Learning Multilingual Meta-Embeddings for Code-Switching Named Entity Recognition

Genta Indra Winata, Zhaojiang Lin, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{giwinata,zlinao}@connect.ust.hk, pascale@ece.ust.hk

Abstract

In this paper, we propose Multilingual Meta-Embeddings (MME), an effective method to learn multilingual representations by leveraging monolingual pre-trained embeddings. MME learns to utilize information from these embeddings via a self-attention mechanism without explicit language identification. We evaluate the proposed embedding method on the code-switching English-Spanish Named Entity Recognition dataset in a multilingual and cross-lingual setting. The experimental results show that our proposed method achieves state-of-the-art performance on the multilingual setting, and it has the ability to generalize to an unseen language task.

1 Introduction

Learning a representation through embedding is a fundamental technique to capture latent word semantics (Clark, 2015). Practically, word-level representation has been extensively explored to improve many downstream natural language processing (NLP) tasks (Mikolov et al., 2013; Pennington et al., 2014; Grave et al., 2018). A new wave of "meta-embeddings" research aims to learn how to effectively combine pre-trained word embeddings in supervised training into a single dense representation (Yin and Schütze, 2016; Muromägi et al., 2017; Bollegala et al., 2018; Coates and Bollegala, 2018; Kiela et al., 2018). This method is known to be effective to overcome domain and modality limitations. However, the generalization ability of previous works has been limited to monolingual tasks, so we aim to extend the method to multilingual contexts which benefits the processing of code-switching text.

In multilingual societies, speakers tend to move back and forth from one language to another during the same conversation, which is commonly called "*code-switching*". Code-Switching is produced in both written text and speech in a discourse. Recent studies in code-switching has been mainly focused on natural language tasks, such as language modeling (Winata et al., 2018a; Pratapa et al., 2018; Garg et al., 2018), named entity recognition (Aguilar et al., 2018), and language identification (Solorio et al., 2014; Molina et al., 2016; Barman et al., 2014). Code-Switching is considered as a challenging task because words from different languages may co-exist within a sequence, and models are required to recognize the context of mixed-language sentences. Meanwhile, some words with the same spelling may have entirely different meanings (e.g., cola in English and Spanish) (Winata et al., 2018b). Language identifiers were commonly used to solve the word ambiguity issue in mixed-language sentences. However, it may not reliably cover all code-switching cases, and it creates a bottleneck that would require large-scale crowdsourcing to annotate language identifiers in code-switching data correctly.

To overcome the code-switching problem, we introduce a multilingual meta-embedding model learned from different languages. Our approach can be seen as a method to create a universal mul-

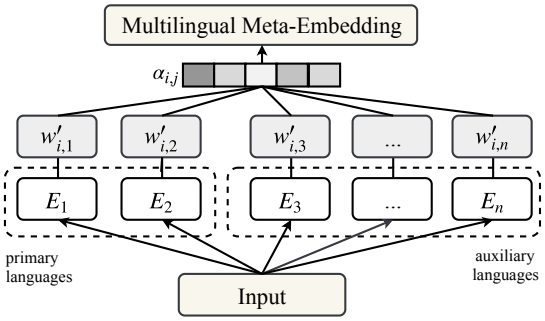

Figure 1: Multilingual Meta-Embeddings. The inputs are word embeddings and the output is a single word representation.

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 181–186
Florence, Italy, August 2, 2019.

tilingual meta-embedding learned in a supervised way with code-switching contexts by gathering information from monolingual sources. Concurrently, this is a language-agnostic approach where it does not require any language information of each word. We show the possibility of transferring information from multiple languages to unseen languages, and this approach can also be useful for a low-resource setting. To effectively leverage the embeddings, we use FastText subwords information to solve out-of-vocabulary (OOV) issues. By applying this method, our model can align the words with the corresponding languages. Our contributions are two-fold:

- We propose to generate multilingual meta-representations from pre-trained monolingual word embeddings. The model can learn how to construct the best word representation by mixing multiple sources without explicit language identification.
- We evaluate our multilingual meta-embedding on English-Spanish code-switching Named Entity Recognition (NER). The result shows the effectiveness of the method on multilingual setting and demonstrates that our meta-embedding can generalize to unseen languages in a cross-lingual setting.

2 Meta-Embeddings

Word embedding pre-training is a well-known method to transfer the knowledge from previous tasks to a target task that has fewer high-quality training data. Word embeddings are commonly used as features in supervised learning problems. We propose to generate a single word representation by extracting information from different pre-trained embeddings. We extend the idea of meta-embeddings from Kiela et al. (2018) to solve a multilingual task. We define a sentence that consists of m words $\{\mathbf{x}_j\}_{j=1}^{m}$, and $\{\mathbf{w}_{i,j}\}_{j=1}^{n}$ word vectors from n pre-trained word embeddings.

2.1 Baselines

We compare our method to two baselines: (1) concatenation and (2) linear ensembles.

Concatenation We concatenate word embeddings by merging the dimensions of word representations. This is the simplest way to utilize all sources of information; however, it is very inefficient due to the high-dimensional input:

$$\mathbf{w}_i^{CONCAT} = [\mathbf{w}_{i,1}, ..., \mathbf{w}_{i,n}]. \quad (1)$$

Linear Ensembles We sum all word embeddings into a single word vector with an equal weight. This method is efficient since it does not increase the dimensionality of the input. We apply a projection layer through $\mathbf{w}_{i,j}$ to have equal dimension before we sum:

$$\mathbf{w}_i^{LINEAR} = \sum_{j=0}^{n} \mathbf{w'}_{i,j}, \quad (2)$$

$$\mathbf{w'}_{i,j} = \mathbf{a}_j \cdot \mathbf{w}_{i,j} + b_j, \quad (3)$$

where $\mathbf{a}_j \in \mathbb{R}^{l \times d}$ and $b_j \in \mathbb{R}^d$ are trainable parameters, and l and d are the original dimensions of the pre-trained embeddings and projected dimensions respectively.

2.2 Multilingual Meta-Embedding

We generate a multilingual vector representation for each word by taking a weighted sum of monolingual embeddings. Each embedding $\mathbf{w}_{i,j}$ is projected with a fully connected layer with a non-linear scoring function ϕ (e.g., tanh) into a d-dimensional vector, and an attention mechanism to calculate attention weight $\alpha_{i,j} \in \mathbb{R}^d$:

$$\mathbf{w}_i^{MME} = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{w'}_{i,j}, \quad (4)$$

$$\alpha_{i,j} = \frac{e^{\phi(\mathbf{w'}_{i,j})}}{\sum_{j=1}^{n} e^{\phi(\mathbf{w'}_{i,j}))}}. \quad (5)$$

3 Named Entity Recognition

Our proposed model is based on a self-attention mechanism from a transformer encoder (Vaswani et al., 2017) followed by a Conditional Random Field (CRF) layer (Lafferty et al., 2001).

Encoder Architecture We apply a multi-layer transformer encoder as our sentence encoder:

$$h_0 = \text{Concat}(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_m)\mathbf{W}_t + \mathbf{W}_p, \quad (6)$$

$$h_l = \text{Transformer\_blocks}(h_0), \quad (7)$$

$$o = h_l \mathbf{W}_o + b_o, \quad (8)$$

where $\mathbf{W}_t$ is the projection matrix, $\mathbf{W}_p$ is the positional encoding matrix, $\mathbf{W}_o$ is the output layer, h_0 is the first layer hidden states, and h_l is the output representation from the final transformer layer. The output of the final layer is logits o.

Conditional Random Field This model calculates the dependencies across tag labels. NER requires a stronger constraint where `I-PERSON` should follow only after `B-PERSON`. We use CRF to learn the correlations between the current label and its neighbors (Lafferty et al., 2001). We consider $\mathbf{A} \in \mathbb{R}^{(k+2)\times(k+2)}$ as a trainable matrix, transition scores of the tags, where k is the number of tags. $\mathbf{A}_{i,j}$ denotes the transition score from tag i to tag j. We include a start tag and an end tag in the matrix, and calculate the score of a tag sequence $\mathbf{y}$ given o as follows:

$$s(o, \mathbf{y}) = \sum_{i=0}^{n} \mathbf{A}_{\mathbf{y}_i, \mathbf{y}_{i+1}} + \sum_{i=0}^{n} \mathbf{P}_{i, \mathbf{y}_i}, \quad (9)$$

where $\mathbf{P}_{i,\mathbf{y}_i} \in \mathbb{R}^{n\times k}$ represents the output probability of the tags. We use the Viterbi algorithm to select the best sequence.

4 Experiments

4.1 Dataset

For our experiment, we use English-Spanish tweets data provided by Aguilar et al. (2018). There are nine entity labels. The labels use IOB format, where every token is labeled as a `B-label` in the beginning and then an `I-label` if it is a named entity, or `O` otherwise.

4.2 Experimental Setup

We use pre-trained FastText [1] English *(EN)* and Spanish *(ES)* word embeddings (Grave et al., 2018) as our primary language embeddings, and pre-trained FastText Catalan *(CA)* and Portuguese *(PT)* word embeddings as our auxiliary language embeddings. We opt for *CA* and *PT* because they come from the same Romance language family as Spanish. We also include GloVe Twitter English embedding *(GLOVE_EN)* (Pennington et al., 2014).[2] Experiments are conducted in two different settings. In the multilingual setting, we learn our meta-embedding from primary languages and auxiliary languages, while in the cross-lingual setting only auxiliary languages are used. We run all experiments five times and calculate the average and standard deviation. To improve our final predictions, we ensemble all five experiments and take the results from a majority consensus.

[1] https://fasttext.cc/docs/en/crawl-vectors.html
[2] https://nlp.stanford.edu/projects/glove/

Approaches	F1
Trivedi et al. (2018) (Single)	61.89
Wang et al. (2018) (Single)	62.39
Wang et al. (2018) (Ensemble)	62.67
Winata et al. (2018b) (Single)	62.76
Trivedi et al. (2018) (Ensemble)	63.76
MONOLINGUAL	
EN	62.75 ± 0.66
ES	62.91 ± 1.07
CONCAT	
EN + ES	65.30 ± 0.38
EN + ES + CA	65.36 ± 0.85
EN + ES + PT	65.53 ± 0.79
EN + ES + CA + PT	64.99 ± 1.06
LINEAR	
EN + ES + CA + PT (Single)	65.33 ± 0.87
EN + ES + CA + PT (Ensemble)	67.03
MME	
EN + ES	65.43 ± 0.67
EN + ES + CA	65.69 ± 0.83
EN + ES + PT	65.65 ± 0.48
EN + ES + CA + PT (Single)	**66.63** ± **0.94**
EN + ES + CA + PT (Ensemble)	**68.34**

Table 1: Multilingual results (mean and standard deviation from five experiments). *EN*: both English FastText and GloVe word embeddings.

Implementation Details Our model is trained using a Noam optimizer with a dropout of 0.1 for multilingual setting and 0.3 for the cross-lingual setting. Our model contains four layers of transformer blocks with a hidden size of 200 and four heads. We start the training with a learning rate of 0.1. We replace user hashtags (#user) and mentions (@user) with `<USR>`, and URL (https://domain.com) with `<URL>`, similarly to Winata et al. (2018b).

5 Results

Multilingual experimental results are shown in Table 1. Interestingly, both concatenation and linear ensemble are strong baselines since they can achieve higher performance compared to any existing works that use more complicated features, such as character-based features using a bidirectional long short-term memory (LSTM) (Winata et al., 2018b; Wang et al., 2018) or a convolutional neural network (CNN) with additional gazetteers (Trivedi et al., 2018). Overall, our transformer encoder using a single word embedding achieves better performance compared to the LSTM encoder

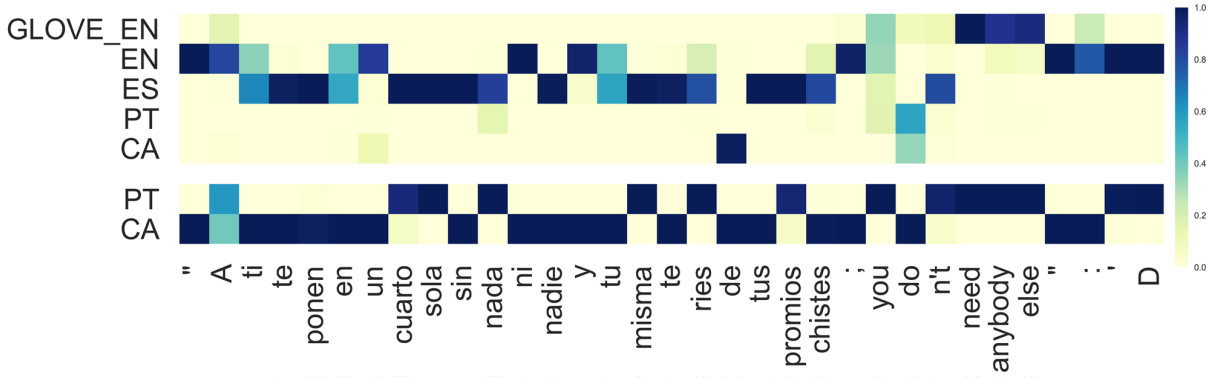

Figure 2: An example of attention weights on a development sample evaluated from a multilingual model **(top)** and a cross-lingual model **(bottom)**. Darker color shows higher attention scores.

Approaches	**F1**
MONOLINGUAL	
CA	53.96 ± 1.42
PT	54.86 ± 4.10
CONCAT	
CA + PT	58.28 ± 2.66
LINEAR	
CA + PT (Single)	60.72 ± 0.84
CA + PT (Ensemble)	62.9
MME	
CA + PT (Single)	**61.75 ± 0.56**
CA + PT (Ensemble)	**63.66**

Table 2: Cross-lingual results (mean and standard deviation from five experiments).

structure used by Winata et al. (2018b); Trivedi et al. (2018); Wang et al. (2018). More importantly, MME outperforms the two baselines on different language combinations, which shows its effectiveness. The results also show that the two baselines cannot effectively exploit the information from auxiliary languages. Here we note that the main advantage of MME is that it dynamically weights the different language pre-trained embeddings for each input token, while the concatenation and linear ensemble approaches always score the weights equally.

In the cross-lingual setting, our model does not perform well when we only use one auxiliary language, as seen in Table 2. A significant improvement is shown after we combine both languages, and MME shows a similar performance to the previous state-of-the-art result (Trivedi et al., 2018). This implies that our approach can effectively generalize word representations on an unseen language task by transferring information from languages that come from the same root as the primary languages.

We inspect the assigned weights on word embeddings to see which embedding our model attends. Figure 2 visualizes the weights for the multilingual and cross-lingual cases. It appears that our model can align words to their languages (e.g., Spanish words, such as "*ti*", "*te*", and "*ponen*" attend to *ES*) with strong confidences. In most cases, our model strongly attends to a single language and takes a small proportion of information from other languages. It shows the potential to automatically learn how to construct a multilingual embedding from semantically similar embeddings without requiring any language labels.

6 Related Work

Early studies on named entity recognition heavily relied on language-specific knowledge resources, such as hand-crafted features or gazetteers (Lafferty et al., 2001; Ratinov and Roth, 2009; Tsai et al., 2016). However, this approach was costly for new languages and domains. Thus, end-to-end approaches that do not rely on any external knowledge were proposed. Sobhana et al. (2010) proposed to use a CRF without any external resources, to leverage the label dependencies. Then, neural-based approaches, such as LSTM with a CRF (Lample et al., 2016; Lin et al., 2017; Greenberg et al., 2018) and LSTM with a CNN (Chiu and Nichols, 2016) showed a significant improvement in performance. Liu et al. (2018); Trivedi et al. (2018) proposed a character-level LSTM to capture the underlying style and structure, such as word boundaries and spellings. Finally, word-embedding ensemble techniques and preprocessing techniques, such as tokenization and normal-

ization have been introduced to reduce OOV issues (Winata et al., 2018b; Wang et al., 2018).

7 Conclusion

In this paper, we propose a novel approach to learn multilingual representations by leveraging monolingual pre-trained embeddings. MME solves the dependencies on the language identification in code-switching Named Entity Recognition task since it utilizes more information from semantically similar embeddings. The experiment results show that our method surpasses previous works and baselines, achieving the state-of-the-art performance. Moreover, cross-lingual setting experiments demonstrate the generalization ability of MME to an unseen language task.

Acknowledgments

We want to thank Samuel Cahyawijaya for insightful discussions about this project. This work has been partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, the Hong Kong SAR Government, and School of Engineering Ph.D. Fellowship Award, the Hong Kong University of Science and Technology, and RDC 1718050-0 of EMOS.AI.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Danushka Bollegala, Kohei Hayashi, and Ken-Ichi Kawarabayashi. 2018. Think globally, embed locally: locally linear meta-embedding of words. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3970–3976. AAAI Press.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Stephen Clark. 2015. Vector space models of lexical meaning. *Handbook of Contemporary Semantics*, 10:9781118882139.

Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding–computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 194–198.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual rnns and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. 2018. Marginal likelihood training of bilstm-crf for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2824–2829.

Douwe Kiela, Changhan Wang, and Kyunghyun Cho. 2018. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1477.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Bill Y Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 160–165.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Avo Muromägi, Kairit Sirts, and Sven Laur. 2017. Linear ensembles of word embedding models. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 96–104.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1543–1553.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.

N Sobhana, Pabitra Mitra, and SK Ghosh. 2010. Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):143–147.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Shashwat Trivedi, Harsh Rangwani, and Anil Kumar Singh. 2018. Iit (bhu) submission for the acl shared task on named entity recognition on code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 148–153.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228.

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

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018a. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67.

Genta Indra Winata, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2018b. Bilingual character representation for efficiently addressing out-of-vocabulary words in code-switching named entity recognition. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 110–114.

Wenpeng Yin and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1351–1360.