GoURMET – Machine Translation for Low-Resourced Languages

Peggy van der Kreeft
Deutsche Welle, Bonn, Germany
peggy.van-der-kreeft@dw.com

Sevi Sariisik
BBC, London, UK
sevi.sariisik@bbc.co.uk

Wilker Aziz
Univ. of Amsterdam, Netherlands
w.ferreiraaziz@uva.nl

Alexandra Birch
Univ. of Edinburgh, UK
A.Birch@ed.ac.uk

Felipe Sánchez-Martínez
Univ. of Alicante, Spain
fsanchez@dlsi.ua.es

Abstract

The GoURMET project, funded by the EU H2020 research and innovation action (under grant agreement 825299), develops models for machine translation, in particular for low-resourced languages. Data, models and software releases as well as the GoURMET Translate Tool are made available as open source.

1 The Project

GoURMET (Global Under-Resourced Media Translation) started in January 2019 and runs until 30 June 2022.

The consortium consists of five partners: The University of Edinburgh (coordinator), University of Alicante, University of Amsterdam, and user partners BBC and Deutsche Welle (DW). [1]

The aim is to significantly improve the robustness and applicability of neural machine translation (NMT) for low-resourced language pairs and domains. This is in particular important because machine translation (MT) is increasingly used as a technology for supporting communication in a globalized world. The two international broadcasters participating in GoURMET are faced with the need to use MT to support their editorial work, especially for languages for which such tools are currently hard to find or lack quality.

The main objectives of the project are:

- to advance deep-learning for natural language applications
- to arrive at high-quality MT for low-resourced languages and diverse language pairs and domains
- to develop tools for media analysts and journalists in the form of a sustainable and maintainable platform and services.

The work is built around three use cases. The first is *global content creation*, where we use MT in multilingual content production, with editorial control. The second use case is *media monitoring* for low-resourced and especially strategically important languages. The third use case focuses on *a specific topic*, and the health sector, in particular COVID, was selected for this purpose, fitting the news requirements over the past two years. The objective in the last use case is to apply transfer learning between topical domains.

2 Languages Covered

MT models were selected for sixteen low-resourced languages, jointly by user and technology partners and developed in different phases of the project. These languages are: Amharic, Bulgarian, Burmese, Gujarati, Hausa, Igbo, Kyrgyz, Macedonian, Pashto, Serbian, Swahili, Tamil, Tigrinya. Turkish, Urdu, and Yoruba – all of them from and into English. Pashto was selected as a "surprise language" and

[1] https://gourmet-project.eu/

developed as a special case upon request in a period of two months.

Different factors were taken into account for the selection process, including strategic importance for the news partners, proximity of languages, research interest and complexity for the development of the models.

3 Research and Development of the Models

Each selected language was assigned to one technology partner, who developed the model. Different methods were used among the consortium, allowing a comparison and evaluation of pros and cons of each development method, encouraging enhancement of processes and exchange among research partners.

Research was done as to the availability of data. Data was gathered for each language, from external sources and user partner content. Bilingual datasets were established and manually annotated by editors from BBC and/or DW in terms of their level of equivalence.

Novel approaches were used to enhance the results. One such approach is the multi-task learning data augmentation (MTL DA), in which we generate additional parallel sentences which, despite being completely unlikely under the data distribution, systematically improve the quality of the resulting NMT system. The output proves to be more robust against domain shift and produce less hallucinations.

We also produced a survey covering the state of the art in low-resource MT research.[2]

4 Evaluation and Benchmarking

The user partners evaluated the MT models using a customized evaluation process, including direct assessment (by native speakers of the low-resourced languages), gap filling (looking at English-language MT output) and post editing. Specific assessment user interfaces (UI) and test sets were developed for this purpose.

Technical benchmarking provided a comparative analysis of GoURMET models with Google MT models using BLEU-scores and chrF-scores. In addition, user partners benchmarked the MT output from an editorial point of view, including considering the usefulness and adequacy of the respective models in the field and for different purposes (e.g. understanding or multilingual text production).

5 Applications for the Models

The models are trialed and implemented in several applications by the user partners in the project. First of all, an open-source GoURMET Translate Tool has been developed as a customized UI for text translation in all GoURMET languages.

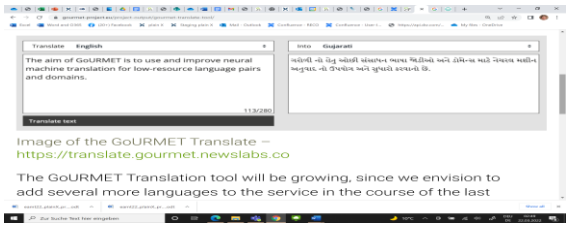

Figure 1: GoURMET Online Translation UI

BBC has implemented some of the GoURMET models in three prototypes, including its multilingual MT prototype Frank[3].

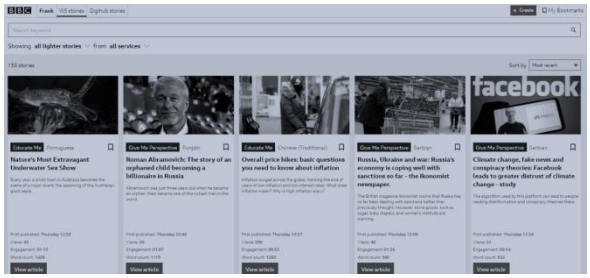

Figure 2: BBC's Frank Multilingual Prototype

DW has incorporated it in the plain X (semi-)automated translation and subtitling platform and as an application of the SELMA[4] project.

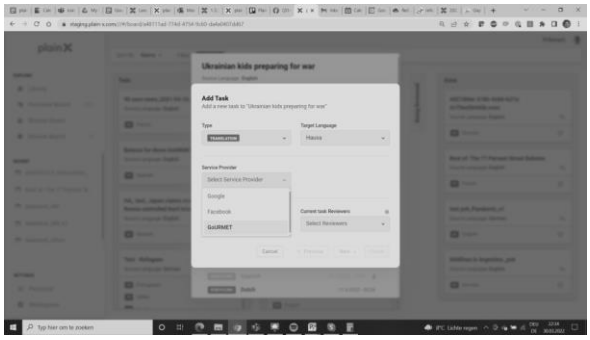

Figure 3: GoURMET in DW's plain X HLT platform

[2] https://arxiv.org/abs/2109.00486

[3] https://bbcnewslabs.co.uk/projects/Frank/

[4] https://selma-project.eu/