Sakura at SemEval-2023 Task 2: Data Augmentation via Translation

Alberto Poncelas **Maksim Tkachenko** **Ohnmar Htun**
Rakuten Institute of Technology
Rakuten Group, Inc.
{first.last}@rakuten.com,

Abstract

We demonstrate a simple yet effective approach to augmenting training data for multilingual named entity recognition using machine translation. The named entity spans from the original sentences are transferred to the translations via word alignment and then filtered with the baseline recognizer to retain high quality annotations. The proposed data augmentation approach improves the baseline performance of XLM-Roberta on the multilingual dataset.

1 Introduction

Named Entity Recognition (NER) is an essential task in natural language processing (NLP) that involves identifying and classifying entities such as people, organizations, locations, and more recently, complex entities such as events, creative works, diseases. Multilingual Complex Named Entity Recognition (MultiCoNER) shared task (Malmasi et al., 2022a,b) is organized to promote research in multilingual NER for complex entities. MultiCoNER II (Fetahu et al., 2023b) features one multilingual track and twelve monolingual tracks for the following languages: English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, and German. The datasets contain fine-grained tagsets that mainly cover three domains (Wiki sentences, questions, and search queries) and include six top-level tags: Location (LOC), Creative Work (CW), Group (GRP), Person (PER), Product (PROD), and Medical (MED).

Complex named entities are highly ambiguous and can take the form of any linguistic constituency: "Sir" (*Creative-Work/Visual-Work*), "Catch Me If You Can" (*Creative-Work/Visual-Work*), "Jeremy Corbyn" (*Person/Politician*). Thereby, making a granular distinction between entity classes (e.g., *Politician* vs. *Scientist*) and identifying rare entities solely from the context are challenging problems. To alleviate this challenge, we propose to augment multilingual training data with translations, which we believe can help the NER model to disambiguate more entities by learning from diverse and noisy contexts. Table 1 shows a few examples of such translations with the aligned entity annotations.

The proposed augmentation strategy involves three steps: translating training example, transferring entity annotations from the source sentences to the target, and filtering out low-quality annotations. To produce translations, we use publicly-available mBART-50 (Tang et al., 2020). To transfer entity annotations, we compute the token alignment between the source and its translation using SimAlign (Jalili Sabet et al., 2020) and then build an entity alignments on top of the token alignments. To ensure high-quality annotations, we use XLM-RoBERTA (XLM-R) (Conneau et al., 2020) fine-tuned on the original training data to filter out low-quality annotations from the translated data. We find that XLM-R fine-tuned on the augmented dataset outperforms the same model fine-tuned on the original training dataset.

2 Related Work

NER is one of the most popular NLP task and is widely explored across domains: news (Tjong Kim Sang and De Meulder, 2003), biomedical (Settles, 2004), social media (Ritter et al., 2011; Ashwini and Choi, 2014). With the recent advances in representation learning, models such as XLM-R (Conneau et al., 2020) and BERT (Devlin et al., 2019) achieve the state-of-the-art NER performance with minimal feature-engineering efforts. In this work, we use XLM-R as our base model for fine-tuning on the NER task.

Data augmentation via translation has been widely used. An example of this is back-translation (Sennrich et al., 2016), a methodology of obtaining parallel data by translating sentences from the target language to the source language.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1718–1722

Sentence	Translation
він також володіє трьома гелікоптерами [eurocopter | AerospaceManufacturer] .	También tiene tres helicópteros [eurocopter | AerospaceManufacturer]
il a étudié au [conservatoire royal de la haye | MusicalGRP] auprès de [louis andriessen | Artist] .	er studierte am [royal haye conservatory | MusicalGRP] mit [louis andriessen | Artist] .

Table 1: Translation examples.

Although such data augmentation approaches may introduce noise, increasing the lexical diversity has been effective (Vanmassenhove et al., 2019; Poncelas et al., 2019; Burchell et al., 2022). Translation have been successfully applied to cross-lingual NER problem under unsupervised transfer settings with limited resources (Xie et al., 2018; Mayhew et al., 2017). Schäfer et al. (2022) use machine translation and word alignment techniques to transfer inferred named entity annotations from a rich resource language to a low resource language. We draw inspiration from these works and employ the same approach to augmenting the training data for NER, in addition we assume that we have enough data in the target language to train a filtering model.

3 Data Augmentation via Translation

We employ mBART-50[1] (Tang et al., 2020) to generate translations for every sentence in the original dataset (Fetahu et al., 2023a). mBART-50 is trained on 50 languages including the 12 languages of the shared task. We pivot translations via English, as we generally observe that pivoting produces better translations than the direct approach, e.g., Hindi $\rightarrow$ English $\rightarrow$ Chinese vs. Hindi $\rightarrow$ Chinese. The translations are generated from every source language to every target language except English, as our focus is on languages with less resources.

After generating the translations we align the tokens using SimAlign (Jalili Sabet et al., 2020). Through the token alignment we transfer the entity annotations from the source sentences to their translations. Although the word order varies in each language, name entities comprise mostly indivisible spans in different languages, which makes transfer generally feasible. To align the entity boundaries, we collect all the target tokens aligned to any of the tokens in the source entity and confirm the transfer only if the target tokens form a continuous span in the target sentence (see Figure 1).

Although mBART-50 supports all the task languages, the translation quality is not guaranteed especially for the low-resource languages, besides the alignment quality also contributes to the compounding error. Thus, to ensure the quality of the augmented dataset, we first retain the translations that only have matching number of entities with the sources. Then, we apply the baseline NER model fine-tunted on the original dataset to the translated data and retain the sentences that agree with baseline on the number of entities and their boundaries. Table 2 shows the data stats before and after augmentation and filtering. By adding the translations, we can increase the size of the data between 30K and 102K sentences depending on the language. Note that in the augmented dataset, the artificially-generated sentences are the majority.

4 Named Entity Recongition

We formulate Named Entity Recognition (NER) as a sequential labelling task, where the goal is to assign each token in a given text a label that corresponds to its entity type. For example, in the sentence "John Smith works at New York", the label "B-PER" would be assigned to "John", the label "I-PER" would be assigned to "Smith", the label "B-LOC" would be assigned to "New", and the label "I-LOC" would be assigned to "York".

The sequential labeling is a well-studied tasks and enables us to easily leverage standard model architectures. We use XLM-Roberta (Conneau et al., 2020) as our baseline model. As the premise of our exploration is the data augmentation strategy only, we make no modifications to the XLM-R architecture and use it as-is for fine-tuning on the data variants.

Table 3 shows the macro measures (precision, recall, and F1) computed on the multilingual test set (Fetahu et al., 2023a) for the model trained with the original data and after our augmentation. The data augmentation via transation improves the baseline by 0.85 F1. Table 4 shows fine-grained results of the F1 of each named entity class for the model trained on the original and augmented data. We see

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual#mbart50-models

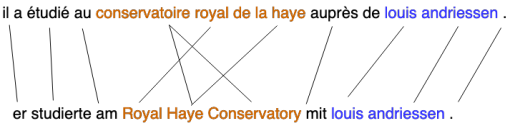

Figure 1: Example of word aligment

Dataset	Original	+ Translation
Bangla	10K	+ 30K
German	10K	+ 58K
Spanish	16K	+ 101K
Farsi	16K	+ 71K
French	17K	+ 102K
Hindi	10K	+ 52K
Italian	17K	+ 76K
Portugese	16K	+ 80K
Swedish	16K	+ 82K
Ukranian	16K	+ 72K
Chinese	9K	+ 39K

Table 2: MultiCoNER and augmented datasets, number of sentences. The + symbol indicates the number of sentences used for augmenting the original size. For example, the total number of sentences for *Bangla* language is 40K in total.

that the F1 of the augmented model is higher for most of the name entities. Only for four of entities, i.e., *HumanSettlement*, *Station*, *MusicalWork*, *Food* show marginal deterioration in the performance.

We fine-tune XLM-R on a machine with 8 x NVIDIA A100 80Gb with effective batch size of 1024, learning rate of $2 \cdot 10^{-5}$. Every model is trained for 250 epoch using the cross entropy loss. During the inference malformed tag sequences are removed. The output of the final model is submitted for evaluation.

Table 5 show the results of our proposed approach in each track.[2] This table includes not only the overall F1 scores but also a fine-grained evaluation on the clean and noisy subsets. In the column *Rank Status* we also indicate the position our system in each track.

Training Data	Precision	Recall	F1
Original	72.22	73.91	72.97
Augmented	**73.56**	**74.22**	**73.82**

Table 3: Macro measures on multilingual test set.

[2]Details of the final ranking for all the 13 tracks can be found in https://multiconer.github.io/results

Fine-Grained Class	Original (F1)	Augmented (F1)
Facility	73.20	**73.93**
OtherLOC	74.38	**75.84**
HumanSettlement	**89.15**	89.05
Station	**81.50**	81.35
VisualWork	82.84	**83.37**
MusicalWork	**77.93**	77.86
WrittenWork	74.11	**74.33**
ArtWork	58.08	**60.32**
Software	82.24	**82.40**
MusicalGRP	80.63	**81.33**
PublicCorp	77.39	**77.96**
PrivateCorp	68.09	**74.07**
AerospaceManufacturer	72.51	**73.31**
SportsGRP	87.53	**87.94**
CarManufacturer	77.59	**78.70**
ORG	74.69	**75.25**
Scientist	52.26	**54.31**
Artist	81.50	**81.88**
Athlete	81.20	**81.88**
Politician	67.06	**68.35**
Cleric	66.75	**68.31**
SportsManager	64.55	**65.96**
OtherPER	56.72	**57.25**
Clothing	64.03	**64.40**
Vehicle	65.97	**66.84**
Food	**66.16**	65.69
Drink	69.24	**70.11**
OtherPROD	67.50	**67.92**
Medication/Vaccine	79.51	**79.57**
MedicalProcedure	73.67	**74.08**
AnatomicalStructure	75.09	**75.74**
Symptom	66.99	**68.52**
Disease	78.12	**78.26**

Table 4: Breakdown comparison on multilingual test set.

5 Conclusion

In this work we have shown that data augmentation via translation is an useful approach to improve the performance of NER models. On the multilingual test set we observe an improvement over the baseline XLM-R by 0.85 F1. Our approach uses open-source components and can potentially benefit NER in various domains.

In the future we plan to explore how different translation configurations can boost NER models. Reducing the noise of augmented sentences should have a positive impact on the performance. This can be achieved by improving the performance of the translation model. Alternatively, the dataset could be augmented with translations from several models. Therefore, using a selection of sentences from different sources (Poncelas and Way, 2019; Soto et al., 2020) to augment the data could also be investigated.

Track	Language	Clean Subset F1	Noisy Subset F1	Overall Macro F1	Rank Status
1	English	71.86	64.06	70.16	11/34
2	Spanish	75.42	67.39	72.85	5/18
3	Swedish	76.74	68.12	73.79	7/16
4	Ukrainian	72.31	-	72.31	6/14
5	Portuguese	72.74	64.76	69.98	9/17
6	French	75.58	66.86	72.86	7/17
7	Farsi	64.88	-	64.88	8/14
8	German	75.74	72.56	76.24	6/17
9	Chinese	68.79	51.2	64.61	9/22
10	Hindi	78.37	-	78.37	6/17
11	Bangla	77.2	-	77.2	6/18
12	Italian	76.67	69.03	74.19	7/15
13	Multilingual	73.82	-	73.82	7/18

Table 5: The official ranking results of Sakura in 13 tracks.

References

Sandeep Ashwini and Jinho D. Choi. 2014. Targetable named entity recognition in social media. *CoRR*, abs/1408.0782.

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online.

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.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark.

Alberto Poncelas, Maja Popović, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining SMT and NMT back-translated data for efficient NMT. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 922–931, Varna, Bulgaria.

Alberto Poncelas and Andy Way. 2019. Selecting Artificially-Generated Sentences for Fine-Tuning Neural Machine Translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 219–228, Tokyo, Japan.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural

machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, Seattle, WA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 107–110, Geneva, Switzerland.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 3898–3908, Seattle, USA.

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 preprint arXiv:2008.00401*.

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

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium.