

# Machine Translation

MVA - Speech and Language Processing

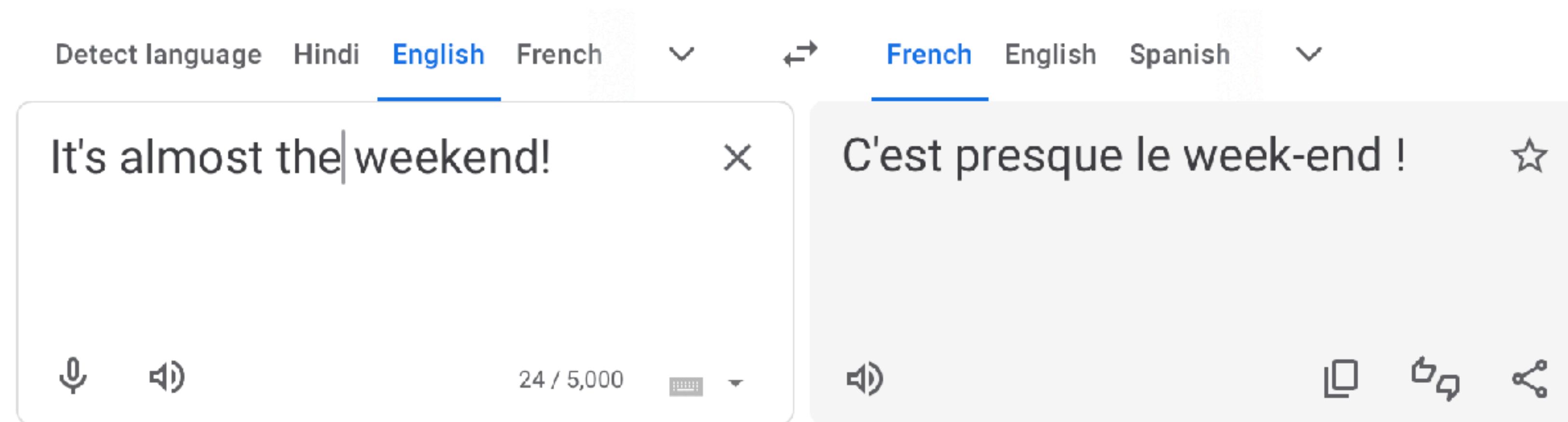
Rachel Bawden  
Inria

[rachel.bawden@inria.fr](mailto:rachel.bawden@inria.fr)



ChatGPT 4 (11/3/24): Can you draw me a photorealistic image of machine translation, with illustrations of languages in it?

# What is Machine Translation (MT)?



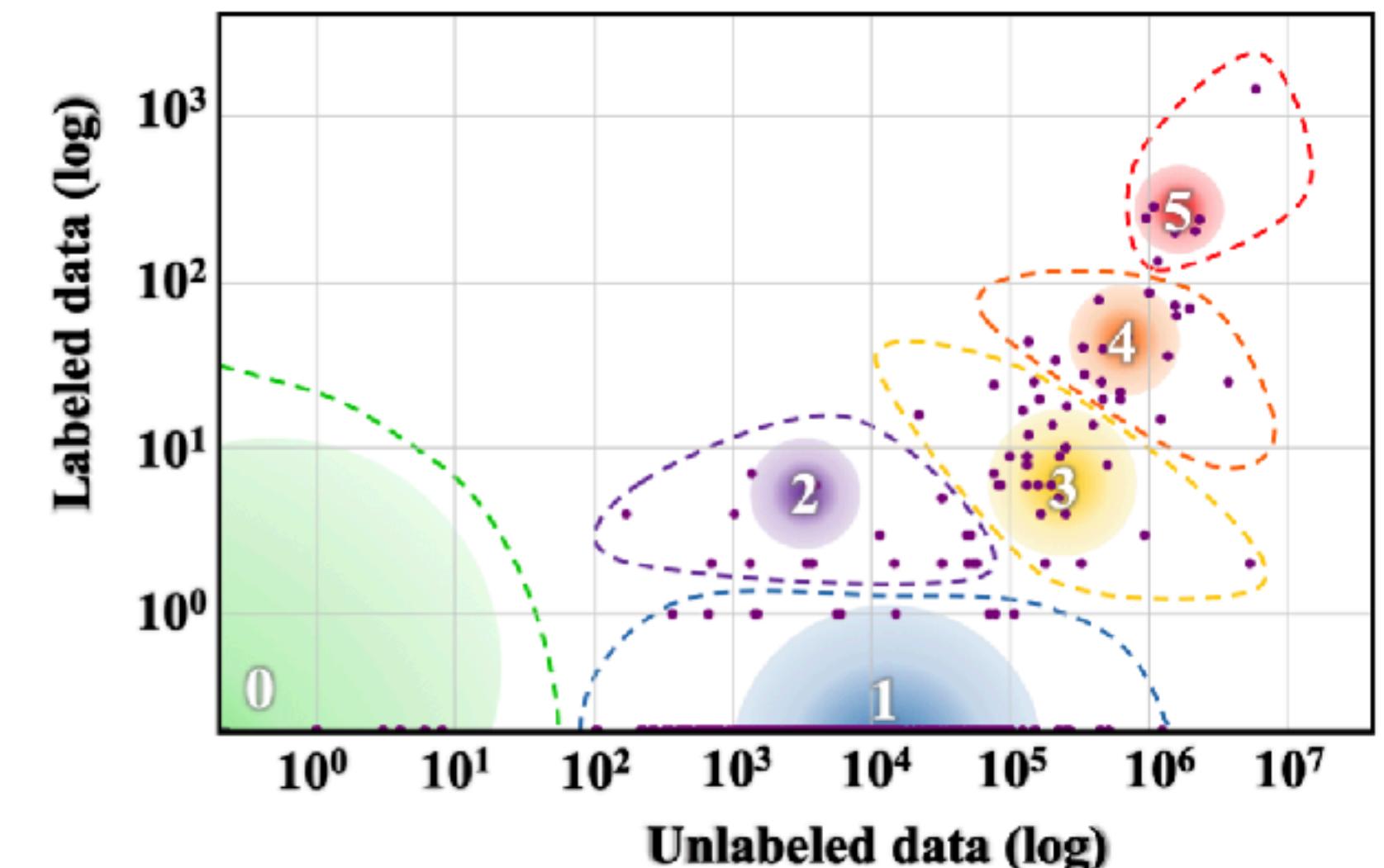
- Translating from one language into another
- Preserving meaning and intention of the original utterance

# Why is it difficult?

- **Linguistic diversity**
- Ambiguity
- Linguistic variation
- Unknown words

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



Table/Figure from (Joshi et al., 2020)

# Why is it difficult?

- Linguistic diversity
- **Ambiguity**
- Linguistic variation
- Unknown words, unknown constructions



🇫🇷 *J'ai demandé de l'aide à un **avocat***

🇬🇧 *I asked a lawyer for help*



🇬🇧 *He saw her duck.*

🇫🇷 *Il a vu son canard.*

🇫🇷 *Il l'a vue baisser la tête.*

# Why is it difficult?

- Linguistic diversity
- **Ambiguity**
- Linguistic variation
- Unknown words, unknown constructions



🇬🇧 *The owl was sleeping. It was tired.*

# Why is it difficult?

- Linguistic diversity
- **Ambiguity**
- Linguistic variation
- Unknown words, unknown constructions



🇬🇧 The owl was sleeping. It  
was tired.

# Why is it difficult?

- Linguistic diversity
- **Ambiguity**
- Linguistic variation
- Unknown words, unknown constructions



🇬🇧 The owl was sleeping. It was tired.

🇫🇷 Le hibou dormait. Il était fatigué.

# Why is it difficult?

- Linguistic diversity
- **Ambiguity**
- Linguistic variation
- Unknown words, unknown constructions



🇬🇧 The **owl** was sleeping. **It** was tired.

🇫🇷 Le **hibou** dormait. **Il** était fatigué.

🇫🇷 La **chouette** dormait. **Elle** était fatiguée.

# Why is it difficult?

- Linguistic diversity

- Ambiguity

- Linguistic variation

- Unknown words, unknown constructions

Not log Article Talk Read Edit

**Machine translation**

From Wikipedia, the free encyclopedia

I don't understand why there are Toffee Pennies in Quality Street. Are they for people who don't like chocolate but really fancy buying a tin of Quality Street?

7:35 pm · 9 Nov 2021 · Twitter for iPhone

42 Retweets 35 Quote Tweets 2,541 Likes

Richard Osman @richardosman · 2h  
Replying to @richardosman  
Am shocked and saddened by some of the pro-Toffee Penny responses.

126 5 745

@ · 2h  
Replying to @richardosman  
Stopped buying them the day they ousted Coffee Creams.

6

@ · 2h  
Have they ousted co  
Street since I retired  
Tell me the green tri

:") ofcfc n i cAN TELL HOW MUCH TIME N EFFORT u put into them cuz  
they always turn out lookin frikKIN ??4!4 AMAZING even when u think  
they arent ehHEHEHE LIKE EVEN IF ISS TERRAFORMING N NOT LIKE ?? A  
BUILDING OR A CASTLE OR A MONUMENT OR A VILLAGE OR WOTEVA,  
ITS ALWAYS !2!!!2!

# Why is it difficult?

- Linguistic diversity
- Ambiguity
- Linguistic variation
- **Unknown words, unknown constructions**

## greedflation



A general [increase](#) in prices during [a time](#) of [record](#) profits

[Grocery shopping](#) is a nightmare now that greedflation [doubled grocery](#) prices.

by [The Random Loon Publisher](#) July 16, 2023

## bed rotting



The gen z term for [staying in](#) bed for days on end, [binging](#) on Netflix, Tik Tok and [Hinge](#)

*Hey man, where have you been [these days](#)? Sorry, [brother](#), [I've](#) been bed rotting!*

by [Chriso!](#) June 2, 2023

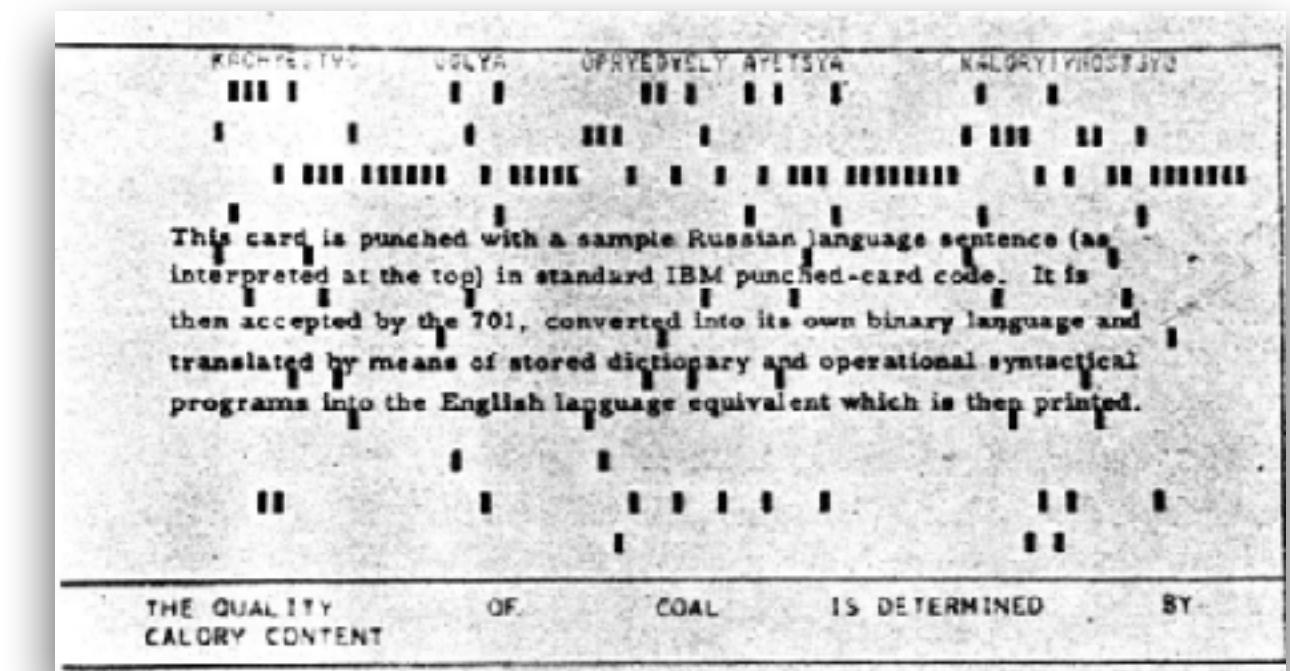
Examples from [urbandictionary.com](#)

# MT: a field with a lot of history!

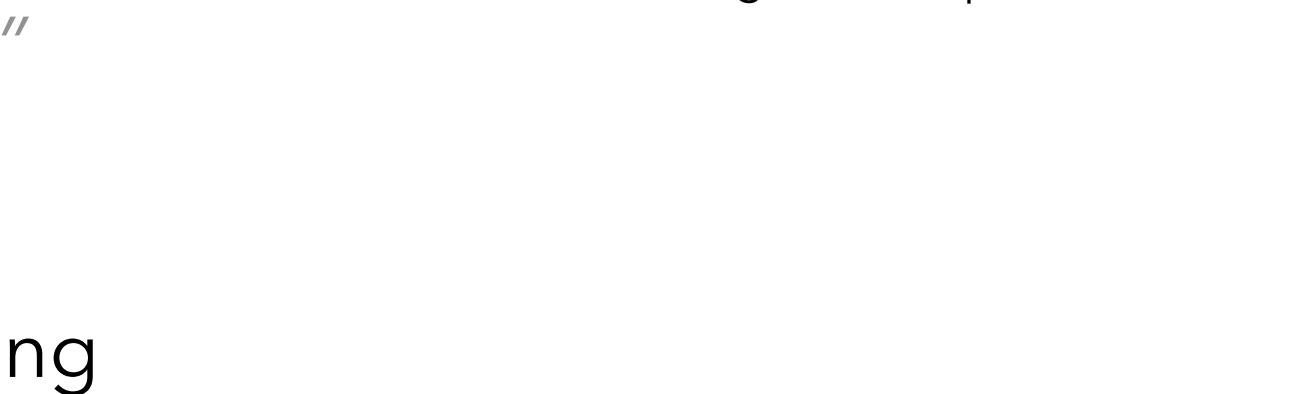
- 1933 2 patents for “translation machines”: Georges Artsrouni “mechanical brain”, Peter Petrovitch Trojanskii
- 1940 MT as one of the 1st applications of informatics ([Booth, Weaver, Richens, 1946-1948](#))
- 1949 Warren’s Weaver’s memorandum *Translation*, outlining 4 propositions, incl.  
 (i) need for context to resolve ambiguity, (ii) formal methods can be used,  
 (iii) use of cryptography and (iv) meaning independent of language
- 1954 IBM-Georgetown experiment: ~60 Russian sentences  
 translated into English from (250 words and 6 syntactic rules)  
*“five, perhaps three, years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact.”* (IBM press release)
- 1960 Bar Hillel’s report on MT  
*“The unreasonableness of aiming at fully-automatic high-quality translation is stressed”*
- 1966 ALPAC report  
 Consequence: less funding for MT and more focus on linguistic analysis such as parsing



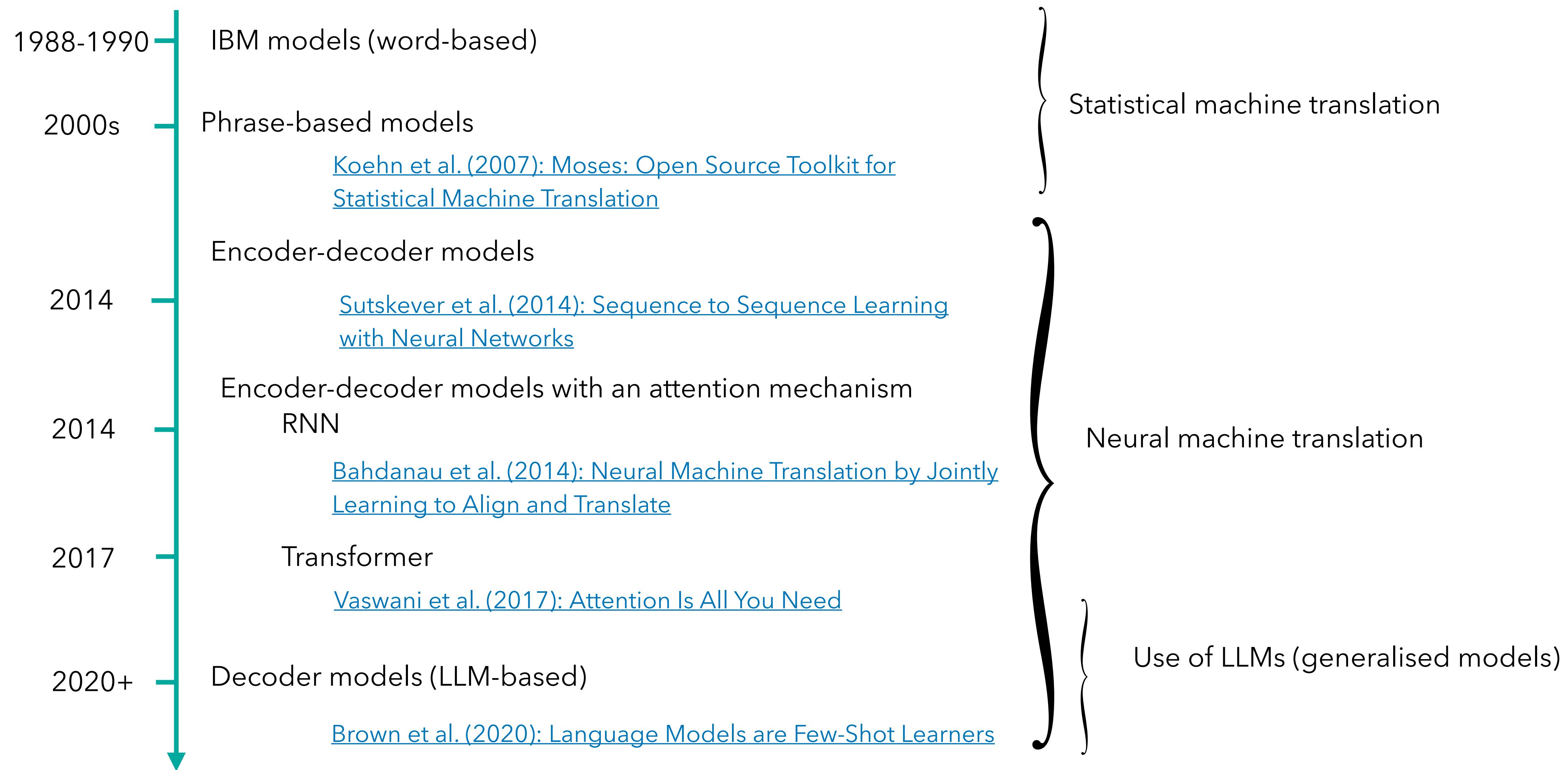
Artsrouni’s second machine (1933), Musée des Arts et Métiers



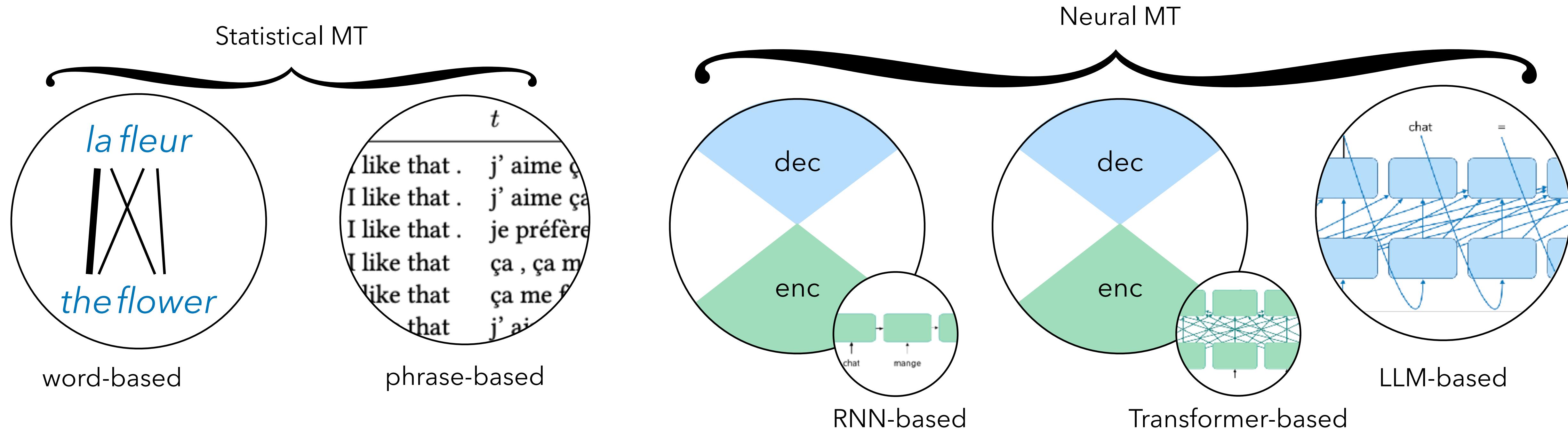
Punchcard from the IBM-Georgetown experiment (1954)



# More recent history



# The evolution of MT



- Increasing use of context for units being represented → → → → →
- Better handling of long-term dependencies (and ability to reorder and rephrase)
- Discrete (word- and phrase-based) to continuous representations

# Statistical MT (SMT)

ChatGPT 4 (11/3/24): Can you draw what machine translation looked like in the pre-neural era?



# Rule-based vs. data-driven

<b>Rule-based</b>	<b>Data-driven (statistical and neural)</b>
<p>✗ Requires complex rules and linguistic knowledge</p>	<p>✗ Requires large quantities of parallel data (corpus)</p>
<p>✗ Maintaining and updating rules can be costly and complex</p>	<p>✗ Influencing a model change through adding new data does not offer a guarantee of positive change</p>
<p>✓ Allows complete control and some interpretability</p>	<p>✗ Lack of interpretability</p>
<p>✗ Difficult to cover all phenomena and vocabulary</p>	<p>✓ Once the data exists, new ways of exploiting it can be developed (training can easily be rerun)</p>

# Statistical MT

- Main idea: find the translation candidate  $\hat{t}$  that maximises the probability of being the translation of a source sentence  $s$

Go through possible translations (the search space must be defined, because it is not possible to test all possible sentences)

$$\hat{t} = \operatorname{argmax}_{t \in T} P(t | s)$$

- Bayes reformulation

$$= \operatorname{argmax}_{t \in T} \frac{P(s | t)P(t)}{P(s)}$$

$$= \operatorname{argmax}_{t \in T} P(s | t)P(t)$$

translation model

language model

Equivalence between these two formulae.

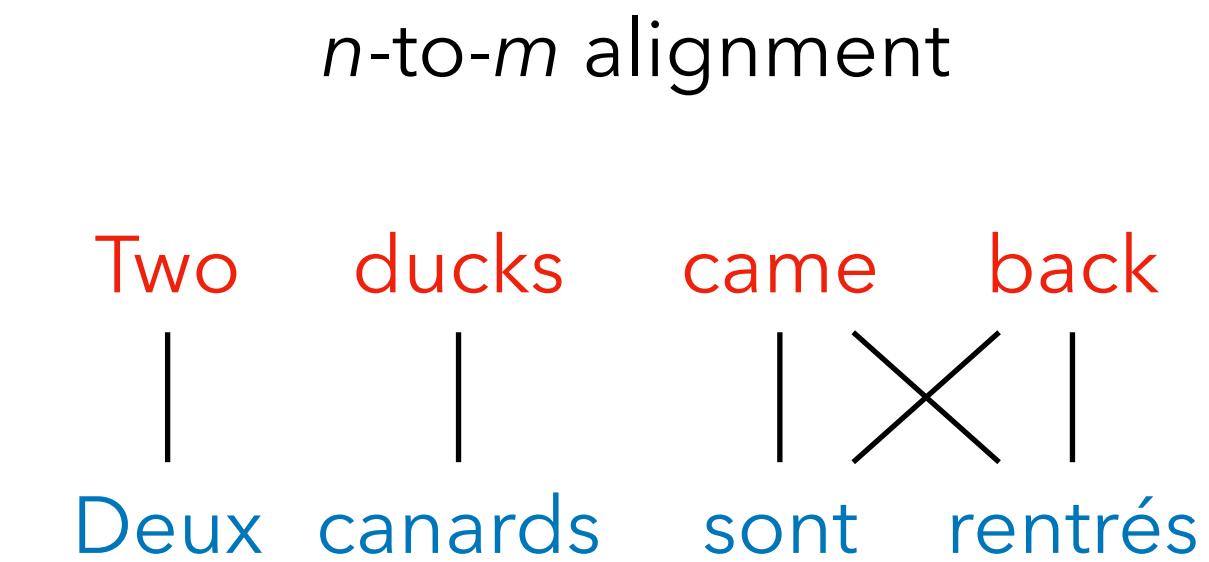
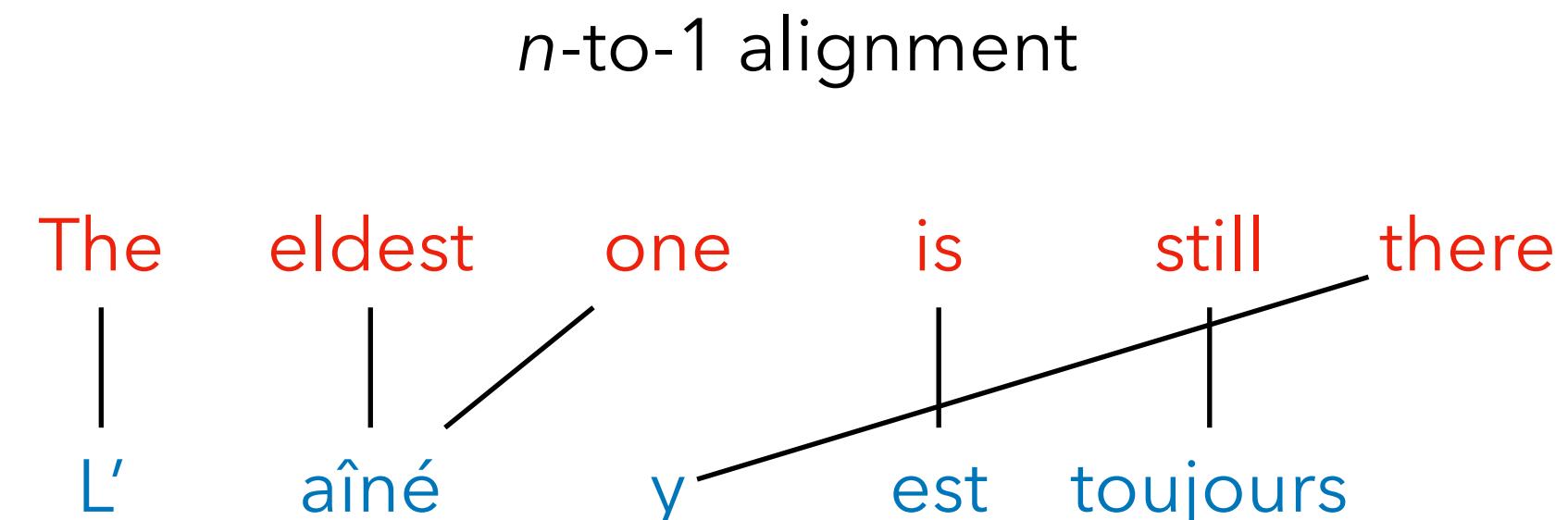
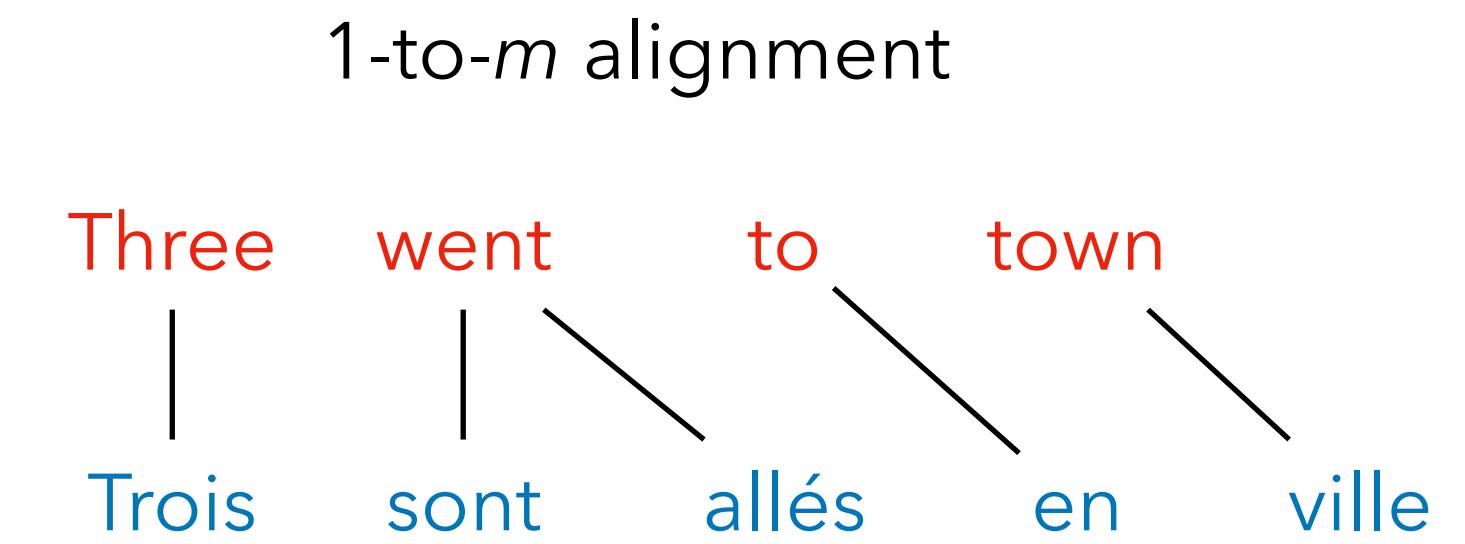
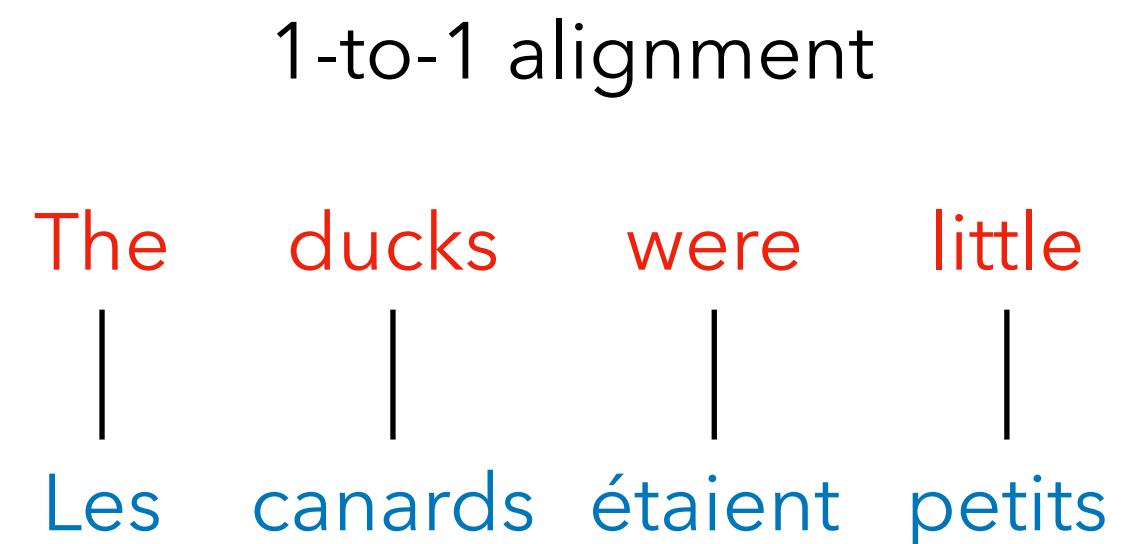
The reformulation helps to decompose the framework and to ensure the output is good quality (the language model can be trained on lots more data!)

# Word-based MT

E.g. IBM models 1-6 of increasing complexity ([Brown et al., 1993](#))

- Translation probabilities calculated at the word-level: decompose  $P(s|t)$  into *lexical* translation probabilities (i.e. what is the probability that word  $a$  is a translation of word  $b$ ?)
- Requires the notion of word alignments

Word alignments =  
more complex than simple  
1-to-1 mapping!



# Calculating lexical probabilities using alignments

- Alignment is a latent variable (i.e. we don't have the alignments, but they can be made to emerge)
- The translation probability is calculated by summing over all possible alignments:

$$P(s | t) = \sum_a P(a | t) \prod_i P(s_i | t_{a_i})$$

- You can initially set all alignments as equally likely, use expectation-maximisation to iteratively change the alignment weights, and then compute lexical probabilities based on the alignment probabilities.

*la maison*  
  
*the house*

*la maison bleue*  
  
*the blue house*

*la fleur*  
  
*the flower*

# Calculating lexical probabilities using alignments

- Alignment is a latent variable (i.e. we don't have the alignments, but they can be made to emerge)
- The translation probability is calculated by summing over all possible alignments:

$$P(s | t) = \sum_a P(a | t) \prod_i P(s_i | t_{a_i})$$

- You can initially set all alignments as equally likely, use expectation-maximisation to iteratively change the alignment weights, and then compute lexical probabilities based on the alignment probabilities.

*la maison*  
  
*the house*

*la maison bleue*  
  
*the blue house*

*la fleur*  
  
*the flower*

# Calculating lexical probabilities using alignments

- Alignment is a latent variable (i.e. we don't have the alignments, but they can be made to emerge)
- The translation probability is calculated by summing over all possible alignments:

$$P(s | t) = \sum_a P(a | t) \prod_i P(s_i | t_{a_i})$$

- You can initially set all alignments as equally likely, use expectation-maximisation to iteratively change the alignment weights, and then compute lexical probabilities based on the alignment probabilities.

*la maison*  
  
*the house*

*la maison bleue*  
  
*the blue house*

*la fleur*  
  
*the flower*

# Calculating lexical probabilities using alignments

- Alignment is a latent variable (i.e. we don't have the alignments, but they can be made to emerge)
- The translation probability is calculated by summing over all possible alignments:

$$P(s | t) = \sum_a P(a | t) \prod_i P(s_i | t_{a_i})$$

- You can initially set all alignments as equally likely, use expectation-maximisation to iteratively change the alignment weights, and then compute lexical probabilities based on the alignment probabilities.

*la maison*  
  
*the house*

*la maison bleue*  
  
*the blue house*

*la fleur*  
  
*the flower*

# Calculating lexical probabilities using alignments

- Alignment is a latent variable (i.e. we don't have the alignments, but they can be made to emerge)
- The translation probability is calculated by summing over all possible alignments:

$$P(s | t) = \sum_a P(a | t) \prod_i P(s_i | t_{a_i})$$

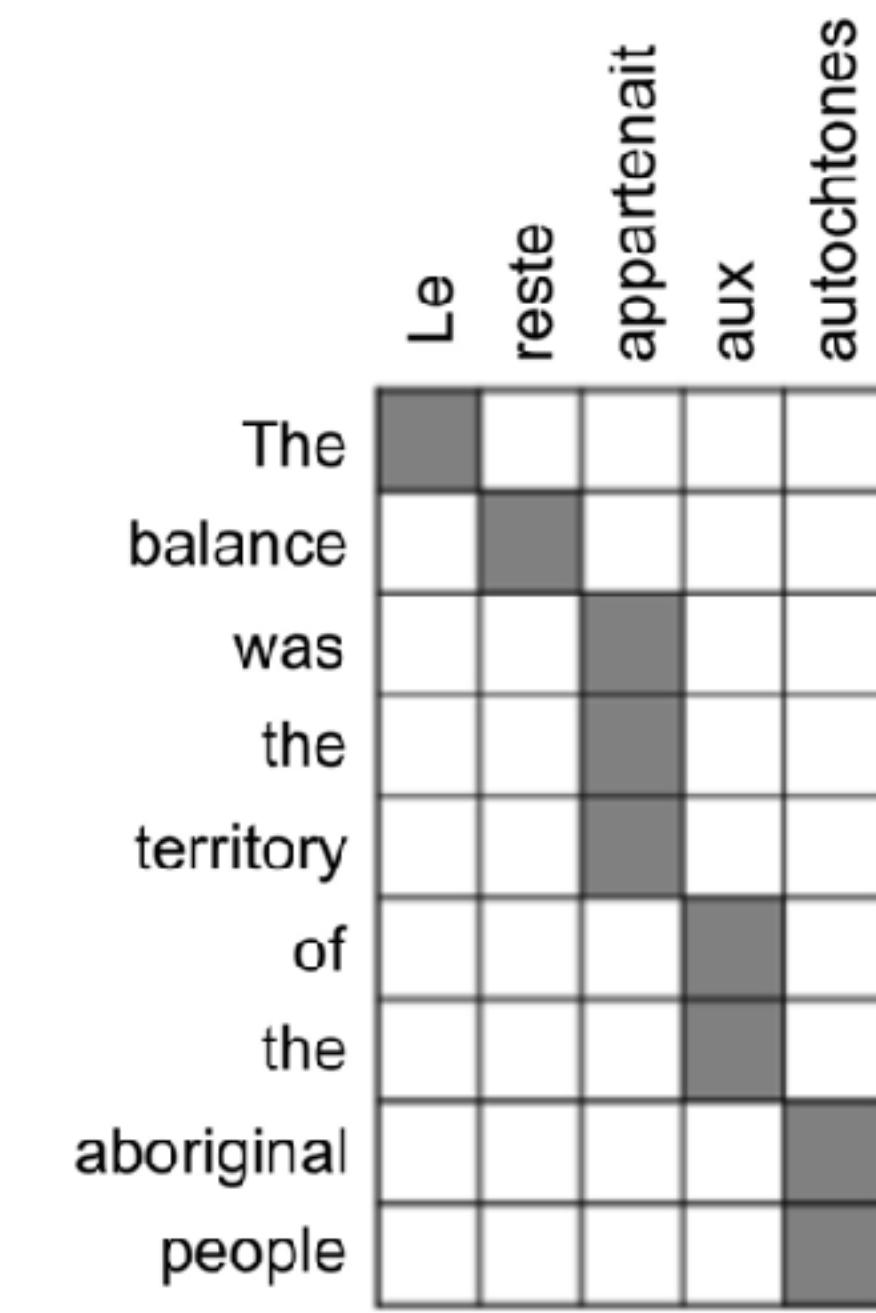
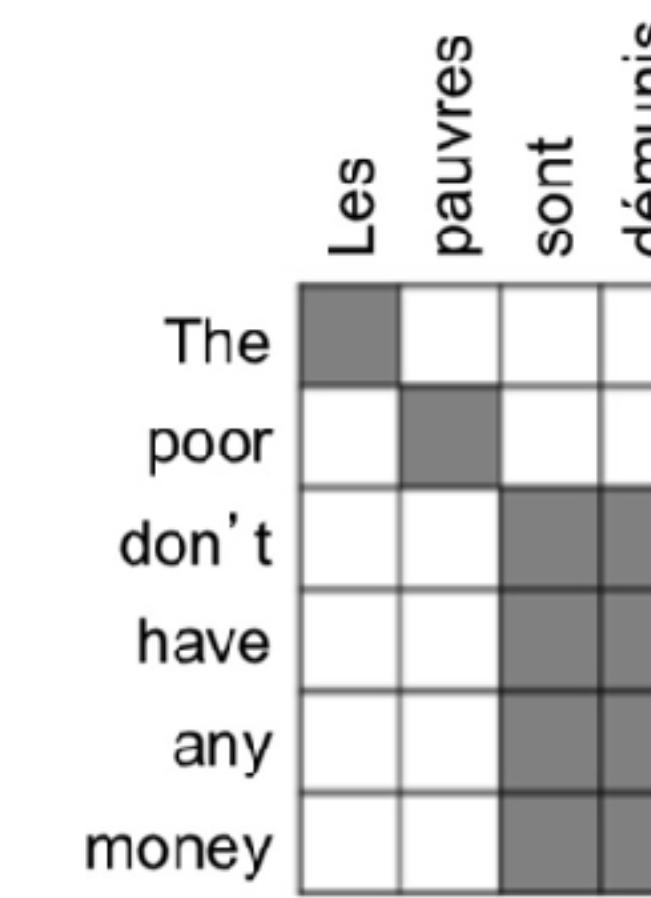
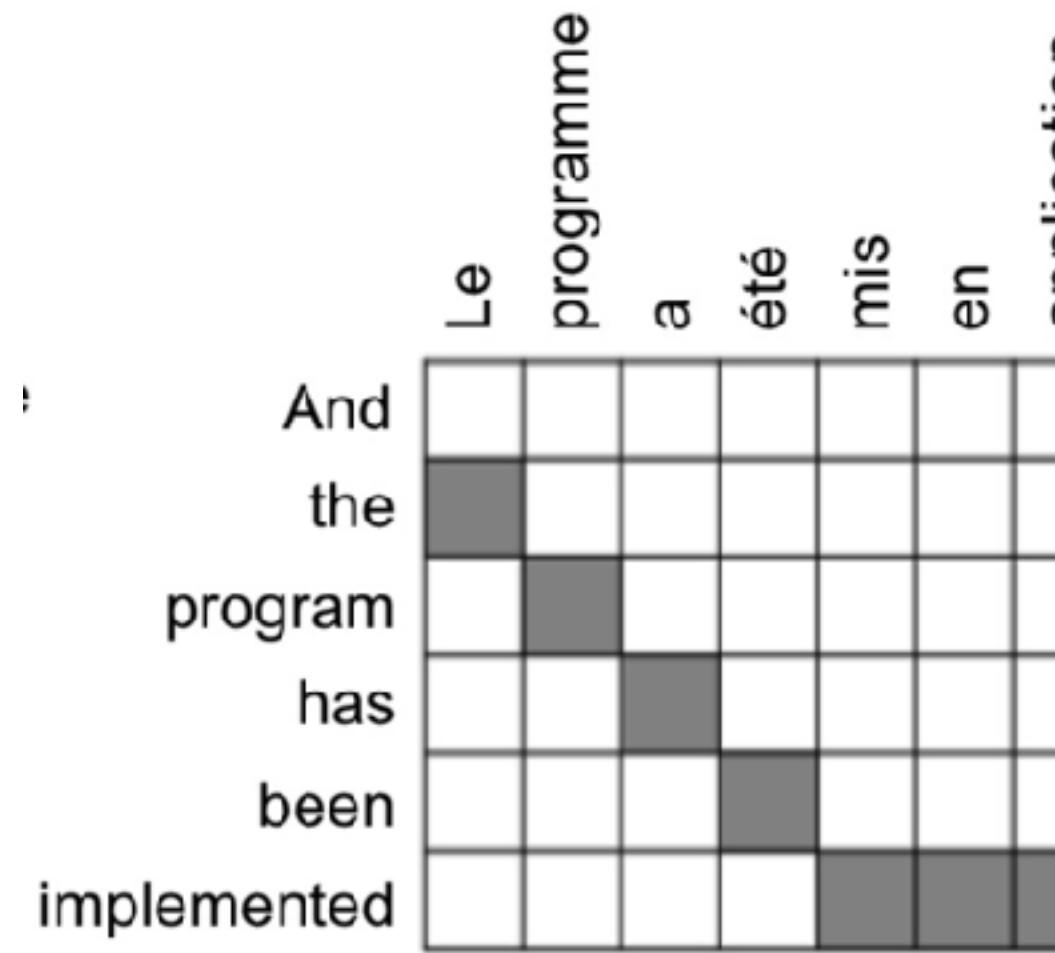
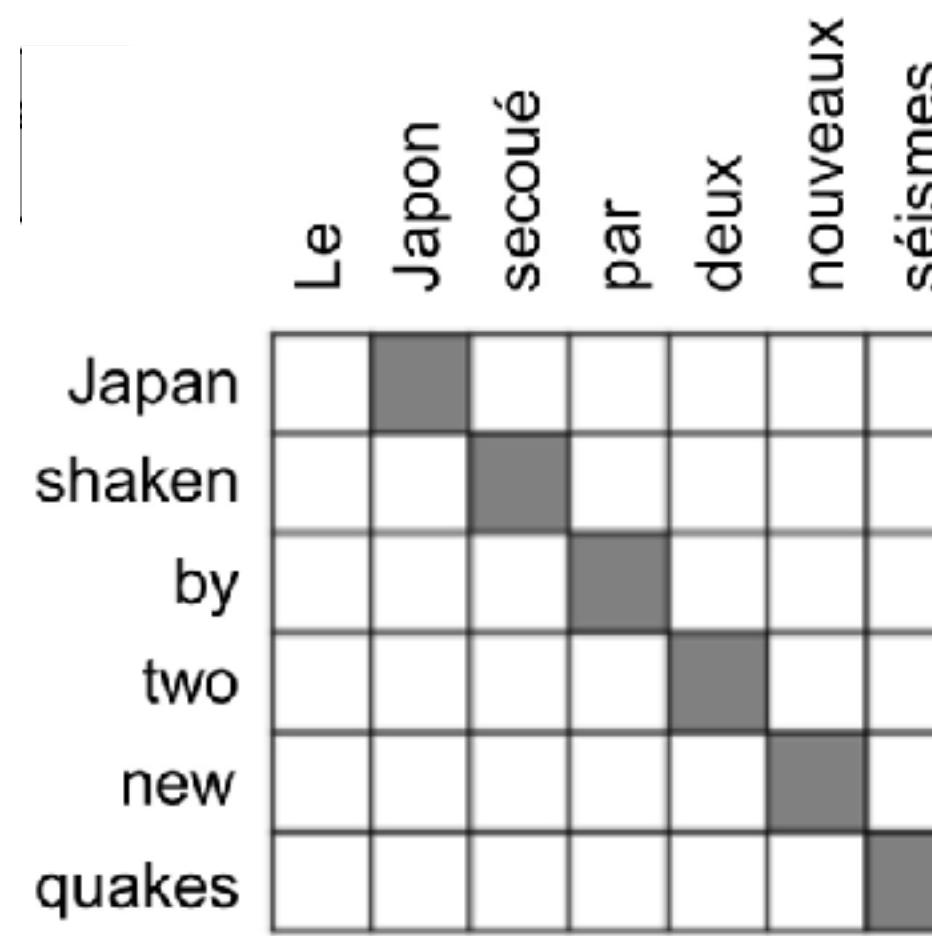
- You can initially set all alignments as equally likely, use expectation-maximisation to iteratively change the alignment weights, and then compute lexical probabilities based on the alignment probabilities.



P(*la|the*) = 0.45  
P(*le|the*) = 0.33  
P(*maison|house*) = 0.88  
P(*bleu|blue*) = 0.56  
...

# Induced word alignments

- The IBM models enable us to induce word alignments in corpora



- Can be useful in itself! (and was so for the next type of MT models →)

# Phrase-based MT

Most used toolkit: [Koehn et al. \(2007\): Moses: Open Source Toolkit for Statistical Machine Translation](#)

- Translation probabilities at the *phrase* level (sequences of tokens)
- Use an alignment model to extract “phrases”, their translations and the probabilities
- Combination of several scoring models (including the important translation model and a language model)
- Weights of the components *tuned* according to a metric (e.g. BLEU) to give optimal results

$$P(t|s) \propto \exp \left( \underbrace{\lambda_{TM} \text{TM}(s|t)}_{\text{Phrase translation model}} + \underbrace{\lambda_{LM} \text{LM}(t)}_{\text{Language model}} + \underbrace{\lambda_{DM} \text{DM}(s, t)}_{\text{Distortion model}} + \underbrace{\text{WP}(t)}_{\text{Word penalty}} \right)$$

# Phrase-based MT

- Probabilities in the phrase translation model:

$s$	$t$	$\log P_t(t s)$	$\log P_t(s t)$	$\log P_{lex}(t s)$	$\log P_{lex}(s t)$
I like that .	j' aime ça !	0.25	0.03336	0.00689	7.79e-05
I like that .	j' aime ça , tiens .	1	0.01779	0.00344	2.51e-08
I like that .	je préfère ça .	0.0185	0.01348	0.00344	0.000136
I like that	ça , ça me plaît .	0.1	0.00175	0.00344	2.90e-08
I like that	ça me fait plaisir .	0.0277	0.00021	0.00344	4.10e-09
I like that	j' aime cette idée .	0.25	0.04379	0.00344	6.06e-08

- Language model: mostly  $n$ -gram models (see course 1)

$$P(\text{I like red roses}) = P(\text{I} \mid \text{BOS})P(\text{like} \mid \text{I})P(\text{red} \mid \text{like})P(\text{roses} \mid \text{red})P(\text{EOS} \mid \text{roses})$$

One of the most used packages: [KenLM \(Heafield, 2011\)](#)

# Training and test data

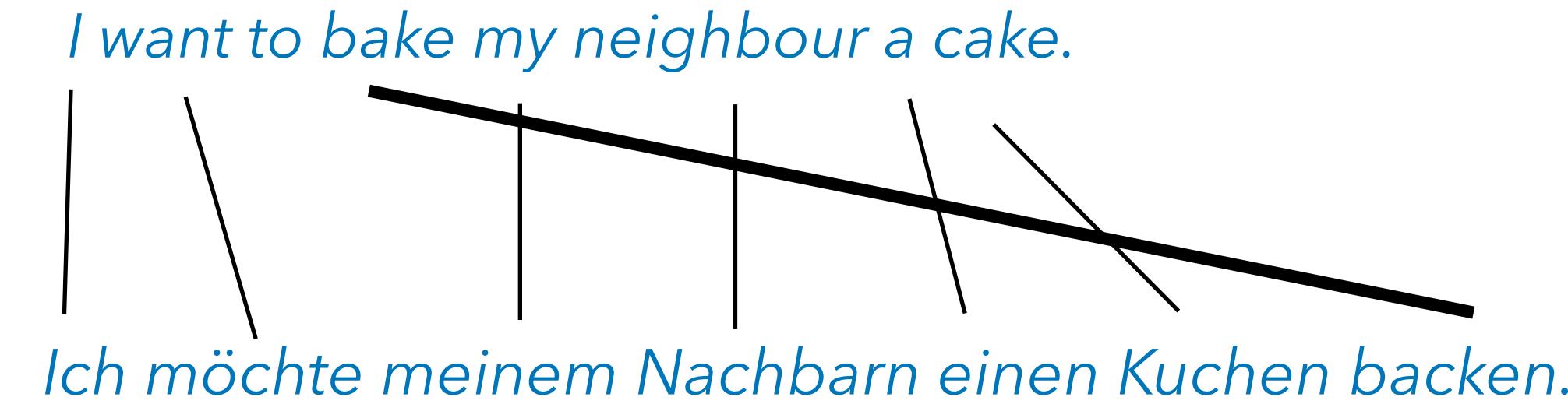
Collection of parallel corpora: [Tiedemann \(2012\): Parallel Data, Tools and Interfaces in OPUS](#)

- Parallel data: (generally) sentences in one language aligned with translations
- Generally translations (*references*) produced by humans
  - Translations exhibit different properties from original texts ("translationese")
  - The quality of translations is variable
- Training data and evaluation data

- **Institutional corpora**
  -  [Europarl](#) (22 languages)
  -  [MultiUN](#) (7 languages)
- **Subtitles**  [OpenSubtitles](#) (1782 language pairs)
- **Mined corpora**
  -  [ParaCrawl](#) (42 language pairs)
  -  [CCMatrix](#) (1200 language pairs)
  -  [NLLB](#) (1600 language pairs)

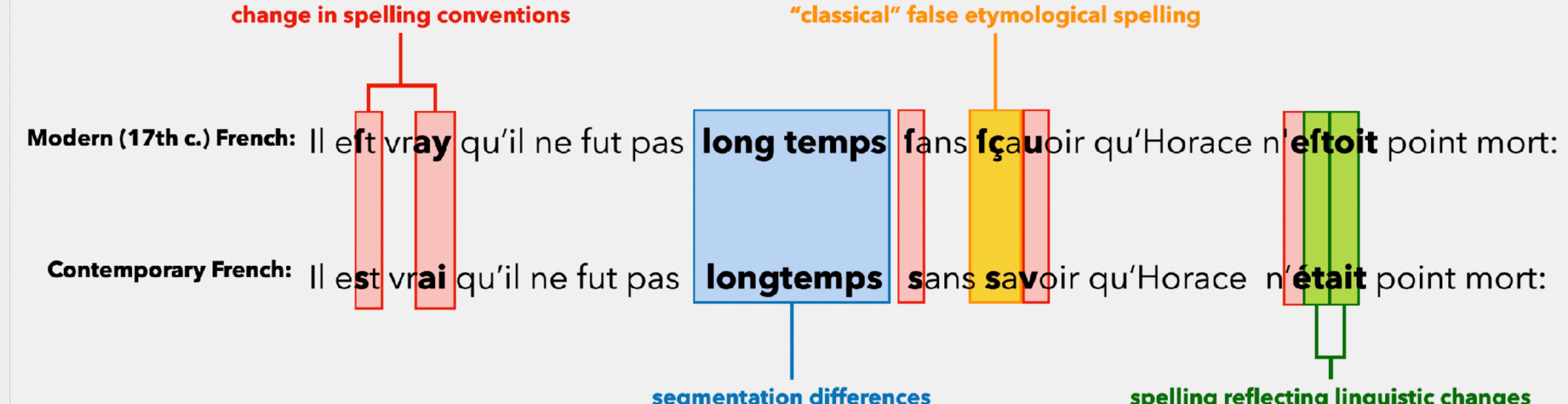
# Advantages and disadvantages of SMT

- More context taken into account than word-based MT, but still limited (inability to handle long-range changes)
- Some reordering is possible, but little reformulation is possible (so problems for certain language pairs with different word orders, e.g. English-to-German).



- Scores were beginning to plateau in the mid-2010s (it was difficult to get more gains). Slight improvements required complex refinements.

# Is statistical MT still relevant?



(Bawden et al., 2022; Gabay et al., 2022)

- For certain “translation-like” tasks that do not require long-distance context, e.g. normalisation of old texts (with minor spelling changes)

# Neural MT models

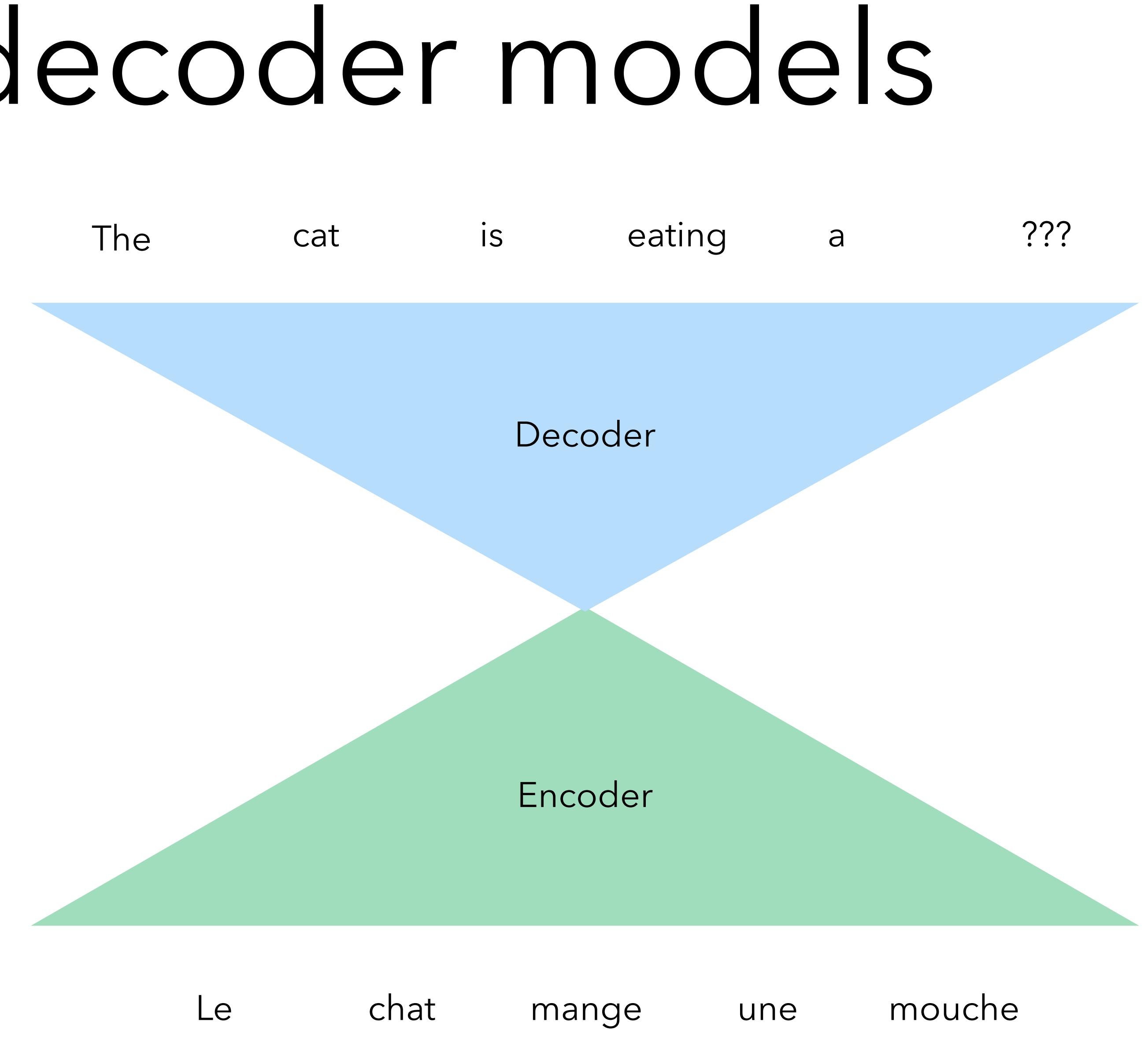
ChatGPT 4 (11/3/24): Can you now draw neural MT?



# Encoder-decoder models

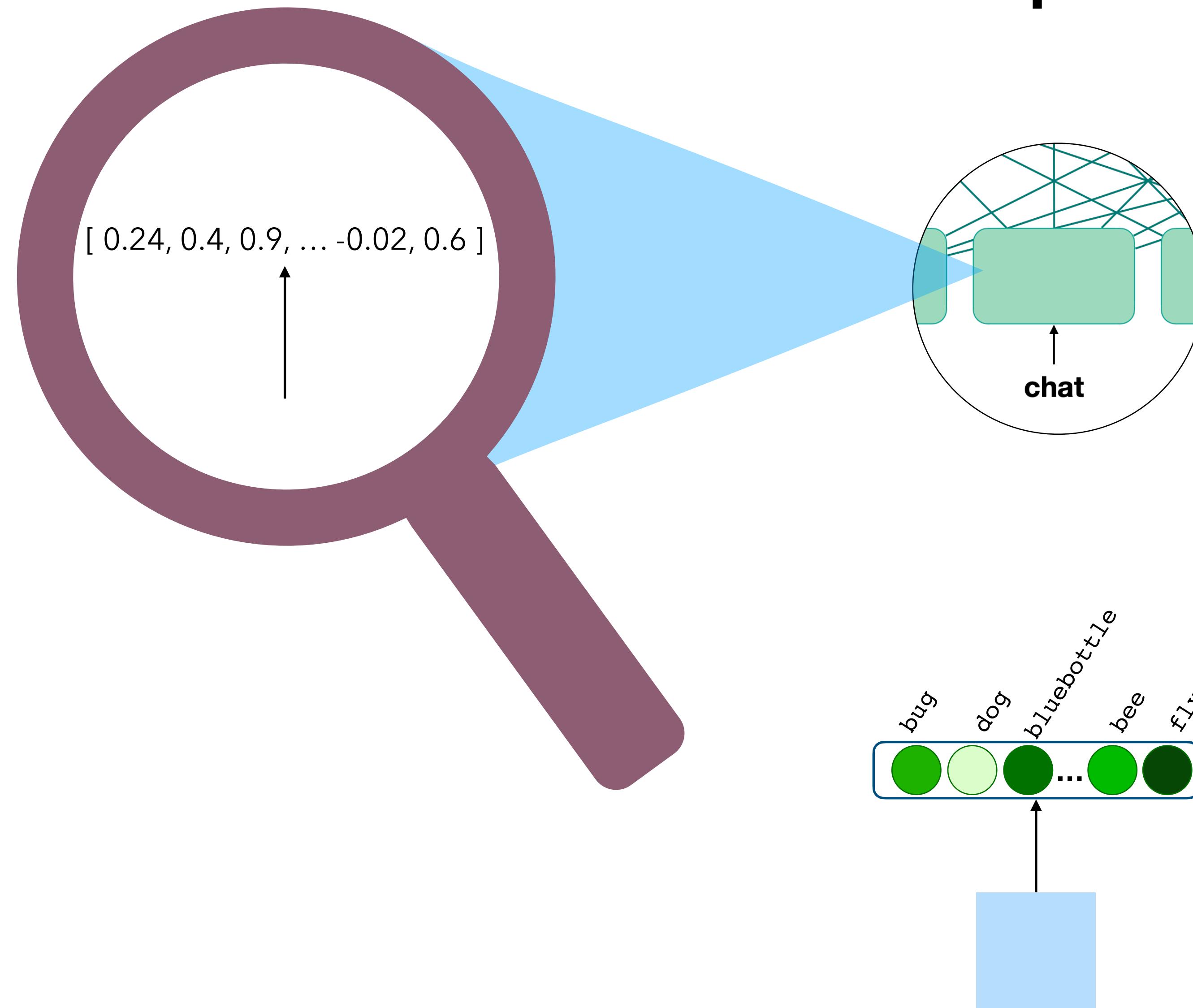
# Encoder-decoder models

- Encoder: representations of the source sequence
- Decoder: uses the encoder representations to generate the target sequence
- Decoding strategy:
  - **Auto-regressive: predict one token after another conditioned on the previous predictions.**



Not a new idea: [Allen \(1987\)](#), [Pollack \(1990\)](#)

# Reminder: inputs and outputs



- Inputs = sequences of tokens
- Each token represented as a 1-hot vector, mapped to an embedding within the model
- Softmax - predict a probability distribution over tokens

# Tokens: Subword segmentation

- Subword tokenisation:

- Increases generalisation capacity (especially when little data)
- Avoid unknown words at decoding time
- Can be trained on monolingual data

- BPE, sentence piece (unigram)

- Vocabulary size is important:

- Smaller = more splitting (fewer whole words)
- Larger = less splitting (more whole words)

Le chat a mangé une mouche

\_Le \_chat \_a \_mangé \_une \_mouche

Le chat a mangé une mouche domestique

\_Le \_chat \_a \_mangé \_une \_mouche \_domestique

Le chat a mangé un *musca domestica*

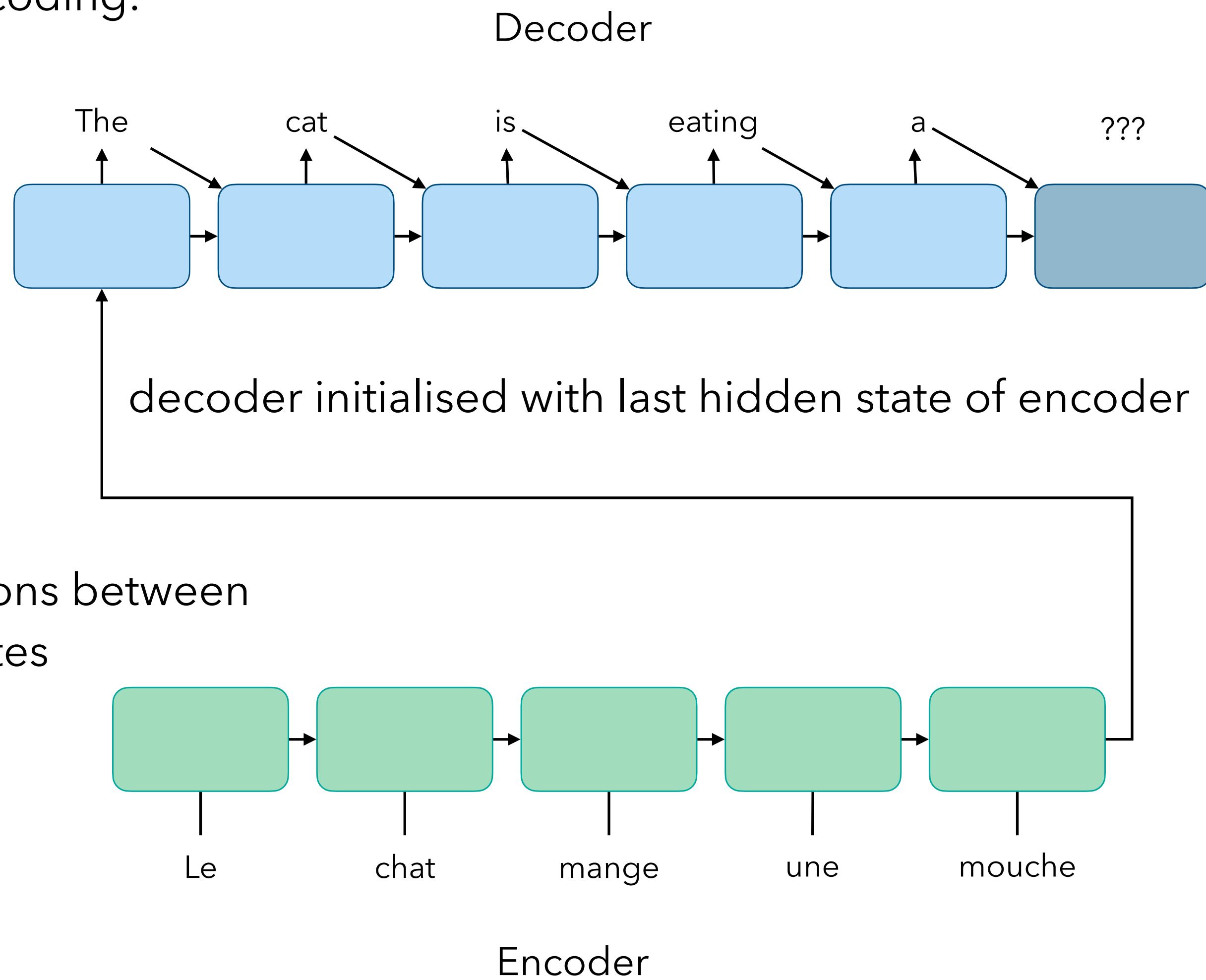
\_Le \_chat \_a \_mangé \_un \_musc a \_d om est ica

BPE subword tokenisation: [Sennrich et al. \(2016\)](#)

Sentencepiece tokenisation: [Kudo and Richardson \(2018\)](#)

# RNN models

auto-regressive decoding:  
token by token

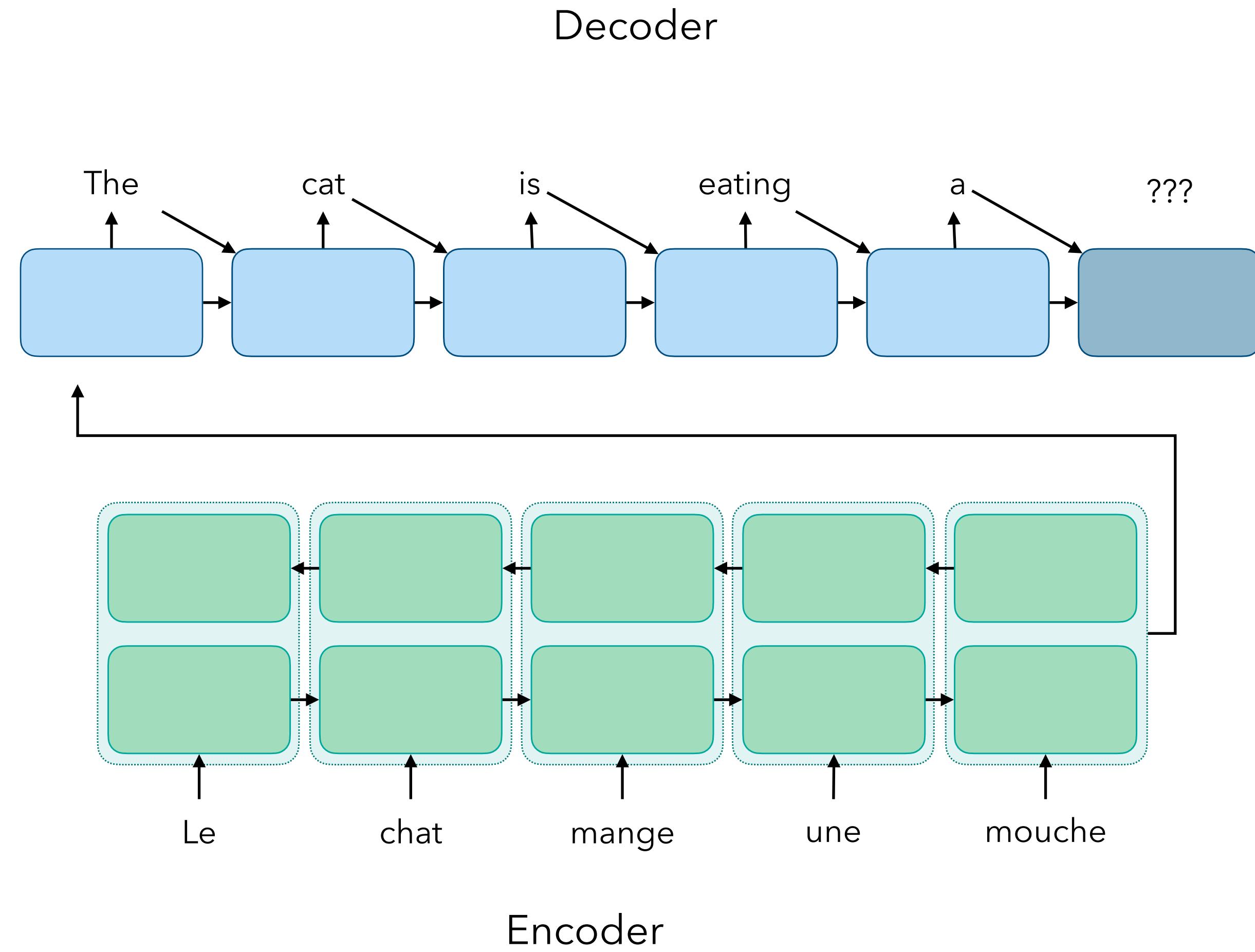


sequential connections between  
adjacent hidden states

(Cho et al., 2014; Sutskever et al., 2014)

- Main idea:
  - models the sequential nature of text
  - representation of each token dependent on its past context
- Some problems:
  - difficult to transfer information long distance:
    - LSTM units (long short-term memory)
    - GRU (gated recurrent units)
  - only access to left context
    - bi-directional RNNs ➔

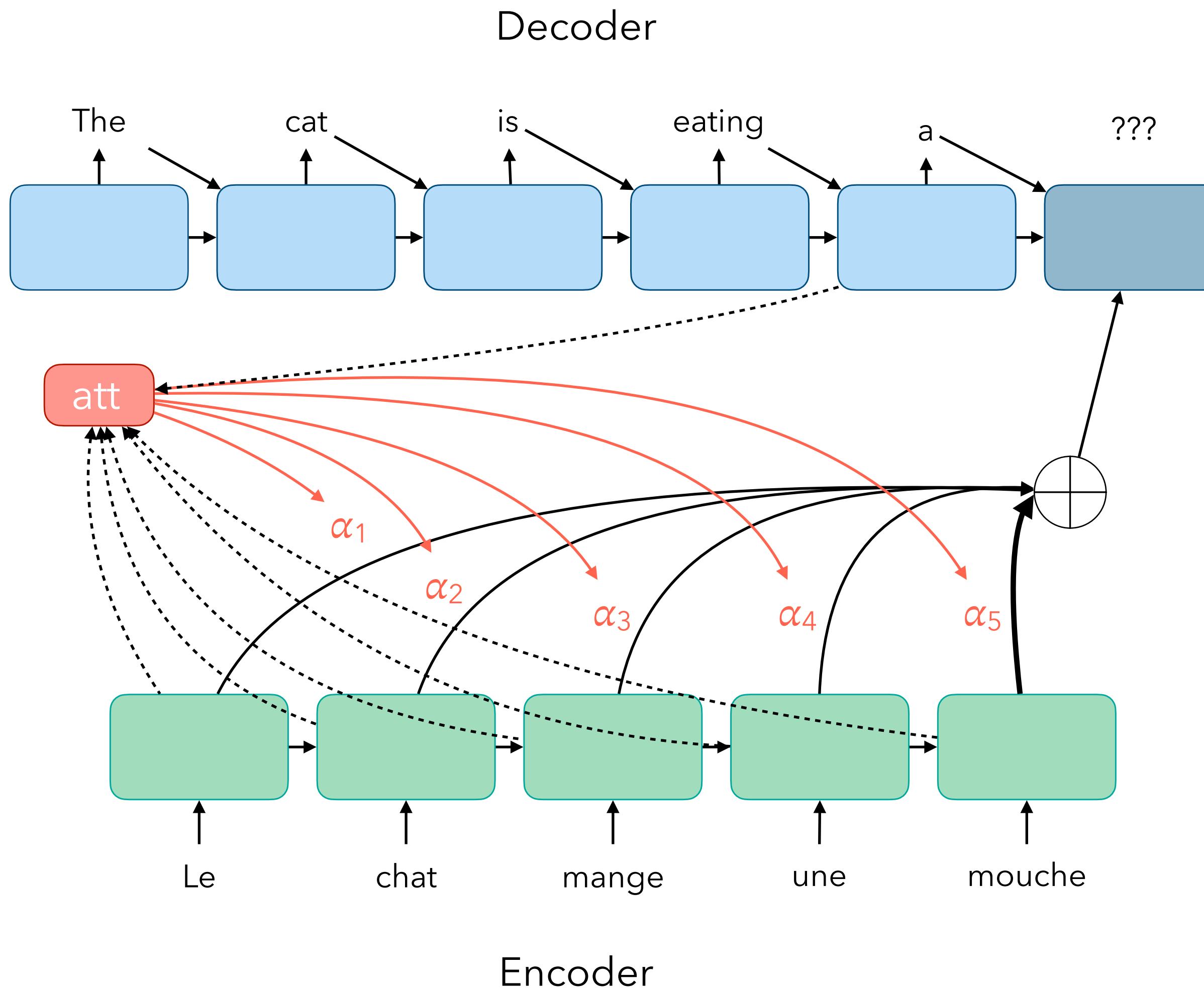
# Bidirectional RNN models



(Schuster and Paliwal, 1997)

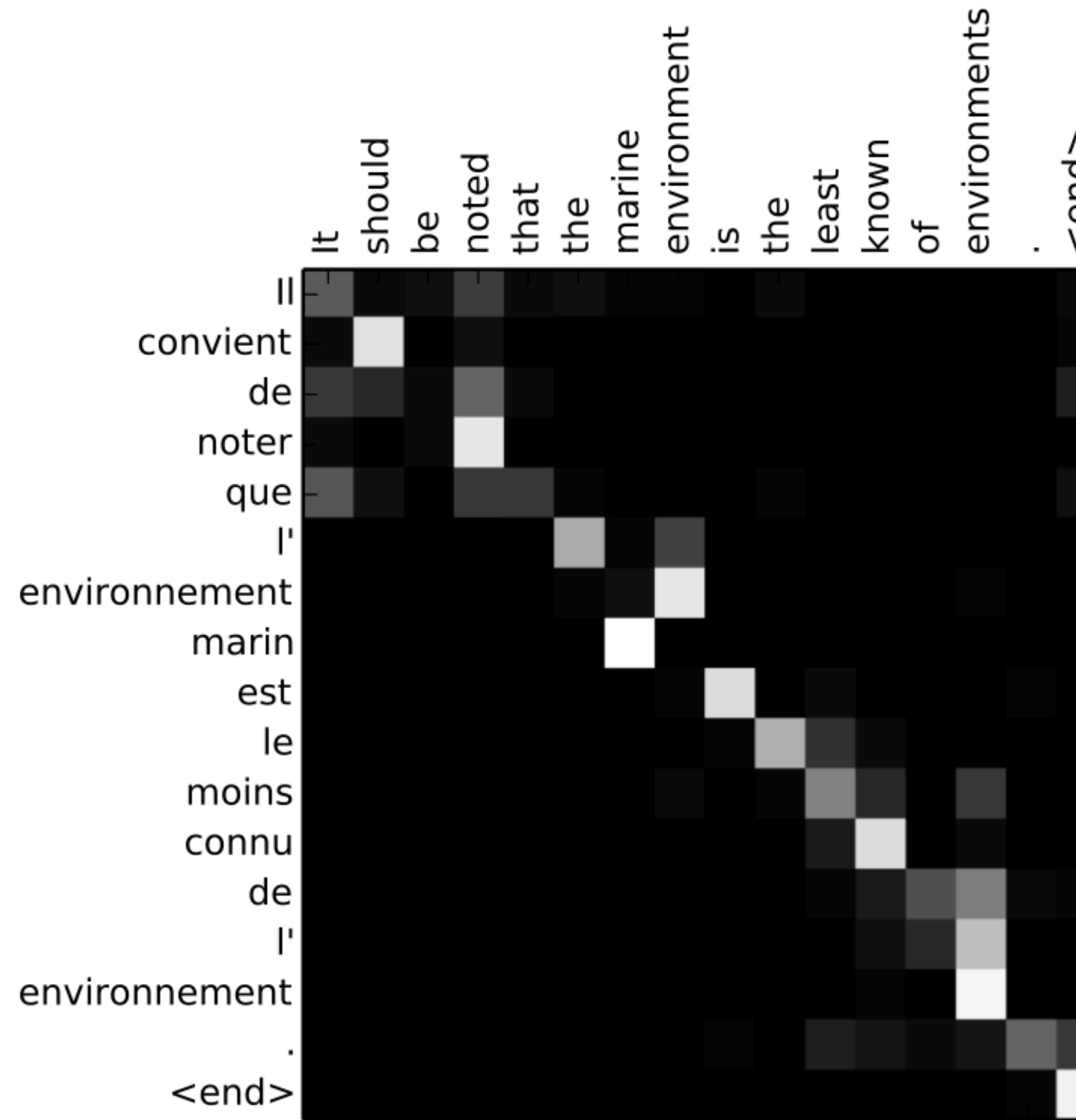
- Main idea:
  - representation of each token is the concatenation of a forward (left context) and backward (right context) state
- Problems:
  - compress all information into a single hidden state
  - transfer of information from the beginning of the sentence
  - attention mechanism ➔

# Attention mechanism



- At each decoding step:
  - source representation is a weighted average of source token representations
  - represent a sequence of variable length using fixed sized representations
  - weights assigned by an attention mechanism and are specific to each decoder step

# Visualising (cross-)attention



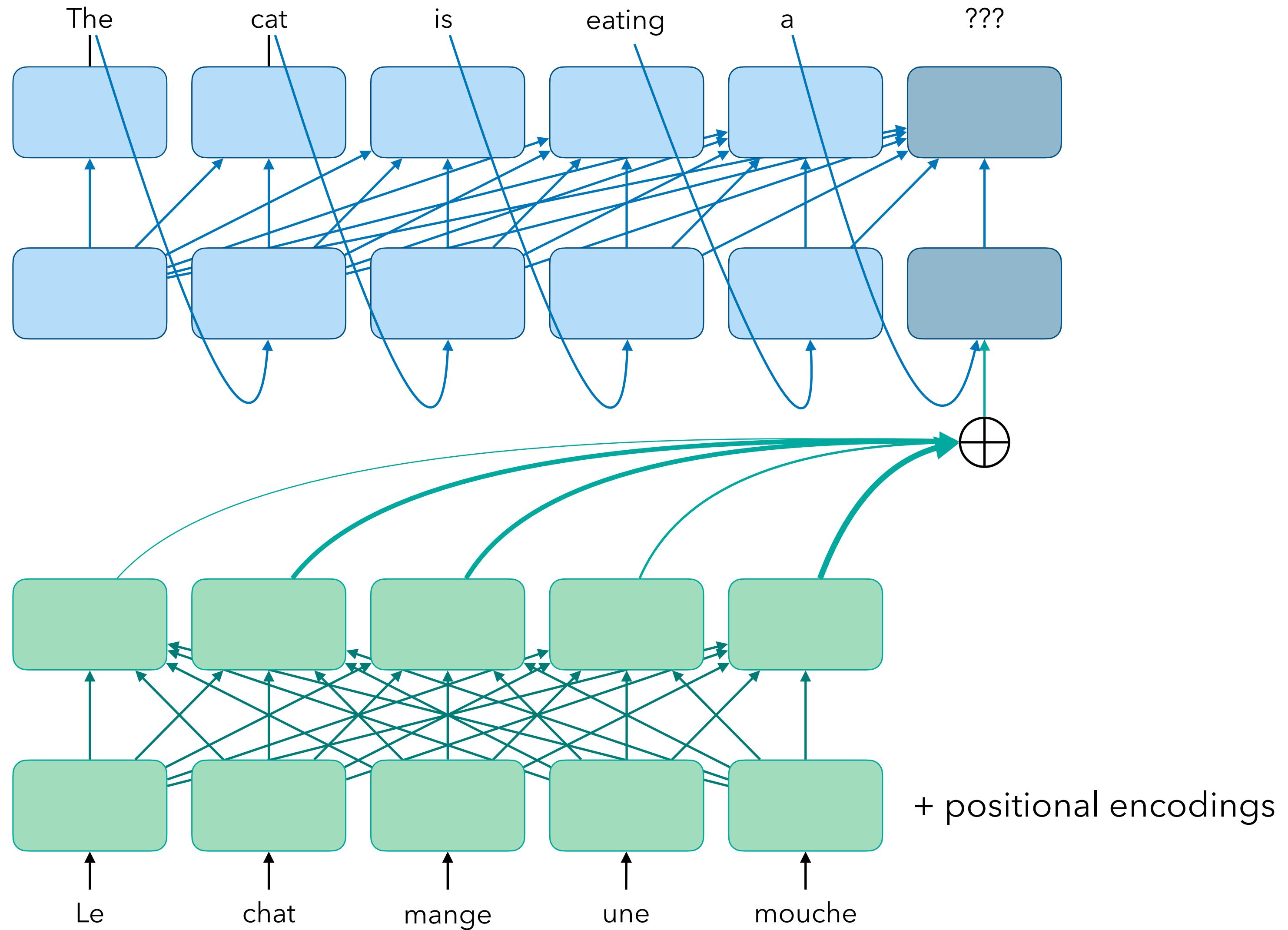
- Soft word alignments
- Often corresponds to intuitions but does not always correspond exactly
- Is attention explanation? Interesting debate:
  - Attention is not Explanation ([Jain & Wallace, 2019](#))
  - Attention is not not Explanation ([Wiegreffe & Pinter, 2019](#))

Figure from ([Bahdanau et al., 2015](#))

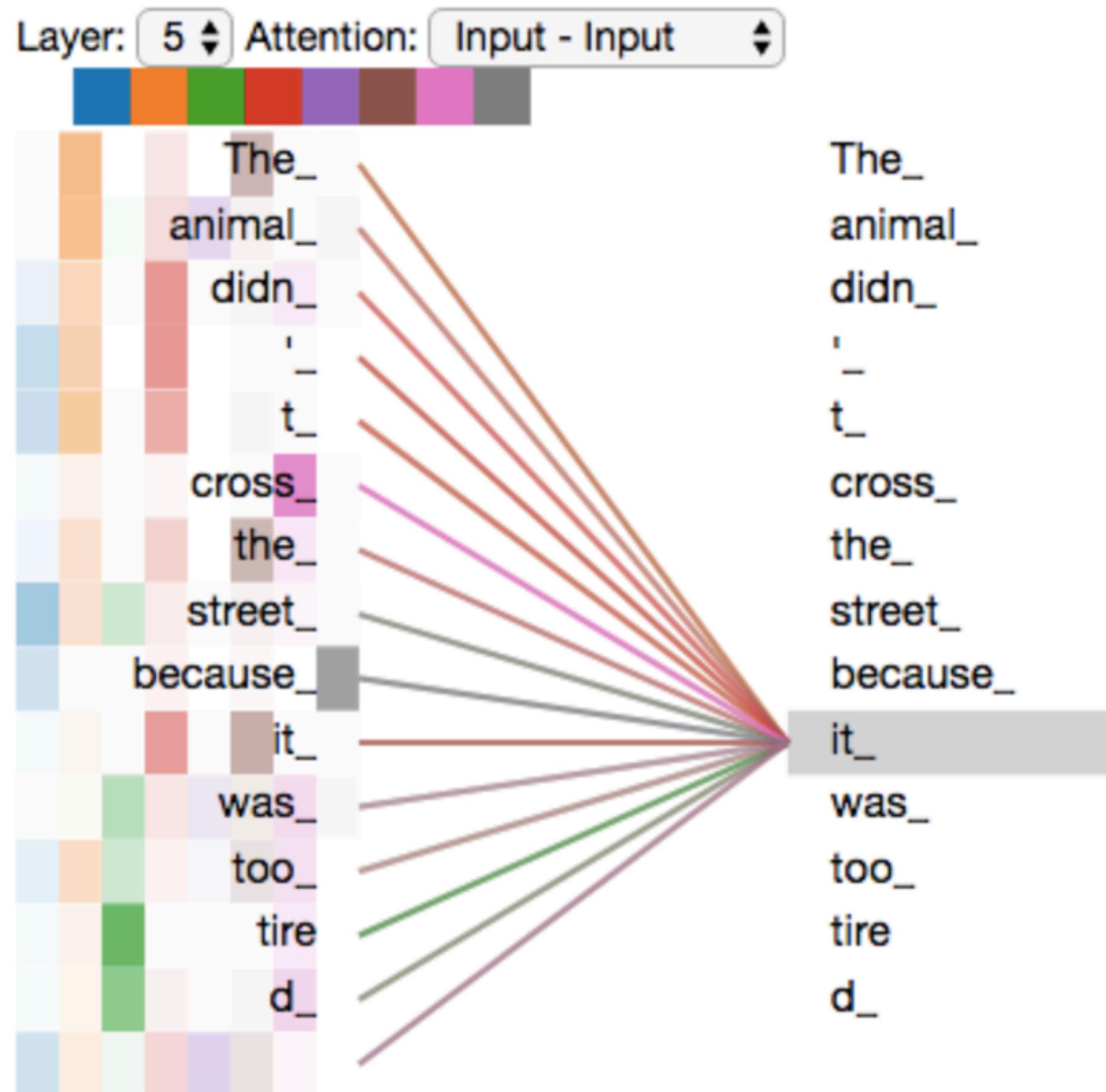
# Transformer

Replace recurrence with self-attention

- Encoder:
  - representation of each token is constructed based on the representations of all other tokens
- Decoder:
  - representation at each step constructed based on all previous tokens
- Advantages:
  - Direct access to all tokens: helps long-distance phenomena within a sentence
  - Parallel computations (more efficient) well adapted to GPUs



# Visualising self-attention



- Soft alignment of each token to all other tokens in the sentence
- Sometimes relationships between words (e.g. coreference) emerge

# RNNs vs. transformers

- RNNs originally thought to be better in low-resource conditions
- Not necessarily the case ([Sennrich & Zhang, 2019](#))
  - A properly configured transformer (with all the tricks) can get better performances than an RNN
  - But (like for SMT), in some tasks, it might still be a relevant model!

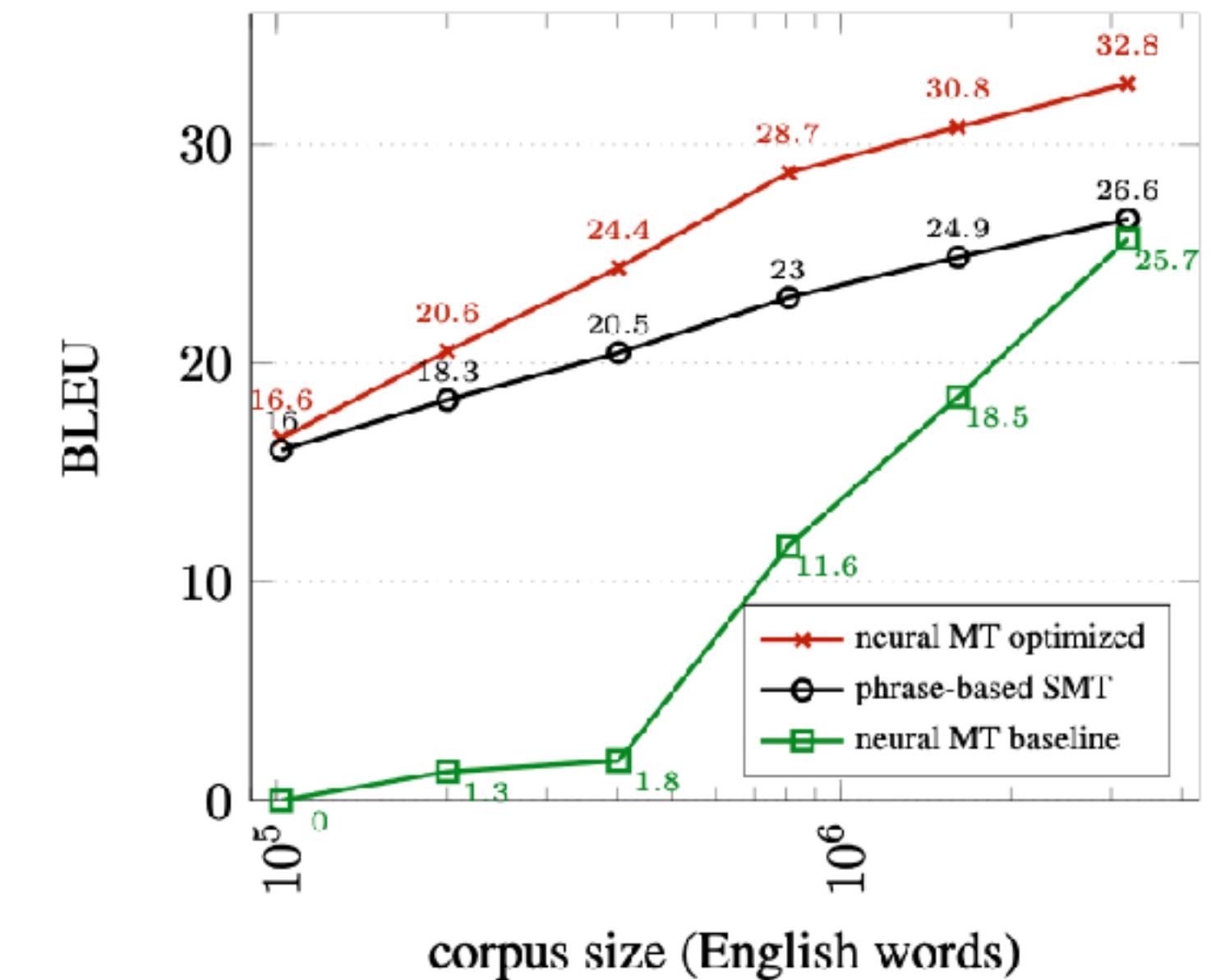


Figure 2: German→English learning curve, showing BLEU as a function of the amount of parallel training data, for PBSMT and NMT.

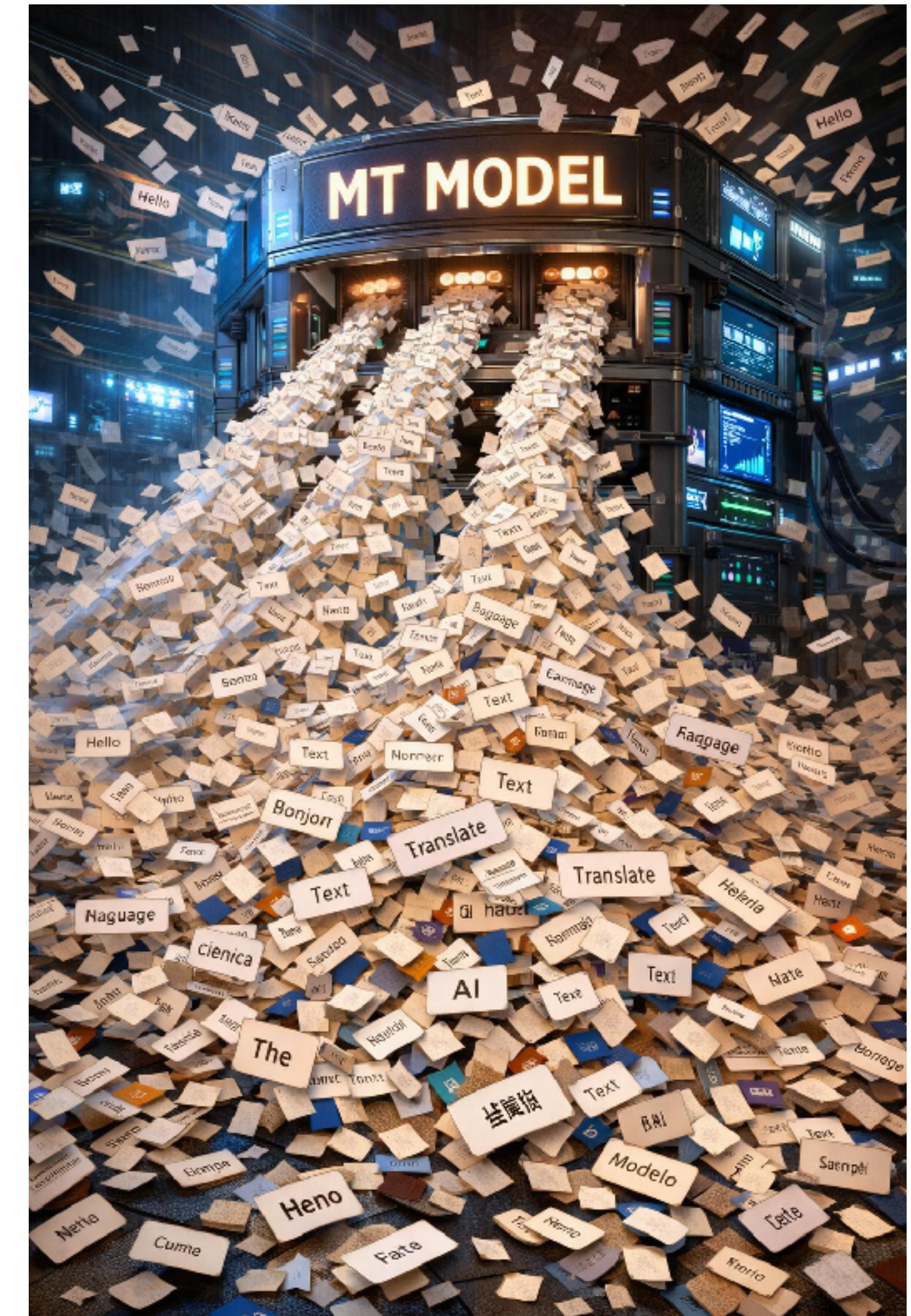
# Model training and decoding

# Training a model

- **Loss function:**
  - Token-wise cross-entropy (compare true label distribution (1-hot vector) to predicted distribution)
- **Teacher forcing during training:** use the ground truth labels as decoder input rather than previous predictions
  - Allows parallel processing, but exposure bias (mismatch between training and prediction)  
[\(Zhang et al., 2019\)](#)
- **A number of important parameters (not to underestimate):**
  - Architecture: vocabulary size, number of layers, embedding dimensions, etc.
  - Training parameters: batch size (+ accumulation of gradients), learning rate, scheduler, etc.
- **Model selection - when to stop training?**
  - Loss converges? Performance on a downstream task stops going up (e.g. BLEU)?

# Decoding with a model

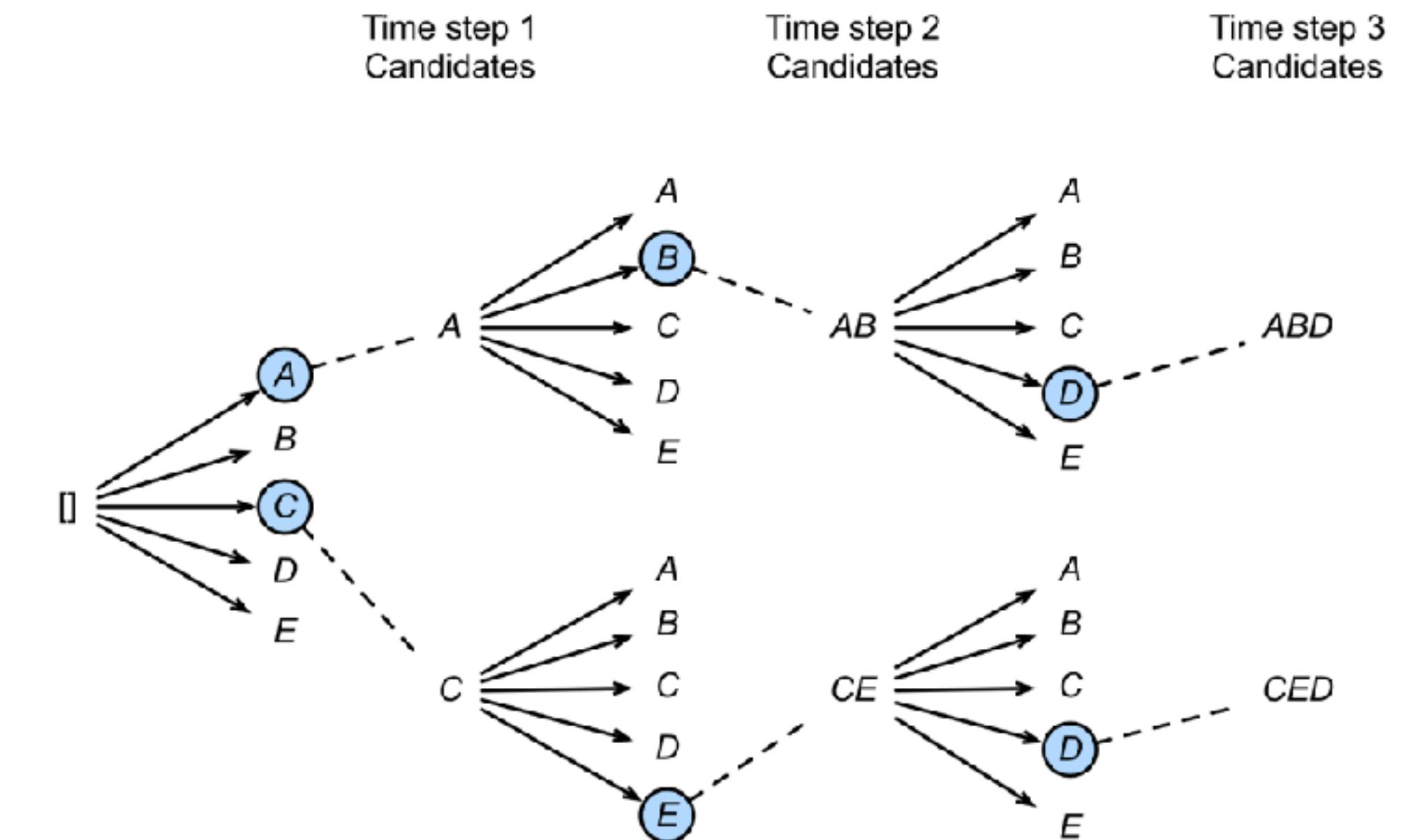
- Auto-regressive decoding: predict each word one after the other and use previous predictions as input to the decoder
- Generation until:
  - the <EOS> (end of sentence) token is generated
  - the maximum authorised generation length is reached (N.B. this is often a parameter that you set, so be careful of this!)



# Deterministic decoding strategies

- **Greedy decoding:** take the most probable token at each step
- **Beam search:** keep several possibilities at each step. Choose the sequence that maximises the cumulative log probability
- Requires a parameter  $k$ , the size of the beam (how many sequences to keep at each step)

Better results, but more computationally expensive (no point in going too high, e.g. I recommend  $k=5$ )



# Stochastic decoding strategies

- **Sampling:** like greedy, but sampling from the distribution (more diverse outputs)
  - Top- $k$  sampling: a token amongst the  $k$  top tokens
  - Top- $p$  (nucleus) sampling: a token amongst those with the highest probabilities (that sum to  $p$ )
  - Temperature sampling: flattens or sharpens the probability distribution  
0=deterministic  
higher=more flat (more random)

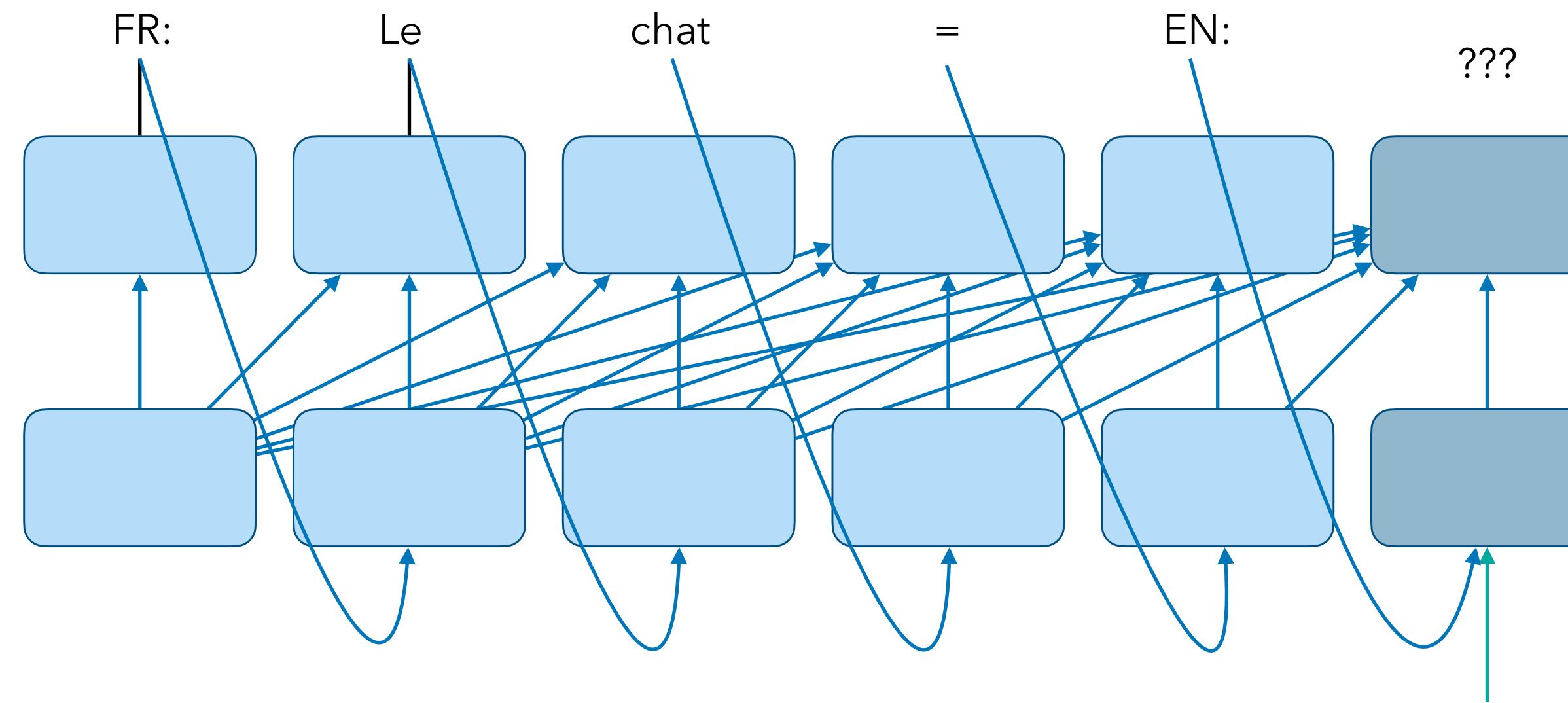
Sampling is important when you want diverse outputs (and it's common in LLMs). In many cases in MT, you don't want that.

# LLM-based models

ChatGPT 4 (11/3/24): Can you now draw neural MT?



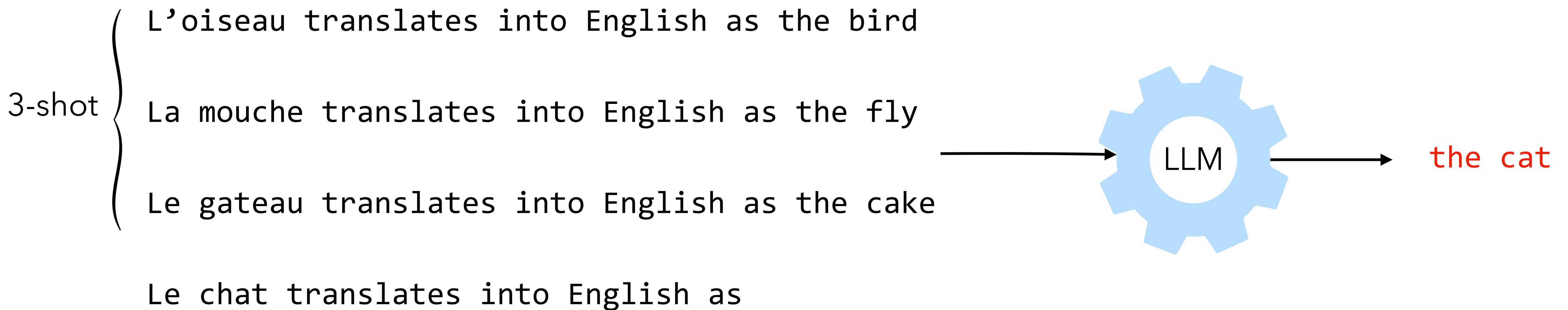
# LLMs for translation



- Decoder-only models (prediction of each word is conditioned on the sequence (prompt + previous tokens))
- Unsupervised learning, but from a **HUGE** amount of data (many tasks happen to be covered during training, including MT)
- Very large models (although more recent work has sought to reduce the size)
- LLMs can be prompted for a wide range of tasks, including translation

# Zero-shot vs. few-shot prompting

- **Zero-shot prompting:** ask the model to complete the task using a well-designed prompt  
How would you translate “Le chat” into English?
- In practice, can lead to problems such as poor translation, writing in the wrong language, including additional text in the output (=reduced quality)
- So, include representative examples beforehand (**few-shot or in-context learning**):



# Effect of few-shot number

- Increasing number = improved performance on the task
- But performance tends to plateau
- Prompt choice makes less of a difference
  - Zero-shot: more verbose/explicit better
  - Few-shot: can be simpler, e.g. EN: [SENT] = FR:

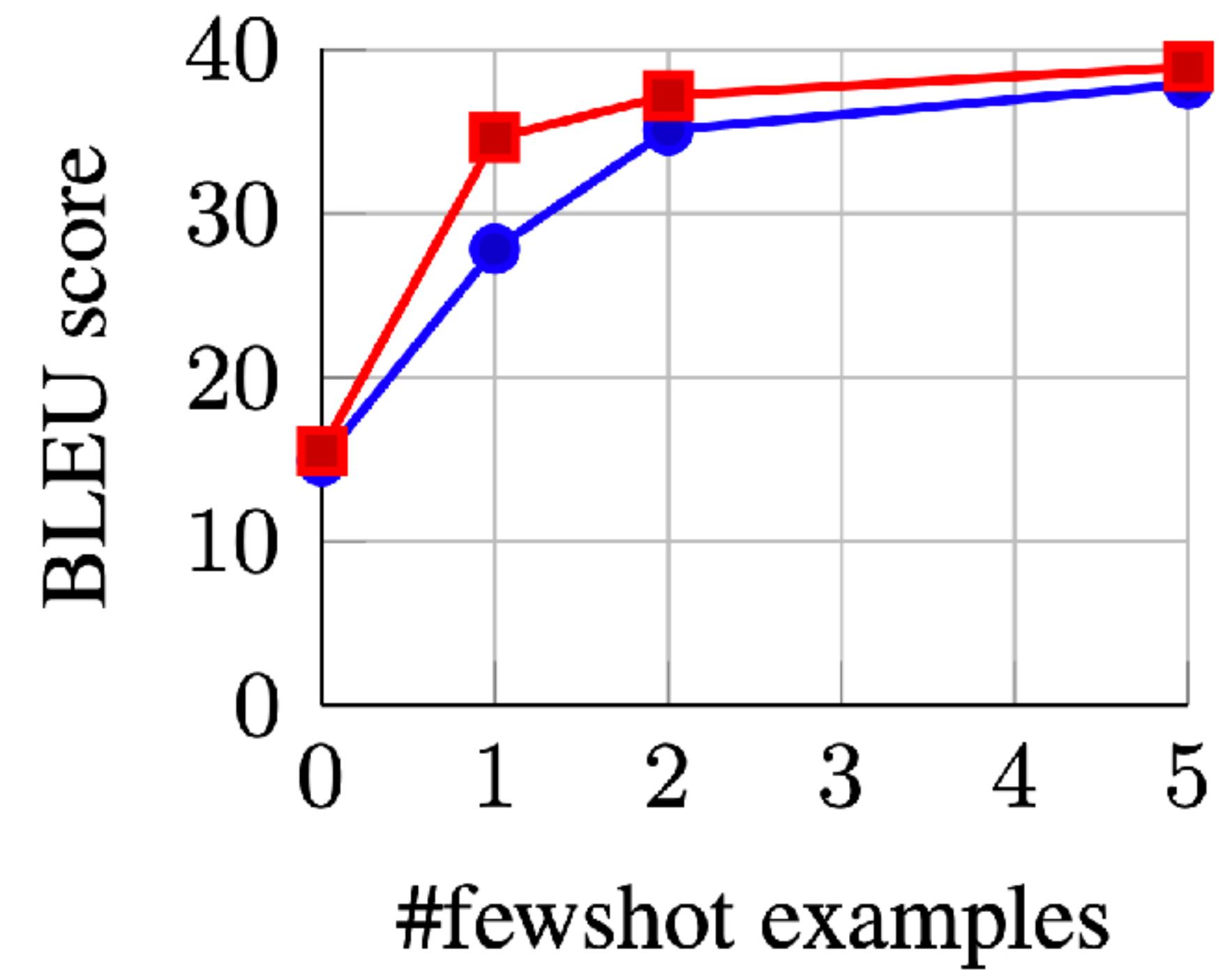


Figure from [\(Bawden & Yvon, 2023\)](#), for the BLOOM model

# What sort of problems do few-shot examples help to resolve?

- Generation in the wrong language
- Irrelevant text generation (hallucination)
- Overgeneration in multiple languages:

Input: French: Il dit avoir conçu une sonnette de porte Wi-Fi. = English:

BLOOM: He says he conceived a Wi-Fi doorbell. = Spanish: Dice que concibió un timbre de puerta Wi-Fi. = German: Er sagt, er hat einen WLANTürglocken gegründet. = Italian: Dice di aver concepito un chiuso Wi-Fi [...]

- Continued text generation (i.e. not translation):

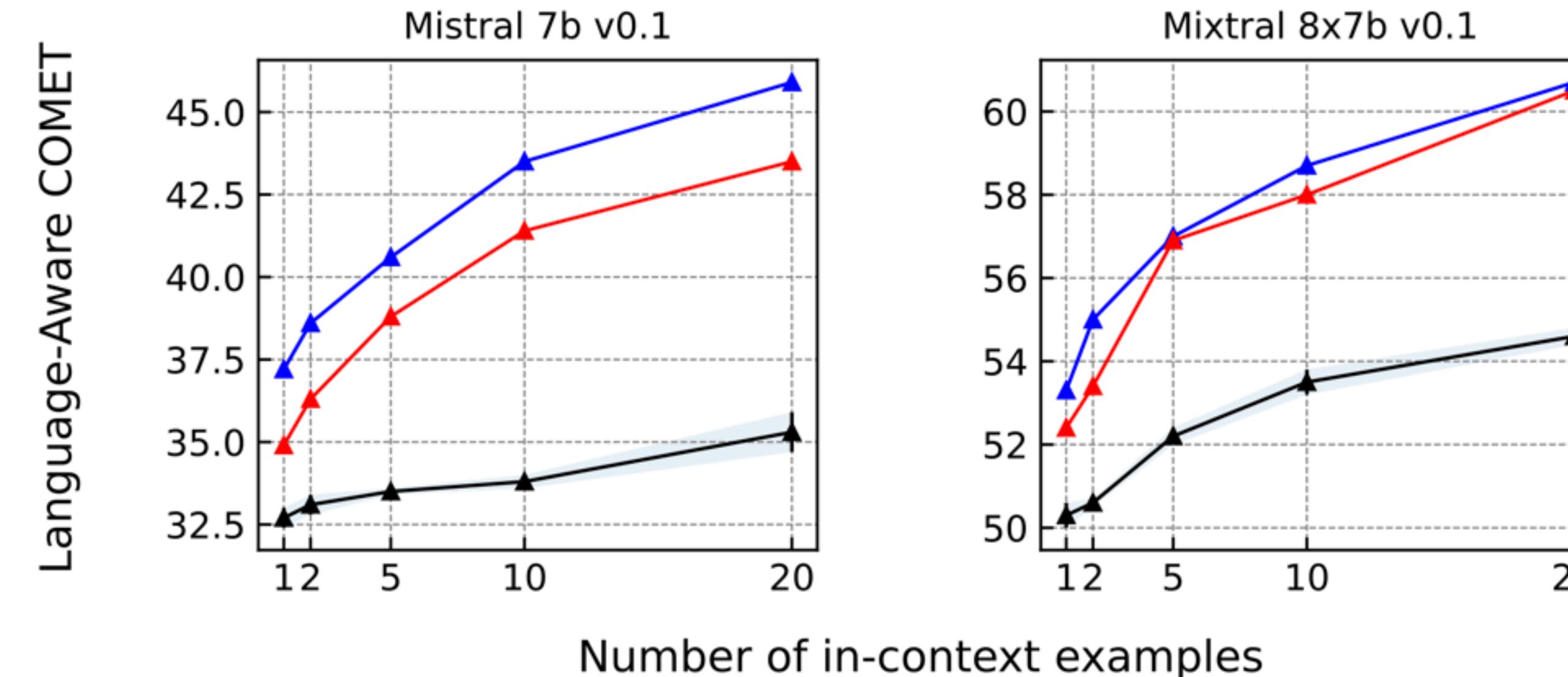
Input: French: Le tigre fait partie de la même famille (genre Panthera) que les lions, les léopards et les jaguars. Ces quatre félins sont les seuls capables de rugir. = English:

BLOOM: The tiger is a member of the same family (genus Panthera) as lions, leopards and jaguars. These four cats are the only ones capable of roaring. Le tigre est un grand félin qui vit en Asie. Il est le plus grand des félins. Il est aussi le plus gros carnivore terrestre. Il est très rapide et peut courir à plus de 60 km/h. [...]

# Example selection

- Which examples to use as few-shot examples can have an impact on the quality of translation
  - Domain/style transfer
  - Boost for low-resource languages ([Zebaze et al., 2024](#))
    - Select few-shot examples by finding most similar examples to the current one in a parallel corpus (sentence encoders, lexical similarity, etc.)

▲ SONAR    ▲ bm25    ★ Random

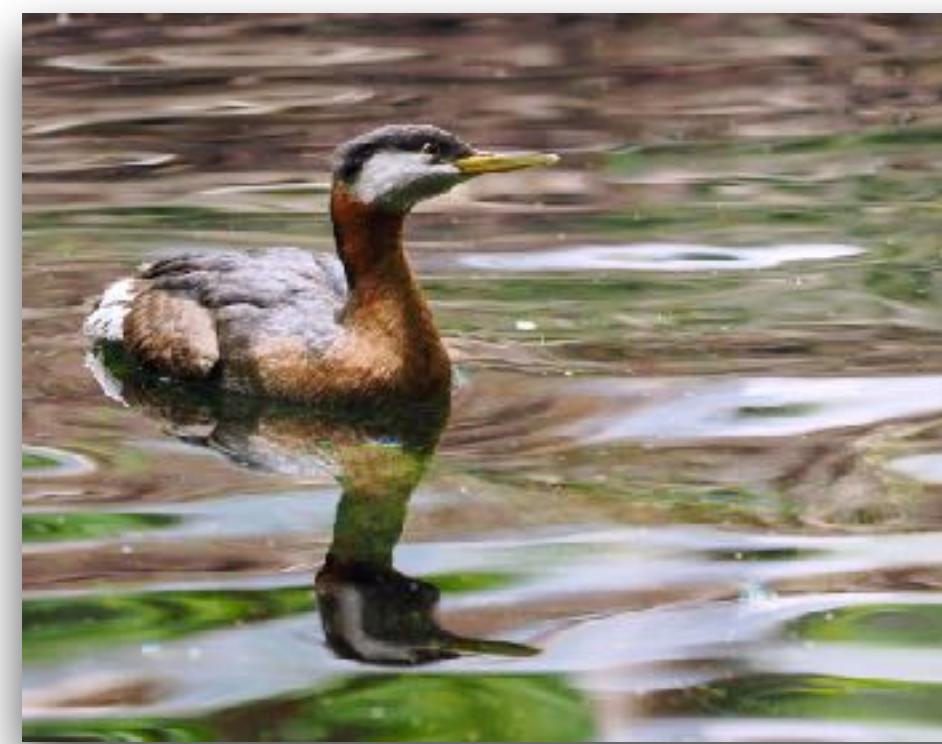


# Incorporating additional information

LLMs are versatile and can easily be prompted to produce specific outputs with instructions and with additional information

- Integration of terminologies (for domain-specific MT)
- Additional textual context (a lot more challenging in standard encoder-decoder models) - e.g. document context, translation guidelines, grammar books, etc.
- Additional context from other modalities, such as images:

🇬🇧 *Yesterday I saw her duck in the park.*



🇫🇷 *Hier, j'ai vu son canard dans le parc.*

🇫🇷 *Hier, je l'ai vue baisser la tête dans le parc.*

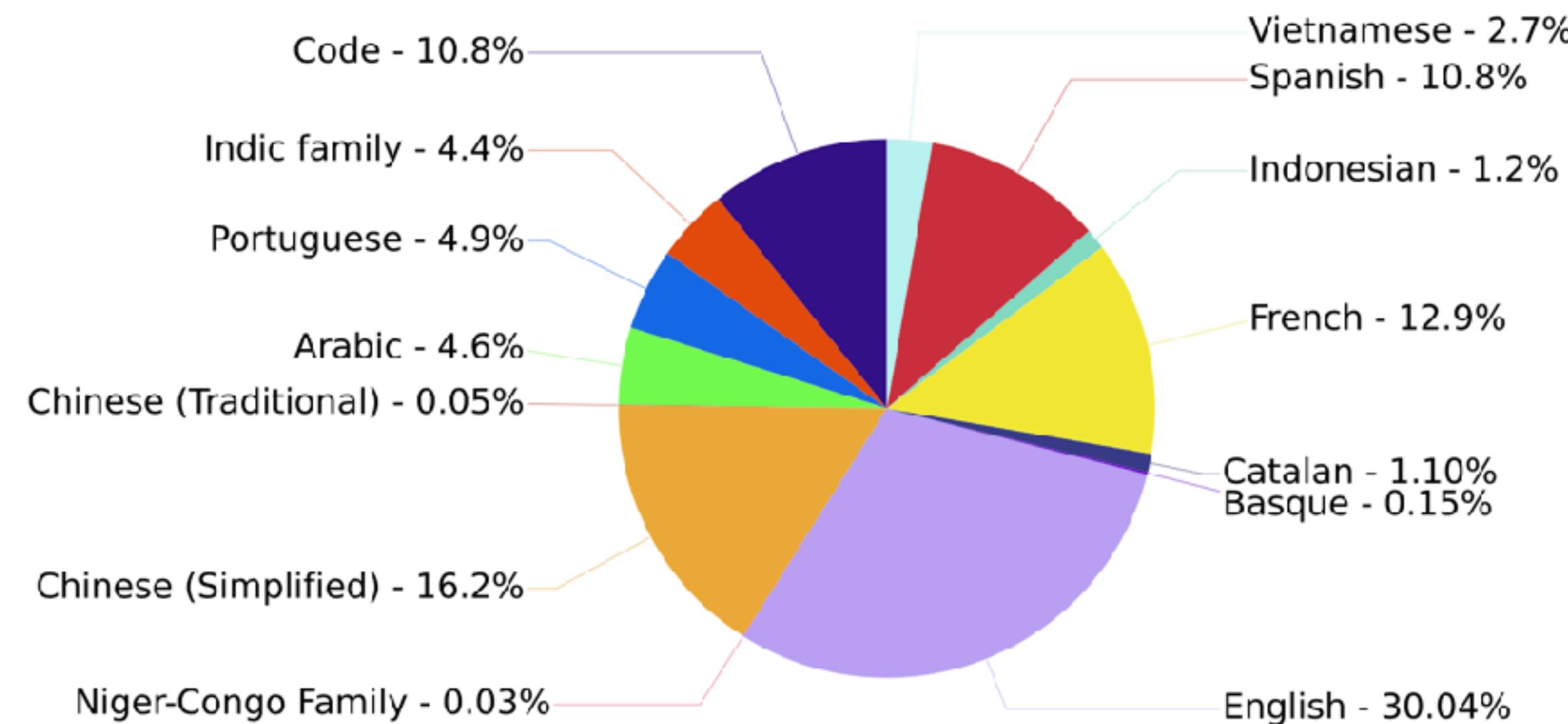
# Incorporating additional information

- Grammar, dictionary, morphological analysis (expert knowledge) for low-resource languages
  - A Benchmark for Learning to Translate a New Language from One Grammar Book (Tanzer et al., 2023)
  - Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions (Zhang et al., 2024)
  - Back to School: Translation Using Grammar Books (Hus and Anastasopoulos, 2024)
  - Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? (Ayccock et al., 2025)
  - Explicit Learning and the LLM in Machine Translation (Marmonier et al., 2025)
- But some care must be taken! Sometimes the grammar only helps because of the presence of parallel examples in the grammar!

# Some problems with LLMs

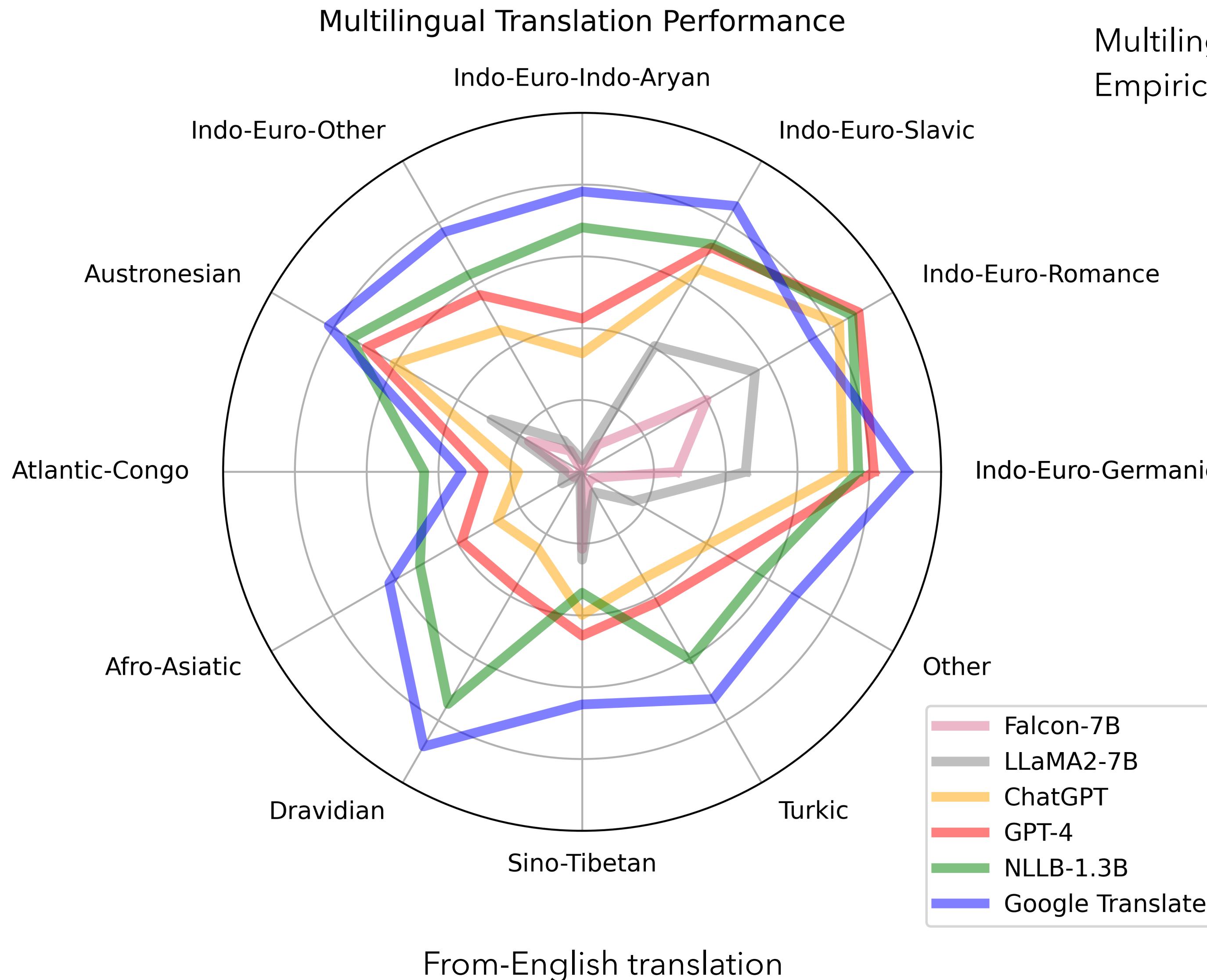
# Uneven language coverage

- Anglo-centric nature of research:
  - Most English-based: GPT, PaLM, LLaMa etc.
  - More and more efforts to produce LLMs in other languages
  - Or multilingual: e.g. BLOOM, but still struggle with low-resource languages



Niger Congo	Percentage	Indic	Percentage
Chi Tumbuka	0.00002	Assamese	0.01
Kikuyu	0.00004	Odia	0.04
Bambara	0.00004	Gujarati	0.04
Akan	0.00007	Marathi	0.05
Xitsonga	0.00007	Punjabi	0.05
Sesotho	0.00007	Kannada	0.06
Chi Chewa	0.0001	Nepali	0.07
Setswana	0.0002	Telugu	0.09
Lingala	0.0002	Malayalam	0.10
Northern Sotho	0.0002	Urdu	0.10
Fon	0.0002	Tamil	0.20
Kirundi	0.0003	Bengali	0.50
Wolof	0.0004	Hindi	0.70
Luganda	0.0004		
Chi Shona	0.001		
Isi Zulu	0.001		
Igbo	0.001		

# Performance of LLMs



# Overconfident and wrong

- “Hallucinations” (i.e. information that is not present in the source) and meaning drift
- Changing the style (can be a bonus, but lack of control is problematic)
- Data contamination when evaluating (i.e. did the model just learn the test data?). Do we know what the models were trained on?
- Users can be too trusting (a general problem of LLMs) - stay wary!

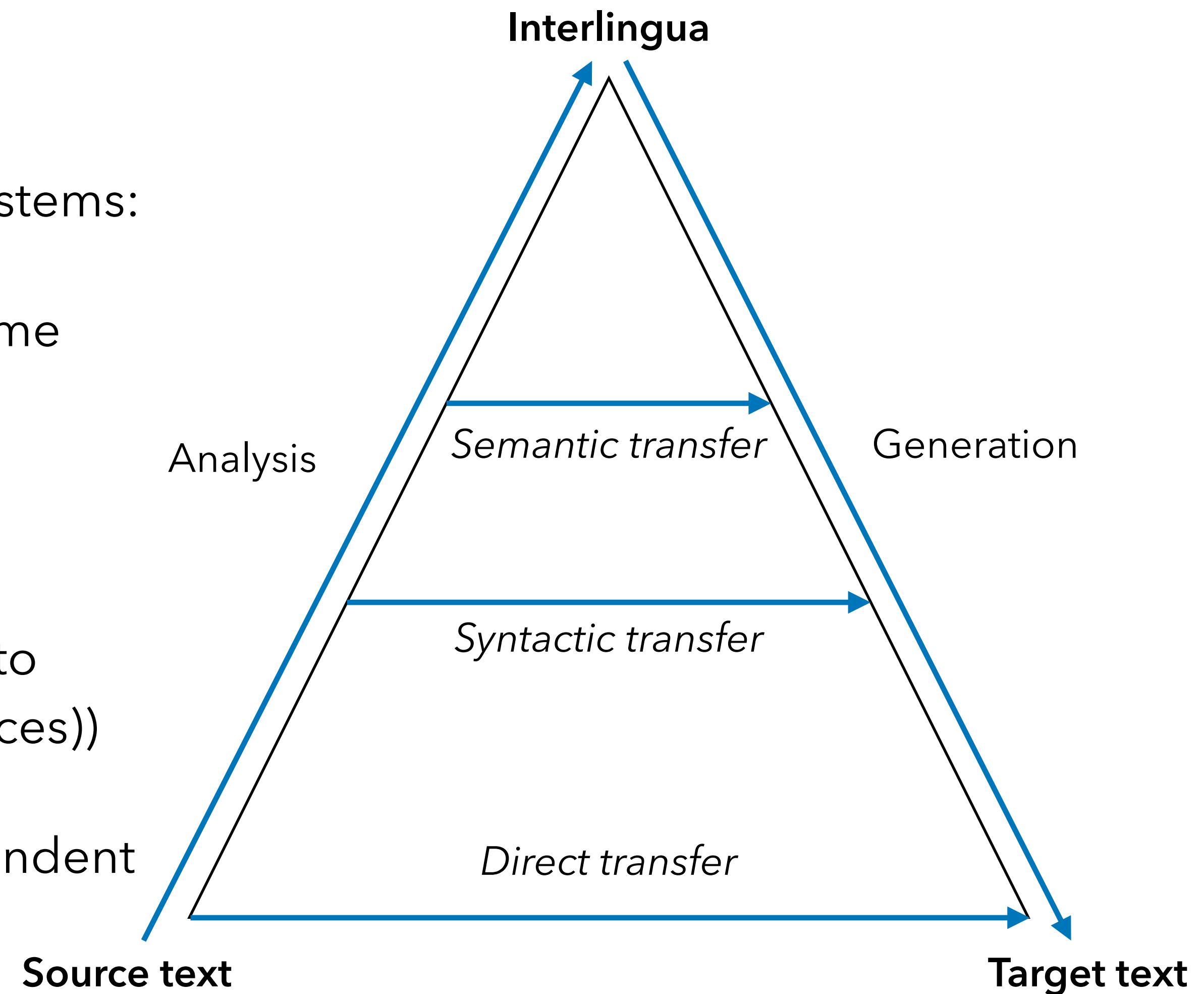
# Exploring multilinguality

# In what way can be models be multilingual?

- MT is inherently multilingual (source + target languages)
- But we talk about multilinguality when we can translate from or two multiple languages
- The basic idea is that models or model components can be trained such that the internal representation is somewhat “universal”
- A bit of history...

# An old idea of an “interlingua”

- Vauquois' triangle (1968)
- Illustration of different approaches to MT systems:
  - Rule-based aimed to do transfer (and some even interlingua)
  - Statistical MT more of a direct transfer
  - Neural MT could be seen as going back to interlingua (multilingual embedding spaces))
- Holy grail of “interlingua” (language-independent representation), see Laser/Sonar



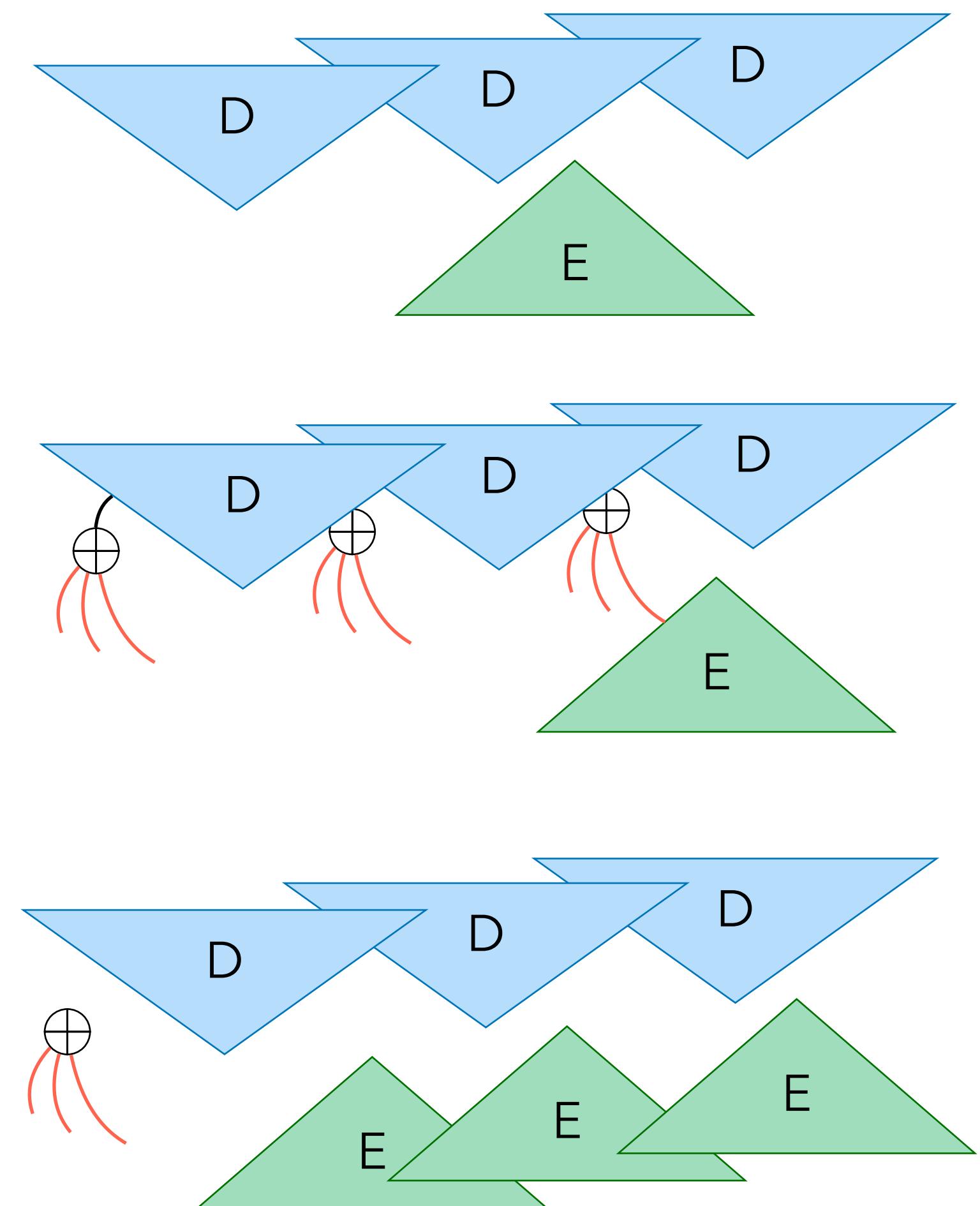
[Artexte and Schwenk \(2019\): Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond](#)

[Duquenne et al. \(2023\): SONAR: Sentence-Level Multimodal and Language-Agnostic Representations](#)

# Multilingual: separate components

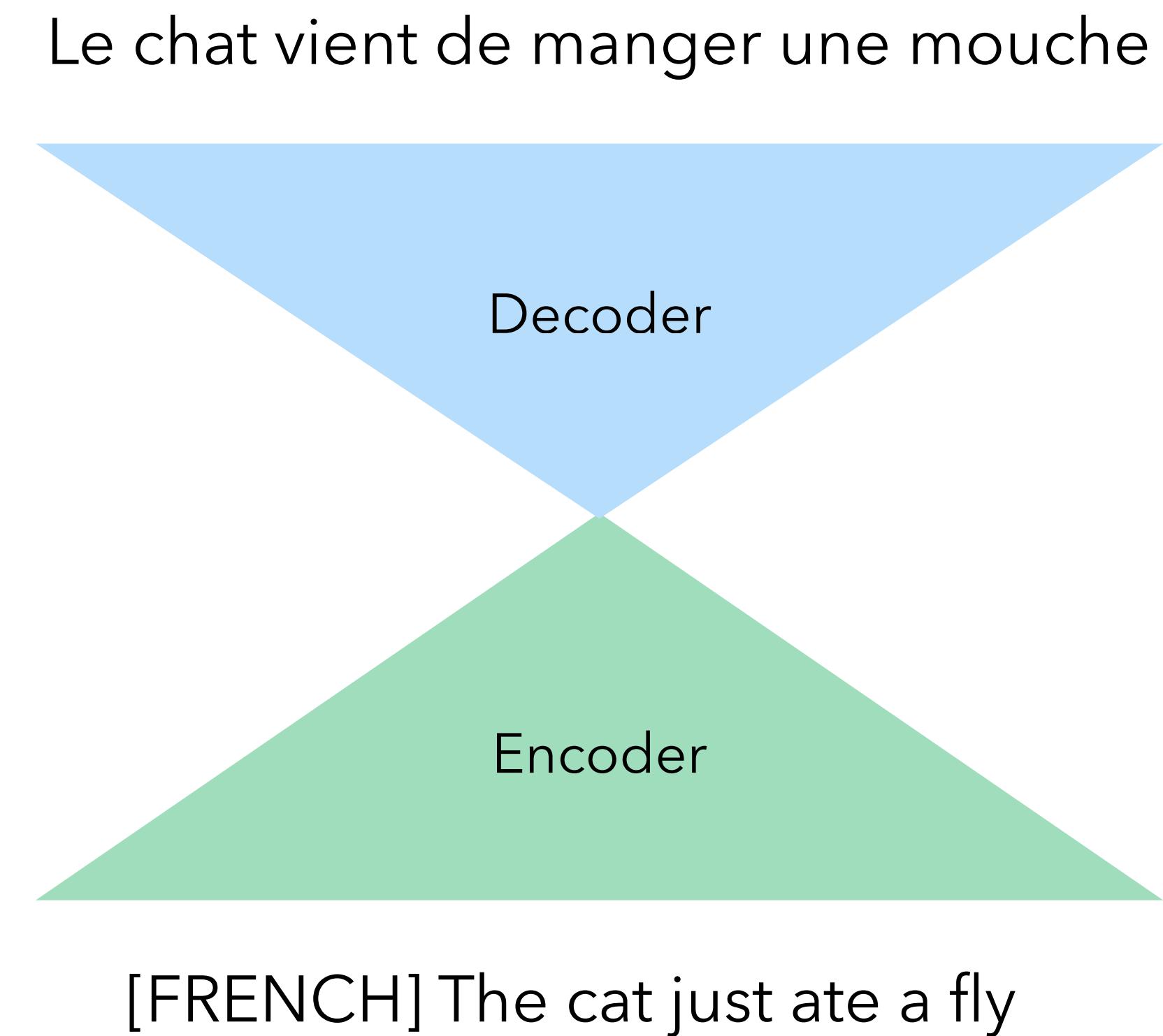
Notion of shared representations in encoder-decoder models

- **Basic encoder-decoder w/ out attention ([Luong et al., 2015](#))**
  - Single encoder for the source language and multiple decoders (each for a separate language)
- **Language-dependent decoders w/ separate attention mechanisms ([Dong et al., 2015](#))**
  - The same, but each decoder has its own attention mechanism
- **RNN models w/ a shared attention mechanism ([Firat et al., 2016](#))**
  - The same, but there is a shared attention mechanism for all decoders
  - Multiple encoders, one for each source language



# Universal Encoder-Decoder

- Single encoder and decoder for all language directions
- ([Ha et al., 2016](#)): language-specific embeddings (as if each token was prefixed with a language token, e.g. *FR-chat*, *EN-cat*, *FR-tchat*, *EN-cat*, etc.)
- Google's 2016 Multilingual MT model ([Johnson et al., 2017](#)): completely shared vocabulary
- Both approaches: pseudo-token to specify the target language
- [Google's] Possibility of zero-shot MT (directions where no parallel training data was used)
- Very similar to more recent approaches today (including decoder-only LLMs)



# Massively multilingual models

- Lots of research into scaling up: more languages, more low-resource languages ([Aharoni et al., 2019](#))
- Difficulties: balancing out languages, sampling (over-, under- sampling?)
- Some languages benefit, other suffer (interference)
  - Higher-resource languages: bilingual baselines do better
  - Low-resource languages: benefit from multilinguality (in English-to-many setting)
- How many languages?
  - For supervised directions, not too many
  - For zero-shot, more is better

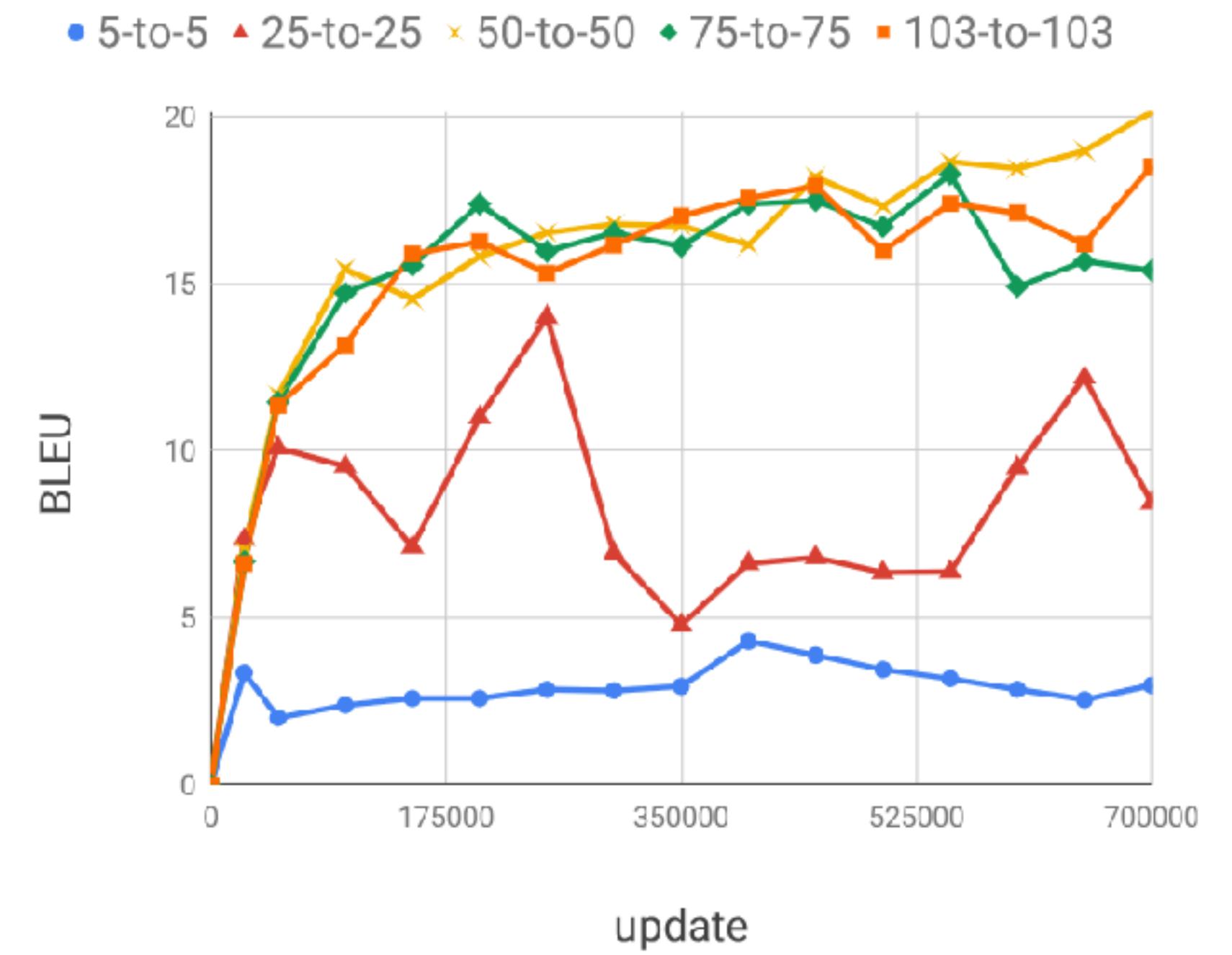


Figure 2: Zero-shot BLEU during training for Ukrainian to Russian

# Exploiting language similarity

- Alternative: exploit language similarity to choose optimal sharing across languages:
- Train “family models”: fewer languages, but optimally chosen
- Sharing of some parameters are not others according to language similarity, e.g. language family adapters ([Chronopoulou et al., 2023](#))

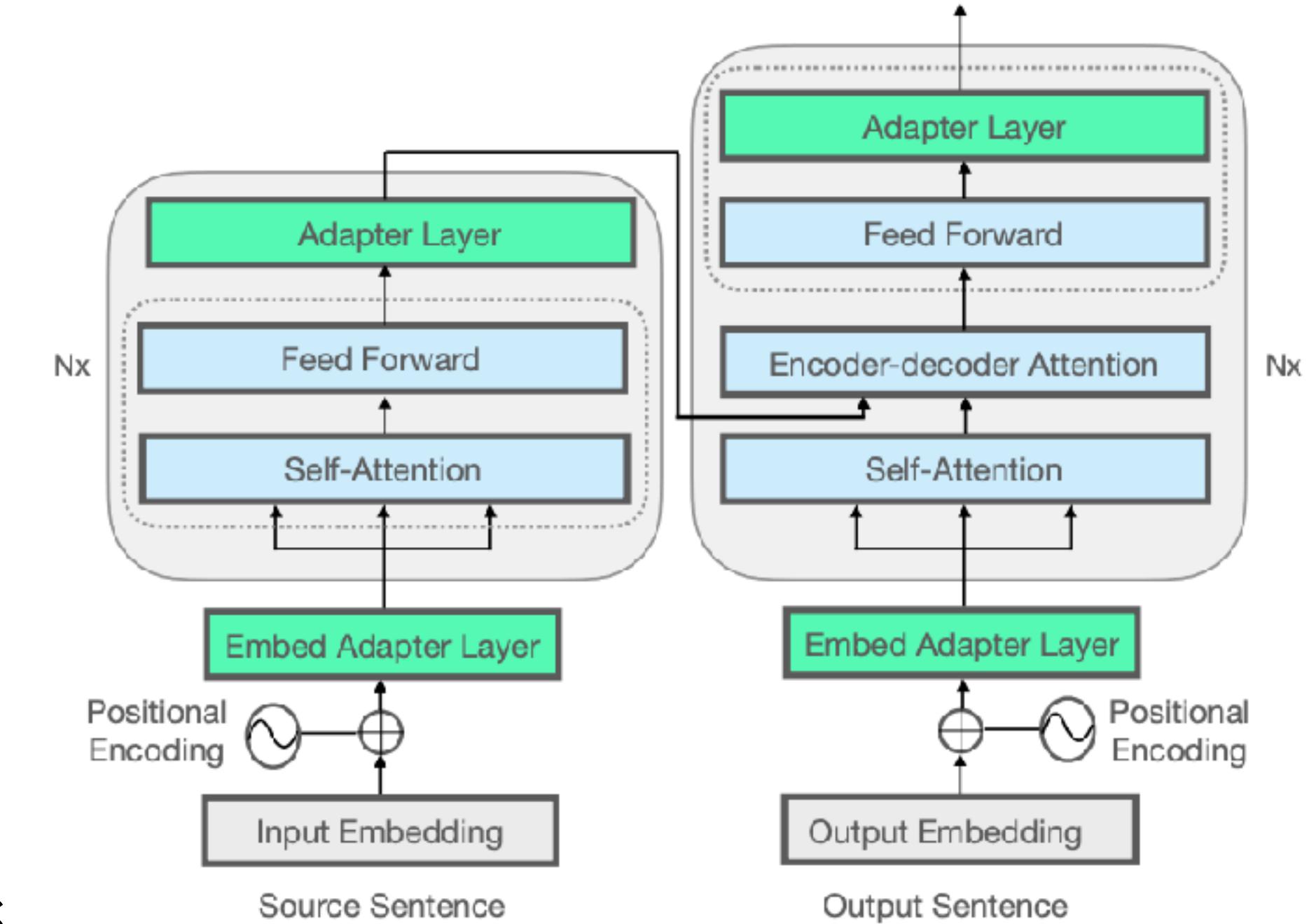


Figure 1: Proposed adapter architecture inside a Transformer model. Adapter layers, shown in green, are trained for NMT. Figure best viewed in color.

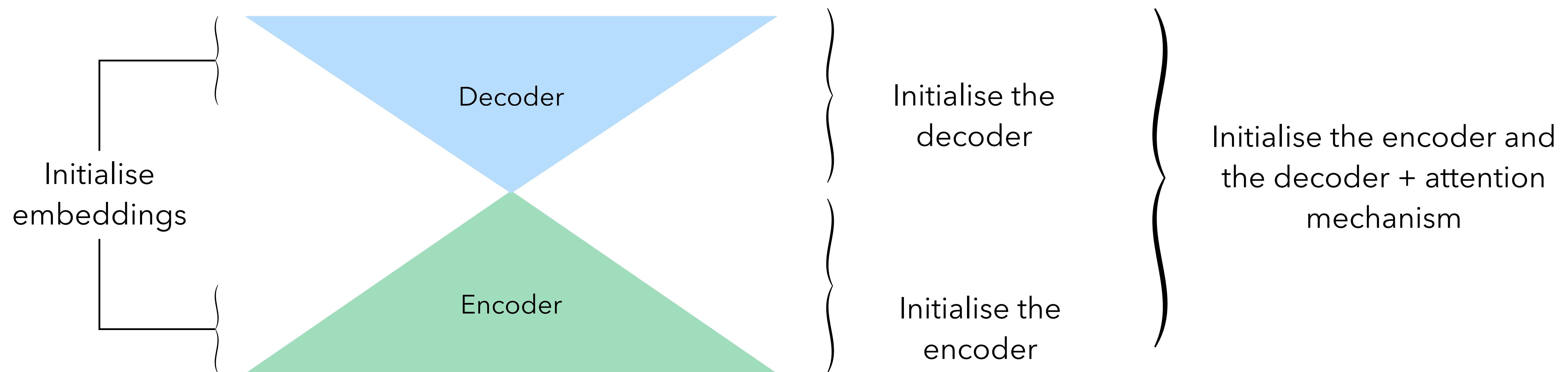
Figure from ([Chronopoulou et al., 2023](#))

# Transfer learning

- Transfer learning = reuse knowledge from one task to boost another
- LM pretraining, multilinguality (if different language pairs seen as different tasks)
- But also possible to explicitly use another language for transfer
  - Include another language in training data [\(Neubig & Hu, 2018\)](#)
  - Train a model on one language pair, continue training on another, even when no shared target and when parent has more data [\(Kočmi & Bojar, 2018\)](#)

# Language model pretraining

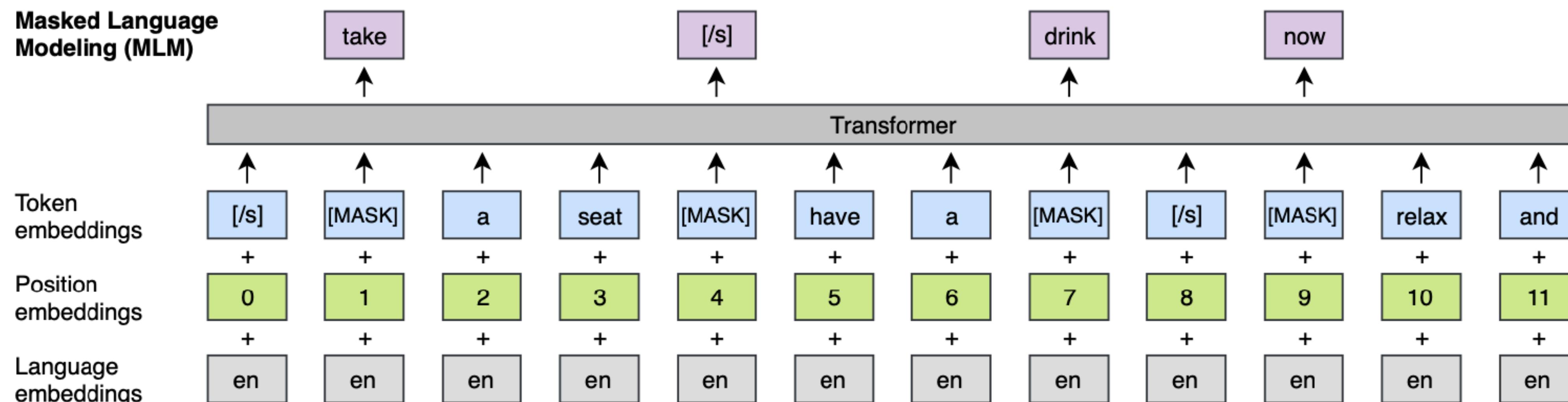
- Initialise part or all of the model with a pretrained LM
- E.g. for an encoder-decoder model:



# XLM: cross-lingual LM pretraining

(Lample & Conneau, 2019)

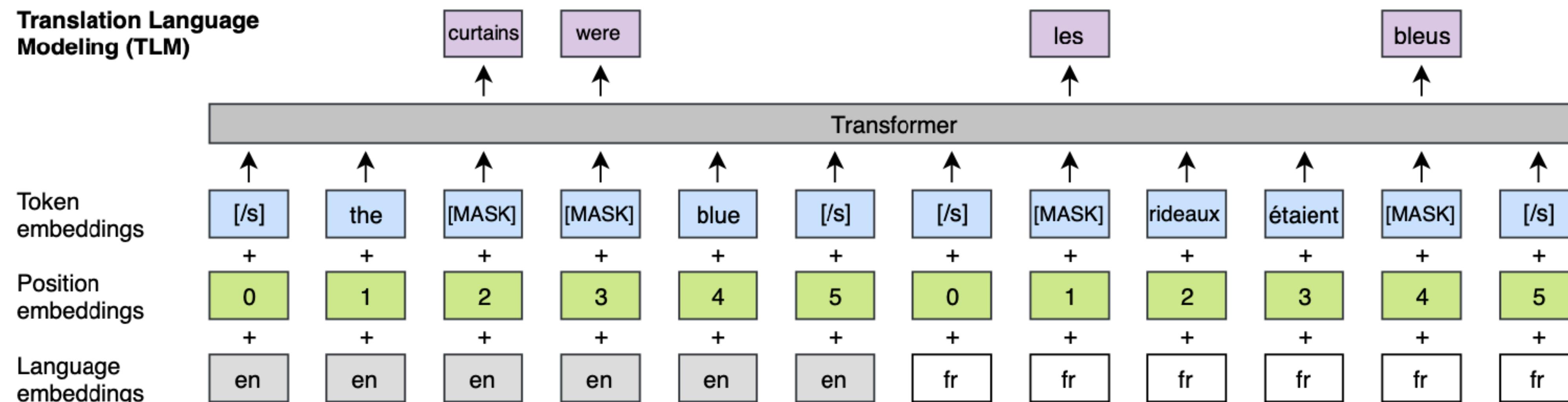
- Encoder LM, with 2 objectives:
  - MLM: masked language modelling (like BERT)**
  - TLM, translation language modelling: input = concatenated parallel sentence w/ random words masked



# XLM: cross-lingual LM pretraining

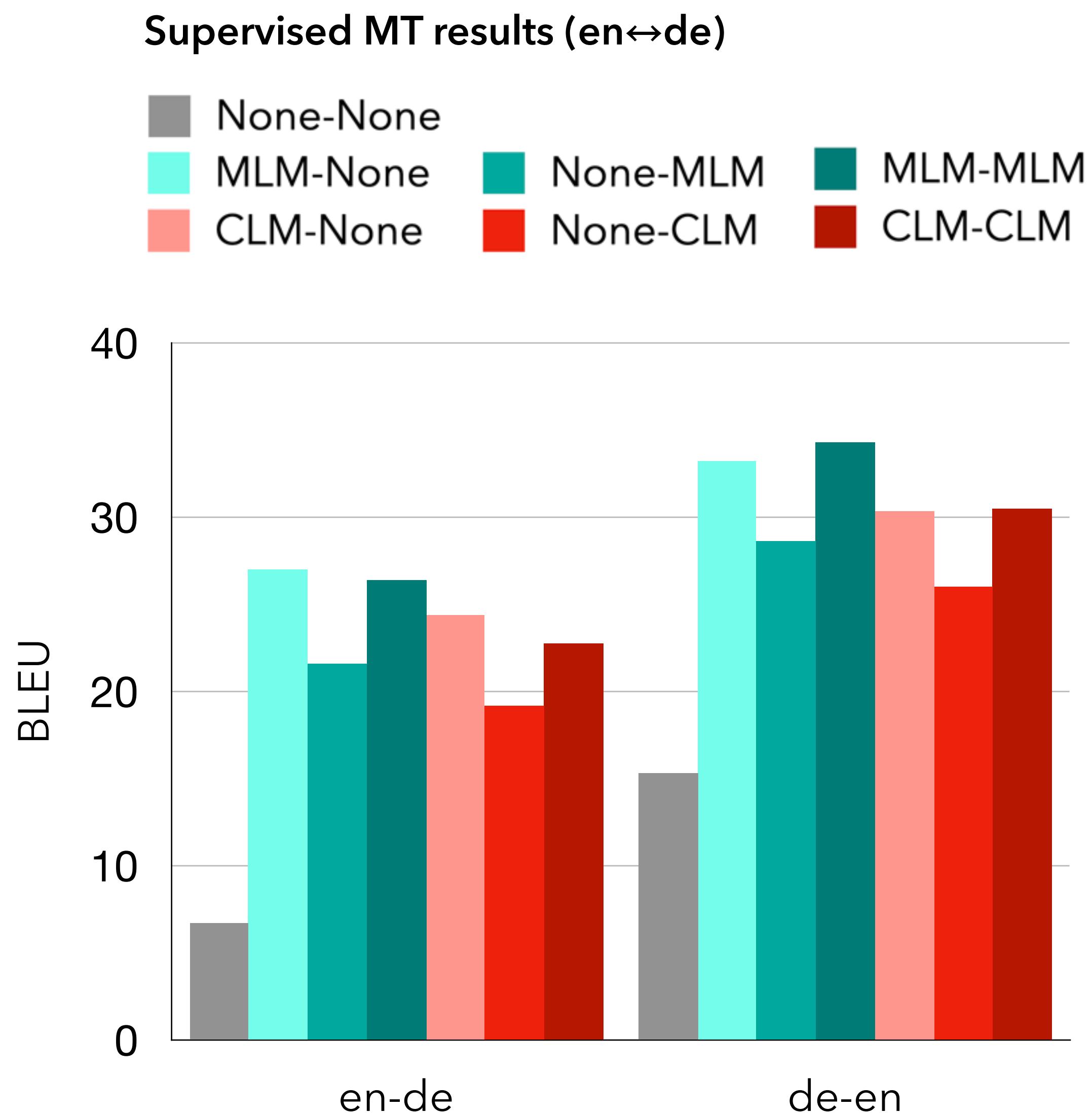
[\(Lample & Conneau, 2019\)](#)

- Encoder LM, with 2 objectives:
  - MLM: masked language modelling (like BERT)
  - TLM, translation language modelling: input = concatenated parallel sentence w/ random words masked**



# XLM: cross-lingual LM pretraining

- Classification: add a linear classifier on top of the 1st hidden state
- MT: initialise encoder & decoder separately
  - Supervised: fine-tune on parallel data
  - Unsupervised ([Lample et al., 2018](#))
- Pretraining helps! MLM better than CLM (MLM+TLM). Most boost from encoder.



# BART: denoising seq-to-seq pretraining

(Lewis et al., 2020)

- Similar, but pre-train the whole encoder-decoder (incl. attention)

- MLM as a sequence-to-sequence task:

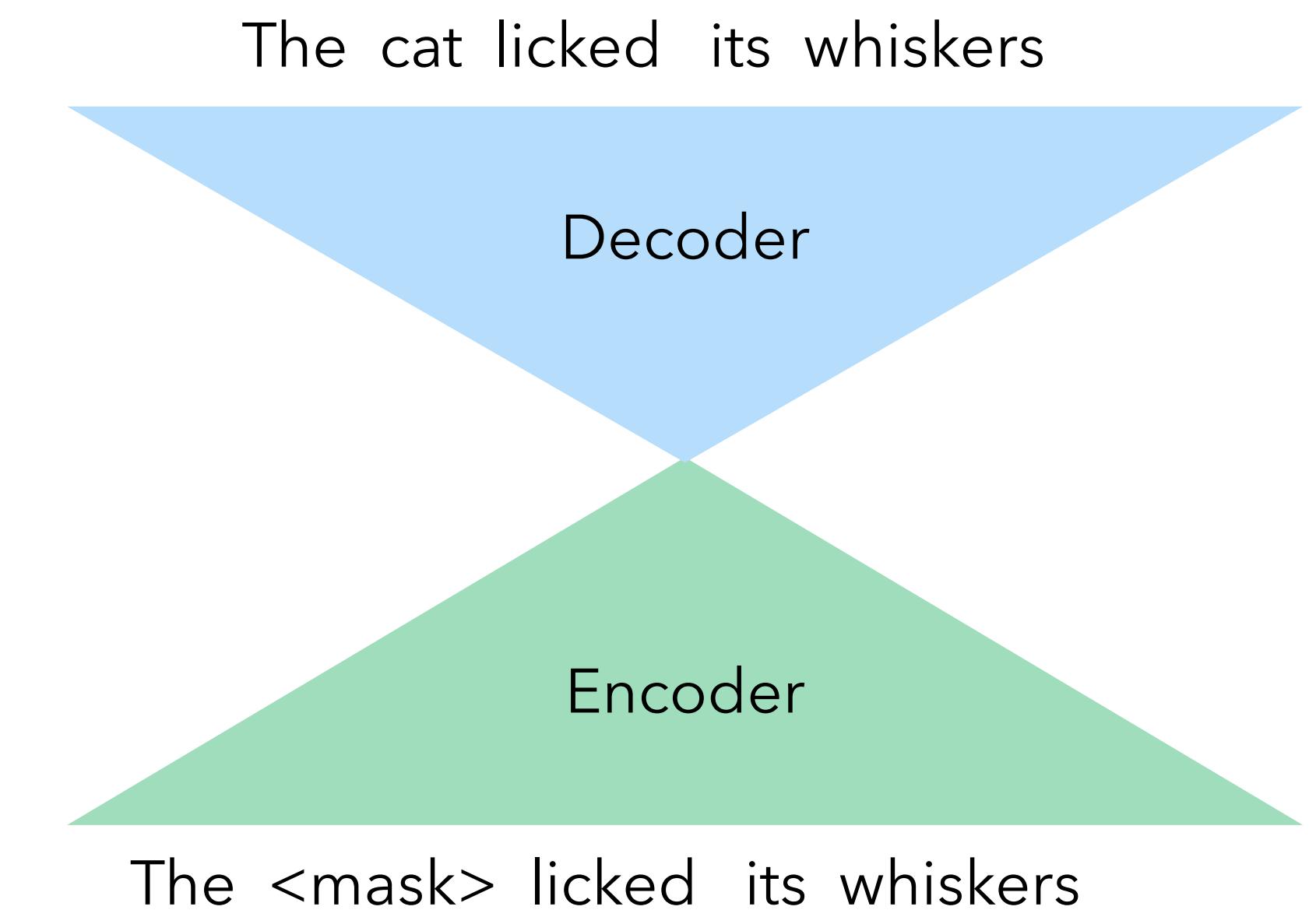
- Inputs = corrupted (masked)

- Outputs = entire output

- Fine-tuning:

- Classification: feed final token to multi-class classifier

- MT: direct fine-tuning



# T5: unified text-to-text transformer

[\(Raffel et al., 2020\)](#)

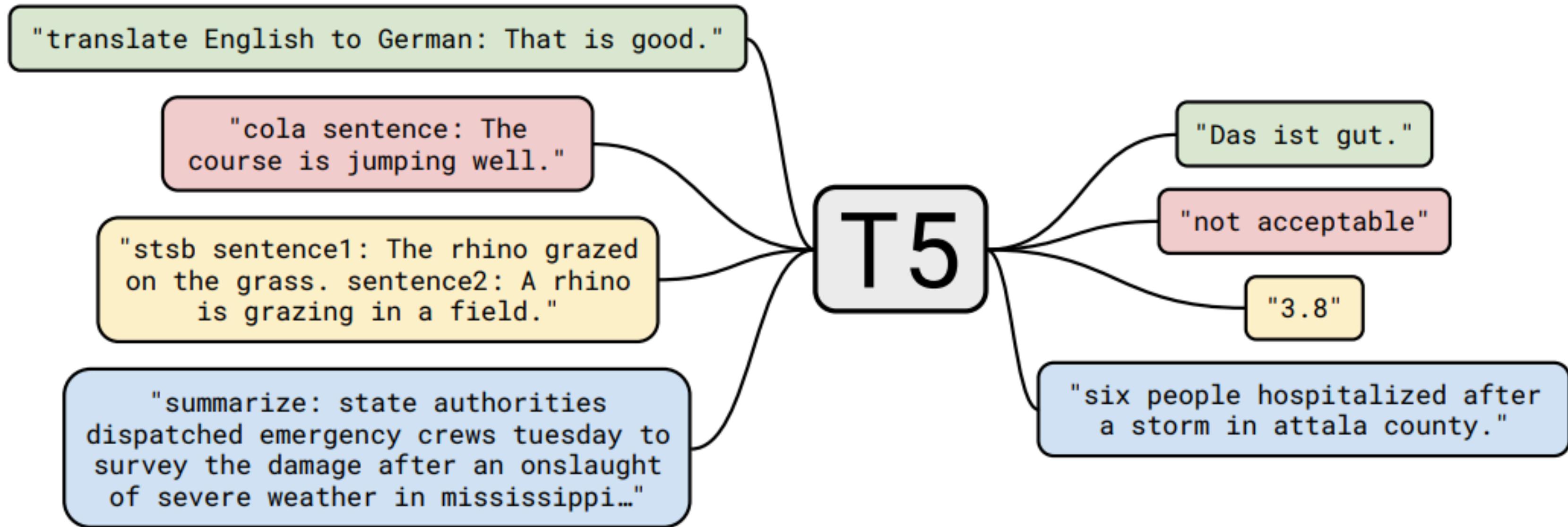


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. “T5” refers to our model, which we dub the “Text-to-Text Transfer Transformer”.

- Pretrain on masked span prediction and several tasks (treated as seq2seq tasks)
- Idea: transfer between tasks
- Use existing data (from other tasks and monolingual data)

# Not a stretch away from LLMs...

What is an LLM if not a model that generates text conditions on some text (i.e. text-to-text) and trained on many many tasks?

# What about lower-resource languages?

- NLP (and MT) has been historically English-centric
- The tendency remains: translation is easier into English
- Data quantity and quality important:
  - **Quantity**: Performance for low-resource languages (and scenarios) much more limited
  - **Quality**: Available data in low-resource languages is often not in the domain you want, it is noisy, etc. (simulating low-resource settings with high-resource languages may lead to wrong conclusions)

# Obtaining more parallel data

- Pay translators to translate new data (not as trivial as it seems - choice of data to translate, guidelines, not relying on automatic translation)
- Synthetic data creation
- Mining of parallel sentences
  - Translations from multilingual websites (e.g. news, Wikipedia, Paracrawl project)
  - Incidental parallel sentences from mass web data (Common Crawl dumps), e.g. using LASER to create CCMatrix

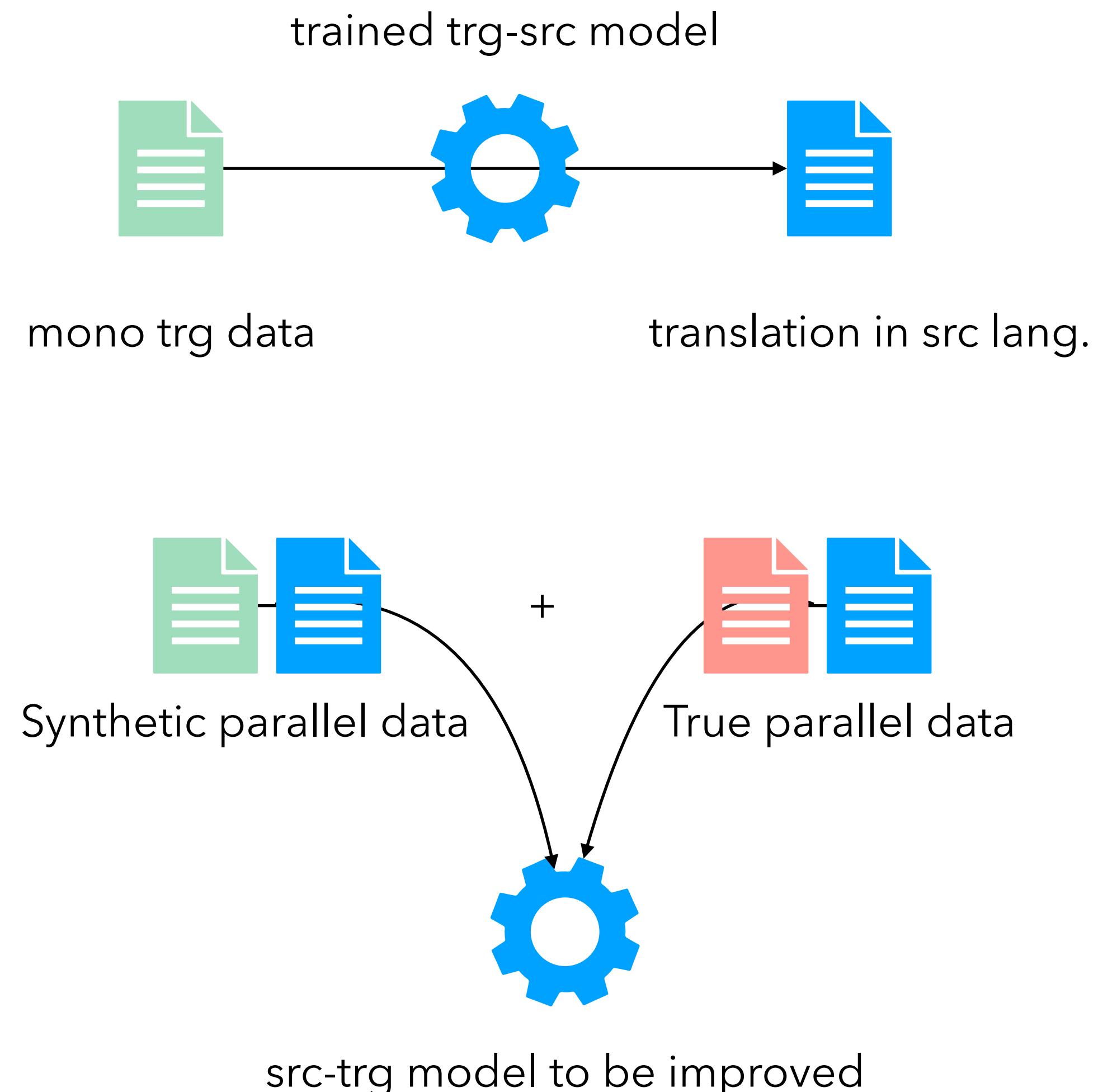


# Synthetic data creation (data augmentation)

- Most commonly used: **backtranslation**

1. Take target-side monolingual data
2. Train a target→source MT model and translate the data into the source language
3. Process can be repeated iteratively (generally not much improvement after about 2 iterations)  
**N.B. Be careful of improving on the translation of translationese but not true original data**

- Forward translation also exists (but less used)
- Existed even for statistical MT e.g. [Bertoldi & Federico, 2009](#)) as well as neural MT ([Sennrich et al., 2016](#))



# Mining with LASER: sentence embeddings

- Embed the whole sequence into a single, sentence embedding [\(Artexte & Schwenk 2018\)](#)
- Max pooling over BiLSTM (RNN) representations
- No (cross-)attention mechanism

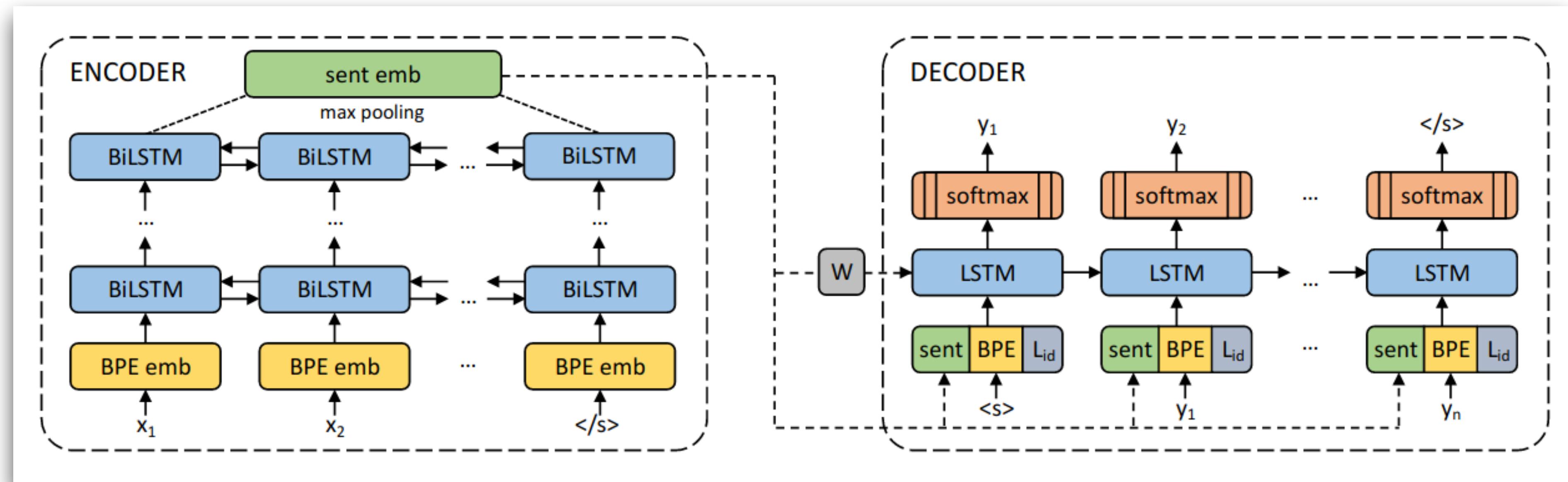


Figure from [\(Artexte & Schwenk 2018\)](#)

Can be seen as a sort of interlingua (although in reality not completely language-independent)

# LASER: data mining

- LASER used to mine data from online dumps
- Embed sentences, compare distance and find parallel pairs in many languages
- Necessary to filter:
  - Margin threshold: ratio between cosine distance between 2 sentences and their average cosine similarities with other sentences
- Result: Wikimatrix, CCMATRIX ([Schwenk et al. 2019a](#); [Schwenk et al., 2019b](#))

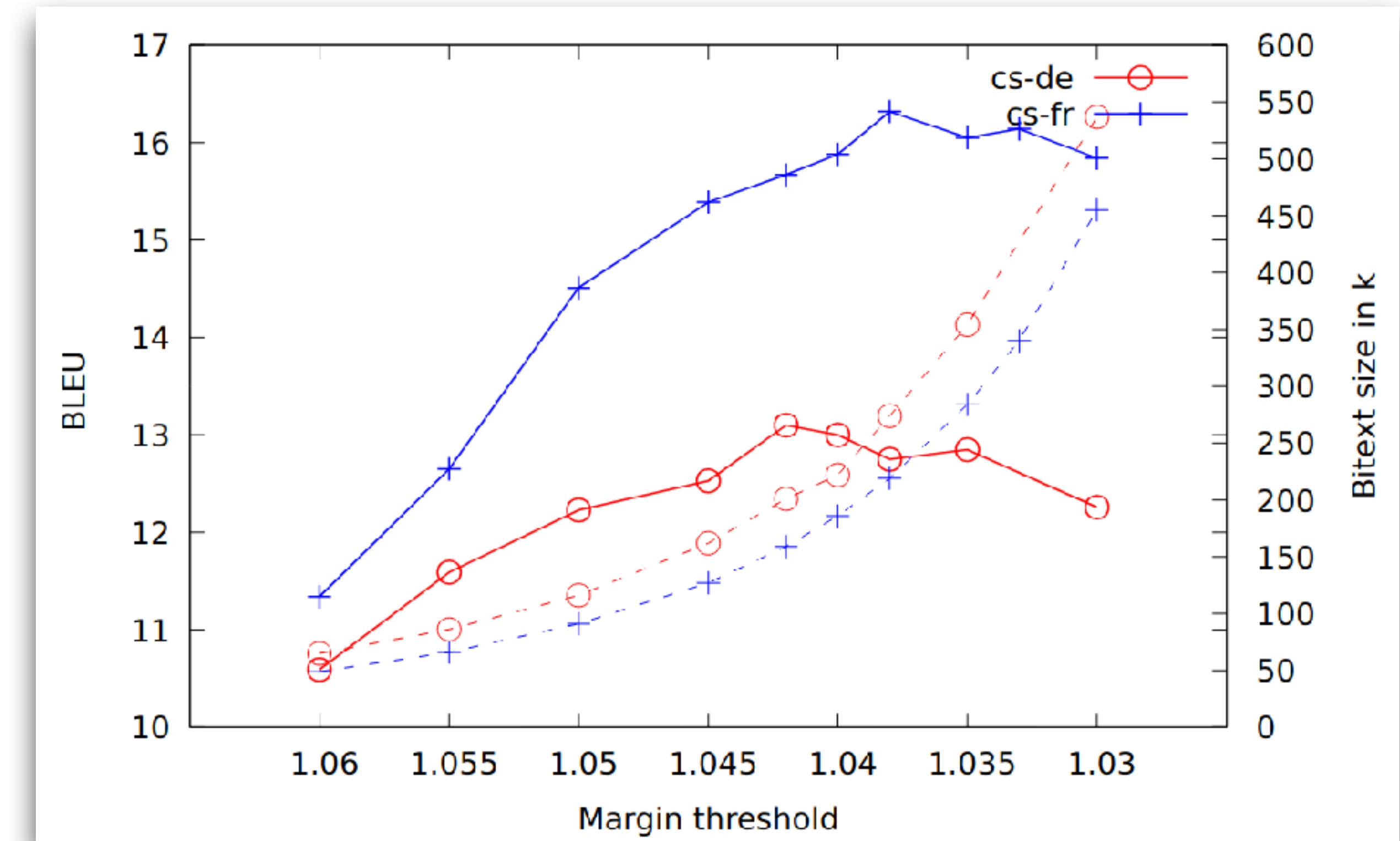
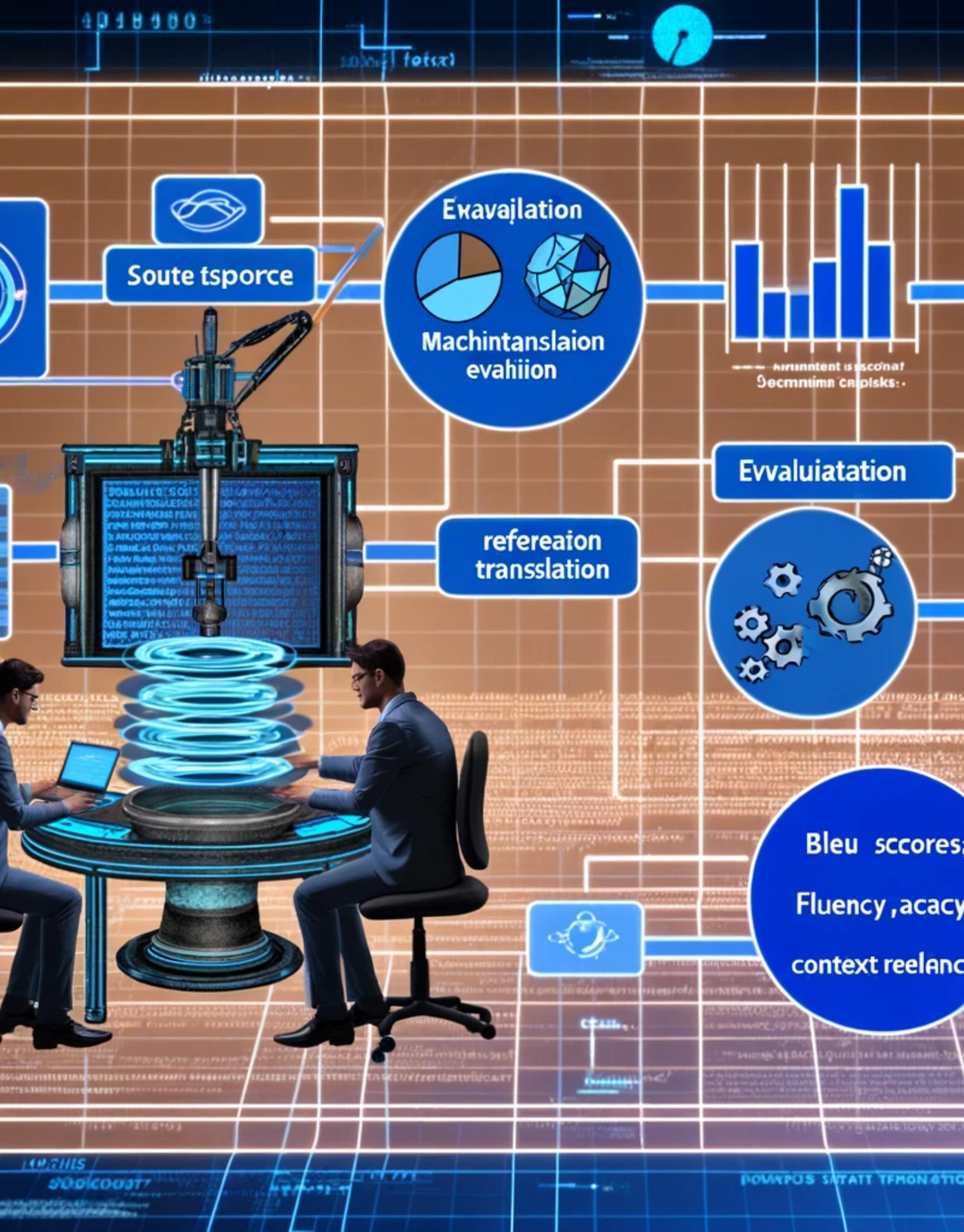


Figure from (Schwenk et al. 2019a)

# Evaluation

ChatGPT 4 (11/3/24): Can you draw the process of machine translation evaluation?



# Why evaluate?

- **How good was a system compared to others and/or baseline models?**
  - To validate an initial research hypothesis
  - To test whether the system is deployable
- **What is a system doing well and what it is doing poorly?**
  - Guides future research (which topics need investigating)
  - Helps to choose between competing systems (depending on what sort of model you are looking to use)

# Claims of human parity in MT

## Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research

### Abstract

Machine translation has made rapid advances in recent years. Millions of people rely on it today in online translation systems and mobile applications in order to communicate across language barriers. The question naturally arises whether such systems can approach or achieve parity with human translations. In this paper, we first address the problem of how to define and accurately measure human parity in translation. We then describe Microsoft's machine translation system and measure the quality of its translations on the widely used WMT 2017 news translation task from Chinese to English. We find that our latest neural machine trans-

that the translation quality is at human parity when compared to professional human translations. We also find that it significantly exceeds the quality of crowd-sourced non-professional translations.



[Submitted on 30 Aug 2018]

## Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation

Antonio Toral, Sheila Castilho, Ke Hu, Andy Way

[Submitted on 21 Aug 2018]

## Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

Samuel Läubli, Rico Sennrich, Martin Volk

[Submitted on 24 Jun 2025]

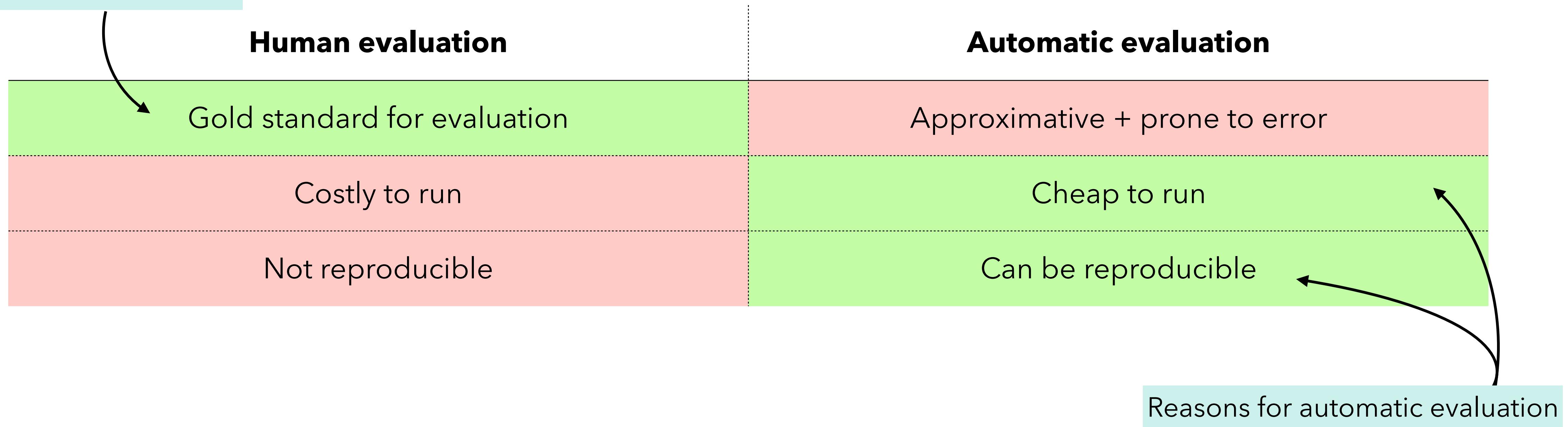
## Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress

Lorenzo Proietti, Stefano Perrella, Roberto Navigli

In Machine Translation (MT) evaluation, metric performance is assessed based on agreement with human judgments. In recent years, automatic metrics have demonstrated increasingly high levels of

# Human vs. automatic evaluation

What we want to aim for



In practice you will need both:

- Automatic evaluation for development and final evaluation (to compare against others)
- Human evaluation to verify the final automatic evaluation and to provide qualitative analysis

# Designing an evaluation protocol (human or automatic)

- **What benchmark?** (Domain/genre of the text, how many examples?)
- **What unit to be evaluated?** (Individual sentence, whole document, paragraph, sentence with additional context, etc.)
- **How is the quality of the unit determined?** (Comparison to reference translation(s), comparison to the source text, using what criteria?)
- **What sort of score is provided?** (Individual quality score for a unit, ranking with another translation, average score over an entire benchmark, etc.)

# Human evaluation

# Human evaluation protocols

- Relative ranking
- Direct assessment  
(Graham et al., 2013)
- SQM (Scalar Quality Metrics)  
(Freitag et al., 2021)
- MQM (Multidimensional quality metrics)  
(Freitag et al., 2021)
- ESA (error span analysis)  
(Kocmi et al., 2025)

Appraise   Overview   Status   cfedermann ▾

Până la mijlocul lui iulie, procentul a urcat la 40%. La începutul lui august, era 52%.

— Source

By mid-July, it was 40 percent. In early August, it was 52 percent.

— Reference

**Best** ← Rank 1  Rank 2  Rank 3  Rank 4  Rank 5  → **Worst**

Until the middle of July, the percentage rose to 40%.

**Best** ← Rank 1  Rank 2  Rank 3  Rank 4  Rank 5  → **Worst**

Until mid-July, the percentage rose to 40%.

**Best** ← Rank 1  Rank 2  Rank 3  Rank 4  Rank 5  → **Worst**

By mid-July, the percentage climbed to 40 per cent.

**Best** ← Rank 1  Rank 2  Rank 3  Rank 4  Rank 5  → **Worst**

Until mid-July, the percentage climbed to 40%.

**Best** ← Rank 1  Rank 2  Rank 3  Rank 4  Rank 5  → **Worst**

Until the middle of July, the figure climbed to 40%.

Submit   Reset   Skip Item

Figure 2 from (Bojar et al., 2016)

# Human evaluation protocols

- Relative ranking
- Direct assessment  
[\*\*\(Graham et al., 2013\)\*\*](#)
- SQM (Scalar Quality Metrics)  
[\*\*\(Freitag et al., 2021\)\*\*](#)
- MQM (Multidimensional quality metrics)  
[\*\*\(Freitag et al., 2021\)\*\*](#)
- ESA (error span analysis)  
[\*\*\(Kocmi et al., 2025\)\*\*](#)

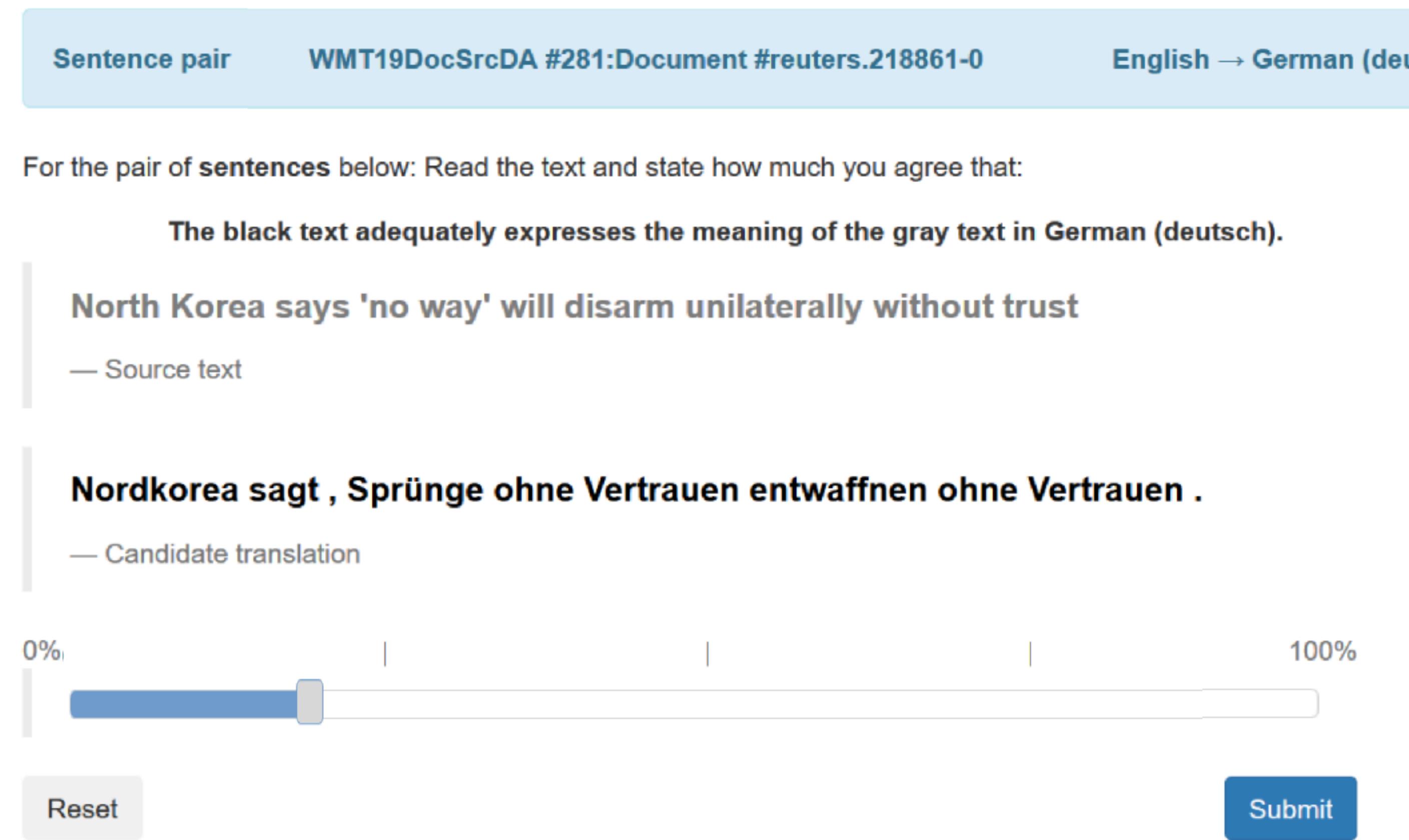


Figure adapted from (Barrault et al., 2019)

# Human evaluation protocols

- Relative ranking
- Direct assessment  
[\(Graham et al., 2013\)](#)
- SQM (Scalar Quality Metrics)  
[\(Freitag et al., 2021\)](#)
- MQM (Multidimensional quality metrics)  
[\(Freitag et al., 2021\)](#)
- ESA (error span analysis)  
[\(Kocmi et al., 2025\)](#)

*Below you see a document with 10 sentences in English (left columns) and their corresponding candidate translations in Czech (čeština) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking at a source text.*

*Assess the translation quality on a continuous scale using the quality levels described as follows:*

- 0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.
- 2: Some meaning preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- 4: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.
- 6: Perfect meaning and grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.

*The numeric labels on the slider are there to help you to adjust the score more precisely, but the slider can be stopped at any position along the track. Try to use the full range of the scale when scoring segments and not limit yourself only to the values around the numeric labels.*

Expand all items Expand unannotated Collapse all items

**↑ The Case Against a Biden Run Is Obvious - and Weak**

0      1      2      3      4      5      6

0: Nonsense/ No meaning preserved      2: Some meaning preserved      4: Most meaning preserved and few grammar mistakes      6: Perfect meaning and grammar

**Reset**

**The Case Against a Biden Run Is Obvious – and Weak**

**Submit**

**▼ Why is this important?**

▼ While Biden and his fellow Democrats can't do much in the way of passing laws with the GOP in control of the House, they can still spend the next two years setting an example.

▼ Collectively, everyone on the team should be seeking out opportunities to play Gallant to the Republicans' weird Goofus impulses.

Proč je to důležité?

I když Biden a jeho kolegové demokraté nemohou dělat mnoho v cestě při schvalování zákonů s GOP v ovládání sněmovny, stále mohou strávit příští dva roky dávat příklad.

Kolektivně, každý v týmu by měl vyhledávat příležitosti hrát Gallant na republikány „divné Goofus impulsy.

# Human evaluation protocols

- Relative ranking
- Direct assessment  
(Graham et al., 2013)
- SQM (Scalar Quality Metrics)  
(Freitag et al., 2021)
- MQM (Multidimensional quality metrics)  
(Freitag et al., 2021)
- ESA (error span analysis)  
(Kocmi et al., 2025)

ref	我希望大家能花点时间考虑一个非常简单的事情那就是，到目前为止， 我们对宇宙的大部分了解都来自于光。  MQM: 20.000	I want to ask you all to consider for a second the very simple fact that, by far, most of what we know about the universe comes to us from light.	[by far] Major Accuracy/Mistranslation rater3
refB	我希望大家能花点时间考虑一个非常简单的事情那就是，到目前为止， 我们对宇宙的大部分了解都来自于光。  MQM: 1.000	I hope you can take some time to consider a very simple fact, that is, so far, most of our knowledge about the universe comes from light.	[that is] Minor Fluency/Grammar rater8
Borderline	我们站在地球上仰望夜空用肉眼就能看到天上的繁星。  MQM: 1.000	We stand on the earth and look up at the night sky. With the naked eye, we can see the stars in the sky.	[earth] Minor Fluency/Spelling rater3
DIDI-NLP	我们站在地球上仰望夜空用肉眼就能看到天上的繁星。  MQM: 0.000	When we stand on the earth and look up at the night sky, we can see the stars in the sky with our naked eyes.	No-error No-error rater7

Data from <https://github.com/google/wmt-mqm-human-evaluation>, visualised using the Marot visualisation tool

# Human evaluation protocols

- Relative ranking
- Direct assessment  
[\(Graham et al., 2013\)](#)
- SQM (Scalar Quality Metrics)  
[\(Freitag et al., 2021\)](#)
- MQM (Multidimensional quality metrics)  
[\(Freitag et al., 2021\)](#)
- ESA (error span analysis)  
[\(Kocmi et al., 2025\)](#)

document 6 / 28 segment 12/100 Hide instructions English → German (Deutsch)

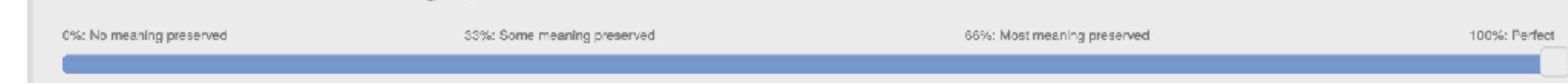
**Highlighting errors:**  
Highlight the text fragment where you have identified a translation error (drag or click start & end). Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.  
**Minor Severity:** Style/grammar/lexical choice could be better/more natural.  
**Major Severity:** Seriously changed meaning, difficult to read, decreases usability.  
 If something is missing from the text, mark it as an error on the [MISSING] word.  
 The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears.  
 Each error should have a separate highlight.

**Score:** After highlighting all errors, please set the overall segment translation scores. The quality levels associated with numerical scores on the slider:  
 0%: No meaning preserved: Nearly all information is lost in the translation.  
 33%: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. The narrative is hard to follow due to errors. Grammar may be poor.  
 66%: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.  
 100%: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

*Getting my eyes checked, because there was a decent deal for it here, and my left eye's been weird for some time now, so maybe they can tell me something what's happening.*

Ich lasse meine Augen untersuchen, denn es gab hier ein gutes Angebot dafür, und mein linkes Auge ist seit einiger Zeit komisch, also vielleicht können sie mir sagen, was los ist. [MISSING]

0%: No meaning preserved      33%: Some meaning preserved      66%: Most meaning preserved      100%: Perfect



*I'm splurging on a new set of frames, these red ones I reeeeally like.*

Ich verschwende mein Geld für eine neue Brillenfassung, diese rote mag ich seeehr. [MISSING]

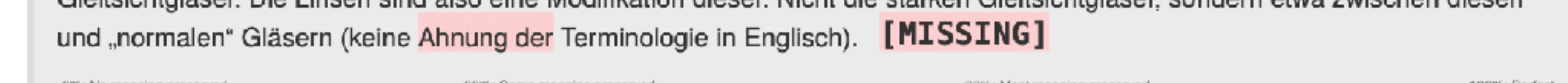
0%: No meaning preserved      33%: Some meaning preserved      66%: Most meaning preserved      100%: Perfect



*So, apparently the ghost image my left eye sees isn't too much of a concern. But I need basically varifocals too. So the lenses have a combo of those. Not the full on varifocals but something in between those and "normal" glasses (no idea of the terminology in English).*

Also ist das Geistsrbild, das mein linkes Auge sieht, nicht so sehr Anlass zum Sorge. Aber grundsätzlich brauche ich auch Gleitsichtgläser. Die Linsen sind also eine Modifikation dieser. Nicht die starken Gleitsichtgläser, sondern etwa zwischen diesen und „normalen“ Gläsern (keine Ahnung der Terminologie in Englisch). [MISSING]

0%: No meaning preserved      33%: Some meaning preserved      66%: Most meaning preserved      100%: Perfect



# Automatic evaluation

# What makes a good automatic metric?

- Best imitate the way in which humans would evaluate translations
  - Evaluate translations from multiple MT systems and calculate the correlation between metrics and human quality judgments, (see WMT metrics shared tasks (e.g. [Freitag et al., 2024](#)))
  - Typically two types of scoring: segment-level (scores for individual units) and corpus/system-level (average over all units in the benchmark)
  - Of course this depends on the quality of the human judgments!

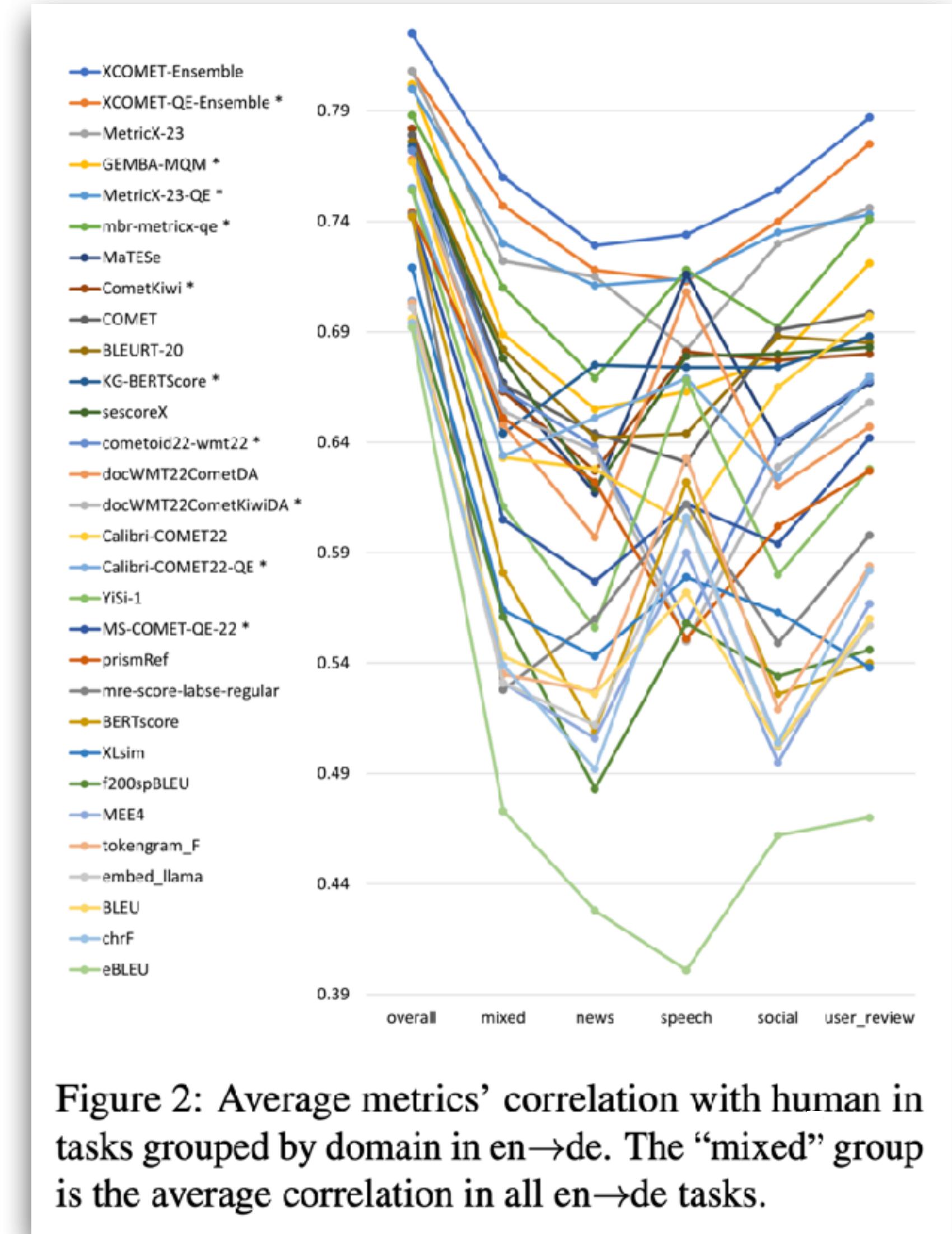


Figure 2: Average metrics' correlation with human in tasks grouped by domain in en→de. The “mixed” group is the average correlation in all en→de tasks.

**Let's look at the most commonly used metric...**

# BLEU

👉 [Papineni et al., 2002. Bleu: a Method for Automatic Evaluation of Machine Translation](#)  
[Standard implementation \(sacrebleu\): A Call for Clarity in Reporting BLEU Scores](#)

- **BiLingual E**valuation **U**nderstudy: the most used metric (at least until very recently)

- Main principle: compare MT outputs to human reference translations

Count the number of overlapping  $n$ -grams between the two

$n$ -gram = sequence of  $n$  things (here, tokens)

**Reference** **Nobody was** bothered **by this problem** .

**Hypothesis** **Nobody was** troubled **by this problem** .

n=1      Nobody **was** **by** **this** **problem** .

n=2      Nobody **was** **by** **this** **this** **problem** **problem** .

n=3      **by** **this** **problem** **this** **problem** .

n=4      **by** **this** **problem** .

# Calculating BLEU

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N \log p_n \right) \quad (\text{usually } N=4)$$

BP (Brevity penalty): penalises overly short translations

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{(1 - \frac{r}{c})} & \text{if } c \leq r, \end{cases}$$

Where  $c$  = candidate length  
and  $r$  = reference length

1 if the candidate translation is longer  
Smaller than 1 if it is shorter or equal in length

Geometric mean of  $n$ -grams

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{\text{n-gram} \in C} \text{Count}_{\text{clip}}(\text{n-gram}, \text{ref})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{\text{n-gram}' \in C'} \text{Count}(\text{n-gram}')}$$

For each value of  $n$ , how often each  $n$ -gram in a candidate translation also appears in the associated reference (clipped to the number of times it appears in the reference)

# Why is this good?

- A translation that is very similar to a reference translation is likely to have the same meaning (and therefore be good)
- Averaged over many examples, this is correlated with human judgments

# Why is this bad?

- Over-reliance on comparing exact surface form similarity with a human reference
  - Problems capturing synonyms, paraphrases and insensitive to small but important changes (e.g. adding a negation marker)
- At the sentence level, it is very poor at imitating human judgments

*Je suis fan de crêpes.*  
↓  
*I love pancakes.*  
*I really like crêpes.*  
*I am a fan of crêpes.*  
*etc.*

# Improving on BLEU

- METEOR ([Banerjee & Lavie, 2005](#)): adding language-specific thesauri and stemming to reward synonyms and near-matches
- ChrF ([Popović, 2015](#)): F-score at the character-level (can reward smaller units of text) - interesting for morphologically rich languages! And ChrF++: calculated both character and token-level  $n$ -grams.

This doesn't solve the major problem of over-reliance on exact formulation of the translation!

# Designing a metric

So you've translated your text and want to evaluate it? Now what?

## 1. **What unit to score?**

Sentence-level

Context-aware and document-level

Easiest! Most metrics do this. Human judgments to evaluate metrics more available

Better capture meaning (and human judgments can be better), but more complicated to implement, and tiring for human evaluators

## 2. **How is the quality determined?**

Comparison to a reference

Comparison to the source and a reference

Comparison to a source sentence

What method to get the quality (lexical, embedding-based, learned metric)

(like BLEU) Comparison in the same language (easier). But requires (costly) references, and likely to be biased towards a particular formulation...

No need for references! But the task is more difficult.

## 3. **What score is provided?**

Average for a corpus (like BLEU)

Sentence-level

Lexical metrics (like BLEU) over-rely on exact formulations. Continuous representations are better!

Both are useful! A single score to compare systems against each other, but sentence-level scores to get finer judgments and to carry out analysis.

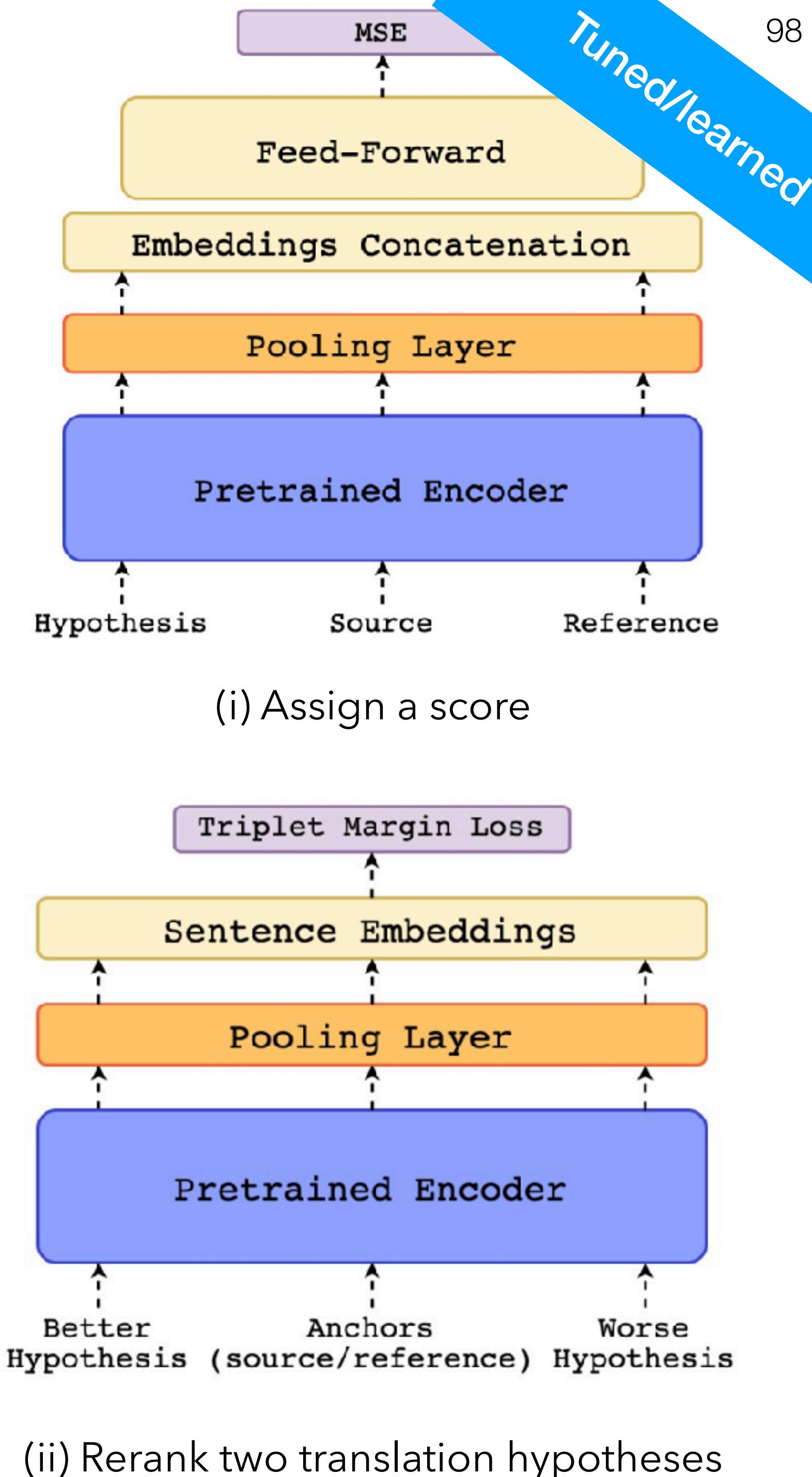
# What approaches (beyond BLEU)?

- **Lexical metrics (comparing surface forms)**
  - BLEU ([Papineni et al., 2002](#)), METEOR ([Banerjee and Lavie, 2005](#)), ChrF ([Popović, 2015](#)), ChrF+  
+ ([Popović, 2017](#))
- **Embedding-based metrics (comparing aligned word embeddings)**
  - BERTScore ([Zhang et al., 2020](#))
- **Tuneable metrics (tune a language model to the task of imitating human judgments)**
  - BLEURT ([Sellam et al. 2020](#)), COMET ([Rei et al., 2020](#)), MetricX ([Juraska et al., 2024](#))
- **LLM-based metrics (ask an LLM to do the task, error detection, etc.)**
  - GEMBA ([Kocmi and Federmann, 2023a](#)), GEMBA-MQM ([Kocmi and Federmann, 2023b](#))

In the following slides, I will list a few that I believe are the most widely used today, but there are more to check out if you are interested!

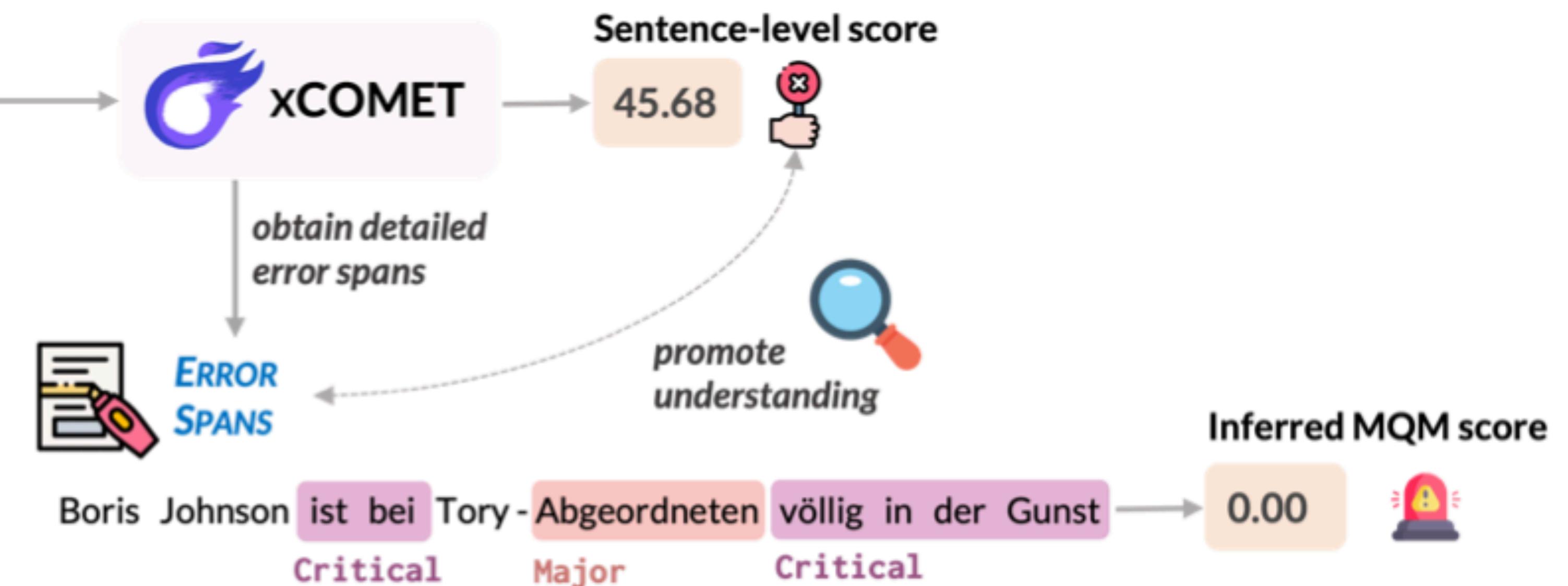
# Learned metric: COMET

- Learned metric: XLM-R fine-tuned on human quality judgments from WMT shared tasks
- Embed source, reference and hypothesis and learn to (i) assign a quality score, or (ii) rerank two translations
- Also exists in a reference-less version! (COMETKiwi - like QE (quality estimation))
- Good correlations with human judgments



# xCOMET

## INPUT TRIPLE



- Train a model to evaluate (like COMET)
- Unified reference-based and reference-free (QE) model
- But also provides annotated error spans

# MetricX

- mT5 base (multilingual version of T5)
- Reference-based and reference-free versions
- Trained on direct assessment (DA) scores then a mixture of MQM and DA scores
- Synthetic data examples to make the model robust to frequent failure cases (fluent but unfaithful translation, under translation)

# LLM-based metrics

- GEMBA ([Kocmi and Federmann, 2024a](#)): GPT-based metric

Score the following translation from {source\_lang} to {target\_lang} **with respect to the human reference** on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

```
{source_lang} source: "{source_seg}"
{target_lang} human reference: {reference_seg}
{target_lang} translation: "{target_seg}"
Score:
```

- GEMBA-MQM ([Kocmi and Federmann, 2024b](#)): predict multiple dimensions of quality (based on the MQM framework)

(System) You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the translation.

```
(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
```

Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them. The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling),

**locale convention (currency, date, name, telephone, or time format)**

style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.\n

Each error is classified as one of three categories: critical, major, and minor.

Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

```
(assistant) {observed error classes}
```

# Alternative methods: Challenge sets

- Test sets containing specific examples to test particular phenomena (e.g. things that MT models struggle to do)
- Human evaluation of those examples
- OR contrastive evaluation: provide correct and incorrect translations and use your MT model to score each of them. Can your model score the correct translation higher than the incorrect one?
- E.g. L'avocat était en colère → The lawyer was angry  
OR The avocado was angry



# Evaluation: The big picture

- Evaluation is the starting point: only train models you can evaluate
- Think about what you want to evaluate and which is the best metric for your task
  - E.g. tuned metrics and LLMs might not work well for all languages
- **In practice:**
  - COMET/MetricX models if your language is well represented, if not check the performance (but be careful if your model is trained using feedback from these metrics)
  - ChrF/BLEU for other cases (especially if differences in scores are large)
  - XCOMET, LLM-based metrics for more detailed insights
  - ALWAYS do some human evaluation (on a subset of your predictions) and look at what doesn't work in addition to what does. Design challenge sentences to test this. Choose an evaluation protocol adapted to your resources (budget, expertise, number of annotators, time)

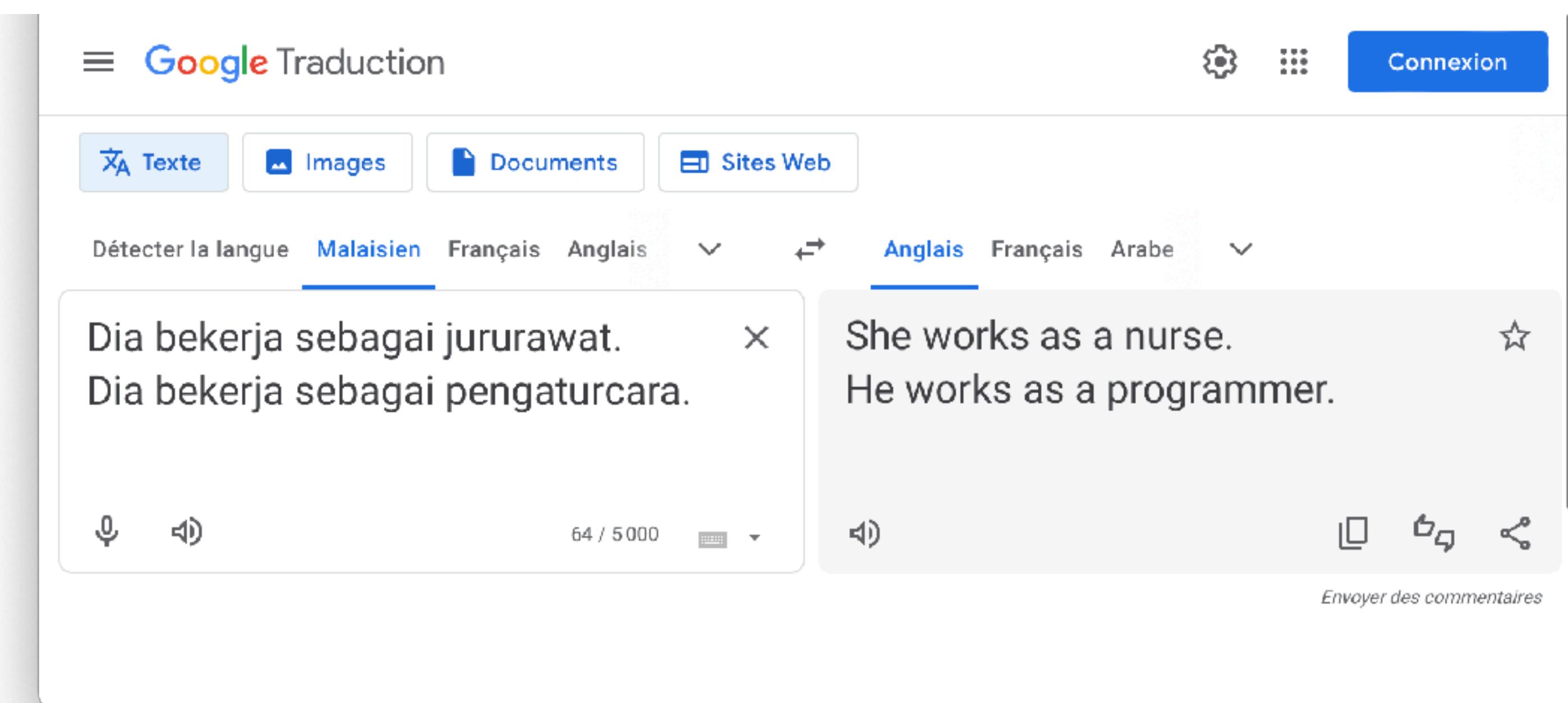
# Some challenges to think about

# Bias

- Bias = priors the model has, particularly when amplifying tendencies seen in the data
- Problematic when for example reinforcing of stereotypes, creating social harm

# Gender bias

- Visible for translation from languages where the pronouns are gender neutral and translating into a language where gender is expressed.



# Gender bias

The screenshot shows a Mac OS X-style window for the ChatGPT 4 interface. The top bar includes standard window controls (red, yellow, green buttons, minimize, maximize, close) and a tab for 'chat.openai.com'. The main area has a light gray background.

**You** (indicated by a teal circular icon with 'BE') asks: "Translate from Malay into English the following sentences: 'Dia bekerja sebagai jururawat. Dia bekerja sebagai pengaturcara.'"

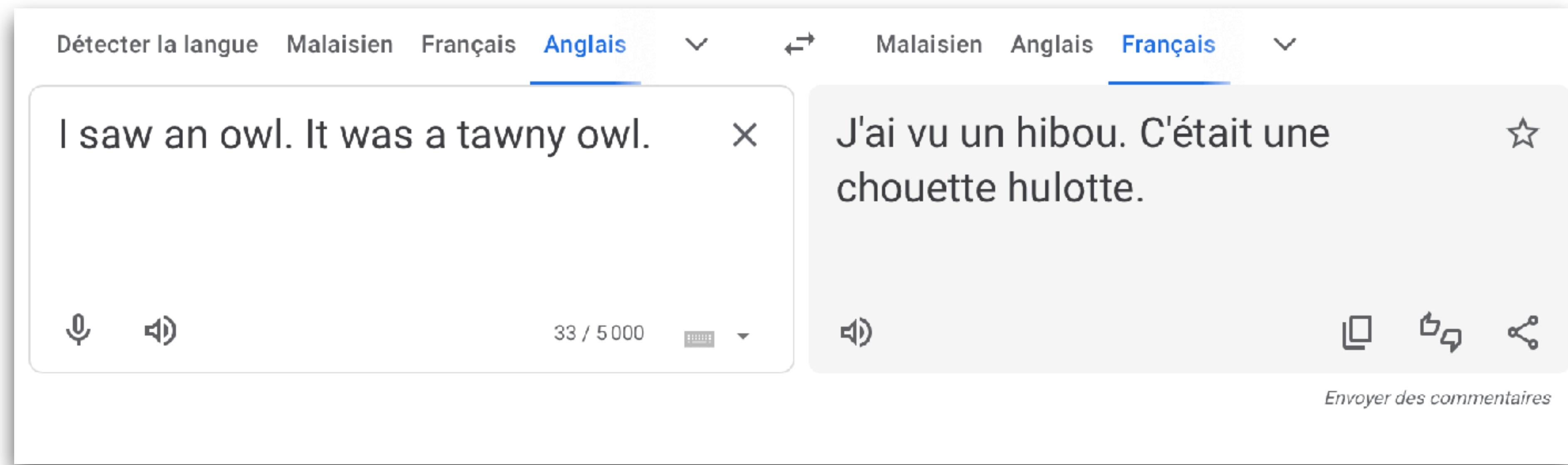
**ChatGPT** (indicated by a purple circular icon with a neural network symbol) responds: "The translation from Malay to English for the given sentences is:

1. "Dia bekerja sebagai jururawat." - "She works as a nurse."  
2. "Dia bekerja sebagai pengaturcara." - "He works as a programmer."

At the bottom, there are five small icons for sharing or rating: a clipboard, a thumbs up, a thumbs down, and two other symbols.

# Taking into account context

- Coherence and consistency of translation

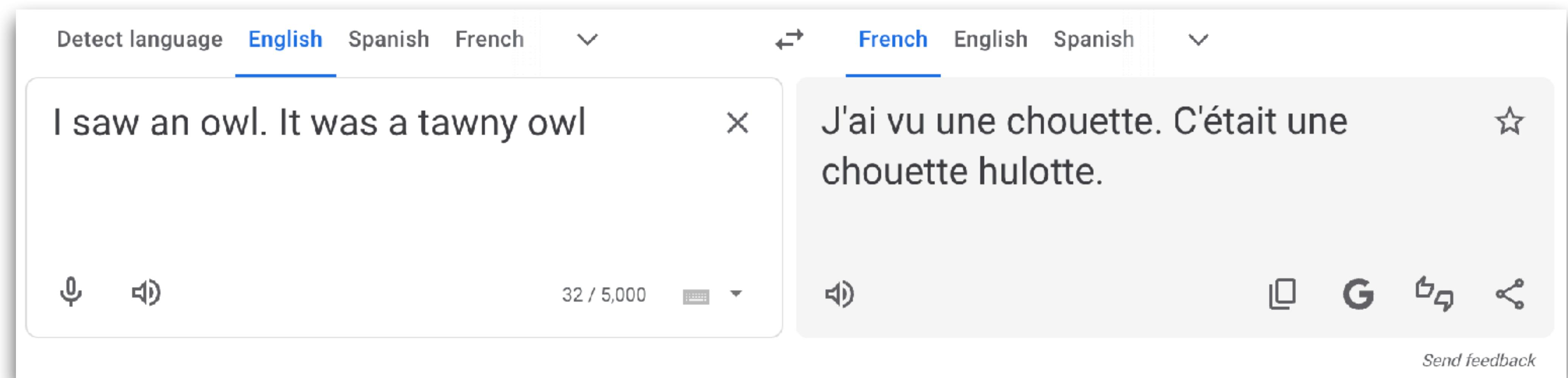


Google Translate (28/12/2023)

# Taking into account context

- Coherence and consistency of translation

Improvements have been made in recent years!



# Taking into account context

- Coherence and consistency of translation

It's something LLMs also do quite well (although not completely)

The screenshot shows a web browser window for [chat.openai.com](https://chat.openai.com). The interface is a conversational AI system.

**You:** BE  
Translate from English into French the following sentences: "I saw an owl. It was a tawny owl."

**ChatGPT:**  
"I saw an owl. It was a tawny owl." translates into French as "J'ai vu une chouette. C'était une chouette hulotte."

At the bottom, there are icons for a file, a thumbs up, a thumbs down, and a refresh arrow.

# But problems remain...

- Flexibility can lead to problems that are more difficult to anticipate
- Additional context can have undesirable effects too!

The screenshot shows a translation interface with two panels. The left panel has "English" as the source language and "French" as the target language. The right panel has "French" as the source language and "English" as the target language. Both panels show a list of translations.

English Input	French Output
It was pretty.	C'était joli.
It was pretty important.	C'était très important.

Below the list are download and audio icons, and a progress bar indicating 39 / 5,000. There are also star and share icons.

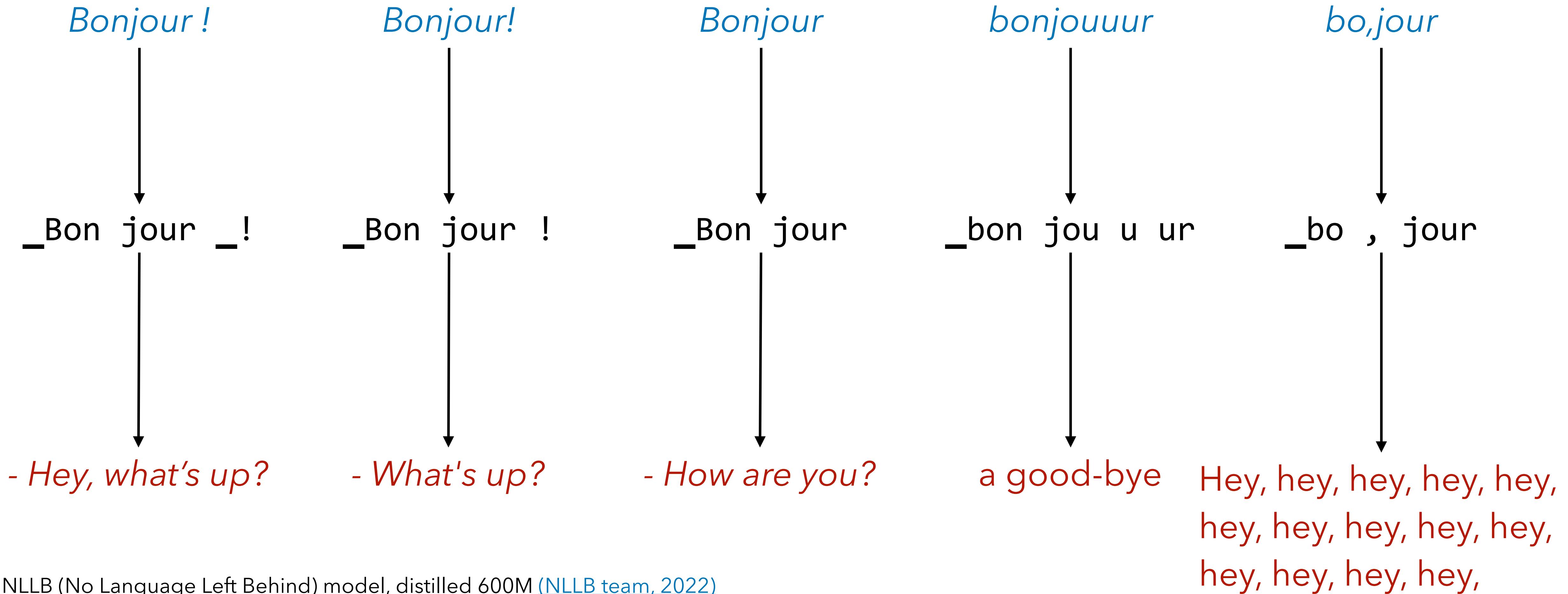
This screenshot shows the same translation interface as the first one, but with different input and output language settings. The left panel has "English" as the source language and "French" as the target language. The right panel has "French" as the source language and "English" as the target language. The results are identical to the first screenshot.

English Input	French Output
It was pretty important.	C'était assez important.
It was pretty.	C'était plutôt.
Important is a good word.	Important est un bon mot.

Below the list are download and audio icons, and a progress bar indicating 65 / 5,000. There are also star and share icons.

# Robustness to language variation

Example from the distilled NLLB model (600M) ([NLLB team, 2022](#))



# Robustness to language variation

NOT a soul gone run me **!!** I'm grown & I'm literally gone do  
**wtf** I want when I want... & dats just my attitude 

[Traduire le Tweet](#)

15:42 · 01/05/2022 · [Twitter for iPhone](#)

# Robustness to language variation

The screenshot shows a Google Translate interface on a Mac OS X desktop. The URL in the address bar is `translate.google.fr`. The main content area displays a comparison between English (Text) and French (Text) versions of the same sentence.

**English (Text):**

NOT a soul gone run me !! I'm grown & I'm  
literally gone do wtf I want when I want... &  
dats just my attitude 🤷

**French (Text):**

PAS une âme n'est partie pour me diriger !! Je  
suis adulte et je suis littéralement parti faire ce  
que je veux quand je veux... et c'est juste mon  
attitude 🤷

Below the text boxes, there is a note in blue text:

Essayez avec cette orthographe : NOT a soul gone run  
me !! I'm grown & I'm literally gonna do wtf I want  
when I want... and that's just my attitude 🤷

The interface includes standard translation controls like microphone, speaker, and share icons at the bottom.

# Robustness to language variation

Detect language English Japanese Spanish ↴ German English French ↴

NOT a soul gone run me !! I'm grown & I'm literally gone do wtf I want when I want... & dats just my attitude 🤘

Did you mean: NOT a soul gone run me !! I'm grown **and i'm** literally **gonna** do wtf I want when I want... **and that's** just my attitude 🤘

🔊 🔊 114 / 5,000 ⌨️

Personne ne me contrôlera !! Je suis adulte et je ferai littéralement ce que je veux quand je veux... c'est tout simplement mon état d'esprit 🤘

🔊 ⌨️ G ↗ 🔍

Detect language English Japanese Spanish ↴ Occitan German Czech ↴

NOT a soul gone run me !! I'm grown & I'm literally gone do wtf I want when I want... & dats just my attitude 🤘

Showing translation for NOT a soul gone run me !! I'm grown **and i'm** literally **gonna** do wtf I want when I want... **and that's** just my attitude 🤘

.Translate instead NOT a soul gone run me !! I'm grown & I'm literally gone do wtf I want when I want... & dats just my attitude 🤘

🔊 🔊 114 / 5,000 ⌨️

PAS una anma partida corre me !! Soi grand e vau literalament far wtf que vòli quand vòli... e aquò's just mon actitud 🤘

🔊 ⌨️ G ↗ 🔍

# How we use MT remains important

Look at your data, look at your outputs!



≡ Google Traduction

Texte Images Documents Sites Web

DéTECTER la langue Gallois Norvégien Anglais Français Arabe

Nid wyf yn y swyddfa ar hyn o bryd. Anfonwch unrhyw waith i'w gyfieithu.

Je ne suis pas au bureau pour le moment. Envoyez n'importe quel travail pour traduction.

# Questions?

# Link to the quiz

