CSECU-DSG at SemEval-2022 Task 11: Identifying the Multilingual Complex Named Entity in Text Using Stacked Embeddings and Transformer based Approach

Abdul Aziz, Md. Akram Hossain, and **Abu Nowshed Chy**
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

Recognizing complex and ambiguous named entities (NEs) is one of the formidable tasks in the NLP domain. However, the diversity of linguistic constituents, syntactic structure, semantic ambiguity as well as differences from traditional NEs make it challenging to identify the complex NEs. To address these challenges, SemEval-2022 Task 11 introduced a shared task MultiCoNER focusing on complex named entity recognition in multilingual settings. This paper presents our participation in this task where we propose two different approaches including a BiLSTM-CRF model with stacked-embedding strategy and a transformer-based approach. Our proposed method achieved competitive performance among the participants' methods in a few languages.

1 Introduction

Named entity recognition (NER) is a popular sequence labeling task in the natural language processing (NLP) arena. It has numerous applications in several computational linguistic tasks including designing efficient search systems, data mining, and document indexing. However, prior studies mostly focused on identifying traditional named entities (NEs) recognition e.g. person names, locations, and organizations names (Murthy et al., 2018).

The ever-growing generation of unstructured social media data contains a huge amount of complex (Meng et al., 2021) and ambiguous(Fetahu et al., 2021) named entities. This is because social media data is severely induced by noise as well as the linguistic constituent, syntactic and semantic ambiguity exists in this data source. Besides, the social media data mostly have multilingual data. Therefore it poses new challenges to the traditional named entity recognizer system (Aguilar et al., 2019; Ashwini and Choi, 2014). To address the challenges of recognizing such complex and ambiguous named entities in multilingual settings, (Malmasi et al., 2022b) introduced a shared task at SemEval-2022 named as MultiCoNER. The task is composed of three category of tracks including multi-lingual, mono-lingual, and code-mixed tracks. A multilingual dataset (Malmasi et al., 2022a) containing data from 11 languages is used to assess the participants' system. To illustrate a clear view of the task definition, we articulate two examples from English languages and corresponding labels in Table 1.

English
Text#1: Adaptation of seinen series by kenichi sonoda. **Tag:** [O, O, B-CW, O, O, B-PER, I-PER, O]
Text#2: It was designed by kohn pedersen fox. **Tag:** [O, O, O, O, B-CORP, I-CORP, I-CORP, O]

Table 1: Data sample.

We articulate the rest of the contents as follows: Section 2 describes our proposed approach whereas, in Section 3, we present our experimental setup and conduct performance analysis against the various settings and participants' methods. Finally, we conclude our work with some future directions in Section 4.

2 Proposed Framework

In this section, we describe our proposed approach for the MultiCoNER shared task. Our goal is to identify the complex and ambiguous named entities in multilingual settings. The task is articulated into multi-lingual, mono-lingual, and code-mixed tracks. To address the task challenges, we employ two different approaches including a BiLSTM-CRF

**The first two authors have equal contributions.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1549 - 1555
July 14-15, 2022 ©2022 Association for Computational Linguistics

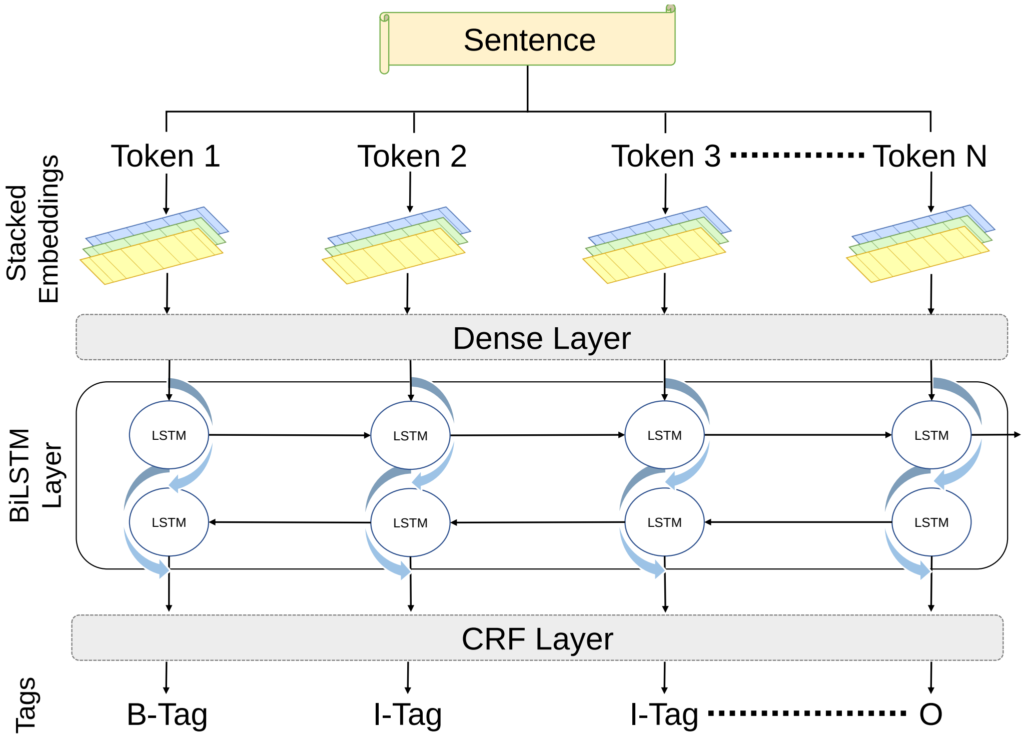

Figure 1: Overview of our proposed BiLSTM-CRF based framework.

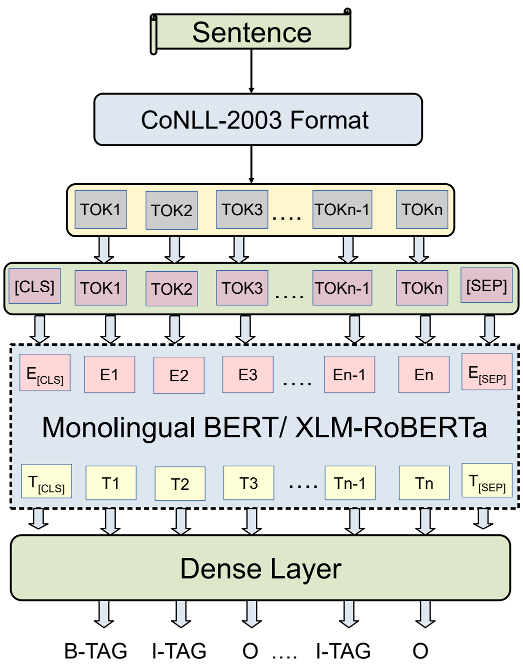

Figure 2: Overview of our proposed transformer-based NER framework.

based approach and a transformer-based approach. The overview diagrams of these approaches are depicted in Figure 1 and Figure 2, respectively. In the BiLSTM-CRF model, we employ the Flair's (Akbik et al., 2019) implementation of stacked embedding technique for effective word representation. Whereas in our transformer-based approach, we employ the available monolingual BERT for each of the languages. However, for some of the languages and for the multilingual and code-mixed settings, we employ the XLM-RoBERTa.

2.1 BiLSTM-CRF with Stacked Embedding

The BiLSTM-CRF model is well-known for the named entity recognition (NER) task. For the training purpose, we use the sentence represented as CoNLL -U format containing BIO tag for each token. A token tagged as O means it is not part of an entity, B-X denotes the first token of an X entity, I-X denotes the token is within the X type entity having multiple tokens. The tokens are then sent to the embedding layer. We employ Flair's (Akbik et al., 2019) implementation of stacked embedding strategy that concatenates embedding vectors of different models together for the effective representation of tokens. To choose the optimal ones in the stacking, we explore various embedding models. The list of embedding models used in our stacked-embedding approach is presented in Table 2. The embedding vectors are concatenated and send to the BiLSTM encoder to distill the contextual dimension of each token. The BiLSTM encoder is followed by a linear-chain conditional random fields (CRF) classifier that generate predictions with the

Track	Word Embeddings Used in Stack Embedding Model (BiLSTM-CRF based System)	Used Transformer Model (Transformer based System)
English (en)	Globe, News-forward, News-backward (Akbik et al., 2018)	bert-base-uncased (Devlin et al., 2019)
Spanish (es)	FastText (es), es-forward, es-backward	dccuchile/bert-base-spanish-wwm-cased (Cañete et al., 2020)
Dutch (nl)	FastText (nl), nl-forward, nl-backward	GroNLP/bert-base-dutch-cased (de Vries et al., 2019)
Russian (ru)	FastText (ru), Byte Pair Embedding (ru), Character Embedding	DeepPavlov/rubert-base-cased-sentence
Turkish (tr)	FastText (tr), Byte Pair Embedding (tr)	dbmdz/bert-base-turkish-128k-cased
Korean (ko)	FastText (ko), Byte Pair Embedding (ko), Character Embedding	klue/bert-base (Park et al., 2021)
Farsi (fa)	FastText (fa), fa-forward, fa-backward	HooshvareLab/bert-fa-base-uncased (Farahani et al., 2020)
German (de)	FastText (de), de-forward, de-backward	-
Chinese (zh)	FastText (zh), Byte Pair Embedding (zh), character Embedding	bert-base-chinese
Hindi (hi)	FastText (hi), hi-forward, hi-backward	xlm-roberta-base
Bangla (bn)	Byte Pair Embedding (bn), Character Embedding	xlm-roberta-base
Multi-linguall (multi)	Byte Pair Embedding (multi), multi-Forward, multi-Backward	xlm-roberta-base
Code-mixed (mix)	-	xlm-roberta-base

Table 2: Used word embeddings and transformer models in our CSECU-DSG system.

BIO tagging scheme.

2.2 Transformer based System

In our transformer-based approach, we employ the monolingual BERT model for each of the languages that are available in the Huggingface repository (Wolf et al., 2019). To choose the optimal transformers, we explore various embedding models. However, for some of the languages, XLM-RoBERTa performed better compared to the monolingual BERT. In such cases and also for the code-mixed and multilingual data, we employ the XLM-RoBERTa model.

The Facebook AI launched the XLM-RoBERTa as an upgrade to their initial XLM-100 model (Conneau et al., 2020). It is a scaled cross-lingual sentence encoder. Using self-supervised training approaches, it offers state-of-the-art performances in cross-lingual understanding where a model is taught in one language and then applied to multiple languages with no additional training data. This model showed increased performance on numerous

NLP applications. Utilizing just monolingual data, XLM-RoBERTa was trained with a multilingual masked language model (MLM) objective.

With users publishing content in over 160 languages on Facebook, XLM-RoBERTa is a major leap toward the goal of offering the greatest possible experience on this platform for everybody, regardless of their native language. XLM-RoBERTa creates the possibility for a one-model-for-many-languages approach rather than a single model per language. There are two versions of XLM-RoBERTa. The base version of XLM-RoBERTa contains 250M parameters, whereas the large version has 560M. The vocabulary in both versions is 250K. In our framework, we use the XLM-RoBERTa-base version to extract the effective transfer learning features. The list of transformer models used in our transformer-based approach is presented in Table 2.

3 Experiment and Evaluation

3.1 Dataset Description

The organizers of the SemEval-2022 MultiCoNER shared task 11 (Malmasi et al., 2022b,a) provided a benchmark dataset to evaluate the performance of the participants' systems. The dataset comprises data of 11 languages along with a code-mixed dataset. The dataset statistics are summarized in Table 3.

Track	Training	validation	Test
BN-Bangla	15,300	800	133,119
DE-German	15,300	800	217,824
EN-English	15,300	800	217,818
ES-Spanish	15,300	800	217,887
FA-Farsi	15,300	800	165,702
HI-Hindi	15,300	800	141,565
KO-Korean	15,300	800	178,249
NL-Dutch	15,300	800	217,337
RU-Russian	15,300	800	217,501
TR-Turkish	15,300	800	136,935
ZH-Chinese	15,300	800	151,661
MULTI-Multilingual	168,300	8,800	471,911
MIX-Code_mixed	1500	500	100000
Total	338,100	18,100	2,567,509

Table 3: The statistics of the datasets used in different lingual track.

3.2 Experimental Settings

We now describe the the set of parameters that we have used to design our proposed CSECU-DSG systems for the MultiCoNER task. We employ two different approaches including a BiLSTM-CRF based approach and a transformer-based approach. We employ Flair's (Akbik et al., 2019) implementation of the BiLSTM-CRF approach and Huggingface (Wolf et al., 2019) implementation of the monolingual and multilingual transformer models with fine-tuning. We finetune these models with the provided training data. We also tune several hyper-parameters to obtain the optimal performances. Our hyper-parameters search space is presented in Table 4 and default settings are used for the others.

Hyper-parameters	Search Space
Training batch size	$\{8, 16, 32\}$
Learning rate	$\{0.1, 1e-3, 1e-5, ..., 3e-6\}$
Number of epochs	$\{3, 5, ..., 100\}$
BiLSTM output size	$\{64, 128, 256, 512\}$

Table 4: The hyper-parameters search space.

3.3 Evaluation Measures

To assess the performance of the participants' systems, SemEval-2022 MultiCoNER shared task 11 (Malmasi et al., 2022b) used different strategies and metrics. Since the evaluation file contains instances from all 11 languages and two other language settings including multilingual and code-mixed, the macro averaged F1 score of all these languages is used as the primary evaluation metric to rank the participants' systems. However, the organizers reported the results based on precision and recall evaluation measures too.

3.4 Results and Analysis

In this section, we analyze the performance of our proposed approaches in the MultiCoNER shared task. We have employed two different approaches including a BiLSTM-CRF based approach and a transformer-based approach. The dataset comprises of 11 different languages and two other language settings including multilingual and code-mixed. The overall performance of the system is estimated considering the macro average F1 score obtains in each languages dataset. Considering this,

https://huggingface.co/xlm-roberta-base

https://huggingface.co/models

Track	BiLSTM-CRF based System			Transformers based System		
	F1-macro	Precision	Recall	F1-macro	Precision	Recall
EN	0.6403	0.6871	0.6033	**0.6924**	**0.6872**	**0.6981**
ES	**0.6562**	**0.6894**	**0.6313**	0.6138	0.6069	0.6232
NL	**0.6794**	**0.7174**	**0.650**	0.5981	0.5985	0.6026
RU	0.6177	**0.6971**	0.5578	**0.6308**	0.626	**0.639**
TR	**0.553**	**0.6457**	0.4925	0.538	0.5285	**0.558**
KO	0.6128	**0.681**	0.5616	**0.6205**	0.6227	**0.6214**
FA	0.5454	**0.5941**	0.5094	**0.5581**	0.558	**0.5617**
DE	**0.7249**	**0.7493**	**0.7047**	-	-	-
ZH	0.387	0.5461	0.3372	**0.6722**	**0.6855**	**0.6761**
HI	**0.5768**	**0.6146**	0.5477	0.5563	0.5638	**0.5551**
BN	0.428	0.4858	0.395	**0.5055**	**0.5221**	**0.4942**
MULTI	0.3505	0.4926	0.3065	**0.644**	**0.6479**	**0.652**
MIX	-	-	-	**0.6403**	**0.6423**	**0.6436**

Table 5: CSECU-DSG results of both systems on all tracks.

Model Type	Parameter	Track Name												
		ES	NL	TR	DE	HI	EN	RU	KO	FA	ZH	BN	MULTI	MIX
BiLSTM-CRF based system	learning rate	0.1	0.1	0.1	0.1	0.1	-	-	-	-	-	-	-	-
	epoch	50	100	100	100	100	-	-	-	-	-	-	-	-
	batch size	32	32	32	32	32	-	-	-	-	-	-	-	-
	hidden size	256	128	256	128	256	-	-	-	-	-	-	-	-
Transformer based system	learning rate	-	-	-	-	-	3e-5	3e-5	3e-5	3e-5	3e-5	3e-5	3e-5	3e-5
	epoch	-	-	-	-	-	7	10	10	7	10	10	6	25
	batch size	-	-	-	-	-	16	16	16	16	16	16	16	16

Table 6: Experimental settings best performing system of all track.

we analyze the performance of our proposed systems, based on each language. The corresponding results are reported in Table 5.

Results showed that the transformer-based approach obtained better performances in most of the languages compared to the BiLSTM-CRF with stacked embedding approach in terms of primary evaluation measure F1-macro. However, in terms of precision BiLSTM-CRF performed better and in terms of recall transformer-based system performed better in most of the languages dataset. Considering the diverse performances in these two approaches the optimal parameter settings of the best-performing ones are articulated in Table 6 and default settings used for the other parameters.

However, comparative performance analysis in most of the languages dataset showed that our proposed system did not perform well to address the task challenges. However, the best performing results of our proposed CSECU-DSG system in the MultiCoNER shared task for the Chinese(ZH) and Hindi (HI) languages along with other top performing and competitive participants systems are articulated in Table 7. Following the benchmark of the MultiCoNER shared task, participants' systems are ranked based on the primary evaluation measures F1-macro. For these two languages, our proposed CSECU-DSG system obtained comparatively better performances.

4 Conclusion and Future Directions

In this paper, we presented our proposed systems to address the challenges of the MultiCoNER shared task. We employed a BiLSTM-CRF with a stacked embedding-based approach and a transformer-based approach. Experimental results demonstrate

Team (Rank)	F1-macro
Chinese (ZH)	
CSECU-DSG (9th)	0.6722
Performance of other participants'	
USTC-NELSLIP (1st)	0.8169
OPDAI (3rd)	0.7954
QTrade AI (8th)	0.7400
RACAI (16th)	0.6270
MarSan_AI (19th)	0.5664
Hindi (HI)	
CSECU-DSG (11th)	0.5768
Performance of other participants'	
DAMO-NLP (1st)	0.8623
RACAI (3rd)	0.6808
YNUNLP (8th)	0.6339
silpa_nlp (14th)	0.5149
Enigma (17th)	0.4862

Table 7: Comparative results with other selected participants on Chinese and Hindi track.

that transformer-based approach performed better compared to the other approach in most of the languages dataset.

In the future, we have a plan to incorporate the task-specific features and technologies to address the challenges properly. We also have a plan to explore the existing NER technologies and fuse them in a unified architecture to overcome the limitations of the current approaches.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2019. A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Sandeep Ashwini and Jinho D Choi. 2014. Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515.

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.

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.

Rudra Murthy, Mitesh M Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–20.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.