WebNLG Challenge 2023: Domain Adaptive Machine Translation for Low-Resource Multilingual RDF-to-Text Generation

Kancharla Aditya Hari
IIIT Hyderabad

Bhavyajeet Singh
IIIT Hyderabad

Anubhav Sharma
IIIT Hyderabad

Vasudeva Varma
IIIT Hyderabad

Abstract

This paper presents our submission to the WebNLG Challenge 2023 for generating text in several low-resource languages from RDF-triples. Our submission focuses on using machine translation for generating texts in Irish, Maltese, Welsh and Russian. While a simple and straightfoward approach, recent works have shown that using monolingual models for inference for multilingual tasks with the help of machine translation (*translate-test*) can outperform multilingual models and training multilingual models on machine-translated data (*translate-train*) through careful tuning of the MT component. Our results show that this approach demonstrates competitive performance for this task even with limited data.

1 Introduction

The multilingual RDF-to-Text generation task involves generating verbalizations of RDF triples in different languages. It represents an important problem, as it can allow for the rapid generation and enrichment of informative texts in different languages by leveraging existing knowledge graphs and databases. This is of particular relevance for low-resource languages which are typically less represented.

Recent methods for RDF-to-Text generation rely on neural methods, which require availability of high quality data for training. However, this is challenging for low-resource languages due to the lack of such data. Manual curation of such datasets requires expert human supervision, and is expensive and time-consuming. Thus, utilizing machine translated data available in different languages is naturally motivated. The *translate-train* method relies of translating the training data to the target language, while the *translate-test* method translate the test data to English and using a monolingual model for inference.

Recent works have shown that through careful adaption of the MT component, the *translate-test* method can perform significantly better than previously shown results, even outperforming *translate-train* and multilingual models across different classification tasks (Artetxe et al., 2023). Our submission investigates the performance of *translate-test* method for generation. In particular, we use domain adaption to improve the MT component and generate high quality texts in the target language with very little training data. We first generate text in English using a monolingual, English generation model, and then translate the generated text to the target text using the MT component.

2 Experimental Setup

2.1 Pre-trained model

We use the T5 model (Raffel et al., 2020) for RDF-to-Text generation in English. The task is thus modeled as a text-to-text generation task, with the input RDF triples serialized to strings. The small, 60M parameter variant of T5 is used. The model is finetuned on the training split of WebNLG 2020 challenge's English dataset.

2.2 MT component

For the MT component, we use the distilled variant of NLLB (Team et al., 2022) with 600M parameters. Beam search with a beam size of 5 is used as the decoding strategy.

In (Artetxe et al., 2023), the authors investigate augmenting the training data using rount-trip translation. This is motivated by the idea that the model at inference takes translated data as input. However, due to the crosslingual nature of this problem, the input is always in English, rendering this process irrevelant. Hence we only investigate domain adaption. For domain adaption, we finetune this model on translating the text pairs made available in the development set of WebNLG challenge 2023 for Irish, Maltese, Russian, and Welsh. We do not use

Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 93–94
September 12, 2023 ©2023 Association for Computational Linguistics

Language	BLEU	chrF++	TER
Irish	15.66	0.44	0.73
Maltese	16.49	0.47	0.7
Russian	36.01	0.57	0.53
Welsh	20.97	0.49	0.67

Table 1: Results of translate-test with domain adaption

Breton as this language is not present in NLLB's pretraining corpus. We translate at a document level, without segmenting the texts into individual sentences. At test time, the generations of the T5 model are translated to the target text by the fine-tuned NLLB model, again at a document level.
We also make available the training code and model checkpoints[1]

2.3 Dataset

The T5 model for generating text in English prior to MT is trained on the training split of the English language data made available in WebNLG Challenge 2020.
The MT component is finetuned on only the development set of all languages except Breton made available in WebNLG Challenge 2023. We withold 20% of this dataset for evaluating the MT component. In total, 1332 samples are used for training and 333 samples are used for evaluation.

3 Results

Table 1 shows the automatic evaluation results of our approach. Despite relying on just the few samples available for the low-resource languages, this approach demonstrates strong performance.

References

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

[1] https://shorturl.at/hsN04