Minanto at SemEval-2023 Task 2: Fine-tuning XLM-RoBERTa for Named Entity Recognition on English Data

Antonia Höfer and **Mina Mottahedin**
Eberhard Karls Universität Tübingen
Geschwister-Scholl-Platz
72074 Tübingen
antonia-roswitha-sophie.hoefer@student.uni-tuebingen.de
mina.mottahedin@student.uni-tuebingen.de

Abstract

Within the scope of the shared task MultiCoNER II our aim was to improve the recognition of named entities in English. We as team Minanto fine-tuned a cross-lingual model for Named Entity Recognition on English data and achieved an average F1 score of 51.47% in the final submission. We found that a monolingual model works better on English data than a cross-lingual and that the input of external data from earlier Named Entity Recognition tasks provides only minor improvements. In this paper we present our system, discuss our results and analyze the impact of external data.

1 Introduction

Nowadays, language models are capable of solving various problems, amongst others grammar checking, speech (pattern) recognition and translation. But there are still some challenges. The recognition of complex named entities is one of them. Named entities (NEs) occur in various domains and can be, for instance, the name of a person, a city or the title of a book or a movie. As these NEs can be ambiguous like "On the Beach", which can either be a preposition or a movie title, their recognition is more challenging than that of a simple noun. Furthermore, the amount of NEs is not as stable as the vocabulary of a language, but increases fast. The focus of the MultiCoNER shared task (Fetahu et al., 2023b) is to address this challenge in monolingual cases as well as in the multilingual case. The main challenges this year are the fine-grained entity taxonomy with over 30 classes and the induced noise in the test dataset, whereby the task becomes considerably harder.

2 Related Work

The same task was already organized last year at SemEval-2022. Malmasi et al. (2022b) presented SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER), which is a shared task focusing on the identification of complex named entities like media titles, products, and groups in text documents across multiple languages and domains. They also discussed the evaluation metrics used in the task, including the F1 score, and presented baseline results for each language and domain.
For each language in the dataset the participants developed models that can identify the boundaries of complex named entities, classify them into predefined entity types, and recognize nested entities and entity modifiers.
In general, the top-performing systems for each language and domain used neural network models, such as transformers like BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), and ensemble models that combine multiple neural network models and made use of multilingual pre-training or language-specific pre-training to improve the performance of the model. The baseline result for English in MultiCoNER 2022 was a F1 score of 61.36% and the best performing team outperformed the baseline with a score of 91.22%. They used a knowledge retrieval module to retrieve K most relevant paragraphs from a knowledge base (i.e., Wikipedia). These paragraphs were concatenated together with the input, and token representations were passed through a Conditional Random Field to predict the labels. They employed multiple such XLM-RoBERTa models with random seeds and then used a voting strategy to make the final prediction (Malmasi et al., 2022b).

The key challenge of this year's MultiCoNER dataset (Malmasi et al., 2022a) is to deal with complex named entities with limited context.

As explained by Fetahu et al. (2021) Named Entity Recognition (NER) for web queries can be very challenging because of the limited context, especially for code-mixed queries. They found that the

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1127–1130

infusion of multi-lingual gazetteers helps to identify NEs in code-mixed queries more accurately. But, according to Meng et al. (2021), the use of Gazetteer features can also lead to poor generalization. Therefore, the appropriate use of gazetteer knowledge integration is still challenging.

3 Data

The MultiCoNER dataset (Fetahu et al., 2023a) used in the SemEval-2023 Task 2 includes 12 languages (Bangla, German, English, Spanish, Farsi, French, Hindi, Italian, Portuguese, Swedish, Ukrainian, Chinese). Its fine-grained taxonomy contains 33 named entity classes from six categories (Location, Creative work, Group, Person, Product, Medical). The data was collected from three domains: Wikipedia sentences, questions and search queries.

The MultiCoNER dataset was designed to address the NER challenges for complex entities as described before. This task can be completed using either a monolingual track, a multilingual track, or a code-mixed track where the entity is from one language and the rest of the text is written in another language. In addition to that, we also trained our model with external data from the CoNLL 2003 ++ and WNUT 2016 datasets [1]. As the labels in the external data were different than those of the training data, we needed to translate them and lost some named entities, because there was no comparable label to translate it to.

The test data additionally included noisy instances with typing errors in the context tokens and entity tokens.

4 Methodology

In this section, the methodology of our proposed model is presented. It consists of two main parts: the pre-processing of the data and the fine-tuning of the pre-trained XLM-RoBERTa-base model for the NER task. The third part is the training on external data.

4.1 Pre-processing

We converted our data into a standard format for representing labeled sequences of words. Then, the data is processed, such that each sentence is represented as a list of tokens and their corresponding labels. Finally, we obtain a dictionary with keys "tokens" and "tags". The prepared dataset was then passed to the neural network model for training.

4.2 Fine-tuning

To fine-tune XLM-RoBERTa-base for the NER task we made use of the Hugging Face library, which allowed us to perform Named Entity Recognition using pre-trained models (in our case XLM-RoBERTa-base). XLM-RoBERTa developed by Conneau et al. (2020) is a transformer-based language model that extends RoBERTa to multiple languages. The model is pre-trained on large amounts of monolingual and parallel data from multiple languages, allowing it to learn cross-lingual representations. This means that it can be fine-tuned on one language and then applied to other languages without the need for extensive language-specific training data. We chose this model to remain open to the possibility of applying it to other languages. The XLM-RoBERTa-base model can be fine-tuned on a labeled NER dataset in a specific language to create a language-specific NER model. This involves encoding the tokens in the text with the XLM-RoBERTa tokenizer and then feeding the resulting token embeddings into a neural network classifier, such as a linear layer, to predict the named entity labels. Therefore, the input data was tokenized and formatted in the required model format. Then, the model was trained on the training data and its performance was evaluated on the development data. Finally, the predictions for the test set were generated.

We trained the model using a batch size of 16, three training epochs, and the common learning rate of 3.0×10^{-5}. These hyperparameters were chosen based on empirical experiments and hyperparameter tuning to achieve optimal performance on our dataset. For the evaluation we computed precision, recall, and F1 score.

4.3 Training on External Data

As external data led to good performance last year, we tried to improve our model with it. Therefore, we translated the NER tags of the external data, such that it matched the NER tags of the MultiCoNER dataset as shown in Table 1. As we could not assign 'MISC' to any class, we translated it to 'O'. To obtain a consistent format, we removed the unnecessary tags in the CoNLL 2003 ++ data and added _ _ before the IOB tags in both datasets. Then, we added the external data to the training data and trained our model on both of them.

[1] https://metatext.io/datasets-list/ner-task

Dataset	Original tag	New tag
WNUT 2016	facility	Facility
	geo-loc	OtherLOC
	company	PublicCorp
	product	OtherPROD
	movie	VisualWork
	person	OtherPER
	sportsteam	SportsGRP
	musicartist	Artist
	tvshow	VisualWork
	other	O
CoNLL 2003 ++	LOC	OtherLOC
	ORG	PublicCorp
	PER	OtherPER
	MISC	O

Table 1: Translation of NER tags of the WNUT 2016 and CoNLL 2003 ++ data

5 Results

Clean Subset	Noisy Subset	Overall Macro
53.43	47.0	51.47

Table 2: Final submission F1 scores in %

In the final ranking we obtained place 28 out of 34 teams with an overall fine-grained macro-averaged F1 score of 51.47% as shown in Table 2. The best team achieved a score of 85.53%. The noise included in the test dataset could explain why there were no better scores.

Table 3 shows the fine-grained per-class performance. Looking at the F1 scores the most accurately predicted class label is by far *HumanSettlement* with 0.8396 followed by *Athlete* (0.7466) and *Artist* (0.7381). The worst predicted class label is definitely *PrivateCorp* with 0.0197 followed by *Symptom* (0.1561), *AerospaceManufacturer* (0.3205), *Scientist* (0.3473), *ArtWork* (0.3758) and *OtherPROD* (0.3845). Analysing the training dataset we found that the three best performing classes were also the ones that occurred the most in the training set (10.28%, 7.05% and 14.59%). Likewise, *Symptom* (0.79%) and *PrivateCorp* (0.78%) were among the four least occurring classes. Though, this correlation does not always exist. *OtherProd*, for instance, occurs twice as often in the training dataset than *Vehicle*, but was less certainly predicted (F1 score of 38.45% vs. 41.77%). The reason for this phenomena could be that some classes are more specific and less ambiguous than others.

Class	**Precision**	**Recall**	**F1**
Facility	0.6073	0.5886	0.5978
OtherLOC	0.5747	0.3793	0.457
HumanSettlement	0.831	0.8485	0.8396
Station	0.7426	0.6915	0.7162
VisualWork	0.6449	0.62	0.6322
MusicalWork	0.6718	0.6697	0.6707
WrittenWork	0.6165	0.5417	0.5767
ArtWork	0.3886	0.3638	0.3758
Software	0.6324	0.5724	0.6009
MusicalGRP	0.5856	0.541	0.5624
PublicCorp	0.4556	0.5327	0.4912
PrivateCorp	0.0857	0.0111	0.0197
AerospaceManufacturer	0.2567	0.4266	0.3205
SportsGRP	0.7048	0.7653	0.7338
CarManufacturer	0.4664	0.4186	0.4412
ORG	0.5533	0.5161	0.5341
Scientist	0.475	0.2737	0.3473
Artist	0.6858	0.7991	0.7381
Athlete	0.7134	0.7829	0.7466
Politician	0.6116	0.4252	0.5017
Cleric	0.5747	0.3138	0.406
SportsManager	0.6368	0.426	0.5105
OtherPER	0.4194	0.4293	0.4243
Clothing	0.4702	0.4679	0.4691
Vehicle	0.4215	0.414	0.4177
Food	0.4625	0.4683	0.4654
Drink	0.4635	0.4016	0.4303
OtherPROD	0.4309	0.3471	0.3845
Medication/Vaccine	0.6304	0.6514	0.6407
MedicalProcedure	0.5796	0.5322	0.5549
AnatomicalStructure	0.6051	0.606	0.6056
Symptom	0.3865	0.0978	0.1561
Disease	0.5624	0.6852	0.6178

Table 3: Fine-grained performance per class label

Our coarse-grained performance was significantly better with precision of 71.75%, recall of 69.82% and a F1 score of 70.75%. Table 4 shows the coarse-grained performance per class label. Obviously our system performs best on the class *PER* with 0.9019 and worst on class *PROD* with 0.5143. Apparently the name of a person is easier to recognize for our system than that of a location, which is still much better than the other categories.

As mentioned before, we tried to enhance our model with external data. During evaluation on the development dataset, we discovered that its incorporation reduced the F1 score instead of increasing it. As a result, the external data was ultimately not used in the final model. However,

Class	Precision	Recall	F1
Medicine	0.6552	0.6491	0.6521
GRP	0.6911	0.6846	0.6879
LOC	0.8158	0.7963	0.806
PER	0.8962	0.9077	0.9019
CW	0.7052	0.6622	0.683
PROD	0.5417	0.4896	0.5143

Table 4: Coarse-grained performance per class label

afterwards, evaluating on the labeled test data, the training with external data increased the F1 score a little bit by 1.07%. After the translation the external data we used only comprised the classes *OtherPER*, *OtherLOC*, *PublicCorp*, *Facility*, *OtherPROD*, *VisualWork*, *Artist* and *SportsGRP*. Therefore its impact on the fine-grained performance is very limited.

6 Conclusion

Participating at SemEval-2023 Task 2 as team Minanto we fine-tuned XML-RoBERTa-base for Named Entity Recognition on English data. In the final submission we obtained as rank 28 out of 34 teams an overall fine-grained macro-averaged F1 score of 51.47%. After the submission, we realized that our model performed better based on RoBERTa-base than on XLM-RoBERTa-base, because RoBERTa-base is trained on English data only.

Furthermore, we found that the additional training on external data from CoNLL 2003 ++ and WNUT 2016 did not increase the performance much, because these datasets are not as fine-grained as the MultiCoNER dataset. As every external dataset we found had different class labels, we needed to translate them, which led to loss and potential errors in translation. Standardized fine-grained labels would make future work with external data much more easier. Besides, the use of noisy external data could be much more helpful with regard to the noisy test data.

References

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.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

A Appendix

The code for our model can be found here: https://github.com/cicl-iscl/SemEval-Task2.