03 – Machine translation

IA161 Advanced Techniques of Natural Language Processing

V. Baisa

NLP Centre, FI MU, Brno

October 2, 2017

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 strojového 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. Moses je implementace statistického (nebo řízeného daty) přístupu k strojového překladu (MT). To je převládajícím přístupem v oblasti v současné době, a je využit pro on-line překladové systémy nasazené v Google a Microsoft.

Statistical Machine Translation

- rule-based systems motivated by linguistics
- SMT inspired by information theory and statistics
- millions of pages translated by SMT on daily basis
- Google Translate, Bing Translator, Moses
- **gisting**: the most frequent usage of MT on Internet
- in fact, MT output is always post-edited
- neural networks: boom in the last few years

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)

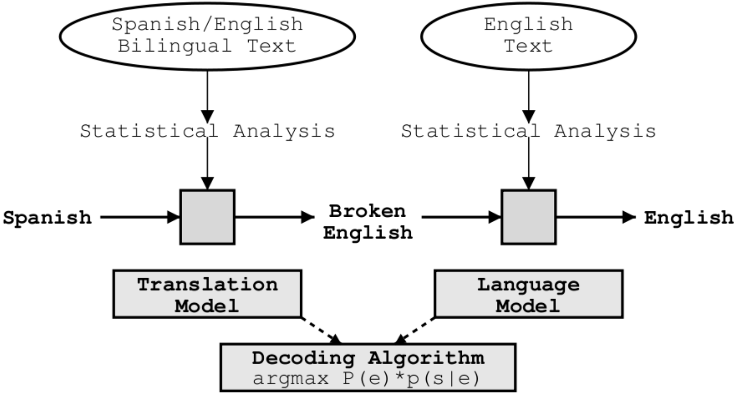
Machine translation nowadays

- intense collecting of data
- development of systems driven by evaluation metrics
- the West: English as target language
- EU: 24 official languages (EuroMatrix)
- software companies focuses on English as source language
- large language pairs (En↔Sp, En↔Fr): fairly high-quality translation
- Google Translate as gold standard
- morphologically rich languages: worse results
- En-* and *-En pairs prevail
- Moses: freely available statistical machine translation [Koehn, 2007]

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)
- DGT translation memory [Steinberger et al., 2013], MyMemory
- InterCorp – manually aligned fiction books (ČNK, FFUK)
- freely available corpora are usually of size of 10–100 million words
- multilingual webpages (Wikipedia)
- comparable corpora: texts from the same domain

Schema of SMT



SMT – noisy channel

Developed by Shannon (1948) [Shannon, 1956] for self-correcting codes, for corrections of coded signals transferred through noisy channels based on information about a source message and types of errors occurring in the channels.

Another application: OCR, Optical Character Recognition. It is messy, but we can estimate what was in the source text from a language model and frequent errors: `l-1-I`, `rn-m` etc.

$$
\begin{aligned}
e^* &= \arg\max_e p(e|f) \\
&= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\
&= \arg\max_e p(e)p(f|e).
\end{aligned}
$$

We will speak about language models later.

Lexical translation

Standard translation dictionary does not contain translation probabilities for word meanings.

key $\rightarrow$ *klíč, tónina, klávesa*

How often are the individual equivalents used?

key $\rightarrow$ *klíč* (0.7), *tónina* (0.18), *klávesa* (0.08), ...

We need a lexical probability distribution p_f with the property:

$$\sum_e p_f(e) = 1$$

$$\forall e : 0 \leq p_f(e) \leq 1$$

$$p_{\text{klíč}}(\text{key}) \ ? \ p_{\text{mrkev}}(\text{carrot})$$

Computing the translation probability

We need to know the value of the translation probability for all words in a sentence.

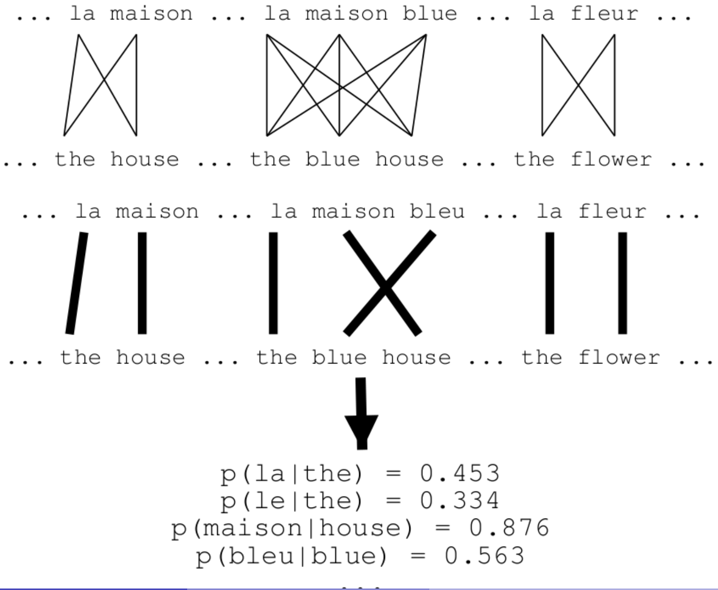
Parallel corpora with aligned sentences are used for this. The alignment is usually on document level, so **sentence-alignment** is needed.

These properties are exploited usually:

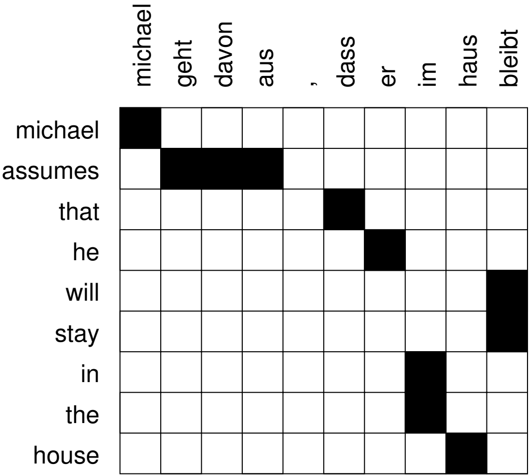
- sentence length comparison,
- translation dictionary or
- cooccurrence of names, numbers, characters and low frequency words.

Word alignment

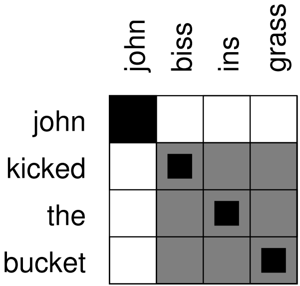
GIZA++ is the most widely used tool. [Och and Ney, 2003]

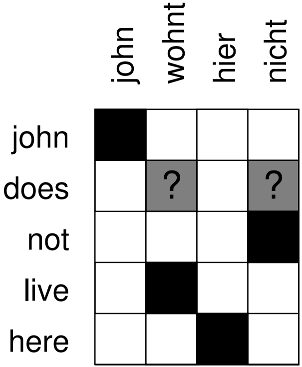


Word Alignment Matrix



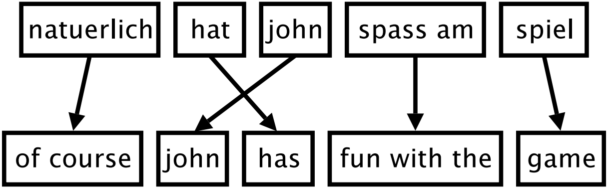
Lexical translation problems





Phrase-based translation model

State-of-the-art of SMT. Not only words, but whole phrases are translated at a time. [Koehn et al., 2003] [Chiang, 2005]

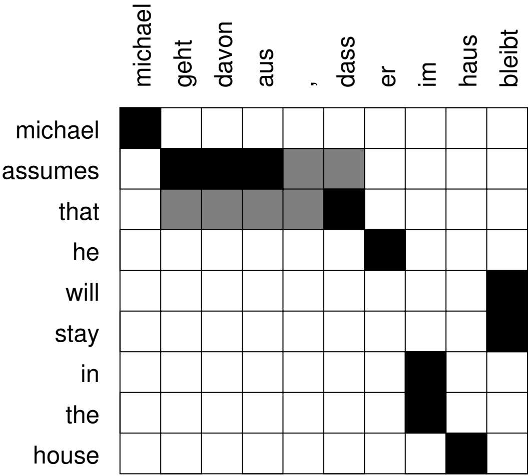


Phrases are not linguistically motivated. German *am* is usually not translated by one word *with*. Statistically significant context *spass am* helps with a proper translation. Common phrases would be segmented in a different way: *(fun (with (the game)))*.

Advantages of PBMT

- we often translate $n : m$ words, a word is unsuitable element
- the translation of groups of words helps with translation ambiguity
- and also fluency
- systems can learn longer phrases, ad infinitum, if data is available
- the model is simpler: fertility, NULL tokens are not needed

Phrase extraction



Automatic evaluation of translation

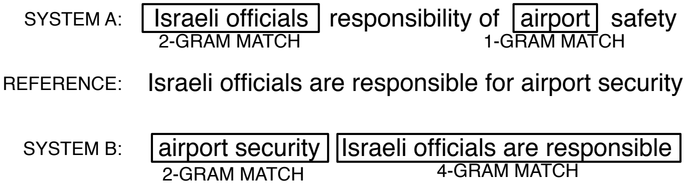
- advantages: speed, price; disadvantages: do we measure quality of translation?
- gold standard: manually prepared reference translations
- candidate c is compared with n reference translations r_i
- various approaches: n-gram agreement between c and r_i, edit distance, ...
- BLEU: the most widely used [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

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

BLEU – an example



metrics	system A	system B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0 %	52 %

Translation quality according to language pairs

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

`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.

Chiang, D. (2005).
A hierarchical phrase-based model for statistical machine translation.
In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Koehn, P. (2005).
Europarl: A parallel corpus for statistical machine translation.
In *MT summit*, volume 5, pages 79–86. Citeseer.

References II

Koehn, P. (2007).
Euromatrix–machine translation for all european languages.
Invited Talk at MT Summit XI, pages 10–14.

Koehn, P., Och, F. J., and Marcu, D. (2003).
Statistical phrase-based translation.
In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003).
A systematic comparison of various statistical alignment models.
Computational Linguistics, 29(1):19–51.

References III

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.

Shannon, C. E. (1956).
The zero error capacity of a noisy channel.
Information Theory, IRE Transactions on, 2(3):8–19.

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.

References IV

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.

Tiedemann, J. and Nygaard, L. (2004).
The opus corpus-parallel and free: http://logos. uio. no/opus.
In *LREC*. Citeseer.