Language Clustering for Multilingual Named Entity Recognition

Kyle Shaffer
Language Weaver (RWS Group)
kshaffer@rws.com

Abstract

Recent work in multilingual natural language processing has shown progress in various tasks such as natural language inference and joint multilingual translation. Despite success in learning across many languages, challenges arise where multilingual training regimes often boost performance on some languages at the expense of others. For multilingual named entity recognition (NER) we propose a simple technique that groups similar languages together by using embeddings from a pre-trained masked language model, and automatically discovering language clusters in this embedding space. Specifically, we fine-tune an XLM-Roberta model on a language identification task, and use embeddings from this model for clustering. We conduct experiments on 15 diverse languages in the WikiAnn dataset and show our technique largely outperforms three baselines: (1) training a multilingual model jointly on all available languages, (2) training one monolingual model per language, and (3) grouping languages by linguistic family. We also conduct analyses showing meaningful multilingual transfer for low-resource languages (Swahili and Yoruba), despite being automatically grouped with other seemingly disparate languages.

1 Introduction

Large transformer language models (Vaswani et al., 2017; Devlin et al., 2019) have shown impressive progress on tasks across different languages, including joint multilingual learning. Many works have focused on cross-lingual transfer from high- to low-resource languages in a zero- or few-shot setting (Hu et al., 2020). However recent work has also highlighted that small amounts of data may be available for some low-resource languages, and even very few examples for fine-tuning on a target language can be effective (Lauscher et al., 2020). Given these insights and the scarcity of studies that present a middle ground between monolingual and multilingual learning, we investigate methods for clustering languages to boost multilingual performance on named entity recognition (NER).

One transformer model that has shown particularly strong performance on multilingual tasks is XLM-Roberta (Conneau et al., 2020), a variant of the Roberta model (Liu et al., 2019) that adapts the multilingual training regime of XLM (Lample and Conneau, 2019) to a CommonCrawl corpus containing 100 languages. This model can be adapted to tasks in multiple languages, and we take this as the base model for NER fine-tuning. Additionally, inspired by work in multilingual neural machine translation (NMT) (Tan et al., 2019), we investigate a method for grouping similar languages using an automated clustering method. We provide a focused evaluation of this method on 15 languages from the WikiAnn corpus (Pan et al., 2017) following the train-test splits from Rahimi et al. (2019) and show that NER models trained on language clusters largely outperform (a) individual monolingual models trained for each language, (b) multilingual models trained on languages that are grouped by linguistic family, and (c) a single multilingual model trained on all available languages.

2 Related Work

Mueller et al. (2020) fine-tune multilingual NER models monolingually on individual target languages, showing this technique to be effective in boosting F1 scores in all considered languages in their study. In a similar vein, Lauscher et al. (2020) test the effectiveness of few-shot adaptation of multilingual models to new languages, finding that even including as few as 10 samples from the target language increases performance over zero-shot transfer.

Similar to our work, Chung et al. (2020) explore grouping languages by similarity, but focus on optimally constructing multilingual sub-word vocabularies, and show that these inputs perform bet-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 40–45
November 7–11, 2021.

ter on tasks such as XNLI and WikiAnn NER. In a more focused work, Arkhipov et al. (2019) investigate NER performance on four related Slavic languages, and demonstrate the advantages of pre-training multilingual BERT on the unsupervised language modeling task. Finally, while not focusing on NER, Tan et al. (2019) show performance gains in multilingual NMT using clustering based on language tag embeddings. We take most direct inspiration from this work, though our embedding technique differs.

3 Clustering Languages for Multilingual NER

While many of the works above provide insight into multilingual NER performance in both broad and narrow contexts, many focus on zero- or few-shot transfer, or linguistically similar language groups. Our work seeks to fill a gap by studying multilingual NER performance for several diverse languages where data is available (though not evenly distributed) to understand how to best group languages for multilingual NER training. Here we present our proposed automatic clustering approach to address this problem.

To obtain input representations for a clustering algorithm, we use a pre-trained XLM-R model.[1] For each sentence in our corpus, we obtain a single vector for the sentence as the output from XLM-R. We then input these vectors to a clustering algorithm to obtain cluster-assigned labels for each sentence. To obtain the final cluster label for an entire language, we simply compute the majority vote of the clustering labels for all sentences within a language.

While the base XLM-R model provides a good starting point for downstream tasks, we found that when clustering in this model's embedding space most languages were assigned to the same cluster regardless of the number of desired clusters.[2] Thus, we fine-tune XLM-R on a language identification task where the model is trained to classify sentences into one of the 15 languages in the dataset. We then use the `[CLS]` token embedding that is fed to the classification layer during fine-tuning as the input for clustering. This language identification model is fine-tuned for 3 epochs on the WikiAnn training set with a batch size of 20, and achieves an overall accuracy of 90% across all languages. Figure 1 shows qualitative evidence of strong grouping of languages such as overlap between Chinese and Japanese that is reflected in assigned clusters in Section 4 below.

To automatically group languages, we follow Tan et al. (2019) in choosing bottom-up agglomerative clustering, which assigns each data point its own cluster and iteratively merges clusters such that the sum of squared distances between points within all clusters is minimized. Similar to k-means, agglomerative clustering uses a k hyper-parameter for the number of clusters, and after experimentation with $k \in \{3, 4, 5, 6\}$ and noting sub-optimal groupings for many values of k, we set this parameter to 4.

4 Experimental Setup

For training NER models with this method, we group all sentences from languages that are assigned the same cluster and train and evaluate on these languages from the WikiAnn dataset. We compare these models against monolingual models for each language, a single multilingual model trained on all languages, and another set of grouped models using linguistic family as the assigned group. Language groupings for the automated clustering method and the linguistically-informed method are shown in tables 1 and 2 respectively.

We note several observations from these groupings. First, several languages appear in their own individual clusters when grouped by linguistic family (ja, ko, zh) or our clustering method (ar). In these cases results for grouped models are identical to those for monolingual models. Second, we note differences between the automated and linguistic grouping methods, most notably the inclusion of Yoruba and Swahili in an otherwise Indo-European cluster. This may be the result of few examples for these two languages in this dataset[3], however we show in Section 5 that this grouping is beneficial to these languages in our experiments despite being counter-intuitive from a linguistic perspective. Finally, we note the grouping of Chinese and Japanese under the automatic clustering method, consistent with qualitative evidence from overlap in semantic space of the fine-tuned language classifier discussed above.

[1] We use the following pre-trained weights: https://huggingface.co/xlm-roberta-base

[2] We experimented with using the `[CLS]` token and max-pooling over the final hidden states as input for the clustering algorithm.

[3] WikiAnn contains 100 yo and 1,000 sw training examples compared to 20,000 for most other languages studied here.

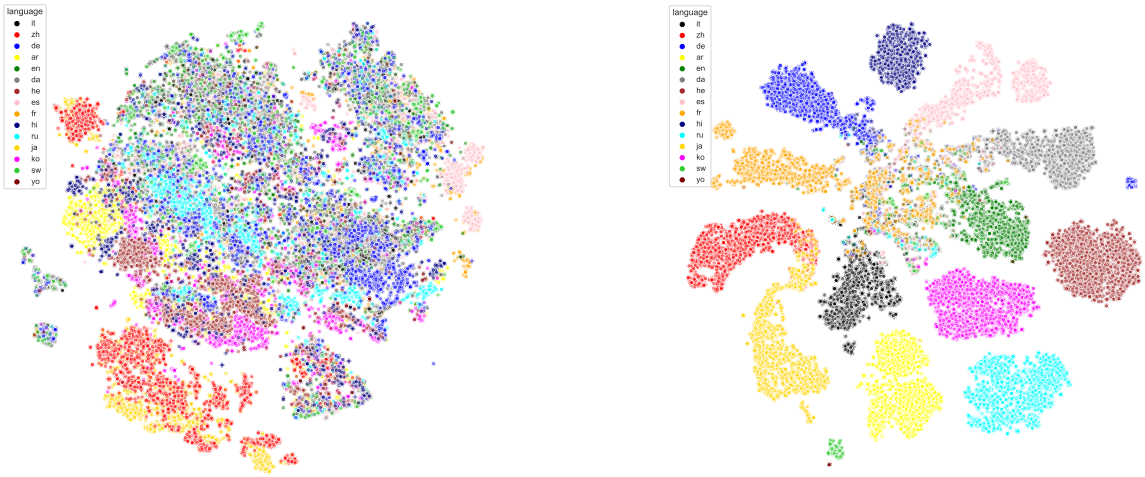

Figure 1: Two-dimensional TSNE projection of sentences for max-pooled XLM-R hidden states **(left)**, and for XLM-R after fine-tuning on the language classification task **(right)**.

Cluster Number	Languages
Cluster 1	ar
Cluster 2	da, de, en, es, fr, hi, it, sw, yo
Cluster 3	he, ko, ru
Cluster 4	ja, zh

Table 1: Assignments for languages based on clustering method.

Language Family	Languages
Indo-European	da, de, en, es, fr, hi, it, ru
Afro-Asiatic	he, ar
Niger-Congo	sw, yo
Koreanic	ko
Japonic	ja
Sino-Tibetan	zh

Table 2: Assignments for languages based on linguistic family.

We initialize all NER models from the pre-trained XLM-R checkpoint available from the Huggingface Transformers library (Wolf et al., 2020) and train all models for 3 epochs, with a batch size of 20, and maximum input sequence length of 300 sub-tokens. We evaluate with span-based F1 score as in the CoNLL-2003 evaluation script (Sang and Meulder, 2003), and report this metric for the three classes available in the dataset - location, organization, and person.

5 Results

Table 3 presents an overview of results from our experiments. For each language grouping we train five models, each newly initialized from the XLM-R weights except for the token classification head, whose weights are randomly initialized. Table 3 reports mean scores over these five training runs with standard deviation in parentheses. We first note fairly strong performance across all methods and languages except Swahili and Yoruba in the monolingual and language family settings. This is unsurprising given that these languages have significantly less data in the WikiAnn dataset. For most classes and languages, best performance is observed when using the proposed language clustering technique. We note slightly better performance using multilingual training for some languages, however these differences are typically less than one F1 point when compared to the clustering based models. Most notably, for Arabic we see best performance across all classes under the fully multilingual grouping, suggesting a need for improvement in our clustering method which assigns Arabic to its own cluster. Overall, these results show evidence that grouping languages together for multilingual NER provides a strong alternative to training a monolingual model for each language or a single multilingual model for all languages.

Additional information about these results is plotted in Figure 2 below.[4] Here we use box plots to show the distribution of the class-averaged F1

[4]Note that y-axes are separately scaled for each sub-plot to show detail for each language.

	Monolingual			Language Family			Clustering			Multilingual		
	LOC	**ORG**	**PER**	**LOC**	**ORG**	**PER**	**LOC**	**ORG**	**PER**	**LOC**	**ORG**	**PER**
yo	13.13 (5.67)	4.32 (0.42)	5.17 (2.98)	24.59 (13.68)	7.17 (4.24)	29.15 (6.94)	**83.89 (1.18)**	**79.42 (0.43)**	**90.61 (0.68)**	74.99 (3.26)	73.05 (5.08)	82.10 (1.06)
sw	60.74 (0.51)	67.54 (0.89)	61.92 (2.26)	52.48 (6.04)	59.25 (1.68)	53.90 (5.72)	**88.62 (0.36)**	**88.87 (0.62)**	**92.57 (0.33)**	86.68 (0.99)	86.49 (0.29)	91.10 (0.26)
it	87.88 (0.37)	79.25 (1.96)	89.92 (1.19)	89.58 (0.16)	82.70 (0.62)	92.61 (1.15)	**91.02 (0.26)**	**85.31 (0.10)**	**94.43 (0.12)**	89.94 (0.08)	83.41 (0.61)	93.63 (0.21)
de	85.48 (0.60)	74.63 (0.35)	86.04 (0.37)	86.26 (0.28)	76.86 (0.72)	88.42 (0.44)	**87.65 (0.28)**	**79.42 (0.28)**	**89.26 (0.08)**	86.33 (0.48)	77.36 (0.21)	88.59 (0.17)
en	81.97 (1.24)	68.26 (1.45)	83.82 (1.58)	84.36 (0.11)	72.78 (0.72)	86.98 (0.19)	**86.17 (0.01)**	**75.61 (0.05)**	**88.58 (0.12)**	84.81 (0.37)	73.50 (0.85)	87.9 (0.11)
da	90.56 (0.47)	81.72 (1.26)	89.45 (1.28)	91.45 (0.11)	84.39 (0.42)	92.36 (0.25)	**92.46 (0.02)**	**85.05 (0.22)**	93.19 (0.05)	91.41 (0.07)	84.73 (0.41)	**93.23 (0.32)**
es	84.99 (3.97)	76.23 (5.63)	83.68 (4.41)	89.96 (0.09)	84.45 (0.63)	91.57 (0.37)	**91.10 (0.19)**	**85.71 (0.01)**	92.26 (0.41)	90.33 (0.28)	85.60 (0.40)	**92.31 (0.13)**
fr	86.04 (2.09)	75.38 (3.57)	85.71 (3.94)	89.01 (0.31)	81.79 (0.52)	91.60 (0.35)	**90.95 (0.11)**	**84.54 (0.17)**	**93.04 (0.14)**	89.37 (0.32)	82.32 (0.35)	92.04 (0.10)
hi	80.62 (0.28)	82.69 (0.18)	88.27 (0.89)	81.63 (1.49)	83.62 (1.14)	90.61 (0.89)	**84.15 (0.72)**	**85.29 (0.56)**	91.46 (0.04)	84.14 (1.37)	85.14 (0.35)	**91.96 (0.29)**
ar	88.36 (0.09)	80.06 (0.48)	87.10 (0.13)	87.83 (0.53)	79.83 (0.32)	86.86 (0.12)	88.36 (0.08)	80.06 (0.48)	87.10 (0.13)	**88.75 (0.38)**	**81.76 (0.36)**	**89.20 (0.26)**
he	**87.12 (0.01)**	77.63 (0.05)	87.93 (0.06)	85.48 (0.26)	75.24 (0.64)	85.76 (0.38)	86.99 (0.14)	**78.45 (0.26)**	**88.31 (0.08)**	86.2 (0.38)	78.04 (0.39)	87.70 (0.24)
ru	88.77 (0.07)	82.04 (0.19)	94.47 (0.17)	88.29 (0.28)	81.74 (0.27)	94.02 (0.15)	**89.44 (0.10)**	**82.78 (0.18)**	94.05 (0.31)	88.84 (0.21)	82.29 (0.46)	**94.59 (0.05)**
ko	91.20 (0.03)	79.19 (0.10)	87.60 (0.19)	90.72 (0.25)	76.69 (0.53)	84.41 (0.68)	**91.58 (0.14)**	**79.86 (0.25)**	**87.65 (0.15)**	90.72 (0.39)	79.03 (0.69)	86.48 (0.25)
zh	76.61 (0.74)	63.78 (0.59)	80.61 (1.21)	76.61 (0.74)	63.78 (0.59)	80.61 (1.21)	**81.52 (0.51)**	71.45 (0.17)	84.40 (0.02)	81.22 (0.26)	**71.99 (0.44)**	**84.93 (0.20)**
ja	71.63 (0.72)	57.06 (1.46)	71.85 (2.29)	71.63 (0.72)	57.06 (1.46)	71.85 (2.29)	**75.26 (0.28)**	62.39 (0.38)	76.73 (0.50)	73.55 (0.90)	**62.68 (0.38)**	**77.45 (0.25)**

Table 3: Comparison of F1 score results on WikiAnn test set. Each score is the mean over five training runs, with standard deviation reported in parentheses.

score for each language, with each box representing a different language grouping. This visualization highlights interesting differences in the spread of scores, including comparatively large spread for monolingual training of languages such as Italian, French and Spanish. Conversely, we see relatively little spread in scores for the clustered language grouping within each language. This may be evidence of increased training stability when grouping similar languages together, although further work is needed to better understand these trends.

We also note drastic performance improvement for Swahili and Yoruba when trained in a single multilingual model compared to monolingual training, consistent with previous findings for low-resource languages in multilingual settings (Rahimi et al., 2019; Hu et al., 2020; Mueller et al., 2020; Conneau et al., 2020). However, we observe best performance for these two languages when grouped using our proposed clustering method, which is somewhat surprising given the counter-intuitive grouping with mostly European languages, though this grouping is also observed in previous work (Chung et al., 2020).

This raises a question as to whether this improvement is due to effective learning of shared multilingual representations or whether it is primarily due to availability of more data of any kind. To test this, we evaluate NER models in a zero-shot framework where we train a multilingual model on all languages in Cluster 2 with Swahili and Yoruba removed and evaluate this model on these two held-out languages. These results are presented in Table 4 below. While we see that this transfer beats performance from monolingual models for some classes in these languages, we see that F1 scores for all classes are well below both the cluster models and the single multilingual model. This suggests that some of the increased performance on these languages in the clustering setting is due to advantageous multilingual transfer.

	LOC	**ORG**	**PER**
yo	24.30 (-59.59)	1.46 (-77.96)	61.13 (-29.48)
sw	55.38 (-33.24)	37.35 (-51.52)	87.11 (-5.46)

Table 4: Zero-shot transfer from Cluster 2 to Yoruba and Swahili. Parentheses show difference in F1-score compared to clustering model results.

Finally, as a test of generalization of our method we evaluate on the English test set of the ConNLL-2003 NER dataset. We present results from training on four different training sets: the WikiAnn training set containing languages in Cluster 2 (denoted WikiAnn in our results), the CoNLL-2003 English training set (denoted CoNLL-2003), the combination of all WikiAnn and CoNLL-2003 training data (denoted All), and finally the combination of the WikiAnn language Cluster 2 training set with the CoNLL-2003 training set (denoted as Cluster Combined). We train each model for a single run with the same settings described above, and present our results in Table 5 below.

WikiAnn	**CoNLL-2003**	**All**	**Cluster Comb.**
58.06	90.27	89.02	**90.83**

Table 5: Average F1 scores on CoNLL-2003 English test set.

We first note poor performance from the model trained solely on WikiAnn data, which is unsurprising given the domain mismatch and idiosyncrasies in each of the datasets. Performance improves substantially in all cases where CoNLL training data

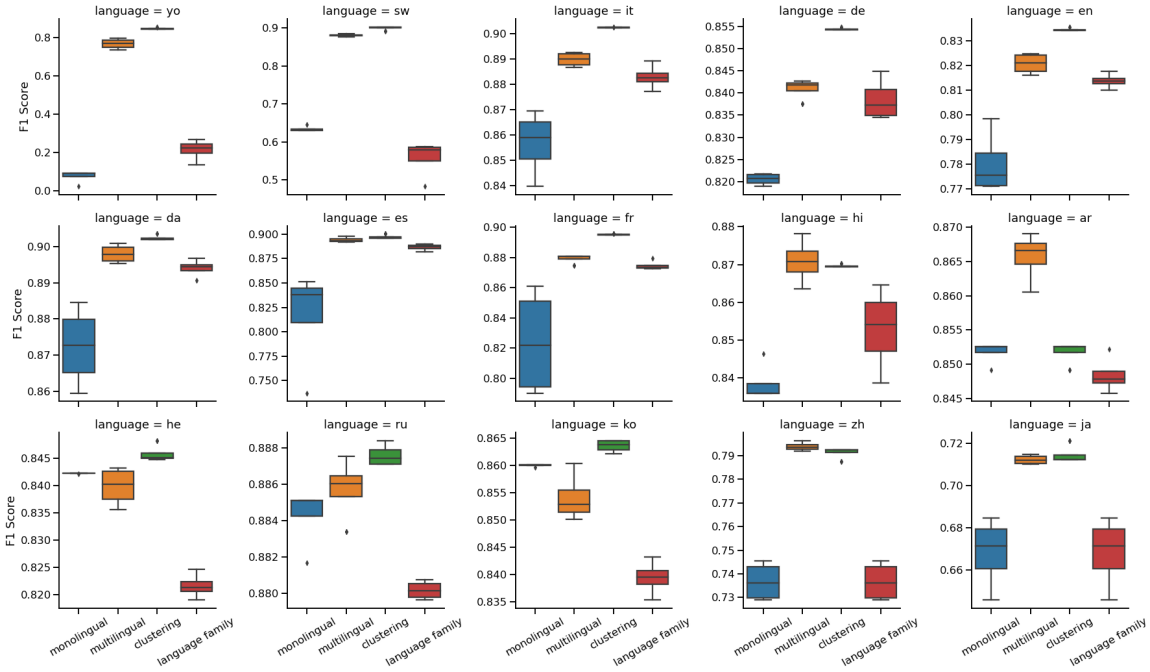

Figure 2: Distribution of average F1 scores over five runs for each language and language grouping.

is used, with best performance noted in the "Cluster Combined" model, which slightly outperforms using all available training data from both datasets. This suggests that even in a new domain the multilingual representations of closely related languages may be helpful, and that utilizing related languages is more useful than simply combining all available multilingual training data as in the "All" setting. We finally note that, despite not being extensively tuned on this dataset, we achieve results within 3.5 F1 points of previously reported state of the art results on this test set (Yamada et al., 2020).

6 Conclusion

We have presented a simple data-driven clustering technique for improving performance on multilingual NER, and showed that this technique largely outperforms naive combination of all languages studied here within a single model, as well as outperforming monolingual models and models for languages grouped by linguistic family. We further tested whether improved performance for low-resource languages in the Niger-Congo family was solely the result of more available data and showed evidence of multilingual transfer via a focused zero-shot experiment. We believe this straightforward method can be easily applied to other multilingual settings as has been shown in previous work in NMT.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for named entity recognition on slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pages 89–93.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics.

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

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.

2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

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 (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

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*.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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