SemEval 2023: MultiCoNER II Project Proposal

Asal Shavandi Chahyon Ku Josh Spitzer-Resnick London Lowmanstone IV
University of Minnesota
{shava006,ku000045,spitz123,lowma016}@umn.edu

Abstract

The MultiCoNER II shared task aims to detect noisy and out-of-domain complex named entities for multiple languages. While knowledge-retrieval-based systems such as DAMO-NLP and USTC-NELSLIP were successful for the previous version of the task, MultiCoNER I, it still had shortcomings such as sensitivity to noise (typos) and lower performance on out-of-knowledge-base named entities. To alleviate these shortcomings, we propose data augmentation techniques to augment training data with noisy and diverse-domained entities.

1 Introduction

The named entity recognition task is a critical part of information extraction in which every word in a sentence is classified to named entity types such as names of people, organization, location, etc. (Nadeau and Sekine, 2007). Since it was first organized in 1996 at the Sixth Message Understanding Conference, many mono- and multilingual tasks, such as the CoNLL 2003 (Sang and Meulder, 2003), Ontonotes corpus v5 (Pradhan et al., 2013), and WNUT 2017 Emerging Entities (Derczynski et al., 2017) were organized to tackle its challenges.

Among named entities, complex named entities are the more syntactically complex named entities–often names of creative works–that existing systems have a hard time recognizing (Ashwini and Choi, 2014). Complex named entities are more challenging to detect than traditional ones, because they are rarer in data, newly created more frequently, and have more diverse syntactic structures.

Even though knowledge-retrieval-based systems achieved great results at the MultiCoNER shared task from SemEval 2022 (Malmasi et al., 2022), they are sensitive to noisy and out-of-domain entities. The MultiCoNER II shared task proposes new tasks to address the shortcomings of top performing models from the MultiCoNER shared task.

Our group, Team Polygots, attempt to replicate the method DAMO-NLP, the best group from MultiCoNER, used and propose improvements by replacing the backbone language model, modifying the knowledge retrieval system, and augmenting the training data.

2 Related Works

2.1 Named Entity Recognition

Named Entity Recognition (NER) is a core natural language processing (NLP) task (Chen et al., 2022) that has a lot of applications in academia, marketing, medical and security domains. Transformer-based pretrained language models have achieved great success in almost every NLP task (Kalyan et al., 2021) including NER. These models learn universal language representations from large volumes of text data using self-supervised learning and transfer this knowledge to downstream tasks (Kalyan et al., 2021). Multilingual BERT (M-BERT), released by (Devlin et al., 2019) as a single language model pre-trained from monolingual corpora in 104 languages, is shown to be very good at cross-lingual model transfer (Pires et al., 2019). XLM-RoBERTa (Conneau et al., 2019a) is another pretrained multilingual language model at scale that has led to significant performance gain for a wide range of cross-lingual transfer tasks.

2.2 Multilingual Language Models

Fine-tuning pretrained contextual embedding is a useful and effective approach to many NLP tasks (Wang et al., 2022) and recently many researchers have put their effort into training fine-tuned multilingual models such as mBERT and XLM-RoBERTa to improve their model's performance. (Malmasi et al., 2022) designed a NER system using XLM-RoBERTa on MultiCoNER I dataset that computes a representation for each token which was then used to predict the token tag using a Conditional Random Field (CRF) classification

layer (Sutton and McCallum, 2010). Their system resulted in an F1 score of 0.478.

2.3 Best Models from MultiCoNER I

The first ranked team from MultiCoNER I shared task, DAMO-NLP (Wang et al., 2022), which got the highest F1 score for the multilingual task, took a different approach and introduced a knowledge-based NER system which used (Malmasi et al., 2022) system as their baseline and added a knowledge retrieval module to enhance their performance. The knowledge retrieval module takes an input sentence as a query and retrieves top-k related paragraphs from Wikipedia which will be then concatenated and fed into the NER module. The output of the NER module which is a token representation of the input sentence will be fed into a linear-chain CRF to produce the label predictions. This method has shown an F1 score of 0.853.

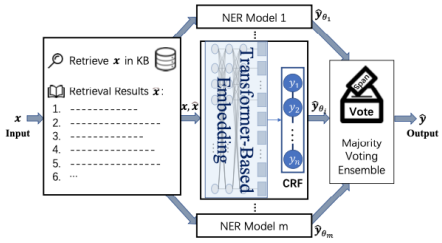

Figure 1: The architecture of DAMO-NLP knowledge-based NER system

The second-ranked team from MultiCoNER I, USTC-NELSLIP (Chen et al., 2022), also used the same concept of a knowledge retrieval system but instead of retrieving top-related paragraphs from Wikipedia, they introduced a gazetteer-adapted integration network (GAIN). This system first adapts the representation of the gazetteer network built from Wikidata to the XLM-RoBERTa model by minimizing KL divergence between them. After adaptation, these networks will be integrated for backend supervised NER training. This method has also shown an F1 score of 0.853.

3 Problem Formulation

MultiCoNER II is a Multilingual Complex Named Entity Recognition shared task, offered as part of SemEval-2023, The 17th International Workshop on Semantic Evaluation. Given a sentence, the task is to detect and categorize all named entities in the sentence. To be more concrete, each word

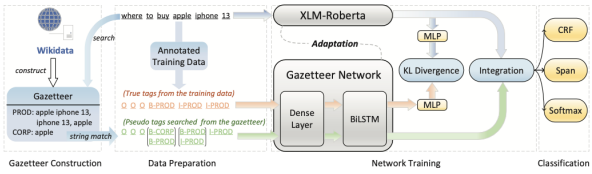

Figure 2: The architecture of USTC gazetteer-adapted integration network

in the sentence is tagged with Beginning-Inside-Outside tags. Beginning tag denotes the first word in a named entity, inside denotes other words in a named entity, and outside means the word is not a part of a named entity. Then, each beginning and inside tags are categorized in to one of 6 categories according to the WNUT 2017 (Derczynski et al., 2017) taxonomy entity types: person, group, corporation, location, product, and creative work. The dataset consists of 11 languages, where each language has 15300 training sentences and 800 development sentences.

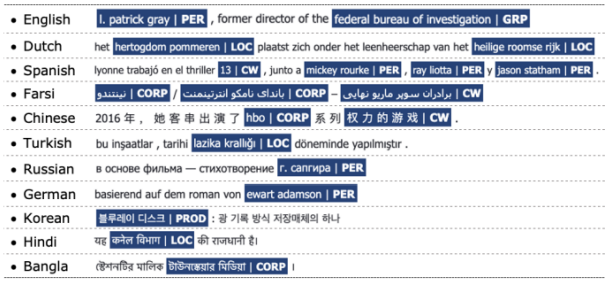

Figure 3: Example sentences and labeled named entities from MultiCoNER (Malmasi et al., 2022)

4 Proposed Idea

4.1 Baseline

To evaluate the strength of our proposal, we will fine-tune an XLM RoBERTa model on the MultiCoNER II dataset (Conneau et al., 2019b). All of top-performing models from the previous MultiCoNER competition utilized XLM RoBERTa (**?**). So, by demonstrating performance improvements to XLM RoBERTa, especially via data augmentation techniques, we hope to provide general techniques that can improve any model previously submitted on the task.

4.2 Proposal

Our proposal is to use two types of data augmentation techniques. The first is standard data augmentation techniques for adding spelling errors and

replacing synonyms within the text. The second is to use a technique called MulDA (Multilingual Data Augmentation). specific to NER. which (Liu et al., 2021) translates sentences into different languages, increasing the diversity of NERs in new contexts.

4.2.1 Standard Data Augmentation

Python Package - NoiseMix NoiseMix is a Python package which quickly and easily modifies text to have typos, including repeated characters, word swaps, and errors based on the QUERTY keyboard layout (Bittlingmayer, 2018a). These sorts of data augmentation have been shown to increase performance on NLP tasks (Bittlingmayer, 2018b). We include NoiseMix specifically because Multi-CoNER II aims to target NER in "noisy scenarios like the presence of spelling mistakes and typos," which this package simulates.

Python Package - NLP Data Augmentation NLP Data Augmentation operates at the word and sentence level rather than at the character level that NoiseMix operates at. This package allows us to replace words with synonyms or remove adjectives that don't affect the meaning of the sentence (Gurung, 2022). Applying these sorts of augmentations will need to be done carefully so as not to edit the NERs or create contexts that a human would reasonably judge to affect the labels. Thus, we will be carefully auditing the results of using this package.

4.2.2 MulDA

MulDA is a data augmentation technique that focuses on multilingual NER, which aligns quite well for this competition (Liu et al., 2021).

The main idea behind MulDA is to keep the NERs the exact same while doing language translation. This enables higher diversity of NERs, as they are found in languages that they normally would not be in. For example, the name "James" would not usually be found in French text, but should still be correctly recognized as a person, regardless of the language. We expect that this augmentation will help with the recognition of complex NERs, such as titles of works, in a multilingual setting.

5 Broader Impact

Multilingual complex named entity recognition has a number of practical impacts, which we will break this down into three primary domain categories: marketing, academia, and areas with domain-specific knowledge.

First, we would be remiss if we did not acknowledge that the authors of this shared task are employed at Amazon. In marketing and global e-commerce, complex entities can occur as long product names (including creative works such as book titles) in unstructured language such as product reviews or references in social media. For global companies, the ability to maintain this understanding across the languages and markets they operate in is just as critical. Named entity recognition is also a feature in Amazon Web Services' managed Comprehend product for natural language understanding, which corporate clients of AWS can find equally valuable.

In academia, research papers often have long and complex names, and many other papers, theorems, and other specific named concepts may be referenced in the body of the text. The ability to identify and understand the relationships between ideas referenced in papers is applicable to efforts such as Allen AI's Semantic Scholar product.

Lastly, other areas such as medical, legal, and political domains with specific vocabulary can benefit greatly. Medical vocabulary such as diagnoses may be difficult to identify but are crucial to understand regardless of the language in which care is provided or records are kept. Providers of electronic medical record systems such as Epic are uniquely positioned to derive insights from such data. The vast corpora of legal and political texts likewise contain references to innumerable laws, concepts, and other ideas that legal professionals make careers out of understanding how each is connected to or builds upon one another. Thomson Reuters' Westlaw product for legal research has much to gain from understanding the connections between legal documents that reference each other. Other companies operating in domains with particular vocabularies can likewise benefit from understanding the connections between domain specific concepts, as well as company products. In a regulated context, identifying and understanding complex phrases such as security questions for account authentication, as well as legal disclosures are important from a quality assurance perspective.

Role Assignments

1. Asal Shavandi: Literature review on recent and related methods
2. Chahyon Ku: Introduction and task formulation
3. Josh Spiter-Resnick: Broader impact
4. London Lowmanstone IV: Proposed ideas

References

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

Adam Bittlingmayer. 2018a. NoiseMix. Original-date: 2018-05-09T21:43:30Z.

Adam Bittlingmayer. 2018b. NoiseMix - data generation for natural language.

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Unsupervised Cross-lingual Representation Learning at Scale.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Pema Gurung. 2022. NLP Data Augmentation. Original-date: 2022-02-17T05:27:43Z.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. 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 5834–5846, Online. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. 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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields.

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.