IXA at SemEval-2023 Task 2: Baseline Xlm-Roberta-base Approach

Edgar Andrés Santamaría
University of the Basque Country (UPV/EHU)
HiTZ Center, IXA Research group
Datu(a) IA
edgar.andres.santamaria@gmail.com

Abstract

IXA proposes a Sequence labeling fine-tune approach, which consists of a lightweight baseline (10e), the system takes advantage of transfer learning from pre-trained Named Entity Recognition and cross-lingual knowledge from the LM checkpoint. This technique obtains a drastic reduction in the effective training costs that works as a perfect baseline, future improvements in the baseline approach could fit: 1) Domain adequation, 2) Data augmentation, and 3) Intermediate task learning.

Practical Information

The system is composed of two main stages, first of all, fine-tuning which consists in learning IOB sequences via sequence classification task, and then the inference phase when unannotated examples are given to guess the correct IOB sequence.

We learn the correct way to label sequences using real examples from the training set and back-propagate the labeling errors over the development set compared to the intended gold standard. At this point, we are taking advantage of already pre-trained checkpoints over MLM and Sequence Classification tasks.

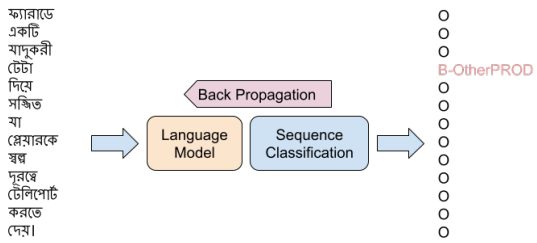

Figure 1: Trainning Schema.

Finally, we infer the proper IOB sequences for raw examples, those inferences are led by Language Model expertise on previous sequence classification plus the fine-tuning on MulticoNER this property is called transfer learning, additionally the Masked Language Modelling allows the unseen token representation into the already known word space via substrings in this case, this property is called cross linguality.

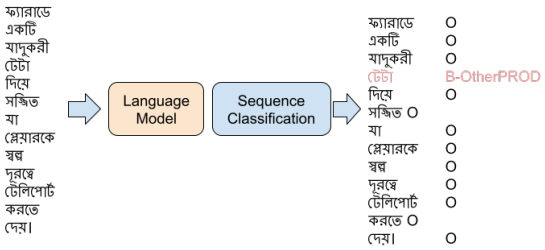

Figure 2: Inference Schema.

1 Introduction

The tag set of MultiCoNER is a fine-grained taxonomy 3 with 36 tags divided into 6 main groups. Further relevant information related to SemEval shared tasks is provided in guidelines (Fetahu et al., 2023).

In addition to multilingualism, there were two key challenges in this year's task: (1) a fine-grained entity taxonomy which mainly penalizes the precision of systems as the ability to guess the correct tag, and (2) simulated errors added to the test set to make the task more realistic and difficult which mainly penalizes the recall of systems as the ability to guess the correct chunk of text to tag.

2 Related Work

SemEval-2023 Task 2 follows the multilingual NER task started in 2022 (Malmasi et al., 2022b). critical challenges in the 2022 datasets were dealing with complex entities with limited context (Malmasi et al., 2022a). This is commonly understood as a low-resourced specific domain problem e.g "Medical Entity Recognition", main

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 379–381

issues come from the fact that only a few of the total examples are annotated and the provided contexts show specific lexical usually unknown for language models e.g "in particular he worked on the laser photocoagulation of threshold retinopathy of prematurity".

The challenges of Sequence Classification for recognizing complex entities in low-context conditions was recently outlined by Meng et al. (2021). Other work has extended the content to multilingual and code-mixed settings (Fetahu et al., 2021). In our experience, the multilingual constraints force the necessity of data to improve the learning results, here we identify three main paths: 1) Domain adequation, which consists in learning MLM task with the raw texts provided in the task, 2) data augmentation, which is generating synthetic data with a sustainable lexical difference according to real examples but maintaining the possibility to align the annotated IOB chains, and 3) Intermediate task learning, which feeds Language Model with more expertise via Silver Dataset learning (Similar tasks in other formats), this is usually pursued in more generic fine tune tasks such as Natural Language Inference.

3 Data

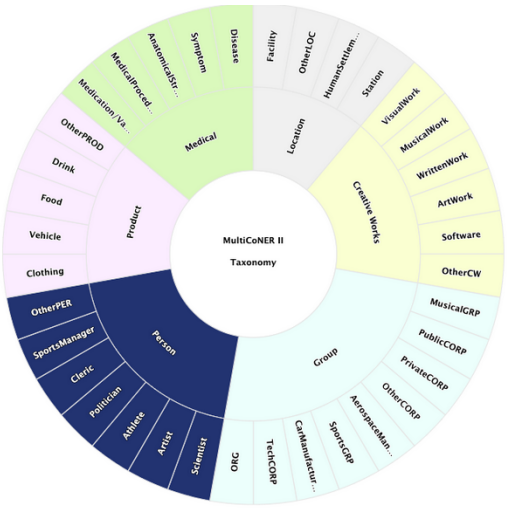

Figure 3: MultiCoNER fine-grained taxonomy.

As we can appreciate we are provided a fine-grained taxonomy with 36 labels divided into 6 main groups. The data collection methods used to compile the dataset used in MultiCoNER will be described in a paper to be published shortly (Malmasi et al., 2022a). We are provided a dataset divided into train, dev, and test sets, in this case, mainly unannotated corpora (test set) as shown in table 1.

Language	Training	Validataion	Test
Bangla	9,708	507	19,859
German	9,785	512	20,145
English	16,778	871	249,980
Spanish	16,453	854	246,900
Farsi	16,321	855	219,168
French	16,548	857	249,786
Hindi	9,632	514	18,399
Italian	16,579	858	247,881
Portuguese	16,469	854	229,490
Swedish	16,363	856	231,190
Ukrainian	16,429	851	238,296
Chinese	9,759	506	20,265

Table 1: Data Distribution.

Language	Time	F1	Precision	Recall
Chinese	15m	0.488	0.483	0.515
Ukranian	33m	0.631	0.622	0.655
Swedish	25m	0.622	0.622	0.637
Italian	30m	0.597	0.606	0.609
Portuguese	31m	0.598	0.606	0.604
Farsi	37m	0.529	0.550	0.545
Hindi	10m	0.622	0.653	0.632
French	30m	0.552	0.557	0.556
English	32m	0.526	0.538	0.531
Bangla	13m	0.639	0.649	0.647
Spanish	32m	0.597	0.635	0.584
German	13m	0.562	0.563	0.580

Table 2: Development Results per language.

4 Results

We pursued 12 experiments one per language, each xlm-roberta-base (Conneau et al., 2019) model was trained with 32 batch sizes for 10 epochs. The training process took on average around 25 minutes using NVIDIA Tesla V100 GPU. We took around 13h using 400W that is 5.20 KW which could be understood as 7,84 CO2e (lbs) generated by the experimentation.

We provide two result tables to describe the proposed system performance, development evaluation shown in table 2 was performed via seqeval framework (Nakayama, 2018), this evaluates the performance of chunking tasks. And test evaluation is shown in table 3. The ranking is computed using Macro averaged F1 scores. In the held-out test

Language	F1	Precision	Recall
Chinese	8.06	4.49	6.93
Ukranian	22.81	-	-
Swedish	27.96	21.72	25.96
Italian	20.05	14.82	18.41
Portuguese	18.4	13.91	16.97
Farsi	15.87	-	-
Hindi	26.13	-	-
French	18.9	13.89	17.4
English	16.88	11.8	15.39
Bangla	18.49	-	-
Spanish	17.65	12.16	16.01
German	16.09	-	-

Table 3: Official Test Results (Fine-grained).

set, we had a noisy subset for 8 languages where the sentences were corrupted with noise either on context tokens or entity tokens. The test sets were corrupted for EN, ZH, IT, ES, DE, FR, PT, SV. In this case fine-grained metrics were assessed.

5 Conclusion

These brief experiments are considered baseline approaches to the task. Those techniques applied in various domains/tasks show that we can effectively approach the baseline system as a sequence labeling fine-tune over the dataset. Additionally, future improvements in the baseline approach could fit: 1) Domain adequation, 2) Data augmentation, and 3) Intermediate task learning.

References

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. *arXiv preprint arXiv:1911.02116*.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. 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.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. Multiconer: A large-scale multilingual dataset for complex named entity recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.