



HAL
open science

L3i++ at SemEval-2023 Task 2: Prompting for Multilingual Complex Named Entity Recognition

Carlos-Emiliano González-Gallardo, Tran Thi Hong Hanh, Nancy Girdhar, Emanuela Boros, Jose G. Moreno, Antoine Doucet

► **To cite this version:**

Carlos-Emiliano González-Gallardo, Tran Thi Hong Hanh, Nancy Girdhar, Emanuela Boros, Jose G. Moreno, et al.. L3i++ at SemEval-2023 Task 2: Prompting for Multilingual Complex Named Entity Recognition. 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Jun 2023, New Mexico, United States. hal-04676751

HAL Id: hal-04676751

<https://hal.science/hal-04676751>

Submitted on 23 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

L3i++ at SemEval-2023 Task 2: Prompting for Multilingual Complex Named Entity Recognition

Carlos-Emiliano González-Gallardo¹ and Thi Hong Hanh Tran^{1,2} and Nancy Girdhar¹
and Emanuela Boros¹ and Jose G. Moreno^{1,3} and Antoine Doucet¹

¹ University of La Rochelle, L3i, La Rochelle, France

² Jožef Stefan Institute, Ljubljana, Slovenia

³ University of Toulouse, IRIT, Toulouse, France

{firstname.lastname}@univ-lr.fr

jose.moreno@irit.fr

Abstract

This paper summarizes the participation of the L3i laboratory of the University of La Rochelle in the SemEval-2023 Task 2, *Multilingual Complex Named Entity Recognition* (MultiCoNER II). Similar to MultiCoNER I, the task seeks to develop methods to detect semantic ambiguous and complex entities in short and low-context settings. However, MultiCoNER II adds a fine-grained entity taxonomy with over 30 entity types and corrupted data on the test partitions. We approach these complications following prompt-based learning as (1) a ranking problem using a seq2seq framework, and (2) an extractive question-answering task. Our findings show that even if prompting techniques have a similar recall to fine-tuned hierarchical language model-based encoder methods, precision tends to be more affected.

1 Introduction

In SemEval-2022, *Multilingual Complex Named Entity Recognition* (MultiCoNER) I (Malmasi et al., 2022a) (Task 11) aimed at developing complex NER systems for 11 languages. The task focused on detecting semantically ambiguous and complex entities in short, lowercase, low-context monolingual, and multilingual settings. These challenges needed to be approached, while testing the domain and language adaption capability of the NER systems proposed by the participants. Thus, our participation consisted of a robust approach for tackling the following challenges for the language-specific, multilingual, and mixed sub-tasks (Boroş et al., 2022): (1) *short texts*: we built multilingual contexts with and without entities, in order to improve the focus of the NER system more on the context and less on the other surrounding entities; (2) *lowercase texts*: we prioritized the usage of uncased language models; and (3) *code-mixed and*

low-resource languages: we added a multilingual language model along with a language-specific language model. Our findings showed that, while adding contexts (from train partitions), with and without entities, was promising, the topics or domains overlap influenced the performance in both directions.

In SemEval-2023, MultiCoNER II (Fetahu et al., 2023b) (Task 2) featured complex NER in these languages (slight changes in comparison with the previous MultiCoNER): English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, and German. In addition to multilingualism, there were two new key challenges: (1) a fine-grained entity taxonomy with over 30 different classes, and (2) simulated errors added to the test set to make the task more realistic and difficult. Approaching these challenges would require an ability to easily adapt to the taxonomies of entities that are constantly evolving and to new scenarios with noise (for which there are few or no labelled data). Recently, a new paradigm based on prompt-based learning (*prompt engineering*) has achieved great success in information extraction tasks by reformulating classification tasks as close questions (Li et al., 2019; Liu et al., 2022; Boros et al., 2022).

Inspired by the advent of prompt engineering (Liu et al., 2023) or specifically template-based prompting methods (Cui et al., 2021; Liu et al., 2022), we propose to treat the NER task as (1) a ranking problem using the sequence-to-sequence (seq2seq) framework, and (2) an extractive question-answering (QA) task. In the scope of our research (and due to several constraints), we focused only on five languages.

In Section 2, we detail the data in the five languages we chose to participate in this year and the *prompt engineering*-based methods for approaching NER in Section 3. We then point out our findings and perform an error analysis in Section 4. We

present several similar works in Section 5. Finally, in Section 6 we present our conclusions, future work and discuss the limitations of the proposed methods.

2 Data

We utilized the MultiCoNER II corpus built by the organizers to challenge systems on the complex NER task (Fetahu et al., 2023a), where entities are not only standard entity classes such as locations, persons, organizations, etc., but also titles of creative works or medical-related terms. The corpus comprises documents in twelve languages and a multilingual dataset corresponding to the concatenation of all monolingual datasets. Entities are organized using a taxonomy of two levels, with six entries at the top level (coarse-grained) and 36 at the bottom level (fine-grained). However, in practice, only 33 are used.

Table 1 presents the dataset statistics for the five languages (Bangla, German, English, Hindi, and Chinese) and the multilingual dataset in which we participated. Test partitions are generally twice as large as train partitions, except for the English dataset, where the test partition is almost 15 times larger than the train one. For all datasets, train partitions are 19 times larger than the development ones. We detected the existence of repeated sentences at different rates among all train partitions. A detailed description and statistics of the MultiCoNER II corpus can be consulted in Fetahu et al. (2023a).

Language		# Sentences	# Tokens	# Entities
Bangla	train	9,708	129,244	13,046
	dev	507	6,729	668
	test	19,859	254,802	25,013
German	train	9,785	145,454	15,457
	dev	512	7,361	813
	test	20,145	287,007	28,877
English	train	16,778	252,352	25,038
	dev	871	13,298	1,278
	test	249,980	3,764,453	377,805
Hindi	train	9,632	158,483	12,737
	dev	514	8,592	673
	test	18,399	293,552	23,199
Chinese	train	9,759	244,715	15,188
	dev	506	12,770	769
	test	20,265	487,740	27,999
Multilingual	train	170,824	2,672,490	254,919
	dev	8,895	139,823	13,305
	test	358,668	5,537,072	522,691

Table 1: MultiCoNER II corpora statistics for train, development, and test partitions.

3 Methodology

In this section, we investigate the repercussion of semantically ambiguous and complex named entities on two prompt-based NER methods and their feasibility to be used in noisy and multilingual settings. To position the performance of the explored methods, we set a strong baseline, which has shown to be effective even on noisy datasets.

3.1 Stacked Transformers (StackedNER)

Our baseline model is the StackedNER architecture that corresponds to the BERT+2×T model (Boroş et al., 2022) from our previous participation in MultiCoNER I (Malmasi et al., 2022b). StackedNER comprises a hierarchical structure, with a pre-trained and fine-tuned BERT-based encoder. The encoder consists of a stack of Transformer blocks and a conditional random field (CRF) prediction layer that is applied to decode the best tag path from all possible tag paths. Enhancing the model by leveraging transfer learning with additional Transformer layers has proven to be effective in handling noisy and varied input lengths, as well as detecting coarse-grained and fine-grained named entities (Boros et al., 2020a,b).

Hyperparameters We used the same hyperparameters as by Boroş et al. (2022). As for the pre-trained language model, we used bert-base-multilingual-uncased¹.

3.2 Templates for NER (TemplateNER)

The first prompt-based NER method considers the task as a ranking problem using the sequence-to-sequence (*seq2seq*) framework, where the original sentences are the source sequence and the templates filled by the candidate named entities span are the target sequence during the training. During inference, each candidate span is classified based on the template score.

We chose BART (Lewis et al., 2019), a denoising autoencoder for pretraining seq2seq models and fine-tuned it for English, while for other languages, we applied mBART (Liu et al., 2020), the multilingual version that was pre-trained on large-scale monolingual corpora in many languages using the same BART objective. The templates that we proposed are shown in Figure 1.

Furthermore, we tested the scenarios of using an English template and apply to another language

¹<https://huggingface.co/bert-base-multilingual-uncased>

Template	Fine-grained			Coarse-grained		
	P	R	F1	P	R	F1
(1) <i>is a/an</i> (2) <i>entity</i>	0.193	0.208	0.169	0.235	0.327	0.250
(1) <i>belongs to</i> (2) <i>category</i>	0.153	0.202	0.151	0.178	0.303	0.207
(1) <i>should be tagged as</i> (2)	0.151	0.188	0.143	0.158	0.280	0.190
(1) <i>gehört zur Kategorie</i> (2)	0.166	0.004	0.007	0.583	0.008	0.015
(1) <i>sollte als</i> (2) <i>gekennzeichnet werden</i>	0.121	0.113	0.110	0.391	0.167	0.222

Table 2: Evaluation of different template languages and templates in German, where (1) refers to candidate span and (2) refers to the entity type.

Bangla	(1) (2) এর অন্তর্গত
German	(1) gehört zur Kategorie (2)
English	(1) belongs to (2) category
Hindi	(1) (2) से संबंधित है
Chinese	(1) 屬於 (2) 類

Figure 1: The template form for each language, where (1) refers to candidate span and (2) refers to the entity type. All the templates are translated from the original English version (1) *belongs to* (2) *category*.

(e.g., German) as well as manually selecting different templates for a language (some variants include (1) *is a/an* (2) *entity* or (1) *should be tagged as* (2)). We also tested with different templates and transfer templates from one language to another one. The results are demonstrated in Table 2. The design of templates helps the models better fit our specific task. Further approaches to selecting the proper templates (e.g., gradient-based discrete searching, language model generating) will be left for future work.

Hyperparameters We experimented with BART² (Lewis et al., 2019) for English and mBART³ (Tang et al., 2020) for the other languages, with a batch size of 32, 5 epochs, and a max sequence length of 70.

3.3 Question Answering for NER (QaNER)

The second approach treats NER as an extractive question-answering (QA) task (QaNER). Given a context C and a question Q , the goal is to extract

²<https://huggingface.co/facebook/bart-large>

³<https://huggingface.co/facebook/mbart-large-50-many-to-one-mmt>

the answer A as a sub-string of C . Thus, a sample is composed of the triple (C, Q, A) . Similar to Liu et al. (2022), we utilize a pre-trained BERT-based encoder to represent the concatenated Q and C . Then, a start and end token classifier layer maps the encoded text to a predicted span corresponding to the start and end of the answer A .

To obtain named entities from an extractive QA architecture, questions such as *Where is the location?*, *Who is the person?* or *What is the product?* are needed to be prompted. In these examples, the adverbs *where*, *who*, and *what* depend on the entity types *location*, *person*, and *product*. Liu et al. (2022) explored different ways of constructing these prompts and concluded that manually deciding which adverb must precede an entity type produced better results for coarse-grained taxonomies with no more than 12 classes. The case of the MultiCoNER II corpus with more than 30 entity types is different. Our preliminary experiments over the English dataset showed that automatically prompting the template [MASK] *is the* [ENTITY] ? ([ENTITY] is replaced by each one of the 36 fine-grained entity types) into the BERT-based encoder to obtain the corresponding 5W1H (i.e., *who*, *what*, *when*, *where*, *why*) question word produced the best results. We thus decided to follow this procedure to perform all our experiments.

To train our QA system, we first transformed each phrase from the MultiCoNER II corpus into multiple triples (C, Q, A) , one for each entity type. It is possible for a phrase to have multiple answers for one entity type, in those cases, multiple triples for the same question are needed. For those entity types not present in the phrase, we assigned the classification [CLS] token for the start and end of the answer span. At the end of this process, we obtained at least 36 triples from one phrase of the original training set. Then, for each sample (C, Q, A) with Q and C concatenated by [CLS],

Approach	Fine-grained			Coarse-grained		
	P	R	F1	P	R	F1
Bangla						
StackedNER	0.626	0.471	0.498	0.737	0.665	0.696
QaNER	0.487	0.607	0.506	0.531	0.764	0.623
TemplateNER	0.180	0.286	0.203	0.380	0.367	0.345
German						
StackedNER	0.628	0.542	0.547	0.779	0.748	0.763
QaNER	0.481	0.584	0.516	0.621	0.776	0.686
TemplateNER	0.107	0.226	0.122	0.339	0.442	0.359
English						
StackedNER	0.282	0.223	0.229	0.668	0.471	0.520
QaNER	0.517	0.677	0.573	0.601	0.820	0.686
TemplateNER*	0.320	0.423	0.310	0.393	0.676	0.483
Hindi						
StackedNER	0.489	0.398	0.415	0.669	0.585	0.621
QaNER	0.461	0.595	0.505	0.535	0.736	0.611
TemplateNER	0.216	0.258	0.207	0.471	0.370	0.395
Chinese						
StackedNER	0.586	0.485	0.503	0.780	0.709	0.736
QaNER	0.372	0.519	0.425	0.490	0.682	0.565
TemplateNER	0.127	0.117	0.101	0.246	0.181	0.199
Multilingual						
StackedNER	0.705	0.683	0.686	0.809	0.811	0.810
QaNER	0.413	0.627	0.480	0.476	0.785	0.582
TemplateNER	0.128	0.312	0.170	0.336	0.440	0.351

Table 3: Evaluation of different approaches to extract complex named entities (dev). The best score per language is marked in bold. * BART (mBART for the other languages)

the training procedure consists of maximizing the likelihood that two tokens located at the right side of [CLS] correspond to the beginning and end of the answer A.

During entity prediction of a given phrase C , we first concatenated the corresponding Q of all entity types with C , then we prompted the system to obtain the tokens indicating the beginning and end of each entity type or the [CLS] token representing the non-existence of that named entity in the phrase. Multiple answer spans were possible for one entity type, therefore we checked for overlaps and retain the spans that obtained the highest prediction scores without overlapping. When two or more entity types were predicted for the same span, we kept the entity type with the highest score.

Hyperparameters We experimented with the multilingual XLM-RoBERTa⁴ (Conneau et al., 2019) pre-trained model for all languages except English in which we used BigBird⁵ (Zaheer et al., 2021). We set to 64 tokens the maximum sequence length with a stride of 16, learning rate to 2×10^{-5} , weight decay equal to 0.01, and 4 epochs.

⁴<https://huggingface.co/xlm-roberta-base>

⁵<https://huggingface.co/google/bigbird-roberta-base>

Approach	Fine-grained			Coarse-grained		
	P	R	F1	P	R	F1
Bangla						
StackedNER	0.526	0.412	0.425	0.697	0.632	0.655
QaNER	0.364	0.557	0.413	0.489	0.714	0.574
TemplateNER	0.141	0.291	0.180	0.401	0.385	0.373
German						
StackedNER	0.571	0.529	0.540	0.779	0.729	0.753
QaNER	0.413	0.578	0.466	0.577	0.756	0.649
TemplateNER	0.193	0.208	0.169	0.235	0.327	0.250
English						
StackedNER	0.571	0.413	0.449	0.747	0.642	0.685
QaNER	0.476	0.638	0.530	0.572	0.784	0.655
TemplateNER*	0.392	0.563	0.450	0.476	0.702	0.560
Hindi						
StackedNER	0.492	0.416	0.437	0.710	0.610	0.651
QaNER	0.382	0.581	0.436	0.522	0.734	0.603
TemplateNER	0.173	0.287	0.206	0.474	0.380	0.390
Chinese						
StackedNER	0.500	0.473	0.479	0.707	0.657	0.680
QaNER	0.296	0.482	0.353	0.420	0.647	0.500
TemplateNER	0.105	0.127	0.110	0.281	0.199	0.215
Multilingual						
StackedNER	0.653	0.658	0.654	0.781	0.769	0.775
QaNER	0.368	0.612	0.444	0.449	0.767	0.557
TemplateNER	–	–	–	–	–	–

Table 4: Evaluation of different approaches to extract complex named entities (test). The best score per language is marked in bold. * BART (mBART for the other languages)

4 Results

The performance of the systems for the NER task was evaluated using macro precision (P), recall (R), and F-score (F1). Tables 3 and 4 display the performance of all systems for both fine- and coarse-grained entities over the development and the test partitions, respectively.

4.1 General Observations

StackedNER ranks first for most of the results on the development partition for the coarse-grained (five out of six). Moreover, QaNER manages to outperform StackedNER in half of the languages for the fine-grained entities. TemplateNER does not perform well regardless of the language or the kind of entity. However, recall performances are interestingly dominated by QaNER in coarse- and fine-grained for most languages, except for Chinese and Multilingual. From these results, we can evidently conclude that StackedNER is a strong baseline, but QaNER is an interesting alternative that may be combined with a more precise NER to positively impact the final result. Similar results are observed on the test partition (Table 4), but with a clear dominance by StackedNER in all

configurations.

4.2 Error Analysis

Impact of Entity Type One new challenge proposed in MultiCoNER II is the introduction of a fine-grained entity taxonomy with 33 different classes. From Tables 3 and 4 we can conclude that this class diversity clearly impacts the systems’ performance, especially in terms of precision. To deepen the analysis, we display in Figure 2 the behaviour of QaNER for each class, and language in terms of precision, recall and F1.

We can observe that classes such as OtherLOC, Sympton, AnatomicalStructure, MedicalProcedure, OtherPROD, Drink, Food, Vehicle, Clothing, OtherPER, and ArtWork are problematic for all languages and present precision scores lower than 0.5. The case of Bangla is interesting, the system gives the impression of being precise for PrivateCorp (a precision of 0.750), however, it misses more than 83% of the entities corresponding to this class. On average, HumanSettlement, SportsGRP, and Station obtained the best scores among all languages in terms of F1, while PrivateCorp, Sympton, and Artwork were poorly classified.

Impact of Noise The introduction of noisy subsets for some of the languages (i.e., English and Chinese) consisted in the corruption of context or entity tokens in 30% of the sentences. Figure 3 shows the performance of the QaNER system for clean and noisy subsets in terms of F1. We observe that the noisy tokens have a negative impact on both languages, nevertheless, it is more evident in Chinese. A clear example is the ArtWork entity type, which decreases abruptly from 0.303 to 0.028. We interpret this behaviour as the result of two factors. First, 82% of the noisy tokens are present on entity tokens, compared to 76% for English. Second, QaNER obtained the lowest performance for Chinese with clean sentences, showing its limited capacity to handle it.

Impact of Domain Shift MultiCoNER is focused on challenging NER models with complex named entities, multilingualism, and cross-domain adaptation capabilities (Malmasi et al., 2022b; Fetahu et al., 2023b). To quantify the domain distribution among partitions, we obtained topic vectors by a joint representation of sentences and token semantic embeddings with Top2Vec (Angelov, 2020). We first projected all words and sentences

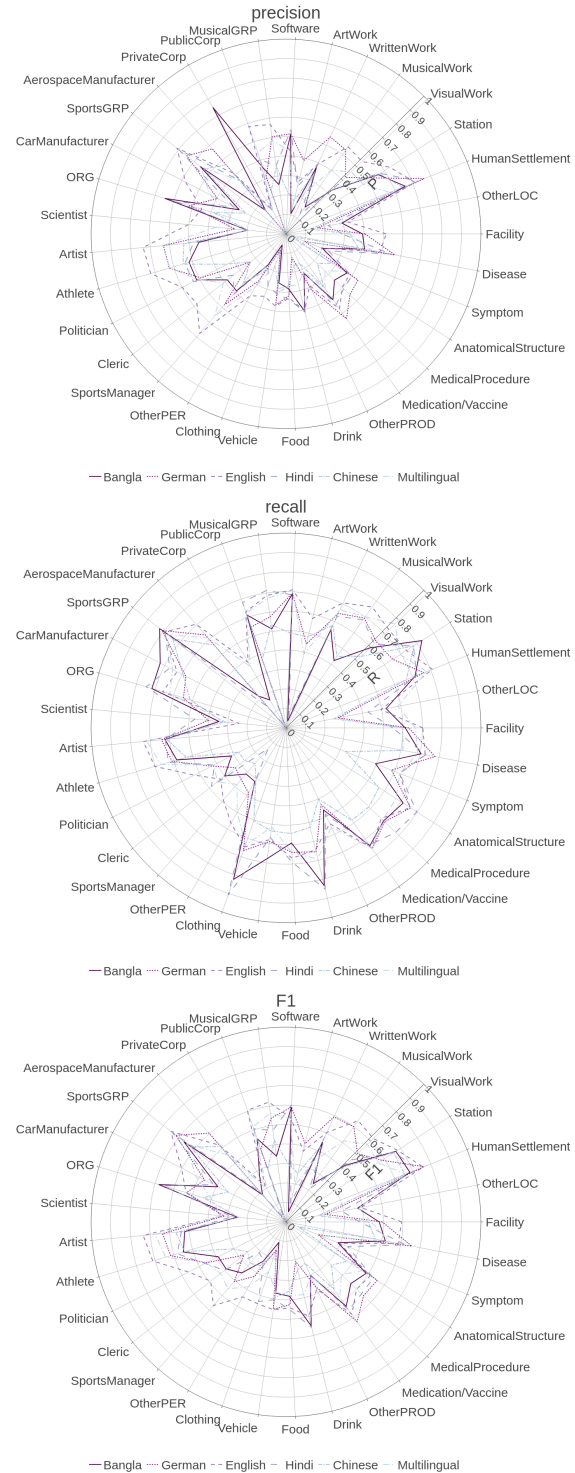


Figure 2: Fine-grained QaNER performance (test).

into a common multidimensional space with a multilingual Sentence-BERT (Reimers and Gurevych, 2020) pre-trained model (paraphrase-multilingual-MiniLM-L12-v2⁶).

Then, to reduce dimensionality, we applied UMAP to the sentence vectors (McInnes et al.,

⁶<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

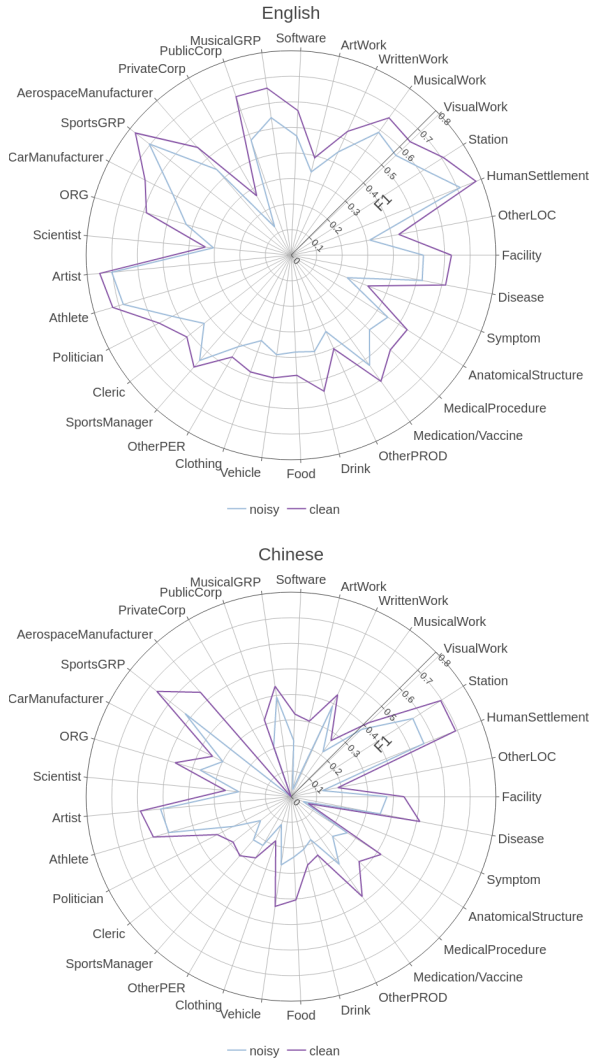


Figure 3: QaNER F1 performance over noisy and clean subsets (test).

2018) followed by HDBSCAN (Campello et al., 2013) to find dense areas of sentences. Finally, to obtain the topic vectors, we calculated the centroid of each dense area. A topic is therefore characterized by the words and sentences present in the vicinity of its centroid. We kept the same hyperparameters as set by Boroş et al. (2022). From Tables 1 and 5 we can notice that the topic variation present on each partition depends on the number of sentences that constitute them.

In order to measure the presence of out-of-domain sentences between data partitions, we computed the topic overlap between (train & dev) and (train & test) partitions. First, for each topic vector in a train partition, we calculated its cosine similarity against all topic vectors from development and test partitions. We then compared the topics by calculating the mean of the cosine similarities

Language	Domains / Topics		
	train	dev	test
Bangla	28	3	49
German	95	2	124
English	78	7	782
Hindi	94	7	158
Chinese	51	5	107
Multilingual	617	52	1,116

Table 5: Number of topics per dataset partition.

between the sets of topics. The resulting topic similarities are shown in Figure 4. Multilingual presents the smallest similarities, which is an expected behaviour given the mixture of languages and the number of sentences. Bangla presents the highest topic similarity while having the smallest number of topics. This shows uniformity in the distribution of its topics which has shown to be beneficial for certain entity types such as PrivateCorp, ORG and Station.

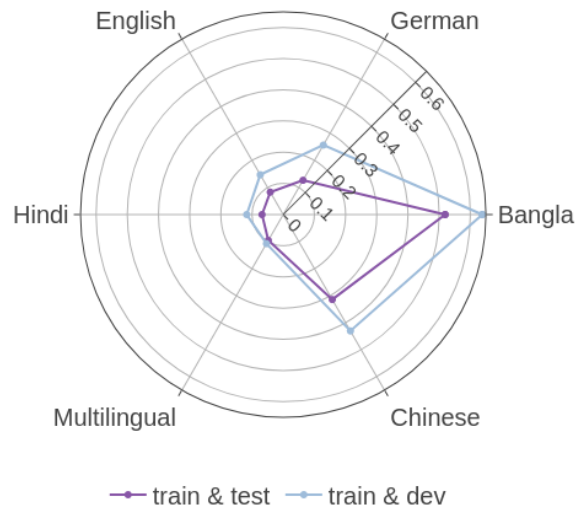


Figure 4: Topic similarity between train & test (purple) and train & dev (blue) partitions.

5 Related Work

Similar to our work, Chen et al. (2022) approached NER as a generation problem and introduced LightNER, a tuning paradigm for NER that adopts a pluggable prompting method to improve NER performance in low-resource settings by freezing most of the generation model parameters. Their experimental results showed that LightNER obtained comparable performance in the standard supervised setting and outperformed strong baselines in low-

resource settings.

Template-based NER was explored by Cui et al. (2021), where NER was approached as a language model ranking problem under a sequence-to-sequence framework (generative pre-trained language model). The source sequences corresponded to the original sentences, while the target sequences were statement templates filled by the candidate named entity spans. During entity prediction, the model classified each span based on the corresponding template scores. Their experiments demonstrated that the proposed method performed significantly better than fine-tuning BERT.

Concerning QA, Liu et al. (2022) explored a prompted-based learning NER method based on extractive QA. In this approach, each entity type was formulated into a question and prompted into the NER system to obtain the offset within the input context corresponding to the tokens of the entity type. The method showed to be faster and more robust in low-resource conditions compared to other prompting methods.

6 Conclusions

In this paper, we investigated the performance of two prompt-based NER methods over five monolingual and multilingual datasets provided by the SemEval-2023 Task 2, *Multilingual Complex Named Entity Recognition* (MultiCoNER II). Our findings show that semantically ambiguous and complex entities in short and low-context settings are challenging for prompting techniques, affecting mostly their precision. These techniques were also influenced by the entity type taxonomy and context size, which increased the training and inference time. In future work, we will explore other methods to select proper templates and include knowledge bases to add informative context that helps discern the correct entity type.

Limitations

The TemplateNER approach has several limitations due to two aspects. First, it is time-consuming to generate the template for all the possible spans at the sentence level. Here, we only consider the named entities that constitute up to four words (1-4 grams) to reduce the time cost of the models. As the number of entities in a sentence is much fewer than non-entities, there is a significant imbalance between the number of templates enumerated for entities and the non-entities, which might affect the

performance. Second, if the length of the sentence increases, the decoding time would rise catastrophically. Further improvements are left for future development.

Regarding QaNER, the number of samples is multiplied by the number of entity types, which could be expensive, especially for cases when a fine-grained entity prediction is needed. It is also important to take into consideration the word or words used to replace [ENTITY] in the template [MASK] is the [ENTITY]?. Entity types such as ArtWork present possible substitutions (i.e., [MASK] is the artwork?, [MASK] is the work of art?) which have an impact on the encoding done by the pre-trained language model.

Ethics Statement

We noticed that MultiCoNER II has taken a political stance. Due to this reason and also, the size of some language-specific datasets, we did not participate in some of the languages.

Acknowledgements

This work has been supported by the ANNA (2019-1R40226) and TERMITRAD (2020-2019-8510010) projects funded by the Nouvelle-Aquitaine Region, France. We would also like to thank our colleagues Wenjun Sun, Nicolas Sidère, Mickaël Coustaty, Beatriz Martínez Tornés, and Souhail Bakkali.

References

- Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#).
- Emanuela Boros, Carlos-Emiliano González-Gallardo, José G Moreno, and Antoine Doucet. 2022. L3i at semeval-2022 task 11: Straightforward additional context for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1630–1638.
- 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, José G Moreno, and Antoine Doucet. 2022. Exploring entities in event detection as ques-

- tion 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.
- 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.
- Xiang Chen, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, Huajun Chen, and Ningyu Zhang. 2022. [LightNER: A lightweight tuning paradigm for low-resource NER via plugable prompting](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2374–2387, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- 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](#). *CoRR*, abs/1911.02116.
- Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using bart. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.
- 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, 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.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. [Entity-relation extraction as multi-turn question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.
- Andy T Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. Qaner: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543*.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. 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.
- 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)*, pages 1412–1437.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.
- Nils Reimers and Iryna Gurevych. 2020. [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.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv e-prints*, pages arXiv–2008.
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. [Big bird: Transformers for longer sequences](#).