CSCI 544: Applied Natural Language Processing

Advances in Transformer & NMT

Xuezhe Ma (Max)

USC University of Southern California

Course Organization

NLP Tasks

Models

Text Classification	Sequence Labeling	Syntactic Parsing	Seq2seq & MT	Language Modeling & Generation
Naive Bayes Logistic Regression SVM …	HMMs MEMMs CRFs	Tree-structured Models/Algorithms	……	……
Word Embedding & Neural Networks	RNN LSTM	Neural Structured Networks	Attention & Transformers	Generative Models & LLMs

Advances In NMT

- Semi-Supervised NMT
- Multilingual NMT
- Context-Aware NMT
- Non-Autoregressive NMT
- Evaluation beyond BLEU

Semi-Supervised NMT

Semi-Supervised NMT: Motivation

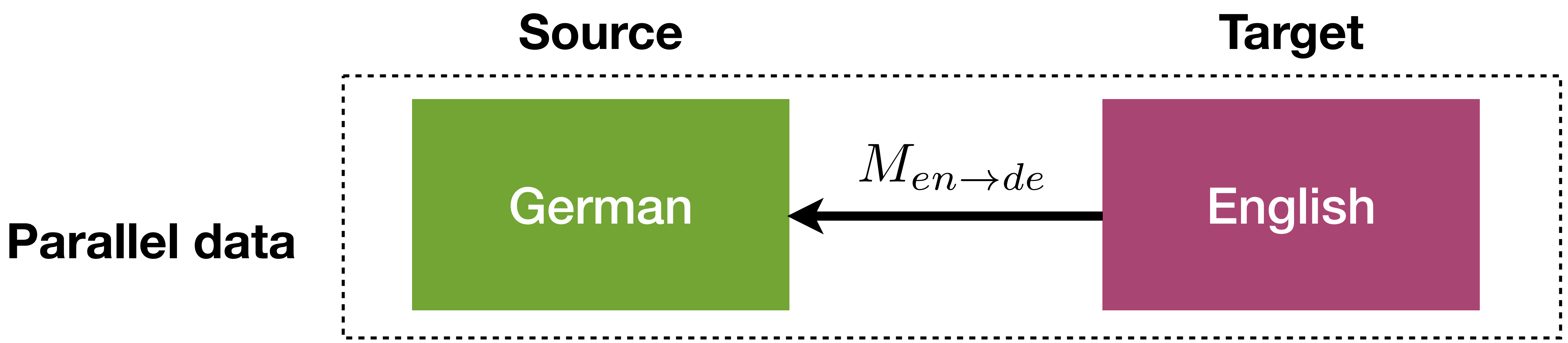
- First train a translation model with limited amount of parallel sentences
- Use this model to generate more synthetic sentence pairs with monolingual corpus

Semi-Supervised NMT: Scenarios

- **Monolingual data from target side**
  - Back-translation
- **Monolingual data from source side**
  - Self-learning

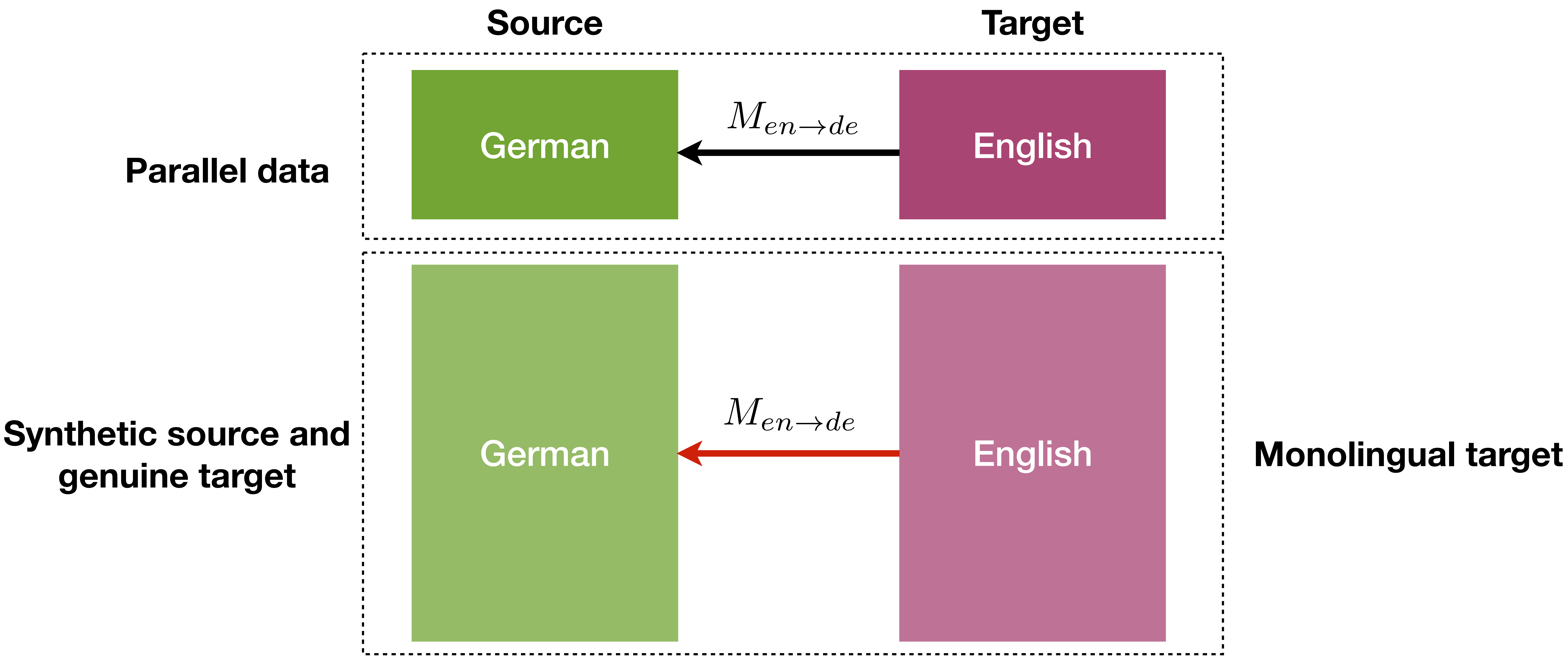
Back Translation

- **<u>Back-translation:</u> using monolingual target side data (Sennrich et al., 2016, Edunov et al., 2018)**



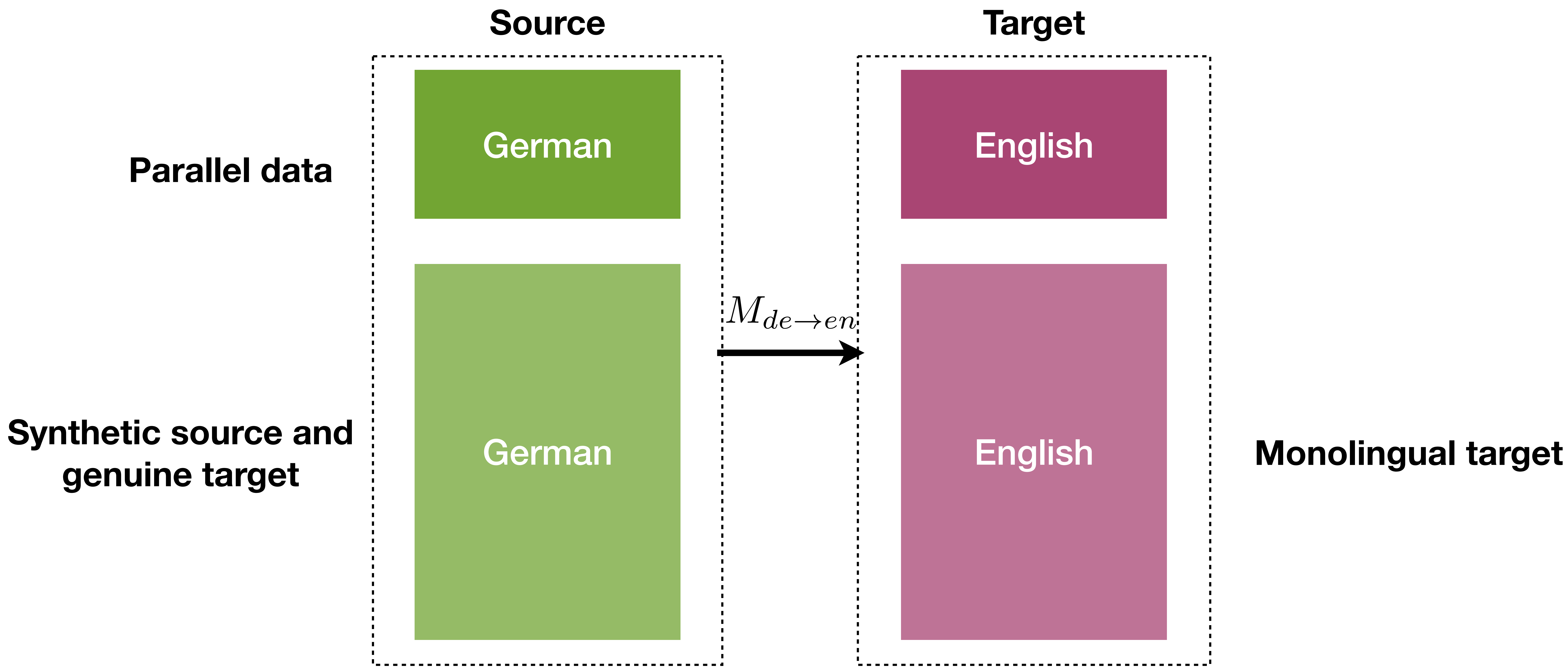
Back Translation

- <u>Back-translation:</u> using monolingual target side data (Sennrich et al., 2016, Edunov et al., 2018)



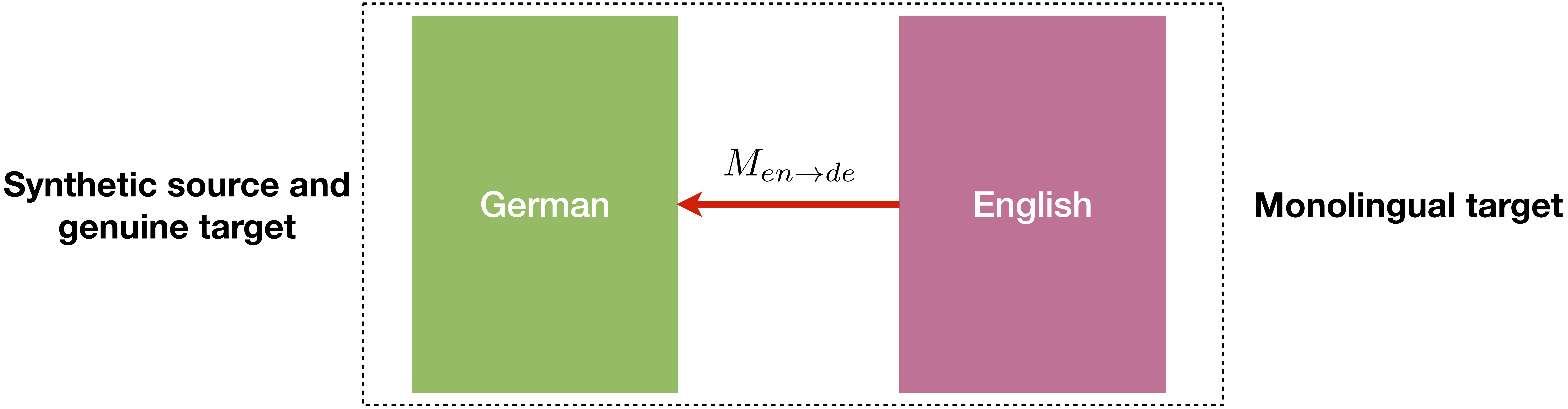
Back Translation

- <u>Back-translation:</u> using monolingual target side data (Sennrich et al., 2016, Edunov et al., 2018)



Back Translation

- **How to generate syntactic data?**
  - Beam search
  - Greedy search
  - Top k
  - Sampling from $p_\theta(Y|X)$
  - +noise



Back Translation: Results

- **English-German**
- **Data**
  - Parallel: WMT-18 (5.2M sentence pairs)
  - Monolingual: 24M German sentences

	news2013	news2014	news2015	news2016	news2017	Average
bitext	27.84	30.88	31.82	34.98	29.46	31.00
+ beam	27.82	32.33	32.20	35.43	31.11	31.78
+ greedy	27.67	32.55	32.57	35.74	31.25	31.96
+ top10	28.25	33.94	34.00	36.45	32.08	32.94
+ sampling	28.81	34.46	34.87	37.08	32.35	33.51
+ beam+noise	29.28	33.53	33.79	37.89	32.66	33.43

Back Translation: Explanations

- **More sentences in target language improves decoder**
  - Better language model in target language
- **Synthetic sentences (with noise) improves encoder**
  - More robust against imperfect source sentences

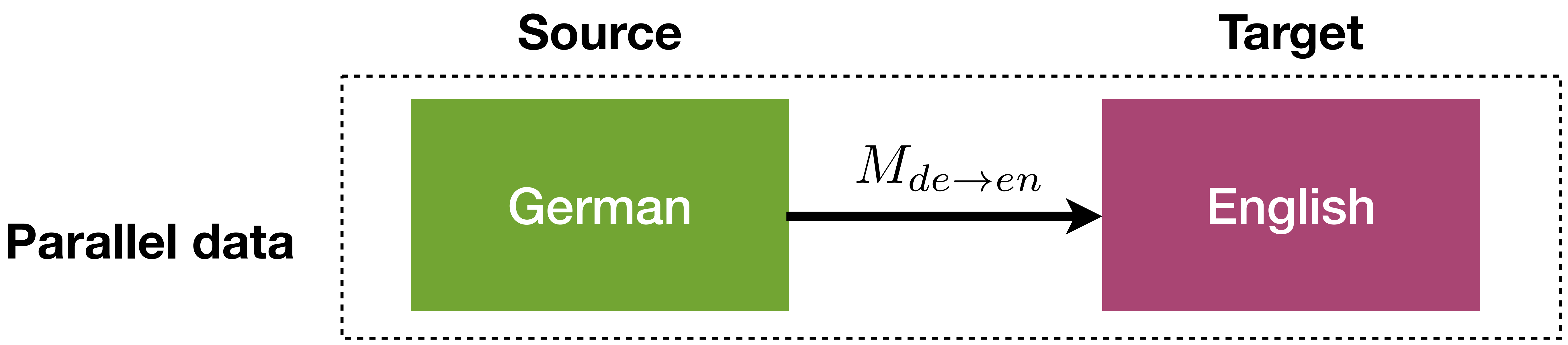
	news2013	news2014	news2015	news2016	news2017	Average
bitext	27.84	30.88	31.82	34.98	29.46	31.00
+ beam	27.82	32.33	32.20	35.43	31.11	31.78
+ greedy	27.67	32.55	32.57	35.74	31.25	31.96
+ top10	28.25	33.94	34.00	36.45	32.08	32.94
+ sampling	28.81	34.46	34.87	37.08	32.35	33.51
+ beam+noise	29.28	33.53	33.79	37.89	32.66	33.43

Semi-Supervised NMT: Scenarios

- **Monolingual data from target side**
  - Back-translation
- **Monolingual data from source side**
  - Self-learning

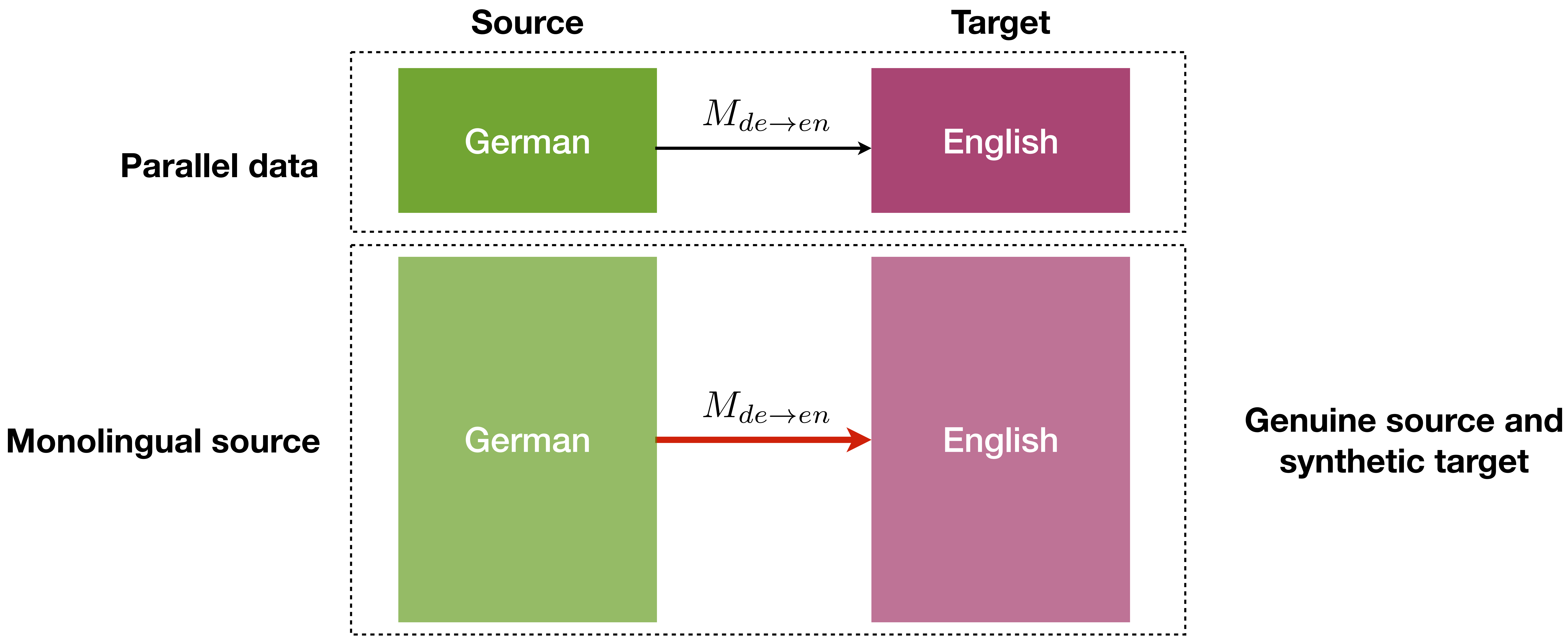
Self-Learning

- <u>Self-training:</u> using monolingual source side data (Scudder 1965, He et al., 2020)



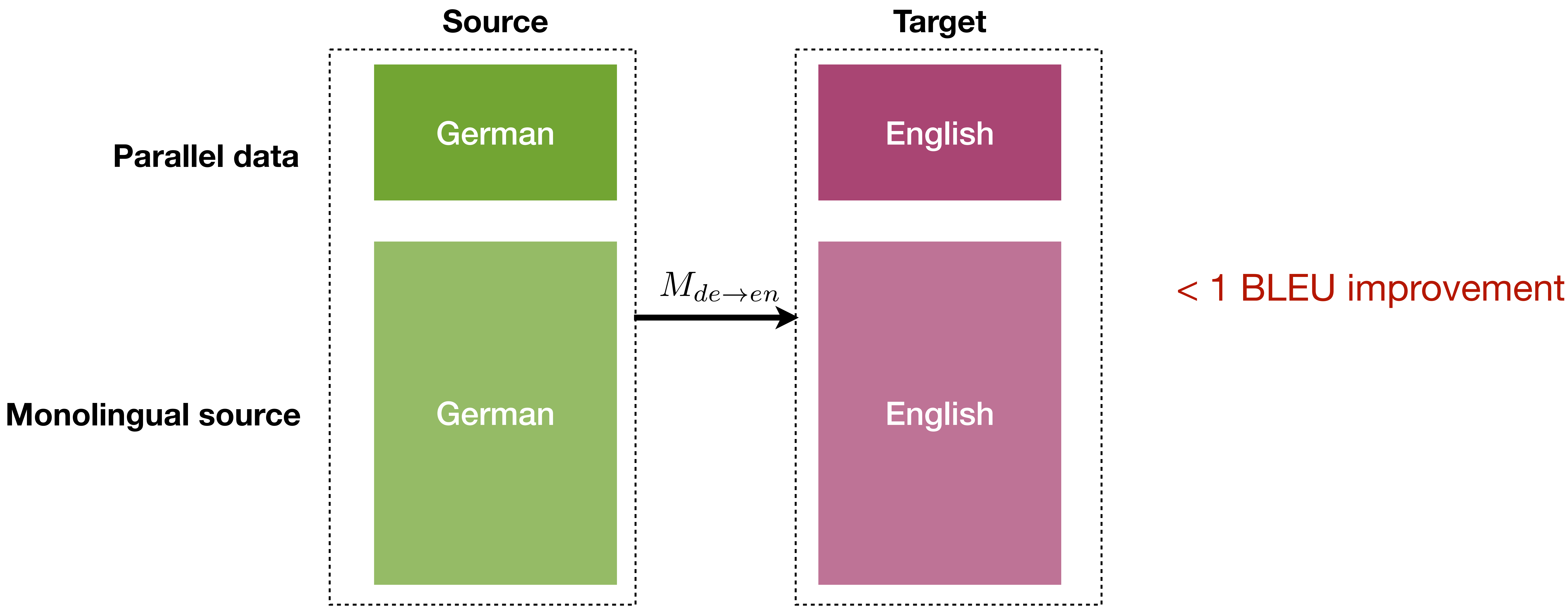
Self-Learning

- **Self-training:** using monolingual source side data (Scudder 1965, He et al., 2020)



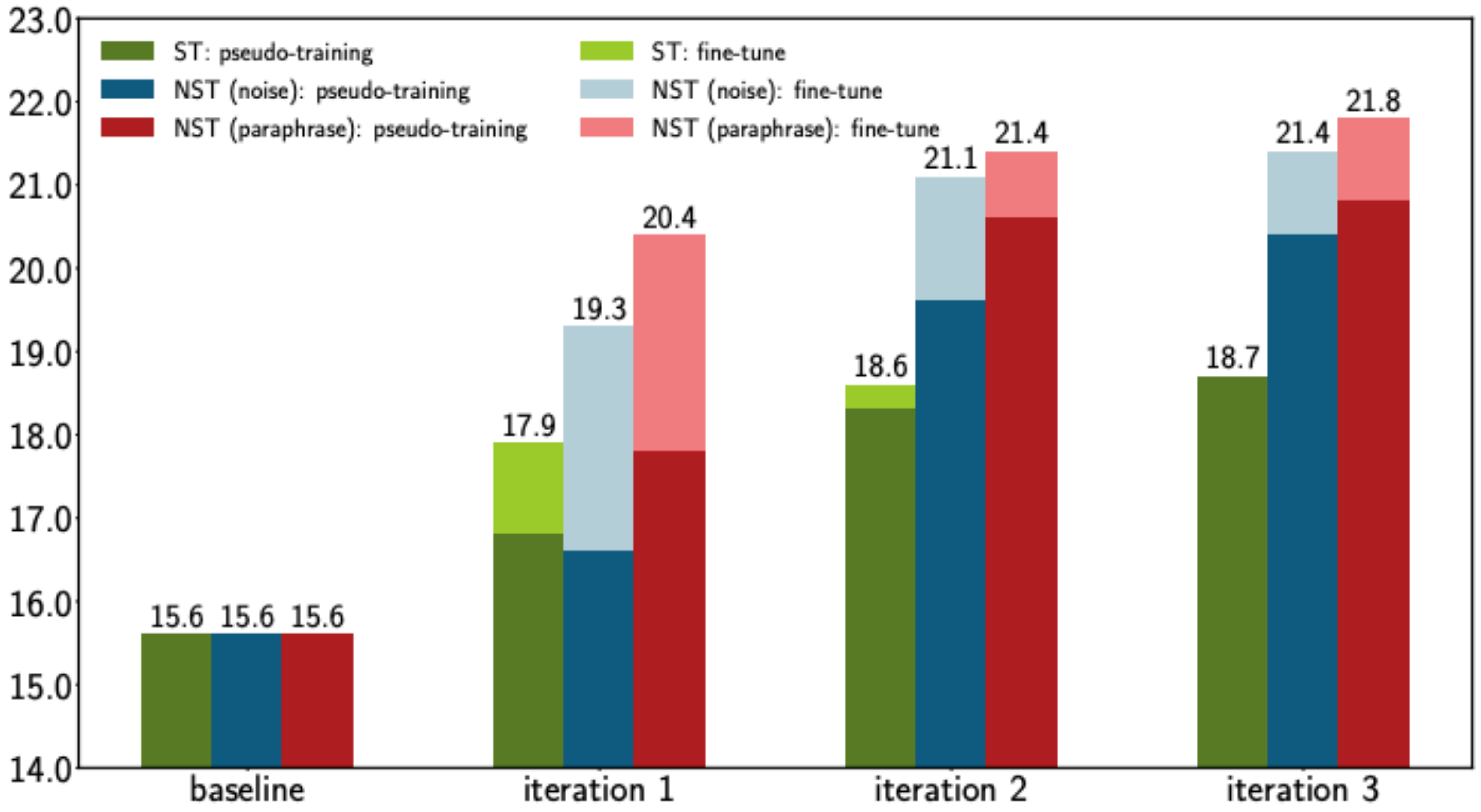
Self-Learning

- **Self-training:** using monolingual source side data (Scudder 1965, He et al., 2020)



Noisy Self-Learning

- **We should add noise to encoder, but genuine sentences to decoder!**
- **Adding noises to source inputs** (He et al., 2020)
  - Word dropout (masks)
  - Permutations
  - Paraphrasing



Semi-Supervised NMT: Summary

- **Motivation**
  - Leveraging large-scale monolingual data to improve MT models
  - Monolingual target sentences: back-translation
  - Monolingual source sentences: self-learning
- **Empirical Evidences**
  - Genuine sentences helps decoder: better language model
  - Noisy sentences helps encoder: robust against noise

Multilingual NMT

Multilingual NMT

- **Many languages are left behind**
  - There are not enough monolingual data for many languages
  - Even less annotated data for NMT

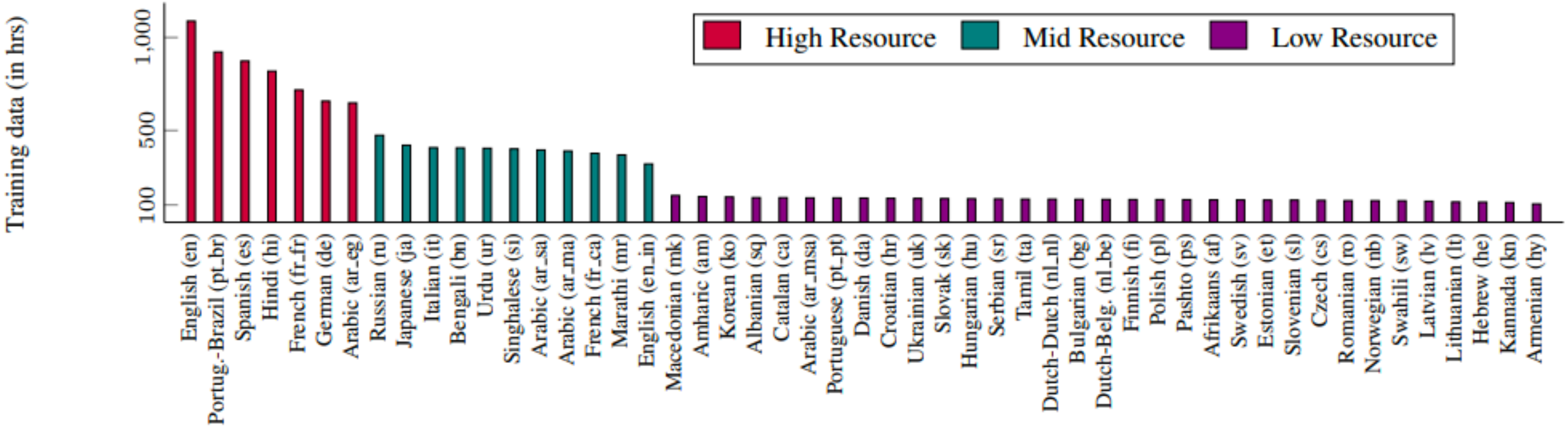
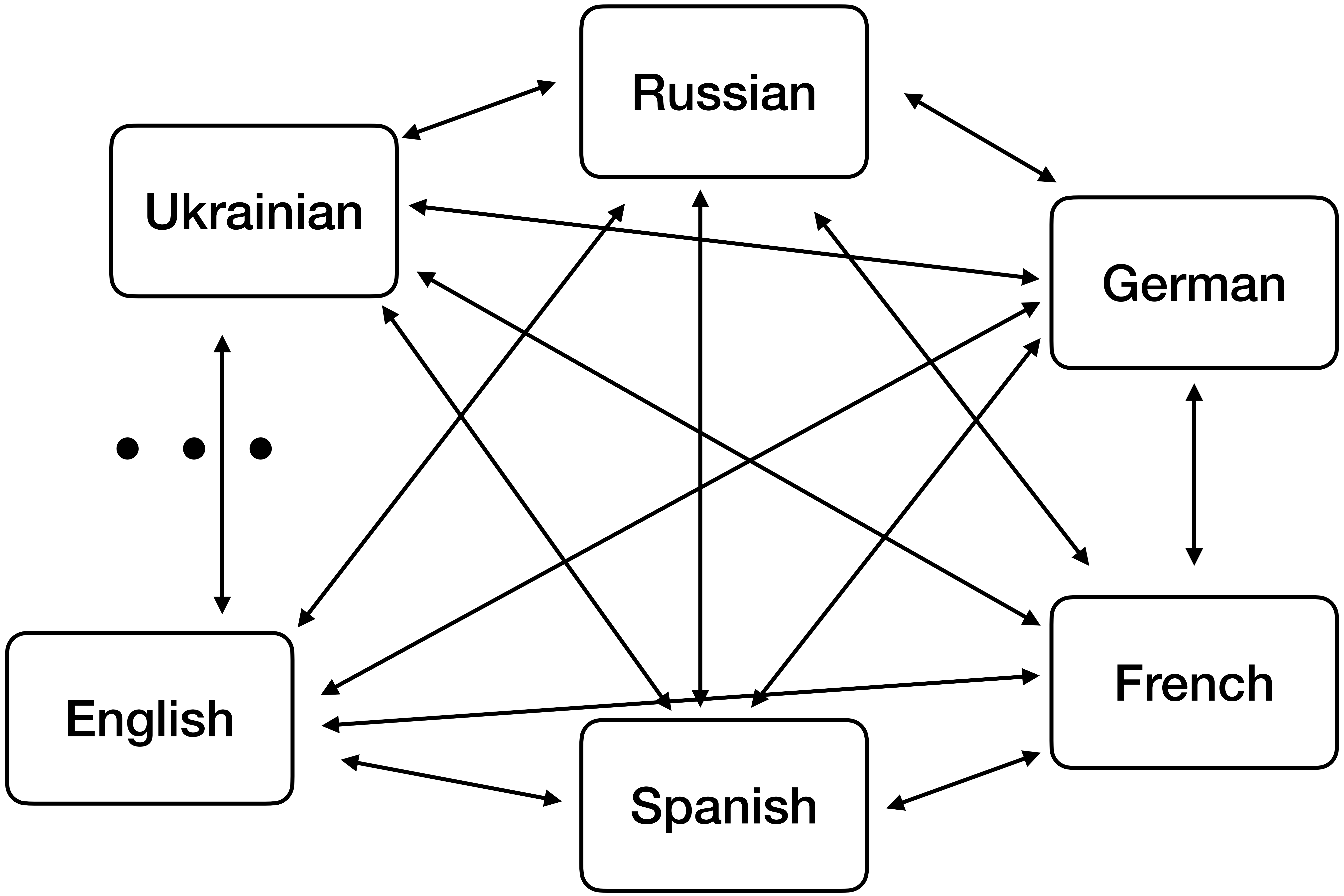


Figure : *Training data distribution across different languages*

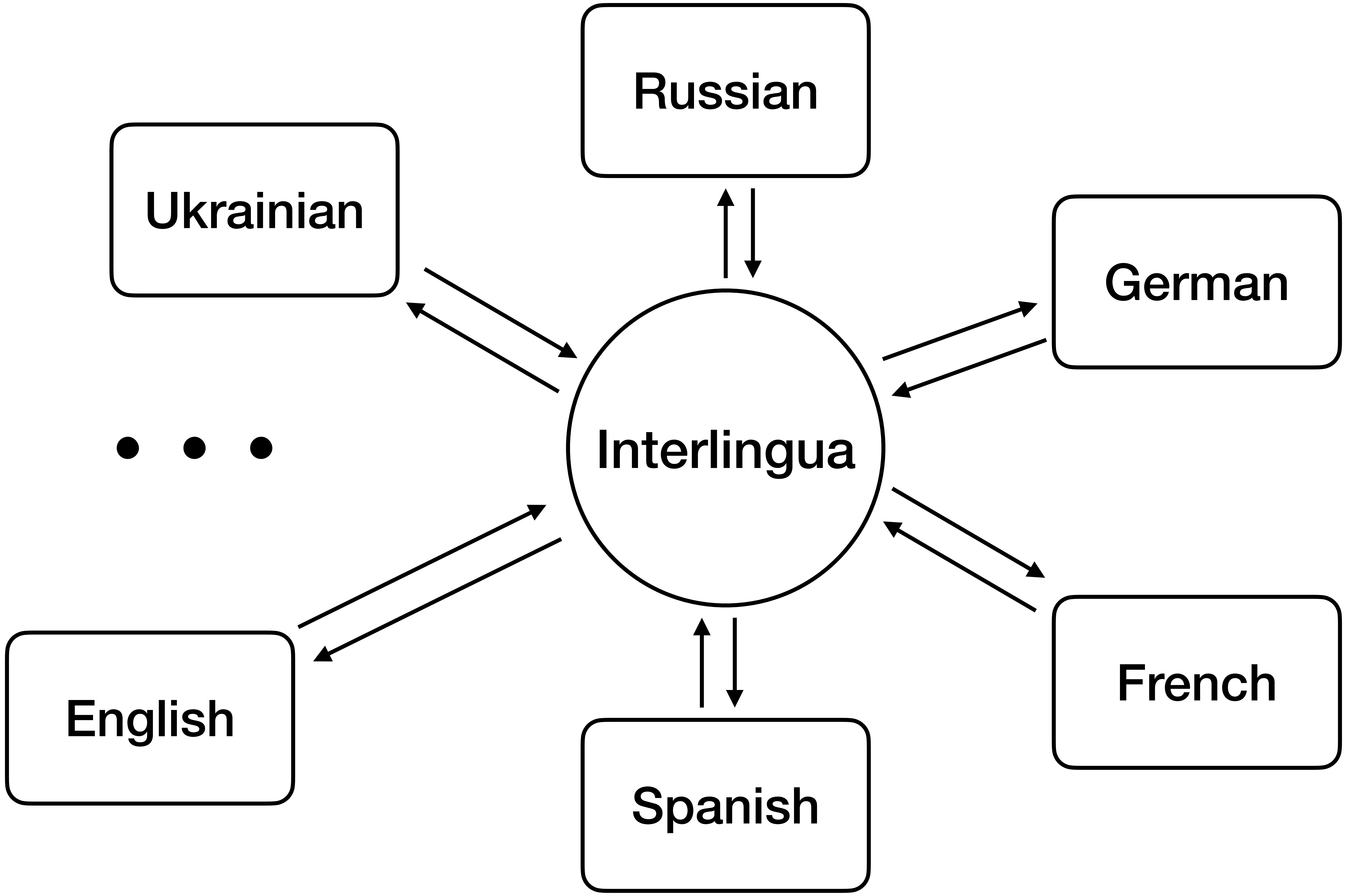
Multilingual NMT

- **Supporting multiple languages could be tedious**
  - Supporting translating from n languages requires $n \times (n-1)$ NMT models



Multilingual NMT

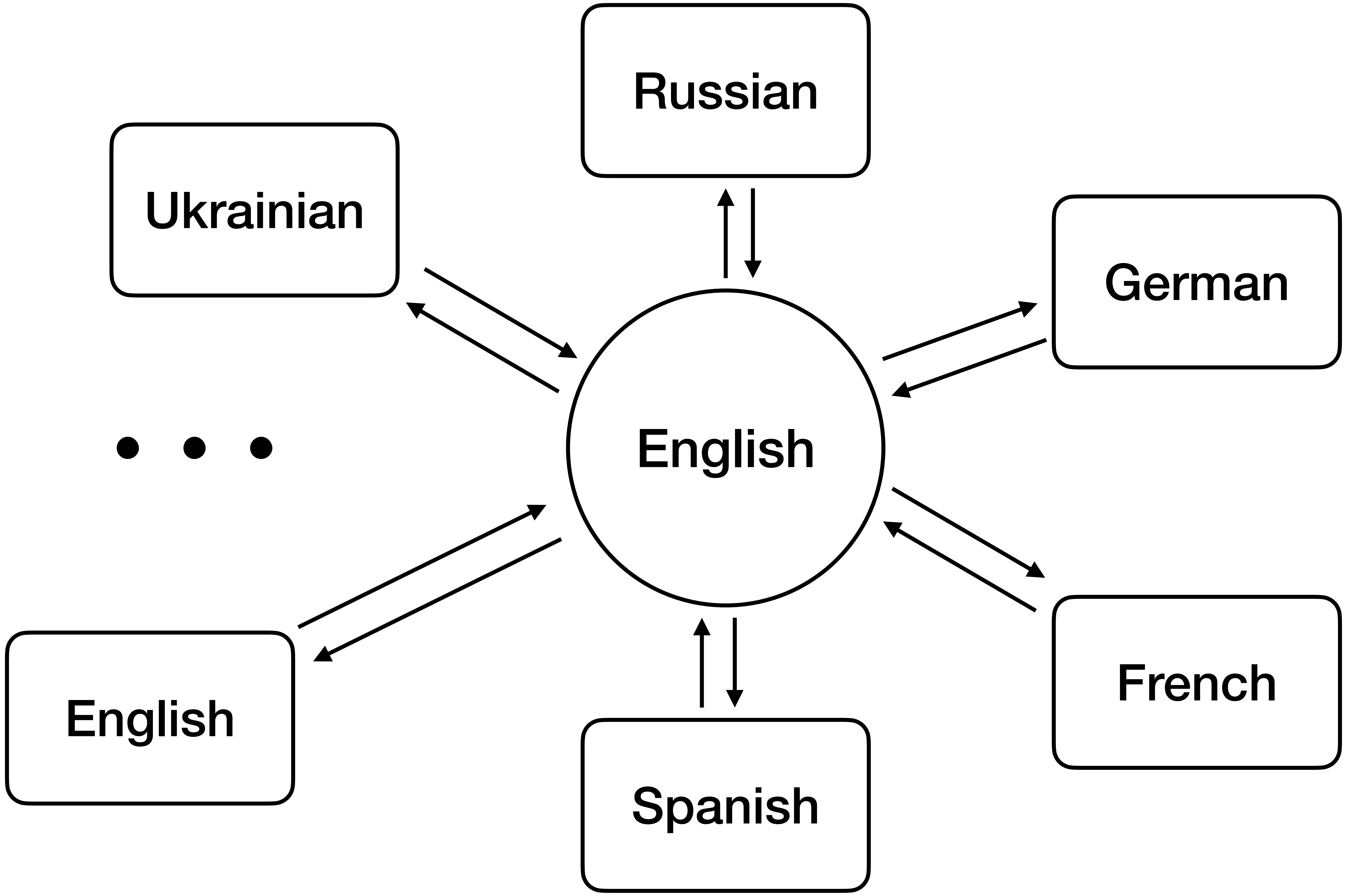
- **Interlingual representation for NMT**



Small languages benefit from big ones that are in the same language family

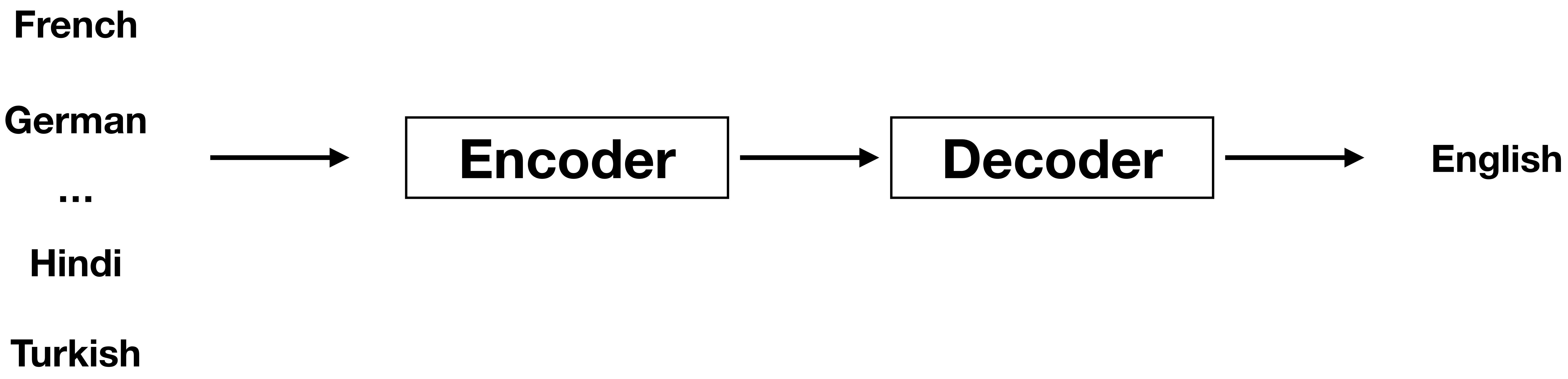
Many-to-One NMT

- **English as a pivot language**



Many-to-One NMT

- Training a single Encoder-Decoder model from multiple languages to English



We need a shared vocabulary across multiple languages!

Vocabulary across Multiple Languages

- **Lexical Divergences**
  - Wall in English corresponds to two words in German, *Wand* (walls inside a building) and *Mauer* (walls outside a building)
- **Morphological Divergences**
  - Number of morphemes per word
    - Isolating languages: Chinese and Vietnamese
    - Polysynthetic languages: Eskimo
  - Morphological boundary
    - Agglutinative languages: relatively clean boundaries
      - Turkish
    - Fusion languages: no clean boundaries
      - Russian: *stolom* (table-SG-INSTR-DECL1)
      - -om: singular (SG), instrumental (INSTR) and first declension (DECL1)

Vocabulary across Multiple Languages

- **Combination of individual vocabularies**
  - Too many different words
  - No shared information
- **Character- or Byte- level vocabulary**
  - Too long sentences
  - Too difficult contextual information

Trade-off between these two ideas?

Byte Pair Encoding

- First split each word into characters (bytes)
- Count the frequency of each consecutive byte pair, find out the most frequent one and merge the two byte pair tokens to one item

	V={all chars/bytes}	V=V + {es}	V=V + {est}	
low: 5	l o w </w>: 5	l o w </w>: 5	l o w </w>: 5	
lower: 2	l o w e r </w>: 2	l o w e r </w>: 2	l o w e r </w>: 2	......
newest: 6	n e w e s t </w>: 6	n e w es t </w>: 6	n e w est </w>: 6	
widest: 3	w I d e s t </w>: 3	w I d es t </w>: 3	w I d est </w>: 3	

Byte Pair Encoding

- First split each word into characters (bytes)
- Count the frequency of each consecutive byte pair, find out the most frequent one and merge the two byte pair tokens to one item
- Iterate from the longest token from learned vocabulary to the shortest one, trying to replace the substring in each of the word to tokens.

V={est</w>, gh, *chars/bytes*}

highest ⟶ h i g h e s t </w> ⟶ h i g h est</w>

Byte Pair Encoding

- First split each word into characters (bytes)
- Count the frequency of each consecutive byte pair, find out the most frequent one and merge the two byte pair tokens to one item
- Iterate from the longest token from learned vocabulary to the shortest one, trying to replace the substring in each of the word to tokens.

V={est</w>, gh, *chars/bytes*}

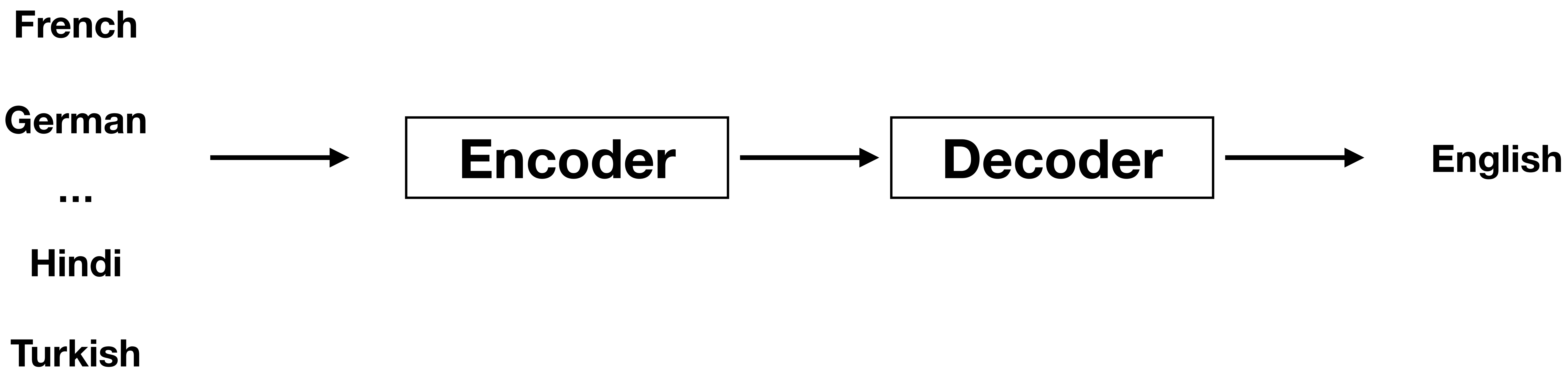
highest ⟶ h i g h e s t </w> ⟶ h i g h est</w> ⟶ h i gh est</w>

Byte Pair Encoding: Pros and Cons

- **Pros**
  - Trade-off between char/byte -level tokens and original words
  - Capturing shared morphemes/sub-words across similar languages
  - Usually no *unknown* words, unless meeting special/uncommon characters
  - Not only for multilingual tasks, but also for monolingual ones
- **Cons**
  - Shallow similarity, working well only on similar languages
  - Over-segment low-resource or morphologically rich languages

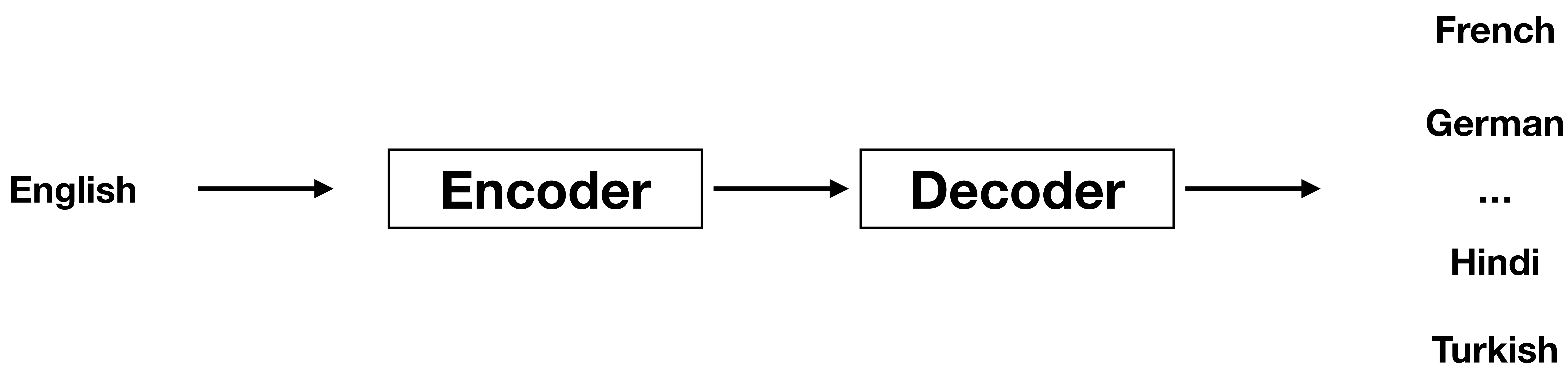
Many-to-One NMT

- Training a single Encoder-Decoder model from multiple languages to English



One-to-many?

One-to-Many NMT



Given an English sentence, how could we know which language we want to translate to?

One-to-Many NMT

English → Encoder → Decoder →

French

German

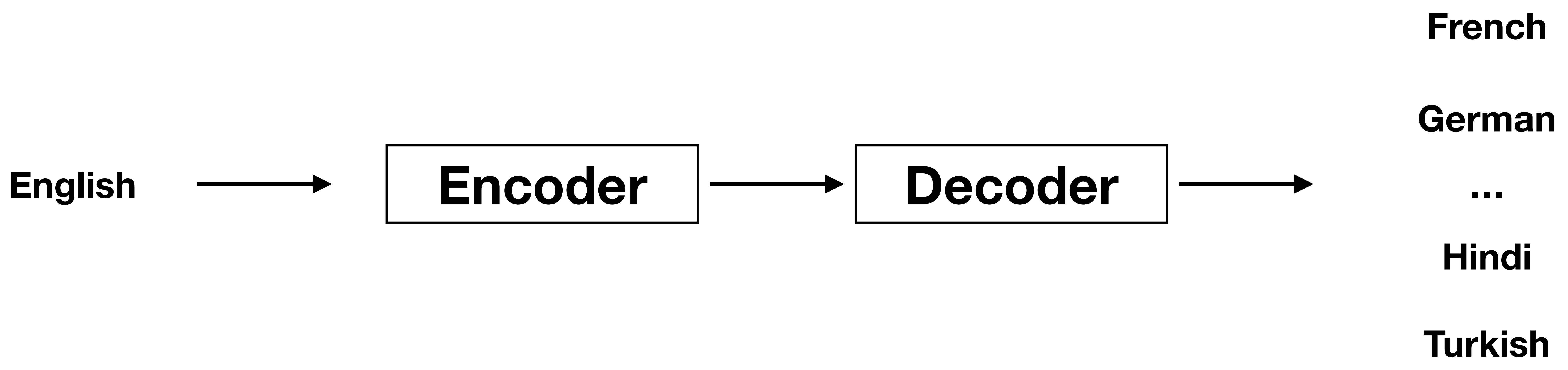
...

Hindi

Turkish

	French	Comment ça va?
How are you?	Spanish	cómo estás?
	Turkish	nasılsın?

One-to-Many NMT



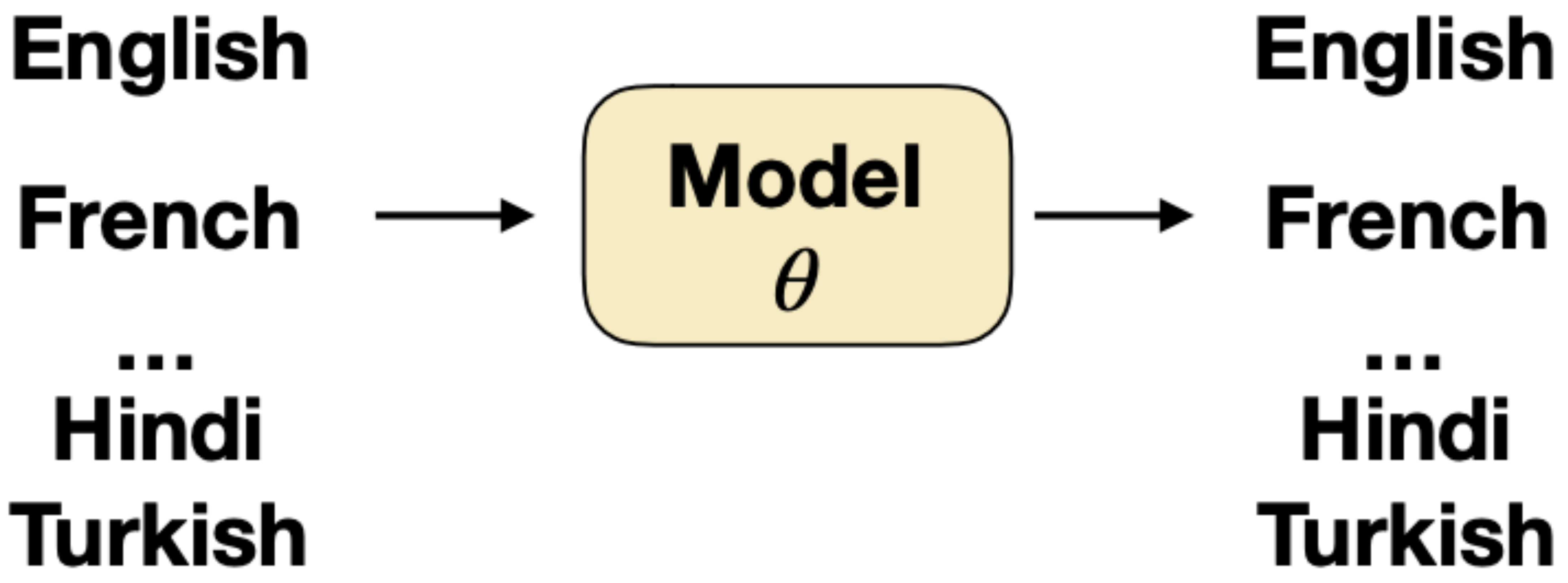
<2fr> How are you? French Comment ça va?

<2es> How are you? Spanish cómo estás?

<2tr> How are you? Turkish nasılsın?

Many-to-Many NMT

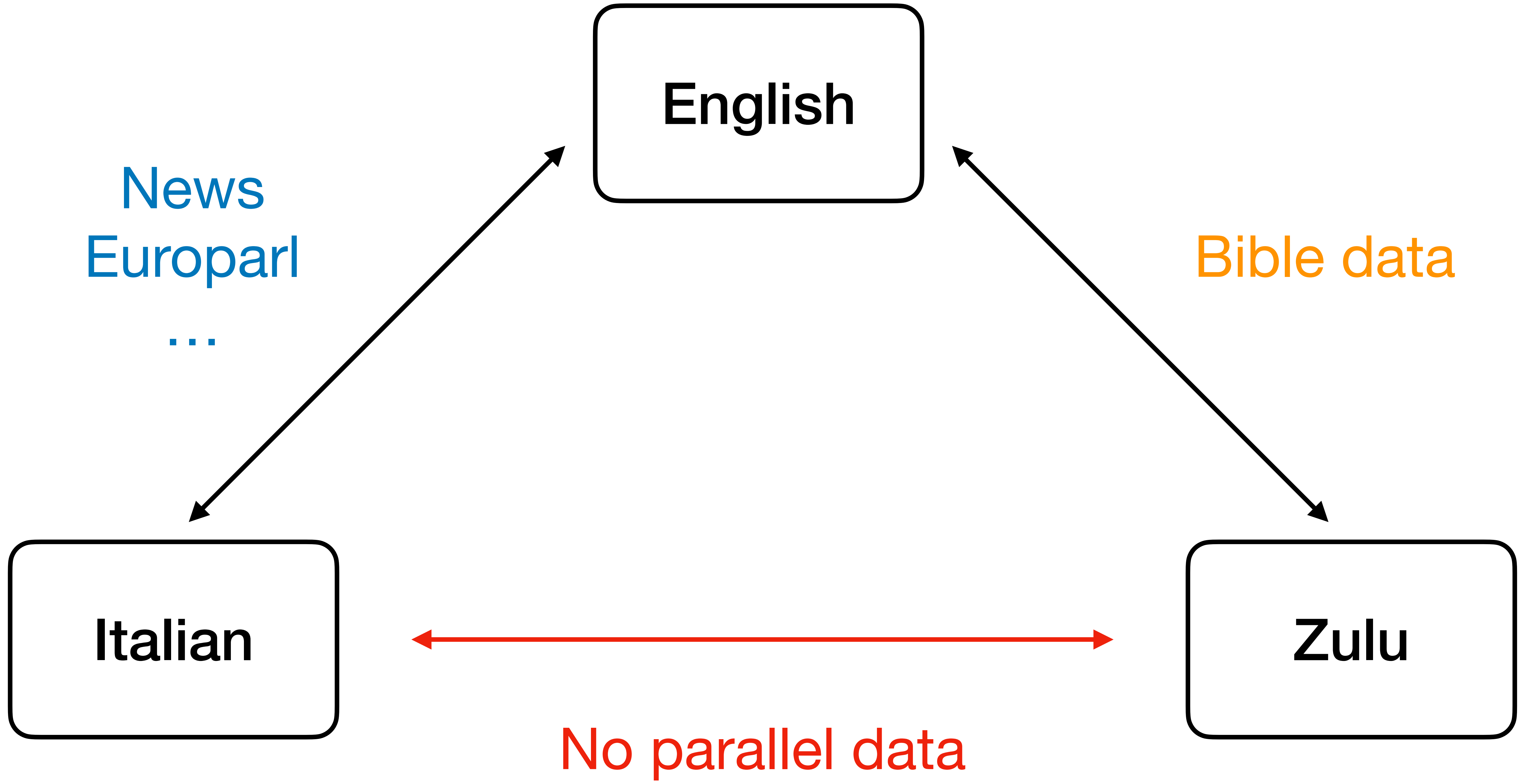
- **Is English always a good pivot language?**
  - Chinese-Japanese
  - Spanish-Portuguese
- **Can we do many-to-many translation?**
  - Training a single model on a mixed dataset from multiple language pairs



Google's multilingual neural machine translation system. (Johnson et al., 2016)

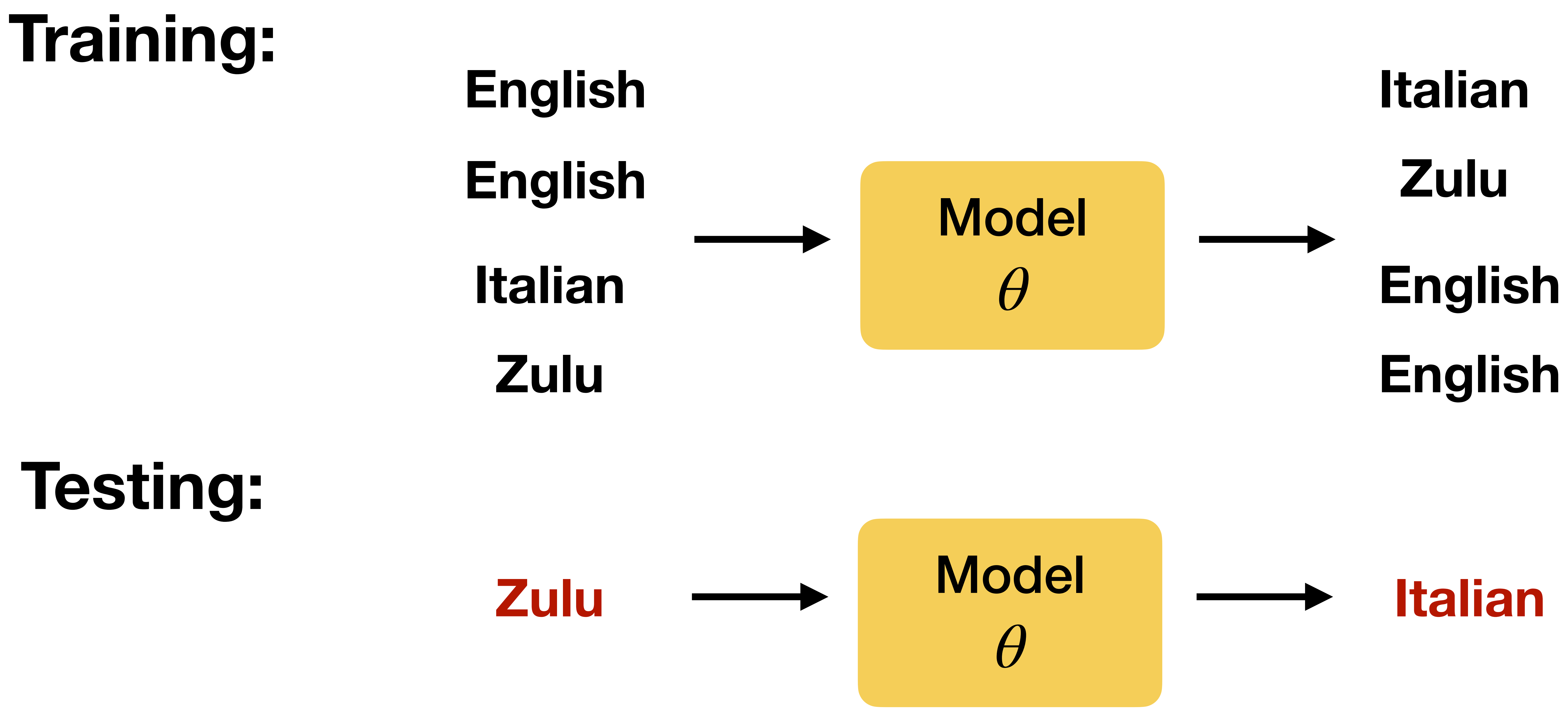
Many-to-Many NMT: Zero-shot Transfer

- Not all language pairs have parallel data



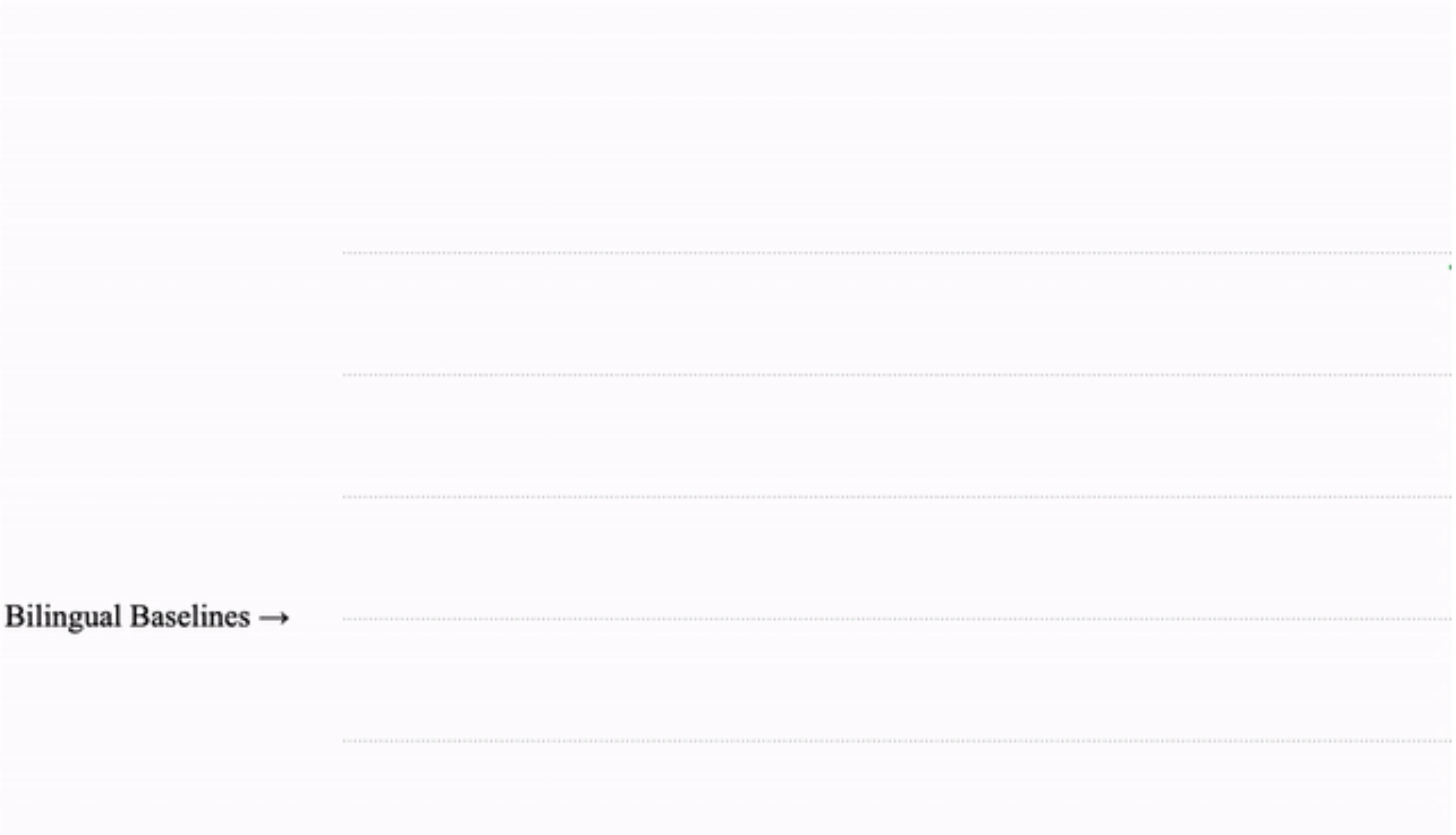
Many-to-Many NMT: Zero-shot Transfer

- Training on {English-Zulu, Zulu-English, English-Italian, Italian-English}
- Zero-shot transfer: the model can translate directly between Zulu and Italian



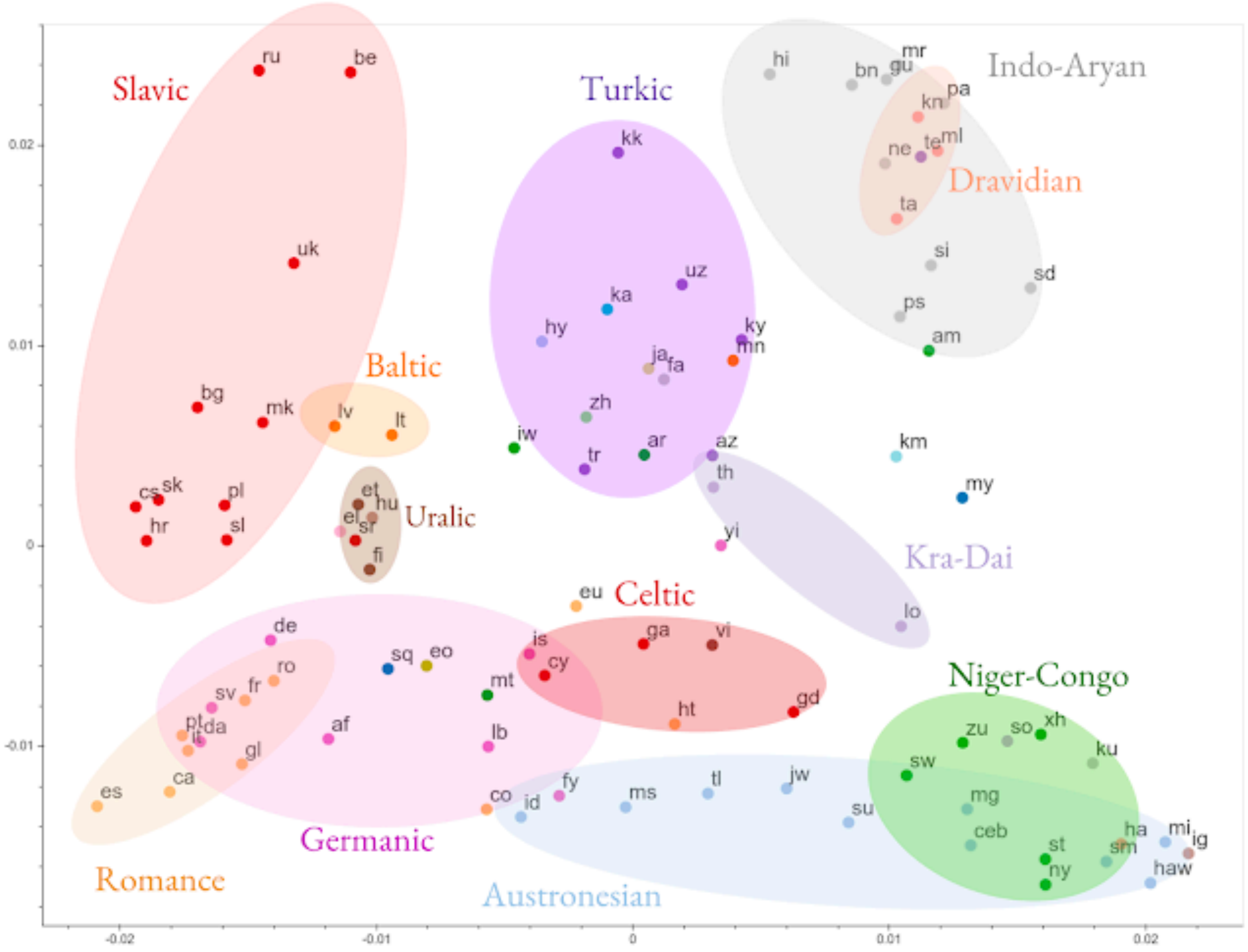
Google's multilingual neural machine translation system. (Johnson et al., 2016)

Many-to-Many NMT



Google's multilingual neural machine translation system. (Arivazhagan et al., 2019)

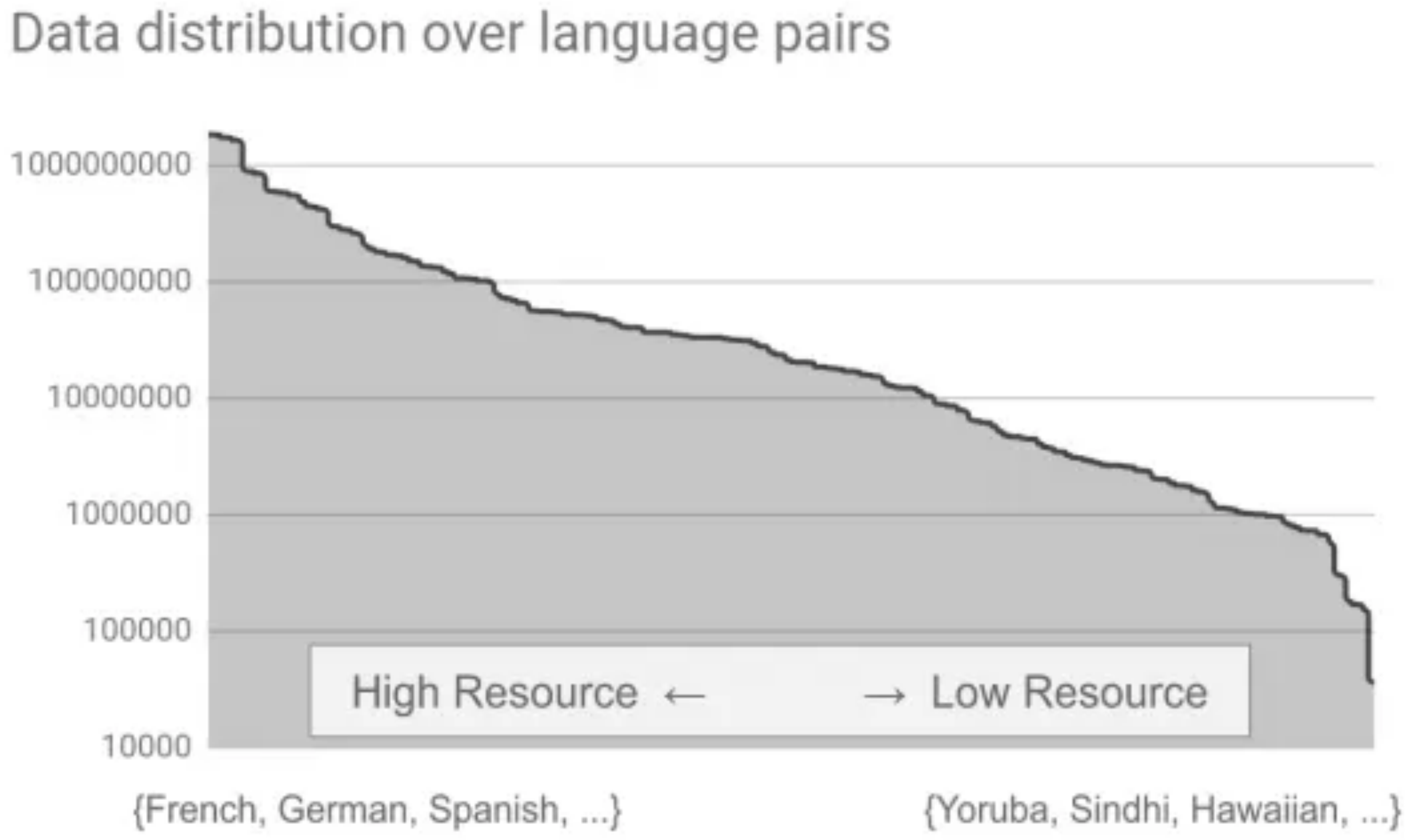
Many-to-Many NMT



Google's multilingual neural machine translation system. (Arivazhagan et al., 2019)

Open Problems for Multilingual NMT

- **Imbalanced training data**
  - Important to upsample low-resource data



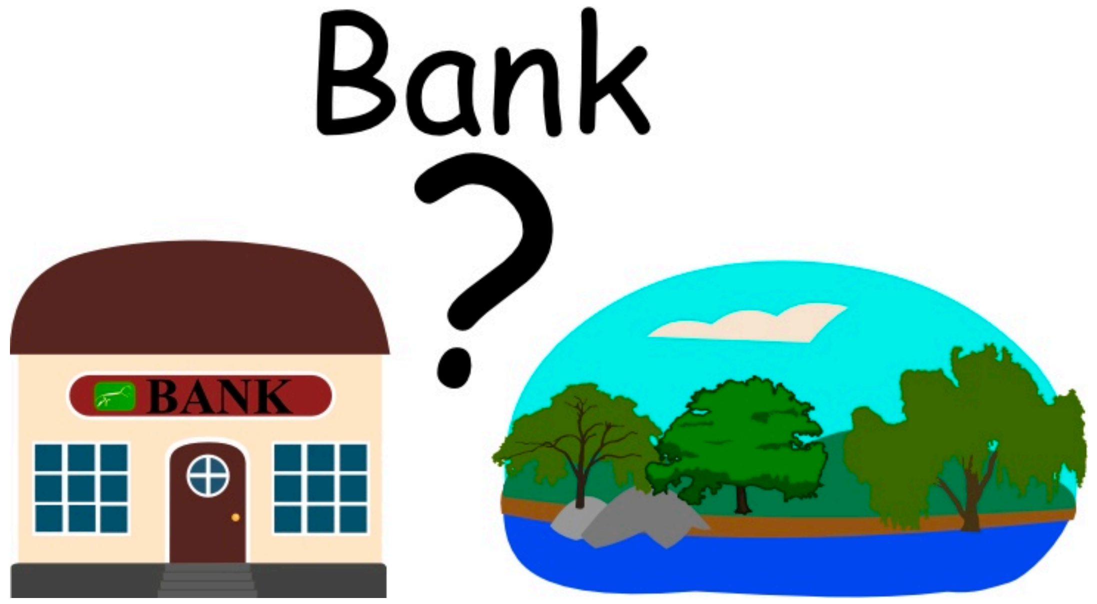
- **Underperforming bilingual models**
  - Degrading high-resource languages (Arivazhagan et al., 2019)
- **Vocabulary shared across many languages**
  - Upsampling low-resource languages and run joint BPE on all languages
  - Over-segment low-resource or morphologically rich languages
- **Multilingual Evaluation**
  - Average BLEU over all languages or BLUE for the worst case?
  - Are BLUE scores between two languages comparable?

Context-Aware NMT

Why Context-Aware NMT?

He is now at the **bank**. → 他正在银行。

He is now at the **bank**. → 他正在岸边。



Ambiguous Word in Translation!

Why Context-Aware NMT?

Ambiguous Word in Translation!

Why Context-Aware NMT?

The cat, **it** is tired	→	**Die** Katze, **sie** ist müde.
The dog, **it** is tired.	→	**Der** Hund, **er** ist müde.

Animals can be masculine or feminine in German

the table, **it** is nice	→	**der** Tisch, **er** ist schön.
the sun, **it** is warm	→	**die** Sonne, **sie** ist warm.
the museum, **it** is open	→	**das** Museum, **es** ist offen.

English objects don't have a gender, German ones do!

Discourse Phenomenon

Coreference
The cat and the actor were hungry, **it** was hungrier. → …… , **Sie** was hungrier.

Ambiguity
他睡过了 → a) He overslept b) He already slept

Lexical Cohesion
repetition of the same named entities

Pronoun Ellipsis
(她)来了 → She came

Discourse Marker Ellipsis

Tense

What is Context-Aware NMT?

Given source document **X**, target document **Y**,

Problem: $P_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i, C_i) = \prod_{j=1}^{N} P_\theta(y_i^j|\boldsymbol{x}_i^j, y_i^{<j}, C_i),$ context in the document

- **Context can be all available sentences in the document.**
- **Or a few neighboring sentences.**
  - 1-2: $C_i = \{y_{i-1}\}$
  - 2-2: $C_i = \{x_{i-1}, y_{i-1}\}$
  - 3-1: $C_i = \{x_{i-1}, x_{i+1}\}$
  - …..

SOURCE	*TARGET*
sieh , Bob ! _BREAK_ -Wo sind sie ?	look , Bob ! _BREAK_ - Where are they ?

(Tiedemann, 2017)

Challenges in Context-Aware Neural Machine Translation

Linghao Jin[†1] **Jacqueline He**[†2] **Jonathan May**[1] **Xuezhe Ma**[1]
[1]Information Sciences Institute, University of Southern California
[2]University of Washington
{linghaoj, jonmay, xuezhema}@isi.edu jyyh@cs.washington.edu

Challenges in Context-Aware NMT

Discourse phenomena is sparse in surrounding context.

Context does not help disambiguate certain discourse phenomena.

The context-agnostic baseline performs comparably to context-aware settings.

Advanced model architectures do not meaningfully improve performance.

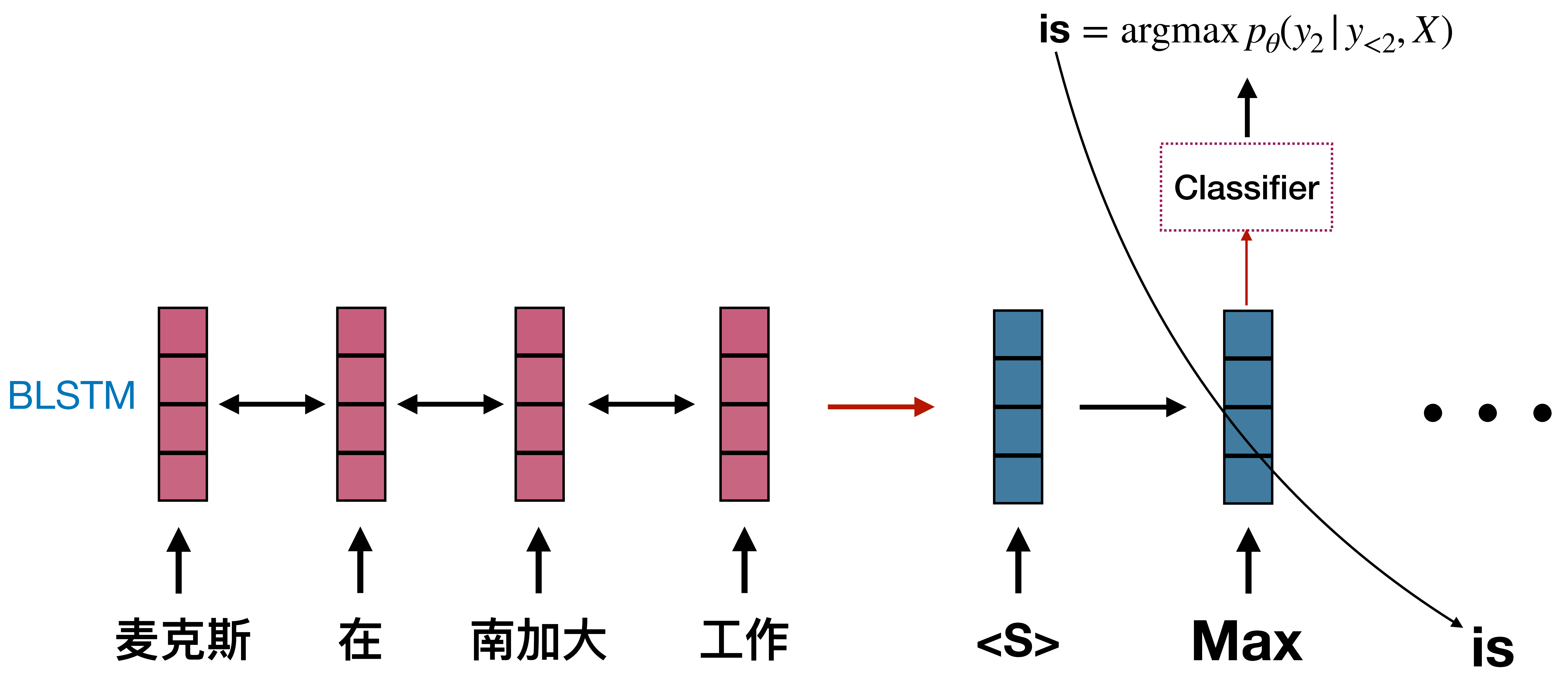
There is a need for an appropriate document-level translation metric.

Non-Autoregressive NMT

Auto-Regressive Decoding

- **Greedy decoding:**

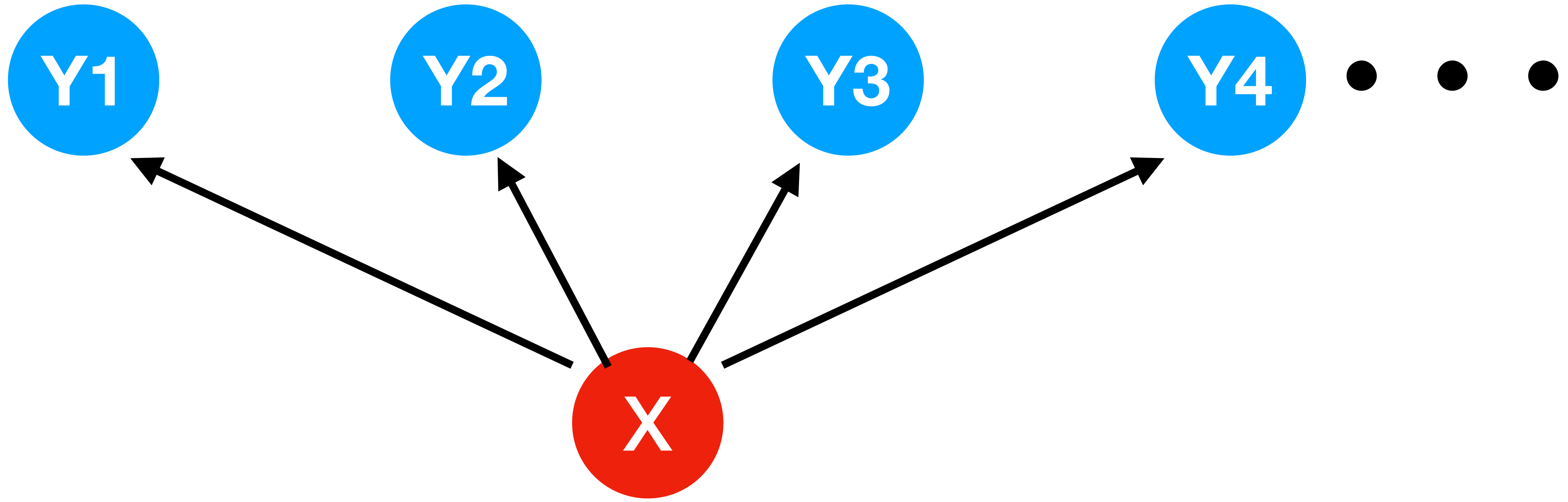
$$y_t{}^* = \arg\max_{y_t} p_\theta(y_t \mid y_{<t}, X), \forall t$$



Non-Autoregressive MT?

- A naïve solution:

$$p_\theta(Y|X) = \prod_{t=1}^{T} p_\theta(y_t|X)$$



Too Strong Independent Assumption!

Non-Autoregressive MT

- **Iterative Decoding**
  - Mask-predict (Ghazvininejad et al., 2019)
- **Latent Variable Model**
  - Flowseq (Ma et al., 2019)

Iterative Decoding

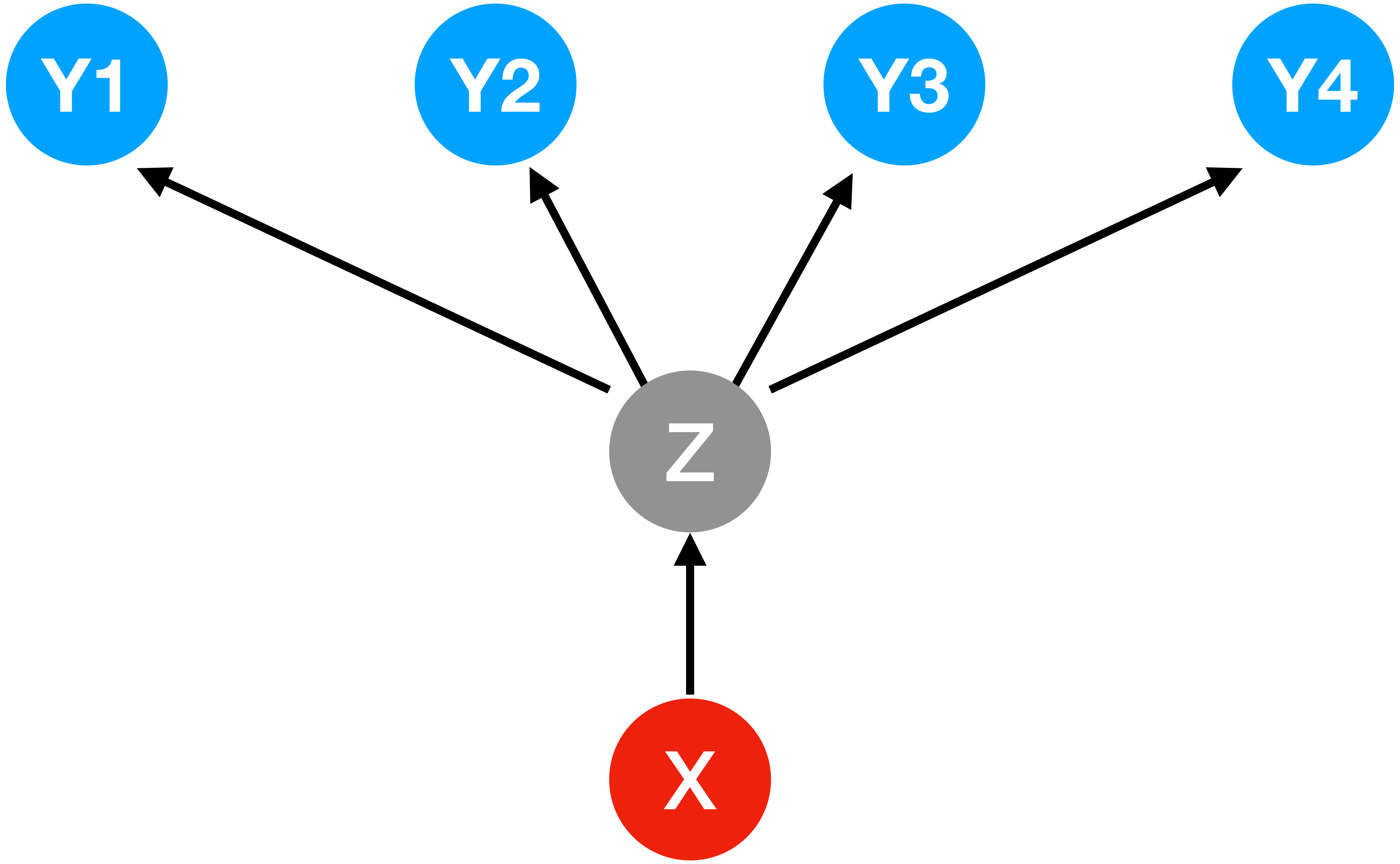
- **Key Idea: iteratively refine translations**
  - First using the naive model to obtain a low-quality translation
  - Detecting and deleting the words with low confidence
  - Re-predicting the deleted words with remaining words in the translation as context

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Latent Variable Models

Latent Variable Z $$p_\theta(Y|X) = \int_Z p_\theta(Y|Z,X)p_\theta(Z|X)dz,$$

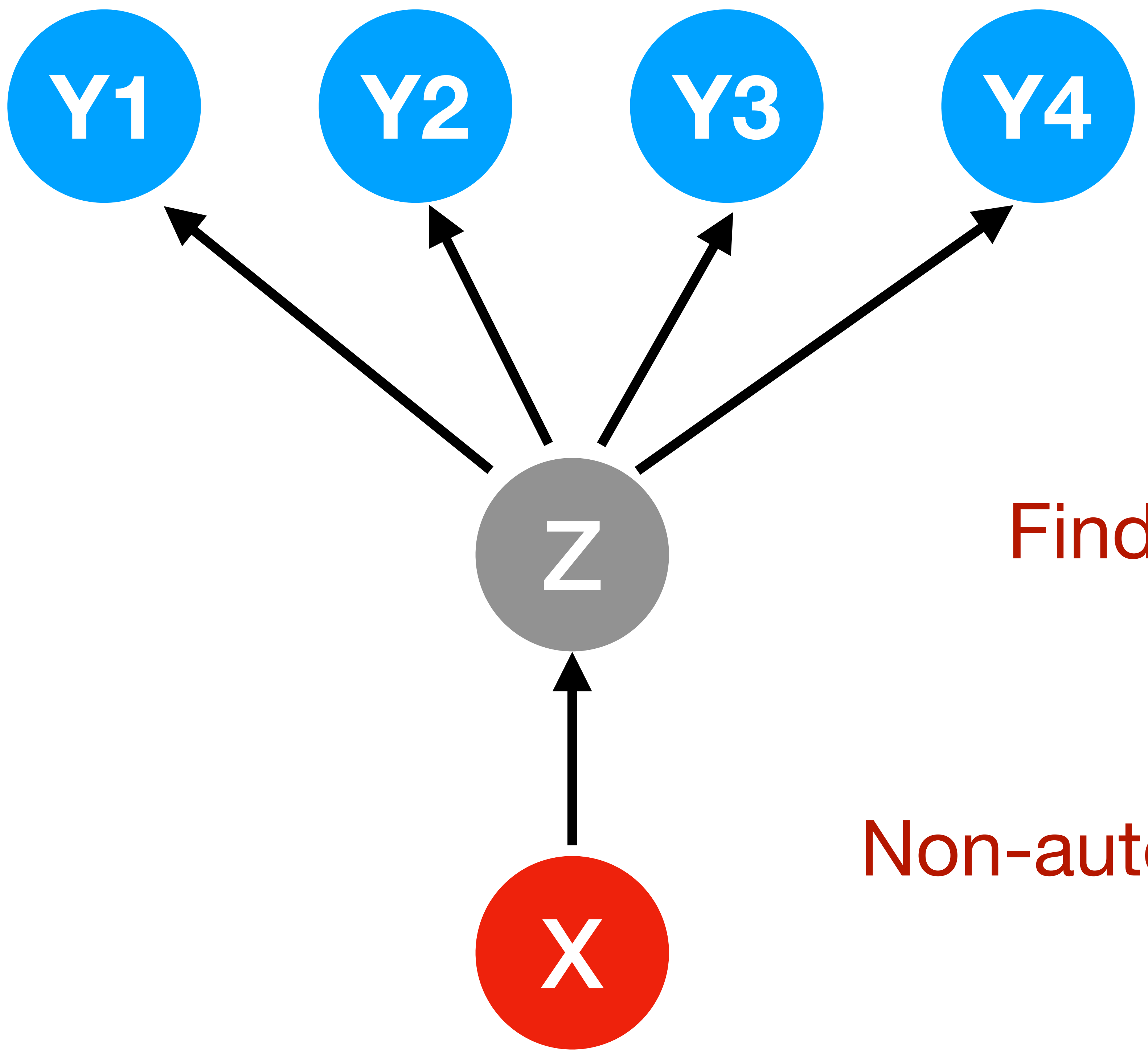
Non-Autoregressive $$p_\theta(Y|Z,X) = \prod_{t=1}^{T} p_\theta(y_t|Z,X)$$



Latent Variable Models

Latent Variable Z $$p_\theta(Y|X) = \int_Z p_\theta(Y|Z,X)p_\theta(Z|X)dz,$$

Non-Autoregressive $$p_\theta(Y|Z,X) = \prod_{t=1}^{T} p_\theta(y_t|Z,X)$$



Advantages:

- No direct independent assumptions between X and Y
- Efficient Decoding:

Find optimal Z $$z^* = \text{argmax}_{z\in\mathcal{Z}} p_\theta(z|x)$$

$$y^* = \text{argmax}_{y\in\mathcal{Y}} p_\theta(y|z^*,x)$$

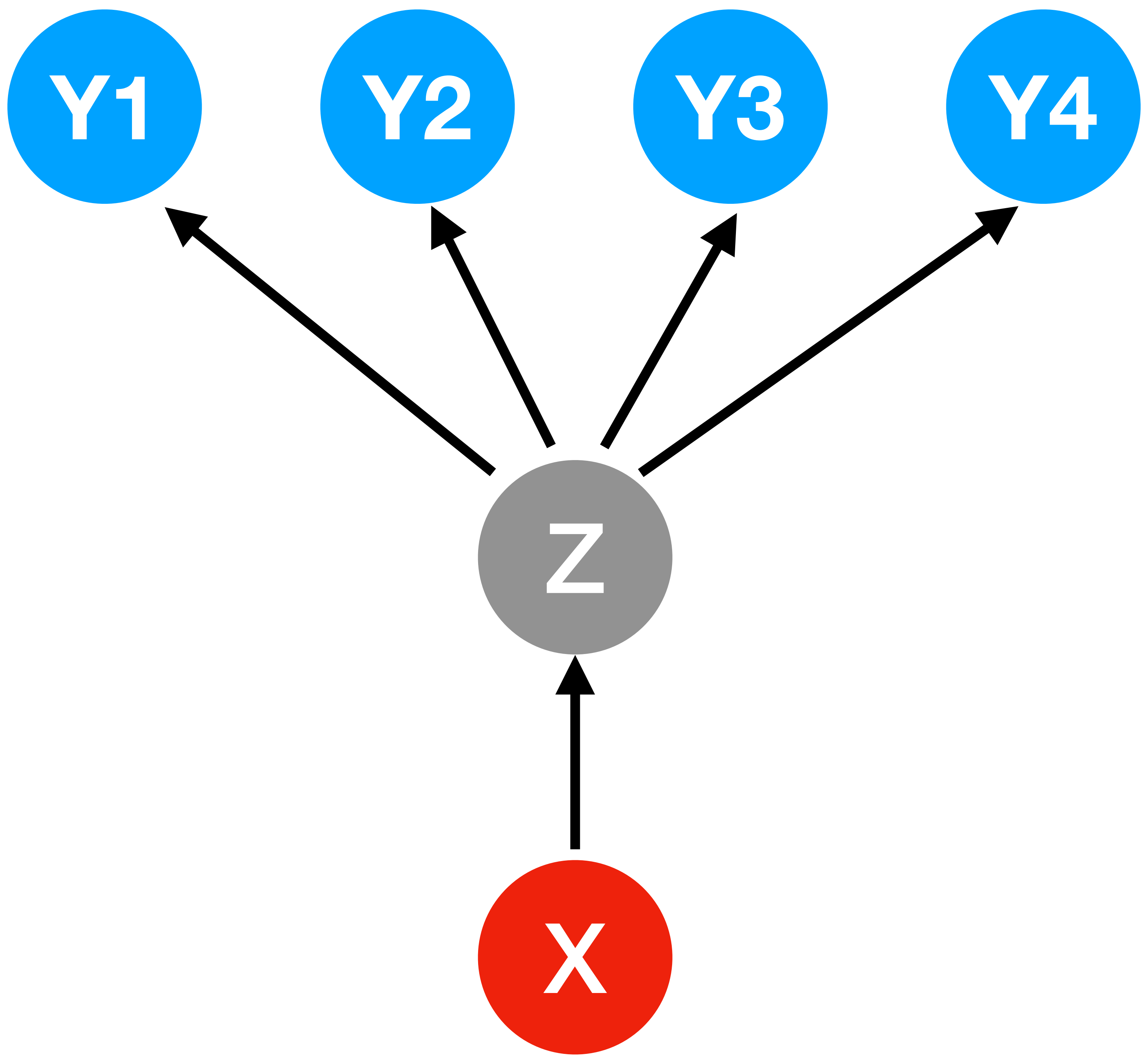
Non-autoregressive $\Longleftrightarrow$

$$y_t^* = \text{argmax}_{y_t\in V} p_\theta(y_t|z^*,x), \forall t$$

Latent Variable Models

Latent Variable Z $\quad p_\theta(Y|X) = \boxed{\int_Z p_\theta(Y|Z,X)p_\theta(Z|X)dz,}$

Non-Autoregressive $\quad p_\theta(Y|Z,X) = \boxed{\prod_{t=1}^{T} p_\theta(y_t|Z,X)}$



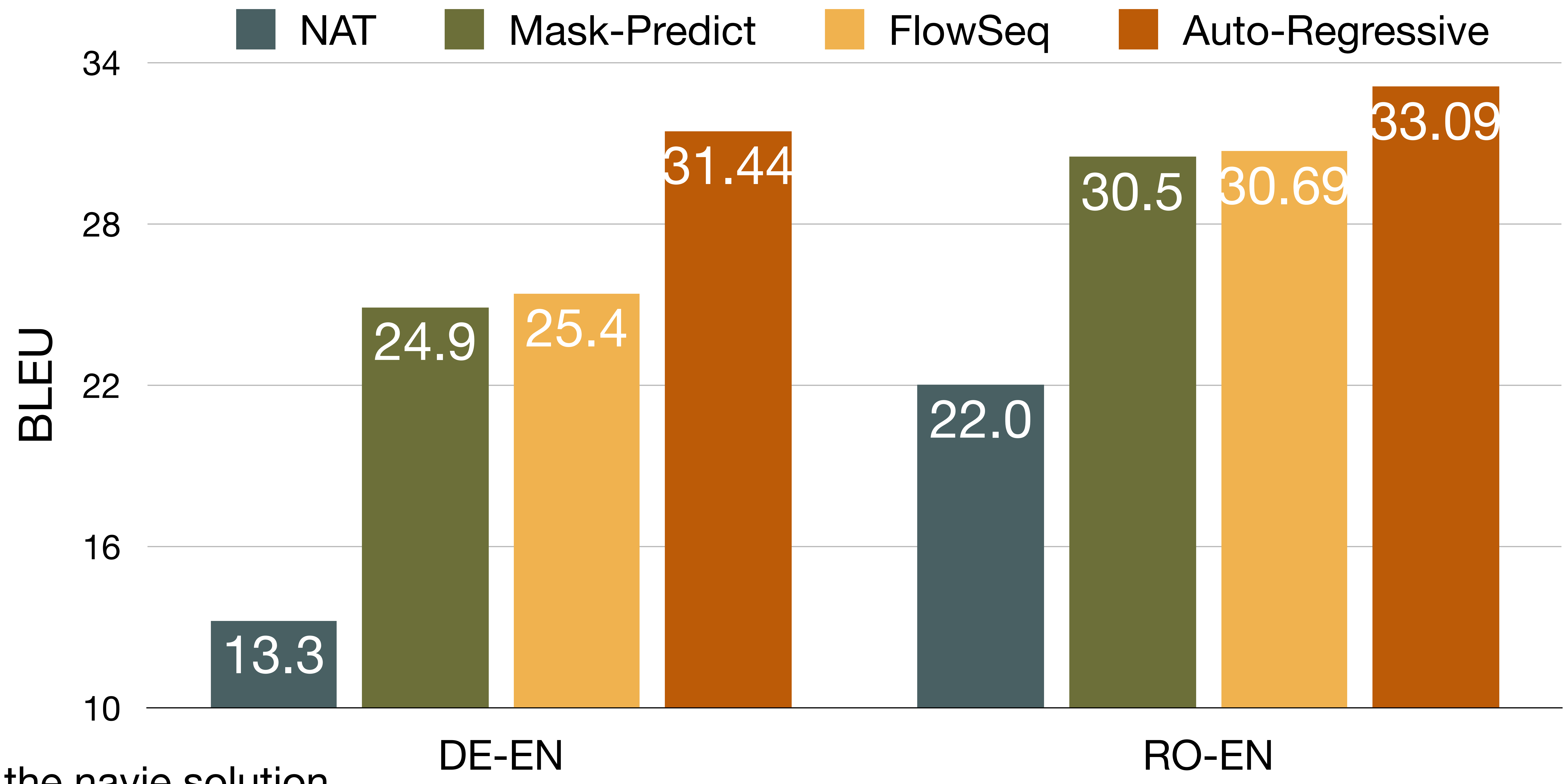
Problems:

- How to compute the integral of $p_\theta(Y|X)$?
  - Variational Inference
- Z needs to be sufficiently expressive to encode all the structured dependencies of Y
  - Generative Flow

Not Today!

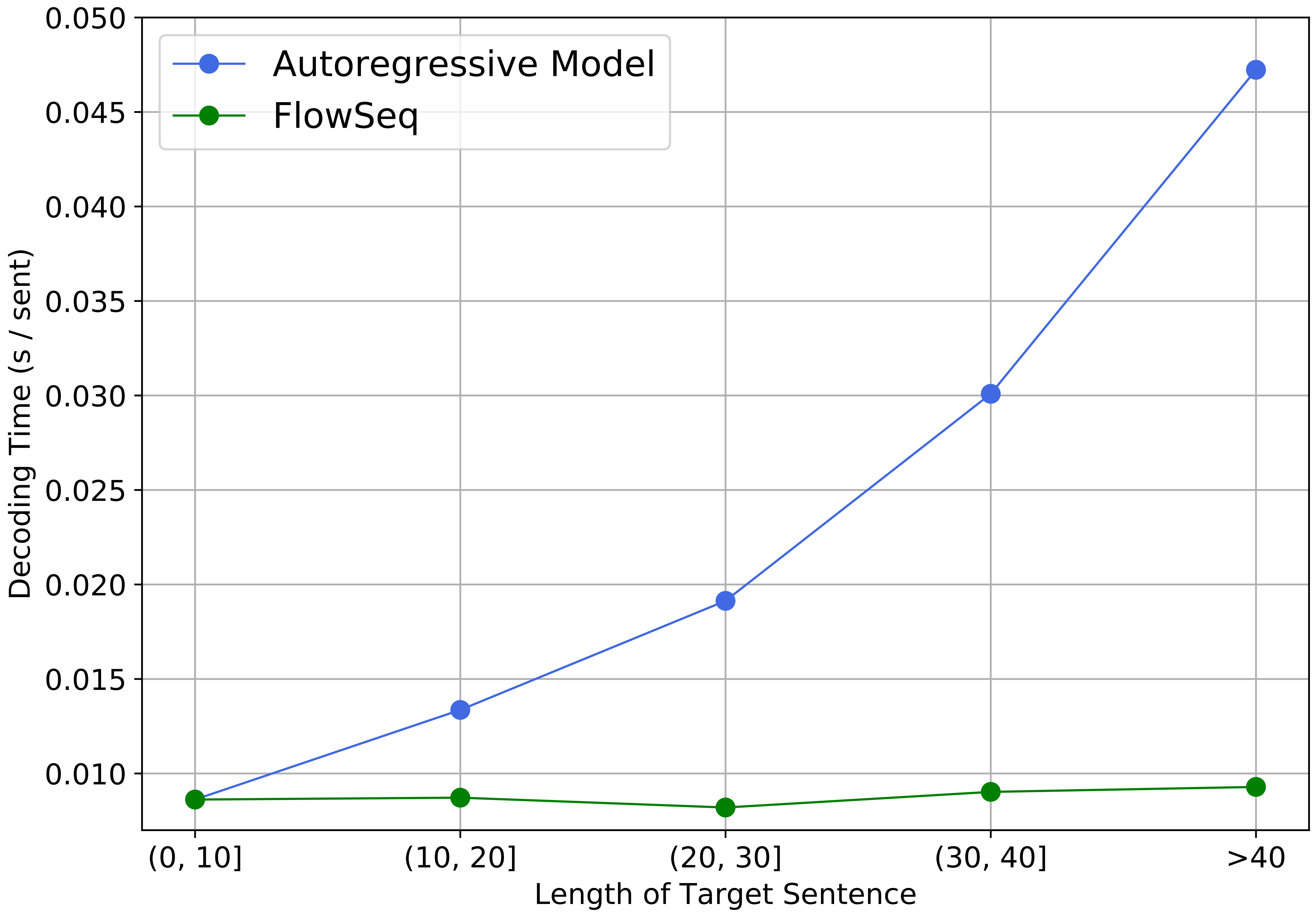
Evaluation: Machine Translation

- **Data**
  - WMT14: German to English (DE-EN)
  - WMT16: Romanian to English (RO-EN)

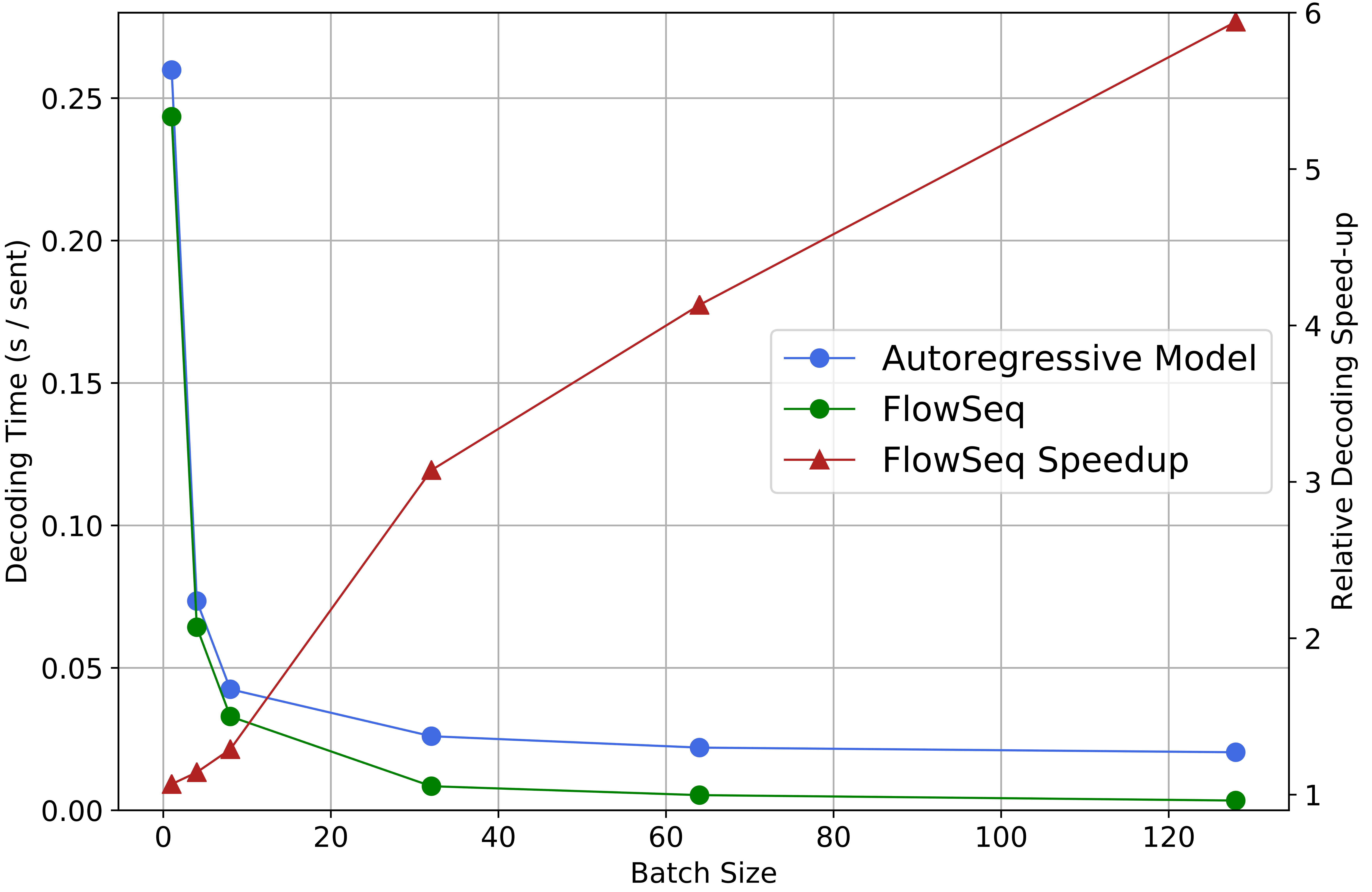


NAT: similar to the navie solution

Translation Speed: Sentence Length



Translation Speed: Batch Size



Evaluation beyond BLEU

MT Evaluation

- **Criterion:**
  - Adequacy: measure of correctness
  - Fluency: measure of naturalness
  - Other aspects: hallucination? Coverage?

MT Evaluation

- **Criterion:**
  - Adequacy: measure of correctness
  - Fluency: measure of naturalness
  - Other aspects: hallucination? Coverage?
- **Automatic Evaluation**
  - BLEU score (Papineni et al., 2002): n-gram based metric
    - N-gram precision
    - Brevity penalty
  - Other metrics:
    - METEOR (Denkowski et al., 2014)
    - Translation Edit Rate (TER) (Snover et al., 2006)

MT Evaluation

- **Drawbacks of n-gram based metrics (Zhang et al., 2020):**
  - Penalize semantically-correct paraphases due to string matching
    - e.g. “No worries!” and “Don’t worry!”
  - Fail to capture distant dependencies and penalize semantically-critical ordering changes or word drops(Isozaki et al., 2010)
    - e.g. “A because B” and “B because A”
    - e.g. “A loves B” vs. “A hates B”, ref: “A likes B”

MT Evaluation

- **Drawbacks of n-gram based metrics:**
  - Penalize semantically-correct paraphases due to string matching (Banerjee & Lavie, 2005)
  - Fail to capture distant dependencies and penalize semantically-critical ordering changes or word drops(Isozaki et al., 2010)
- **Recent Proposed Metrics: contextualized embedding based metrics**
  - BERTScores (Zhang et al., 2020)
    - Computes token similarity using contextual embedding between candidate and reference
  - BLEURT (Stellam et al., 2020)
    - A learned evaluation metric based on BERT
  - Better aligned with human preferences
    - Not interpretable