Machine Translation

IN4080

Natural Language Processing

Yves Scherrer

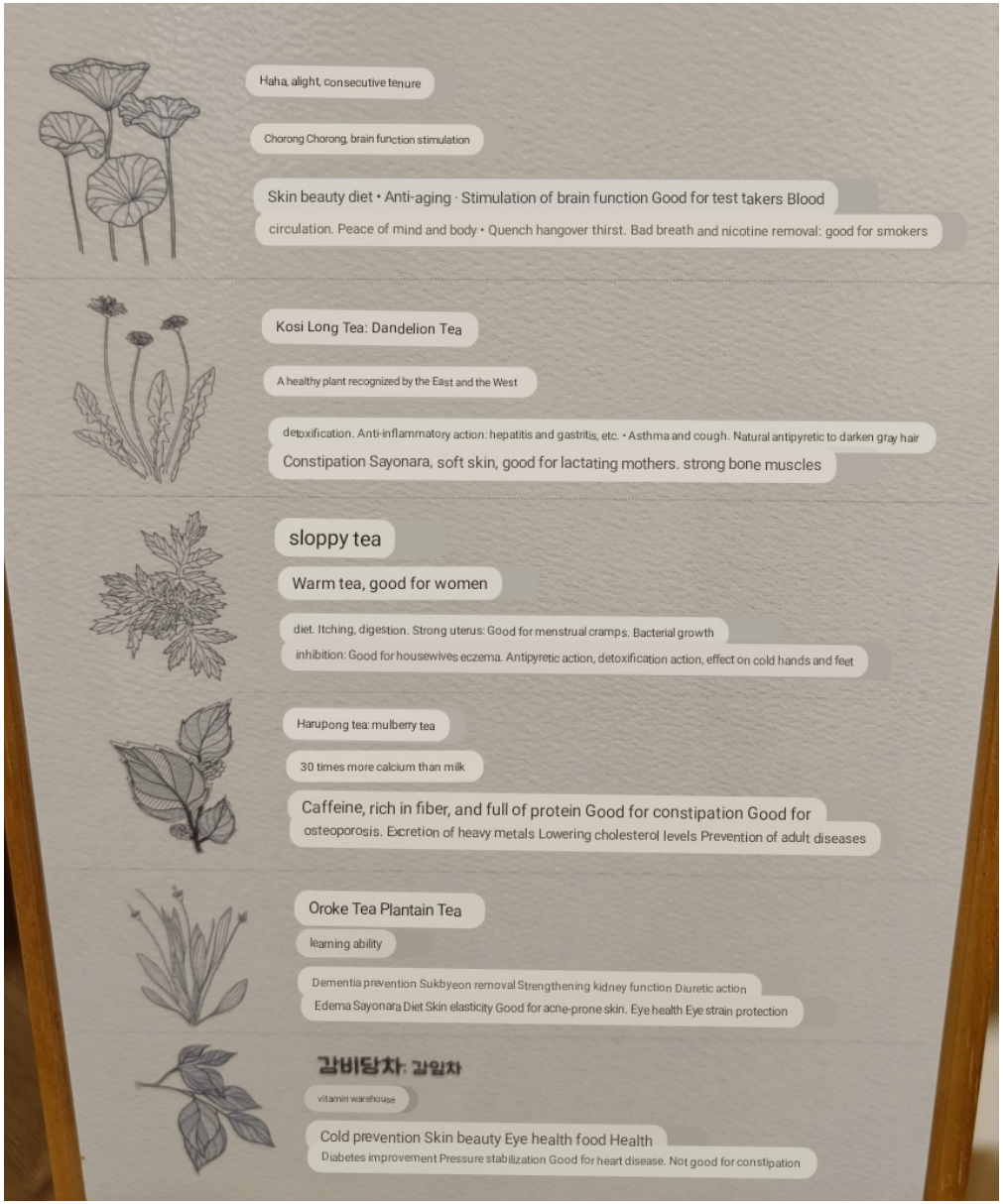
Motivation

Machine translation has made a lot of progress in recent years and has become truly useful in many situations and contexts.

Motivation

Machine translation has made a lot of progress in recent years and has become truly useful in many situations and contexts.

Google Lens, 1.11.2022



Machine translation

is a complex but natural task:

- Natural input, natural output
- Not fixed to particular linguistic theories or formalisms
- “The research field lacks ideological battles but is rather characterized by a friendly competitive spirit.” (Ph. Koehn)

is tied to intrinsic properties of language:

- What are the differences between languages?
- How can we preserve meaning when translating?

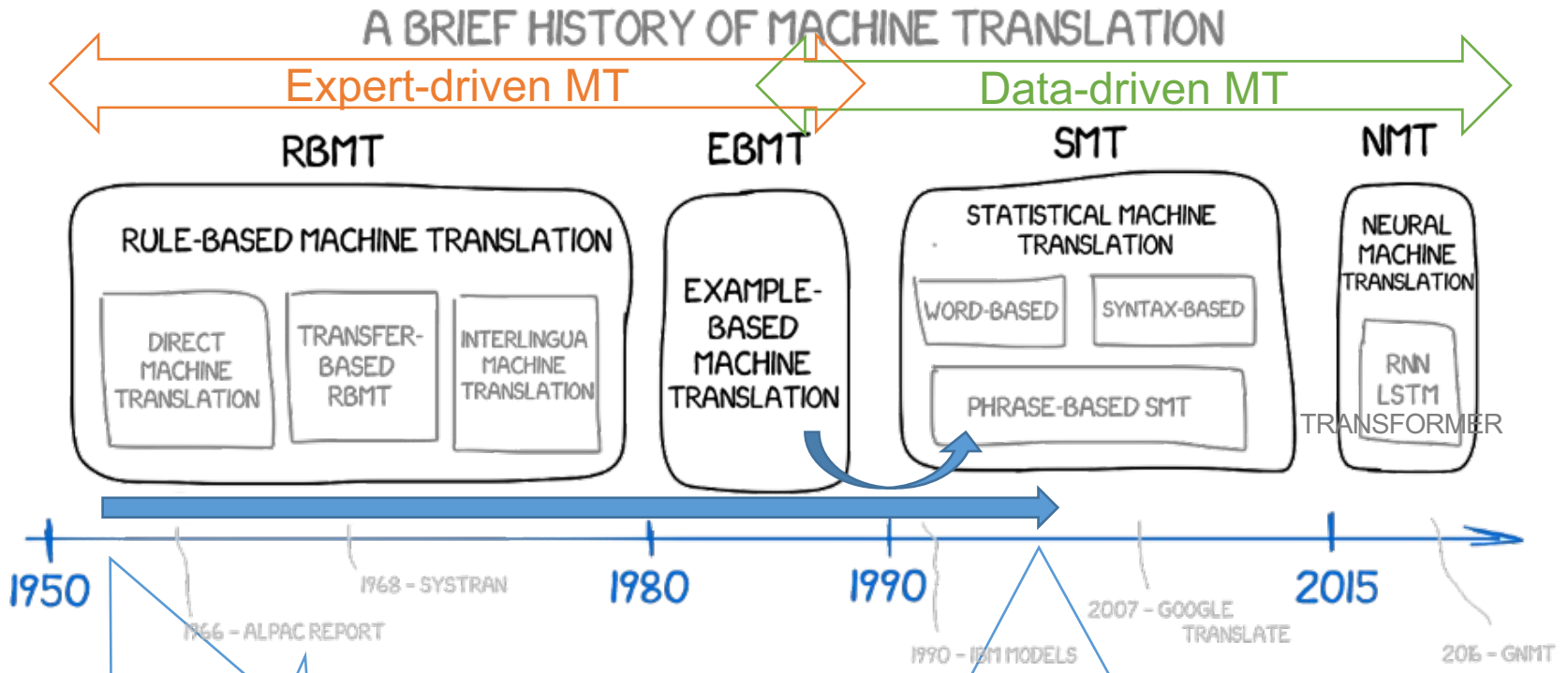
A bit of history

MT is an old idea...

In fact, one of the first language technology tasks.

Warren Weaver (1947):

- *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”.*
- Success of code-breaking in WWII
- Beginning of cold war



1954 – IBM-Georgetown experiment: a machine could translate 250 words and 6 grammar rules

1966 – ALPAC report:
MT is expensive, inaccurate, and unpromising

2002 – BLEU evaluation metric
2005 – Moses open-source system
2005 – Europarl corpus
2006 – First workshop on machine translation

A long history characterized by radical and disruptive methodological changes

https://vas3k.com/blog/machine_translation/

Huge interest (and pressure) from non-experts, unrealistic expectations

Facebook's AI Just Set A New Record In Translation And Why It Matters

Linguists, update your resumes because Baidu thinks it has cracked fast AI translation

Microsoft AI translates news as well as humans, takes on Google Translate

SDL Cracks Russian to English Neural Machine Translation

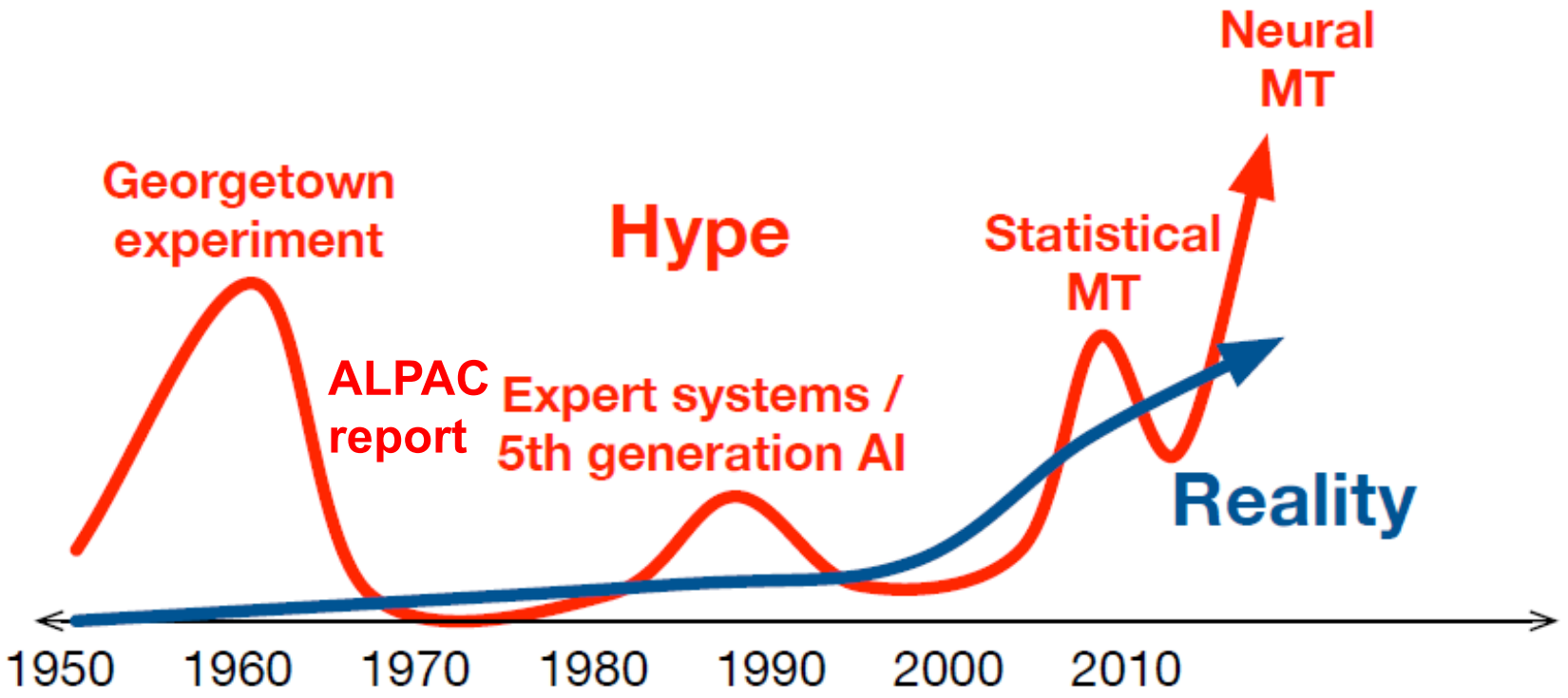
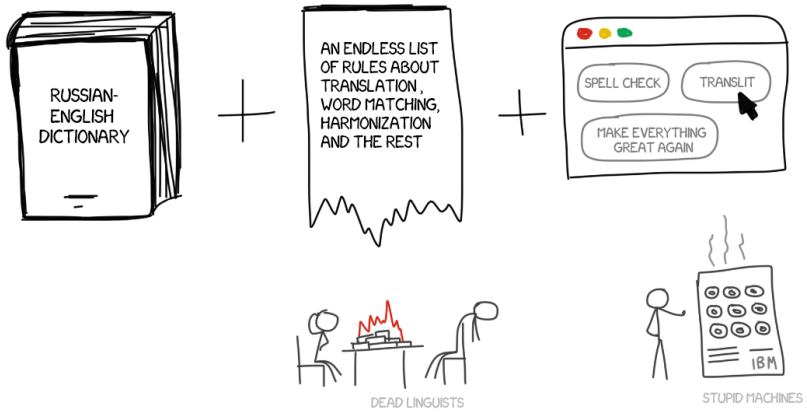


Figure: Philipp Koehn

Rule-based MT

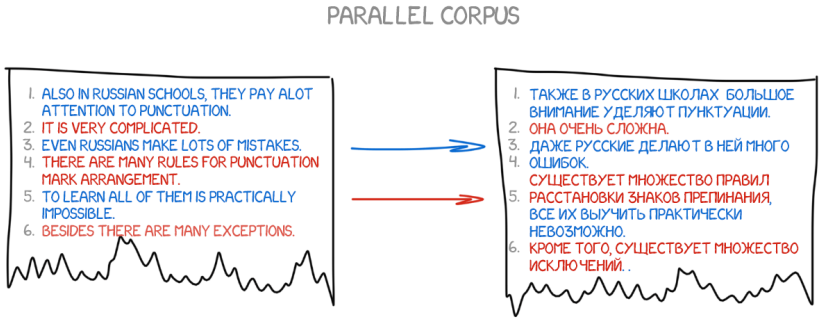
- Build dictionaries
- Write transformation rules
- Refine, refine, refine



Commercial applications:

- 1976: weather forecast translations French-English
- 1968: Systran

Statistical MT



1990: “IBM models”

- Idea: learn everything from a parallel corpus

Mid-2000s: phrase-based models

- A lesson from EBMT: Translating each word separately is harder than it needs to be
- Keep frequent word sequences (“phrases”) together and translate them as a whole

Late 2000s: syntax-based models

2010: Commercial viability (Google Translate...)

Neural MT

Late 2000s: successful use of neural models for computer vision

2012, 2013: first neural models for MT proposed

Since 2016: NMT is the new state of the art

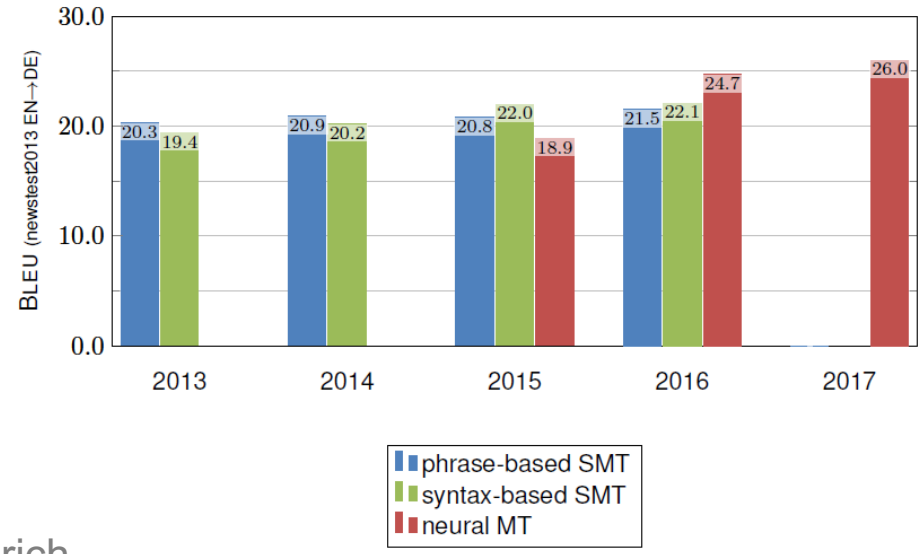
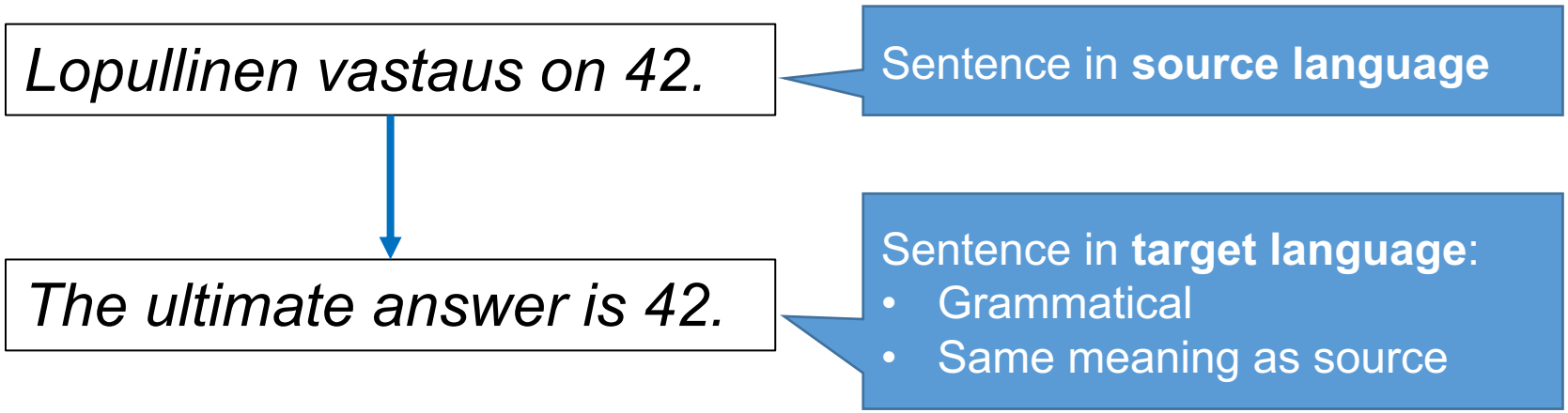


Figure: Rico Sennrich

*NMT 2015 from U. Montréal: https://sites.google.com/site/acl16nmt/

Data-driven MT – The task



Training data:

- A **parallel corpus** or **bitext**, i.e. a (rather large) set of sentence pairs with the same meaning.
  - Typically, one side is the original and the other side is produced by a human translator.
- Where can such data be found?

Data-driven MT – The data

Potential data sources:

- International political organizations, and national organizations of multilingual countries
  - European Parliament (and other EU institutions)
  - United Nations
- Movie subtitles
- Multilingual web sites

OPUS corpus collection: https://opus.nlpl.eu/

- Rule of thumb: at least 1 million sentence pairs

Model architectures

NMT model architecture

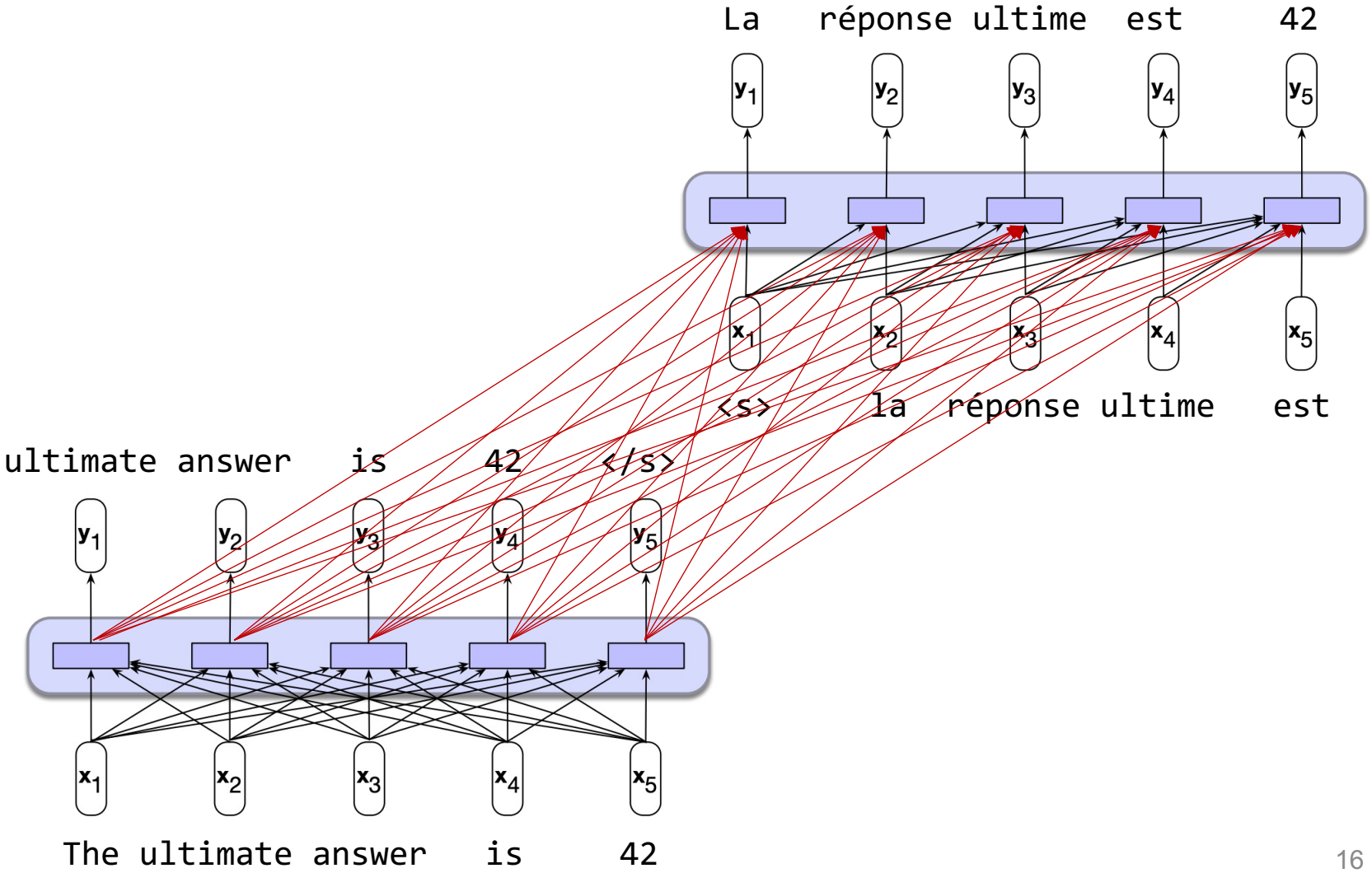
Encoder-decoder with attention:

Neural machine translation models

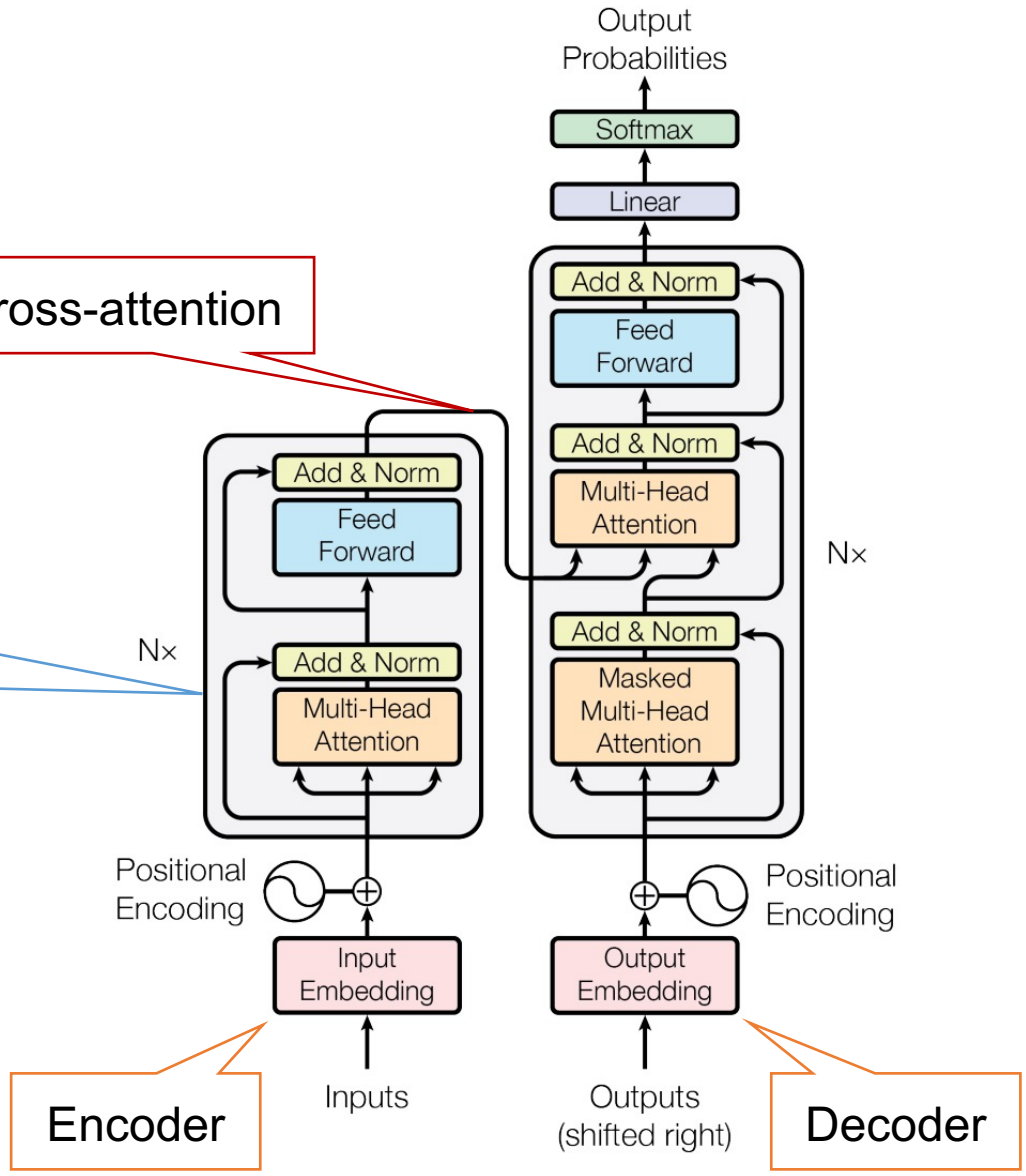
- encode the source sentence, then
- decode the target sentence by attending the most relevant source tokens.

Most popular architecture today: Transformer

Encoder-decoder model with attention



Transformer



https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Cross-attention weights

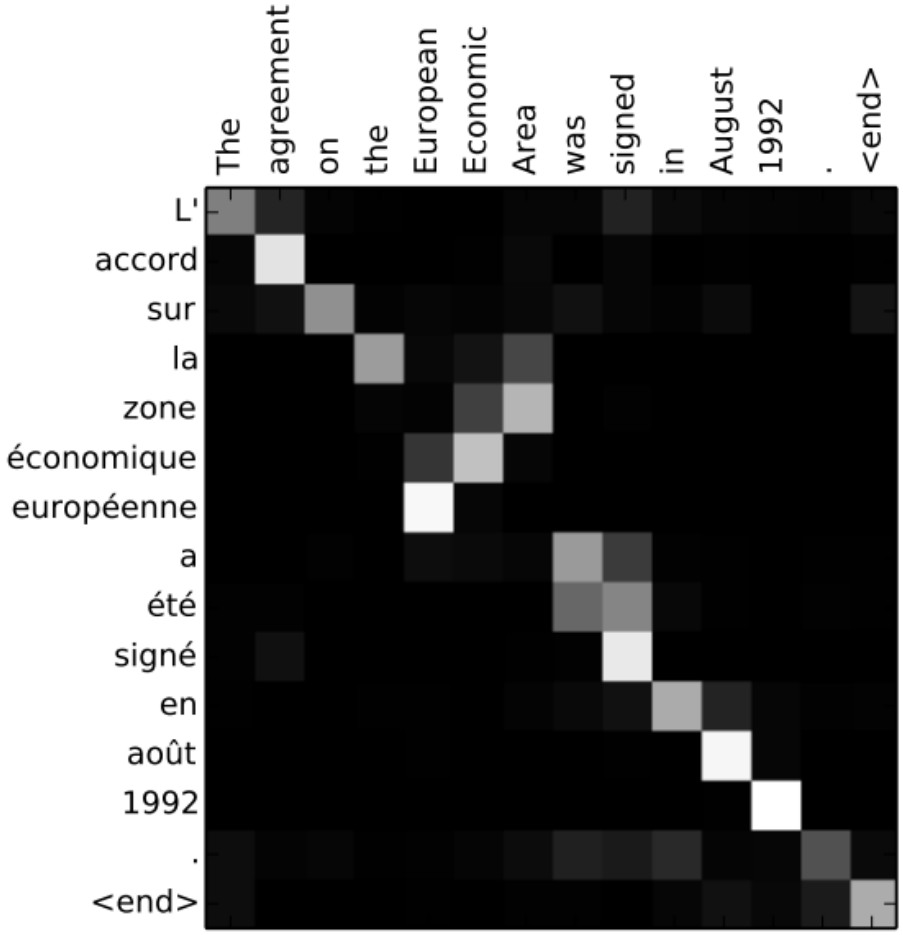


Figure: Bahdanau et al. (2014)

Experimental setup

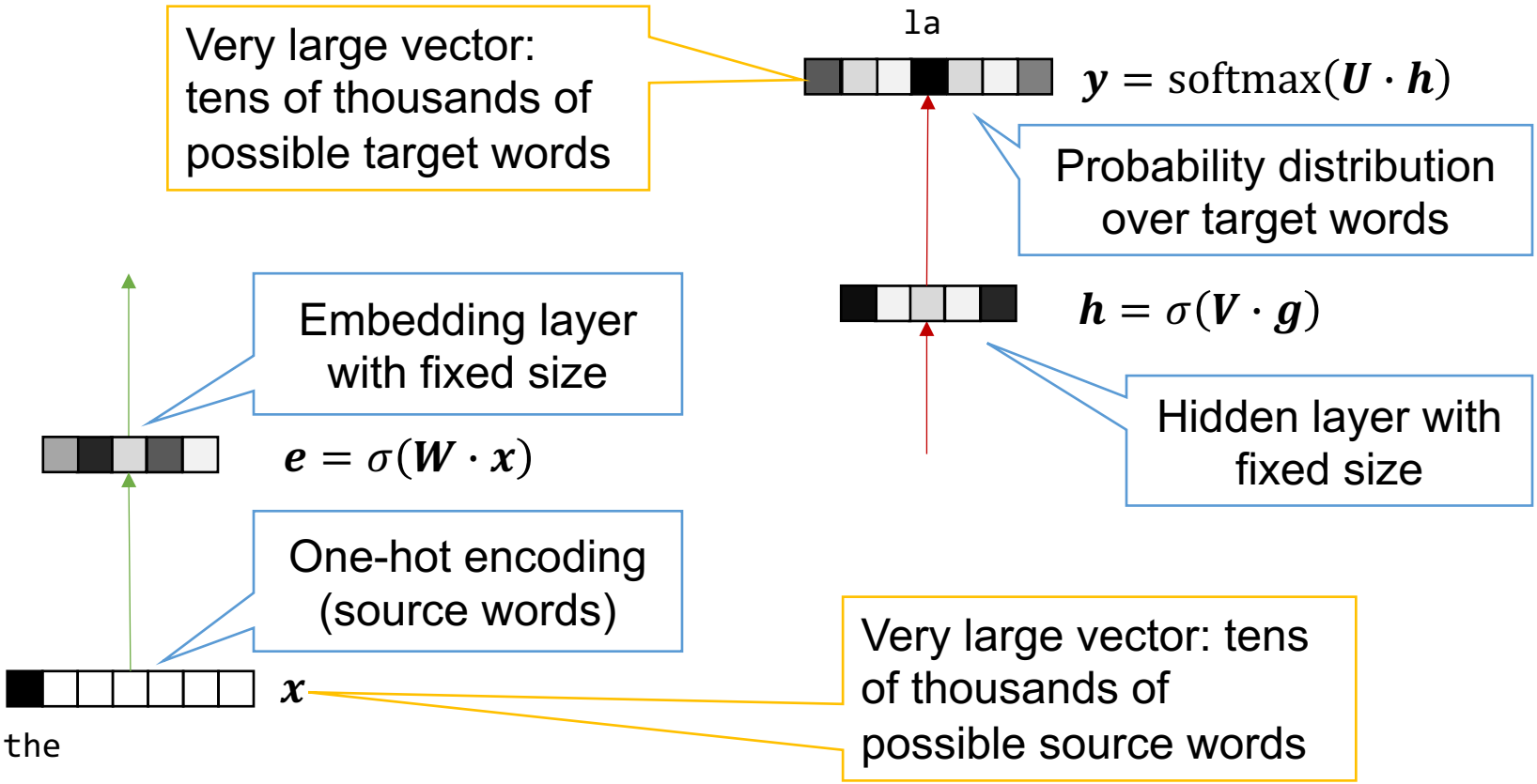
- Find parallel training data for language pair
  - OPUS is a good starting point, but some resources there are really noisy…
  - Segment the data into subwords
- Train a model
  - Transformer architecture, default parameters are generally fine
  - There are several easy-to-use MT toolkits: OpenNMT, Sockeye, fairseq
- Produce new translations with the trained model
  - Evaluate the quality of the translations

Open-vocabulary translation

Representing words

Encoder:

Decoder:



Open-vocabulary translation

- Training corpora typically contain millions of word types
- Morphological processes (inflection, derivation, compounding) allow formation and understanding of unseen words
- No training corpus contains all existing names, numbers, etc.

Translation is an open-vocabulary problem

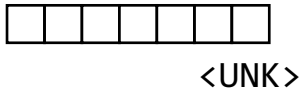
Open-vocabulary translation

Large vocabularies:

- Increase the memory requirements
- Decrease training and decoding speed
- Are still not large enough…

Solution (well, not really…):

- Limit the vocabulary vectors to the *n* most frequent words
- Add one reserved cell for all rare and unknown words: **<UNK>**



Translation with <UNK> token

Some examples from English-Czech translation:

- The author Stephen Jay Gould died 20 years after diagnosis.
  Autor <UNK> <UNK> <UNK> zemřel 20 let po <UNK>.
- As the Reverend Martin Luther King Jr. said fifty years ago:
  Jak řekl reverend Martin <UNK> King <UNK> před padesáti lety:
- Her 11-year-old daughter, Shani Bart, said it felt a "little bit weird" [..] back to school.
  Její <UNK> dcera <UNK> <UNK> řekla, že je to "trochu divné", [..] vrací do školy.

- **That's ok for a start, but we'll need a better solution…**

Open vocabulary translation

Solution 1 – Back-off models:

- Replace rare words with <UNK> at training time
- When system produces <UNK>, translate with a back-off method, for example a dictionary

- Limitations?
  - Compounds: hard to model 1-to-many relationships
  - Morphology: hard to predict inflection with back-off dictionary
  - Names: if alphabets differ, we need transliteration
  - It is quite hard to determine the source word for a given target position (attention helps less than expected)

Open vocabulary translation

Solution 2 – Character-level translation:

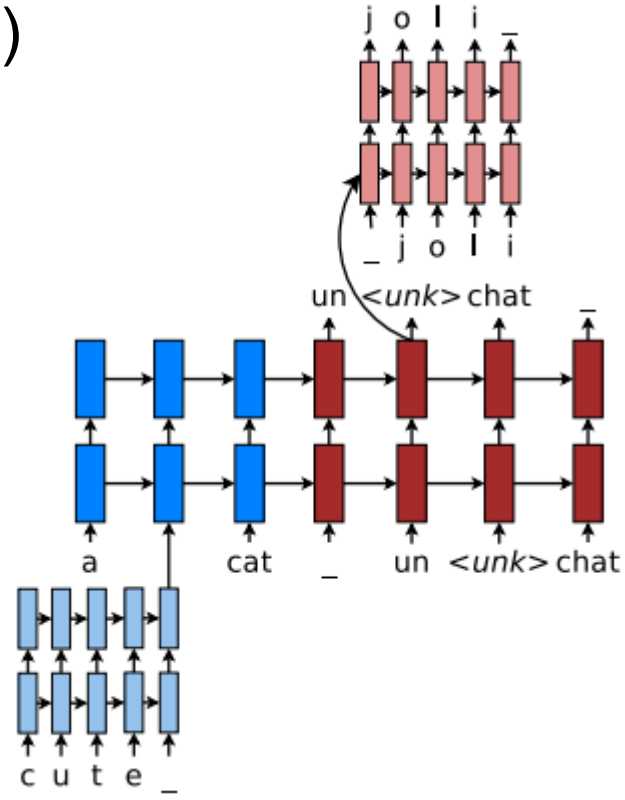
- T h e _ a u t h o r _ S t e p h e n _ J a y _ G o u l d _ d i e d _ 2 0 _ y e a r s _ a f t e r _ d i a g n o s i s .
- <UNK> is only used for unknown characters

- Limitations?
  - Still many UNKs for languages with large character sets (e.g. Chinese)
  - Increasing sequence length slows training/decoding
  - On which level should we represent meaning? Characters don't have any meaning by themselves…

Open vocabulary translation

Solution 2b – Hierarchical models:

- E.g. Luong & Manning (2016)
- Encode each word as a sequence of characters
- When the model produces <UNK> in the output, back off to a character-level decoder.



Open vocabulary translation

Solution 3 – Subword units:

- Split words into pieces of variable length
  - In most cases, longer than single characters
  - In most cases, shorter than entire words
- Ideally in a morphologically sensible way
  - Such that any unseen word can be decomposed into subwords
- Without using any external resources
  - Shouldn’t rely on morphological analyzers, dictionaries, grammars, …

Subword segmentation schemes

- Byte pair encoding (BPE) ~2016-2019

  R. Sennrich, B. Haddow & A. Birch (2016). *Neural Machine Translation of Rare Words with Subword Units*. Proceedings of ACL 2016.

- Morfessor

  S. Grönroos, S. Virpioja & M. Kurimo (2020). *Morfessor EM+Prune: Improved Subword Segmentation with Expectation Maximization and Pruning*. Proceedings of LREC 2020.

- SentencePiece unigram model Since 2019

  T. Kudo & J. Richardson (2018). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. Proceedings of EMNLP 2018 System Demonstrations.

Byte pair encoding

- General algorithm:
  - Start with a vocabulary of single characters
  - Merge frequent n-gram pairs into a new n-gram
  - Stop after k merging steps (this controls the final vocabulary size)
- Example:

Dictionary	
l o w	5
l o w e r	2
n e w e s t	6
w i d e s t	3

Vocabulary
l, o, w, e, r, n, w, s, t, i, d

Byte pair encoding – Example

Dictionary	
l o w	5
l o w e r	2
n e w **es** t	6
w i d **es** t	3

Vocabulary
l, o, w, e, r, n, w, s, t, i, d, **es**

The pair e+s has a frequency of 9. Merge it.

Dictionary	
l o w	5
l o w e r	2
n e w **est**	6
w i d **est**	3

Vocabulary
l, o, w, e, r, n, w, s, t, i, d, es, **est**

The pair es+t has a frequency of 9. Merge it.

Dictionary	
lo w	5
lo w e r	2
n e w est	6
w i d est	3

Vocabulary
l, o, w, e, r, n, w, s, t, i, d, es, est, **lo**

The pair l+o has a frequency of 7. Merge it.

Byte pair encoding – Example

Suppose we have the following new word: *lowest*

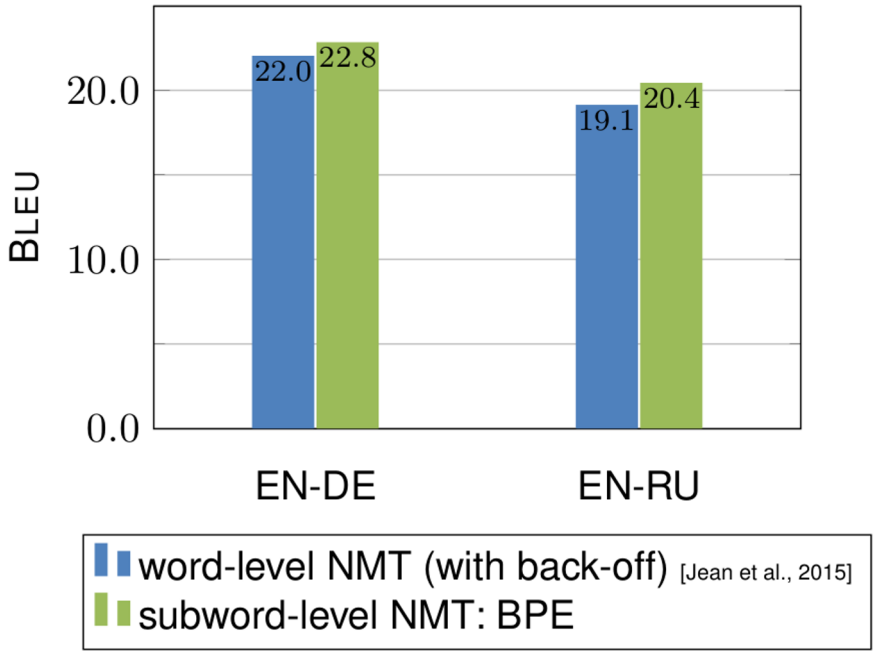
- l o w e s t (split into characters)
- l o w **es** t (apply merge *e s* → *es*)
- l o w **est** (apply merge *es t* → *est*)
- lo w **est** (apply merge *l o* → *lo*)
- Final segmentation: *lo w est*

Byte pair encoding

- Automatically creates fixed-size sub-word vocabularies for NMT
- The operations learned on training set can be applied to unknown words
  - How well it really generalizes is still an open question
- Compression of frequent character sequences improves efficiency
  - Trade-off between text length and vocabulary size
  - Vocabulary size is determined by a *hyperparameter*: the number of merging steps

Byte pair encoding

Translation quality:



Byte pair encoding

Examples:

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
word-level (with back-off)	Forschungsinstitute
BPE	Gesundheits|forsch|ungsin|stitute
source	rakfisk
reference	ракфиска (rakfiska)
word-level (with back-off)	rakfisk → UNK → rakfisk
BPE	rak|f|isk → рак|ф|иска (rak|f|iska)

SentencePiece

- SentencePiece uses a probabilistic model of subword segmentation
  - It can provide a list of possible segmentations together with their probabilities
- SentencePiece does not require word boundaries in its input. Whitespace is modelled in the same way as any other character
  - Useful for languages that do not use whitespace between words
  - Does not require tokenization (separate punctuation signs from words)

BPE vs SentencePiece

- BPE marks word continuation:
  - Hello wor@@ ld@@ .
  - Postprocessing: remove @@<space> sequences

- SentencePiece marks whitespace:
  - Hello _wor ld .
  - Postprocessing: delete all spaces, then replace _ by space

Evaluation of machine translation

Evaluation

Human evaluation:

- ultimately what we are interested in
- very time consuming (and boring), costly
- not re-usable
- subjective

Automatic evaluation:

- cheap and re-usable
- not necessarily reliable

Human evaluation criteria

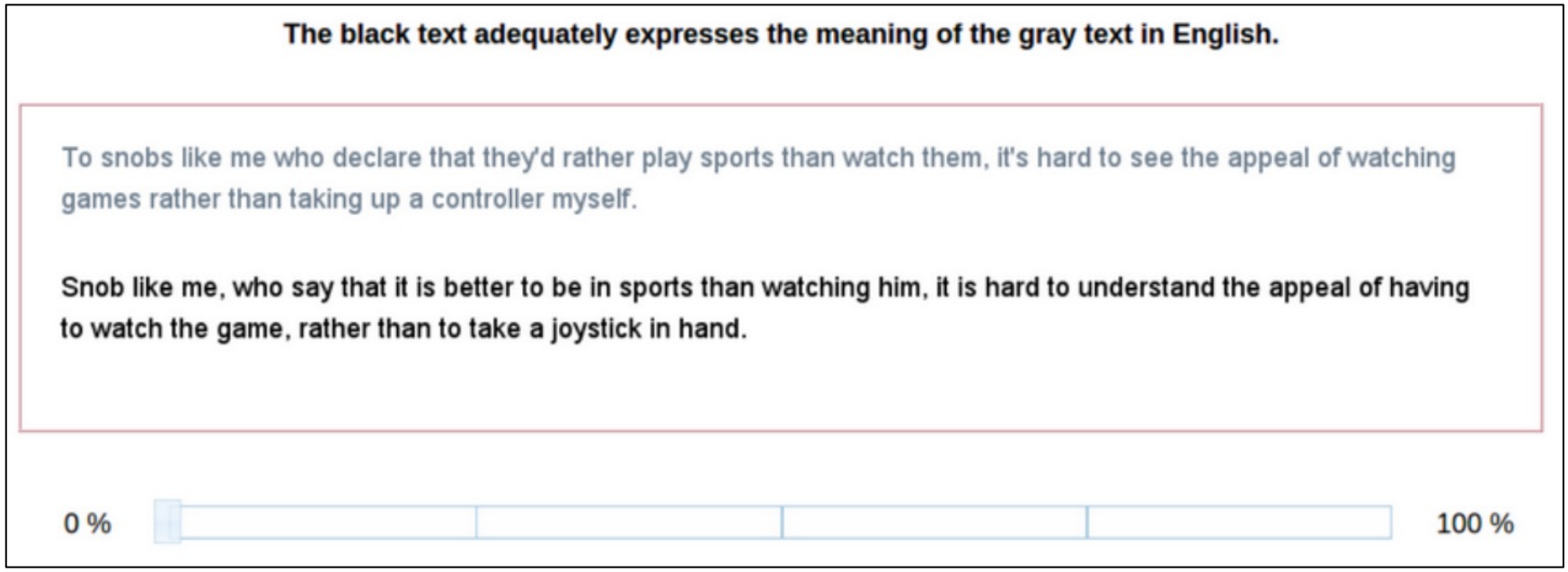
- **Adequacy:**
  - Does the output convey the same meaning as the input or reference sentence?
  - Is part of the message lost, added, or distorted?
  - Requires access to either source or reference.
- **Fluency:**
  - Is the output good fluent English?
  - This involves both grammatical correctness and idiomatic word choices.
  - Can be judged without other resources.

Adequacy	
5	all meaning preserved
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Human evaluation criteria

- **General translation quality**
  - It is sometimes difficult to distinguish between fluency errors and adequacy errors.



Evaluation

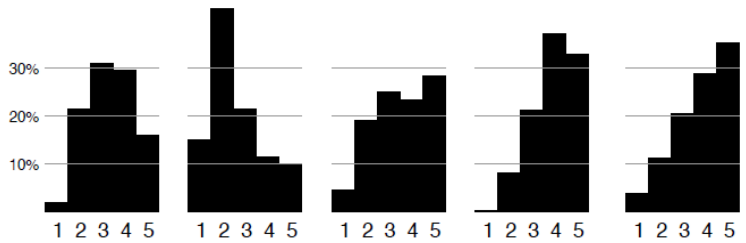
Human evaluation:

- ultimately what we are interested in
- very time consuming (and boring), costly
- not re-usable
- subjective

Automatic evaluation:

- cheap and re-usable
- not necessarily reliable

Histogram of adequacy judgements by different human evaluators



(from WMT 2006 evaluation)

Automatic evaluation

- Provide a human reference translation
- Compute similarity between reference translation and machine translation output

Source:

La France a-t-elle bénéficié d'informations fournies par la NSA concernant des opérations terroristes visant nos intérêts ?

System output:

Did France profit from furnished information by the NSA concerning of the terrorist operations aiming our interests?

Reference:

Has France benefited from the intelligence supplied by the NSA concerning terrorist operations against our interests?

Automatic reference-based evaluation

The BLEU score

- The most popular (and most criticized) evaluation metric for MT…
  - Stands for *BiLingual Evaluation Understudy*
- Main idea: compute n-gram overlap ($n = 1 \ldots 4$) between system output and reference
  - Add a brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{Len_{Sys}}{Len_{Ref}}\right) \cdot Prec_1 \cdot Prec_2 \cdot Prec_3 \cdot Prec_4$$

Brevity penalty

BLEU score – An example

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	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%

BLEU is a meaningful measure for an entire corpus, not for a single sentence.

Figure: Philipp Koehn

Issues with BLEU score

- Assumes that there is only one correct solution
  - Multiple references could be used, but rarely done in practice
- Ignores the relevance of words
  - Some words contribute more to the meaning than others
- Does not account for morphology, typos, etc.
  - A word is considered wrong as soon as one character is off
- Operates on local level
  - Does not consider overall grammaticality of the sentence or sentence meaning
- Scores are meaningless
  - Scores are very test-set specific, their absolute value is not informative
- Human translators score low on BLEU
  - Possibly because of higher variability, different word choices

chrF: character-level score

Break down words into character n-grams

- Give partial credits for matching stems without requiring a stemmer / lemmatizer

words	This is an example.
characters	T h i s i s a n e x a m p l e .
+space	T h i s _ i s _ a n _ e x a m p l e .

3-gram 4-gram

Computation:

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}}$$

- $\beta = 2$ for best results
- chrP: percentage of n-grams in output that are also present in reference ($n = 1 \ldots 6$)
- chrR: percentage of n-grams in reference that are also present in output ($n = 1 \ldots 6$)

Semantic similarity metrics (e.g. BERTScore)

If pre-trained embeddings or language models are available for the target language:

- Produce a sentence embedding of the system output
- Produce the sentence embedding of the reference
- Compute the cosine similarity between the two

- Depends on the quality of the embedding model
- Unclear to what extent this accounts for differences in fluency

Trained metrics (e.g. COMET)

15 years of MT evaluation campaigns have produced large datasets of human evaluations:

$$\langle \text{source}, \text{system\_output}, \text{reference}, \textit{score} \rangle$$

- These datasets can be used to train a supervised classifier that predicts evaluation scores for unseen examples.
- If $\text{source}, \text{system\_output}, \text{reference}$ are encoded as sentence embeddings by a multilingual language model, the classifier can also be applied to languages not seen in training data.

Applications and challenges

Sequence-to-sequence models

Machine translation is an instance of a sequence-to-sequence transformation task.

There are other similar tasks:

Input	Output	Task
English text	Japanese text	Machine translation
Old English text	Modern English text	Modernization / normalization
Colloquial English	Formal English	Style transfer
Entire document	Short description	Summarization
Inflected word form	Base form	Lemmatization
Speech signal	Transcription	Speech recognition

Multilingual translation models

- One model can learn to translate between multiple language pairs and translation directions.
- Append **language labels** to each source sentence to inform the model about the pair:

Training data	
<FROM_ES> <TO_FR> Visitaré a los niños.	Je viendrai voir les enfants.
<FROM_EN> <TO_ES> You did well, you did very well.	Bien hecho. Genial.
<FROM_ES> <TO_EN> Llegaremos enseguida.	We will be arriving soon.
<FROM_FR> <TO_ES> C'est la voix de notre âme qui parle.	Es la voz del alma que habla.

Multilingual translation models

- The model automatically learns to make use of the language labels when deciding which target words to generate.

Test data	
<FROM_EN> <TO_ES> It's the only way to achieve victory.	

Note: We have seen <FROM_EN> <TO_ES> examples during training.

Multilingual translation models

- **Zero-shot translation:** Translate from a known source language to a known target language without having seen training data for this particular language pair.

Zero-shot test data	
<FROM_EN> <TO_FR> It's the only way to achieve victory.	

Note: We have seen <FROM_EN> examples and <TO_FR> examples during training, but not <FROM_EN> <TO_FR> examples.

Readings

- Jurafsky & Martin, chapter 13
- Mikel Forcada (2017): Making sense of neural machine translation. Translation Spaces 6(2).
- Philipp Koehn (2020): Neural machine translation. Cambridge University Press.
  - That's a whole book – for reference only…

Reminder: there are still several MT-related thesis topics available!