Hierarchical Meta-Embeddings for Code-Switching Named Entity Recognition

Genta Indra Winata, Zhaojiang Lin, Jamin Shin, Zihan Liu, 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,jmshinaa,zliucr}@connect.ust.hk,
pascale@ece.ust.hk

Abstract

In countries that speak multiple main languages, mixing up different languages within a conversation is commonly called *code-switching*. Previous works addressing this challenge mainly focused on word-level aspects such as word embeddings. However, in many cases, languages share common subwords, especially for closely related languages, but also for languages that are seemingly irrelevant. Therefore, we propose *Hierarchical Meta-Embeddings* (HME) that learn to combine multiple monolingual word-level and subword-level embeddings to create language-agnostic lexical representations. On the task of Named Entity Recognition for English-Spanish code-switching data, our model achieves the state-of-the-art performance in the multilingual settings. We also show that, in cross-lingual settings, our model not only leverages closely related languages, but also learns from languages with different roots. Finally, we show that combining different subunits are crucial for capturing code-switching entities.

1 Introduction

Code-switching is a phenomenon that often happens between multilingual speakers, in which they switch between their two languages in conversations, hence, it is practically useful to recognize this well in spoken language systems (Winata et al., 2018a). This occurs more often for entities such as organizations or products, which motivates us to focus on the specific problem of Named Entity Recognition (NER) in code-switching scenarios. We show one of the examples as the following:

- *walking dead* le quita el apetito a cualquiera
- **(translation)** *walking dead* (a movie title) takes away the appetite of anyone

For this task, previous works have mostly focused on applying pre-trained word embeddings from each language in order to represent noisy mixed-language texts, and combine them with character-level representations (Trivedi et al., 2018; Wang et al., 2018; Winata et al., 2018b). However, despite the effectiveness of such word-level approaches, they neglect the importance of subword-level characteristics shared across different languages. Such information is often hard to capture with word embeddings or randomly initialized character-level embeddings. Naturally, we can turn towards subword-level embeddings such as FastText (Grave et al., 2018) to help this task, which will evidently allow us to leverage the morphological structure shared across different languages.

Despite such expected usefulness, there has not been much attention focused around using subword-level features in this task. This is partly because of the non-trivial difficulty of combining different language embeddings in the subword space, which arises from the distinct segmentation into subwords for different languages. This leads us to explore the literature of Meta-Embeddings (Yin and Schütze, 2016; Muromägi et al., 2017; Bollegala et al., 2018; Coates and Bollegala, 2018; Kiela et al., 2018; Winata et al., 2019), which is a method to learn how to combine different embeddings.

In this paper, we propose **Hierarchical Meta-Embeddings (HME)** [1] which learns how to combine different pre-trained monolingual embeddings in word, subword, and character-level into a single language-agnostic lexical representation without using specific language identifiers. To address the issue of different segmentations, we add a Transformer (Vaswani et al., 2017) encoder

[1] The source code is available at https://github.com/gentaiscool/meta-emb

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3541–3547,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

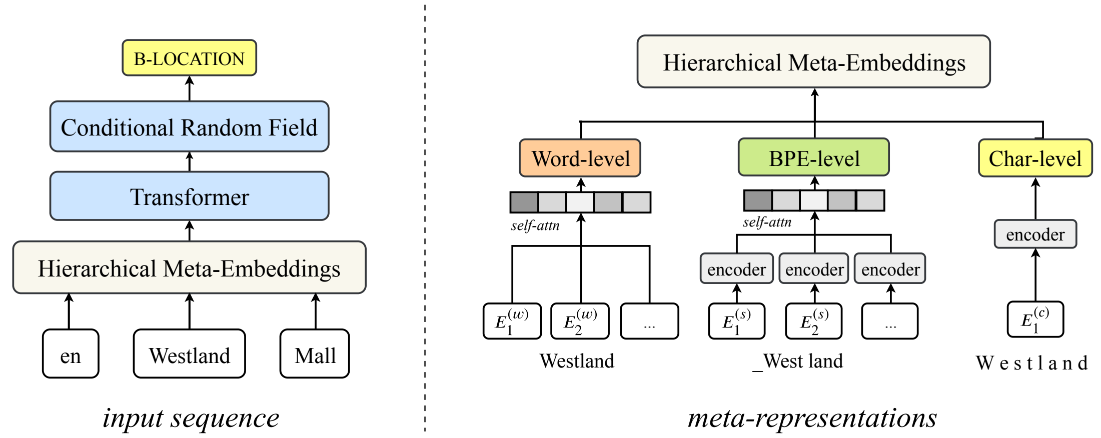

Figure 1: Hierarchical Meta-Embeddings (HME) architecture for Named Entity Recognition (NER) task. **Left:** Transformer-CRF architecture for Named Entity Recognition. **Right:** HME accept words, BPEs, and characters inputs.

which learns the important subwords in a given sentence. We evaluate our model on the task of Named Entity Recognition for English-Spanish code-switching data, and we use **Transformer-CRF**, a transformer-based encoder for sequence labeling based on the implementation of Winata et al. (2019). Our experimental results confirm that HME significantly outperforms the state-of-the-art system in *absolute F1 score*. The analysis shows that in the task of English-Spanish mixed texts not only similar languages like Portuguese or Catalan help, but also seemingly distant languages from Celtic origin also significantly increase the performance.

2 Related Work

Embeddings Previous works have extensively explored different representations such as word (Mikolov et al., 2013; Pennington et al., 2014; Grave et al., 2018; Xu et al., 2018), subword (Sennrich et al., 2016; Heinzerling and Strube, 2018), and character (dos Santos and Zadrozny, 2014; Wieting et al., 2016). Lample et al. (2016) has successfully concatenated character and word embeddings to their model, showing the potential of combining multiple representations. Liu et al. (2019) proposed to leverage word and subword embeddings into the application of unsupervised machine translation.

Meta-embeddings Recently, there are studies on combining multiple word embeddings in pre-processing steps (Yin and Schütze, 2016; Muromägi et al., 2017; Bollegala et al., 2018; Coates and Bollegala, 2018). Later, Kiela et al. (2018) introduced a method to dynamically learn word-level meta-embeddings, which can be effectively used in a supervised setting. Winata et al. (2019) proposed an idea to leverage multiple embeddings from different languages to generate language-agnostic meta-representations for mixed-language data.

3 Hierarchical Meta-Embeddings

We propose a method to combine word, subword, and character representations to create a mixture of embeddings. We generate a multilingual meta-embeddings of word and subword, and then, we concatenate them with character-level embeddings to generate final word representations, as shown in Figure 1. Let $\mathbf{w}$ be a sequence of words with n elements, where $\mathbf{w} = [w_1, \ldots, w_n]$. Each word can be tokenized into a list of subwords $\mathbf{s} = [s_1, \ldots, s_m]$ and a list of characters $\mathbf{c} = [c_1, \ldots, c_p]$. The list of subwords $\mathbf{s}$ is generated using a function f; $\mathbf{s} = f(\mathbf{w})$. Function f maps a word into a sequence of subwords. Further, let $E^{(w)}$, $E^{(s)}$, and $E^{(c)}$ be a set of word, subword, and character embedding lookup tables. Each set consists of different monolingual embeddings. Each element is transformed into a embedding vector in $\mathbb{R}^d$. We denote subscripts $_{\{i,j\}}$ as element and embedding language index, and superscripts $^{(w,s,c)}$ as word, subword, and character.

3.1 Multilingual Meta-Embeddings (MME)

We generate a meta-representations by taking the vector representation from multiple monolingual pre-trained embeddings in different subunits such as word and subword. We apply a **projection matrix** $\mathbf{W}_j$ to transform the dimensions from the original space $\mathbf{x}_{i,j} \in \mathbb{R}^d$ to a new shared space $\mathbf{x'}_{i,j} \in \mathbb{R}^{d'}$. Then, we calculate **attention weights** $\alpha_{i,j} \in \mathbb{R}^{d'}$ with a non-linear scoring function ϕ (e.g., tanh) to take important information from each individual embedding $\mathbf{x'}_{i,j}$. Then, MME is calculated by taking the weighted sum of the projected embeddings $\mathbf{x'}_{i,j}$:

$$\mathbf{x'}_{i,j} = \mathbf{W}_j \cdot \mathbf{x}_{i,j}, \tag{1}$$

$$\alpha_{i,j} = \frac{\exp(\phi(\mathbf{x'}_{i,j}))}{\sum_{k=1}^{n} \exp(\phi(\mathbf{x'}_{i,k}))}, \tag{2}$$

$$\mathbf{u}_i = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{x'}_{i,j}. \tag{3}$$

3.2 Mapping Subwords and Characters to Word-Level Representations

We propose to map subword into word representations and choose byte-pair encodings (BPEs) (Sennrich et al., 2016) since it has a compact vocabulary. First, we apply f to segment words into sets of subwords, and then we extract the pre-trained subword embedding vectors $\mathbf{x}_{i,j}^{(s)} \in \mathbb{R}^d$ for language j.

Since, each language has a different f, we replace the projection matrix with Transformer (Vaswani et al., 2017) to learn and combine important subwords into a single vector representation. Then, we create $\mathbf{u}_i^{(s)} \in \mathbb{R}^{d'}$ which represents the subword-level MME by taking the weighted sum of $\mathbf{x'}_{i,j}^{(s)} \in \mathbb{R}^{d'}$.

$$\mathbf{x'}_{i,j}^{(s)} = \text{Encoder}(\mathbf{x}_{i,j}^{(s)}), \tag{4}$$

$$\mathbf{u}_i^{(s)} = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{x'}_{i,j}^{(s)}. \tag{5}$$

To combine character-level representations, we apply an encoder to each character.

$$\mathbf{u}_i^{(c)} = \text{Encoder}(\mathbf{x}_i) \in \mathbb{R}^{d'}. \tag{6}$$

We combine the word-level, subword-level, and character-level representations by concatenation $\mathbf{u}_i^{HME} = (\mathbf{u}_i^{(w)}, \mathbf{u}_i^{(s)}, \mathbf{u}_i^{(c)})$, where $\mathbf{u}_i^{(w)} \in \mathbb{R}^{d'}$ and $\mathbf{u}_i^{(s)} \in \mathbb{R}^{d'}$ are word-level MME and BPE-level MME, and $\mathbf{u}_i^{(c)}$ is a character embedding. We randomly initialize the character embedding and keep it trainable. We fix all subword and word pre-trained embeddings during the training.

3.3 Sequence Labeling

To predict the entities, we use Transformer-CRF, a transformer-based encoder followed by a Conditional Random Field (CRF) layer (Lafferty et al., 2001). The CRF layer is useful to constraint the dependencies between labels.

$$h = \text{Transformer}(\mathbf{u}), h' = \text{CRF}(h). \tag{7}$$

The best output sequence is selected by a forward propagation using the Viterbi algorithm.

4 Experiments

4.1 Experimental Setup

We train our model for solving Named Entity Recognition on English-Spanish code-switching tweets data from Aguilar et al. (2018). There are nine entity labels with IOB format. The training, development, and testing sets contain 50,757, 832, and 15,634 tweets, respectively.

We use FastText word embeddings trained from Common Crawl and Wikipedia (Grave et al., 2018) for English (*es*), Spanish (*es*), including **four Romance languages**: Catalan (*ca*), Portuguese (*pt*), French (*fr*), Italian (*it*), and **a Germanic language**: German (*de*), and **five Celtic languages** as the distant language group: Breton (*br*), Welsh (*cy*), Irish (*ga*), Scottish Gaelic (*gd*), Manx (*gv*). We also add the English Twitter GloVe word embeddings (Pennington et al., 2014) and BPE-based subword embeddings from Heinzerling and Strube (2018).

We train our model in two different settings: **(1) multilingual setting**, we combine main languages (*en-es*) with Romance languages and a Germanic language, and **(2) cross-lingual setting**, we use Romance and Germanic languages without main languages. Our model contains four layers of transformer encoders with a hidden size of 200, four heads, and a dropout of 0.1. We use Adam optimizer and start the training with a learning rate of 0.1 and an early stop of 15 iterations. We replace user hashtags and mentions with `<USR>`, emoji with `<EMOJI>`, and URL with `<URL>`. We evaluate our model using *absolute F1 score* metric.

4.2 Baselines

CONCAT We concatenate word embeddings by merging the dimensions of word representations. This method combines embeddings into a high-dimensional input that may cause inefficient com-

Model	Multilingual embeddings				Cross-lingual embeddings		
	main languages	+ closely-related languages		+ distant languages	closely-related languages		distant languages
	en-es	ca-pt	ca-pt-de-fr-it	br-cy-ga-gd-gv	ca-pt	ca-pt-de-fr-it	br-cy-ga-gd-gv
Flat word-level embeddings							
CONCAT	65.3 ± 0.38	64.99 ± 1.06	65.91 ± 1.16	65.79 ± 1.36	58.28 ± 2.66	64.02 ± 0.26	50.77 ± 1.55
LINEAR	64.61 ± 0.77	65.33 ± 0.87	65.63 ± 0.92	64.95 ± 0.77	60.72 ± 0.84	62.37 ± 1.01	53.28 ± 0.41
Multilingual Meta-Embeddings (MME) (Winata et al., 2019)							
Word	65.43 ± 0.67	66.63 ± 0.94	66.8 ± 0.43	66.56 ± 0.4	61.75 ± 0.56	63.23 ± 0.29	53.43 ± 0.37
Hierarchical Meta-Embeddings (HME)							
+ BPE	65.9 ± 0.72	67.31 ± 0.34	67.26 ± 0.54	66.88 ± 0.37	63.44 ± 0.33	63.78 ± 0.62	60.19 ± 0.63
+ Char	65.88 ± 1.02	67.38 ± 0.84	65.94 ± 0.47	66.1 ± 0.33	61.97 ± 0.6	63.06 ± 0.69	57.5 ± 0.56
+ BPE + Char	66.55 ± 0.72	**67.8 ± 0.31**	67.07 ± 0.49	67.02 ± 0.16	63.9 ± 0.22	**64.52 ± 0.35**	60.88 ± 0.84

Table 1: Results (percentage F1 mean and standard deviation from five experiments). **Multilingual**: with main languages, **Cross-lingual**: without main languages.

Model	F1
Trivedi et al. (2018)	61.89
Wang et al. (2018)	62.39
Wang et al. (2018) (Ensemble)	62.67
Winata et al. (2018b)	62.76
Trivedi et al. (2018) (Ensemble)	63.76
Winata et al. (2019) MME	66.63 ± 0.94
Random embeddings	46.68 ± 0.79
Aligned embeddings	
MUSE (es → en)	60.89 ± 0.37
MUSE (en → es)	61.49 ± 0.62
Multilingual embeddings	
Our approach	67.8 ± 0.31
Our approach (Ensemble)†	**69.17**
Cross-lingual embeddings	
Our approach	64.52 ± 0.35
Our approach (Ensemble)†	**65.99**

Table 2: Comparison to existing works. **Ensemble:** We run a majority voting scheme from five different models.

putation.

$$\mathbf{x}_i^{CONCAT} = [\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,n}]. \tag{8}$$

LINEAR We sum all word embeddings into a single word vector with equal weight. This method combines embeddings without considering the importance of each of them.

$$\mathbf{x}_i^{LINEAR} = \sum_{j=0}^{n} \mathbf{x'}_{i,j}. \tag{9}$$

Random Embeddings We use randomly initialized word embeddings and keep it trainable to calculate the lower-bound performance.

Aligned Embeddings We align English and Spanish FastText embeddings using CSLS with two scenarios. We set English (en) as the source language and Spanish (es) (en → es) as the target language, and vice versa (es → en). We run MUSE by using the code prepared by the authors of Conneau et al. (2017). [2]

5 Results & Discussion

In general, from Table 1, we can see that word-level meta-embeddings even without subword or character-level information, consistently perform better than flat baselines (e.g., CONCAT and LINEAR) in all settings. This is mainly because of the attention layer which does not require additional parameters. Furthermore, comparing our approach to previous state-of-the-art models, we can clearly see that our proposed approaches all significantly outperform them.

From Table 1, in Multilingual setting, which trains with the main languages, it is evident that adding both closely-related and distant language embeddings improves the performance. This shows us that our model is able to leverage the lexical similarity between the languages. This is more distinctly shown in Cross-lingual setting as using distant languages significantly perform less than using closely-related ones (e.g., *ca-pt*). Interestingly, for distant languages, when adding subwords, we can still see a drastic performance increase. We hypothesize that even though the characters are mostly different, the lexical structure is similar to our main languages.

On the other hand, adding subword inputs to the model is consistently better than characters. This is due to the transfer of the information from the pre-trained subword embeddings. As shown in Ta-

[2] https://github.com/facebookresearch/MUSE

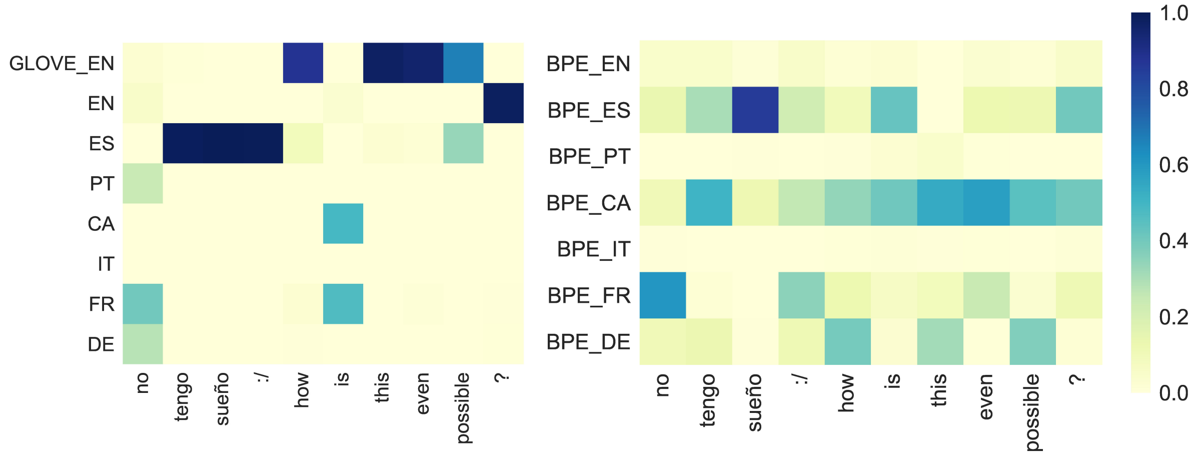

Figure 2: Heatmap of attention over languages from a validation sample. **Left**: word-level MME, **Right**: BPE-level MME. We extract the attention weights from a multilingual model (*en-es-ca-pt-de-fr-it*).

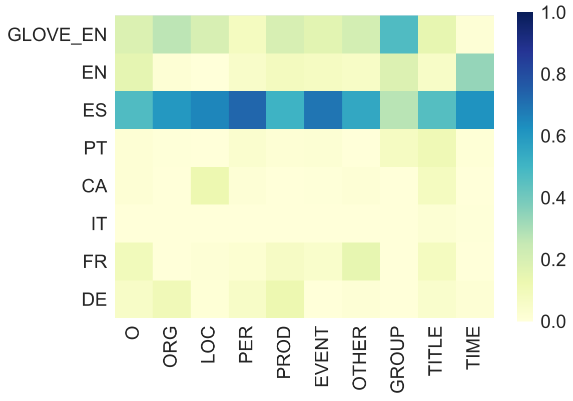

Figure 3: The average of attention weights for word embeddings versus NER tags from the validation set.

ble 1, subword embeddings is more effective for distant languages (Celtic languages) than closely-related languages such as Catalan or Portuguese.

Moreover, we visualize the attention weights of the model in word and subword-level to interpret the model dynamics. From the left image of Figure 2, in word-level, the model mostly chooses the correct language embedding for each word, but also combines with different languages. Without any language identifiers, it is impressive to see that our model learns to attend to the right languages. The right side of Figure 2, which shows attention weight distributions for subword-level, demonstrates interesting behaviors, in which for most English subwords, the model leverages *ca*, *fr*, and *de* embeddings. We hypothesize this is because the dataset is mainly constructed with Spanish words, which can also be verified from Figure 3 in which most NER tags are classified as *es*.

6 Conclusion

We propose *Hierarchical Meta-Embeddings* (HME) that learns how to combine multiple monolingual word-level and subword-level embeddings to create language-agnostic representations without specific language information. We achieve the state-of-the-art results on the task of Named Entity Recognition for English-Spanish code-switching data. We also show that our model can leverage subword information very effectively from languages from different roots to generate better word representations.

Acknowledgments

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. We sincerely thank the three anonymous reviewers for their insightful comments on our paper.

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.

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.

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.

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

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)*.

Benjamin Heinzerling and Michael Strube. 2018. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282, Florence, Italy. Association for Computational Linguistics.

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.

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.

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

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.

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.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2019. Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186.

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.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multitask training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

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.