KDDIE at SemEval-2023 Task 2: External Knowledge Injection for Named Entity Recognition

Caleb Martin, Huichen Yang, and **William Hsu**
Computer Science of Kansas State University
Manhattan, Kansas 66502
{calebjm288, huichen, bhsu}@ksu.edu

Abstract

This paper introduces our system for the SemEval 2023 Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER II) competition. Our team (KDDIE) focused on the sub-task of Named Entity Recognition (NER) for the language of English in the competition and reported our results. To achieve our goal, we utilized transfer learning by fine-tuning pre-trained language models (PLMs) on the competition dataset. Our approach involved combining a BERT-based PLM with external knowledge to provide additional context to the model. In this report, we present our findings and results.

1 Introduction

There is a large and ever-increasing supply of unstructured data present in today's world. Much of this data is in the form of free text. Named Entity Recognition (NER) is the task of labeling named entities within the text. This allows us to gather structured data from the free text. SemEval 2023 Task 2 is a competition in which groups compete to build the best NER system for the provided data (Fetahu et al. (2023b)). This task contains 13 different tracks: one for each of 12 different languages and a multilingual track, which combines all the other languages (Fetahu et al. (2023b)). In this paper, we will focus on Track 1 which is monolingual English NER. The task requires NER systems to identify 36 different labels. The six coarse labels to be used were person, location, group, medical, product, and creative work (Fetahu et al. (2023b)). Each of these coarse-level labels is then split into several fine-grain labels. For example, location is split into facility, human settlement, station, and other locations. Figure 1 shows some examples of words being labeled with their corresponding tags.

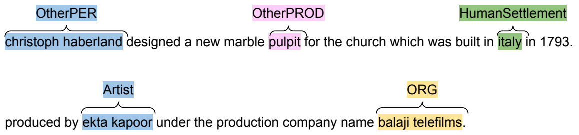
OtherPER
OtherPROD
HumanSettlement
christoph haberland designed a new marble pulpit for the church which was built in italy in 1793.
Artist
ORG
produced by ekta kapoor under the production company name balaji telefilms.

Figure 1: An example of labeling some entities with the SemEval 2023 Task 2 tagging scheme. These two sentences are taken from the training set of SemEval 2023 Task 2 (Fetahu et al. (2023a)).

In this paper, we used external knowledge injection from outside data sources to provide additional context to our PLMs, as described in (Wang et al. (2022)). Our approach involved training multiple transformer models from the HuggingFace library (Wolf et al. (2020)) on the provided training data that was enriched with additional context. We then combined the predictions from several of these models to further improve their performance on the dataset. We fine-tuned several training parameters, including the number of epochs, learning rate, batch size, and other parameters for all the models to achieve the best possible scores.

We have found several key findings through our experiments. Firstly, using the external knowledge injection as outlined in Wang et al. (2022) significantly improved performance even on the new 2023 dataset (Fetahu et al. (2023a)). Additionally, we found that combining the prediction from several different transformer models can lead to better performance overall.

2 Related Work

The work in this competition is a continuation of the research done in the previous multilingual NER task done last year (Malmasi et al. (2022b)). This year's competition is more difficult because there are many more possible labels. Our team also submitted a report (Martin et al. (2022)) to last year's competition where we made use of DeBERTa (He et al. (2021)) for the NER competition. There are many different challenges that make NER difficult. It explained in Meng et al. (2021) that identifying named entities is especially difficult when there is

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1498–1501
July 13-14, 2023 ©2023 Association for Computational Linguistics

not much context or in situations where the entities are exceptionally complex. These issues were key problems that made last year's competition dataset (Malmasi et al., 2022a) so difficult to get good scores on. This problem is partially solved by using external knowledge injection to add more context as done by Wang et al. (2022) which we make use of in the work for this paper. As explained in Li et al. (2020) NER also requires large amounts of well-annotated data. This can be a problem because it can be quite expensive to annotate data.

3 Data

The competition data was given in CoNLL format (Fetahu et al. (2023a)). For the English data, there were 16778 examples for training and 871 examples for validation (Fetahu et al. (2023a)). The data was provided in BIO format. BIO is a labeling scheme in which, if a word is at the beginning of an entity it is labeled B, if the word is inside the entity it is labeled I, if the word is outside of an entity it is labeled O.

We started processing the data by splitting the data in each example and then splitting each example into lists of labels and tokens. Next, we assigned each label a number, 0-72, to represent it. Finally, we create a Hugging Face dataset object from these lists of lists (Wolf et al. (2020)). For our experiments, we used the default train/eval splits which have 16778 examples for training and 871 examples for validation.

For this paper, we used the metrics of macro precision, recall, and f1-score to evaluate our models. To calculate these metrics, we make use of seqeval (Nakayama (2018)) a Python library. Seqeval makes it simple to calculate these metrics, we just need to give it the predicted values and the ground truth. The equations seqeval (Nakayama (2018)) uses to calculate these metrics are shown below with tp meaning true positives, fp meaning false positives, and fn meaning false negatives:

$$Macro\ Precision = \frac{tp}{tp + fp}$$

$$Macro\ Recall = \frac{tp}{tp + fn}$$

$$Macro\ F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4 Methodology

Our main approach for this paper is to fine-tune large feature-based transformer models on the provided training data, using external knowledge to enhance our approach. We then combine the predictions from all models to get the final results. We used the HuggingFace library (Wolf et al. (2020)) to download and train the transformer models. We followed in the work of Wang et al. (2022) to extract external knowledge to provide our models with extra context.

Figure 2 shows a pipeline diagram of our system. The first step is feeding the input sentence into the knowledge injection module, which produces a new sentence with extra context appended on the end. Then as seen in Figure 2 this new sentence is given to three different transformer models to predict an output on. After this is done the output of the three models is then combined to produce a final prediction. Each of these steps is described in more detail in the upcoming sections.

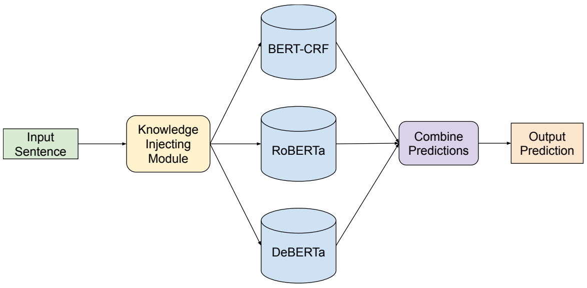

Figure 2: A pipeline diagram of our system implemented for this paper.

4.1 External Knowledge Injection

All models discussed in this paper incorporate external knowledge injection to enhance performance. Following in the work of Wang et al. (2022), we make use of Wikimedia[1] and use Elasticsearch[2] to search them through a large amount of data. The search results are appended on to the original example from the training or test data. Then when this sentence is fed to the transformer model it has more context to work with. As seen in Wang et al. (2022) this significantly improves the performance of the model for the Named Entity Recognition task. For this competition, we experimented with training several different models on this knowledge-enhanced training data.

[1] https://www.wikimedia.org/

[2] https://www.elastic.co

4.2 BERT-CRF

The first model we tried was a BERT model trained on the data provided (Devlin et al. (2018)). BERT was one of the earlier transformer models that a lot of others built upon, but it still performs quite well. Following the research in (Yang and Hsu (2021)), we used a CRF layer to help the BERT model learn the specific parameters of the task, and thus help the performance. To do this we took a BERT model (Devlin et al. (2018)) from HuggingFace (Wolf et al. (2020)) with the CRF layer and trained it for 5 epochs at a learning rate of 2e-5 on the competition training data with the added context.

4.3 RoBERTa

As described in Wang et al. (2022) they used a RoBERTa model to train on their dataset because they believed it to be the best for their situation. In this work, we will also train a RoBERTa model (Liu et al. (2019)) on our knowledge-enhanced training data. In their work Liu et al. (2019), the creators of RoBERTa describe that they improved on BERT by making numerous improvements to the training step such as training longer and on more data. For our final result we trained the RoBERTa model on our knowledge-enhanced data for 10 epochs with a starting learning rate of 2e-5.

4.4 DeBERTa

The last model we used was a DeBERTa pre-trained language model. DeBERTa is also a BERT (Devlin et al. (2018)) based transformer model and it uses enhanced decoding and disentangled attention to improve performance (He et al. (2021)). Within DeBERTa they used a two-vector approach where they split the token encoding and the position encoding into two separate vectors. They used enhanced decoding where they provided the model with both their relative word positions within the sentence and the absolute word positions. These improvements allow DeBERTa to outperform BERT in many different scenarios (He et al. (2021)). For this paper, we took a DeBERTa model and trained it on the SemEval 2023 Task 2 training data with the added external knowledge. We got the best results when training this DeBERTa model for 5 epochs with a learning rate of 2e-5.

4.5 Combining Predictions

Once we had the predictions for all the different models, we designed a system to combine the predictions from all the different models. Essentially this system would average out the predictions from all the models. For example, if you have 3 models and 2 of them predict the tag PERSON and one predicts the tag LOCATION then the final result will have PERSON. If there is a tie, then it also looks at the confidence each model had that it was correct and will use the prediction with the highest combined confidence. Upon using this system, we can see that combining predictions does improve the scores over using each individual model. Each model will have slightly different predictions with strengths in different areas, by combining their predictions we can attempt to get the best of each model.

5 Results

As seen in Table 1, our best scores came from combining results from all 3 of our different models. At an individual level, the DeBERTa model performed the best, followed by the RoBERTa model, and lastly, the BERT-CRF model was the worst.

Model	Precision	Recall	F1
BERT-CRF	0.739	0.770	0.743
RoBERTa	0.774	0.759	0.758
DeBERTa	0.797	0.777	0.781
All 3 Combined	**0.828**	**0.808**	**0.809**

Table 1: Summary of the scores of all the models tested in this paper. All the scores are from testing on the SemEval 2023 Task 2 validation dataset (Fetahu et al. (2023a)).

For the official results on the English track, our team placed 5th out of 34 teams with a macro f1 score of 78.06 on the test dataset. We used the combined results from the 3 different models for that final prediction on the test data.

6 Conclusion

Throughout the competition, we experimented with several different models and found that the DeBERTa model performed the best individually. Aditionally we found that adding the extra context improved the performance of all the models. Moreover, we also discovered that even though the DeBERTa model performed the best on its own, we could improve its scores by combining the predictions with the other models. In future work, we could explore using different data sources for external knowledge injection to determine whether they can further enhance our results.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, 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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

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*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Caleb Martin, Huichen Yang, and William Hsu. 2022. KDDIE at SemEval-2022 task 11: Using DeBERTa for named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1531–1535, 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

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 *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

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.

Huichen Yang and William Hsu. 2021. Named entity recognition from synthesis procedural text in materials science domain with attention-based approach. In *SDU@ AAAI*.