Data collection for Neural MT within CEF eTranslation

Andreas Eisele
MT Engines Chief Scientist
European Commission, DGT

20 November 2018

Questions to address

- Background: MT@EC and CEF eTranslation
- Neural MT: Benefits and challenges
- Training data and approaches to data curation
- What next?

Some Background: MT@EC

- In 2010, DGT started developing a Statistical MT solution for translators and end-users in EU and member state administrations, which went officially into production mid 2013 (after a "real-life trial" within DGT since 2011)
- MT@EC was designed to cover all 24 official EU languages and possibly more
- MT@EC evolved from SMT to hybrid MT, using rule-based pre- and post-processing around a Moses-based statistical core
- Demand has grown steadily over the years, both from end-users and translators (covering ~80 out of 556 language pairs*, tens of millions of pages translated), and feed-back from translators was very helpful to improve the system
- Quality depends strongly on the complexity of the target language; inflection and free word order seriously impair SMT quality
- Hence, for more than half of the EU languages, a purely SMT-based solution turned out to be not sufficient

* 24x23 EU languages + EN ←→ {NB,IS}

More Background: CEF eTranslation

Within the "Connecting Europe Facility", DG CNECT asked us to evolve our service into a building block for CEF, providing

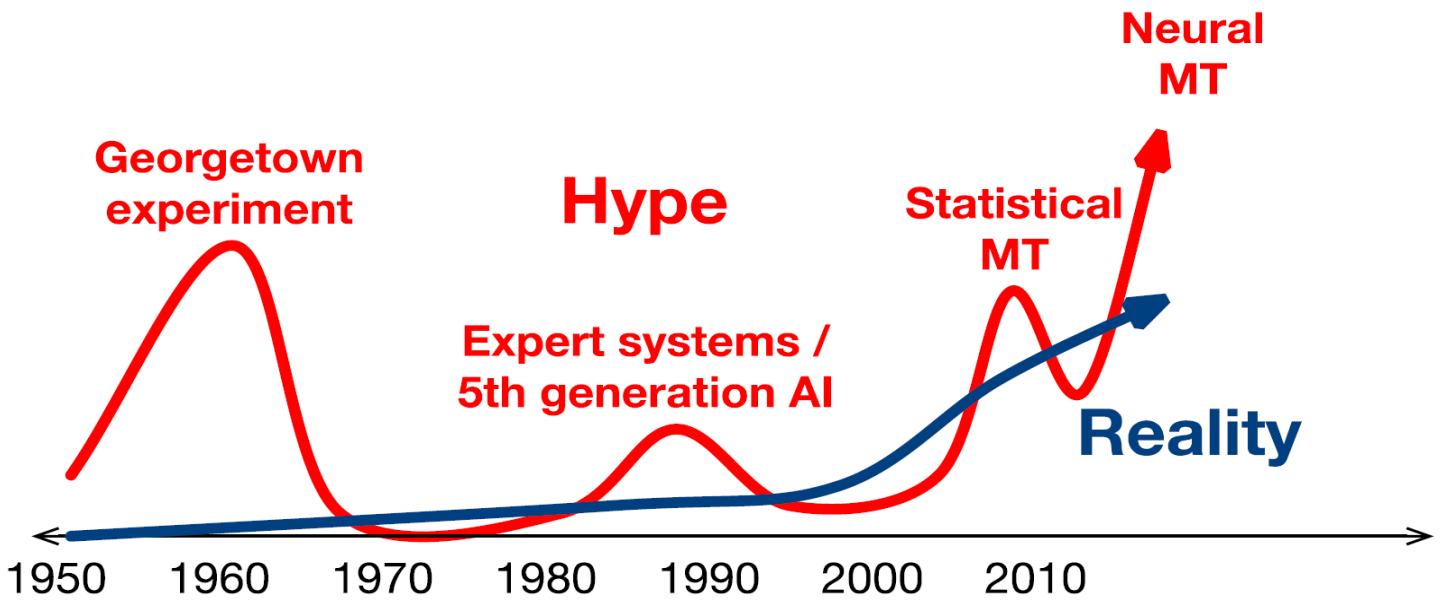
- Faster and better machine translation
- Support for all EU/ EEA languages (24 + Norwegian and Icelandic)
- Neural machine translation
- Running on a secured cloud
- Translating full documents or snippets
- Additional services to support multilinguality

NMT within CEF eTranslation

- Exploration of NMT since 2016
- László Tihanyi and Csaba Oravecz presented first results at the Conference on Hungarian Computational Linguistics in Jan. 2017*
- Prototype EN→HU running from early 2017 to mid 2018
- NMT incorporated into eTranslation in November 2017 (6 LPs)
- Additional LPs and new versions in various releases since December 2017, covering 2*23 LPs (EN ←→ X)
- Starting with Nematus, various toolkits have been explored, currently in use: Marian, but evaluating many others
- Users within DGT tend to prefer NMT over SMT (strongly in some languages), but it has to be used with care!
- MT volume is growing steadily, sometimes over 1M pages/week

* They received the best paper award

Hype and Reality



Omniscien™ TECHNOLOGIES

Thanks to Philipp Koehn for this nice summary

SMT vs. NMT: Some Important Differences

Statistical MT	Neural MT
Reuses translations of word groups	Reconstructs words from simulated "neural activations"
Can handle very large vocabularies, but no complex linguistic constructions	Limited vocabulary, but better in handling complex sentence structures
Does not generalise from an observation to "similar" cases	Generalisation is possible, but somewhat hard to control
Good in adequacy, not so good in fluency	Much better fluency, but problems with adequacy, e.g. omissions, distortions, inventions. Fluency make them hard to detect
Modular: models focusing on certain aspects can be improved separately	Holistic: More difficult to modify system behaviour
Incorporating new data via incremental training is complicated	Re-training with new data can facilitate updates and adaptation

Building Neural MT Engines

General approach:

- Focus on language pairs where NMT gives biggest quality boost
- Use existing open source toolkits like NEMATUS, OpenNMT, Marian
- Adapt them to our needs and embed them into existing infrastructure
- Try to stay close to developments in the research community

Engines delivered so far:

- **EN ↔ DE, HU; EN → ET, FI** *(15 November 2017)*
- **ET → EN, FI → EN** *(12 December 2017)*
- **EN → LT, LV** *(27 February 2018)*
- **EN → CS, GA, PL** *(14 March 2018)*
- **EN → BG, SK, SL; GA, LV, LT → EN** *(4 April 2018)*
- **EN → HR** *(25 April 2018)*
- **FR ↔ EN; EN → DA, ET (rebuild), FI (rebuild), HR (rebuild), NL, SV; BG, CS, DE (rebuild), ET (rebuild), FI (rebuild), HR (rebuild), PL, SK, SL → EN** *(22 June 2016)*
- **EN ↔ EL, ES, IT, MT, PT, RO ; DA, NL, SV → EN** *(5 July 2018)*

Data sets from Euramis

(Still excluding data from some EU institutions, all numbers in million pairs of segments EN ↔ X)

1.7 M	GA
6.9 M	HR
12.5 M	LV
13.0…16.1 M	SL,RO,CS,MT,HU,ET,LT,SK
	BG,DA,PL,NL,EL,FI,IT,SV
16.6 M	ES
18.3 M	PT
20.1 M	DE
25.7 M	FR

Challenges:

Using NMT in large scale poses interesting challenges related to MT quality:

- Measuring MT quality in a meaningful way is very hard, so numerical scores like BLEU and TER are routinely used as replacements ("understudy").

But:

- These scores are of limited help when comparing between different types of MT systems, trying to make fine-grained distinctions, or assessing fluency and adequacy separately
- Human evaluation of MT quality is very expensive and cannot easily be used as part of the development process

- NMT output looks very fluent , sometimes deceptively so
  - ➔ **Mistranslations are harder to find for translators**
  - ➔ **End users may get the wrong message and believe it**

- NMT adapts better to complex patterns in training data, but also adapts to errors
  - ➔ **NMT requires extra care in cleaning the training data**

Challenges in Data Curation:

- **Even EU translation memory contains some "noise": EN texts by non-native speakers, mixed languages, … How to find and cope with it?**

DE	EN
Anfangsrelativpermeabilität (initial relative permeability) größer/gleich 120 000 und Dicke kleiner/gleich 0,05 mm;	Initial relative permeability of 120 000 or more and a thickness of 0,05 mm or less;
PND Particle Number Diluter (Partikelanzahlverdünner)	PND Particle number diluter

- **ELRC is collecting parallel data from national authorities, but small sizes and need for quality control pose challenges**
- **We need to extend coverage to broader domains, started to use OPUS data with some success, but it may introduce inadequacies**
- **Results from ParaCrawl etc. will make things more challenging**
- **Automatic detection of non-parallel parts in bitexts can make human effort for data cleaning much more effective**

Ideas/Questions on Data Curation and beyond

- A lot of EU data is multilingual, some parts are 24-way parallel
- If two languages partially disagree, a third or further language(s) may help to isolate the mismatch, remove noise in a more surgical manner, distinguishing senses, …
- N-way alignment is more challenging, but potentially also more promising
- There is already interesting work on cross-lingual modeling (zero-shot MT, multilingual word embeddings, …)
- Mappings from many languages to joint representation spaces may be a good basis for tools addressing multiple purpose:
  - Finding and cleaning parallel data
  - Selecting task-specific subcorpora
  - Finding errors in MT output (aka confidence estimation)
- The next WS in this series is supposed to address these and related questions

Questions?

andreas.eisele@ec.europa.eu
DGT-MT@ec.europa.eu