NamedEntityRangers at SemEval-2022 Task 11: Transformer-based Approaches for Multilingual Complex Named Entity Recognition

Amina Miftahova
Innopolis University, Russia
noteisenheim@gmail.com

Alexander Pugachev
HSE University, Russia
avpugachev@edu.hse.ru

Artem Skiba
Ershov Institute of
Informatics Systems, Russia
efildream@gmail.com

Ekaterina Artemova
HSE University
Huawei Noah's Ark lab, Russia
elartemova@hse.ru

Tatiana Batura
Moscow State University
Novosibirsk State University, Russia
t.batura@g.nsu.ru

Pavel Braslavski
HSE University
Ural Federal University, Russia
pbras@yandex.ru

Vladimir Ivanov
Innopolis University
Kazan Federal University, Russia
v.ivanov@innopolis.ru

Abstract

This paper presents the two submissions of NamedEntityRangers Team to the MultiCoNER Shared Task, hosted at SemEval-2022. We evaluate two state-of-the-art approaches, of which both utilize pre-trained multi-lingual language models differently. The first approach follows the token classification schema, in which each token is assigned with a tag. The second approach follows a recent template-free paradigm (Ma et al., 2021), in which an encoder-decoder model translates the input sequence of words to a special output, encoding named entities with predefined labels. We utilize RemBERT and mT5 as backbone models for these two approaches, respectively. Our results show that the oldie but goodie token classification outperforms the template-free method by a wide margin. Our code is available at: https://github.com/Abiks/MultiCoNER.

1 Introduction

This paper describes two submissions to the Multilingual Complex Named Entity Recognition (MultiCoNER) Shared Task, held by SemEval-2022 (Malmasi et al., 2022b). This shared task aims at recognizing named entities with the ambitious end goal of building systems that support up to 11 languages. Multilingual setups are complicated when the dataset mixes languages from different groups. This way, the MultiCoNER dataset comprises 11 languages from different families, and multiple scripts (Malmasi et al., 2022a). This setup hayj s become increasingly popular recently since it allows to test for transfer learning across languages within a single pre-trained model. The dataset, proposed in the shared task, has several unique features previously neglected in NER evaluation: syntactically complex entities, ambiguous entities, divers, and low-frequency (aka long-tail) entities. This makes the shared task setup closer to real-life settings, where datasets are way noisier and nonuniformly distributed.

Our solution consists of two state-of-the-art approaches adopted to the task. First, we use a mainstream NER technique, token classification. Under this approach, the model is trained to assign a label to each input token. The second approach falls into the group of prompt-based techniques, at the core of which are the capabilities of auto-regressive language models to memorize and reproduce input texts. In this case, we train an encoder-decoder model to replace entities in the input sentence with predefined labels. We utilize RemBERT (Chung et al., 2021) and mT5 (Xue et al., 2021) as backbone models for these two approaches, respectively. These two models provide state-of-the-art results for the common test-beds of cross-lingual experiments (Hu et al., 2020).

In other words, we explore the following questions: 'How do pre-trained transformer-based models perform in the Multilingual Complex Named Entity Recognition task?' and 'Which of the two approaches perform better?'.

Our results show that plain fine-tuning of the above-mentioned state-of-the-art multilingual transformer-based models can give moderate re-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1570 - 1575

sults. Results are within 10% margin of the top solution for the multilingual NER task. We analyzed the errors of both models. As expected, for two novel 'complex' entity types ('Creative Work' and 'Product') error rate is higher than for 'common' types ('Group', 'Location', 'Person'). However, some of the errors might be caused by inconsistencies in the labeling of the MultiCoNER dataset. The performance of the Template-free approach suffers from such inconsistencies more than the performance of the Token classification approach.

2 Related Work

Named Entity Recognition is one of the central tasks in Natural Language Processing, which attracts a lot of research efforts. Yang et al. (2016) encode morphology and context information via character and word embeddings. Recent studies (Ghaddar and Langlais, 2018; Jie and Lu, 2019; Liu et al., 2019; Meng et al., 2021) employ syntactic dependencies, lexical similarity, gazetteers, etc. in the word representations before feeding them to context encoding layers. The authors show that additional information may lead to improvements in NER performance. However, NER still faces multiple challenges (Li et al., 2022) such as detection of fine-grained and nested named entities (Kim and Kim, 2021; Ringland et al., 2019; Loukachevitch et al., 2021), NER in domain-specific areas (Weber et al., 2021), NER from noisy data (Derczynski et al., 2017) and code-mixed data (Fetahu et al., 2021).

Recent research in this area considers not only standard types of entities (*person, location, organization*) but also semantically ambiguous and complex entities (Hanselowski et al., 2018). For example, a system has to recognize the titles of movies, books, or songs, which may contain verbs, adverbs, prepositions, etc. Cui et al. (2021) propose a template-based method, treating NER as a language model ranking problem in a seq2seq manner. Original sentences and statement templates filled by a candidate named entity span are the source sequence and the target sequence, respectively. Ma et al. (2021) induces a language model to predict label words at entity positions during fine-tuning. This method demonstrates the effectiveness under the few-shot setting.

The SemEval-2020 shared task MultiCoNER (Malmasi et al., 2022a) focuses on a more exciting and challenging problem of building a NER system for multiple languages. The results of a recent Multilingual Named Entity Challenge in six Slavic languages (Piskorski et al., 2021) have also confirmed the complexity and significance of the task. Training competitive multilingual NER systems requires either manually labeled text collections or large automatically annotated datasets (Nothman et al., 2013).

3 Experiments

3.1 Token Classification Approach

Our baseline is to treat the NER task as a token classification problem. The base model is a pre-trained RemBERT (Chung et al., 2021) with a linear layer on top of the hidden-states output.

RemBERT is based on a multilingual BERT bidirectional transformer architecture. It uses decoupled embeddings, which allows changing the size of input and output embeddings. The input embeddings are reduced in size, thus making the fine-tuning process faster without performance loss compared to BERT.

In the first experiment on a tokenization step the original label is propagated to all of the word tokens. We fine-tuned the model on multilingual data for three epochs to predict the labels in BIO format. The batch size is 32, Adam optimizer is used with the learning rate of 10^{-5} and the scheduler decreases the learning rate by 0.1 each epoch. The metrics obtained on the development set for this approach are presented in Table 1.

In the following experiment we investigated how the performance changes if we combine the models into an ensemble. The ensemble consists of three models trained with different seeds as described above. The final predictions are made based on a hard or soft voting scheme. The models perform very similarly; the differences in evaluation scores are insignificant.

The confusion matrix in Figure 1 is built for the multilingual model. The largest ratio of erroneously assigned O labels is PROD (Product) or CW (Creative work). Almost for all individual languages, the confusion matrix is very similar to the aggregated one. The picture differs a bit for Farsi and Russian languages: the number of mislabelled entities as O is higher for each entity tag. This might be due to the difference in the language structure. For the Chinese language, 23% of the GRP (Group) entities were classified as CORP (Corporation). This behavior is unique to the Chinese

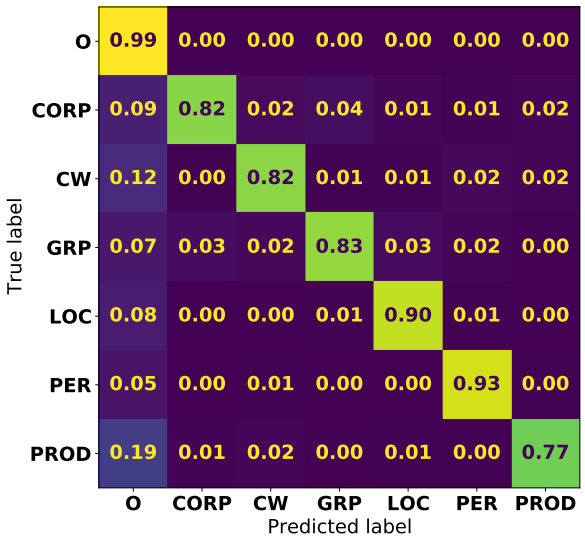

Figure 1: Confusion matrix for fine-tuned RemBERT

language.

The ensemble model does not introduce much improvement. However, training models with different seed values converge almost to the exact predictions. On the dev set, only one model's prediction out of three base models differs in 10% of the cases, and all three models have different predictions in 1% of the cases.

3.2 Template-free Approach

The second approach which we considered was Template-free (Ma et al., 2021). This approach showed decent results on CoNLL03 (Sang and Meulder, 2003) and MIT-Movie (Liu et al., 2013) datasets, so we decided to apply it to the multilingual data. The language model is trained to substitute named entity text spans with several predefined label words. For instance, given the sentence "*its headquarters are in sandy springs, united states of america*" we require the model to replace "*sandy springs*" and "*united states of america*" text spans with a predefined label word "*germany*" since these spans are considered as LOC (location) entities. In this case, the target for this example would be: "*its headquarters are in germany, germany*". After the model has substituted several text spans with the label words, we need to reconstruct these spans based on surrounding tokens to match each initial token with its predicted label. The label word for a specific entity class was chosen as the most frequent token of this class in the training data. Since we were working with multilingual data, we created a unique label words mapping for each language.

As the backbone model for the Template-free approach, we considered the mT5 pre-trained language model (Xue et al., 2021). mT5 is a multilingual variant of the T5 model that is pre-trained on a new Common Crawl-based dataset covering 101 languages. We chose two variants of the mT5 model for our experiments: mT5-Large and mT5-XL; the latter model showed better results. Due to the computational complexity, we could not run similar experiments with the mT5-XXL language model. Both models were trained with batch size equal to 8 and optimizer AdamW with learning rate $5 \cdot 10^{-5}$.

During the experiments with the Template-free approach, we encountered several challenges. Firstly, if the input phrase has two or more consecutive entity spans with a token length of more than one, it was impossible to reconstruct these spans unambiguously based on the sequence of predicted label words. In this case, we assigned the first $k-1$ predicted label words to $k-1$ input tokens, and the last label word was assigned to the rest of the tokens. Secondly, because of the punctuation issues occurring in some languages described in more detail in Section 4, it was difficult to generalize the reconstruction rules for all languages.

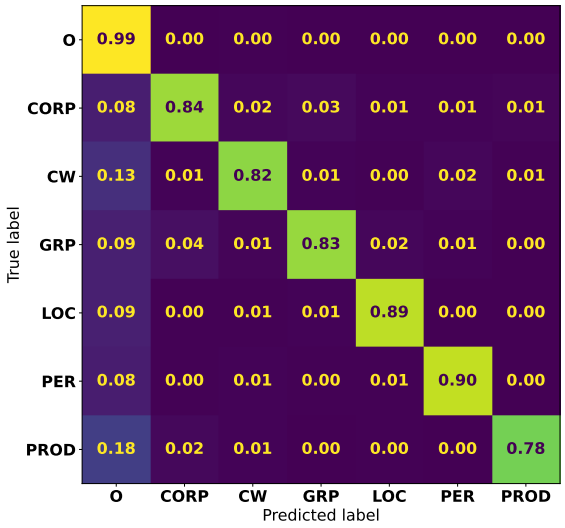

Figure 2: Confusion matrix for Template-free mT5-XL

The token-level confusion matrix for the Template-free approach on the development set is shown on Figure 2. The results for the mT5-XL fine-tuned model on the dev set is presented in Table 1.

Based on Figure 2 we can mention that the most

Table 1: Results for Token Classification and Template-free approaches on dev and test sets

Approach	Dev			Test		
	P	R	F1	P	R	F1
Token Classification	0.81	**0.84**	**0.82**	0.83	0.80	0.81
Template-free	**0.82**	0.77	0.79	0.58	0.51	0.54

common model's error is predicting the O-tag for tokens which are actually considered as parts of the entities. In other words, the model more often fails to recognize a named entity than confuses two different entities. We can also mention that in terms of languages the fine-tuned model shows the best results for English, Dutch and German and the worst results for Farsi and Russian according to F1 score. The dramatic performance downgrade on the test set compared to the dev set for the Template-free model may occur due to the significant distribution shift in the test data. For instance, the average sentence length in terms of tokens in the train and dev data is approximately equal to 16.4. However, the average length in the test data is equal to 9.6, we assume that the Template-free model could not be resistant for such changes.

4 Discussion

We provide the performance analysis of the models that we developed. The confusion matrices for both approaches demonstrate similar commonly occurring errors. Moreover, while working on experiments, we noticed several dataset issues which we suppose are worth mentioning. The problems potentially contribute to the low performance of the Template-free approach.

1. Classification decisions for tokens from PROD and CW entity types are more often confused with the O labeled tokens (Fig. 1, 2). These entities are examples of complex entities, which the competition was focused on. At the same time, the labeling for the two entity types in the provided dataset shows low consistency. For example, in the sentence[1] "*rice - long, medium, or short-grain white; also popcorn rice*" the first occurrence of "*rice*" is labelled as PROD. However, "*popcorn rice*" is not labeled as a named entity. The precise definitions of the complex entities and consistent labeling might be crucial for developing high-performing models for complex entity recognition. Another example is country names. In one sentence a country name can be considered as a named entity, but in another sentence a country name is annotated as O token. For instance, in the sentence[2] "*spain has an embassy ...*" the token "*spain*" has O label, but in the sentence "*in madrid, spain ...*" the same token is considered as a LOC (location) entity despite the fact that in both contexts this token refers to country name. Despite, we call such cases "mislabeled-as-O entities", not all of them can be considered wrong labels. As far as we have noticed, this problem is very common for the Farsi language. To be precise, we calculated the average number of mislabeled-as-O entities for each language (see Table 2). We suppose that this issue can influence the performance of both models.

Language	Mislabeled-as-O
BN-Bangla	0.691
DE-German	0.474
EN-English	1.002
ES-Spanish	3.061
FA-Farsi	20.234
HI-Hindi	1.995
KO-Korean	5.489
NL-Dutch	3.269
RU-Russian	1.606
TR-Turkish	3.494
ZH-Chinese	1.094

Table 2: Mean number of mislabeled-as-O entities

2. The second issue is that we noticed while working on the Template-free approach is punctuation labeling. For example, in the sentence[3] "*thomas earnshaw, inventor of ...*" the comma is presented as a separate token. However, in the sentence "*... museum of fine arts, houston ...*" the comma is part of token "*arts,*". This issue was crucial on the reconstruction entity spans step since the tokenizer always considers any punctuation token as a separate one. This issue is common for MultiCoNER data in many languages, especially in German and Bangla.

[1]The example sentence ID: 2e9d398c-a956-4419-a48d-f53790d2d237 (file en_train.conll)

[2]The example sentence IDs are df4360c4-a483-493b-bd93-87814db0104c and 475fb6b2-b9aa-4ec8-8b58-443f5e2774e8 (file en_train.conll)

[3]The example sentence IDs are be16705b-7f6e-4c28-b086-5eabf5950d29 and ea916f1b-b9a5-4959-9537-aeb875c1faf1 (file en_train.conll)

3. For the purposes of more detailed performance analysis we considered the dependence between prediction accuracy and entity length (in number of tokens). The Figure 3 shows this dependence for both Template-free and Token Classification approaches considering the dev dataset. On the one hand, the mT5-XL model, trained according to the Template free approach, performs better on longer entities compared to the RemBert model. On the other hand, the RemBert model, trained that exemplifies the Token Classification approach, shows better results on shorter entities. This could be the reason why the Token Classification method outperforms the Template free approach on the test set since the average sentence length in the test data is dramatically less than the average sentence length in the train and dev datasets.

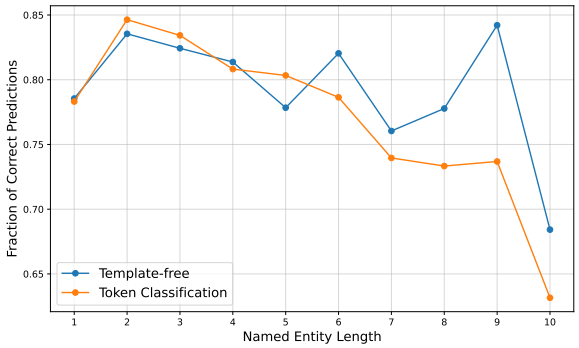

Figure 3: The dependence of prediction accuracy on the entity lengths

5 Conclusion

In the paper, we implemented and evaluated two straightforward yet different approaches to the multilingual NER subtask with complex types of entities (MultiCoNER). The first approach uses a state-of-the-art transformer-based model and fine-tuning for token classification, while the second one applies a template-free information extraction paradigm. Our results are within the 10% margin of the top solution for multilingual NER tasks and much higher than organizers' baseline performance. Therefore, we can conclude that out-of-the-box approaches generalize quite well for complex NER tasks and provide a viable alternative. The performance of the template-free approach can suffer from inconsistent annotation. The second finding is the relatively low performance of an ensemble model, but this issue needs further investigation. Evaluation of systems for multilingual NER with complex entity types is still challenging. Our study analyzed the dataset and found several issues that can be addressed in future versions of MultiCoNER Track or in similar competitions.

Acknowledgements

The project is supported by the Russian Science Foundation, grant # 20-11-20166. This research was supported in part through computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

References

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

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.

Abbas Ghaddar and Phillippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1896–1907. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task

benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Zhanming Jie and Wei Lu. 2019. Dependency-guided lstm-crf for named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 3862–3872.

Hongjin Kim and Harksoo Kim. 2021. Fine-grained named entity recognition using a multi-stacked feature fusion and dual-stacked output in korean. *Applied Sciences*, 11(22):10795.

PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. 2021. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing.

J. Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34:50–70.

Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and James Glass. 2013. Query understanding enhanced by hierarchical parsing structures. pages 72–77.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307. Association for Computational Linguistics.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. In *Proceedings of Recent Advances in Natural Language Processing*, page 876–885.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. 2021. Template-free prompt tuning for few-shot ner. *arXiv preprint arXiv:2109.13532*.

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.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133. Association for Computational Linguistics.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A dataset for nested named entity recognition in english newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.