Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: GoURMET

Project name: Global Under-Resourced Media Translation

Project coordinator: Alexandra Birch (University of Edinburgh)

Project consortium:

- University of Edinburgh
- Universiteid van Amsterdam
- Universitat d'Alacant
- BBC
- Deutsche Welle

Funding: EU Horizon 2020

Project duration: 39 Months

Main key words: Translation, media, low-resource

Background of the research topic: Machine translation works very well in situations where there are millions of translated sentences for training models. For low-resourced language pairs (this is the list of languages tackled in GoURMET), however, the quality of translation is barely, if at all, usable. Our project is focussed on both collecting and creating low-resource language data, and pushing forward the latest research in machine learning to be able to make the best use of the little data we have.

The media industry is one of the pillars of a functioning democracy and it is increasingly under a range of threats such as political populism and social media content aggregators. Our project will help the media industry to thrive by allowing them to reach a bigger audience with less effort. Our translation models will allow journalists to understand a broad spread of news from countries of interest, and to produce content faster in local languages by leveraging the output of machine translation models.

Goal of the project:

1. Advancing low-resource deep learning for natural language applications

2. Development of high-quality machine translation for low-resource language pairs and domains

3. Development of tools for media analysts and journalists.

Project abstract: Machine translation (MT) is an increasingly important technology for supporting communication in a globalised world. MT technology has gradually increased over the last ten years, but recent advances in neural machine translation (NMT), have resulted in significant interest in industry and have lead to very rapid adoption of the new paradigm (eg. Google, Facebook, UN, World International Patent Office). Although these models have shown significant advances in state-of-the-art performance they are data intensive and require parallel corpora of many millions of human translated sentences for training. Neural Machine translation is currently not able to deliver usable translations for the vast majority of language pairs in the world. This is especially problematic for our user partners, the BBC and DW who need access to fast and accurate translation for languages with very few resources.

The aim of GoURMET is to significantly improve the robustness and applicability of neural machine translation for low-resource language pairs and domains.

The project will focus on two use cases:

Click here to be redirected to the virtual room of the Project Expo.

- Global content creation - managing content creation in several languages efficiently by providing machine translations for correction by humans;
- Media monitoring for low resource language pairs - tools to address the challenge of international news monitoring problem.

The outputs of the project are being field-tested at partners BBC and DW.

Publications:

Authors	**Title**	**Venue**	**Date**	**Status (submitted, accepted, published)**
Antonio Valerio Miceli-Barone, Alexandra Birch, Rico Sennrich	Distributionally Robust Recurrent Decoders with Random Network Distillation	Arxiv	2021	https://arxiv.org/abs/2110.13229
Alexandra Birch, Barry Haddow, Antonio Valerio Miceli-Barone, Jindřich Helcl, Jonas Waldendorf, Felipe Sánchez-Martínez, Mikel L Forcada, Víctor Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady, Sevi Sariisik, Peggy van der Kreeft, Kay Macquarrie	Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months	MTSummit	2021	Accepted https://aclanthology.org/2021.mtsummit-research.8/
Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, Alexandra Birch	Survey of Low-Resource Machine Translation	Computational Linguistics	2021	Submitted https://arxiv.org/abs/2109.00486
Víctor M.Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez	Rethinking data augmentation for low-resource neural machine translation: a multi-task learning approach	EMNLP	2021	Accepted https://arxiv.org/abs/2109.03645
F Arthaud, R Bawden, A Birch	Few-shot learning through contextual data augmentation	EACL	2021	accepted https://www.aclweb.org/anthology/2021.eacl-main.90/

Click here to be redirected to the virtual room of the Project Expo.

Christos Baziotis, Ivan Titov, Alexandra Birch, Barry Haddow	Exploring Unsupervised Pretraining Objectives for Machine Translation	ACL	2021	submitted (under review)
Gonçalo M. Correia, Vlad Niculae, Wilker Aziz, and André F. T. Martins	Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity	NeurIPS	2020	https://papers.nips.cc/paper/2020/hash/887caadc3642e304ede659b734f79b00-Abstract.html Spotlight paper
Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de-Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann and Lana Yeganova	Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages	WMT	2020	Accepted
Nikita Moghe, Christian Hardmeier and Rachel Bawden	The University of Edinburgh-Uppsala University's Submission to the WMT 2020 Chat Translation Task	WMT	2020	Accepted
Rachel Bawden, Biao Zhang, Andre Tättar and Matt Post	ParBLEU: Augmenting Metrics with Automatic Paraphrases for the WMT’20 Metrics Shared Task	WMT	2020	Accepted
Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli Barone and Philip Williams	The University of Edinburgh’s English-Tamil and English-Inuktitut Submissions to the WMT20 News Translation Task	WMT	2020	Accepted

Click here to be redirected to the virtual room of the Project Expo.

Wilker Aziz	Demystifying Deep Generative Language Models under review	ACL	2020	rejected at ACL2020
Mathijs Pieters and Wilker Aziz	Interpretable Text Classification using Latent Vocabularies under review	ACL	2020	rejected at ACL2020
Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, Felipe Sánchez-Martínez	Bicleaner at WMT 2020: Universitat d'Alacant-Prompsit's submission to the parallel corpus filtering shared task	WMT	2020	accepted
Víctor Manuel Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez	Understanding the effect of morphological tags in under-resourced neural machine translation	COLING	2020	accepted
Christos Baziotis, Barry Haddow, Alexandra Birch	Language Model Prior for Low-Resource Neural Machine Translation	EMNLP	2020	accepted at EMNLP https://arxiv.org/abs/2004.14928
Arturo Oncevay, Barry Hadddow, Alexandra Birch	Bridging linguistic typology and multilingual machine translation with multi-view language representations	EMNLP	2020	accepted at EMNLP https://arxiv.org/abs/2004.14923
Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar and Matt Post	A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing	EMNLP Findings	2020	accepted at EMNLP Findings https://arxiv.org/abs/2004.14989
Bryan Eikema and Wilker Aziz	Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation	Coling	2020	https://www.aclweb.org/anthology/2020.coling-main.398/ Best paper
Nicola De Cao, Michael Schlichtkrull, Wilker Aziz, Ivan Titov	How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking	EMNLP	2020	https://www.aclweb.org/anthology/2020.emnlp-main.262/
Barry Haddow, Faheem Kirefu	PMIndia -- A Collection of Parallel Corpora of Languages of India		2020	preprint https://arxiv.org/abs/2001.09907

Click here to be redirected to the virtual room of the Project Expo.

Susie Coleman, Andrew Secker, Rachel Bawden, Barry Haddow and Alexandra Birch	Architecture of a Scalable, Secure and Resilient Translation Platform for Multilingual News Media	IWLTP	2020	accepted Paper
António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang and André T. Martins	Document-level Neural MT: A Systematic Comparison	EAMT	2020	accepted Paper (proceedings are a single PDF)
Radina Dobreva, Jie Zhou and Rachel Bawden	Document Sub-structure in Neural Machine Translation	LREC	2020	accepted
Duygu Ataman, Wilker Aziz, Alexandra Birch	A Latent Morphology Model for Open-Vocabulary Neural Machine Translation	ICLR	2020	accepted https://openreview.net/forum?id=BJxSI1SKDH Spotlight paper before that: rejected at EMNLP2019
Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen and Alexandra Birch	Toward Making the Most of Context in Neural Machine Translation	IJCAI	2020	accepted
Andrea Zaninello, Alexandra Birch	MultiWord Expression Aware Neural Machine Translation	LREC	2020	accepted
Biao Zhang, Philip Williams, Ivan Titov, Rico Sennrich	Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation	ACL	2020	accepted
Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, Julie Wall	An English-Swahili parallel corpus and its use for neural machine translation in the news domain	EAMT	2020	accepted
Víctor M. Sánchez-Cartagena, Mikel L.	A multi-source approach for Breton-French hybrid machine translation	EAMT	2020	accepted

Click here to be redirected to the virtual room of the Project Expo.

Forcada, Felipe Sánchez-Martínez				
Tom Pelsmaeker and Wilker Aziz	Effective Estimation of Deep Generative Language Models	ACL	2020	accepted rejected at EMNLP
De Cao, Nicola, Ivan Titov, and Wilker Aziz	Block neural autoregressive flow	Arxiv	2019	accepted
Emelin, Denis; Titov, Ivan; Sennrich, Rico;	Widening the Representation Bottleneck in Neural Machine Translation with Lexical Shortcuts	WMT	2019	accepted
Carolina Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia	Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality	16th International Workshop on Spoken Language Translation (IWSLT 2019)	2019	accepted
Loïc Barrault , Ondřej Bojar |,Marta R. Costa-jussà , Christian Federmann , Mark Fishel , Yvette Graham , Barry Haddow ,Matthias Huck , Philipp Koehn , Shervin Malmasi , Christof Monz , Mathias Müller , Santanu Pal |, Matt Post and Marcos Zampieri	Findings of the 2019 Conference on Machine Translation (WMT19)	Conference on Machine Translation (WMT19)	2019	published
Arturo Oncevay, Barry Haddow and Alexandra Birch	Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features	TyP-NLP https://typology-and-nlp.github.io/	2019	accepted, best paper

Click here to be redirected to the virtual room of the Project Expo.

Duygu Ataman	On the Importance of Word Boundaries in Character-level Neural Machine Translation	WNGT	2019	accepted
Bryan Eikema and Wilker Aziz	Auto-Encoding Variational Neural Machine Translation	RepL4NLP	2019	accepted
Joost Bastings, Wilker Aziz, and Ivan Titov	Interpretable Neural Predictions with Differentiable Binary Variables	ACL	2019	accepted
Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, Alexandra Birch	The University of Edinburgh's Submissions to the WMT19 News Translation Task	WMT	2019	accepted
Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez.	The Universitat d'Alacant submissions to the English-to-Kazakh news translation task at WMT 2019	WMT	2019	accepted