Multilingual BERT Named Entity Recognition

Alida Sefada 802668, Or Tzofi 208468546

Submitted as final project report for the Deep Learning course,
IDC, 2020

1 Introduction

Named Entity Recognition (NER) is one of the most common tasks of natural language processing (NLP). Its goal is locating and classifying named entities in unstructured text into pre-defined categories such as person names, locations, organizations and more. There is a great deal of research on NER in the literature, most of which focuses on monolingual tasks and data. In this paper, we will be exploring the BERT multilingual model for NER across multiple languages.

We have chosen this topic for the final project in deep learning due to our interest in languages, the opportunity to learn how multilingual NLP models work and the results they can achieve in the light of the linguistic similarities and differences between certain languages. In this work, we seek to train Multilingual BERT NER models on different combinations of languages, test them on the same sets of languages and use them for zero-shot NER on other languages. That is, learning with different source and target domain (Pan and Yang, 2010). Then we will compare the results of the different models and find the best training languages for each target language.

1.1 Related Works

Prior multilingual works include Wu and Dredze (2019) who showed the zero-shot transfer learning potential of Multilingual BERT on 5 NLP tasks. Pires, Schlinger, and Garrette (2019) used Multilingual BERT to obtain state-of-the-art results for zero-shot on the four CoNLL languages (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Baumann (2019) used Multilingual BERT for NER on English and German. In these works, the model was fine-tuned on one language at a time for the NER task.

Moon et al. (2019) investigated a Multilingual BERT model that was trained jointly on many languages and was able to achieve state-of-the-art results for zero-shot on three CoNLL languages. Karthikeyan et al. (2020) provided a study on the cross-lingual abilities of Multilingual BERT covering the linguistic properties and the similarities of target and source languages, the network architecture, and the input and the learning objective.

Aside from these papers, we used some other projects included in the references.

2 Solution

2.1 General approach

The goal of the experiment is investigating the effect the different combinations of source languages (i.e., the ones used for training) have on the target languages, considering the linguistic similarities and differences between them. We build 10 different Multilingual BERT models, each fine-tuned for NER and trained on different languages: Three models are trained on one language, four models are trained jointly on two languages, and three models are trained jointly on three languages. Then, we evaluate each model on 17 languages, some of which do not share the same script and structure, with zero-shot inference and analyze the models' results.

2.2 Design

2.2.1 Datasets

We use the English CoNLL 2003 dataset and the Spanish and Dutch CoNLL 2002 dataset (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) which are widely used in NER research. We also use the German GermEval 2014 dataset (Benikova et al., 2014), the Danish DaNE dataset (Hvingelby et al., 2020) and the Afrikaans NER Corpus developed by North-West University, South Africa. All these datasets use the CoNLL annotation standards and have four entity types (PER, LOC, ORG, MISC).

In addition, we used the Chinese MSRA dataset (Levow, 2006) and the WikiANN dataset (Rahimi et al., 2017), from which we used the French, Italian, Portuguese, Swedish, Turkish, Russian, Hebrew, Arabic, Japanese and Korean datasets. These datasets have three entity types (PER, LOC, ORG).

2.2.2 Model Architecture

We use the pretrained bert-base-multingual-cased and set the number of NER tags to 9, that is [O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, B-MISC, I-MISC]. Note that this BERT model has 12 hidden layers. We add a dropout layer with a probability of 0.3, similarly to other projects, after comparing the results of 0.1 and 0.3 dropouts. Finally, an additional linear layer of size (768, 9) for the NER tags classification is added on top of the model (768 is BERT's hidden dimension). We use the Cross Entropy loss function.

The data is preprocessed using Google's WordPiece tokenization as suggested by Devlin et al. (2018) and Karthikeyan et al. (2020). We set the maximum length to 128 and truncate longer tokenized inputs as most samples will not be affected. The tokenized data is then converted into a BERT input feature consisting of token ids, NER tags, segment mask (token type ids) and attention mask. The input going through BERT and the linear layer results in the NER classification.

All the datasets except for German use the same NER tags encodings. For example, while label 'O' normally corresponds to 0, it corresponds to 8 in the German dataset. Hence, we have to convert the German dataset to the same encodings format.

As not all datasets share the MISC label, we feared that might be an issue for the model evaluation. We solve that problem by always including data of at least one language with this tag in the training set and ignoring that label when evaluating on datasets without it. Therefore, the evaluation on languages without this label may not be perfect, but it should still be good enough and provide us with insight for the purpose of this paper, as it is just one label. We believe that this solution is better than completely removing the MISC label or further limiting the datasets for this research.

2.2.3 Fine-tuning

We follow the settings suggested by the aforementioned related works: We use a batch size of 32 a learning rate of 2e-5 for Adam Optimizer. This learning rate was chosen after comparing the results of 1e-5, 2e-5 and 3e-5 on the English dataset. It was decided to train each model for 3 epochs, after comparing the results of 3 and 5 epochs on the English dataset. Although the results of 5 epochs were about 0.4 higher in terms of F1 score, it took more than 150% of the time of 3 epochs and we chose training time in this trade off. The training of each model took approximately 15 to 30 minutes, depending on the size of the training set.

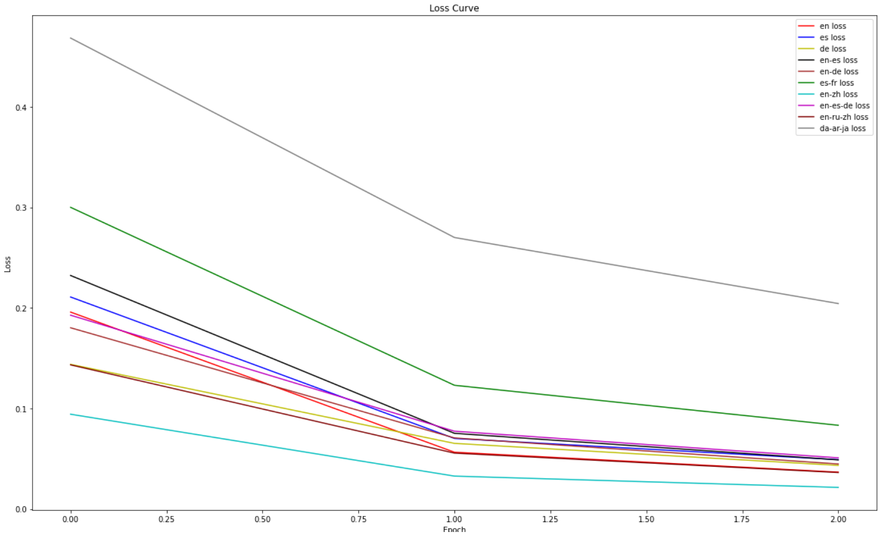


Figure 1: Loss curves of each model

3 Experimental results

Each dataset we use already contains a train and a test set except for the German dataset that we split in a ratio of 4:1 and the Afrikaans dataset, in which we use 15% of the data for testing. Dev sets are not included in the experiment. F1 score, which is commonly used in NER, is the measurement metric used for evaluating the models.

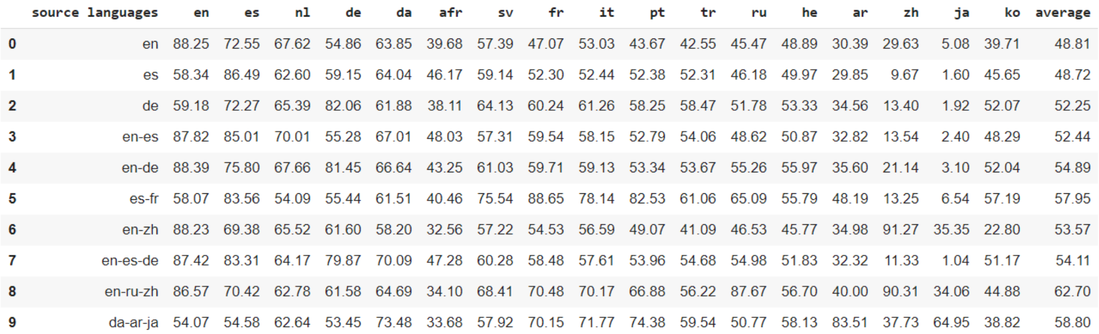
We stick to the notations used by Karthikeyan et al. (2020): We denote Multilingual BERT model trained on language A and B as A-B, e.g., Multilingual BERT trained on English (*en*) and Spanish (*es*) as *en-es*. The languages abbreviations are documented below.

	Language	Abbreviation
1	Afrikaans	afr
2	Arabic	ar
3	Danish	da
4	German	de
5	English	en
6	Spanish	es
7	French	fr
8	Hebrew	he
9	Italian	it
10	Japanese	ja
11	Korean	ko
12	Dutch	nl
13	Portuguese	pt
14	Russian	ru
15	Swedish	sv
16	Turkish	tr
17	Chinese	zh

Table 1: Language Abbreviations

Each model is trained on the train set of the source languages and evaluated on the test set of all the languages included in the research. Monolingual models are trained on the whole train set of the source language. Bilingual models are trained on a new train set that consists of half the samples of each of the source languages train sets randomly chosen and mixed. Trilingual models are trained in a similar manner, only with one third of each source language train set. This procedure was also done in order to avoid overfitting and it is similar to the multilingual training process of Moon et al. (2019). Note that the train sets sizes vary.

The whole process of training and evaluating the models took about 6 hours (15-30 minutes for training and 15 minutes for evaluation of each model); models were trained on GPU (Tesla 4) in Google Colab. The results of the experiment are presented in figure 2- that is, the F1 score achieved by each model on the target languages. The result on each target language corresponds to the result of zero-shot inference if no annotated data in it was shown. Otherwise, it is simply the normal evaluation result on the test set of the source language.



	source languages	en	es	nl	de	da	afr	sv	fr	it	pt	tr	ru	he	ar	zh	ja	ko	average
0	en	88.25	72.55	67.62	54.86	63.85	39.68	57.39	47.07	53.03	43.67	42.55	45.47	48.89	30.39	29.63	5.08	39.71	48.81
1	es	58.34	86.49	62.60	59.15	64.04	46.17	59.14	52.30	52.44	52.38	52.31	46.18	49.97	29.85	9.67	1.60	45.65	48.72
2	de	59.18	72.27	65.39	82.06	61.88	38.11	64.13	60.24	61.26	58.25	58.47	51.78	53.33	34.56	13.40	1.92	52.07	52.25
3	en-es	87.82	85.01	70.01	55.28	67.01	48.03	57.31	59.54	58.15	52.79	54.06	48.62	50.87	32.82	13.54	2.40	48.29	52.44
4	en-de	88.39	75.80	67.66	81.45	66.64	43.25	61.03	59.71	59.13	53.34	53.67	55.26	55.97	35.60	21.14	3.10	52.04	54.89
5	es-fr	58.07	83.56	54.09	55.44	61.51	40.46	75.54	88.65	78.14	82.53	61.06	65.09	55.79	48.19	13.25	6.54	57.19	57.95
6	en-zh	88.23	69.38	65.52	61.60	58.20	32.56	57.22	54.53	56.59	49.07	41.09	46.53	45.77	34.98	91.27	35.35	22.80	53.57
7	en-es-de	87.42	83.31	64.17	79.87	70.09	47.28	60.28	58.48	57.61	53.96	54.68	54.98	51.83	32.32	11.33	1.04	51.17	54.11
8	en-ru-zh	86.57	70.42	62.78	61.58	64.69	34.10	68.41	70.48	70.17	66.88	56.22	87.67	56.70	40.00	90.31	34.06	44.88	62.70
9	da-ar-ja	54.07	54.58	62.64	53.45	73.48	33.68	57.92	70.15	71.77	74.38	59.54	50.77	58.13	83.51	37.73	64.95	38.82	58.80

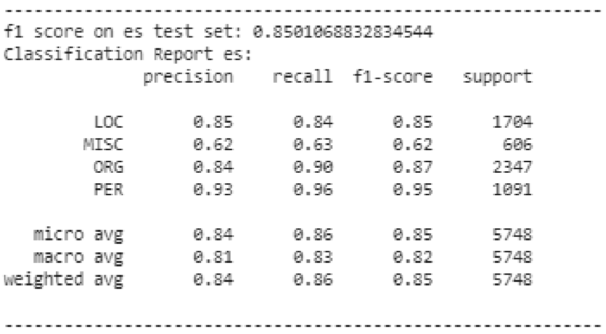
Figure 2: Table showing the F1 score achieved by each model on the target languages, including the average score

We note that the F1 scores of bilingual and trilingual models on their source languages fall short by very little of the scores of the corresponding monolingual models despite having seen only half or third of the data in that language. In the case of English, *en-de* even got a higher score than *en*, meaning the bilingual model surpassed the monolingual one.

Looking at the average F1 scores, we can see that the multilingual models outperform the monolingual ones, with the best three being *en-ru-zh, da-ar-zh* and *es-fr*. As for the first two, we assume the reason might be a better exposure to different language families, scripts and grammatical structures in the fine-tuning stage. Another possible explanation is the higher amount of available training data in our chosen datasets for Chinese and the WikiANN languages.

Most of the results are good and surprising since we expected the performance to be poor when the target and source languages belonged to different language families, and that was not always the case. For example: the results of *da-ar-zh* on Romance languages, the results of *es-fr* on Turkish and Swedish, the fact that most results on Hebrew were in the score range of 50-60, etc.

The most surprising result is how relatively well some of the models perform in zero-shot inference on Korean considering their extremely poor performance on Chinese or Japanese, especially the ones trained on European languages. All models trained on Chinese greatly increased the F1 score of Japanese and vice-versa but that seemed to have no effect on Korean. Although these three languages do not belong to the same language family, they share some similarities due to geographical proximity and we struggle in finding an explanation for that gap. This analysis can be extended to every language included in the research and having a linguist in future works could be beneficial for that purpose.



```
------------------------------------------------------------
f1 score on es test set: 0.8501068832834544
Classification Report es:
              precision    recall  f1-score   support

         LOC       0.85      0.84      0.85      1704
        MISC       0.62      0.63      0.62       606
         ORG       0.84      0.90      0.87      2347
         PER       0.93      0.96      0.95      1091

   micro avg       0.84      0.86      0.85      5748
   macro avg       0.81      0.83      0.82      5748
weighted avg       0.84      0.86      0.85      5748

------------------------------------------------------------
```

Figure 3: *en-es* results on Spanish

Comparing our results to previous works, our monolingual English model received a score of 88.25 on the English CoNLL test set while the state-of-the-art result on this set is 93.5 (Baevski et al, 2019). Moon et al. (2019), whose work is more similar to this one, achieved a score of 91.3 for English, 87.5 on Spanish CoNLL with a monolingual Spanish model and 87.9 (SotA result) with a multilingual model, while we achieved a result of 86.48 on Spanish. This shows that we stand in line with previous works although we didn't replicate SotA results and more could have been done to improve our model. Undoubtedly, the results prove the effectiveness of the cross-lingual transfer of Multilingual BERT for the NER task.

4 Discussion

In this paper we show how the source languages affect the performance of Multilingual BERT in cross-lingual zero-shot inference as well as in predictions on new data in the source languages. We train ten Multilingual BERT models on different sets of languages and test them on 17 languages, some of which do not share the same structure and script, and some belong to different language families.

Our results show that multilingual models outperform the monolingual models or fall short by very little on certain languages. Some of the results were surprising due to the relatively high zero-shot performance some models had on target languages from different language families.

More could have been done to improve the performance of our models. That includes using the development set, training for more epochs or freezing some layers as proposed by previous works. Nonetheless, the effectiveness of the cross-lingual transfer of Multilingual BERT for the NER task was clearly proven and we are satisfied with the results.

5 Code

Link to our code: https://colab.research.google.com/drive/1nbAQNYMCrNKXucV09BqTpdeq_h-1f87Q#scrollTo=nylASdjpWxeG
Link to additional images that we have used: https://colab.research.google.com/drive/1KBudHgQHObfpOW19zqdje2OFb109s3d2#scrollTo=3OrwINl1vzUC

References

- Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition https://www.aclweb.org/anthology/W03-0419.pdf

- Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, Anders Søgaard. DaNE: A Named Entity Resource for Danish, 2020 https://www.aclweb.org/anthology/2020.lrec-1.565.pdf

- Afshin Rahimi, Yuan Li, Trevor Cohn. Massively Multilingual Transfer for NER, 2020 https://arxiv.org/pdf/1902.00193.pdf

- Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019 https://www.aclweb.org/anthology/N19-1423.pdf

- Shijie Wy and Mark Dresde. Beto, bentz, becas: The surprising cross-lingual effectiveness of Bert, 2019 https://www.aclweb.org/anthology/D19-1077.pdf

- Telmo Pires, Eva Schlinger, Dan Garrette. How multilingual is Multilingual BERT?, 2019 https://www.aclweb.org/anthology/P19-1493.pdf

- Antonia Baumann. Multilingual Language Models for Named Entity Recognition in German and English, 2019 https://www.aclweb.org/anthology/R19-2004.pdf

- Sergey Edunov, Alexei Baevski, Michael Auli. Pre-trained Language Model Representations for Language Generation, 2019 https://arxiv.org/pdf/1903.09722.pdf

- Vaswani et al. Attention Is All You Need, 2017 https://arxiv.org/abs/1706.03762

- Gina-Anne Levow. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition, https://faculty.washington.edu/levow/papers/sighan06.pdf

- Taesun Moon, Parul Awasthy, Jian Ni, Radu Florian. Towards Lingua Franca Named Entity Recognition with BERT https://arxiv.org/abs/1912.01389

- Karthikeyan K, Stephen Mayhew, Dan Roth, Zihan Wang. Cross-Lingual Ability of Multilingual BERT: An Empirical Study https://arxiv.org/abs/1912.07840

- Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition https://www.aclweb.org/anthology/W02-2024.pdf

- https://colab.research.google.com/github/huggingface/notebooks/blog/master/examples/token_classification.ipynb#scrollTo=FBiW8UpKIrJW

- https://github.com/google-research/bert/blob/master/multilingual.md

- https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/

- https://medium.com/@yingbiao/ner-with-bert-in-action-936ff275bc73

- https://colab.research.google.com/drive/1Y4o3jh3ZH70tl6mCd76vz_IxX23biCPP#scrollTo=Hba10sXR7Xi6

- https://colab.research.google.com/github/abhimishra91/transformers-tutorials/blob/master/transformers_multi_label_classification.ipynb#scrollTo=Ov1_3R_pAcMo

- https://github.com/huggingface/datasets

- https://huggingface.co/datasets