Machine Translation

Alexandra Birch, Ulrich Germann, Tomasz Dwojak,
Andrei Popescu-Belis and Hassan M. Sajjad

a.birch@inf.ed.ac.uk, ugermann@inf.ed.ac.uk, t.dwojak@amu.edu.pl,
andrei.popescu-belis@idiap.ch, hsajjad@qf.org.qa

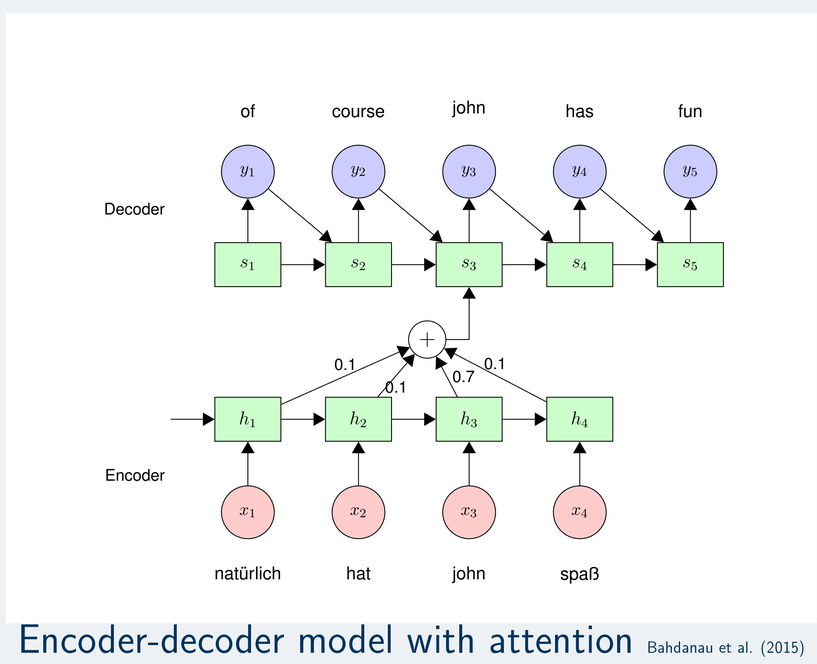
Overview

Translation of media content from across the world into English is a key enabling technology. In the SUMMA platform it allows the monitor to get a broad view of the news and also allows us to apply sophisticated natural language processing tools which have been developed largely for English.
Translation for media monitoring faces the following challenges:

- Large volume of incoming text
- High resource (Arabic, German, Spanish, Portuguese, Russian, Latvian) and low resource (Ukrainian, Farsi) language pairs
- Translating output from speech recognition: potential errors, no segmentation, punctuation or capitalization
- Large variety of text styles and registers: speech, newswire, social media
- Constantly changing media landscape

MT meets Neural Networks

Machine translation has recently undergone a paradigm shift from phrase-based statistical models which combined many hand-engineered features, each applied independently, to one large neural network model where features are implicit and global dependencies are captured.



Encoder-decoder model with attention Bahdanau et al. (2015)

NMT toolkits developed for SUMMA

Nematus is implemented in Theano/Python. The toolkit prioritizes high translation accuracy, usability, and extensibility. Nematus has been used to build top-performing submissions to shared translation tasks at WMT 2016/2017 and IWSLT 2016. It is widely used in academic publications.

Marian is an open source neural network toolkit developed specifically for NMT. Marian provides fast, scalable, and efficient production ready software with no external dependencies. It is written in C++/CUDA and offers distributed GPU and CPU capabilities. Marian's CPU translation speed is nearly as good as Nematus' decoding speed on GPU. If GPUs are available, there is a speed up of a factor of ten.

Links

`https://github.com/rsennrich/nematus`
`https://marian-nmt.github.io/`

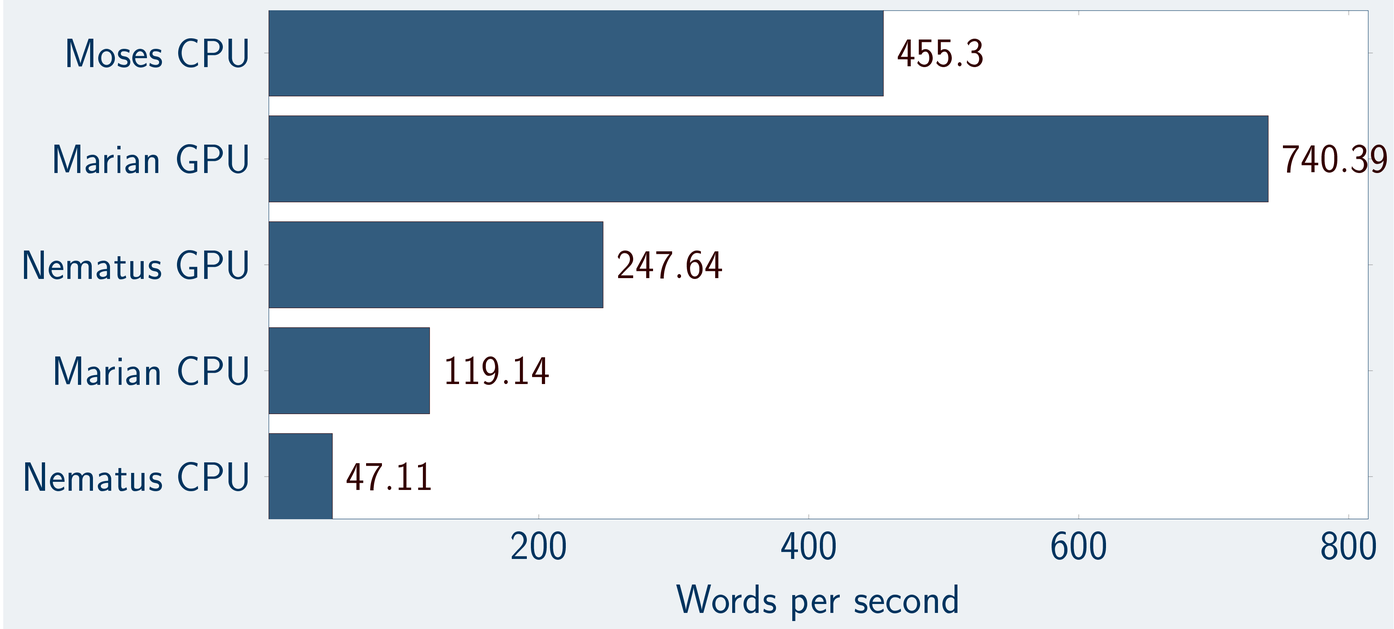
Translation Quality

The main goal of the SUMMA project is to deliver high quality machine translation. We deploy the-state-of-the-art MT models which won numerous tracks at the WMT 2017 shared task "News Translation". Our innovations include:

- Dealing with morphologically rich languages: translating sub-word units (BPE)
- Leveraging in-domain English text: backtranslation
- Deeper models

Translation direction	Shared Task BLEU	Google BLEU
German-English	35.10	32.48
Arabic-English	31.78	30.72
Russian-English	39.14	31.18
Spanish-English	26.83	34.20
Latvian-English	19.00	15.72

Performance



Batch-Decoding Using GPU, Marian can reach performance up to 5,000 wps.

Future Research

Spoken Language Translation

Spoken language translation is a challenge for MT. MT models are trained on written text and they struggle to translate ASR output. We are investigating using further information from the ASR model to improve translation of spoken language. Where ASR models are poor, using potentially more reliable phoneme sequences as additional source information to the MT model, can lead to better translations.

Dialects in Arabic

We want to be able to translate morphologically rich, resource poor Arabic dialects into English. We are investigating language independent tools for segmenting and processing dialects to optimize translation quality.

Low Resource Languages

We will provide translation models for Ukrainian and Farsi. Those languages are "low resourced" as there is very little existing translated corpora for training them. We will develop methods to deal with low resourced languages by leveraging corpora from related languages and applying machine learning techniques such as self-training.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 688139 (SUMMA).