LSJSP at SemEval-2023 Task 2: FTBC: A FastText based framework with pre-trained BERT for NER

Shilpa Chatterjee*
IIT Kanpur
shilpa@cse.iitk.ac.in

Leo Evenss*
IIT Kanpur
leoevenss@cse.iitk.ac.in

Pramit Bhattacharyya*
IIT Kanpur
pramitb@cse.iitk.ac.in

Joydeep Mondal
IIT Kanpur
joydeep20@iitk.ac.in

Abstract

This study introduces the system submitted to the SemEval 2022 Task 2: MultiCoNER II (Multilingual Complex Named Entity Recognition) by the LSJSP team. We propose FTBC, a FastText-based framework with pre-trained Bert for NER tasks with complex entities and over a noisy dataset. Our system achieves an average of 58.27% F1 score (fine-grained) and 75.79% F1 score (coarse-grained) across all languages. FTBC outperforms the baseline BERT-CRF model on all 12 monolingual tracks.

1 Introduction

NER (Named Entity Recognition) is a sequence labeling task that finds spans of text and tags them to a named entity (persons, locations, organizations, and others.). NER finds its applications in various Natural Language Processing (NLP) tasks, including Information Retrieval, Machine Translation, Question Answering, and Automatic Summarization (Jurafsky and Martin, 2021). It also finds application across various domains, including social media, news, e-commerce, and healthcare. Words having multiple meanings increase the complexity of the NER task. In English, the word "Sunny" can refer to both a person and the weather. The problem of ambiguous entities become all the more pronounced for Indian languages. In the Bangla sentence "cañcala caudhurī cañcala bodha karachena" the first "cañcala" refers to a person named cañcala whereas the second "cañcala" (restless) is an adjective.

Pre-trained contextual embeddings such as BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2019), and other transformer-based (Vaswani et al., 2017) models have improved the performance of NER. The embeddings are trained on large corpora, improving the named entities' contextual representations. However, for low-resource languages like Bengali, Hindi, Swedish, and Farsi, the performance of these pre-trained models lag compared to their English counterpart.

MultiCoNER II shared task (Fetahu et al., 2023b) aims at building Named Entity Recognition (NER) systems for 12 languages, including Bangla (BN), German (DE), English (EN), Spanish (ES), Farsi (FA), French (FR), Hindi (HI), Italian (IT), Portuguese (PT), Swedish (SV), Ukrainian (UK), Chinese (ZH). The task has two kinds of tracks, including one multilingual track and 12 monolingual tracks for all the languages. The monolingual track requires the development of NER models for that particular language, whereas the multilingual track requires the development of a single multilingual model that can handle all 12 languages. The sequence of tokens in MultiCoNER II is classified into 30+ classes. Further, including simulated errors, like spelling mistakes, in the test set makes the task more realistic and difficult.

This paper proposes FTBC, a framework comprising FastText (Bojanowski et al., 2017) and BERT for the MultiCoNER II task without using any other external knowledge bases. FTBC achieves an average F1 score of 58.27% on fine-grained and 75.79% on coarse-grained categories across all the 12 monolingual tracks. FTBC outperforms the baseline BERT-CRF for all the languages for both fine and coarse-grained categories.

2 Related Work

Named Entity Recognition is a fundamental task in natural language processing (NLP). Most of the earlier works considered NER as a sequence labelling task. Chiu and Nichols (2016) introduced convolutional neural networks for NER, whereas Žukov-Gregorič et al. (2018) used recurrent neural networks (RNNs) with CRF for NER. However, these models do not perform very well on complex enti-

*These authors contributed equally to this work

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1254–1259

ties like movie names (Meng et al., 2021), which are challenging to tag. Fetahu et al. (2021) has extended this work to multilingual code-mixed settings. Pre-trained language models like ELMo (Peters et al., 2018) or transformer-based (Vaswani et al., 2017) models like BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019) improve the performance of NER on complex entities.

MultiCoNER I (Malmasi et al., 2022c) was a shared task to build NER systems for complex entities in 11 languages. MultiCoNER I shared challenge was conducted using the dataset curated by Malmasi et al. (2022a). The critical challenge of MultiCoNER I was dealing with complex entities with limited context. MultiCoNER II classifies the tokens into 30+ finer classes. Additionally, simulated errors have been inserted in the dataset to make the task more challenging and realistic.

3 Dataset Description

MultiCoNER II shared task (Fetahu et al., 2023b) (Fetahu et al., 2023a) focussed on building complex Named Entity Recognition (NER) systems for 12 languages. The languages include Bangla (BN), German (DE), English (EN), Spanish (ES), Farsi (FA), French (FR), Hindi (HI), Italian (IT), Portuguese (PT), Swedish (SV), Ukrainian (UK), Chinese (ZH). MultiCoNER II is an extension of MultiCoNER I (Malmasi et al., 2022c) (Malmasi et al., 2022b) where the six broader classes, namely Person, Group, Creative Works, Location, Medical, and Product, are further classified into 30+ finer classes. The fine to the coarse level mapping of the classes is as follows:

- **Person (PER):** Scientist, Artist, Athlete, Politician, Cleric, SportsManager, OtherPER.
- **Group (GRP):** MusicalGRP, PublicCORP, PrivateCORP, AerospaceManufacturer, SportsGRP, CarManufacturer, ORG.
- **Creative Work (CW):** VisualWork, MusicalWork, WrittenWork, ArtWork, Software.
- **Location (LOC):** Facility, OtherLOC, HumanSettlement, Station.
- **Medical (MED):** Medication/Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, Disease.
- **Product (PROD):** Clothing, Vehicle, Food, Drink, OtherPROD.

The train data of each of the 12 languages varies from 9,000+ sentences to 16,500+ sentences, and the validation set ranges between 500+ to 800+ sentences. The test set for each track had at least 18,000+ sentences, with the lowest for Hindi (18,349 sentences). The multilingual track comprises data from each of the monolingual tracks. The dataset consists of simulated errors like typographical and spelling mistakes, making the NER task more challenging and realistic.

4 Baseline Systems

We now describe the baseline systems developed for MultiCoNER II shared task (Fetahu et al., 2023b).

4.1 BERT

We tried using only pre-trained BERT models for all the languages in the monolingual and multilingual tracks initially. We fine-tuned the pre-trained BERT models available in Hugging Face on the provided NER dataset. The output vectors by the BERT model are then fed to either a Linear Layer or a CRF layer to get the predictions.

4.2 BERT-Linear

The input token sequences were passed to the BERT model, which generated vectors of dimension d, equal to the fixed vector dimension of BERT. These vectors then go through a classifier layer made of two fully connected layers, followed by a softmax normalization to get the predicted tags of the input tokens. The predictions have a dimension of m, where m denotes the total number of unique labels provided for the task.

4.3 BERT-CRF Layer

Analysis of the results obtained from BERT with linear layer shows that it was important to consider the sequence for the NER task. Conditional Random Fields(CRF) (Sutton and McCallum, 2012) is a class of discriminative models that helps in recognizing how tags can change considering the position of a particular token in a sentence. CRF often finds its use in machine learning tasks like Named Entity Recognition, Parts of Speech Tagging, Object Identification, etc. Instead of using the BERT-Linear model, which predicts tags one at a time, we employ pre-trained BERT models with an extra CRF layer to predict NER tags jointly. This helps in avoiding mistakes like predicting tags

like I-PER after B-LOC. Let us consider an input sequence:

$$X = x_1, x_2, ...x_n$$

and the predicted labels as

$$Y = y_1, y_2, ...y_n$$

. We obtain the token embedding of each of the input sequences, x_i in dimension d where d is the dimension of BERT embedding. These embeddings are then passed through a dense layer, where the dimensions are transformed from d to m (m denotes the number of labels/tags present in the dataset). The output of the dense layer is the emission score for all the tokens in the input sequence X. These emission scores are then passed on to the CRF layer, where we calculate the score of the predicted label for each input token.

We calculate the score, following the (Lample Guillaume and Chris, 2016) as follows: $Score(X, Y) = \sum_{i=0}^{i=n} A_{y_i, y_{i+1}} + \sum_{i=1}^{i=n} P_{i,y_i}.\mathrm{A}_{ij}$ denote the transition score of the i^{th} label to the j^{th} label in the CRF layer and $P \in R^{nXm}$ denotes the emission scores of the input tokens from the dense layer.

We intend to maximize the logarithmic probability of the correct label sequence Y for the input sequence X.

$$\log p(Y|X) = Score(X, Y) - \log \sum_{y' \in Y_x} e^{Score(X, y')}$$

where Y_x denotes all possible label sequence for the input sequence X. While decoding, we take the label sequence obtaining the maximum score as the predicted label.

$$y^* = \underset{y' \in Y_x}{\operatorname{argmax}}\, Score(X, y')$$

Table 3 shows the fine-grained F1 score of all the monolingual tracks obtained by BERT-CRF model. We use this result as the baseline result for our system evaluation.

5 FTBC

This section discusses the FTBC system developed for the MultiCoNER II monolingual tracks. The MultiCoNER II dataset contains simulated errors, including typographical and spelling mismatches. Hence, we devised a two-layered system for the task. The tokens are checked for spelling or typographical errors in the first layer using FastText's nearest neighbor method. Since (Kumar et al., 2020) showed that for inflectional languages such as Bangla, Hindi, etc., FastText performs better than Glove (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013); hence, we used FastText for our framework. We observed that FastText's nearest neighbor could sometimes return garbage values. To avoid those garbage values, we used Levenshtein distance to calculate the distance between the provided input token and the nearest neighbor prediction by the FastText model. If the Levenshtein distance was ≥ 50, we considered the predicted word from the fastText model as our input token. The output of the fastText model is then passed on to the BERT+CRF model to get the predictions on the input tokens.Figure 1 shows a schematic representation of the FTBC system.

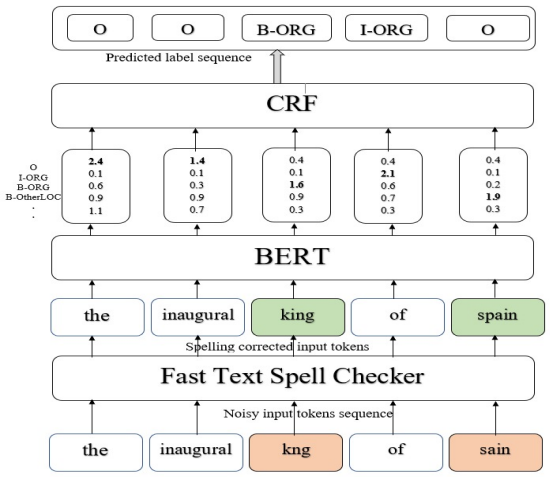

Figure 1: Architecture of System

5.1 Models Description

We used the pre-trained embeddings released by models from FastText project (Bojanowski et al., 2017) for all the languages. The corrected sequences are passed through the BERT-CRF layer. Table 1 describes the hyperparameters for the models used, whereas Table 2 contains links to HuggingFace pre-trained models used for the FTBC system. F1 scores of each of the monolingual track using FTBC is reported in Table Table 4.

Hyperparamter	Value
Batch Size	16
Epoch	20
Learning Rate	5e-5
Loss	CrossEntropy
Optimizer	Adam

Table 1: Hyperparameter Used

Tracks	Bert models
EN	dbmdz/bert-base-finetuned-conll03-english
ES	dccuchile/bert-base-spanish-wwm-cased
SV	KB/bert-base-swedish-cased
UK	Geotrend/bert-base-uk-cased
PT	neuralmind/bert-large-portuguese-cased
FR	dbmdz/bert-base-french-europeana-cased
FA	HooshvareLab/bert-fa-base-uncased
DE	dbmdz/bert-base-german-cased
ZH	bert-base-chinese
HI	muril-base-cased
BN	muril-base-cased
IT	dbmdz/bert-base-italian-cased
All	bert-base-multilingual-cased

Table 2: Pre-trained BERT Models Used

6 Results

In this section, we quantitatively evaluate FTBC against the baseline models. Table 3 shows the performance of FTBC compared to BERT-Linear and BERT-CRF models, whereas Table 4 shows the FTBC system's performance in all the monolingual tracks along with the final rank on the test set. FTBC outperforms the baseline models in all the tracks.

Tracks	Bert-Linear	Bert-CRF	FTBC
EN	0.52	0.57	**0.58**
ES	0.56	**0.60**	**0.60**
SV	0.61	0.62	**0.64**
UK	0.53	0.57	**0.58**
PT	0.38	0.56	**0.58**
FR	0.51	0.55	**0.57**
FA	0.57	0.60	**0.61**
DE	0.50	**0.54**	**0.54**
ZH	0.51	0.57	**0.60**
HI	0.31	0.52	**0.53**
BN	0.34	0.53	**0.56**
IT	0.58	0.60	**0.62**

Table 3: Fine grained F1-score of Bert-Linear and Bert-CRF across all the monolingual tracks

Analysis of the output revealed that fine-grained categories like PrivateCorp, Scientist, and Symptom were misclassified across all the languages. Additionally, categories like Aerospace-Manufacturer, OtherPER, and OtherLOC have F1 scores of less than 0.3 for all the languages. Using external knowledge bases or gazetteer lists can improve the F1 score of these classes.

Tracks	F1 Fine	F1 Coarse	Rank on Test Set
EN	0.58	0.72	24
ES	0.60	0.75	14
SV	0.64	0.81	10
UK	0.58	0.75	13
PT	0.58	0.76	16
FR	0.57	0.72	14
FA	0.61	0.72	9
DE	0.54	0.71	16
ZH	0.60	0.75	24
HI	0.53	0.80	16
BN	0.56	0.83	18
IT	0.62	0.78	14

Table 4: Results of FTBC accross all tracks

Observations also revealed confusion between the categories Politician and Artist and between MusicalGRP and Artist. Analysis shows this confusion is mainly because the train set classifies the same terms into different tags in different sentences. For example, "John" was classified in Spanish training as B-Politician, B-OtherPer, and B-Artist, leading to misclassification by the models. These instances are also found in other languages, including English, German, Italian, Portuguese, and Swedish. Tale 5 reports the sentences having different tags of "John".

Tag	Sentence
B-OtherPER	**john** sterling(born 1948) sportscaster for the new york yankees.
B-Artist	**john** baker saunders founding member and bassist for the grunge rock supergroup mad season.
B-Athlete	he was always a reserve keeper at vale park though behind first alan bosewell and when he left **john** connaughton.

Table 5: Sentences from English validation set for the word "john"

7 Conclusion

We developed FTBC, a FastText-based framework with pre-trained Bert for the MultiCoNER II shared

task. Our system outperforms the BERT-CRF baseline models for all 12 monolingual tracks. FTBC performs well for noisy data, but ambiguous entities affect its overall performance. In the future, we would like to enhance the performance of FTBC using external knowledge bases.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

Besnik Fetahu, 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.

Dan Jurafsky and James H. Martin. 2021. *Speech and Language Processing*.

Saurav Kumar, Saunack Kumar, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020. "a Passage to India": Pre-trained Word Embeddings for Indian Languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages, SLTU/CCURL@LREC 2020, Marseille, France, May 2020*, pages 352–357. European Language Resources association.

Subramanian Sandeep Kawakami Kazuya Lample Guillaume, Ballesteros Miguel and Dyer Chris. 2016. Neural Architectures for Named Entity Recognition . In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 260–270. Association for Computational Linguistics.

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

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022c. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andrej Žukov-Gregorič, Yoram Bachrach, and Sam Coope. 2018. Named entity recognition with parallel recurrent neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–74, Melbourne, Australia. Association for Computational Linguistics.