L3i at SemEval-2022 Task 11: Straightforward Additional Context for Multilingual Named Entity Recognition

Emanuela Boros
University of La Rochelle, L3i
La Rochelle, France
emanuela.boros@univ-lr.fr

Carlos-Emiliano González-Gallardo
University of La Rochelle, L3i
La Rochelle, France
carlos.gonzalez_gallardo@univ-lr.fr

Jose G. Moreno
University of Toulouse, IRIT
Toulouse, France
jose.moreno@irit.fr

Antoine Doucet
University of La Rochelle, L3i
La Rochelle, France
antoine.doucet@univ-lr.fr

Abstract

This paper summarizes the participation of the L3i laboratory of the University of La Rochelle in the SemEval-2022 Task 11, *Multilingual Complex Named Entity Recognition* (MultiCoNER). The task focuses on detecting semantically ambiguous and complex entities in short and low-context monolingual and multilingual settings. We argue that using a language-specific and a multilingual language model could improve the performance of multilingual and mixed NER. Also, we consider that using additional contexts from the training set could improve the performance of a NER on short texts. Thus, we propose a straightforward technique for generating additional contexts with and without the presence of entities. Our findings suggest that, in our internal experimental setup, this approach is promising. However, we ranked above average for the high-resource languages and lower than average for low-resource and multilingual models.

1 Introduction

Named entity recognition (NER) is the task of detecting entities and recognizing their type (e.g., person, location, organization) (Grishman and Sundheim, 1996). Standard benchmarks (e.g., CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003)) for NER are either focused on a high-resource language or on a single domain (e.g., news). However, NER is required in different fields, bringing thus several new challenges regarding: the size of the textual content (e.g., short textual snippets like Web queries (Fetahu et al., 2021) or social media posts (Rajoria, 2021)), the characteristics of the entities to be extracted (i.e., capitalization/punctuation features (Mayhew et al., 2019)), unseen and emerging entities (e.g., entities that have a fast growth rate such as new songs and movies are released weekly (Bernier-Colborne and Langlais, 2020)), complex entities (i.e., complex noun phrases such as particularly long nested person names and dates in historical documents (Boros et al., 2020a,c)). Moreover, multilingual documents in which entities of different languages than the rest of the text are present (i.e., code-mixed documents (Winata et al., 2021; Fetahu et al., 2021)) add another level of complexity to the task. Transformer-based (Vaswani et al., 2017) architectures for NER became popular since the release of the BERT model and they hold the state of the art in NER (Akbik et al., 2018; Nie et al., 2020; Yamada et al., 2020; Wang et al., 2020, 2021). However, while most NER systems have been developed to generally address contemporary news data in high resource languages (e.g. English) (Yamada et al., 2020; Wang et al., 2020, 2021), many challenges remain in NER from short texts with complex entities in low-resource scenarios (Meng et al., 2021).

The SemEval-2022 Task 11, *Multilingual Complex Named Entity Recognition* (MultiCoNER) (Malmasi et al., 2022b) aims at developing complex NER systems for 11 languages. The task focuses on detecting semantically ambiguous and complex entities in short, lowercased, low-context monolingual, and multilingual settings. The languages were: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla. For some languages, an additional track with code-mixed data was offered. The organizers aim at testing the domain and language adaption capability of a NER system. For this reason, NER systems for monolingual settings are also constrained to avoid multilingual prediction models and vice versa.

Multilingual and code-mixed documents have motivated the development of systems based on multilingual Transformer-based architectures (Winata et al., 2021; Fetahu et al., 2021). Also, re-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1630 - 1638

cently, Wang et al. (2021) proposed an approach to NER from short texts by finding external contexts of a sentence. The authors retrieved and selected a set of semantically relevant texts through a search engine, with the original sentence as the query, which proved to be efficient in detecting entities.

In this paper, we propose a robust approach for NER and we tackle the following challenges for the language-specific sub-task, including the multilingual and mixed sub-tasks: (1) short texts: by adding multilingual contexts with and without entities, in order to improve and focus more on the context and less on the other surrounding entities, (2) lowercased texts: we prioritize the usage of uncased language models, and (3) code-mixed and low-resource languages: we add a multilingual language model along with a language-specific language model.

2 Data

For each language, organizers provide fixed-sized training and development sets of $15,300$ and 800 sentences respectively, whereas test sets are composed of an average of $181,418$ ($\sigma = 35,181$) sentences. A multilingual dataset consisting of the concatenation of all monolingual sets is also provided (with a selection of $471,911$ sentences for the test set)[1].

In addition, a smaller code-mixed dataset with entities in a foreign language of the rest of the sentence was given. In this case, $1,500$ sentences constitute the training set, 500 the development set, and $100,000$ the test set. We present a sample annotated example from the English training set in Figure 1.

```
# id d8e7dbc2-6002-452d-8a9e-ec17d6e6d955     domain=train
the _ _ O
vendor _ _ O
members _ _ O
sell _ _ O
street _ _ B-PROD
food _ _ I-PROD
or _ _ O
other _ _ O
goods _ _ O
, _ _ O
such _ _ O
as _ _ O
fresh _ _ O
flowers _ _ O
and _ _ O
toys _ _ O
. _ _ O
```

Figure 1: English annotated example from the training set.

[1]MultiCoNER Dataset can be accessed at https://registry.opendata.aws/multiconer.

To highlight the important difference between the number of phrases in the train and the test sets, we present these numbers in Table 1. The corpus is based on LOWER, MSQ-NER (Bajaj et al., 2016), and ORCAS-NER (Craswell et al., 2020) statements and queries from Wikipedia and Microsoft Bing (Meng et al., 2021). A detailed description and exhaustive statistics of all datasets are presented in Malmasi et al. (2022a,b).

Language	Train	Dev	Test
English			217,818
Spanish			217,887
Dutch			217,337
Russian			217,501
Turkish			136,935
Korean	15,300	800	178,249
Farsi			165,702
German			217,824
Chinese			151,661
Hindi			141,565
Bangla			133,119
Multilingual	168,300	8,800	471,911
Mixed	1,500	500	100,000

Table 1: Number of sentences in the dataset splits per language.

3 Methodology

Our proposed framework consists in augmenting each input sentence with similar contexts taken from the multilingual train collection and a NER model based on a BERT-based pre-trained and fine-tuned language model as an encoder with several Transformer (Vaswani et al., 2017) layers stacked on top. For each language, we consider a language-specific BERT model, as presented in Table 2. We prioritize the use of uncased models, with some exceptions where there were no such models (Dutch, Polish). Finally, we tackle the multilingualism and lack of resources for all languages by concatenating these representations with those provided by a multilingual language model. The methodology is outlined in Figure 2.

Named Entity Recognition Model Our base model was recently proposed for coarse-grained and fine-grained named entity recognition (Boros et al., 2020a,b). The method consists of a hierarchical approach, with a pre-trained and fine-tuned BERT-based encoder (Devlin et al., 2019a). This model consists of a stack of Transformer blocks

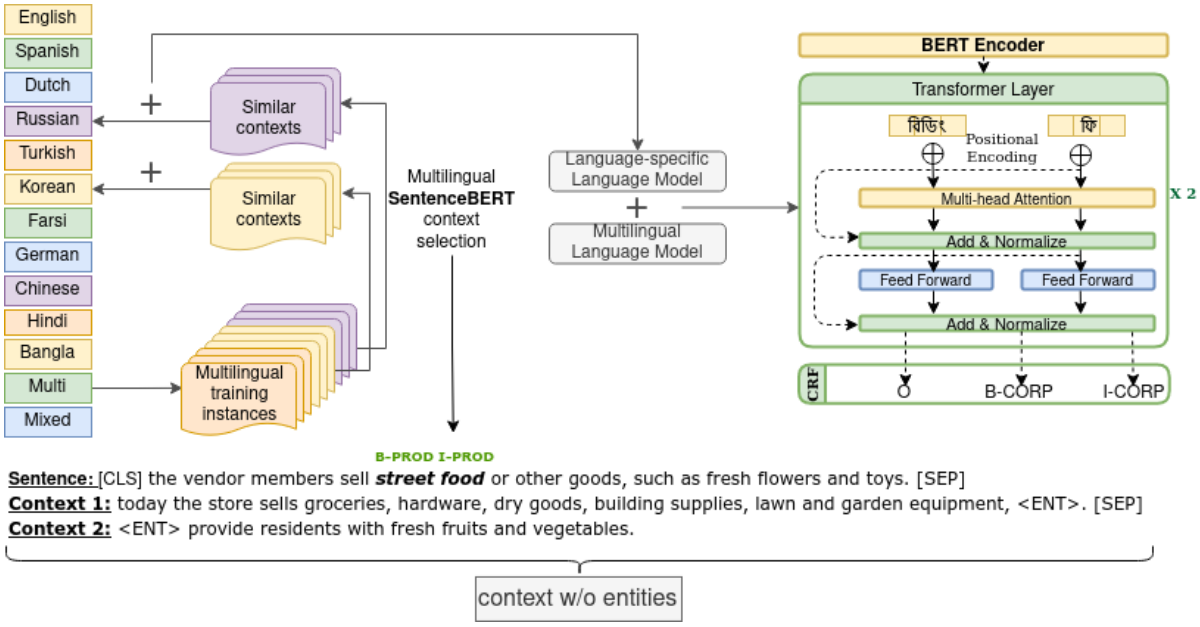

Figure 2: MultiCoNER methodology.

Language	Model
English	`bigbird-roberta-large`
Spanish	`bert-base-spanish-wwm-uncased`
Dutch	`bert-base-dutch-cased`
Russian	`rubert-base-cased`
Turkish	`bert-base-turkish-uncased`
Korean	`bert-kor-base`
Farsi	`bert-fa-base-uncased-persiannews`
German	`bert-base-german-dbmdz-uncased`
Chinese	`bert-base-chinese`
Hindi	`bert-base-multilingual-uncased`
Bangla	`bert-base-multilingual-uncased`
Multilingual	`bert-base-multilingual-uncased`
Mixed	`bert-base-multilingual-uncased`

Table 2: Language-specific models. All models can be found at `https://huggingface.co`.

on top of the BERT encoder, and a conditional random field (CRF) layer to decode the best tag path in all possible tag paths. First, the encoder is already based on a stack of Transformer layers. A Transformer block (encoder) is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings. It is composed of a stack of identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by layer normalization. These language models generally expect the input data in a specific format: a special token, [SEP], to mark the end of a sentence or the separation between two sentences, and [CLS], at the beginning of the text. A Transformer encodes the information contained in a given text in the [CLS] token that flows afterward through the layers. This architecture with additional Transformer layers proved to be efficient when the input is noisy and of varying lengths.

Adding Context First, we encode the collection of training multilingual sentences using a pre-trained multilingual Sentence BERT model (Reimers and Gurevych, 2019, 2020a). We consider that using only sentences from the training set is enough for enriching the contexts, even if they are seen during the training process. The underlying idea behind this is that, since the data is multi-domain, we would expect a very small overlap between the train and the test set, thus the additional contexts could benefit from information from other domains that those existing. Then, for each input sentence (for each dataset split), we rank them using the cosine similarity. From these retrieved texts, we select the first $n \in [1, 10]$ sentences and we consider them as semantically relevant.

Next, as shown in the example in Figure 2 with "street food" as a product (PROD) entity, we concatenate the initial input sentence with the retrieved texts, separated by the special token [SEP]. For this step, we propose two versions of taking advan-

Approach	P	R	F1	P	R	F1
	English			**Spanish**		
Baseline	-	-	61.20	-	-	57.40
BERT+2×T	82.99	86.10	84.52	83.55	80.78	82.14
2×BERT+2×T	86.14	89.92	87.99	85.21	82.31	83.74
2×BERT+2×T-context w/ entities	**87.86**	90.00	88.92	85.80	**85.80**	**85.80**
2×BERT+2×T-context w/o entities	87.83	**90.33**	**89.06**	**86.90**	84.61	85.74
	Dutch			**Russian**		
Baseline	-	-	61.60	-	-	59.10
BERT+2×T	83.73	84.10	83.92	79.33	79.17	79.25
2×BERT+2×T	86.51	86.43	86.47	78.02	**81.77**	**79.85**
2×BERT+2×T-context w/ entities	86.46	**88.24**	**87.34**	**80.00**	77.93	78.95
2×BERT+2×T-context w/o entities	**86.77**	87.29	87.03	78.28	74.38	76.28
	Turkish			**Korean**		
Baseline	-	-	45.70	-	-	54.60
BERT+2×T	84.58	86.35	85.45	83.30	83.49	83.39
2×BERT+2×T	84.67	**87.39**	**86.01**	82.22	84.87	83.52
2×BERT+2×T-context w/ entities	84.11	87.15	85.60	83.17	84.25	83.71
2×BERT+2×T-context w/o entities	**85.03**	86.67	85.84	**86.58**	**87.71**	**87.14**
	Farsi			**German**		
Baseline	-	-	51.80	-	-	63.40
BERT+2×T	77.65	**82.19**	79.86	89.47	89.18	89.33
2×BERT+2×T	**80.69**	81.29	**80.99**	89.86	90.15	90.01
2×BERT+2×T-context w/ entities	78.99	80.87	79.92	**90.24**	**91.04**	**90.64**
2×BERT+2×T-context w/o entities	78.18	82.11	80.10	89.94	90.96	90.45
	Chinese			**Hindi**		
Baseline	-	-	51.10	-	-	46.90
BERT+2×T	**88.20**	88.13	**88.17**	72.47	70.89	71.67
2×BERT+2×T	88.05	86.89	87.47	**77.57**	75.60	**76.57**
2×BERT+2×T-context w/ entities	86.74	86.81	86.77	76.04	75.12	75.58
2×BERT+2×T-context w/o entities	87.67	**88.29**	87.98	74.04	**76.81**	75.40
	Bangla			**Multilingual**		
Baseline	-	-	39.10	-	-	54.10
BERT+2×T	71.94	69.88	70.89	84.54	85.29	84.91
2×BERT+2×T	76.14	75.00	75.57	**85.76**	**86.66**	**86.21**
2×BERT+2×T-context w/ entities	**78.79**	**79.88**	**79.33**	84.91	85.10	85.00
2×BERT+2×T-context w/o entities	77.49	78.75	78.12	84.99	86.09	85.54
	Mixed					
Baseline	-	-	58.10	-	-	-
BERT+2×T	72.33	71.15	71.74	-	-	-
2×BERT+2×T	71.85	71.97	71.91	-	-	-
2×BERT+2×T-context w/ entities	71.97	71.97	71.97	-	-	-
2×BERT+2×T-context w/o entities	**73.92**	**72.95**	**73.43**	-	-	-

Table 3: MultiCoNER results on the development set.

tage of this additional context. First, we consider that simply concatenating the semantically relevant contexts as they are with the sentence in question are sufficient for bringing an improvement to the detection of entities. These models are referred to as *context w/ entities*.

We also examine the scenario where these additional contexts are added without any entity present, as shown in Figure 2 where the entities from the additional contexts are replaced with a special to-

ken <ENT> (Boros et al., 2021, 2022)). The idea behind this choice is that we are trying to improve the contextual information without confusing the detection of entities with the presence of others. We refer to these models as *context w/o entities*.

For finding semantically relevant contexts, we used a multilingual Sentence-BERT (SBERT) (Reimers and Gurevych, 2020b) model[2], a modified pre-trained BERT (Devlin et al., 2019b) that uses a siamese and triplet network structure to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. We keep the top-10 semantically relevant contexts.

Hyperparameters For each language, including the multilingual and the code-mixed cases, we choose a language-specific BERT pre-trained model[3], as presented in Table 2, with the exception for Hindi and Bangla. For the models for which we do not have a language-specific model, we use multilingual BERT. For each language, we additionally use a multilingual model (XLM-RoBERTa-large (Conneau et al., 2020)), approaches referred as BERT×2 in Table 3.

4 Experiments

Next, we perform a detailed error analysis of our approaches[4]. The evaluation is performed in terms of macro precision (P), recall (R), and F1. Our results are presented in Table 3. Each type of approach is detailed with the corresponding pre-trained models.

Table 3 presents the results for our preliminary results on the provided dev set. We, first, observe that all our results considerably outperform the baseline scores provided by the organizers. Also, overall, it is clear that adding a multilingual language model to an already language-specific model brings an increase in performance. Next, we notice that adding *context w/ entities* slightly improved the performance of a limited number of languages (English, Dutch, Korean, German, Bangla) and *context w/o* improved considerably the performance of a larger number of languages (English, Spanish, Dutch, Korean, German, Chinese, Bangla). This allows us to understand that the presence of entities in the additional contexts can influence the detection of the entities of interest by hindering the importance of context with other entities. Thus, adding context brings performance improvements (marginally, with entities, or considerably, without entities). However, the fact that we used the already seen contexts from similar domains or topics (multilingual train data) could also be a factor that contributed to the drop in F1 for some of the languages.

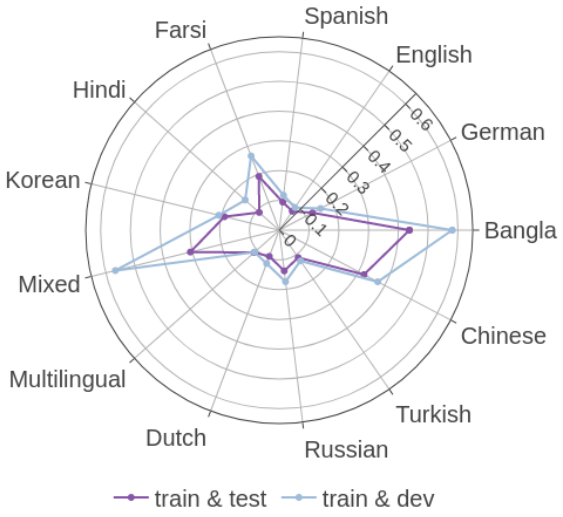

Figure 3: Similarity scores between the topics of the train and test (purple) & train and dev (blue) sets for each language.

Error Analysis Since one of the main challenges and purposes of SemEval-2022 Task 11 is to test not only the language adaption, but also the domain capability of a NER, we analyze the ability of our proposed models to handle multi-domain data. To measure the domain distribution among datasets, we obtain their topic vectors by a joint representation of documents and word semantic embeddings as in Angelov (2020). For this, we first obtain a common embedding of sentences and words with a pre-trained multilingual Sentence-BERT model[5] (Reimers and Gurevych, 2020a). Second, we utilize UMAP (McInnes et al., 2018) to create embeddings of lower dimension for all sentence vectors and find dense areas of sentences with HDBSCAN (Campello et al., 2013). Finally, we calculate the centroid of each dense area that corresponds to the topic vector.

Topic Detection Hyperparameters We set UMAP to a 5-dimensional space with a default

[2]We used the MULTI-QA-MPNET-BASE-DOT-V1 model.

[3]All models can be found at https://huggingface.co.

[4]Our code is available at https://github.com/EMBEDDIA/stacked-ner

[5]We use the PARAPHRASE-MULTILINGUAL-MINILM-L12-V2 model.

Split	Hyperparameter		
		neigh	dist
	2	5	0.0
Train	Mixed	Bangla	Mixed, Dutch, Chinese, Bangla
Dev	Korean, Mixed, Bangla	Farsi, Dutch	Mixed, Dutch, Chinese, Bangla, Korean
Test	-	-	Mixed, Chinese

Table 4: UMAP hyperparameters.

size of the local neighborhood (neigh) of 15, an effective minimum distance between embedded points (dist) equal to 0.1, and cosine distance as a similarity metric. We tuned these hyperparameters depending on the language and split when the obtained number of topic vectors was less than 2. These values are summarized in Table 4. Regarding HDBSCAN, we fix the minimum number of samples in a cluster to 15.

Language	Domains / Topics		
	Train	**Dev**	**Test**
English	72	12	1,188
Spanish	99	8	1,274
Dutch	112	9	1,404
Russian	95	6	1,330
Turkish	101	8	1,088
Korean	107	10	854
Farsi	126	5	1,097
German	93	6	1,194
Chinese	89	8	607
Hindi	105	5	1,209
Bangla	55	6	142
Multilingual	573	55	1,551
Mixed	32	11	644

Table 5: Number of topics in the dataset splits per language.

As seen in Table 5, the number of topics within the train set for each language is on average 96. The multilingual dataset has roughly six times more topics than the monolingual dataset even though it is the concatenation of all monolingual datasets. This suggests that certain portions of the monolingual datasets were obtained by a translation mechanism. Topic diversity from the development set is clearly smaller than the train set given the limited amount of samples it contains. Monolingual test sets present on average 11 times more topics with respect to the train sets, which is explained by the big amount of out-of-domain data organizers added to test sets with the objective of measuring out-of-domain performance.

To estimate the amount of out-of-domain sentences within data sets, we compute the topic overlap between (train and dev) and (train and test) sets. For each topic vector in the train set, we compute the cosine similarity with the topic vectors of the corresponding test or dev set. Then, we compare the topics by calculating the mean of the cosine similarities between the two sets of topics, as shown in Figure 3. We observe that Bangla and Mixed have very similar topics in the train, test, and dev sets. Our results in Table 3 seem to prove otherwise, thus, we interpret this behavior as the result of the lack of training of the multilingual Sentence-BERT model for this language.

Table 3 shows improvements when adding contexts for English, Spanish, Dutch, Korean, Bangla, and Mixed. With the exception of Bangla and Mixed that we just discussed, we notice that for the other languages, the similarity between the topics is rather small (around 0.2). Table 3 also shows a decrease in performance for Russian, Turkish, Farsi, Chinese, and Hindi. While for Farsi and Chinese, the similarity of topics is slightly higher (between 0.3 and 0.4), for the other languages, it plateaus at 0.2, which could indicate that there is a correlation between the number of overlapping topics between the train, test, and dev sets.

Metric	Correlation (train, dev)	Correlation (train, test)
P	-0.3812	-0.4995
R	-0.3450	-0.3266
F1	-0.3648	-0.4070

Table 6: Pearson correlation values between the dev and train topic similarity and evaluation metrics.

We, therefore, decide to compute the Pearson correlation between the precision, recall, and F1, and the similarity between topics. We observe weak negative correlation coefficients, allowing us to understand that the higher the topical similarity, the lower the performance scores. These

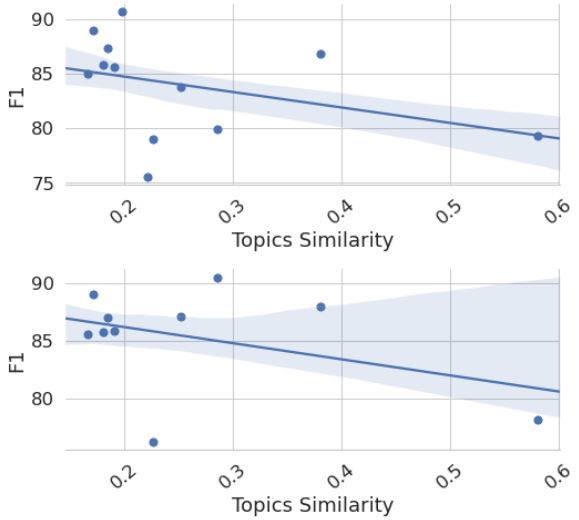

Figure 4: Topics similarity between train and dev set, 2×BERT+2×T-context *w/ entities* (upper figure), train and dev sets, 2×BERT+2×T-context *w/o entities* (lower figure), versus the F1 scores for the context models.

correlation values are shown in Table 6. Moreover, in order to understand if there is a correlation between the number of overlapping topics and the F1 scores for the models that use *context w/* or *w/o entities*, we draw a scatterplot of these two variables, then fit a regression model and plot the resulting regression line and a 95% confidence interval for that regression, in Figure 4. We observe again that, generally, the higher the similarity between topics in test and dev, the lower the F1 scores are.

SemEval-2022 Task 11 In the official SemEval-2022 Task 11, our best results ranked above the average for English, Spanish, and German, close to the average for Dutch, Turkish, Farsi, and Chinese. For the other languages, our approach obtained lower scores (Russian, Korean, Hindi, Bangla, Multilingual, and Mixed). For Hindi and Bangla, we did not have a specialized language model (we used multilingual BERT), which is clearly another reason for which the results were the lowest. Interestingly, we expected higher results for Korean and Spanish, but the size of the test set was most probably another important factor to consider.

5 Conclusions

In this paper, we presented a straightforward approach for adding semantically relevant contexts for NER in the monolingual, multilingual, and code-mixed datasets provided by SemEval-2022 Task 11 *Multilingual Complex Named Entity Recognition* (MultiCoNER). Our findings show that, while adding contexts from the train set, with and without entities, is promising, the topics or domains overlap could influence the performance in both directions. Future work will include the automatic generation of semantically relevant contexts without the presence of entities.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (EMBEDDIA), and by the ANNA and Termitrad projects funded by the Nouvelle-Aquitaine Region.

References

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Gabriel Bernier-Colborne and Philippe Langlais. 2020. Hardeval: Focusing on challenging tokens to assess robustness of ner. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1704–1711.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020a. Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, pages 431–441, Online. Association for Computational Linguistics.

Emanuela Boros, Jose G Moreno, and Antoine Doucet. 2021. Event detection with entity markers. In *European Conference on Information Retrieval*, pages 233–240. Springer.

Emanuela Boros, José G Moreno, and Antoine Doucet. 2022. Exploring entities in event detection as question answering. In *European Conference on Information Retrieval*, pages 65–79. Springer.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. 2020b. Robust named entity recognition and linking on historical multilingual documents. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, pages 1–17, Thessaloniki, Greece. CEUR-WS.

Emanuela Boros, Verónica Romero, Martin Maarand, Kateřina Zenklová, Jitka Křečková, Enrique Vidal, Dominique Stutzmann, and Christopher Kermorvant. 2020c. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International conference on frontiers in handwriting recognition (ICFHR)*, pages 79–84. IEEE.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2983–2989.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, 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.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on 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.

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.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. *arXiv preprint arXiv:1903.11222*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving named entity recognition with attentive ensemble of syntactic information. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4231–4245, Online. Association for Computational Linguistics.

Lakshya Rajoria. 2021. Named entity recognition in tweets. *International Journal of Research in Engineering, Science and Management*, 4(1):43–50.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

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*.