Cross-lingual Named Entity Recognition

Simon Becker, Michael Kozielski, and Shahram Khadivi

ABSTRACT

Named Entity Recognition (NER) is an important step for many NLP tasks inside eBay. NER systems have to be trained for every language, and even for every category, separately, which is very costly, as data has to be labeled by human annotators. For that reason, currently, only an English NER system is available inside eBay. We show that it is possible to build a German NER system without any German labeled data, by using labeled English data and unlabeled German data in a cross-lingual model.

1 INTRODUCTION AND MOTIVATION

For many Natural Language Processing (NLP) tasks, it is vital to classify words of a text segment into different categories. In previous research, Named-Entity Recognition was used to extract the names of people, locations and organizations. The definition of a Named-Entity and a text segment depend on the task itself. For an e-Commerce website the titles of item listings are an important source of information that can be used to improve search, listings, and translation. From these, we may want to extract attributes like color or material, the brand of the item, and the product name (Figure 1).

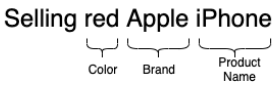

Figure 1: A product title with NER tags

Creating training data for NER is a laborious and expensive task, as examples from each category have to be labeled by human annotators. Additionally, the annotation work has to be repeated for every language that we wish to support. A solution to that is called Cross-lingual Named Entity Recognition, where we use our available labeled data from one language to create a NER model for a different language. For this work, English is our high-resource language, where human-annotated data is available. Additionally, we assume that we have unlabeled data in our target language, which is a realistic assumption for real life scenarios. The target language in our case is German.

2 APPROACH

As our baseline system we will use Multilingual BERT (MBERT). BERT [2] has recently been successfully applied to many NLP tasks, including NER. MBERT (see Figure 2) is a BERT model, which was trained on the combination of more than 100 languages and beside being multilingual, it could also help cross-lingual applications. As our baseline, an MBERT model is trained on labeled English sentences and then evaluated on the German test set.

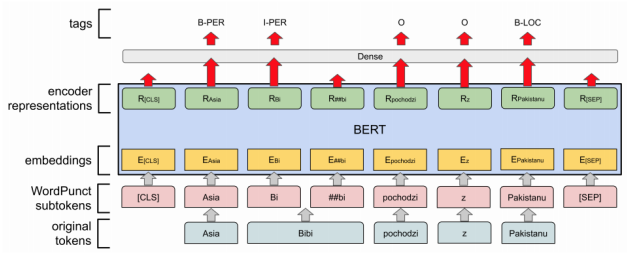

Figure 2: The MBERT architecture [1]

Self-Training

While we only have access to labeled samples from the source language, unlabeled samples from the target language are often available. We can make use of this raw, unlabeled data to further improve our model on the target language. To do this, a method from Semi-Supervised Learning (SSL) is used, called Self-Training or Pseudo Labeling.
The Self-Training algorithm is as follows:

(1) Train a model on source language training data
(2) Predict labels on unlabeled data of the target language
(3) Create a new data set with both original training data and the newly generated pseudo-labeled data
(4) Train a new model on the combined data set
(5) Test the model on the target language

Repeat steps 2-4 until no further improvement is observed.

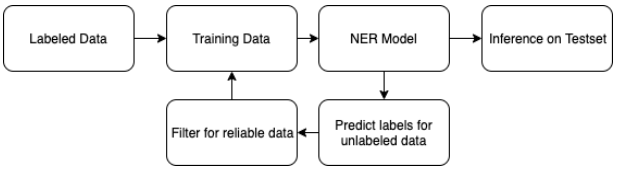

Figure 3: Semi-Supervised Learning Schema

Confidence Filtering

Self-Training can lead the model to become over-confident in its predictions, as it treats its own output as the correct label for training. To overcome this behavior, samples are filtered according to some criterion to avoid adding too many bad

samples to the training set. Commonly, filtering functions make use of the confidence of the model's output. Two simple metrics to assign a confidence value to a sequence of tokens are:

(1) Geometric mean of the probabilities of the tokens
(2) Minimum of the probabilities of the tokens

After assigning a confidence value to each sequence, only sequences with a confidence score above a threshold are added to the training data. This threshold is a hyper-parameter τ. Therefore, Self-Training with filtering involves one more step:

(1) Train a model on source language training data
(2) Predict label on unlabeled data of the target language
(3) Filter the labeled sequences according to a filtering method
(4) Create a new dataset with the original training data and the filtered pseudo-labeled data
(5) Train a new model on the combined dataset
(6) Test the model on the target language

Repeat steps 2-5 until no further improvement is observed.

Label Consistency Filtering

Filtering the newly created samples according to the confidence of the model is difficult. First of all, it requires the proper setting of the threshold parameter. Moreover, using any filtering method that uses the model's own confidence may lead to the model becoming over-confident in its own predictions.

We introduce a new filtering method that is independent of the model's own confidence and also does not need an additional hyper-parameter. The intuition behind this method is that we can reasonably conclude a multilingual NER model is making a correct prediction for a sentence if it would make the same prediction for the translation of the sentence. However, translating a sentence may change the word order, so we have to define a criterion which tells us if the predictions for a sentence and its translation are equal. It is also essential not to filter out too many samples, even if they contain some mistakes. Therefore, we should find a balance between the accuracy and size of the added samples. In our experiments, we discover that the equality of the set of predicted labels in the text and its translation provides the right balance.

3 EXPERIMENTS

As a proof of concept, our datasets come from the ConLL Named Entity Recognition tasks [4], which are the common academic benchmarks for (Cross-lingual) Named Entity Recognition. Specifically, we use the English dataset as our labeled training data and the German dataset as our unlabeled training data. We evaluate our models on the German test set. This setup is called zero-shot, as our models never see any labeled German data, but are evaluated on it. The sentences are taken from newspaper articles and there are four types of named entities, which have to be predicted. The entities to be predicted are organizations, persons, locations and miscellaneous entities. Following the IOB tagging scheme, words are either outside an entity (O), at the beginning of an entity (B), or inside an entity (I).

For the experiments, all models are initialized from an MBERT checkpoint with a zero-shot performance of 70.40 F1 from English to German. The threshold of τ=0.7 was found to be the best performing on the English development set. The translation model was an already existing internal transformer-based [5] model trained on an e-commerce domain.

4 RESULTS

In Table 1 we can see that while all methods outperform the MBERT baseline of 70.40 F1, filtering according to the Label Consistency comes out ahead without any additional hyper-parameter. There were no improvements after three iterations.

Table 1: F1 scores of different filtering methods over 3 runs.

	Iterations		
Method	1	2	3
Mean Confidence τ=0.7	71.93	71.38	71.77
Min Confidence τ=0.7	73.14	74.12	73.52
Label Consistency	**73.20**	**74.65**	**74.88**

Table 2 compares our results with old, and current state-of-the-art methods. While we beat our MBERT baseline, which is the current state-of-the-art for zero-shot NER, it is also interesting to note that the gap between the fully supervised German and our semi-supervised method has been reduced significantly.

Table 2: Comparing our best performing method to other methods.

	Method	EN F1	DE F1
Supervised	Lample et al. 2016 [3]	90.74	78.76
	DeepNER [6]	91.64	79.43
	MBERT (ours)	91	82
Zero-Shot EN ->DE	MBERT		70.40
	+ Label Consistency		74.88

5 CONCLUSION

We have shown that Cross-lingual NER is a promising approach to create NER models in other languages with acceptable level of quality without the need for human annotation. Furthermore, we have shown a simple and stable method to filter for good data outperforming the current state-of-the-art by a large margin.

REFERENCES

[1] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, 89–93. https://doi.org/10.18653/v1/W19-3712

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[3] 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*. Association for Computational Linguistics, San Diego, California, 260–270. https://doi.org/10.18653/v1/N16-1030

[4] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*. Association for Computational Linguistics, USA, 142–147. https://doi.org/10.3115/1119176.1119195

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[6] Yingwei Xin, Ethan Hart, Vibhuti Mahajan, and Jean-David Ruvini. 2018. Learning Better Internal Structure of Words for Sequence Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2584–2593. https://doi.org/10.18653/v1/D18-1279