. Explicit semantic analysis (ESA) is a representative example; it represents a document as a bag of entities, which is a sparse vector in which each dimension is a score reflecting the relevance of the text to each entity (Gabrilovich and Markovitch, 2006; Chang et al., 2008; Negi and Rosner, 2013). More recently, Song et al. (2016) proposed cross-lingual explicit semantic analysis (CLESA), an extension of ESA, to address cross-lingual text classification. CLESA computes sparse vectors from the intersection of Wikipedia entities in the source and target languages using Wikipedia language links. Unlike CLESA's approach, we address cross-lingual text classification by extending state-of-the-art PLMs with a language-agnostic entity-based document representation based on Wikidata.

The most relevant to our proposed approach is the neural attentive bag-of-entities (NABoE) model proposed by Yamada and Shindo (2019). It addresses monolingual text classification using entities as inputs and uses an attention mechanism to detect relevant entities in the input document. Our model can be regarded as an extension of NABoE by (1) representing documents using a shared entity embedding across languages and (2) combining an entity-based representation and attention mechanism with state-of-the-art PLMs.

3 Proposed Method

Figure 1 shows the architecture of our model. The model extracts Wikipedia entities, converts them into Wikidata entities, and computes the entity-based document representation using an attention mechanism. The sum of the entity-based document representation and the text-based document representation computed using the PLM is fed into a linear classifier to perform classification tasks.

3.1 Entity detection

To detect entities in the input document, we use two dictionaries that can be easily constructed from the KB: (1) a mention-entity dictionary, which binds an entity name (e.g., "Apple") to possible referent KB entities (e.g., *Apple Inc.* and *Apple (food)*) by using the internal anchor links in Wikipedia (Guo et al., 2013), and (2) an inter-language entity dictionary, which links multilingual entities (e.g., *Tokyo*, 도쿄, 東京) to a corresponding identifier (e.g., Q7473516) of Wikidata.

All words and phrases are extracted from the given document in accordance with the mention-entity dictionary, and all possible referent entities are detected if they are included as entity names in the dictionary. Note that all possible referent entities are detected for each entity name rather than a single resolved entity. For example, we detect both *Apple Inc.* and *Apple (food)* for entity name

"Apple". Next, the detected entities are converted into Wikidata entities if they are included in the inter-language entity dictionary.

3.2 Model

Each Wikidata entity is assigned a representation $\boldsymbol{v}_{e_i} \in \mathbb{R}^d$. Since our method extracts all possible referent entities rather than a single resolved entity, it often extracts entities that are not related to the document. Therefore, we introduce an attention mechanism inspired by previous work (Yamada and Shindo, 2019; Peters et al., 2019) to prioritize entities related to the document. Given a document with K detected entities, our method computes the entity-based document representation $\boldsymbol{z} \in \mathbb{R}^d$ as the weighted average of the entity embeddings:

$$\boldsymbol{z} = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v}_{e_i}, \tag{1}$$

where $a_{e_i} \in \mathbb{R}$ is the attention weight corresponding to entity e_i and calculated using

$$\boldsymbol{a} = \mathrm{softmax}(\boldsymbol{W}_a^\top \boldsymbol{\phi}), \tag{2}$$

$$\phi(e_i, d) = \begin{bmatrix} \mathrm{cosine}(\boldsymbol{h}, \boldsymbol{v}_{e_i}) \\ p_{e_i} \end{bmatrix} \tag{3}$$

where $\boldsymbol{a} = [a_{e_1}, a_{e_2}, \cdots, a_{e_K}]$ are the attention weights; $\boldsymbol{W}_a \in \mathbb{R}^2$ is a weight vector; $\boldsymbol{\phi} = [\phi(e_1, d), \phi(e_2, d), \cdots, \phi(e_K, d)] \in \mathbb{R}^{2\times K}$ represents the degree to which each entity e_i is related to document d; and $\phi(e_i, d)$ is calculated by concatenating commonness[1] p_{e_i} with the cosine similarity between the document representation computed using the PLM, $\boldsymbol{h} \in \mathbb{R}^d$ (e.g., the final hidden state of the [CLS] token), and entity embedding, $\boldsymbol{v}_{e_i}$.

The sum of this entity-based document representation $\boldsymbol{z}$ and text-based document representation $\boldsymbol{h}$ is fed into a linear classifier[2] to predict the probability of label c:

$$p(c \mid \boldsymbol{h}, \boldsymbol{z}) = Classifier(\boldsymbol{h} + \boldsymbol{z}). \tag{4}$$

[1]Commonness (Mihalcea and Csomai, 2007) is the probability that an entity name refers to an entity in Wikipedia.

[2]In preliminary experiments, we also tested concatenation, but observed worse overall results than with summation.

Dataset	Language	Train	Dev.	Test
MLDoc	8	1,000	1,000	4,000
TED-CLDC	12	936	105	51–106
SHINRA	30	417,387	21,967	30k–920k

Table 1: Number of examples in MLDoc, TED-CLDC, and SHINRA2020-ML datasets.

4 Experimental Setup

In this section, we describe the experimental setup we used for the three cross-lingual text classification tasks.

4.1 Data

We evaluated our model using three datasets: MLDoc (Schwenk and Li, 2018), TED-CLDC (Hermann and Blunsom, 2014), and SHINRA2020-ML (Sekine et al., 2020).

MLDoc is a dataset for multi-class text classification, i.e., classifying news articles into four categories. We used the english.train.1000 and english.dev datasets, which contain 1000 documents for training and validation data. As in the previous work (Schwenk and Li, 2018; Keung et al., 2020), we used accuracy as the metric.

TED-CLDC is a multi-label classification dataset covering 15 topics. This topic classification dataset is exactly like the MLDoc dataset except that the classification task is more difficult because of its colloquial nature and because the amount of training data is small. Following the previous work (Hermann and Blunsom, 2014), we used micro-average F1 as the metric.

SHINRA2020-ML is an entity typing dataset that assigns fine-grained entity labels (e.g., Person, Country, Government) to a Wikipedia page. We used this dataset for multi-label classification tasks; we used all datasets in 30 languages except English for the test data. Following the original work (Sekine et al., 2020), we used micro-average F1 as the metric.

We created a validation set by randomly selecting 5% of the training data in TED-CLDC and 5% of the training data in SHINRA2020-ML. We used English as the source language in all experiments. A summary of the datasets is shown in Table 1.

Model	MLDoc	TED-CLDC	SHINRA2020-ML
M-BERT	32 / 2e-05	16 / 2e-05	128 / 5e-05
XLM-R	32 / 2e-05	16 / 5e-05	64 / 2e-05
M-BoE (M-BERT)	32 / 2e-05	16 / 2e-05	128 / 5e-05
M-BoE (XLM-R)	32 / 2e-05	16 / 5e-05	64 / 2e-05

Table 2: Hyper-parameters used for experiments. In each cell, the left value indicates batch size, and the right value indicates learning rate.

Model	en	fr	de	ja	zh	it	ru	es	target avg.
MultiCCA (Schwenk and Li, 2018)	92.2	72.4	81.2	67.6	74.7	69.4	60.8	72.5	71.2
LASER (Artetxe and Schwenk, 2019)	89.9	78.0	84.8	60.3	71.9	69.4	67.8	77.3	72.8
M-BERT	94.0	79.4	75.1	69.3	68.0	67.1	65.3	75.2	71.4 ± 1.4
XLM-R	94.4	84.9	86.7	78.5	85.2	73.4	71.3	81.5	80.2 ± 0.5
M-BoE (M-BERT)	94.1	84.0	76.9	71.1	72.2	70.0	68.9	75.5	74.1 ± 0.7
M-BoE (XLM-R)	**94.6**	**86.4**	**88.9**	**80.0**	**87.4**	**75.6**	**73.7**	**83.2**	$\mathbf{82.2 \pm 0.6}$

Table 3: Classification accuracy for topic classification on MLDoc dataset; "target avg." indicates average scores for target languages.

4.2 Entity preprocessing

We constructed a mention-entity dictionary from the January 2019 version of Wikipedia dump[3] and an inter-language entity dictionary from the March 2020 version in the Wikidata dump,[4] which contains 45,412,720 Wikidata entities (e.g., Q312). We computed the commonness values from the same versions of Wikipedia dumps in the corresponding language, following the work of Yamada and Shindo (2019).

We initialized Wikidata entity embeddings using pre-trained English entity embeddings trained on the KB. To train these embeddings, we used the open-source Wikipedia2Vec tool (Yamada et al., 2020a). We used the January 2019 English Wikipedia dump mentioned above and set the dimension to 768 and the other parameters to the default values. We initialized an entity embedding using a random vector if the entity did not exist in the Wikipedia2Vec embeddings. Note that we used only English Wikipedia to train the entity embeddings.

4.3 Models

We used M-BERT (Devlin et al., 2019) and XLM-R$_{base}$ (Conneau et al., 2020a) as the baseline multilingual PLMs to evaluate the proposed method. We added a single fully-connected layer on top of the PLMs and used the final hidden state $\boldsymbol{h}$ of the first [CLS] token as the text-based document representation. For the MLDoc dataset, we trained the model by minimizing the cross-entropy loss with softmax activation. For the TED-CLDC and SHINRA2020-ML datasets, we trained the model by minimizing the binary cross-entropy loss with sigmoid activation. For these two tasks, we regarded each label as positive if its corresponding predicted probability was greater than 0.5 during inference.

For topic classification using MLDoc, we compared the performance of the proposed model with those of two state-of-the-art cross-lingual models: LASER (Artetxe and Schwenk, 2019) (see Section 2), and MultiCCA (Schwenk and Li, 2018), which is based on a convolutional neural network with multilingual word embeddings. To ensure a fair comparison, we did not include models that use additional unlabeled text data or a parallel corpus to train models for each target language.

For entity typing, we tested a model that uses oracle entity annotations (i.e., hyperlinks) contained in the Wikipedia page to be classified instead of entities detected using the entity detection method described in Section 3.1. Note that this model also uses attention mechanisms and pre-trained entity embeddings.

4.4 Detailed settings

The hyper-parameters used in our experiments are shown in Table 2. We tuned them on the basis of the English validation set. We trained the model using the AdamW optimizer with a gradient clipping of 1.0.

In all experiments, we trained the models until the performance on the English validation set con-

[3] https://dumps.wikimedia.org/
[4] https://dumps.wikimedia.org/wikidatawiki/entities/

Model	en	fr	de	it	ru	es	ar	tr	nl	pt	pl	ro	target avg.
M-BERT	51.6	47.7	43.9	50.6	47.9	53.1	41.3	44.2	49.4	46.2	45.1	45.4	47.1 ± 1.4
XLM-R	51.5	49.5	49.7	48.7	48.3	51.2	45.6	51.3	48.8	46.3	48.3	48.4	49.1 ± 1.8
M-BoE (M-BERT)	**52.9**	49.5	46.2	**53.3**	49.2	**54.7**	44.7	49.1	51.0	47.6	47.7	48.2	49.6 ± 1.1
M-BoE (XLM-R)	51.7	**50.0**	**53.8**	51.3	**52.3**	52.9	**50.5**	**53.1**	**52.0**	**49.3**	**50.5**	**49.6**	**51.8 ± 0.9**

Table 4: F1 score for topic classification on TED-CLDC dataset.

	fr	de	ja	zh	it	ru	es	ar	tr	nl	pt	pl	ro	hi	no
M-BERT	68.5	84.2	81.3	80.7	85.2	81.4	85.6	57.4	50.7	55.6	80.4	77.7	76.9	81.8	83.6
XLM-R	73.0	82.6	77.4	75.1	84.2	81.0	85.3	58.9	69.1	63.7	79.8	80.0	76.9	83.3	82.4
M-BoE (M-BERT)	69.3	**85.1**	**82.5**	**82.2**	86.4	83.2	**86.6**	61.9	54.0	59.0	81.7	79.4	**80.5**	82.9	84.8
M-BoE (XLM-R)	**77.4**	84.5	79.0	77.0	85.6	83.2	85.8	**63.3**	**72.3**	65.5	80.7	**81.8**	77.8	84.8	84.0
Oracle M-BoE (M-BERT)	75.4	85.2	81.9	81.8	86.5	83.0	86.5	61.9	53.7	61.7	**81.8**	79.7	79.9	83.0	**84.8**
Oracle M-BoE (XLM-R)	76.5	84.8	79.6	77.2	85.5	**83.4**	86.2	63.0	71.8	**67.6**	80.4	81.5	78.8	**84.8**	83.2
	th	ca	da	fa	id	sv	vi	bg	cs	fi	he	hu	ko	uk	target avg.
M-BERT	84.0	81.5	80.1	80.2	72.4	79.4	79.3	74.0	74.6	75.7	74.0	77.1	81.3	78.0	76.6 ± 0.7
XLM-R	81.4	79.0	81.0	82.4	75.5	75.5	80.7	76.0	77.9	74.7	70.5	73.1	82.6	74.3	77.1 ± 1.2
M-BoE (M-BERT)	85.1	83.2	81.4	82.1	75.4	**82.4**	81.2	76.1	76.8	**77.6**	**78.1**	**79.2**	82.9	**80.0**	78.7 ± 0.5
M-BoE (XLM-R)	82.1	80.9	**83.3**	**84.1**	78.2	78.7	81.9	79.1	79.6	76.9	71.9	75.5	**84.0**	77.0	**79.2 ± 0.9**
Oracle M-BoE (M-BERT)	**85.3**	**83.2**	82.3	82.4	75.5	82.0	81.6	76.6	77.4	77.4	77.8	78.7	83.3	79.9	79.0 ± 0.5
Oracle M-BoE (XLM-R)	81.8	81.2	82.9	83.9	**78.3**	78.2	**82.5**	**79.1**	**79.9**	77.1	71.8	75.8	83.92	76.9	**79.2 ± 0.9**

Table 5: F1 score for entity typing on SHINRA2020-ML dataset.

Setting	M-BoE (M-BERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9

Table 6: Results of analysis of our model on MLDoc.

verged. We conducted all experiments ten times with different random seeds, and recorded the average scores and 95% confidence intervals.

5 Results

Tables 3, 4, and 5 show the results of our experiments. Overall, the M-BoE models outperformed their baselines (i.e., M-BERT and XLM-R) for all target languages on all three datasets. Furthermore, there was a significant difference in the mean scores for the target languages for those models in a paired t-test ($p < 0.05$). In particular, the performance of our model clearly exceeded that of the M-BERT baseline by 2.7% in accuracy, 2.5% in F1, and 2.1% in F1, on the MLDoc, TED-CLDC, and SHINRA2020-ML datasets, respectively.

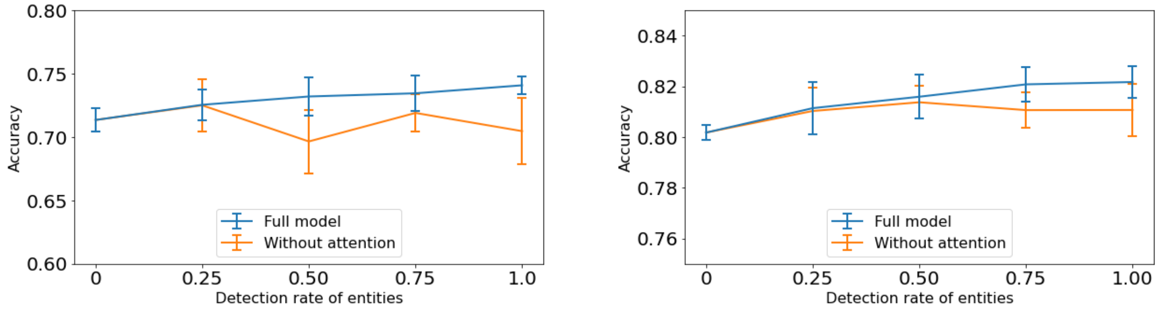
For entity typing, using the entities detected with our simple dictionary-based approach achieved comparable performance to using gold entity annotations (Table 5: Oracle M-BoE) on the SHINRA2020-ML dataset, which clearly demonstrates the effectiveness of our attention-based entity detection method.

6 Analysis

We conducted a series of experiments to analyze the performance of our model on the MLDoc dataset (Table 6). We first analyzed the impact on the performance of each component in the M-BoE model, including the attention mechanism, pre-trained entity embeddings, and entity detection methods. We then evaluated the sensitivity of the model's performance to differences in the number of detected entities for each language. Finally, we conducted qualitative analysis by visualizing important entities.

6.1 Attention mechanism

We examined the effect of the attention mechanism on performance. When the attention mechanism was removed (Table 6: **Attention mechanism**), the performance was substantially lower than with the proposed model. This indicates that the attention mechanism selects the entities that are effective in solving the classification task. Next, we examined the effectiveness of the two features (i.e., cosine and commonness) in the attention mechanism by excluding them one at a time from the M-BoE



(a) M-BoE (M-BERT)

(b) M-BoE (XLM-R)

Figure 2: Classification accuracy for each entity detection rate using MLDoc dataset.

Model	en (train)	fr	de	ja	zh	it	ru	es	avg.
External entity linking	20.0	19.2	14.6	8.15	5.2	11.7	12.7	13.8	13.2
Dictionary-based method (ours)	105.8	97.8	78.9	47.9	34.5	53.2	64.6	72.3	64.2

Table 7: Comparison of the number of detected entities on MLDoc dataset. Numbers indicate average number of entities detected for each example.

model. Table 6 shows that there was a slight drop in performance when either of them was not used, indicating that both features are effective.

6.2 Entity embeddings

To investigate the effect of entity embedding initialization, we replaced Wikipedia2Vec with (1) random vectors and (2) knowledge graph (KG) embeddings (Table 6: **Entity embeddings**). For KG embedding, we used ComplEx (Trouillon et al., 2016), a state-of-the-art KG embedding method. We trained the ComplEx embeddings on the wikidata5m dataset (Wang et al., 2021) using the kge tool (`https://github.com/uma-pi1/kge`). We set the dimension to 768 and used the default hyper-parameters for everything else in the `wikidata5m-complex configuration` in the tool. The results show that using Wikipedia2Vec was the most effective although using KG embeddings was better than using random vectors.

6.3 Entity detection method

To verify the effectiveness of our dictionary-based entity detection method, we simply replaced it with a commercial multilingual entity linking system, Google Cloud Natural Language API[5] (Table 6: **Entity detection method**). All entities were detected with the API and converted into Wikidata entities, as explained in Section 3.1. Note that unlike our dictionary-based method, the entity linking system detects a single disambiguated entity for each entity name.

The results show that our entity detection method outperformed the API. We attribute this to the number of entities detected with our dictionary-based detection method. As shown in Table 7, the number of entities detected with the entity linking system was substantially lower than with our entity detection method because, unlike our method, the system detects only disambiguated entities and does not detect non-named entities. Therefore, we attribute the better performance of our method compared with that of the API to (1) non-named entities also being important features and (2) the inability to use the correct entity if the disambiguation error is caused by entity linking.

Furthermore, as described in Section 5, our entity detection method performed competitively with the human-labeled entity annotations on the SHINRA2020-ML dataset.

Next, we examined the performance impact of the number of detected Wikidata entities. For the full model and no attention model, we observed a change in performance when some percentage of the entities were randomly removed during training and inference. Figure 2 shows that, the higher the entity detection rate, the better the performance of the full model. When the attention mechanism was removed, however, there was no consistent

[5] `https://cloud.google.com/natural-language`

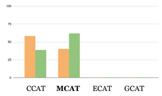
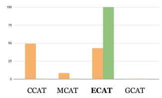
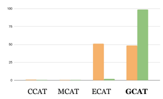
Language	Document	Label	Probability distribution (M-BERT, M-BoE)	Top three entities
Ja	［台北 2日 ロイター］引け前の台湾株式市場で、加権指数が3.28%急落した。フローカーらによると、工業株に売りが集中したため、という。大引け前10分（0350gmt）現在、加権指数は278．07ポイント（3．28%）急落し、8207．59。売買代金は、1090億台湾ドル。	**MCAT** **(Markets)**		**"Stock certificate"** **"Share price"** **"Taiwan Capitalization Weighted Stock Index"**
Zh	〔路透社東京19日電〕日本大藏省一顧問小組週四促請大藏省取消目前只允許被授權外匯銀行進行外匯交易的管制,完全開放外匯市場交易資格的限制. 這項限制的取消將使投資人進出外匯市場更為容易；此外,銀行業也可藉此增進競爭力,並促進市場的流動性及活絡匯市的交易.（完）	**ECAT** **(Economics)**		**"Ministry of the Treasury"** **"Financial transaction"** **"Competition (economics)"**
Ru	москва, 17 мар (рейтер) - президент рф борис ельцин подписал федеральные законы о внесении изменении и дополнении в статьи 100 и 110 закона рф "о государственных пенсиях в рф", сообщила пресс-служба президента рф. статья 100 закона излагается в следующеи редакции: "в заработок для исчисления пенсии включаются все виды выплат (дохода), полученных в связи с выполнением работы, предусмотреннои статьеи 89 закона, на которые начисляются страховые взносы в пенсионный фонд рф". пресс-служба президента рф сообщила, что виды выплат, на которые не начисляются страховые взносы в пенсионный фонд рф, определяются правительством рф.	**GCAT** **(Government Social)**		**"Federal law"** **"Pension Fund of the Russian Federation"** **"Kremlin Press Secretary"**

Figure 3: Example results for MLDoc. "Top three entities" indicates the three most influential entities selected by attention mechanism.

trend. The performance remained the same or even dropped. These results suggest that the more entities detected, the better the performance, and that the attention mechanism is important for this consistent improvement.

6.4 Performance sensitivity to language differences

In our method, the number of detected Wikidata entities during inference differs depending on the target languages. We investigated how this affects performance. For each of the datasets, we computed the Pearson's correlation coefficient between the number of detected entities and the rate of improvement in performance for each language (see Table 8 in the Appendix). As a result, there was no clear trend in the correlation coefficients, which ranged from -0.3 to 0.2. These results indicate that the performance was consistently improved for languages with a small number of detected entities. We attribute this to the ability of our method to detect a sufficient number of entities, even for languages with a relatively small number of entity detections.

6.5 Qualitative analysis

To further investigate how the M-BoE model improved performance, we took the MLDoc documents that our model classified correctly while M-BERT did not and examined the influential entities that were assigned the largest attention weights by the M-BoE model. Figure 3 shows three examples in which the M-BoE model effectively improved performance. Overall, it identified the entities that were highly relevant to the document. For example, the first document is a Japanese document about the Taiwanese stock market, and the M-BoE model correctly identified the relevant entities, including ***Stock certificate***, ***Share price***, and ***Taiwan Capitalization Weighted Stock Index***.

7 Conclusions

Our proposed M-BoE model is a simple extension of multilingual PLMs: language-independent Wikidata entities are used as input features for zero-shot cross-lingual text classification. Since the Wikidata entity embeddings are shared across languages, and the entities associated with a document are further selected by the attention mechanism, a model trained on these features in one language can efficiently be applied to multiple target languages. We achieved state-of-the-art results on three cross-lingual text classification tasks, which clearly shows the effectiveness of our method.

As future work, we plan to evaluate our model on a variety of natural language processing tasks, such as cross-lingual document retrieval. We would also like to investigate whether our method can be combined with other methods, such as using additional textual data in the target language.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grants JP16H06302, JP18H04120, JP20K23355, JP21H04907, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, Japan.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6306–6310.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1541–1544.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5702–5707.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4951.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, page 1301–1306.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1355–1360.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554.

Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. *arXiv preprint arXiv:1909.07009*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4483–4499.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242.

Sapna Negi and Michael Rosner. 2013. UoM: Using explicit semantic analysis for classifying sentiments. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 535–538.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, and Georg Rehm. 2019. Enriching BERT with Knowledge Graph Embedding for Document Classification. In *Proceedings of the GermEval 2019 Workshop*, Erlangen, Germany.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 43–54.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. 2020. Overview of shinra2020-ml task. In *Proceedings of the NTCIR-15 Conference*.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2901–2907.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1952–1957.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020a. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020b. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6442–6454.

Ikuya Yamada and Hiroyuki Shindo. 2019. Neural attentive bag-of-entities model for text classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 563–573.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Dataset		Model	fr	de	it	ru	es	ja	zh	ar	tr	nl	pt	pl	ro	Pearson
MLDoc	#Ent		97.8	78.9	53.2	64.6	72.3	47.9	34.5	-	-	-	-	-	-	
	Rate	M-BERT	5.8	2.4	4.3	5.5	0.4	2.6	6.2	-	-	-	-	-	-	-0.13
		XLM-R	1.8	2.5	3.0	3.4	2.1	1.9	2.6	-	-	-	-	-	-	-0.34
TED-CLDC	#Ent		218.9	223.5	217.8	227.2	227.9	-	-	227.3	185.0	190.7	166.4	134.5	211.2	
	Rate	M-BERT	3.8	5.2	5.7	2.7	3.0	-	-	8.2	11.1	3.2	3.0	5.8	6.2	-0.11
		XLM-R	1.0	8.2	5.3	8.3	3.3	-	-	10.7	3.5	6.6	6.5	4.6	2.5	0.17

Table 8: Pearson correlation coefficient between average number of detected entities (#Ent) and rate of improvement in performance (Rate) for each target language.

Appendix for "A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification"

A Details of performance sensitivity to language differences

As described in Section 6.4, we tested the sensitivity of performance to the number of entities detected in the target languages. Specifically, for each target language, we computed (1) the ratio of performance improvement to the baseline and (2) the average number of detected entities per document and computed the Pearson correlation coefficient between the two variables on the MLDoc and TED-CLDC datasets.

The experimental results (Table 8) do not show any clear trend in the correlation coefficients, indicating that the number of entity detections during inference does not substantially affect the model's performance. For example, even for Chinese on the MLDoc dataset, for which the number of entity detections was the lowest, the performance was consistently higher than that of the baseline, as it was for the other languages.