IITD at SemEval-2023 Task 2: A Multi-Stage Information Retrieval Approach for Fine-Grained Named Entity Recognition

Shivani Choudhary, Niladri Chatterjee, Subir Kumar Saha

{shivani@sire, niladri@maths, saha@mech}.iitd.ac.in Indian Institute of Technology Delhi Hauz Khas, Delhi-110016, India

Abstract

MultiCoNER-II is a fine-grained Named Entity Recognition (NER) task that aims to identify ambiguous and complex named entities in multiple languages, with a small amount of contextual information available. To address this task, we propose a multi-stage information retrieval (IR) pipeline that improves the performance of language models for fine-grained NER. Our approach involves leveraging a combination of a BM25-based IR model and a language model to retrieve relevant passages from a corpus. These passages are then used to train a model that utilizes a weighted average of losses. The prediction is generated by a decoder stack that includes a projection layer and conditional random field. To demonstrate the effectiveness of our approach, we participated in the English track of the MultiCoNER-II competition. Our approach yielded promising results, which we validated through detailed analysis.

1 Introduction

Named entity recognition (NER) is a crucial task in natural language processing (NLP), where the goal is to identify and classify named entities in text, such as people, organizations, locations, and products. NER has a wide range of applications, including information extraction, question answering, and sentiment analysis, among others.

Recently, there has been a growing interest in the development of NER models that can handle ambiguous, complex, and emerging named entities with shorter context. In response to this, the Multi-CoNER challenge was introduced, which aims to evaluate NER models on a diverse range of named entity classes and contexts.

The first iteration of the MultiCoNER challenge, MultiCoNER-I, focused on classification into six classes: *PER, LOC, GRP, CORP, PROD, CW* (Malmasi et al., 2022a). However, the performance of NER systems on complex entities such as creative works, groups, and products was found to be lesser than on entities such as Person (Malmasi et al., 2022b). In response to this, the second iteration of the MultiCoNER challenge, MultiCoNER-II, was introduced, which aimed to address this issue by providing a fine-grained NER task with 36 different classes(Fetahu et al., 2023a).

To make the task more realistic, the organizers added noise to the test dataset. Additionally, a new coarse-grained class, Medical, was introduced in MultiCoNER-II, and the CORP class was subsumed into GRP. The coarse-grained classes were further subdivided into fine-grained entities. MultiCoNER-II consists of 13 different tracks, including 11 monolingual tracks for languages such as English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla, as well as a code-mixed track and a multilingual track(Fetahu et al., 2023b).

To tackle the MultiCoNER-II challenge, we propose an NER system based on the concepts of a single-stage information retrieval (IR) framework. Our information retrieval pipeline uses two document corpora, KILT and MS-MARCO, to retrieve relevant documents. We rely on the BM25 scoring method for document retrieval, and retrieve documents are chunked into small passages so that we can capture the relevant part of the document.

To increase the confidence of the prediction, we have used an ensemble of focal loss and conditional random field (CRF) loss. Our system also incorporates additional contextual information into the sentences to improve the performance of lowperforming classes. Through experiments and evaluations, we demonstrate the effectiveness of our approach in handling the MultiCoNER-II challenge.

2 Related Work

Named Entity Recognition is one of the core tasks of NLP. It powers different real-world applications such as chatbots, and document parsers to identify and extract key information. NER is applied in different domains e.g. medical (Song et al., 2015; Ramachandran and Arutchelvan, 2021), e-commerce(Luo et al., 2020; Fetahu et al., 2021) and sentiment analysis (Batra and Rao, 2010) to name a few.

In the MultiCoNER-I, the dataset for the task has sentences comprised of complex and emerging entities with smaller context (Malmasi et al., 2022a). Challenges associated with the low-context NER and complex entities are outlined by (Meng et al., 2021). They have suggested an approach to use the gazetteer-based information for NER. Document level context all improves the performance of NER system (Yamada et al., 2020; Luoma and Pyysalo, 2020; Chen et al., 2022)

Transformer-based models are the cornerstone of NER. To effectively use transformer-based models, researchers have employed transfer learning. In MultiCoNER-I, teams have relied on transformer (Vaswani et al., 2017) based models - BERT (Devlin et al., 2019), LUKE(Yamada et al., 2020) and RoBERTa (Liu et al., 2019) as their base encoders. In order to enrich low-context sentences with contextual information, participating teams has relied on knowledge base (Wang et al., 2022), gazetteer (Chen et al., 2022) based network to infuse additional information to the model. Almost all the participating teams beat the official baseline model. However, NER systems have produced a lower score on complex entities like CW, and GRP (Malmasi et al., 2022b). It shows the complexity associated with the task.

3 Data

MultiCoNER-II is a named entity recognition (NER) task that focuses on fine-grained classification of 36 different classes. The dataset used for this task is MultiCoNER V2, which was introduced in a paper by Fetahu et al.. The dataset follows the IOB format for annotating entities, and the distribution of instances for English is presented in Table 1. The distribution of B-tag is presented in Figure 1. The train-test ratio for this dataset is approxi-

	Number of instances
Dev set	871
Train set	16778
Test set	249980

Table 1: Details of number of records in different set

mately 1:15, with a large test set that is designed to



Figure 1: Distribution of class labels in train dataset

test the generalization capability of the model. To simulate real-world scenarios, approximately 30% of the test data contains noise that is added to either the context or the entity tokens.

In addition to the MultiCoNER V2 dataset, we have used two other datasets, MS-MARCO (Payal Bajaj, 2016) and KILT (Petroni et al., 2021), for our information retrieval pipeline. MS-MARCO is a dataset of search queries and passages, and the we use the 2019-20 dump of MS-MARCO for our system. KILT is a benchmark for knowledge-intensive NLP tasks, and the we use it for our information retrieval pipeline.

4 System Description

The goal of our system is to use the external context that is retrieved through an information retrieval pipeline. To describe our paper, we will formally define our information retrieval (IR) pipeline and named entity recognition (NER) system. We will then provide details about the retrieval framework and NER system.

Our system relies on the IR pipeline to provide a list of the most relevant passages for a given query. Specifically, given a query \mathbf{Q} , the system retrieves top-k documents $\{d_1, d_2, ..., d_k\}$. Each document is then segmented into passages of length L and ranked in terms of their relevance to the query \mathbf{Q} . The IR system generates a ranked list of passages $\{p1, p2, ..., pr\}$ for each query \mathbf{Q} . NER tagging involves predicting one of the tags for each of the tokens in a given sentence \mathbf{S} , with words $\{w_1, w_2, ..., w_n\}$ and set of tags $\{t_1, t_2, ..., t_n\}$.

In our model, the query ${\bf Q}$ and the top-n $\{p1,p2,...,pn\}$ passages are passed through a

transformer model to generate contextual embeddings for the query. The contextual embeddings, along with an appropriate token mask, are then passed to a CRF layer (Lafferty et al., 2001) to learn transition probabilities for a BIO tagging scheme. The schematic diagram of our system is shown in Figure-2.

4.1 Document Indexing and Retrieval

The IR pipeline is based on the pipeline presented in TREC CAST-21 (Choudhary, 2022). The original proposed pipeline for information retrieval (IR) was a multi-stage pipeline designed to memorize relevant keywords in a conversation and perform multi-stage retrieval and ranking. However, the current implementation is a single-stage retrieval pipeline that aims to evaluate the efficacy of additional context from an IR perspective. As such, it does not store relevant keywords and does not use pseudo-relevance feedback to perform a secondstage retrieval.

The inverted index for the document corpus comprising MS-MARCO (2019/20 dump) (Payal Bajaj, 2016) and KILT (Petroni et al., 2021) was built using the Pyserini¹ toolkit, with an option to keep a copy of raw documents. The passages generated after document retrieval from the corpus were re-ranked using a T5-based pre-trained re-ranking model (Nogueira et al., 2020) and a python wrapper named pyterrier (Macdonald and Tonellotto, 2020). The T5 model was fine-tuned on MS-MARCO dataset. The model is trained to generate target token ("true" or "false") that indicates whether a given query Q is relevant to a document D. During the inference phase, a candidate query Q_c compared against a set of retrieved documents $\{d_1, d_2, ..., d_k\}$ and , and the model computes the probability of the "true" and "false" labels for each document in the set. Probability is generated using logits generated from the first decoding step.

4.2 Named entity recognition stack

The NER stack has three components: an encoder stack, a linear layer, and a CRF layer with focal loss. The transformer-based encoder stack generates a representation in R^{TXH} , where T is the number of tokens in the sentence and H is the dimension of final hidden state of the transformer model. The linear layer projects the H dimension embedding to a 73 dimension vector. It represents

the classes as per BIO scheme. The log-softmax score is further passed to linear chain CRF layer (Huang et al., 2015) with a mask to ignore the tokens associated with added context. To train the encoder stack, we used an ensemble of losses. In addition to CRF loss, we used focal loss in training stage to accommodate class imbalance issue in the train data. Focal loss (FL) is a modified variant of cross-entropy loss (Lin et al., 2017). There are two parameters associated with FL — α and γ . α handles the class imbalance issue by generating a weighted loss and γ controls the loss generated due to incorrectly classified samples.

$$FL(t_i) = -\alpha * (1 - p_{t_i})^{\gamma} \log p_{t_i} \tag{1}$$

4.3 Training and Evaluation

In the first stage of the training, we have extracted top-5 documents from the corpus. Top-5 documents are chunked into small passages using Spacy². Compute and time requirements associated with the re-ranking framework is very high. For official submission, we have sampled 5 passages from the list of chunked passages for every query (sentences as per in data set) for re-ranking. We have selected top-2 relevant passage to add the contextual information.

In the second stage of training, we combine each sentence with the contextual information retrieved in the first step. Transformer models have a limitation on the maximum sequence length which can be processed by the model. To ascertain that the length of input is less then the maximum length, we tokenized the context and selected the number of tokens as per the below equation.

$$\min(CL, (CL - LS))$$

where CL is context length LS is length of tokens in original sentence. We have also generated an addition mask for the CRF layer. Custom mask helps CRF to mask token scores generated for retrieved context from the first step. Custom mask masks the additional context from CRF loss calculation. The model is trained with a weighted combination of focal loss and CRF loss. It should be noted that focal loss is not calculated without the mask. In our system, we have used the focal loss implementation by Kornia³ and CRF implementation by allennlp ⁴. The hyperparameter space used for training the

²https://spacy.io/

³https://github.com/kornia/kornia

⁴https://github.com/allenai/allennlp/tree/master/allennlp



Figure 2: Fine grained NER pipeline. First block is based on information retrieval where corpus index is build using Pyserini, followed by transformer based encoder and CRF layer. Training loss is calculated using a weighted average of Focal loss (Lin et al., 2017) and CRF loss

model is presented in Table 2

The model is trained with warmup schedule followed by a constant learning rate. Different components of NER stack is trained with different learning rate. Transformer model is trained with default learning while CRF and Linear layer is trained with a 100 times of the default learning rate. The model is trained with an early stopping criteria and objective to minimize validation loss.

Base encoder	BERT-large-uncased,	
	BERT-base-uncased	
Dropout rate	0.4 , 0.5, 0.6	
Context length	256	
Focal loss weightage	0.6 , 0.4, 0.0	
Learning rate	1e-5 , 2e-5, 5e-5	

Table 2: Hyper-parameter space used for tuning. Items marked in bold are the parameter for the best performing official submission

During evaluation the additional context generation for test set followed the same procedure as train set.

5 Results and Discussion

5.1 Official submission

According to the official leaderboard, the proposed system ranked 15th in the English track of the MultiCoNER-II competition. The best performing submission metrics are presented in Table 3, and the performance of the system on both cleaned and noisy sets is shown in Table 4.

The document retrieval system utilized the BM25 score to select relevant documents for the proposed system. However, BM25 is a variant of the bag-of-words model, and its effectiveness is

	Fine-grained	Corase-grained
macro@F1	0.6321	0.772
macro@P	0.6427	0.7712
macro@R	0.6289	0.7735

Table 3: Official leader board submissions with bestmacro-F1 score

	Clean set	Noisy set
macro@F1	0.6752	0.5359

Table 4: Macro@F1 score on clean and noisy set.

dependent on the presence or absence of words in the document. If a word is corrupted, it can impact the quality of the retrieved documents, leading to inaccurate or irrelevant results. For example, *it was described by george hmpson [hampson] in 1996.* There are fewer anchors to direct the retrieval system to point towards the relevant records .This limitation is highlighted in the results, where the model's performance on the noisy set was significantly lower than on the cleaned subset of the test data. However, some sentences with noisy data had enough information (keywords) to outweigh the impact of the noise, leading to the retrieval of relevant documents despite the presence of cor-

	LR	DR	F1
BERT-Large	5e-5	0.5	0.6312
BERT-Large†	1e-5	0.4	0.6321
BERT-base	5e-5	0.4	0.6047
BERT-base	1e-5	0.4	0.6123

Table 5: F1-Score from other submissions, All the submissions are using uncased BERT and additional context of length 256, † marks the leader board submission

rupt words. For example, the regular cast was sid **jjames [james]** peggy mount john le mesurier and keith marsh.

The retrieval system also struggled with complex cases in the test data that were not present in the training data. This can be attributed to the lack of training data containing corrupted sentences. Additionally, the computational cost of the system limited the number of documents selected for reranking. This limitation may have affected the model's ability to perform well on the test data.

	Р	R	F1
HumanSettlement	0.854	0.8968	0.8749
SportsGrp	0.8282	0.8429	0.8355
Scientist	0.5631	0.3297	0.4159
PrivateCorp	0.3276	0.4235	0.3694

Table 6: Best and worst performing classes in the official run

As shown in Table 6, the top and bottom two classes had varying levels of precision. Private-Corp and Symptom classes had particularly poor precision. The B-tag confusion matrix revealed that the model classified ORG as both PrivateCorp and PublicCorp, and PublicCorp was also misclassified as PrivateCorp. This confusion between classes with similar semantic features, such as Scientist vs OtherPER, SportsManager vs Athlete, indicates that the model may not have learned to distinguish between them effectively.

5.2 Efficacy of retrieved context

We have analyzed the impact of context on the NER performance. We used BERT-large + CRF stack as our baseline model for comparison. There is a performance gain of 0.0877 over the baseline run.

	F1
Baseline model	0.6129
Baseline+RC+FL(official submission)	0.7003
Δ	0.0877

Table 7: Comparison between the performance of baseline model and our official submission on **validation** set. **RC** means retrieved context and **FL** represents focal loss

5.3 Ablation study

We further analysed the performance of the model with respect to different strategies for focal loss and context length from the passages.

5.3.1 Variation in α in focal loss

We analyzed the impact of weights associated with each of the class labels. We considered B tag and I tag as separate classes. We have considered constant class weight, class weights are inversely proportional to the frequency of the respective classes and softmax over class weight. We observed that there is an increase in micro@F1 score by changing over to variable weight scheme, as presented in Table 8. However, the gain is not significantly higher between softmax and inverse weighting schemes.

	F1
Baseline+RC+FLC	0.7003
Baseline+RC+FL _{inv_weight}	0.714
Baseline+RC+FL _{softmax_weight}	0.718

Table 8: Comparison of different weight scheme for α and added a context of length 256

5.3.2 Effect of passage re-ranking and context length

We evaluated the impact of the number of passages sampled for the re-ranking. We re-ranked 15 passages compared to 5 in the official submission. We have also increased the context length to 384. We wanted to analyze the effect of the quality of the ranked passages and their content. We were not able to run it with full 512 context length due to computing constraints. Comparison of F1 score is presented in Table 9

	F1
Baseline+RC(256)+FL	0.7003
Baseline+RC(384)	0.6975
Baseline+RC(384)+FL _{softmax_weight}	0.7218

Table 9: Comparison of macro@F1 score with context length 384 and 256.

6 Conclusion

This paper presents an information retrieval based pipeline for fine-grained NER. The training data has a class imbalance issue. In order to handle the class imbalance, we have introduced focal loss. The model has generated a good performance on clean subset of test data. However, it was not able to produce similar performance on noisy data. The ablation study points out that the model was able to leverage additional context. It also points out that focal loss has enhanced the performance of the model. The use of focal loss can be explored further with masking of context tokens. The retrieval pipeline can be explored further with a query reformulation step to handle the short context and noise in sentences.

References

- Siddharth Batra and Deepak Rao. 2010. Entity based sentiment analysis on twitter. *Science*, 9(4):1–12.
- Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition.
- Shivani Choudhary. 2022. Iitd-dbai: Multi-stage retrieval with pseudo-relevance feedback and query reformulation. *ArXiv*, abs/2203.17042.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805.
- Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. Multi-CoNER v2: a Large Multilingual dataset for Finegrained 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.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.
- John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2999–3007.

- 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, abs/1907.11692.
- Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghang Wu, Qiang Li, Keping Yang, and Kenny Q. Zhu. 2020. Alicoco: Alibaba e-commerce cognitive concept net. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, SIGMOD '20, page 313–327, New York, NY, USA. Association for Computing Machinery.
- Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In Proceedings of the 28th International Conference on Computational Linguistics, pages 904– 914, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*.
- 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.
- Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings* of the Association for Computational Linguistics: *EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.
- Nick Craswell Li Deng Jianfeng Gao Xiaodong Liu Rangan Majumder Andrew McNamara Bhaskar Mitra Tri Nguyen Mir Rosenberg Xia Song Alina Stoica Saurabh Tiwary Tong Wang Payal Bajaj, Daniel Campos. 2016. Ms marco: A human generated machine reading comprehension dataset. In *InCoCo@NIPS*.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian

Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

- R. Ramachandran and K. Arutchelvan. 2021. Named entity recognition on bio-medical literature documents using hybrid based approach. *Journal of Ambient Intelligence and Humanized Computing*, pages 1 – 10.
- Min Song, Hwanjo Yu, and Wook-Shin Han. 2015. Developing a hybrid dictionary-based bio-entity recognition technique. *BMC Medical Informatics and Decision Making*, 15:S9 – S9.
- 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.
- 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.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Conference on Empirical Methods in Natural Language Processing*.