Machine Translation

Alexandra Birch, Ulrich Germann, Faheem Kirefu

a.birch@inf.ed.ac.uk, ugermann@inf.ed.ac.uk, fkirefu@inf.ed.ac.uk

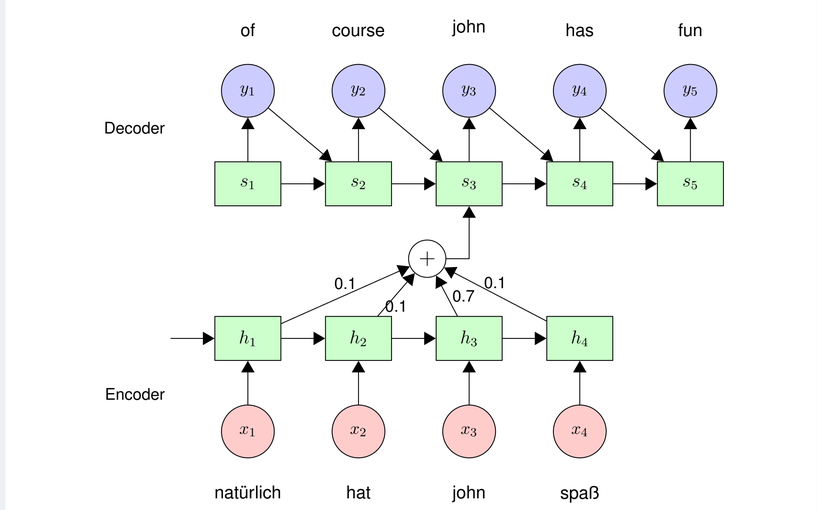
Overview

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 natual 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, Arabic, Russian, Latvian) and low resource (Ukranian, Farsi) language pairs
- Translating output from speech recognition: potential errors, no segmentation, punctuation or capitalisation
- 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.

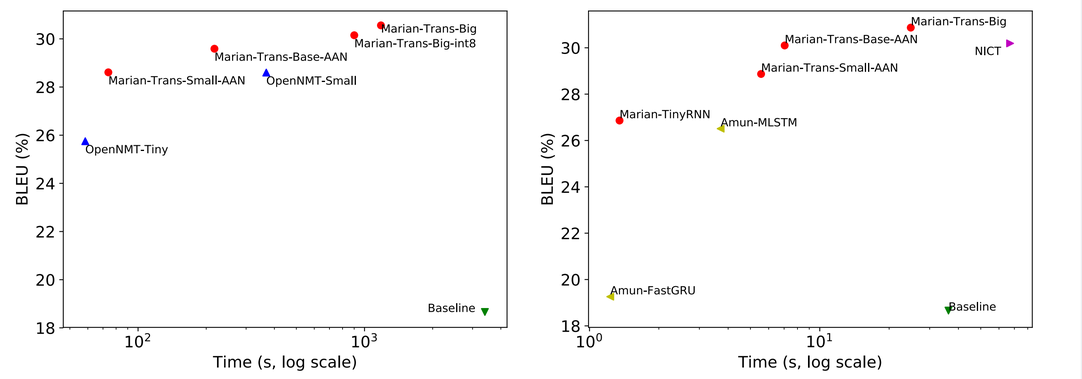
Marian is an open source neural network toolkit developed specifically for NMT. Marian provides fast, scaleable, and efficient production ready software with no external dependencies. It is written in C++/CUDA and offers distributed GPU and CPU capabilities. Marian's CPU translation speed is nearly as 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 Speed: WNMT18



(a) CPU Time vs. Accuracy

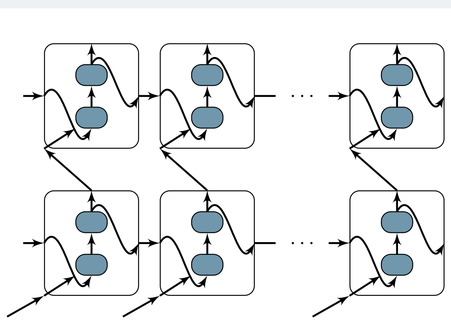
(b) GPU Time vs. Accuracy

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 models which incorporate context

Deeper Models: WMT17

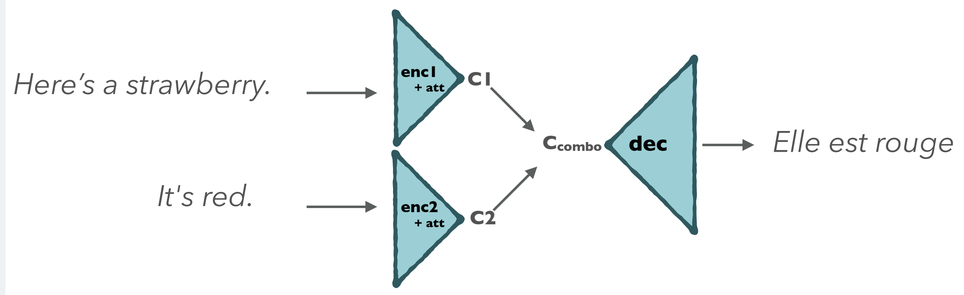


Constrained news task, including ties:

	From English	To English
Turkish	Edinburgh	Edinburgh
Czech	Edinburgh	Edinburgh
Chinese	Edinburgh	Edinburgh
German	Munich	Edinburgh
Russian	Edinburgh	National Research Council
Latvian	Tilde	Edinburgh
Finnish	Helsinki	UPC

(Edinburgh did not participate in Finnish)

Contextual Models



	BLEU	Co-reference	Ambiguity
Baseline	19.52	50.0	50.0
Multis. (prev. source)	20.22	50.0	53.0
Multis. (prev. target)	17.89	47.0	50.5
Concat. (src + tgt)	20.09	63.5	52.0
Multis. and Concat. (tgt)	**20.85**	**72.5**	**57.0**

Future Research

Low Resource Languages

We have used transfer learning and multilingual models to train Ukranian which is a low-resource language. We will extend this reasearch in a new project called GoURMET, translating African and Indian languages.

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