D2KLab at SemEval-2023 Task 2: Leveraging T-NER to Develop a Fine-Tuned Multilingual Model for Complex Named Entity Recognition

Thibault Ehrhart and **Raphaël Troncy**
EURECOM, Sophia Antipolis, France
thibault.ehrhart@eurcom.fr and raphael.troncy@eurecom.fr

Julien Plu
Buster.ai, Paris, France
plu.julien@gmail.com

Abstract

This paper presents D2KLab's system used for the shared task of "Multilingual Complex Named Entity Recognition (MultiCoNER II)", as part of SemEval 2023 Task 2. The system relies on a fine-tuned transformer based language model for extracting named entities. We present the architecture of the system, and we discuss our results and observations. Our implementation is open sourced at https://github.com/D2KLab/multiconer.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that aims to identify and classify named entities in text. While NER has been extensively studied in monolingual settings and with basic named entity types (Usbeck et al., 2015), the challenges of multilingual NER are significant due to variations in language structure, orthography, and morphology. The SemEval-2023 Task 2, "Multilingual Complex Named Entity Recognition (MultiCoNER II)" (Fetahu et al., 2023b), provides a platform for evaluating systems on this challenging task across 12 languages. The key challenges of this year's task include a fine-grained entity taxonomy with over 30 different classes and simulated errors added to the test set to make the task more realistic and difficult.

In this paper, we present D2KLab's system for MultiCoNER II, which relies on a fine-tuned transformer based language model. We also report our system's results and observations, along with a discussion of the challenges and opportunities for multilingual NER. Our system's performance highlights the importance of fine-tuning on target languages and the effectiveness of transformer-based models for multilingual texts.

2 Related Work

In the field of Named Entity Recognition (NER), researchers have made significant progress in recent years, particularly in addressing the challenges associated with recognizing complex entities and limited context situations. This work builds on the progress made in the Multilingual NER task started in 2022 with the first edition of MultiCoNER (Malmasi et al., 2022b), where the key challenges included dealing with complex entities with limited context (Malmasi et al., 2022a). Meng et al. (2021) outlined the challenges of NER in recognizing complex entities and in low-context situations. Furthermore, Fetahu et al. (2021) extended this work to multilingual and code-mixed settings. This extension was also included in the MultiCoNER dataset, and our work builds on this research.

3 Data

The data collection methods used to compile the dataset used in MultiCoNER is described in (Fetahu et al., 2023a). The dataset comprises annotated sentences for 12 languages, namely *Bangla*, *German*, *English*, *Spanish*, *Farsi*, *French*, *Hindi*, *Italian*, *Portuguese*, *Swedish*, *Ukrainian*, and *Chinese*, along a multilingual track. The data is in the CoNLL-2002 format, which consists of one token per line with tab-separated columns indicating the word, part of speech, and named entity tag. Each sentence is annotated with named entity labels issued from a tagset of 34 fine-grained labels.

We also used a combination of publicly available datasets listed in Table 1 which include TweetNER7 (Ushio et al., 2022), TweeBank NER (Jiang et al., 2022), MIT Restaurant, BioNLP 2004 (Collier and Kim, 2004), WNut2017 (Derczynski et al., 2017), OntoNotes5 (Hovy et al., 2006), BC5CDR (Wei et al., 2016), FIN (Salinas Alvarado et al., 2015), BTC (Derczynski et al., 2016), and ConLL2003 (Tjong Kim Sang and De Meulder, 2003). We selected these datasets based on their popularity and availability, as well as their diversity in terms of domain and language. By training our model on these datasets, we aimed to increase the

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 836–840
July 13-14, 2023

model's ability to recognize named entities across various domains and languages. This approach allowed us to train our model on a large amount of labeled data from various sources, which can improve the model's generalization ability to new and unseen data.

4 Methodology

In this section, we describe the methodology used in our system.

4.1 System Architecture

We used T-NER (Ushio and Camacho-Collados, 2021), an open-source Python library, for fine-tuning a transformer-based language model for named entity recognition. T-NER provides an easy-to-use interface that allows for rapid experimentation with different language models, training data, and evaluation metrics. We fine-tuned our language model on a diverse range of named entity recognition (NER) datasets, including the Multilingual Complex Named Entity Recognition (MultiCoNER) 2022 dataset, as well as other publicly available datasets (see Section 3).

4.2 Experiments

T-NER offers a hyperparameters search approach in order to find the best hyperparameters across a set of given values. By using this feature, we have set up several experiments in order to know how much adding more data can improve a NER model and see until when it stops improving. The set of hyperparameters in T-NER was the same for all the experiments:

- learning rate: $1e^{-4}, 5e^{-4}, 1e^{-5}, 5e^{-5}, 1e^{-6}, 5e^{-6}$
- batch size: 8 , 16 , 32
- CRF: with (1) , without (0)
- gradient accumulation: 1 , 2 , 4
- weight decay: 0 , $1e^{-6}, 1e^{-7}, 1e^{-8}$
- max gradient normalization: 0 , 5 , 10 , 15
- learning rate warmup: 0 , 0.1 , 0.2 , 0.3

The CRF parameter is for using a CRF layer on top of output embedding or not. The selected model for the experiments on English data was DeBERTaV3-large [7]. The reason we have selected this model is because DeBERTaV3 is currently the state-of-the-art encoder model on many downstream tasks[1].

All the experiments have been done on 2 RTX 3090 GPUs. For the first experiments we have started to evaluate, in English, the concatenation of all the datasets cited above. The first run took 4 days to compute for 15 epochs. The best combination of hyperparameters was:

- learning rate: $5e^{-5}$
- batch size: 16
- CRF: with (1)
- gradient accumulation: 1
- weight decay: $1e^{-7}$
- max gradient normalization: 10
- learning rate warmup: 0.1

We reached an average of 84% of F1 with this run over all the test datasets of the datasets we used to train this model. Next, we launch another run to see if it improves with a bigger number of epochs increased to 20. No changes in terms of results compared to the previous run and the set of best hyperparameters stay the same. In order to see how this model behaves compared to a model trained over a single dataset only, we trained one model for each dataset. The final comparison shows that this model improves the results up to 5% of F1 on few of these datasets. Thereafter, we decided to conduct the same experiment in a multilingual context with the Wiki-ANN, MultiNERD and the WikiNeural datasets by selecting only the languages proposed in MulticoNER. The selected model was the multilingual version of DeBERTaV3. We used the values of the best hyperparameters computed during the previous experiment to train this model. Once the model was finished to train we used it as a pre-trained model to fine-tune for our experiments on the MulticoNER 2023 dataset. The final model was trained with the same hyperparameters search values than the pre-trained model and finally the best hyperparameters stay the same as well.

5 Results

Our system generated a model that participated in all MultiCoNER tracks, with macro-averaged F1 being the official ranking metric. Table 2 displays the performance for all tracks in alphabetical order of languages, with the multilingual track presented at the end of the table. The highest average F1 score was achieved in the German track (67.1%). While other tracks have similar scores, we observe that the Farsi and Chinese tracks obtain lower scores (respectively 54.2% and 54.9%). The worst results

[1]https://paperswithcode.com/paper/debertav3-improving-deberta-using-electra

Dataset name	Nb. of entities	Nb. of entity types	Languages	Year
tner/tweetner7 [20]	11,380	7	English	2022
tner/tweebank_ner [9]	3,550	4	English	2022
tner/mit_restaurant	9,181	8	English	2014
tner/wnut2017 [3]	4,691	6	English	2017
tner/bionlp2004 [1]	22,402	5	English	2004
tner/ontonotes5 [8]	76,714	8	English	2006
tner/bc5cdr [23]	16,423	2	English	2016
tner/fin [14]	1,467	4	English	2015
tner/btc [2]	9,339	3	English	2016
tner/conll2003 [17]	20,744	3	English	2003
tner/wikiann [13]	-	3	282 languages	2017
tner/multinerd [16]	13,048	17	9 languages	2022
tner/wikineural [15]	-	16	9 languages	2021

Table 1: List of datasets used for training the model used by D2KLab's system.

are with the English track which has only a F1 score of 42.1%.

Track	F1	P.	R.
BN	0.614	0.590	0.667
DE	**0.671**	**0.642**	**0.715**
EN	0.421	0.400	0.470
ES	0.632	0.617	0.667
FA	0.542	0.517	0.596
FR	0.641	0.635	0.652
HI	0.633	0.610	0.684
IT	0.648	0.638	0.685
PT	0.608	0.592	0.657
SV	0.630	0.610	0.688
UK	0.641	0.620	0.699
ZH	0.549	0.526	0.586
Multi	0.638	0.620	0.664
Avg.	*0.605*	*0.585*	*0.648*

Table 2: Results of D2KLab's system using a multilingual model. The metrics reported are the regular precision, recall and F1.

During the SemEval 2022 Task 2, Wang et al. (Wang et al., 2022) proposed a knowledge-based system for multilingual NER using a multi-stage fine-tuning approach. The first stage refers to training a multilingual model on data from different languages. In the second stage, this fine-tuned multilingual model is used as a starting point for training a monolingual model. AdaSeq introduces two baselines for the task[2]. The first baseline is based on Bert-CRF and uses XLM-R large as the embedding for all languages. The second baseline called RaNER is a variant of Bert-CRF, where the retrieved data act as extra contexts for encoder but are ignored when calculating loss (Wang et al., 2021). We compared the results obtained between these baselines and our system in Table 3. While our system gets close and even outperforms the Bert-CRF solution in both English (+0.61 f1) and French (+2.68 f1), it underperforms compared to the RaNER solution.

Track	Bert-CRF	RaNER	D2KLab
BN	77.06	89.12	61.43
DE	73.17	76.78	67.09
EN	60.68	71.32	61.29
ES	65.04	68.24	63.17
FA	59.40	76.76	54.2
FR	61.41	74.61	64.09
HI	83.80	88.78	63.29
IT	71.12	83.43	64.77
PT	63.94	76.7	60.79
SV	68.4	77.06	62.98
UK	65.71	78.26	64.14
ZH	72.60	75.84	54.92
Avg.	*68.53*	*78.08*	*61.84*

Table 3: Comparison of results between Bert-CRF, RaNER, and D2KLab systems, using the macro-averaged F1 scores as the metric (in %).

[2]https://github.com/modelscope/AdaSeq/tree/master/examples/SemEval2023_MultiCoNER_II

6 Conclusion

In this paper, we presented the D2KLab system which achieved a reasonable performance for the Multilingual Complex Named Entity Recognition task, as part of SemEval 2023 Task 2. The use of a fine-tuned transformer-based language model for extracting named entities proved to be effective, with a macro F1-score of 0.605. However, our results also demonstrated that the system's performance varied across different languages. Our implementation is available at `https://github.com/D2KLab/multiconer`.

References

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. 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 *44th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 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 *17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building NLP models for social media analysis. *CoRR*, abs/2201.07281.

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)*, pages 1412–1437, Seattle, United States. 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 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pages 1499–1512.

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, Vancouver, Canada. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Australasian Language Technology Association Workshop*, pages 84–90, Parramatta, Australia.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. 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. In *Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL: General Entity Annotator Benchmarking Framework. In *24^{th} International Conference on World Wide Web*, pages 1133—-1143. International World Wide Web Conferences Steering Committee.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *16^{th} Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*, pages 53–62. Association for Computational Linguistics.

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco. Barbieri, and Jose Camacho-Collados. 2022. Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *59th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1812, Online. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016.