

# 02 – Machine translation

## IA161 Natural Language Processing in Practice

P. Rychlý

NLP Centre, FI MU, Brno

October 21, 2022

- 1 Introduction
- 2 Neural Machine Translation
- 3 Machine translation evaluation

# Translation: English→Czech

Moses is an implementation of the statistical (or data-driven) approach to machine translation (MT). This is the dominant approach in the field at the moment, and is employed by the online translation systems deployed by the likes of Google and Microsoft.

- 1 Mojžíš je implementace statistické (nebo řízené daty) přístupu k strojovému překladu (MT). To je převládajícím přístupem v oblasti v současné době, a je zaměstnán pro on-line překladatelských systémů nasazených likes Google a Microsoft.
- 2 Moses je implementace statistického (nebo daty řízeného) přístupu k strojovému překladu (MT). V současné době jde o převažující přístup v rámci strojového překladu, který je použit online překladovými systémy nasazenými Googlem a Microsoftem.
- 3 Mojžíš je provádění statistické (nebo aktivovaný) přístup na strojový překlad (mt). To je dominantní přístup v oblasti v tuto chvíli, a zaměstnává on - line překlad systémů uskutečněné takové, Google a Microsoft.
- 4 Mojžíš je implementace statistického (nebo datově řízeného) přístupu k strojovému překladu (MT). To je v současné době dominantní přístup v oboru a je využíván online překladatelskými systémy, které používají společnosti Google a Microsoft.

# Statistical Machine Translation

- rule-based systems motivated by linguistics
- SMT inspired by information theory and statistics
- Google Translate (before 2016), Bing Translator, Moses
- **gisting**: the most frequent usage of MT on Internet
- in fact, MT output is always post-edited for final production

# Neural Machine Translation

- neural networks: boom in the last few years
- current state-of-the-art
- *all* research and production systems use NMT
- big improvements over SMT
- end-to-end systems, almost no knowledge about languages needed

# Machine translation: what is translated

- web pages
- technical manuals, how-tos
- scientific documents, papers, articles
- leaflets, flyers, catalogues
- texts from limited domains in general
- Wikipedia articles (CS–SK)

# Machine translation nowadays

- intense collecting of data
- development of systems driven by evaluation metrics
- EU: 24 official languages (EuroMatrix)
- software companies focus on English as source language (i18n)
- large language pairs (En↔Sp, En↔Fr): fairly high-quality translation
- Google Translate as a base standard  
(there are better systems in specialized domains)
- morphologically rich languages: worse results
- En-\* and \*-En pairs prevail

## Data: parallel corpora

- Europarl: a collection of texts from the European Parliament [Koehn, 2005]
- OPUS: parallel texts of various source, one of the biggest resources [Tiedemann and Nygaard, 2004]
- Acquis Communautaire: EU laws [Steinberger et al., 2006]
- EUR-Lex: access to European Union law
- DGT translation memory [Steinberger et al., 2013], MyMemory
- freely available corpora are usually of size of 10–100 million words
- multilingual webpages (Wikipedia)
- comparable corpora: texts from the same domain

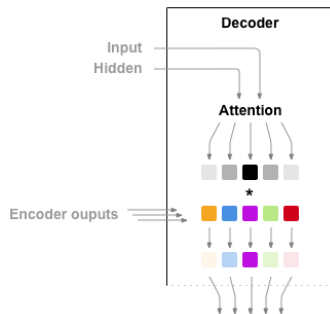
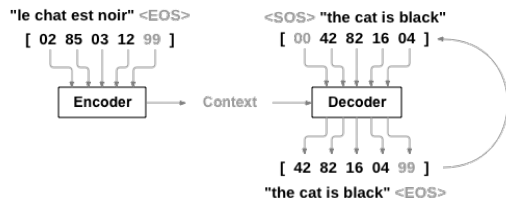


# Sentence alignment

- crucial part of training data
- 1:1, 1:0, 0:1, 1:2, 2:1, ... alignments
- Gale-Church (sentence lengths)
- Hunalign (with a dictionary, G-Ch is a fallback)
- BLEUalign (MT-based sentence alignment) [Sennrich and Volk, 2011]
- Vecalign (multilingual sentence embeddings)  
[Thompson and Koehn, 2019]

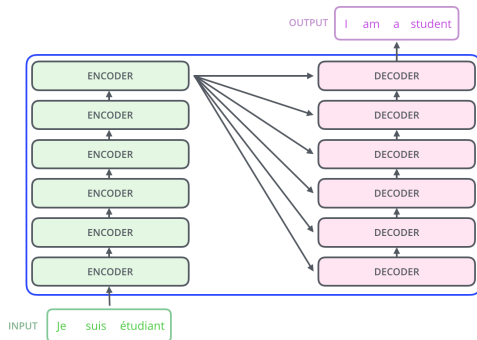
# Neural Machine Translation

- deep neural nets
- encoder-decoder architecture



# Neural Machine Translation

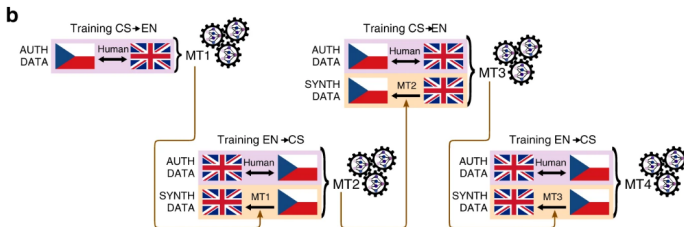
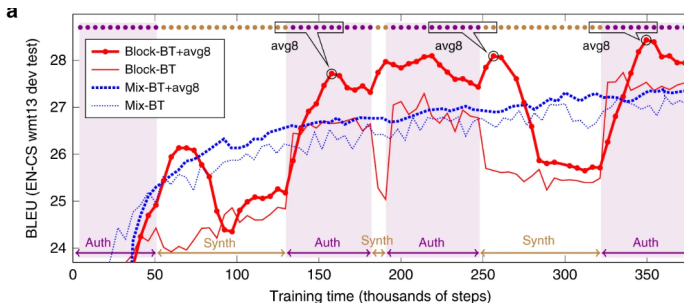
- current trend: transformers



<http://jalammar.github.io/illustrated-transformer/>

# Neural Machine Translation

- CUBBITT system for EN to CS
- better than human in adequacy in certain circumstances



## Word Alignment Matrix

Could be generated from attentions in NMT, useful for phrase extraction.

	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael	█									
assumes		█	█	█						
that						█				
he							█			
will										█
stay										█
in								█		
the								█		
house									█	

# Automatic evaluation of translation

- advantages: speed, price; disadvantages: do we measure quality of translation?
- gold standard: manually prepared reference translations
- candidate  $c$  is compared with  $n$  reference translations  $r_i$
- various approaches: n-gram agreement between  $c$  and  $r_i$ , edit distance, ...
- BLEU: the most widely used [Papineni et al., 2002]
- METEOR: correlates best with human evaluation [Banerjee and Lavie, 2005]

# BLEU

- the most popular (a standard), the most widely used, the oldest (2001)
- IBM, Papineni [Papineni et al., 2002]
- n-gram agreement between references and candidates
- precision for 1–4-grams
- brevity penalty

$$\text{BLEU} = \min \left( 1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$





# BLEU computing: sacreBLEU

- BLEU depends on tokenization  
different tokenization = different results
- sacreBLEU[Post, 2018]: WMT standard tokenization
- automatically downloads common WMT test

## Example (Python usage)

```
from sacrebleu.metrics import BLEU
refs = [['The dog bit the man.', ...], ...]
sys = ['The dog bit the men.', ...]
bleu = BLEU()
bleu.corpus_score(sys, refs)
```

# Translation quality according to language pairs

2015 vs 2019

		output language					
		Czech	German	English	Finnish	French	Russian
input language	Czech			26.2			
	German			29.3			
	English	18.8	24.9		15.5	33.6	24.3
	Finnish			19.7			
	French			33.1			
	Russian			27.9			
	Chinese						

		output language										
		Czech	German	English	Finnish	French	Gujarati	Kazakh	Lithuanian	Russian	Chinese	
input language	Czech	19.3										
	German	20.1	42.8		37.3							
	English	29.9	44.9	27.4		28.2	11.1	20.1	36.3	44.6		
	Finnish			33.0								
	French		35.0									
	Gujarati			24.9								
	Kazakh			30.5								
	Lithuanian			36.3								
	Russian			40.2								
	Chinese			39.9								

<http://matrix.statmt.org/> [Koehn, 2007]

# References I



Banerjee, S. and Lavie, A. (2005).

Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.

*In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.



Koehn, P. (2005).

Europarl: A parallel corpus for statistical machine translation.

*In MT summit*, volume 5, pages 79–86. Citeseer.



Koehn, P. (2007).

Euromatrix–machine translation for all european languages.

*Invited Talk at MT Summit XI*, pages 10–14.

## References II



Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).  
Bleu: a method for automatic evaluation of machine translation.  
In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.






Post, M. (2018).  
A call for clarity in reporting BLEU scores.  
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.



Sennrich, R. and Volk, M. (2011).  
Iterative, MT-based sentence alignment of parallel texts.  
In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182.

## References III

-  Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2013).  
Dgt-tm: A freely available translation memory in 22 languages.  
*arXiv preprint arXiv:1309.5226.*
-  Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006).  
The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.  
*arXiv preprint cs/0609058.*
-  Thompson, B. and Koehn, P. (2019).  
Vecalign: Improved sentence alignment in linear time and space.  
*In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages*

## References IV

1342–1348, Hong Kong, China. Association for Computational Linguistics.



Tiedemann, J. and Nygaard, L. (2004).

The OPUS corpus-parallel and free: <http://logos.uio.no/opus>.  
In *LREC*. Citeseer.