Constrained Decoding for Cross-lingual Label Projection

Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu
Georgia Institute of Technology
{dminh6, yangc}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

Abstract

Zero-shot cross-lingual transfer utilizing multilingual LLMs has become a popular learning paradigm for low-resource languages with no labeled training data. However, for NLP tasks that involve fine-grained predictions on words and phrases, the performance of zero-shot cross-lingual transfer learning lags far behind supervised fine-tuning methods. Therefore, it is common to exploit translation and label projection to further improve the performance by (1) translating training data that is available in a high-resource language (e.g., English) together with the gold labels into low-resource languages, and/or (2) translating test data in low-resource languages to a high-source language to run inference on, then projecting the predicted span-level labels back onto the original test data. However, state-of-the-art marker-based label projection methods suffer from translation quality degradation due to the extra label markers injected in the input to the translation model. In this work, we explore a new direction that leverages constrained decoding for label projection to overcome the aforementioned issues. Our new method not only can preserve the quality of translated texts but also has the versatility of being applicable to both translating training and translating test data strategies. This versatility is crucial as our experiments reveal that translating test data can lead to a considerable boost in performance compared to translating only training data. We evaluate on two cross-lingual transfer tasks, namely Named Entity Recognition and Event Argument Extraction, spanning 20 languages. The results demonstrate that our approach outperforms the state-of-the-art marker-based method by a large margin and also shows better performance than other label projection methods that rely on external word alignment. [1]

1 Introduction

Large language models (LLMs) have demonstrated the potential to perform a variety of NLP tasks in zero or few-shot learning settings. This is attractive because labeling data is expensive — annotating fine-tuning data across many languages for each task is not feasible. However, for traditional NLP tasks that involve word/phrase-level predictions, such as named entity recognition or event extraction, the performance of zero and few-shot learning lags far behind supervised fine-tuning methods that make use of large amounts of labeled data (Lai et al., 2023). Prior work has therefore trained multilingual models that support cross-lingual transfer from a high-resource language (e.g., English), where fine-tuning data is available to many low-resource languages where data may not be available (e.g., Bambara, which is spoken primarily in Africa). Encoder-based LLMs such as XLM-RoBERTa (Conneau et al., 2020) or mDeBERTa (He et al., 2021) work surprisingly well for cross-lingual transfer, yet the performance of models that are fine-tuned on target-language data is still significantly better (Xue et al., 2021). Motivated by this observation, we present a new approach to automatically translate NLP training datasets into many languages that uses constrained decoding to more accurately translate and project annotated label spans from high to low-resource languages.

Our approach builds on top of EasyProject (Chen et al., 2023), a simple, yet effective state-of-the-art method for label projection, that inserts special markers into the source sentences to mark annotated spans, then runs the modified sentences through a machine translation (MT) system, such as NLLB (Costa-jussà et al., 2022) or Google Translate. A key limitation of EasyProject, as noted by Chen et al. (2023), is that inserting special markers into the source sentence then translating it degrades the translation quality; nevertheless, EasyProject

[1] **This work has been accepted at the *Twelfth International Conference on Learning Representations*, 2024**

was shown to be more effective than prior work for label projection that largely relied on word alignment (Yarmohammadi et al., 2021). To address the problem of translation quality degradation, in this paper, we present a new approach, **Co**nstraint **De**coding for **C**ross-lingual Label Projection (CODEC), for translating training datasets using a customized constrained decoding algorithm. The training data in the high-resource language is first translated *without markers* followed by a second constrained decoding pass to inject the markers. Since the source sentence does not include markers during the translation phase, the final translated text quality from CODEC is preserved. The second decoding pass of CODEC relies on a translation model that is conditioned on the modified input sentence *with markers* (thus is noisier) in order to find the appropriate positions for inserting markers. Using a specially designed constrained decoding algorithm, however, we can retain the high-quality translation while having the right number of labels projected by enforcing both as constraints during decoding.

In essence, CODEC only explores the search space which contains valid hypotheses, i.e., translated outputs that conform to (i) the high-quality translation from the first decoding pass without markers' interference and (ii) having the correct number of markers inserted. A brute-force enumeration of all possible such hypotheses is intractable, as the number of sequences that would need to be scored using the translation model is $O(n^{2m})$, where n is the sequence length and m is the number of labeled spans to be projected. We therefore design a constrained decoding algorithm based on the branch-and-bound method (Stahlberg and Byrne, 2019), in which a depth-first search is conducted to identify a lower-bound on the best complete hypothesis, and branches that do not have any solutions with a better score than the current lower bound are pruned from the search space.

However, even when pruning branches using this bound, decoding time is still prohibitively long. To speed up decoding, we introduce a new *heuristic lower bound*, which removes branches more aggressively. We also introduce a technique to prune unlikely positions for the opening markers in advance. Putting everything together, compared to exact branch-and-bound search, our proposed method significantly reduces decoding time with only a slight drop in performance in a few languages. For example, for the Bambara language, CODEC is about 60 times faster than exact search, while only losing 0.6 absolute F1, a 1.1% drop in performance.

We conduct extensive experiments to evaluate CODEC on two popular cross-lingual tasks (i.e., Named Entity Recognition and Event Argument Extraction), covering 20 language pairs. In our experiment, CODEC and other label projection baselines are used to project the label from English datasets to their translated version to augment the data in the target language, which is referred to as *translate-train* (Hu et al., 2020). The results demonstrate that, on average, the model fine-tuned on CODEC-augmented data outperforms models fine-tuned on the data produced by other label projection baselines by a large margin. This reinforces our hypothesis that preserving translation quality is essential and constrained decoding can improve the accuracy of label projection. Moreover, since CODEC separates two phases of translation and marker insertion, it can also be used to improve cross-lingual transfer by using machine translation at inference time, sometimes referred to as *translate-test* (Artetxe et al., 2023). This approach translates test data from the low-resource language to the high-resource language, uses a fully-supervised NLP model to automatically annotate the translation on the high-resource side, then projects the annotations back to the original language. Experiments show that, compared to translate-train alone, using CODEC in the translate-test setting further boosts cross-lingual transfer performance in the Named Entity Recognition task. Finally, we evaluate the performance of GPT-4 (Achiam et al., 2023) on the same task and find that, GPT-4 performs well but still falls behind CODEC with a big gap on average.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual

transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. 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. 2020. 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.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

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*, pages 483–498.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, et al. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967.