Deep Learning for Natural Language Processing

Introduction to Machine Translation

UNIVERSITY OF GOTHENBURG

CHALMERS

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

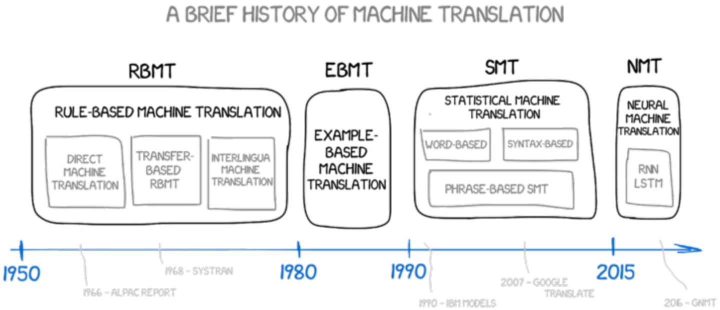
Richard Johansson

richard.johansson@gu.se

introduction to MT

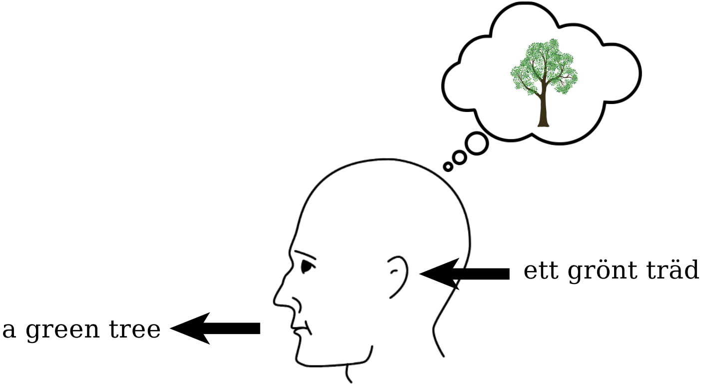
- goal: a computer program that translates a text in one language (**the source**) into another language (**the target**).
- this is one of the most high-profile areas of NLP, and perhaps its most classical problem (Weaver, 1949)

  When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."


A BRIEF HISTORY OF MACHINE TRANSLATION
RBMT
RULE-BASED MACHINE TRANSLATION
DIRECT MACHINE TRANSLATION
TRANSFER-BASED RBMT
INTERLINGUA MACHINE TRANSLATION
EBMT
EXAMPLE-BASED MACHINE TRANSLATION
SMT
STATISTICAL MACHINE TRANSLATION
WORD-BASED
SYNTAX-BASED
PHRASE-BASED SMT
NMT
NEURAL MACHINE TRANSLATION
RNN
LSTM
1950
1966 - ALPAC REPORT
1968 - SYSTRAN
1980
1990
1990 - IBM MODELS
2007 - GOOGLE TRANSLATE
2015
2016 - GNMT

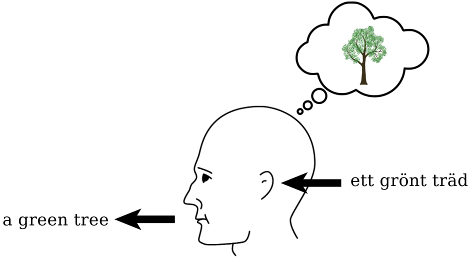
[source]

idealized intuition of the translation process



interlingua-based translation

- can we implement a system based on our intuition?

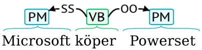


- idea:
  - map the source-language sentence into some "meaning representation" or **interlingua**
  - then convert the representation into the target language

example: interlingua-based translation

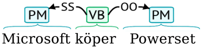
Microsoft köper Powerset

example: interlingua-based translation



example: interlingua-based translation

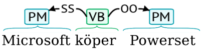
[Microsoft]$_{\texttt{Buyer}}$ köper [Powerset]$_{\texttt{Goods}}$



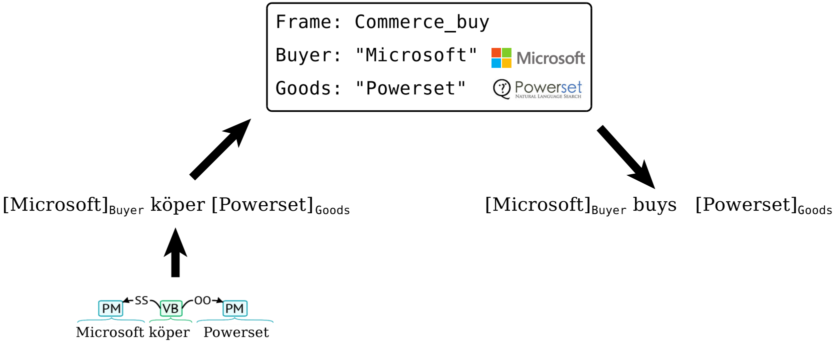
example: interlingua-based translation

```
Frame: Commerce_buy
Buyer: "Microsoft"
Goods: "Powerset"
```

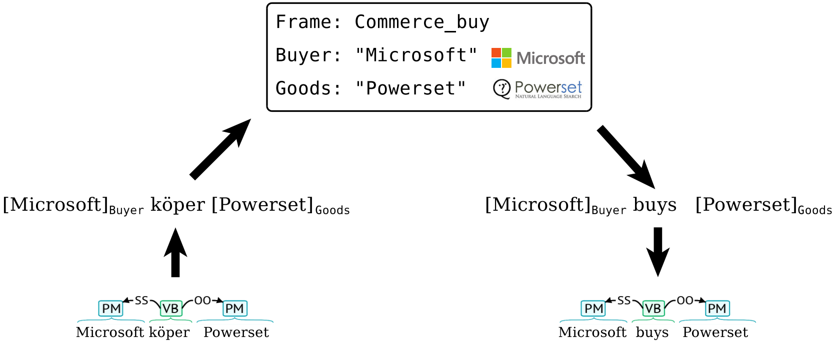
[Microsoft]$_{\texttt{Buyer}}$ köper [Powerset]$_{\texttt{Goods}}$



example: interlingua-based translation



example: interlingua-based translation



data-driven machine translations systems

- instead of writing rules, since the early 1990s, most MT systems are **data-driven**: they are trained on example texts
  - **statistical** MT systems: word-based and phrase-based
  - **neural** MT systems

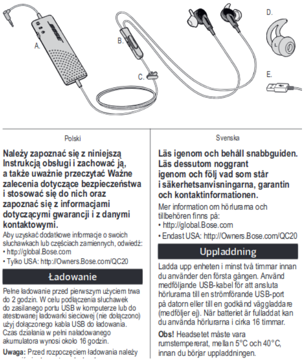
data-driven machine translations systems

- instead of writing rules, since the early 1990s, most MT systems are **data-driven**: they are trained on example texts
  - **statistical** MT systems: word-based and phrase-based
  - **neural** MT systems
- data-driven MT systems are trained on **parallel** text, typically aligned at the sentence level:

  EN: *I should like to know a little more about that.*

  SV: *Jag skulle gärna få det förklarat litet närmare.*

parallel text



Polski

Należy zapoznać się z niniejszą Instrukcją obsługi i zachować ją, a także uważnie przeczytać Ważne zalecenia dotyczące bezpieczeństwa i stosować się do nich oraz zapoznać się z informacjami dotyczącymi gwarancji i z danymi kontaktowymi.

Aby uzyskać dodatkowe informacje o swoich słuchawkach lub częściach zamiennych, odwiedź:

- http://global.Bose.com
- Tylko USA: http://Owners.Bose.com/QC20

Ładowanie

Pełne ładowanie przed pierwszym użyciem trwa do 2 godzin. W celu podłączenia słuchawek do zasilanego portu USB w komputerze lub do atestowanej ładowarki sieciowej (nie dołączono) użyj dołączonego kabla USB do ładowania. Czas działania w pełni naładowanego akumulatora wynosi około 16 godzin.

Uwaga: Przed rozpoczęciem ładowania należy

Svenska

Läs igenom och behåll snabbguiden. Läs dessutom noggrant igenom och följ vad som står i säkerhetsanvisningarna, garantin och kontaktinformationen.

Mer information om hörlurarna och tillbehören finns på:

- http://global.Bose.com
- Endast USA: http://Owners.Bose.com/QC20

Uppladdning

Ladda upp enheten i minst två timmar innan du använder den första gången. Använd medföljande USB-kabel för att ansluta hörlurarna till en strömförande USB-port på datorn eller till en godkänd väggladdare (medföljer ej). När batteriet är fulladdat kan du använda hörlurarna i cirka 16 timmar.

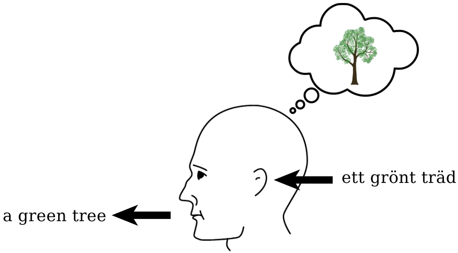
Obs! Headsetet måste vara rumstempererat, mellan 5°C och 40°C, innan du börjar uppladdningen.

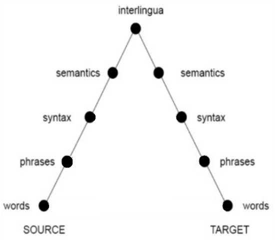
examples of sentence-aligned parallel text data

- first well-known **parallel dataset**: Canadian Hansards, English–French
- **Europarl**, http://www.statmt.org/europarl
- **Opus** http://opus.nlpl.eu/
- the Bible (largest number of languages?), Quran etc

fundamental idea in neural MT

- the architecture used in most neural MT systems:
  - **encoder**: "summarize" the information in the source sentence
  - **decoder**: based on the encoding, generate the target-language output in a step-by-step fashion





Cho's model (Cho et al., 2014)

- the encoder and decoder are both GRUs
- the final state of the encoder is used as the "summary" c
- this summary is accessed by all steps in the decoder

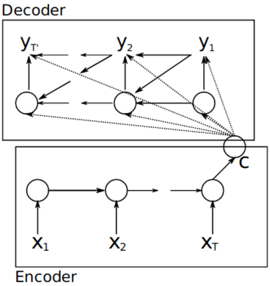
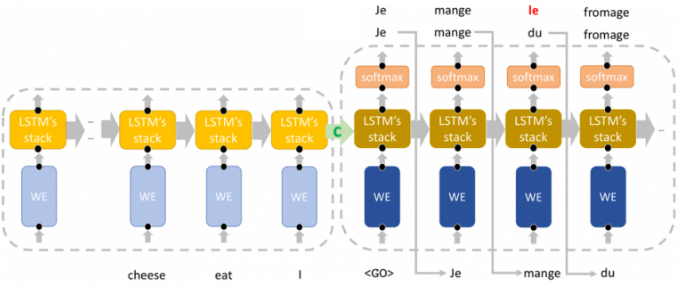


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

Sutskever's model (Sutskever et al., 2014)

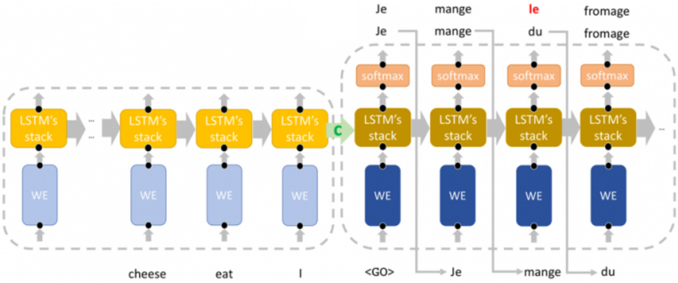
- the encoder and decoder are multilayered LSTMs
- the final state of the encoder becomes the initial state of the decoder
- to make this work, they had to reverse the source sentence...



[source]

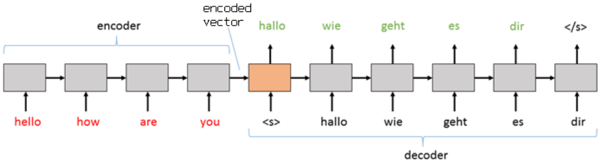
training seq2seq models

- for each decoding step, we compute a softmax over the whole target-language vocabulary
  - and then a cross-entropy loss as usual
  - we're minimizing the word-by-word loss, not maximizing BLEU
- each decoding step uses the output from the previous step
  - during training, we use the **gold-standard output**
  - this is an example of **teacher forcing** that we saw last time



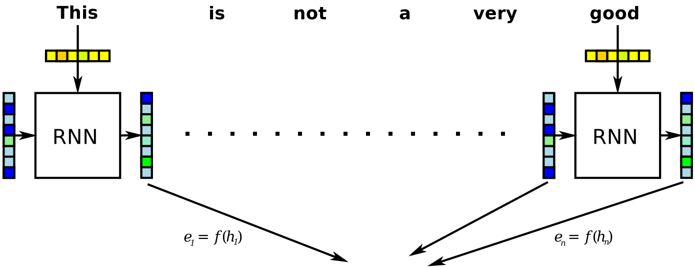
drawbacks of simple seq2seq

- everything that is needed for all the steps of decoding needs to be crammed into a fixed-size vector
- information needs to “flow” through many RNN steps: difficult for long sentences



attention: recap

- first, compute an "importance score" e_i for each state $\boldsymbol{h}_i$



- for the attention weights, we apply the softmax:

$$\alpha_i = \frac{\exp e_i}{\sum_{j=1}^{n} \exp e_j}$$

- finally, the "summary" is computed as a weighted sum

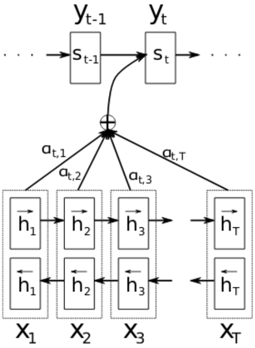
$$s = \sum_{i=1}^{n} \alpha_i \boldsymbol{h}_i$$

attention models in machine translation

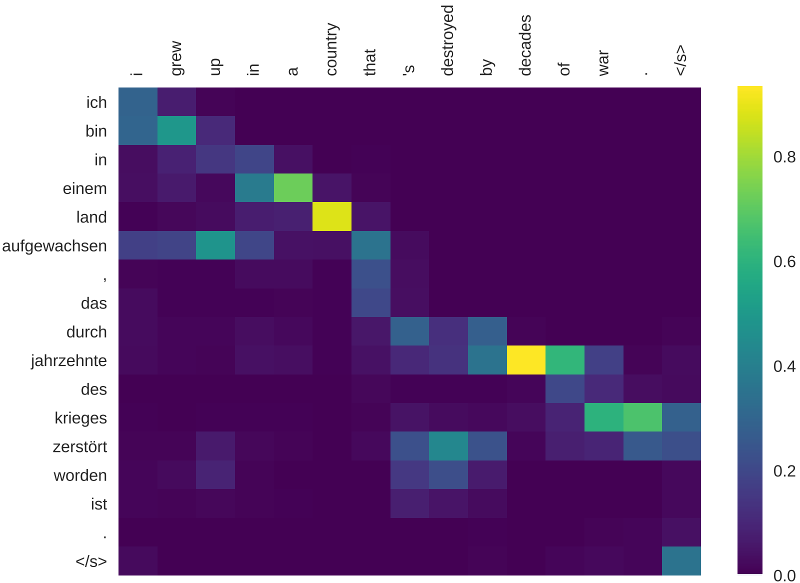
- Bahdanau et al. (2015) proposed **attention** for MT
- their attention model is a straightforward MLP that uses the **previous decoder state**:

$$e_i = f(\boldsymbol{h}_i, \boldsymbol{s}_{t-1})$$

- intuition: the attention mechanisms can decide what is most important **right now**
- the survey by Galassi et al. (2019) gives an overview of implementations of attention



visualizing attention



next up

- **exercise** (Monday): seq2seq with attention
- next MT **lecture**: more advanced techniques in MT

references

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*.

A. Galassi, M. Lippi, and P. Torroni. 2019. Attention, please! A critical review of neural attention models in natural language processing. arXiv:1902.02181.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

W. Weaver. 1949. Translation. In D.A. Booth, editor, *Machine Translation of Languages*, MIT Press.