garNER at SemEval-2023: Simplified Knowledge Augmentation for Multilingual Complex Named Entity Recognition

Md Zobaer Hossain[♣†*], **Averie Ho Zoen So**[♣], **Silviya Silwal**[♣],
H. Andres Gonzalez Gongora[♣], **Ahnaf Mozib Samin**[◇], **Jahedul Alam Junaed**[♠]
Aritra Mazumder[♠], **Sourav Saha**[♠], **Sabiha Tahsin Soha**[♠]
[♣] University of Lorraine, Nancy, France, [◇] University of Malta, Malta
[♠] Shahjalal University of Science and Technology, Sylhet, Bangladesh
[♣]{hossain6u, so2u, silwal1u, gonzal177u}@etu.univ-lorraine.fr
[◇]ahnaf.samin.22@um.edu.mt
[♠]{jahedul25, aritra73, sourav95, sabiha96}@student.sust.edu

Abstract

This paper presents our solution, garNER, to the SemEval-2023 MultiConer task. We propose a knowledge augmentation approach by directly querying entities from the Wikipedia API and appending the summaries of the entities to the input sentence. These entities are either retrieved from the labeled training set (Gold Entity) or from off-the-shelf entity taggers (Entity Extractor). Ensemble methods are then applied across multiple models to get the final prediction. Our analysis shows that the added contexts are beneficial only when such contexts are relevant to the target-named entities, but detrimental when the contexts are irrelevant.

1 Introduction

MultiCoNER is a shared task (Fetahu et al., 2023b) organized as part of the SemEval workshop to assess the performance of Named Entity Recognition (NER) systems across multiple languages. The task specifically addresses the recognition of *complex* named entities in a multilingual setting, such as product names, event names, and creative works. Complex named entities often have form variations and may resemble normal language usage (eg. "Catch Me If You Can"), making them difficult to recognize with traditional NER techniques. To accurately classify these entities, domain-specific knowledge and advanced language modelling techniques are often considered necessary components of the systems submitted to this task.

The introduction of Transformer-based models has resulted in significant advancements in various Natural Language Processing (NLP) tasks, including Named Entity Recognition (NER). The integration of pre-trained embeddings, trained on lengthy texts, has proven to be an effective technique for modelling long-range dependencies and resolving ambiguity in complex named entities within sentences (Peters et al., 2018; Akbik et al., 2018; Straková et al., 2019).

Among the systems submitted in MultiCoNER 2022, the winning system was DAMO-NLP (Wang et al., 2022), which ranked 1st in most languages (Malmasi et al., 2022b). The idea behind their approach is that incorporating a knowledge base system grants NER models a similar advantage as professional annotators, allowing them to retrieve relevant documents and information from the knowledge base to disambiguate challenging samples in the dataset. The underlying principle of incorporating additional context into the original sentences is to enhance the contextual embeddings and thereby improve the contextual representations of the complex entities.

In our revised method, we follow the same principle of extracting relevant information and employing long-range dependencies to establish robust token representations. However, a significant difference between our approach and that of the DAMO-NLP knowledge-based system lies in our proposal for knowledge enhancement without the requirement for a comprehensive knowledge base. The cost of building a knowledge base is significant, especially in a multilingual context, so we propose a streamlined approach in which, given an entity, we use the Wikipedia API to gather a summary from relevant Wikipedia pages, which is then concatenated with the input sentence. This approach considers the computational resource constraints while striving to maintain an equivalent level of performance for complex named entity recognition tasks.

We fine-tune multilingual and monolingual Transformer models and then apply ensembling techniques to obtain the final predictions. Our best system results are consistently positioned in the middle of the rankings, and they rank in the top ten

[*†] Project Lead. All authors have equal contributions

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 823–835
July 13-14, 2023 ©2023 Association for Computational Linguistics

at ten language tracks. Moreover, our error analysis indicates that the added contexts are only advantageous when they are relevant to the target-named entities, whereas they are disadvantageous when the additional context are irrelevant.

2 Related Work

Pre-trained Transformer-based models (Vaswani et al., 2017) can leverage extensive sources of knowledge, such as Wikipedia and social media, to achieve state-of-the-art (SOTA) performance on various NLP tasks. Multilingual Transformer like XLM-RoBERTa (Conneau et al., 2020a) and BERT multilingual base (mBERT) (Devlin et al., 2019a) have undergone extensive training with vast amounts of data drawn from various languages and datasets. Their transfer learning capabilities allow them to identify subtle nuances and patterns common across different languages, leading to superior performance compared to previous approaches. However, monolingual Transformers trained on a large amount of data from a single language can sometimes outperform multilingual models in different NLU tasks, including Named Entity Recognition (NER). This has been observed in languages like Bangla (Bhattacharjee et al., 2022), Portuguese (Souza et al., 2020, 2019), and others.

Wang et al. (2021)'s Google search methodology for information retrieval has been expanded to include the use of pre-existing knowledge sources, such as Wikipedia or DBpedia, which are widely used as general-purpose external knowledge resources across a wide range of natural language processing (NLP) tasks (Chen et al., 2017; Verlinden et al., 2021). It has also been widely used in Named entity recognition (NER) tasks since they contain huge amounts of entities like their types, attributes, and relationships shown by DAMO-NLP (Wang et al., 2022). Also, incorporating domain-specific knowledge such as dictionary features can improve NER performance (Cheng et al., 2021). Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) (Donnelly et al., 2006), Unified Medical Language System (UMLS) have already shown significant enhancement in the performance of NER as demonstrated by UmlsBERT (Michalopoulos et al., 2021).

Ensemble learning, which combines the strengths of multiple models, is a straightforward approach to address nonlinear relationships and minimize data noise. As a result, it demonstrates significant performance in Named Entity Recognition (NER) (Speck and Ngonga Ngomo, 2014), including prior MultiCoNER tasks (Malmasi et al., 2022c). In the recent MultiCoNER task, several ensembling methods were employed, including merging pre-trained language encoders with data augmentation techniques based on translations of the original training data (Pu et al., 2022), multilingual models (such as XLM-RoBERTa-large and Microsoft/infoxlm-large) with a Conditional Random Field (CRF) layer (Hassan et al., 2022), multilingual and monolingual Transformer models (Song and Bethard, 2022; He et al., 2022), and Transformer combined with Reinforcement Learning (Lin et al., 2022). These methods achieved significant performance improvements, demonstrating the effectiveness of ensembling in NER tasks.

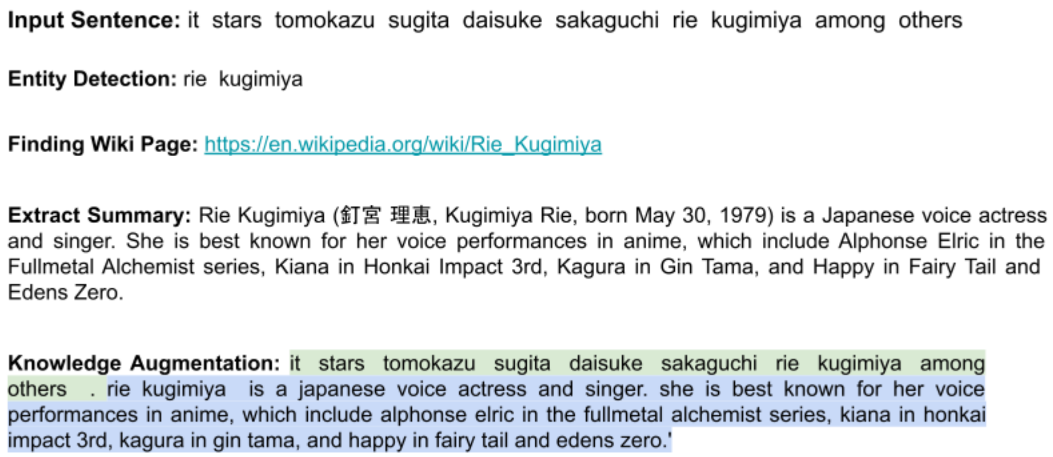
Our study expands on recent developments in external knowledge integration and ensemble learning techniques. While knowledge augmentation has been shown to be highly effective in Named Entity Recognition (NER) tasks, the resulting systems tend to be complex and computationally intensive. Our work introduces a streamlined approach to knowledge augmentation that is combined with ensemble techniques to improve NER performance.

3 Dataset Description

The MultiCoNER dataset has complex NER data with more syntactically challenging scenarios (Malmasi et al., 2022a) for 12 languages: English (EN), Spanish (ES), Hindi (HI), Bangla (BN), Chinese (ZH), Swedish (SV), Farsi (FA), French (FR), Italian (IT), Portuguese (PT), Ukrainian (UK) and German (DE). It also includes a multilingual set, which combines the same monolingual examples from each of the 12 languages (Fetahu et al., 2023a).

This dataset involves identifying and classifying various types of named entities in text, such as location, creative work, group, person, product, medical, etc. It is a fine-grained dataset, which means it contains more specific and detailed labels than coarse-grained alternatives. The dataset employs a fine-grained named entity taxonomy, with 36 fine-grained types and 6 coarse-grained types. Based on various contexts and factors, it can distinguish between different types of named entities. For example, it can distinguish between "scientist" and "athlete", which can be difficult for coarse-grained NER systems.

The MultiCoNER dataset is made up of three



Input Sentence: it stars tomokazu sugita daisuke sakaguchi rie kugimiya among others

Entity Detection: rie kugimiya

Finding Wiki Page: https://en.wikipedia.org/wiki/Rie_Kugimiya

Extract Summary: Rie Kugimiya (釘宮 理恵, Kugimiya Rie, born May 30, 1979) is a Japanese voice actress and singer. She is best known for her voice performances in anime, which include Alphonse Elric in the Fullmetal Alchemist series, Kiana in Honkai Impact 3rd, Kagura in Gin Tama, and Happy in Fairy Tail and Edens Zero.

Knowledge Augmentation: it stars tomokazu sugita daisuke sakaguchi rie kugimiya among others . rie kugimiya is a japanese voice actress and singer. she is best known for her voice performances in anime, which include alphonse elric in the fullmetal alchemist series, kiana in honkai impact 3rd, kagura in gin tama, and happy in fairy tail and edens zero.'

Figure 1: Simplified knowledge augmentation pipeline example. Green highlights the original input and blue is the augmented sentence.

main sources: Wikipedia Sentences (LOWNER), Questions (MSQ-NER), and Search Queries (ORCAS-NER). The train set has roughly 16k instances, whereas the dev set has approximately 800 instances. While each of the DE, HI, ZH, and BN test sets has around 20k instances, each of the remaining languages' test sets has around 250k instances. EN, DE, IT, FR, ES, ZH, PT, and SV have up to 30% of instances with additional noise. In this dataset, the noise consists of typographical errors caused by different keyboard layouts. The final system has been evaluated based on the performance of both noisy and clean dataset.

4 System Description

4.1 Knowledge augmentation

Our simplified knowledge augmentation pipeline consists of the following steps:

1. **Entity Detection:** To identify entities in the input text, we employed two methods. The first method involved using gold labels, while the second method involved utilizing pre-existing entity detector models. For languages other than English, we used a NER system that was developed by fine-tuning XLM-RoBERta on a multilingual dataset[1]. As for English, we relied on the Spacy model[2].

   To extract an entity, we concatenated BI tokens, which is a labeling scheme used to identify entities. When tokens A and B are labeled with the tags B-X and I-X, respectively, they are concatenated as a single identified entity. The augmented dataset generated using gold labels is referred to as ***Gold Entity***, whereas the dataset created using an existing entity detector is referred to as ***Entity Extractor***. The original dataset is named ***No Augmentation*** in the rest of the sections.

2. **Finding Wikipedia Pages:** We searched for a Wikipedia page that precisely matched each detected entity. We only searched within Wikipedia's corresponding language domains, such as `en.wikipedia.com` for English and `zh.wikipedia.com` for Chinese.

3. **Extract Summary:** If a Wikipedia page was discovered, the summary of that page was extracted from the Wikipedia API and cleaned by removing information such as birth dates.

4. **Knowledge Augmentation:** In the final step, the summary of each entity was added to the input sentence. Note that multiple Wikipedia page summaries may be added to the input sentence if multiple entities are detected in the entity detection step.

An example of our knowledge augmentation pipeline is described in Figure 1.

4.2 Fine-tuning pre-trained models

Multilingual We use two Transformer-based masked language models, XLM-RoBERTa (XLM-R) and mBERT, both of which are pre-trained on a

[1] Multilingual Entity Extractor Model

[2] Spacy Model

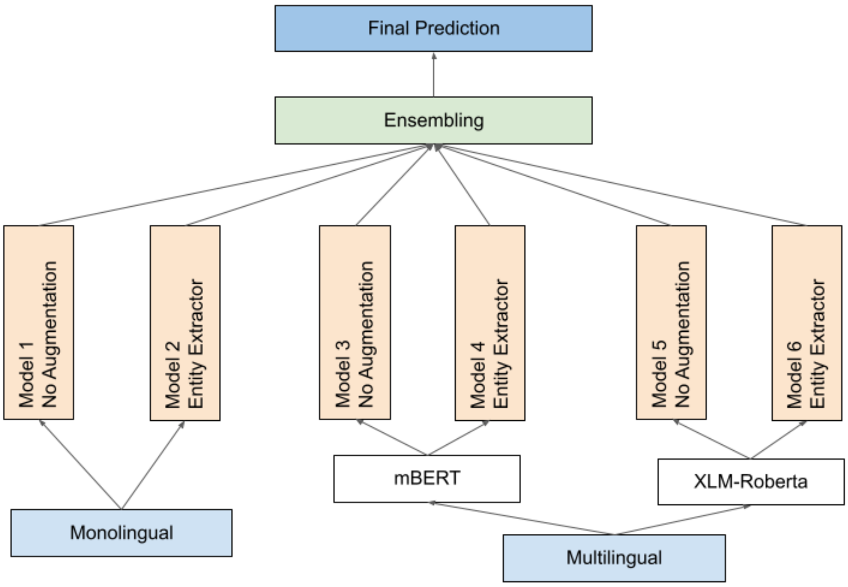


Figure 2: The pipeline of the overall system which makes use of monolingual and multilingual models for an ensemble method. Note that since there are no publicly available monolingual Ukrainian models, we are not using one for the Ukrainian track.

large amount of multilingual textual data (Conneau et al., 2020a; Devlin et al., 2019a). We specifically choose the XLM-R large variant, which is pre-trained with the RoBERTa architecture on 2.5 terabytes of data containing 100 languages. The mBERT base-cased model has been trained on 104 different languages. These multilingual language models have been found to perform well in a variety of tasks, including NER using multilingual contextual representations, particularly in low-resource scenarios (Conneau et al., 2020a). These two large language models have been fine-tuned for each of the 12 single-language NER tasks as well as the multilingual NER task.

Monolingual We employ several monolingual pre-trained Transformer-based large language models for all the languages except for Ukrainian. Appendix A contains a summary of the monolingual models utilized in this study. Although previous studies have confirmed state-of-the-art results using transfer learning across multiple languages (Conneau et al., 2020b), there is an issue with catastrophic forgetting in pre-trained language models (PLMs) that limits the effectiveness of transfer learning by updating most of the weights of the model during full fine-tuning (Pfeiffer et al., 2021; Ramasesh et al., 2022; Thomas et al., 2022). Thus, the benefit of multilingual representations may not be fully leveraged for the downstream task. Furthermore, the majority of the languages studied in this work are high-resource languages for which large-scale monolingual pre-trained models already exist. By taking these into consideration, we conduct our finetuning experiments with monolingual language models as well.

4.3 Ensembling

In the final step of our pipeline, we implement an ensemble method using three different ensembling techniques: majority voting, rank, and weighted. In majority voting, for each token we find which label has the most votes among the predictions across our models. Since we have an even number of models, two or more labels may receive the same number of votes. In the event of a tie, we choose the first label as the output. The rank and weighted ensembling methods are alternative approaches that are expected to alleviate situations where there are ties. In rank ensembling, we first ranked the models based on their f1 score, and the ones with the highest ranks will be allocated more votes in their prediction. For example, if a model m with rank 7 predicted the label y, instead of adding 1 vote to y, the model's rank is added. This means that 7 votes would be added to y, with a higher rank value indicating a better model in this scenario. Finally, the output is determined by whichever prediction received the most votes. We used the same strategy in weighted ensembling, but instead of weighting the labels with rank, we used the model's f1 score to calculate the number of votes.

Figure 2 shows our system with the six fine-tuned models we used per ensembling method.

5 Experimental Setup

We present our approach to enhancing a Named Entity Recognition (NER) model using fine-tuned multilingual and monolingual encoders based Transformer models for a token classification task. To facilitate our NER model pipeline, we first tokenize each sentence and convert it into a fixed-length vector of 256 dimensions, which is then fed into the Transformer model. The final classification layer generates a probability matrix of $256 \times n$ to predict the class label for each token where n is the number of classes. We only consider the output from actual tokens to determine the loss, disregarding those generated from augmented sentences as we lack gold labels for these extra tokens. During fine-tuning, we utilize a learning rate of $1e-4$ and train the model for 7 epochs.

Furthermore, we fine-tune three models for each language and Transformer model, using three different datasets (e.g. No Augmentation, Entity Extractor, and Gold Entity). While the fine-tuned models with the *No Augmentation* and *Entity Extractor* datasets are integrated into our final system, we use the other fine-tuned models with the *Gold Entity* dataset for error analysis. During our experiments, we use the MultiCoNER dev set to evaluate our models and split the train set into 80-20 split to create the training and validation sets for fine-tuning.

6 Results

For our results, we compute macro F1-scores based on token entity predictions.

Comparison with other systems In Table 1, we present the results of our model's performance in each track, along with the macro F1 score for both the clean and noisy categories. Notably, our system ranks in the top 10 out of 10 language tracks.

We can certainly observe that the addition of noise has an impact on the model's performance, but it is encouraging to see that the drop in languages like FR, PT, and IT is not below 60. However, the effects of noise on the ZH track are notable as they are much closer to the ones from top participants like DAMO-NLP and USTC-NELSLIP [3]. Moreover, our model shows greater robustness in some tracks (SV, ZH, FR) compared with participants who outperform us with no noise [3]. Despite this, the EN track is significantly affected by the presence of noise.

Language	Clean	Noisy	Macro	Position
English (EN)	65.25	56.96	62.73	16/34
Spanish (ES)	66.19	58.43	63.73	9/18
Swedish (SV)	70.40	62.19	67.63	9/16
Ukrainian (UK)	65.64	-	65.64	9/14
Portugese (PT)	66.81	60.04	64.51	10/17
French (FR)	68.09	60.22	65.68	10/17
Farsi (FA)	62.12	-	62.12	9/14
German (DE)	63.88	-	63.88	12/17
Chinese (ZH)	67.50	50.17	63.47	10/22
Hindi (HI)	71.23	-	71.23	9/17
Bangla (BN)	73.39	-	73.39	8/18
Italian (IT)	70.16	63.99	68.20	9/15
Multilingual	69.16	-	69.16	12/18

Table 1: F1-scores (%) for the clean and noisy categories as well as the overall macro F1-scores for each track calculated on the test sets. Additionally, our contest position for each entry is displayed.

Overall Table 2 compares the results of our three ensembling models in all of the 13 language tracks where we can observe that in the dev set the majority voting method only outperforms the other ensembling methods in the SV track. On the other hand, weighted average yields the best results in the majority of tracks and rank based performs the best in 5 tracks. Interestingly, both of these methods are the ones we implement to alleviate ties between labels.

Notably, BN reaches impressive results in the validation set with both of the aforementioned ensembling methods, but results remain much closer between the 3 methods on the test set where there is a much more noticeable drop between validation and test sets. This may be the case due to the larger test dataset that is also assessing the level of generalization in our models. Despite the drop, the methods yield good performance and in some tracks, like SV, the gap is not extremely wide (-2.13) between the dev and test sets.

In table 3 we compare the results on the validation set from our monolingual and multilingual models with 3 different knowledge augmentation methods. Note that results for UK are limited to the

[3] SemEval2023 Results

Set	Model	Language												
		EN	ES	SV	UK	PT	FR	FA	DE	ZH	HI	BN	IT	Multilingual
Dev	Majority Voting	68.82	69.73	**69.76**	69.75	69.76	71.19	68.44	74.56	77.32	79.25	86.96	73.59	75.06
	Rank Based	**70.50**	**71.15**	67.11	71.38	**71.60**	**73.02**	**68.67**	74.91	77.32	78.96	90.60	72.20	76.43
	Weighted	68.81	71.13	69.03	**72.13**	70.77	71.21	68.35	**74.93**	**77.89**	**79.27**	**90.90**	**73.95**	77.39
	Best Individual Model	67.47	70.03	66.44	69.40	68.87	68.72	65.57	72.29	75.84	74.66	86.96	70.22	**77.60**
Test	Majority Voting	62.52	62.02	67.37	64.16	64.38	**65.68**	62.09	**63.88**	63.01	70.58	73.34	67.52	65.72
	Rank Based	61.74	63.27	66.11	65.18	63.98	64.94	**62.12**	62.84	61.96	69.37	73.19	67.66	67.50
	Weighted	**62.73**	**63.73**	**67.63**	**65.64**	**64.51**	65.56	61.95	63.83	**63.47**	**71.23**	**73.39**	**68.20**	68.95
	Best Individual Model	58.13	62.24	62.18	63.40	60.87	59.50	58.75	58.07	56.90	62.20	68.92	65.26	**69.16**

Table 2: The table displays macro F1-scores (%) from different ensembling methods for 12 monolingual tracks and a multilingual track. The methods are Majority Voting, Rank Based System, Weighted, and top-performing model scores on both dev and test sets.

		EN	ES	SV	UK	PT	FR	FA	DE	ZH	HI	BN	IT	Multilingual
mBERT	No Augmentation	61.66	67.76	62.75	65.38	63.54	63.95	60.57	**72.29**	73.74	74.58	77.09	68.26	70.22
	Gold Entity	72.14	**74.98**	**72.25**	68.76	69.93	72.10	76.01	74.89	77.00	79.39	83.13	74.23	76.24
	Entity Extractor	61.62	64.87	66.16	66.66	62.85	64.76	60.61	71.14	**75.84**	74.44	79.10	66.51	70.85
XLM-RoBERTa	No Augmentation	61.17	**70.03**	65.05	68.32	68.85	68.14	62.62	65.90	65.83	**74.66**	84.74	**70.22**	68.68
	Gold Entity	72.18	69.88	71.63	**72.34**	**75.60**	**76.64**	76.29	**75.76**	74.40	**84.65**	85.72	**75.25**	**81.84**
	Entity Extractor	65.30	66.35	62.20	**69.40**	**68.87**	**68.72**	64.07	70.72	70.89	72.72	80.39	68.60	**77.60**
Monolingual	No Augmentation	66.06	51.36	66.18	-	65.47	66.14	**65.57**	68.80	72.59	56.75	86.29	67.80	-
	Gold Entity	**73.98**	64.42	70.20	-	73.3	73.77	**79.84**	70.01	**80.08**	66.37	**88.41**	74.46	-
	Entity Extractor	**67.47**	55.15	**66.44**	-	57.81	65.66	65.41	70.72	74.45	65.03	**86.96**	68.40	-

Table 3: The table shows the macro F1-score (%) of the dev set in each language for the monolingual and multilingual models. The scores in **green** represent the best performing model among the monolingual and multilingual models, ignoring the gold entity results. The scores in **orange** represent the overall best performing model for each language.

multilingual models. As expected, the models with the gold entity knowledge perform the best. For the monolingual models, we can see that they benefited from the entity extractor augmentation, even in monolingual tracks like FA where no augmentation yielded the best result, the difference between no augmentation and entity extraction augmentation is marginal.

7 Discussion

Monolingual vs Multilingual Our multilingual models outperform the monolingual ones barring 4 tracks (EN, SV, FA, BN) where the multilingual F1 scores are lower than their monolingual counterparts (Table 3). Interestingly, we can see a similar trend when we compare the performance of the models trained on the gold entity dataset. However, under this setup, the best performing monolingual models are from the EN, FA, ZH, and BN tracks. Furthermore, we can observe that the majority of models that performed the best are the XLM-RoBERTa multilingual ones, since they had the best scores in 7 out of 13 tracks.

However, it is key to note that the difference in the F1 scores between the multilingual and monolingual models is not large except for ES and HI. Although such a large margin in these 2 tracks may seem surprising given that they were trained on a large monolingual dataset, it can be explained by the number of parameters of the ES model (Cañete et al. (2022)'s ALBETO tiny) and the multilingual training on Indian languages of the HI model (Khanuja et al. (2021)'s MuRIL). Our chosen model for the ES track has fewer parameters (5M) than some its counterparts and thus underperforms in NER tasks (Cañete et al., 2022). On the other hand, the Hindi model is only monolingual in regards to the script, as it is in fact a multilingual model trained on multiple languages with a shared script (Bangla, Nepali, Urdu, Hindi). While Khanuja et al. (2021) claims that the chosen Hindi model can take more cues from neighboring words than a multilingual model like mBERT, there are no quantitative results of the model's performance in NER tasks. These results showcase that model selection is crucial to properly assess the performance of multilingual and monolingual models in a given task.

Impact of Knowledge The number of irrelevant entities is larger than the detected relevant entities across all languages, and sometimes close to double the number of Detected Entities (eg. EN)[4]. Given the discrepancy between the Gold Entity and Entity

[4] Table 5 and Figure 3 show the statistics of entity relevance as detected by off-the-shelf entity taggers. Both of these can be found in Appendix B.

Gold tags: **B-Politician** I-Politician I-Politician O O O O O

No Augmentation Prediction: **B-OtherPER** I-OtherPER I-OtherPER O O O O O

joseph le bon guillotined for abuse of power

Entity Extractor Prediction: **B-Politician** I-Politician I-Politician O O O O O

joseph le bon guillotined for abuse of power joseph le bon was a french politician.

(a) Example: The added context has helped the Entity Extractor model to identify that *joseph le bon* is a `POLITICIAN` entity.

Gold tags: **B-VisualWork** I-VisualWork I-VisualWork I-VisualWork O O O O

No Augmentation Prediction: **B-VisualWork** I-VisualWork I-VisualWork I-VisualWork O O O O

currito fo the cross a silent film adaptation

Entity Extractor Prediction: **B-Food** O O I-Food O B-VisualWork I-VisualWork O

currito fo the cross a silent film adaptation boloco is the brand name of an american chain of restaurants founded in 1996.

(b) Example: The added context changed a previously correct output `B-VisualWork` for *currito* to an incorrect one `B-Food`.

Gold tags: O O O O **B-PublicCorp I-PublicCorp** O O O O O O

No Augmentation Prediction: O O O O **B-SportsGRP I-SportsGRP** O O O O O O

translated input: *Both are past employees of McKee's.*

两 者 都 是 麥 記 的 过 去 员 工 .
两 者 都 是 麥 記 的 过 去 员 工 .

Entity Extractor Prediction: O O O O **B-PublicCorp I-PublicCorp** O O O O O O

translated input: *Both are past employees of McKee's. McDonald's (English: mcdonald's) is a multinational fast food chain restaurant originating from Southern California, USA. It is also the largest fast-food chain in the world, mainly selling hamburgers, fries, fried chicken, soft drinks, ice cream, salad, fruits and coffee and other fast food items, currently headquartered in Chicago, USA.*

两 者 都 是 麥 記 的 过 去 员 工 .麥當勞（英語：mcdonald's）是源自美國南加州的跨國連鎖快餐店，也是世界最大的速食連鎖店，主要販售漢堡包及薯條、炸雞、汽水、冰品、沙拉、水果、咖啡等快餐食品，目前總部位於美國芝加哥。

两 者 都 是 麥 記 的 过 去 员 工 .麥當勞（英語：mcdonald's）是源自美國南加州的跨國連鎖快餐店，也是世界最大的速食連鎖店，主要販售漢堡包及薯條、炸雞、汽水、冰品、沙拉、水果、咖啡等快餐食品，目前總部位於美國芝加哥。

(c) Example from Chinese where context has helped to identify a slang usage "麥記/McKee's" as a `PublicCorp` correctly.

Table 4: Contributions of neighbouring tokens on the final classification of the target token from XLM-RoBERTa-large models. The red square indicates the target token. The stronger highlighted tokens indicate the stronger influence of that token on the classification output of the target token. Tokens highlighted in dark blue indicate contradictory evidence to the final prediction.

Extractor models (Table 3), we hypothesise that the relevance of extracted entities has a significant effect on the performance of the Entity Extractor model.

In order to better understand how augmented input affects NER classification, we conduct an analysis of outputs with LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), which provides explanations for the final prediction based on feature importance across the input, in our case the neighbouring tokens in an input sentence. We present example LIME outputs for English and Chinese from the XLM-RoBERTa-large models in Table 4.

Example 4a provides support for our hypothesis that useful contexts can be directly and prominently influential to the final prediction of a complex named entity class. In the original sentence, a neighbour like *guillotined* provides enough information that *joseph* is a `Person` but not enough to distinguish whether *joseph* is a specific kind of person. The augmented context from the Entity Extractor model provides relevant contexts with *politician*, which has shown to be the most influential neighbour to the final prediction of `B-Politician`.

On the other hand, 4b shows that irrelevant or incorrect context can negatively affect model predictions. The non-augmented input has given adequate context for the tag of *currito* as `B-VisualWork`, while the augmented input from the Entity Extractor model is wrongly associated with an incorrect context (*currito* is identified as an alternative name of chain restaurant *Boloco*), causing it to be tagged as `B-Food`. This prediction is mostly influenced by context neighbour *restaurants* despite contradictory predictions from the original sentence such as *silent film* (in dark blue).

In the case of Chinese example 4c, one token in English *"McKee's"* is split into two tokens in Chinese, corresponding to the syllables *mc-kees*. The slang term "麥/*mc* 記/*kee's*" (translated here as *McKee's*) was not recognised as an alternative name for McDonald's in the non-augmented sentence and was tagged incorrectly as `SportsGRP`. In the augmented sentence, this was corrected to `PublicCorp`. The first token '麥/*mc* is influenced by the whole phrase "是源自美國南加州的跨國連鎖快餐店/*is a multinational fast food chain restaurant originating from Southern California, USA*)", while the second token 記/*kee's* is influenced by the original name itself "麥當勞/*mcdonald's*" as well as the English translation of the name *mcdonald's*.

Overall, the above examples have shown that the relevance of the augmented context is a significant factor irrespective of the language. It can correct or improve the original tags (4b and 4c), but it can also override the effects of the original neighbours (4a) leading to detrimental effects.

8 Conclusion and Future Work

In this paper we present our garNER system that shows promising results for models trained with a gold entity dataset, which suggests that relevant context can boost performance in complex NER tasks. Therefore, we consider that training a custom entity extractor model for our knowledge augmentation step can improve the current version of our system. In order to do so, a simpler dataset could be created by removing the fine-grained categories and transforming the tags from B-X to B and I-X to I. Since the current model cannot detect the majority of tags, this is crucial and may improve overall performance. Lastly, we consider that adding a conditional random field (CRF), as it is common practice in token classification tasks, may be relevant because it avoids illegal span predictions. CRF layers could potentially be added on these transformer models to further evaluate NER performance and see if there are any improvements.

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.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Ming Cheng, Shufeng Xiong, Fei Li, Pan Liang, and Jianbo Gao. 2021. Multi-task learning for chinese clinical named entity recognition with external knowledge. *BMC medical informatics and decision making*, 21(1):1–11.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining text encoders as discriminators rather than generators. In *ICLR*.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Fadi Hassan, Wondimagegnhue Tufa, Guillem Collell, Piek Vossen, Lisa Beinborn, Adrian Flanagan, and Kuan Eeik Tan. 2022. Seql at semeval-2022 task 11: An ensemble of transformer based models for complex named entity recognition task. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1583–1592.

JiangLong He, Akshay Uppal, N Mamatha, Shiv Vignesh, Deepak Kumar, and Aditya Kumar Sarda. 2022. Infrrd. ai at semeval-2022 task 11: A system for named entity recognition using data augmentation, transformer-based sequence labeling model, and ensemblecrf. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1501–1510.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Qizhi Lin, Changyu Hou, Xiaopeng Wang, Jun Wang, Yixuan Qiao, Peng Jiang, Xiandi Jiang, Benqi Wang, and Qifeng Xiao. 2022. Pa ph&tech at semeval-2022 task 11: Ner task with ensemble embedding from reinforcement learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1444–1447.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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. 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.

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.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Farahani Mehrdad, Gharachorloo Mohammad, Farahani Marzieh, and Manthouri Mohammad. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. 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. *CoRR*, abs/1802.05365.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Keyu Pu, Hongyi Liu, Yixiao Yang, Jiangzhou Ji, Wenyi Lv, and Yaohan He. 2022. Cmb ai lab at semeval-2022 task 11: A two-stage approach for complex named entity recognition via span boundary detection and span classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1603–1607.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Stefan Schweter. 2020. Italian bert and electra models.

Hyunju Song and Steven Bethard. 2022. Ua-ko at semeval-2022 task 11: Data augmentation and ensembles for korean named entity recognition. Association for Computational Linguistics (ACL).

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Ensemble learning for named entity recognition. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*, pages 519–534. Springer.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Bethan Thomas, Samuel Kessler, and Salah Karout. 2022. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7102–7106. IEEE.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1952–1957, Online. Association for Computational Linguistics.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Model Summary

mBERT: A multilingual model called mBERT (Devlin et al., 2019b) is designed to jointly condition on both left and right context in all layers in order to pretrain deep bidirectional representations from unlabeled text. It was trained on a corpus of 104 languages which makes it useful for cross-lingual applications. Masked language modeling (MLM) and next sentence prediction (NSP) objectives are used to pre-train it.

XLM-RoBERTa: A multilingual version of RoBERTa is called XLM-RoBERTa (Conneau et al., 2020a). 2.5TB of filtered CommonCrawl data containing 100 languages is used as its pretraining material. The model is trained to anticipate the masked tokens in the input using samples of text streams from each language. This model is able to deal with code-switching better because language embeddings are not utilized, in contrast to (Lample and Conneau, 2019).

RoBERTa-large: Using a self-supervised learning method, the transformers-based model RoBERTa (Liu et al., 2019) was pre-trained using a vast collection of English language data. It is both a replication of BERT (Devlin et al., 2019b) and a better method for training BERT models, which takes into account the implications of training set size and hyperparameter adjustment. It is different from BERT (Devlin et al., 2019b) on the sense that this model is trained over a longer period of time with larger batches and more data, the next sentence prediction objective is removed here, it is trained on longer sequence, and also dynamically changing the masking pattern is applied to the training data.

BanglaBERT: The BERT architecture-based Natural Language Understanding (NLU) model known as BanglaBERT (Bhattacharjee et al., 2022) was pre-trained particularly for the Bangla language. Crawling 110 well-known Bangla websites yielded 27.5 GB of Bangla pretraining data (named ”Bangla2B+”) that was used to pretrain BanglaBERT. Using the Replaced Token Detection (RTD) objective, which involves jointly training a generator and a discriminator model, BanglaBERT was pretrained using ELECTRA (Clark et al., 2020).

BERTimbau: A BERT model specifically created for Brazilian Portuguese is named as the BERTimbau (Souza et al., 2020). Three natural language processing tasks—named entity recognition, sentence textual similarity, and textual entailment recognition have been completed by this model with state-of-the-art performance.

CamemBERT: The RoBERTa (Liu et al., 2019) architecture is the foundation of the cutting-edge language model for French called CamemBERT (Martin et al., 2020). CamemBERT uses SentencePiece tokenization (Kudo and Richardson, 2018) and was trained using the French-language subset of the OSCAR corpus (Suárez et al., 2019). To train the model, Masked Language Modeling (MLM) is used.

ParsBERT: The BERT architecture (Devlin et al., 2019b) from Google is used by the monolingual language model ParsBERT (Mehrdad et al., 2021), which has the same configurations as BERT-Base. This model has been pre-trained on a large Persian corpus with more than 2 million documents and a variety of writing styles from a wide range of themes (such as scientific, fiction, and news). This corpus's substantial portion was manually crawled. Masked Language Modeling (MLM) and Next Sentence Prediction(NSP) are the two objectives that the model has been trained on.

MuRIL: The MuRIL (Khanuja et al., 2021) model is based on the BERT (Devlin et al., 2019b) architecture and was trained from scratch using various corpora, including Wikipedia[5] , Common Crawl OSCAR, PMINDIA (Haddow and Kirefu, 2020), and Dakshina (Roark et al., 2020) for 17 Indian languages and their transliterated versions. Masked Language Modeling (MLM) and Translation Language Modeling (TLM) are two language modeling objectives that were used to train the model.

Italian BERT: The OPUS corpora[6] collection of texts plus a recent Wikipedia dump were used to train the Italian BERT (Schweter, 2020) model, producing a final training corpus of 13 GB and 2,050,057,573 tokens. The same training data from OPUS plus additional data from the Italian section of the OSCAR corpus[7] were used for the XXL Italian models, resulting in a final training corpus size of 81GB and 13,138,379,147 tokens.

Swedish BERT: The architecture of BERT (Devlin et al., 2019b) serves as the foundation for the Swedish BERT Model (Malmsten et al., 2020). About 15 to 20 GB of text (200 million sentences and 3 billion tokens) from diverse sources are used to train the model (books, news, government publications, Swedish Wikipedia, and internet forums).

German BERT: The BERT architecture (Devlin et al., 2019b), which was first suggested by Google, was used to create the German BERT language model (Chan et al., 2020). Data was gathered from a variety of sources, including OSCAR (Suárez et al., 2019), German-language Wikipedia dumps, OPUS, Open Legal Data[8], and news items to train this model.

ALBETO: ALBETO (Cañete et al., 2022) is a pre-trained ALBERT model (Lan et al., 2020) variation that has only been trained on Spanish language corpora. It employs the weight-tied approach, which shares all parameters among the model's layers. There are five sizes available for the ALBETO models: tiny, base, large, xlarge, and xxlarge. SentencePiece (Kudo and Richardson, 2018) was used to build a vocabulary of 31K lowercase tokens that was shared by all of these models and used to the training dataset. To obtain the best outcomes, all ALBETO models were trained using the LAMB optimizer (You et al., 2020) in accordance with the authors' recommendations.

Chinese BERT: The objective of this model (Turc et al., 2019) is to achieve gains within a constrained memory and latency budget. A compact student model is trained to imitate the predictions of a highly accurate but resource-intensive teacher model. Model compression is accomplished using the common knowledge distillation technique (Hinton et al., 2015). This approach comprises of three standard training operations: masked LM (MLM) pre-training (Devlin et al., 2019b), task-specific distillation, and optional fine-tuning. The models follow the BERT's input processing (Devlin et al., 2019b) and Transformer design (Vaswani et al., 2017). By fine-tuning a pre-trained BERT$_{\text{LARGE}}$ model on the labeled dataset, a teacher model and 24 students of varying sizes were trained for each end task.

[5] https://www.tensorflow.org/datasets/catalog/wikipedia

[6] https://opus.nlpl.eu/

[7] https://oscar-project.org/

[8] http://openlegaldata.io/research/2019/02/19/court-decision-dataset.html

B Number of entities

	Gold entities			Detected Entities			Irrelevant Entities		
	train	dev	test	train	dev	test	train	dev	test
EN	25448	1295	377802	5266	268	69475	9225	460	142232
BN	13222	674	23731	4745	244	10663	6159	328	13068
ZH	15226	772	26796	6502	344	12790	6612	318	14006
FR	26375	1351	398193	4419	230	64927	7291	413	109559
FA	23656	1235	312097	12250	645	169547	17564	998	245043
ES	23904	1229	356373	3617	166	50892	5170	280	77034
DE	15950	840	28877	2971	166	6221	5381	271	11142
HI	12869	683	23199	4575	235	9675	4882	232	10026
IT	26436	1395	397221	5543	306	82950	8152	460	124370
PT	24438	1290	340751	3898	194	55544	5406	246	72995
SV	25413	1390	361156	4733	254	65893	7530	418	107050
UK	21955	1134	315374	4833	231	70292	6880	376	105183

Table 5: Total number of entities in the dataset (gold), number of entities detected by the entity extractor (detected), and the number of entities detected by entity extractor that are not present in the dataset (irrelevant).

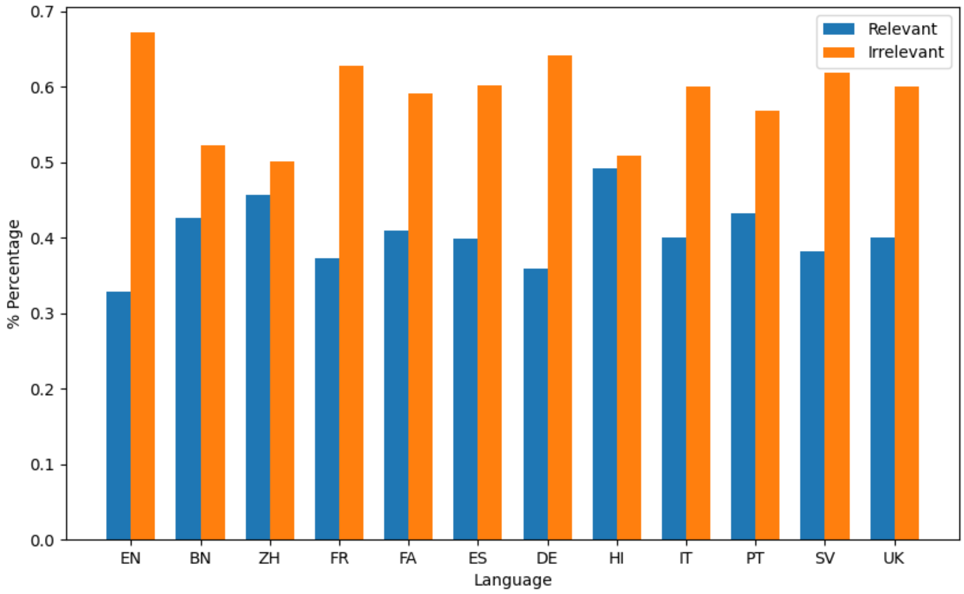


Figure 3: Number of irrelevant entities detected in the test set is greater than the relevant ones across all languages. Interestingly, more irrelevant entities were detected in European languages (EN, FR, ES, DE, IT, PT, SV, and UK) and Persian (FA).