Exploring X-GEAR for Zero-Shot Cross-Lingual Event Argument Extraction

Ramaswamy Iyappan
MS Computer Science
George Mason University
riyappan@gmu.edu

Abhijeet Banerjee
MS Computer Science
George Mason University
abanerj@gmu.edu

Bhargava Canakapally
MS Computer Science
George Mason University
bcanakap@gmu.edu

1 Introduction

1.1 Task / Research Question Description

The objective of this paper is to introduce a study that uses pre-trained generative language models to achieve zero-shot cross-lingual event argument extraction (EAE). The method proposed in this paper formulates EAE as a language generation task that effectively encodes event structures and captures argument dependencies. The paper proposes the use of language-agnostic templates that can represent event argument structures in any language, making cross-lingual transfer possible. The proposed model is trained by finetuning multilingual pre-trained generative language models to generate sentences that fill in the language-agnostic templates with arguments extracted from the input text. Our motivation is to use multilingual pre-trained generative language models that are able to effectively encodes event structures and captures the dependencies between arguments.

1.2 Motivation and Limitations of existing work

The majority of the models discussed in the paper are based on classification techniques, where classifiers are built on top of pre-trained masked language models that can operate in multiple languages. However, some of these models require additional information, like bilingual dictionaries (Liu et al., 2019; Ni and Florian, 2019), translation pairs (Zou et al., 2018), and dependency parse trees (Subburathinam et al., 2019; Ahmad et al.; Nguyen and Nguyen, 2021), to address differences between languages. Nevertheless, according to previous papers (Li et al.; Hsu et al.), classification-based models are not as effective as generation-based models in capturing the relationships between entities. Numerous other studies have proven the effectiveness of generation-based models in achieving high performance on monolingual structured prediction tasks. However, since the targets generated by these models are language-specific, their direct application to zero-shot cross-lingual tasks may result in suboptimal performance. Recently, there has been increased interest in using prompts to guide the behavior of pre-trained language models and elicit knowledge (Peng et al., 2019; Schick and Schutze, 2021). Despite this, there has been limited focus on multilingual tasks (Winata et al., 2021).

1.3 Proposed Approach

The proposed approach is to devise cross-lingual event argument extraction using in zero-shot setting as a language generation task. Along with this we propose, X-GEAR which stands for Cross-lingual Generative Event Argument extractoR(Huang et al., 2022). The X-GEAR model fine-tunes generative models that are pre-trained for multiple languages, such as mBART-50 (Tang et al.) or mT5 (Xue et al., 2021). It also incorporates a copy mechanism to improve its ability to adapt to changes in input language. The paper provides a detailed explanation of X-GEAR's design, including the language-agnostic templates used, the desired output, the format of the input, and the specifics of its training process.

1.4 Likely challenges and mitigations

The previously mentioned approach introduces two notable challenges The training and testing input languages may differ, being the first, and it is necessary for the output strings to be easily parsed into the final predictions, being the second. To address these challenges, the X-GEAR model uses language-agnostic templates. These templates are structured to include a prompt that includes the trigger, event type, and the language-agnostic template. X-GEAR then generates an output string that fills in the template using information extracted from the input passage. The language-agnostic template is designed in a way that makes it easy to parse the final argument and role predictions from the generated output. Since the template is language-independent, it promotes cross-lingual transfer.

2 Related Work

XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation (Ruder et al., 2021), proposes a method for zero-shot cross-lingual transfer learning in structured prediction tasks, where models are trained on high-resource languages and then applied to low-resource languages without additional labelled data. They evaluated their method on named entity recognition and dependency parsing tasks, showing significant improvements over baselines. While they build classifiers on top of pre-trained masked language models, our work focuses specifically on generating structured predictions in a zero-shot cross-lingual setting by incorporating prompts in a fine-tuning process.

A Unified Generative Framework for Various NER Subtasks (Yan et al., 2021), proposes a framework that is easy-to-implement and achieves state-of-the art (SoTA) or near SoTA performance on eight English NER datasets, including two flat NER datasets, three nested NER datasets, and three discontinuous NER datasets. It demonstrates the success of generation-based models for named entity recognition in a monolingual setting. However, their methods are language-dependent and not suitable for zero-shot cross-lingual structured prediction. In contrast, our work is designed for the zero-shot cross-lingual setting and can generate structured predictions for multiple languages with a single model.

Neural Cross-Lingual Event Detection with Minimal Parallel Resources (Liu et al., 2019), proposed a new method for cross-lingual ED, demonstrating a minimal dependency on parallel resources. It is effective in performing cross-lingual transfer concerning different directions and tackling the extremely annotation-poor scenario. It proposed using bilingual dictionaries and translation pairs, respectively, to handle the discrepancy between languages in zero-shot cross-lingual structured prediction. Our work, on the other hand, does not require such external resources and instead leverages fixed prompts and fine-tuning of the language model.

Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference (Schick and Schutze, 2021), introduced Pattern Exploiting Training (PET), a semi-supervised training procedure that reformulates input examples as cloze-style phrases to help language models understand a given task. PET outperforms supervised training and strong semi-supervised approaches in low resource settings by a large margin. It also proposed a prompt-based method for named entity recognition that uses trainable prompts to adapt the language model to specific tasks. In contrast, our work can generate structured predictions for multiple languages without any labelled data for the target language. Additionally, we focus on zero-shot cross-lingual structured prediction tasks, while theirs is applied to monolingual named entity recognition.

Text generation with exemplar-based adaptive decoding (Peng et al., 2019), proposes an exemplar-based adaptive decoding method for text generation that selects exemplars (i.e., similar instances from a training set) to guide the decoding process. The approach was evaluated on several text generation tasks, including machine translation and text summarization, and showed improved performance over traditional decoding methods. Empirical results show that this model achieves strong performance and outperforms comparable baselines. In contrast, our work on Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction focuses specifically on the task of event argument extraction and utilizes a multilingual generative language model to generate cross-lingual event arguments without the need for labeled data in the target language. Additionally, our method employs a prompt-based approach to guide the model's behavior, rather than selecting exemplars from a training set.

3 Experiments

3.1 Datasets

Initially, we planned on utilizing two popular and commonly used event argument extraction datasets: ACE-2005 and ERE. But since they are not freely available from LDC, we were unable to perform any pre-processing steps. However, we thank the authors (Huang et al., 2022), we make use of their pre-processed train / dev / test splits of English (en), Arabic (ar) and Chinese (zh) annotations for ACE-2005.

Language	Sentence	Event	Argument
Train data			
en	17172	4202	4859
ar	2722	1743	2506
zh	6305	2926	5581
Dev data			
en	923	450	605
ar	289	117	174
zh	486	217	404
Test data			
en	832	403	576
ar	272	198	287
zh	482	190	336

Table 1: Dataset Statistics

Table 1 specifically shows more information about the splits. The English and Chinese splits are from (Wadden et al., 2019) and (Lin et al., 2020) respectively. The Arabic part is from the original authors (Huang et al., 2022). As a result, we primarily focus only on ACE-2005 dataset in this work.

3.2 Implementation

We use the code from https://github.com/PlusLabNLP/X-Gear for re-implementing X-Gear (Huang et al., 2022). We use mT5-base (Xu et al., 2021) as our baseline model, with learning-rate 10e-4, batch size 8 and number of training epochs as 25. check https://github.com/ramiyappan/X-Gear for our GitHub repository.

3.3 Results

Table 2 presents the F1-scores of Argument Classification by our basline model X-Gear mT5-base in comparison to the published models X-GEAR (mT5-base and mBART-50-large) (Huang et al., 2022).

3.4 Discussion

Getting the raw dataset of ACE-2005 and ERE from LDC was not possible because it was licensed and not available for free. Hence, we had to search and look-out several Git repositories to get atleast the pre-processed splits to proceed further with the same paper. Moreover, the Hopper cluster was down a few times and allocating jobs to a GPU was very time-consuming.

The results we obtained were similar to the published ones, even though we just ran for 25 epochs unlike (Huang et al., 2022), where they ran it for 60 training epochs. There is no fixed seed for this work, hence it might produce different results during each training period.

3.5 Resources

We used 80GB GPU memory and cpus-per-task=24 as computational resources. It took 4 days and 6 hours for reproducing this paper and the work was evenly split among the authors.

3.6 Error Analysis

To illustrate some instances where the model fails, here are three examples from the paper:

- In Arabic to English transfer, X-GEAR predicted “The EU foreign ministers” while the ground truth was “ministers”. We attribute this error to the fact that entities in Arabic are usually longer than in English, which leads to the model over-generating some more tokens even though they have captured the correct concept. This error falls under the over-generating category.

- In another example from Arabic to English transfer, X-GEAR predicted Vivendi as the Artifact being sold and division as the Seller, while X-GEAR (en ⇒ en) correctly understood that Vivendi was the Seller and division was the Artifact. This error falls under the grammar difference between languages category, where the word order of the sentence in Arabic is reversed with its English counterpart.

- In a Chinese to English transfer, we observe that X-GEAR generates words that do not appear in the passage. This could be because X-GEAR (zh ⇒ en) mixes up singular and plural nouns. This may be because Chinese does not have morphological inflection for plural nouns. For example, the model generates “studios” as prediction while only “studio” appears in the passage.

Model	en ⇓ en	en ⇓ ar	en ⇓ zh	ar ⇓ en	ar ⇓ ar	ar ⇓ zh	zh ⇓ en	zh ⇓ ar	zh ⇓ zh
X-Gear (mBART-50-large) (Huang et al., 2022)	68.3	37.8	48.9	**30.5**	59.8	29.2	45.9	32.3	63.6
X-Gear (mT5-base) (Huang et al., 2022)	67.9	**42.0**	**53.1**	27.6	**66.2**	**30.5**	**52.8**	32.0	**69.4**
X-Gear (mT5-base) **Ours**	**68.7**	37.9	51.9	26.8	65.2	29.1	51.0	**33.5**	66.2

Table 2: Argument Classification (F1%) Scores of ACE-2005 in comparison to the published models. "en ⇒ ar" represents model transferring from en to ar.

Other analysis that the authors could have run include examining the effect of different hyperparameters on the model's performance, examining the impact of tokenization on the model's performance, analysing the distribution of errors across different domains or genres of text, analysing the performance of the model on different domains, and comparing the performance of X-GEAR with other state-of-the-art models.

4 Conclusion

This paper is reproducible only if ACE-2005 and ERE datasets from LDC (or atleast the preprocessed splits) are accessible. Moving further, we are planning to conduct experiments with other pretrained models, hyper-parameter selection, etc.

References

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Peng. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *In Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. Degree: A data-efficient generation-based event extraction model. In *arXiv preprint arXiv:2108.12724*.

Kuan-Hao Huang, I-Hung Hsu, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Minh Van Nguyen and Thien Huu Nguyen. 2021. Improving cross-lingual transfer for event argument extraction with language-universal sentence structures. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Jian Ni and Radu Florian. 2019. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. 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)*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Timo Schick and Hinrich Schutze. 2021. Exploiting cloze questions for few shot text classification and natural language inference. 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)*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. In *arXiv preprint arXiv:2008.00401*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*.

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. Gradual fine-tuning for low-resource domain adaptation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. 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)*.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*.