GoURMET: Global Under-Resourced Media Translation

www.gourmet-project.eu

Alexandra Birch,[1] Barry Haddow,[1] Ivan Titov,[1], Antonio Valerio Miceli Barone,[1] Rachel Bawden,[1] Felipe Sánchez-Martínez,[2] Mikel L. Forcada,[2] Miquel Esplà-Gomis,[2] Víctor M. Sánchez-Cartagena,[2] Juan Antonio Pérez-Ortiz,[2] Wilker Aziz,[3] Andrew Secker,[4] Peggy van der Kreeft[5]

[1]University of Edinburgh (GBR), [2]Universitat d'Alacant (ESP) [3]Universiteit van Amsterdam (NDL), [4]BBC (GBR), [5]Deutsche Welle (DEU)

Rationale

1. Machine translation (MT) is an increasingly important technology for supporting communication in a globalized world.
2. Over the last few years neural machine translation methods have led to significant improvements in translation quality.
3. The aim of GoURMET is to significantly improve the robustness and applicability of neural machine translation for low-resource language pairs and domains.

Objectives

1. Advancing low-resource deep learning for natural language applications;
2. High-quality machine translation for low-resource and diverse language pairs and domains;
3. Development of tools for media analysts and journalists;
4. Sustainable, maintainable platform and services;
5. Dissemination and communication of project results to stakeholders.

Use cases

1. Global content creation – managing content creation in several languages efficiently by providing machine translations for correction by humans;
2. Media monitoring for low-resource language pairs – tools to address the challenge of monitoring media in strategically important languages;
3. International business news analysis – reliably translating and analysing news in the highly specialised financial domain.

Research directions

GoURMET will pursue four complementary research directions and integrate our solutions into our user partner work-flows:

1. **Data Gathering and Augmentation**: gathering existing corpora and language resources, and machine-learned augmentation of corpora.
2. **Modelling Morphological Structure**: learning to model words and to induce their linguistic structure to increase the effectiveness of neural machine translation.
3. **Structure Induction at Sentence Level**: learning and exploiting sentence-level structure in neural translation models to increase their learnability, starting from plain text corpora.
4. **Transfer Learning**: developing new techniques to transfer knowledge from related tasks such as language modelling, word prediction and translation of languages related to the low-resource languages of the project.

Languages

Languages in the project are defined by media partners BBC and Deutsche Welle.

Initial languages: Swahili, Turkish, Gujarati and Bulgarian (bidirectional translation models delivered to media partners for evaluation)

Other languages considered: Bosnian, Croatian, Hausa, Igbo, (North) Korean, Kurdish (Sorani and Kurmanji), Macedonian, Afaan Oromoo, Punjabi, Serbian, Tigrinya, Yoruba.

Surprise language: Mid-way through the project, media partners will select a surprise language for researchers to tackle.

Impact

- The outputs of the project will be field-tested at partners BBC and DW.
- The platform will be further validated through innovation intensives such as the workshops centred around our user group and BBC NewsHacks.

Factsheet

Consortium: **The University of Edinburgh** (coordinator, UK), Universitat d'Alacant (Spain), Universiteit van Amsterdam (the Netherlands), BBC (UK), Deutsche Welle (Germany).

Period: 2019–2021

Funding: European Union's Horizon 2020 research and innovation programme; grant agreement No. 825299.

Contact: info@gourmet-project.eu

Acknowledgements

European Commission | Horizon 2020 European Union funding for Research & Innovation