DS4DH at SemEval-2022 Task 11: Multilingual Named Entity Recognition Using an Ensemble of Transformer-based Language Models

Hossein Rouhizadeh
University of Geneva, Switzerland
hossein.rouhizadeh@unige.ch

Douglas Teodoro
University of Geneva, Switzerland
douglas.teodoro@unige.ch

Abstract

In this paper, we describe our proposed method for the SemEval 2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). The goal of this task is to locate and classify named entities in unstructured short complex texts in 11 different languages. After training a variety of contextual language models on the NER dataset, we used an ensemble strategy based on a majority vote to finalize our model. We evaluated our proposed approach on the multilingual NER dataset at SemEval-2022. The ensemble model provided consistent improvements against the individual models on the multilingual track, achieving a macro F1 performance of 65.2%. However, our results were significantly outperformed by the top ranking systems, achieving thus a baseline performance.

1 Introduction

Named entity recognition (NER) is the process of identifying pre-defined categories of named entities, such as people, places, organizations, from unstructured text. NER usually serves as an important first component in various natural language processing (NLP) tasks, such as question answering (Mollá et al., 2006), information retrieval (Guo et al., 2009) and machine translation (Babych and Hartley, 2003). Thus, the performance of the NER system can influence the quality of many downstream NLP applications. Despite the high performance achieved by the current NER systems, they still face some critical challenges(Augenstein et al., 2017). NER models are typically trained on a well-formed news text containing a variety of entities within a relatively long context. In addition, most of the existing NER datasets usually include a large number of common entities between train set and test set. As a result, the performance of the models drops dramatically in the real world applications as they must deal with unseen entities and noisy texts. Furthermore, previous studies on NER have mostly focused on English and as a result, many other languages specially low-resource ones, such as Turkish, Korean, and Persian, have not been as well studied (Rouhizadeh et al., 2021a,b). In this context, SemEval-2022 proposes the task of Multilingual Complex Named Entity Recognition (MultiCoNER) (Malmasi et al., 2022b), which is concerned with detecting semantically ambiguous and complex entities in short and low-contextual settings for 11 languages (i.e. English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla). In this paper, we present a multilingual NER method based on ensemble of deep neural language models. We first trained multiple NER models on the official training dataset and then utilized an ensemble strategy based on a majority of votes from the top-3 best-performing models. Based on the macro-average F1-score of 65.2, achieved by our model, we placed 20th in the multilingual track of the competition.
The rest of the paper is organized as follows. Section 2 reviews published work related to the NER task. Section 3 and section 4 explain our proposed NER system and the experimental setup respectively. The results and detailed analysis of the model performance are discussed in section 5 and the conclusion and future work are reported in section 6.

2 Related Work

Over the last decade, deep learning approaches have significantly improved the results of different NER tasks (Baevski et al., 2019; Akbik et al., 2018). The most recent works on NER utilize pre-trained language models like BERT in a supervised setting (Yamada et al., 2020; Wang et al., 2020; Schneider et al., 2020; Shaffer, 2021). These models use pre-trained language models that have been trained on a large monolingual or multilingual corpus to fine-tune NER models. Meng et al. (2021) intro-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1543 - 1548

duced a number of current challenges of developed NER datasets and systems. The challenges include the presence of long-tail entities, i.e., entities with large distribution and millions of values, emerging entities, i.e., domains with growing entities, or complex entities, i.e., linguistically complex entities such as gerunds and full clauses, in the context of the systems' inputs. In addition, as discussed in Jayarao et al. (2018) the context of search queries and questions usually include a short amount of words which could be problematic for NER systems. To overcome the above issues, Meng et al. (2021) created three new NER datasets, including short sentences, questions, and search queries, and a novel NER system which uses a contextual gazetteer representation (CGR) encoder and a mixture of experts (MoE) gating network to feed a CRF layer for final predictions. Fetahu et al. (2021) also tackled the challenge of the code-mixed queries in which entities and non-entity query terms coexist simultaneously. They developed a large-scale NER dataset in six languages with four different scripts as well as a novel multi-lingual NER method for code-mixed queries which integrates external knowledge into the multilingual setting.

3 Method

Our multilingual NER system takes sentences in 11 different languages and automatically identifies and classifies named entities within each sentence. For each sentence, the system utilizes three different BERT-like models (fined-tuned on the multilingual NER dataset) to perform entity prediction independently. Next, for each entity, the label with the majority of votes will be chosen as the final prediction. In the following, we provide details on different NER models we used in our pipeline and our ensemble strategy for label prediction in section 3.1 and section 3.2, respectively.

3.1 Training NER Models

To build our NER model, we first fine-tune a number of pre-trained multilingual transformer-based models, i.e., Multilingual-BERT (Pires et al., 2019), XLM-RoBERTa-base, XLM-RoBERTa-Large (Conneau et al., 2019) and Distilbert-Multilingual (Sanh et al., 2019), on the official training dataset (see section 4.1 for more details about the dataset). We fine-tune each particular model by adding (1): a fully connected neural network (FCNN) layer or (2): a conditional random fields (CRF) layer (Lafferty et al., 2001) on the top of the transformer architecture. Transformer-based models usually use the byte-pair encoding for the tokenization. In other words, each token might be divided into more than one sub-token. To deal with this, during training, among the subtokens labels of a given word, the label of the fisrt sub-token has been considered as the label of the word. We also use the BERT-like models to train a simple BiLSTM model with an additional linear classifier on the dataset [1] Following Reimers and Gurevych (2019), we calculate the vector representation for each context word by taking the average of the layer output embeddings of the pre-trained language model and feed them to a BiLSTM neural network as input[2].

As the next step, we select three of the best-performing NER models and use an ensemble strategy (discussed in section 3.2) to finalize our model.

3.2 Ensemble of the NER Models

Having trained multiple NER models, we use an ensemble strategy based on a majority vote to assign the predictions (Copara et al., 2020b,a; Knafou et al., 2020; Naderi et al., 2021). More in detail, for a given sentence S, three NER models infer their predictions independently. Thus, we will have three labeled instances of S associated with several entity labels. Next, for each identified entity, we choose the label that gets the majority of votes (at least two votes) as the final prediction. Note that as we use three different NER models in our pipeline, three different labels might be assigned to a given entity. In such cases, we choose the predicted label of the best-performing model (evaluated on the dev set) as the final prediction.

4 Experimental Setup

This section discusses the dataset we used to conduct our experiments, followed by the parameters we used to train the models.

4.1 Data

Our experiments were conducted using the multilingual dataset provided by the SemEval-2022 Task 11 organizers (Malmasi et al., 2022a). The dataset consists of entity annotated sentences from eleven dif-

[1] We used the code provided by Adelani et al. (2021) to perform BiLSTM experiments.

[2] We only report the results when we feed the BiLSTM with XLM-RoBERTa-large as it performed best compared to the other models

Entity	Train		Dev		Test	
Person	35091	18.4%	8862	18.6%	2342	18.7%
Location	43052	22.6%	10978	23.1%	2932	23.4%
Group	26373	13.8%	6473	13.6%	1638	13.0%
Creative Work	30817	16.2%	7556	15.9%	2015	16.1%
Production	28170	14.8%	6949	14.6%	1848	14.7%
Corporation	26315	13.8%	6575	13.8%	1738	13.8%
All	189818	100%	47393	100%	12513	100%

Table 1: General statistics of the dataset including the number and the distribution of each entity.

Entity / Model	m-BERT	XLM-RoBERTa-base	XLM-RoBERTa-large	m-DistillBERT	BiLSTM	Ensemble
Person	69.2 | 70.8	88.8 | 89.2	90.1 |90.8	83.0 | 82.1	74.3	**91.3**
Location	69.4 | 69.9	86.9 | 87.6	88.0 | 89.3	83.0 | 79.9	75.7	**89.9**
Group	60.7 | 71.1	80.3 | 81.7	84.2 | 85.5	74.0 | 73.4	61.3	**86.2**
Creative Work	58.3 | 59.1	75.0 | 77.4	80.7 | **82.3**	67.0 | 73.2	51.1	81.7
Production	55.0 | 56.6	74.8 | 76.1	79.6 | 80.6	67.0 | 63.5	54.6	**80.6**
Corporation	69.1 | 69.4	82.7 | 83.9	85.5 | 87.1	76.0 | 75.2	61.5	**88.1**
All	63.8 | 64.9	82.5 | 84.0	84.7 | 85.8	75.7 | 75.2	64.2	**86.3**

Table 2: The F1 performance of different multilingual NER models. Each cell include the results when we used a FFCN (the number of the left side) or a CRF layer (the number of the right side) in the model.

ferent languages: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla. The six entity types of the dataset are Person, Location, Production, Corporation, Group, and Creative Work. The organizers provided the competitors with NER-tagged training and development sets, and then released an unlabeled test set for the final prediction. To fine-tune our hyperparameters and evaluate our models in the development phase, we divided the training set into two parts - 0.80% for the train set and 0.20% for the dev set - and used the official dev set (i.e., provided by the organizers) to test the models and analyse our model results as the labels of the official test set are not released. The number and distribution of occurrences of each entity in the training (train and dev) and test (official dev) datasets are reported in Table 1, where we can notice a relatively good class distribution among the training examples.

4.2 Parameters

In our experiments, we fine-tuned different multilingual pre-trained language models including *bert-base-multilingual-uncased*, *XLM-Roberta-base*, *XLM-Roberta-large*, *distilbert-base-multilingual-cased*, and also trained a simple BiLSTM model on the dataset. We trained each particular model for 6 epochs using Adam optimizer (Kingma and Ba, 2014), a batch size of 16, the learning rate of 2e-5, and the maximum sequence length of 256 tokens. We computed the F1 performance of the model on each epoch and finally saved the parameters of the epoch with the best performance to perform NER on the test set.

5 Results and Discussion

5.1 Results

In Table 2, we show the macro-averaged F1 performance of the NER models on the different entities of the unofficial test dataset. We use the three best performing models identified in the dev set, i.e., *XLM-RoBERTa-large + CRF*, *XLM-RoBERTa-base + CRF* and *XLM-RoBERTa-large + FCNN*, to create our ensemble strategy. As shown in Table 2, the ensemble model outperforms the other single transformer-based models, improving the F1-score of the top-performer models by around 1% point. The results also indicate that the models fine-tuned on the XLM-RoBERTa (both large and base) outperform the other models by a wide margin. In addition, a comparison between the results of each particular model with and without CRF on the test set shows that adding a CRF layer to the models could be helpful as it improves the model performance in most cases. The results show that all models perform best in inferring $Person$ and $Location$ entities. This can be due to the large number of instances of both entities in the training set. In Table 1, it is shown that the number of oc-

Sentence Length	$1 \leq N \leq 5$	$6 \leq N \leq 10$	$11 \leq N \leq 15$	$16 \leq N \leq 20$	$N > 20$	All
Number of Sentences	85	1988	2517	1964	2246	8800
Ratio of the sentences	0.1%	22.5%	28.6%	22.3%	25.5%	100%

Table 3: Number and ratio of sentences with different length (in words) in the test set.

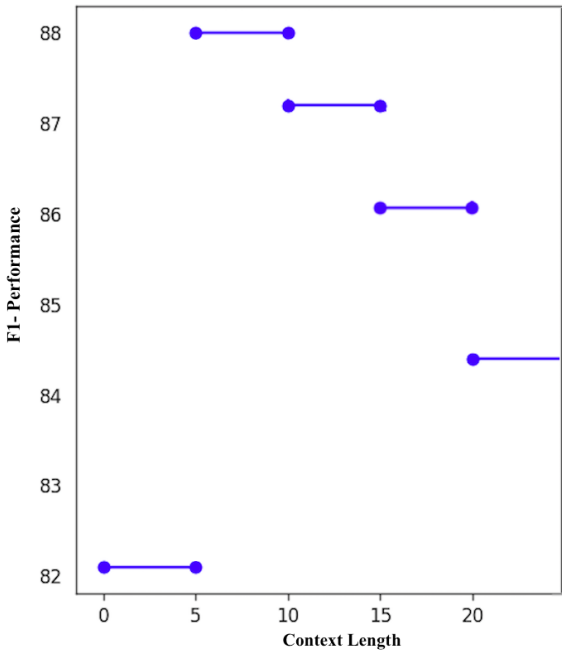

Figure 1: Performance of our ensemble model according to the sentence length (in words).

currences of these entities in the dataset is greater than the other ones. The BiLSTM model also performs significantly worse than the fine-tuned XLM-RoBERTa-large models, despite using the same word vectors.

5.2 Discussion

Effect of the context length One of the most important factors affecting the performance of the NER systems is the context length (Meng et al., 2021). To analyze the effect of the input context on our NER system, we divided the (unofficial) test set into 5 different groups: (1): sentences with five or fewer words, (2): sentences with a context length of at least 6 and less than 11, (3) sentences including at least 10 and less than 15 context words, (4) sentences containing between 15 and 20 words, and (5) sentences containing more than 20 context words. The number and ratio of sentences in each group is reported in Table 3. Figure 1 shows the performance of the ensemble NER model on the different groups of sentences. As it can be seen, the model has the worse performance when the sentences contain 5 or less words. Surprisingly, the model performs best in the second group (sentences containing between 5 and 10 words) showing the strength of the model even in the short the sentences.

6 Conclusion

In this paper, we presented our multilingual NER method that uses an ensemble of different fine-tuned models to identify the named entities in the unstructured texts. Using a variety of multilingual pre-trained language models, we first fine-tuned several NER models and then applied a vote-based ensemble strategy to make the final prediction. Our submission achieved an overall F1 score of 65.2, ranking 20th in the multilingual track of task 11 of SemEval-2022. Our next step would be to examine other possible types of ensemble strategies as it has shown to be effective in the performance of the NER models.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018.

Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020a. Contextualized french language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48.

Jenny Copara, Nona Naderi, Julien Knafou, Patrick Ruch, and Douglas Teodoro. 2020b. Named entity recognition in chemical patents using ensemble of contextual language models. *arXiv preprint arXiv:2007.12569*.

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.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Pratik Jayarao, Chirag Jain, and Aman Srivastava. 2018. Exploring the importance of context and embeddings in neural ner models for task-oriented dialogue systems. *arXiv preprint arXiv:1812.02370*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Julien Knafou, Nona Naderi, Jenny Copara, Douglas Teodoro, and Patrick Ruch. 2020. Bitem at wnut 2020 shared task-1: Named entity recognition over wet lab protocols using an ensemble of contextual language models. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 305–313.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58.

Nona Naderi, Julien Knafou, Jenny Copara, Patrick Ruch, and Douglas Teodoro. 2021. Ensemble of deep masked language models for effective named entity recognition in health and life science corpora. *Frontiers in research metrics and analytics*, 6.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Hossein Rouhizadeh, Mehrnoush Shamsfard, Mahdi Dehghan, and Masoud Rouhizadeh. 2021a. Persian semcor: A bag of word sense annotated corpus for the persian language. In *Proceedings of the 11th Global Wordnet Conference*, pages 147–156.

Hossein Rouhizadeh, Mehrnoush Shamsfard, Vahideh Tajalli, and Masoud Rouhziadeh. 2021b. Persian-wsd-corpus: A sense annotated corpus for persian all-words word sense disambiguation. *arXiv preprint arXiv:2107.01540*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 40–45.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.