

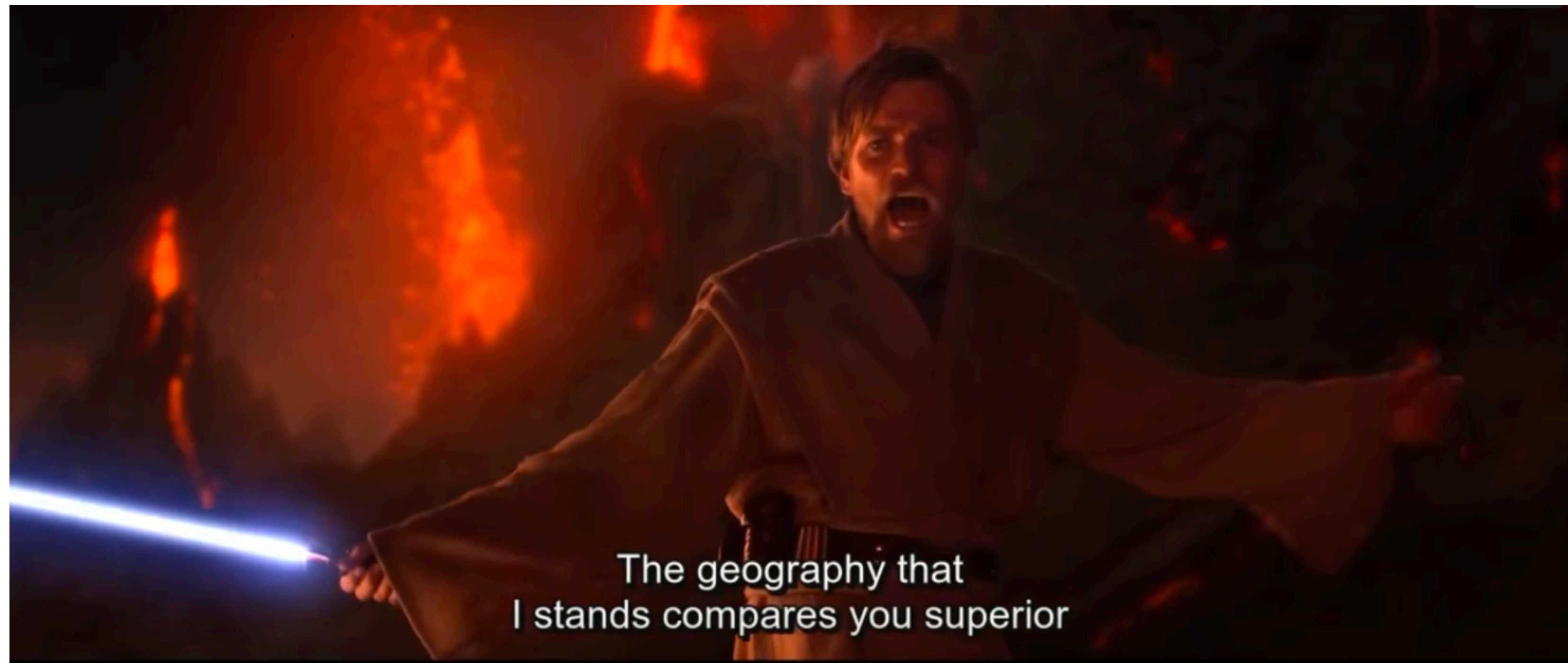
CS371N: Natural Language Processing

Lecture 24: Machine Translation

Greg Durrett



TEXAS
The University of Texas at Austin



Star Wars The Third Gathers: The Backstroke of the West
(subtitles machine translated from Chinese)



Administrivia

- ▶ FP check-ins due Friday
- ▶ Course evaluations: **submit proof for extra credit on final project**
- ▶ Greg's office hours Tuesday at 5pm (normal), Thursday at 1pm (modified this week)
- ▶ A4 back soon, A5 grading underway



Today's Lecture

- ▶ MT basics
- ▶ Phrase-based MT, word alignment
- ▶ Multilingual and cross-lingual models
- ▶ MT frontiers

MT Basics



MT in Practice

- ▶ Bitext: this is what we learn translation systems from. What can you learn?

Je fais un bureau

I'm making a desk

Je fais une soupe

I'm making soup

Je fais un bureau

I make a desk

Qu'est-ce que tu fais?

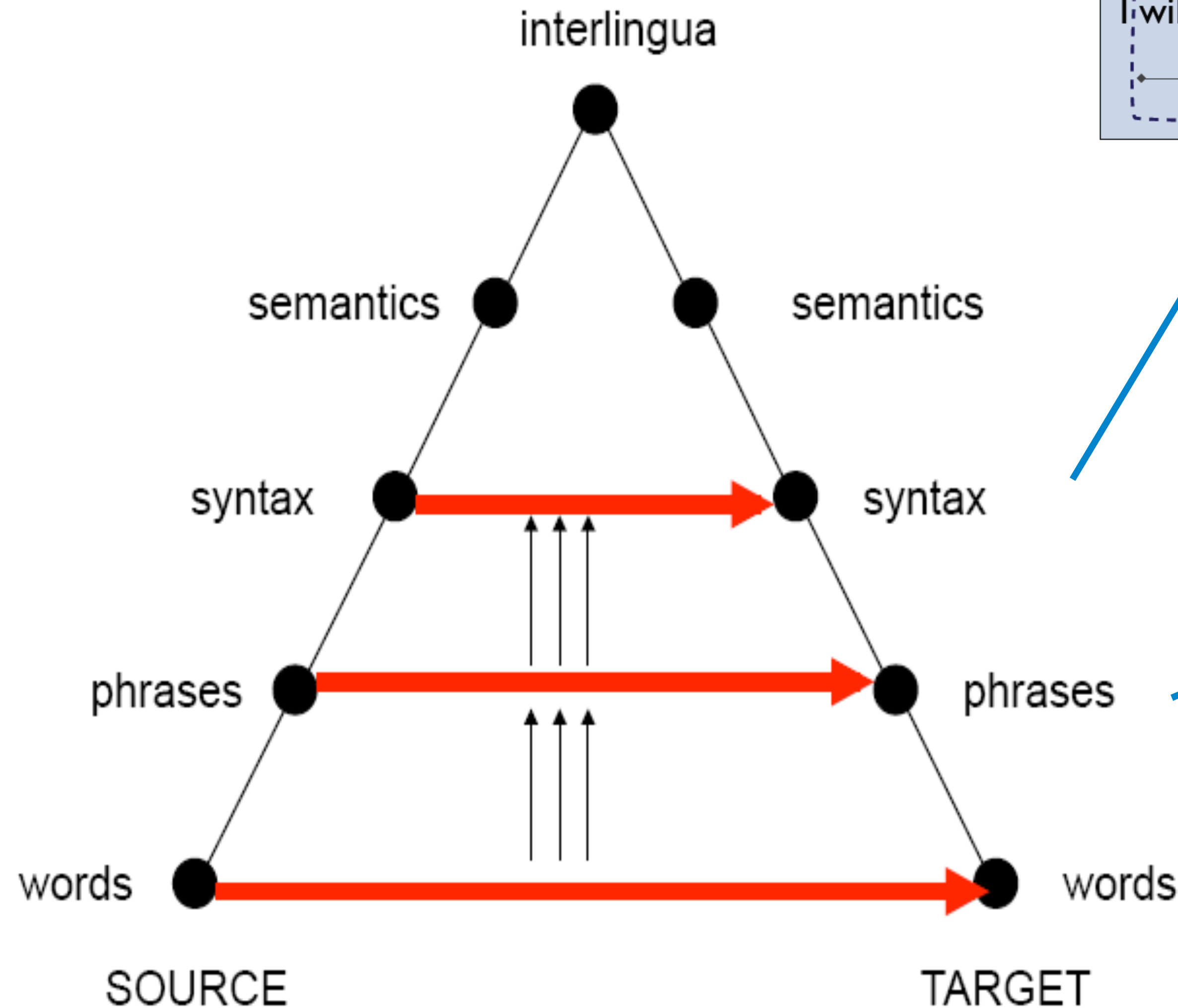
What are you doing?

- ▶ What makes this hard? Not word-to-word translation
Multiple translations of a single source (ambiguous)



Levels of Transfer: Vauquois Triangle

Bernard Vauquois (1968)



Yo lo haré mañana
I will do it tomorrow

NP
VP

$$P\left(\begin{array}{c} \text{MD} \\ | \\ \text{will} \end{array} \begin{array}{c} \text{VP} \\ / \quad \backslash \\ \text{VB} \quad \text{PRN} \quad \boxed{\text{NP}} \\ | \quad | \\ \text{do} \quad \text{it} \end{array} \mid \begin{array}{c} \text{VP} \\ / \quad \backslash \\ \text{lo} \quad \text{haré} \quad \boxed{\text{NP}} \end{array} \right) = 0.8$$

Yo lo haré mañana
I will do it tomorrow

English (E)	$P(E \mid \text{lo haré})$
will do it	0.8
will do so	0.2

Yo lo haré mañana
~~I will do it tomorrow~~

English (E)	$P(E \mid \text{mañana})$
tomorrow	0.7
morning	0.3

- ▶ Classic systems were mostly phrase-based

Slide credit: Dan Klein



Evaluating MT

- ▶ What should our evaluation goals be?



Evaluating MT

- ▶ Fluency: does it sound good in the target language?
- ▶ Fidelity/adequacy: does it capture the meaning of the original?
- ▶ Classic automatic metric: BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram *precision vs.* a reference, multiplied by brevity penalty (penalizes short translations)

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad \text{Typically } n = 4, w_i = 1/4$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad \begin{array}{l} r = \text{length of reference} \\ c = \text{length of prediction} \end{array}$$

- ▶ Which of these criteria does it capture?

Phrase-based MT, Word Alignment



Phrase-Based MT

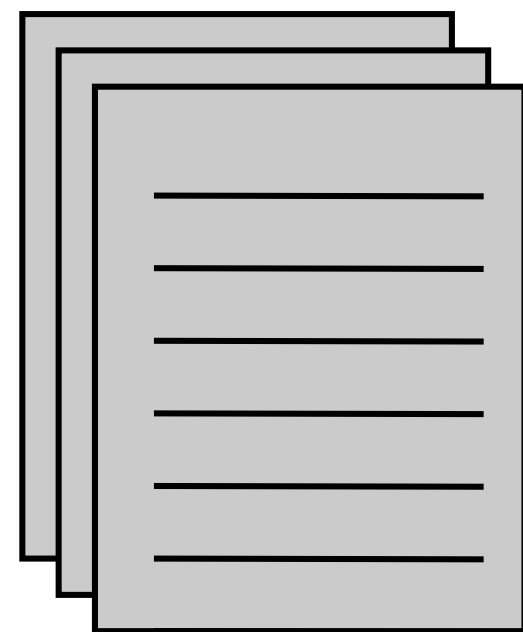
- ▶ Key idea: translation works better the bigger chunks you use
- ▶ Remember phrases from training data, translate piece-by-piece and stitch those pieces together to translate
 - ▶ How to identify phrases? Word alignment over source-target bitext
 - ▶ How to stitch together? Language model over target language
 - ▶ Decoder takes phrases and a language model and searches over possible translations
- ▶ NOT like standard discriminative models (take a bunch of translation pairs, learn a ton of parameters in an end-to-end way)



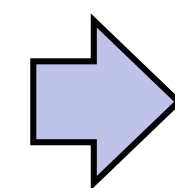
Phrase-Based MT

```
cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
...
```

Phrase table $P(f|e)$



Unlabeled English data



Language model $P(e)$

- ▶ Where does the phrase table come from? First need **word alignment**

$$P(e|f) \propto P(f|e)P(e)$$

Noisy channel model:
combine scores from
translation model +
language model to
translate foreign to
English

“Translate faithfully but make fluent English”



Word Alignment

- ▶ Input: a bitext, pairs of translated sentences

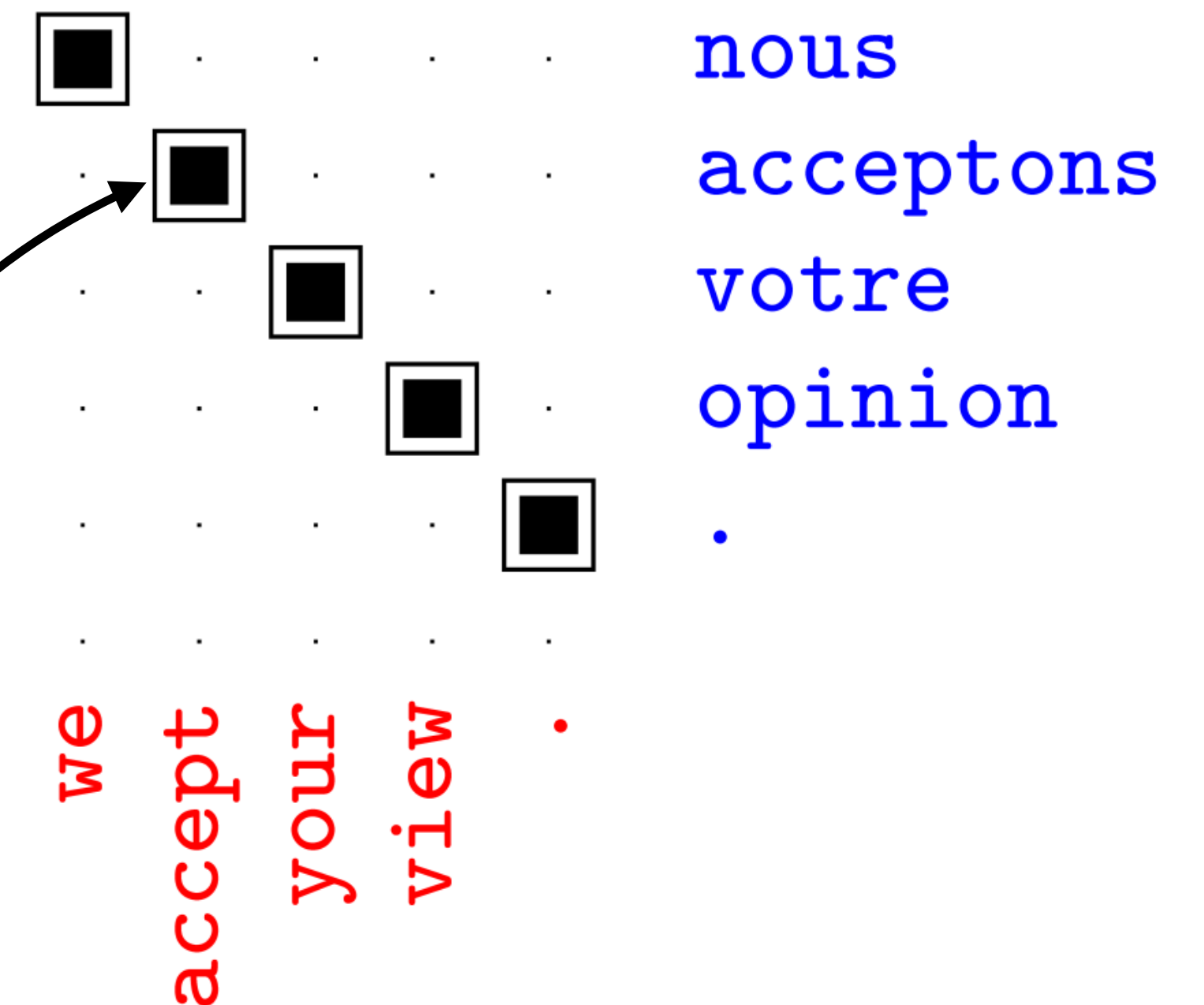
nous acceptons votre opinion . ||| we accept your view

nous allons changer d'avis ||| we are going to change our minds

- ▶ Output: alignments between words in each sentence

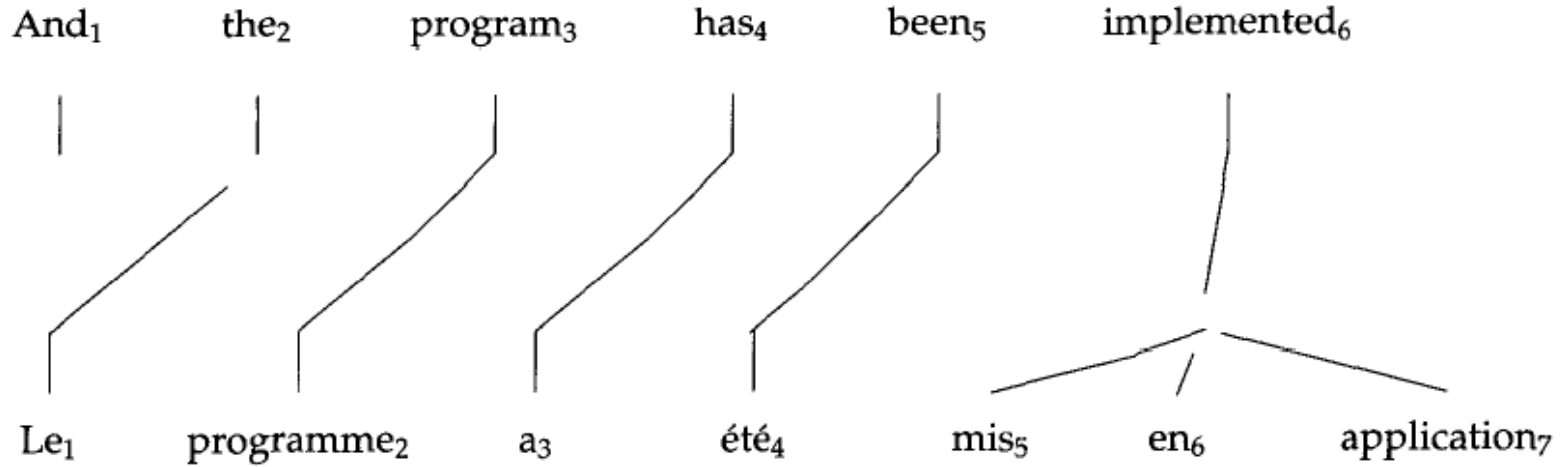
- ▶ We will see how to turn these into phrases

“accept and acceptons are aligned”





1-to-Many Alignments





Word Alignment

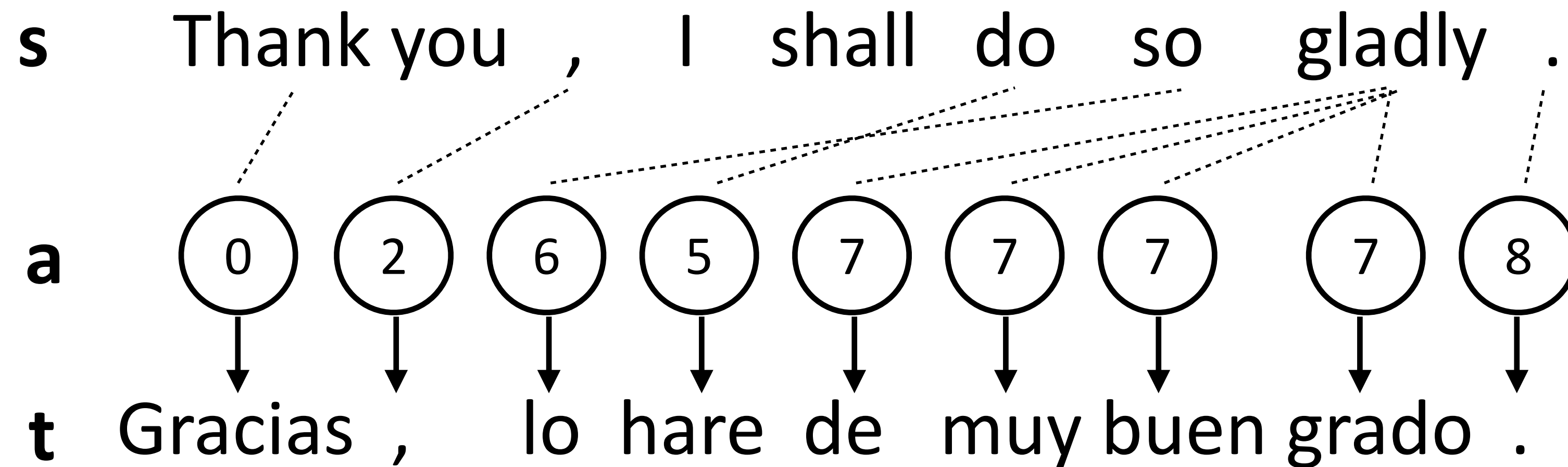
- ▶ Models $P(\mathbf{t}|\mathbf{s})$: probability of “target” sentence being generated from “source” sentence according to a model
- ▶ Latent variable model:
$$P(\mathbf{t}|\mathbf{s}) = \sum_{\mathbf{a}} P(\mathbf{t}|\mathbf{a}, \mathbf{s})P(\mathbf{a})$$
- ▶ Correct alignments should lead to higher-likelihood generations, so by optimizing this objective we will learn correct alignments



IBM Model 1

- ▶ Each target word is aligned to *at most* one source word

$$P(\mathbf{t}, \mathbf{a} \mid \mathbf{s}) = \prod_{i=1}^n P(t_i \mid s_{a_i}) P(a_i)$$



- ▶ Set $P(\mathbf{a})$ uniformly (no prior over good alignments)
- ▶ $P(t_i \mid s_{a_i})$: word translation probability table. Learn with EM
Brown et al. (1993)



IBM Model 1: Example

$$P(\mathbf{t}, \mathbf{a} \mid \mathbf{s}) = \prod_{i=1}^n P(t_i \mid s_{a_i}) P(a_i)$$

	I	like	eat	$\mathbf{s} = \text{Je}$	NULL
Je	0.8	0.1	0.1	$\mathbf{t} = \text{I}$	
J'	0.8	0.1	0.1		
mange	0	0	1.0		
aime	0	1.0	0		
NULL	0.4	0.3	0.3		

What is $P(\mathbf{t}, \mathbf{a} \mid \mathbf{s})$?

What is $P(\mathbf{a} \mid \mathbf{t}, \mathbf{s})$?



IBM Model 1: Example 2

$$P(\mathbf{t}, \mathbf{a} \mid \mathbf{s}) = \prod_{i=1}^n P(t_i \mid s_{a_i}) P(a_i)$$

	I	like	eat	s = J'	aime	NULL
Je	0.8	0.1	0.1	t = I	like	
J'	0.8	0.1	0.1			
mange	0	0	1.0			
aime	0	1.0	0			
NULL	0.4	0.3	0.3			

What is $P(a_1 \mid \mathbf{t}, \mathbf{s})$?



Learning with EM

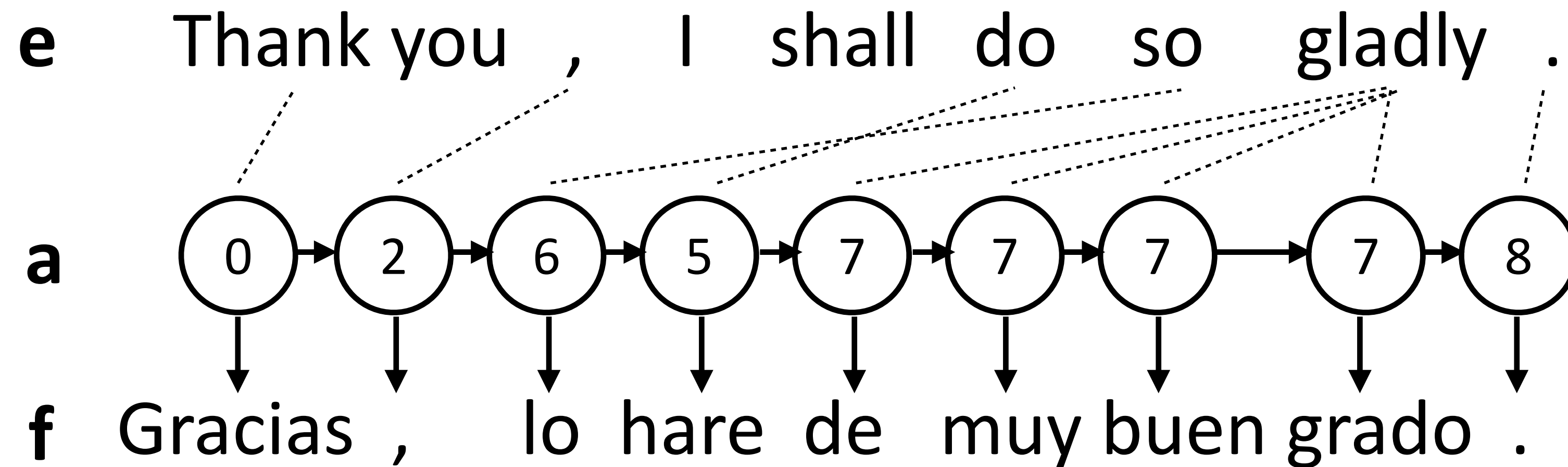
- ▶ E-step: estimate $P(\mathbf{a} \mid \mathbf{t}, \mathbf{s})$
- ▶ M-step: treat $P(\mathbf{a} \mid \mathbf{t}, \mathbf{s})$ as “pseudo-labels” for the data. Read off counts + normalize
- ▶ Common unsupervised learning method for latent variable models



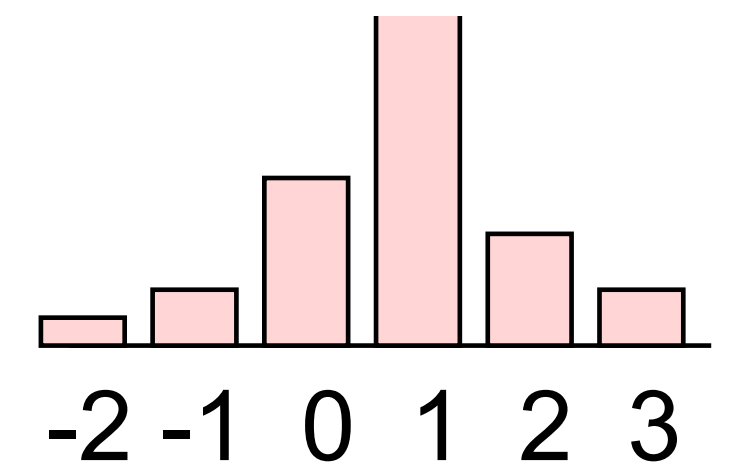
HMM for Alignment

- ▶ Sequential dependence between a's to capture monotonicity

$$P(\mathbf{t}, \mathbf{a} \mid \mathbf{s}) = \prod_{i=1}^n P(t_i \mid s_{a_i}) P(a_i \mid a_{i-1})$$



- ▶ Alignment dist parameterized by jump size: $P(a_j - a_{j-1})$



Vogel et al. (1996)



Phrase Extraction

- Find contiguous sets of aligned words in the two languages that don't have alignments to other words

d'assister à la reunion et ||| to attend the meeting and

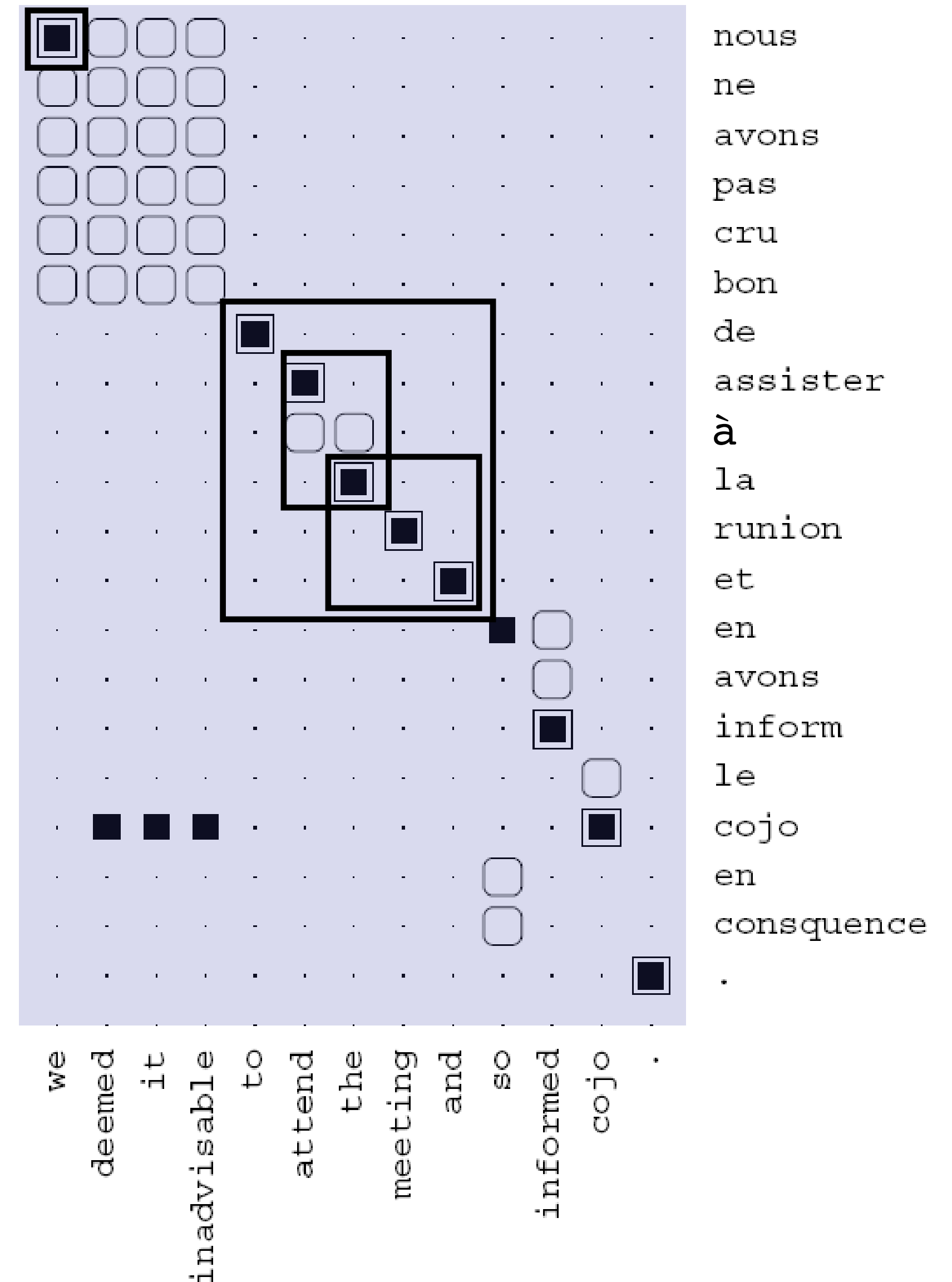
assister à la reunion ||| attend the meeting

la reunion and ||| the meeting and

nous ||| we

...

- Lots of phrases possible, count across all sentences and score by frequency



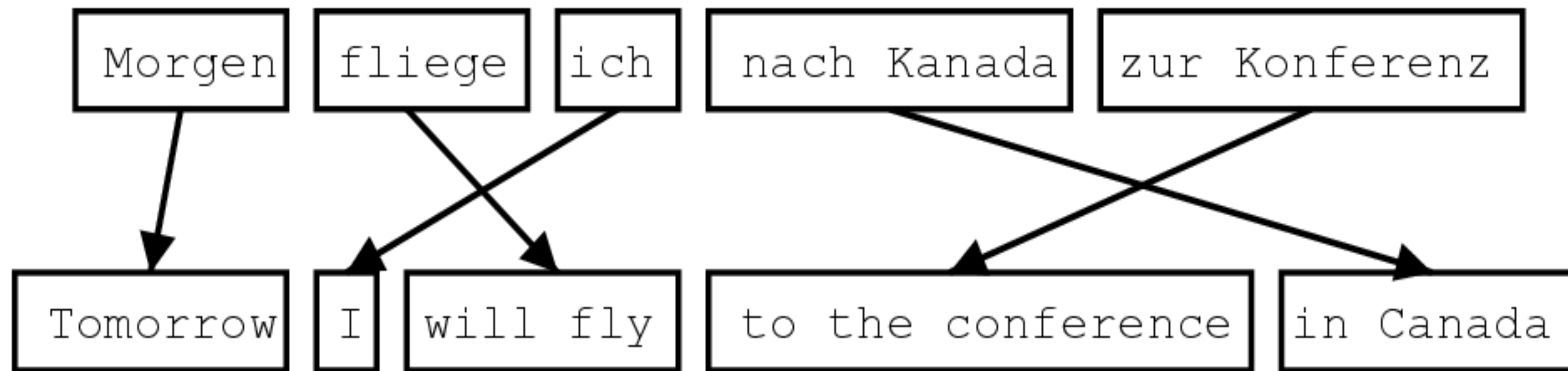


Phrase-Based Decoding

- ▶ Inputs:

- ▶ n-gram language model: $P(e_i | e_1, \dots, e_{i-1}) \approx P(e_i | e_{i-n-1}, \dots, e_{i-1})$
- ▶ Phrase table: set of phrase pairs (\mathbf{e}, \mathbf{f}) with probabilities $P(\mathbf{f} | \mathbf{e})$

- ▶ What we want to find: \mathbf{e} produced by a series of phrase-by-phrase translations from an input \mathbf{f} , possibly with reordering:



- ▶ Uses a beam search algorithm. We will not discuss

Cross-Lingual, Multilingual Word Representations



Multilingual Embeddings

- ▶ MT involves directly mapping between strings in different languages
- ▶ Potentially easier task: learn model that can do the same task in multiple languages? E.g., do POs tagging in both English and French, do a QA in 10 languages, etc.
- ▶ We'll see some neural techniques that can do this, then come back to translation



Multilingual Embeddings

- ▶ Input: corpora in many languages. Output: embeddings where similar words *in different languages* have similar embeddings

I have an apple
47 24 18 427

J' ai des oranges
47 24 89 1981

ID: 24
ai have

ID: 47
I Je J'

- ▶ multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train “monolingual” embeddings over all these corpora

- ▶ Works okay but not all that well

Ammar et al. (2016)



Multilingual BERT

- ▶ Take top 104 Wikipedias, train BERT on all of them simultaneously
- ▶ What does this look like?

Beethoven may have proposed unsuccessfully to Therese Malfatti, the supposed dedicatee of "Für Elise"; his status as a commoner may again have interfered with those plans.

当人们在马尔法蒂身后发现这部小曲的手稿时，便误认为上面写的是“Für Elise”（即《给爱丽丝》） [51]。

Кита́й (официально — Кита́йская Наро́дная Респу́блика, сокращённо — КНР; кит. трад. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэньминь Гунхэго) — государство в Восточной Аз

Devlin et al. (2019)



Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	ES	IT
EN	96.82	89.40	85.91	91.60
DE	83.99	93.99	86.32	88.39
ES	81.64	88.87	96.71	93.71
IT	86.79	87.82	91.28	98.11

Table 2: POS accuracy on a subset of UD languages.

- ▶ Can transfer BERT directly across languages with some success
- ▶ ...but this evaluation is on languages that all share an alphabet

Pires et al. (2019)



Multilingual BERT: Results

	HI	UR		EN	BG	JA
HI	97.1	85.9	EN	96.8	87.1	49.4
UR	91.1	93.8	BG	82.2	98.9	51.6
			JA	57.4	67.2	96.5

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- ▶ Urdu (Arabic/Nastaliq script) => Hindi (Devanagari). Transfers well despite different alphabets!
- ▶ Japanese => English: different script and very different syntax



Scaling Up: XLM-R

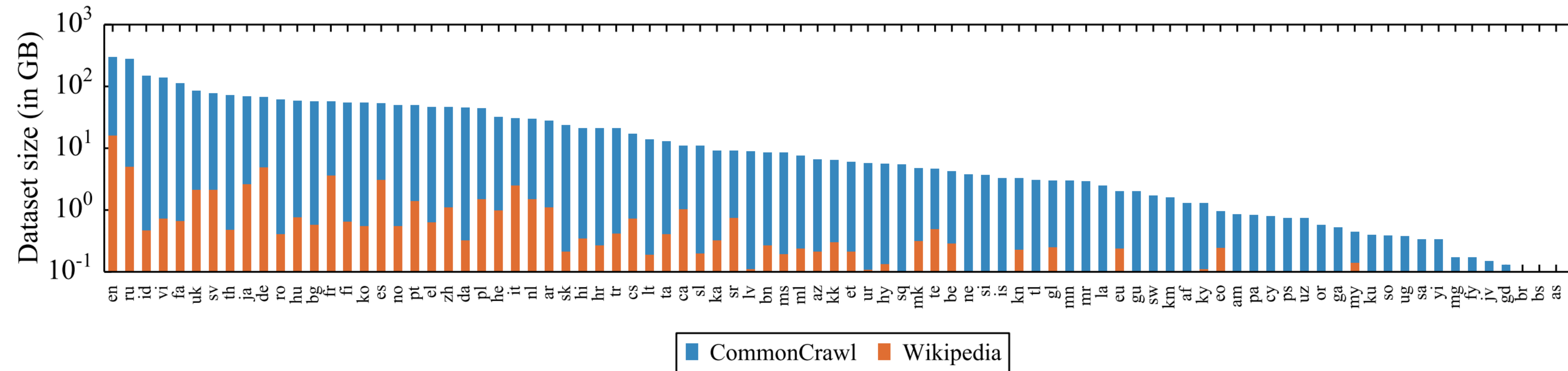


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- ▶ Larger “Common Crawl” dataset, better performance than mBERT
- ▶ Low-resource languages benefit from training on other languages
- ▶ High-resource languages see a small performance hit, but not much



Scaling Up: Benchmarks

Task	Corpus	Train	Dev	Test	Test sets	Lang.	Task
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction
	MLQA			4,517–11,590	translations	7	Span extraction
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval

- ▶ Many of these datasets are translations of base datasets, not originally annotated in those languages
- ▶ Exceptions: POS, NER, TyDiQA

Hu et al. (2021)



TyDiQA

- ▶ Typologically-diverse QA dataset
- ▶ Annotators write questions based on very short snippets of articles; answers may or may not exist, fetched from elsewhere in Wikipedia

Q: Как далеко Уран от Земли?
how far Uranus-SG.NOM from Earth-SG.GEN?

How far is Uranus from Earth?

A: Расстояние между Уран-ом и Земл-ёй меняется от 2,6 до 3,15 млрд км...
distance between Uranus-SG.INSTR and Earth-SG.INSTR varies from 2,6 to 3,15 bln km...

The distance between Uranus and Earth fluctuates from 2.6 to 3.15 bln km...

Clark et al. (2021)

Transformer MT + Frontiers



Transformers

Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	41.29
Transformer (base model)	27.3	38.1
Transformer (big)	28.4	41.8

- ▶ Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved



Frontiers in MT: Small Data

ID	system	BLEU	
		100k	3.2M
1	phrase-based SMT	15.87 \pm 0.19	26.60 \pm 0.00
2	NMT baseline	0.00 \pm 0.00	25.70 \pm 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 \pm 0.62	31.93 \pm 0.05
4	3 + reduce BPE vocabulary (14k \rightarrow 2k symbols)	12.10 \pm 0.16	-
5	4 + reduce batch size (4k \rightarrow 1k tokens)	12.40 \pm 0.08	31.97 \pm 0.26
6	5 + lexical model	13.03 \pm 0.49	31.80 \pm 0.22
7	5 + aggressive (word) dropout	15.87 \pm 0.09	33.60 \pm 0.14
8	7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate)	16.57 \pm 0.26	32.80 \pm 0.08
9	8 + lexical model	16.10 \pm 0.29	33.30 \pm 0.08

- Synthetic small data setting: German \rightarrow English Sennrich and Zhang (2019)



Frontiers in MT: Low-Resource

- ▶ Particular interest in deploying MT systems for languages with little or no parallel data

- ▶ BPE allows us to transfer models even without training on a specific language

- ▶ Pre-trained models can help further

Burmese, Indonesian, Turkish
BLEU

Transfer	My→En	Id→En	Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

Table 3: Investigating the model’s capability to restore its quality if we reset the parameters. We use En→De as the parent.

Aji et al. (2020)



Frontiers in MT: Low-Resource

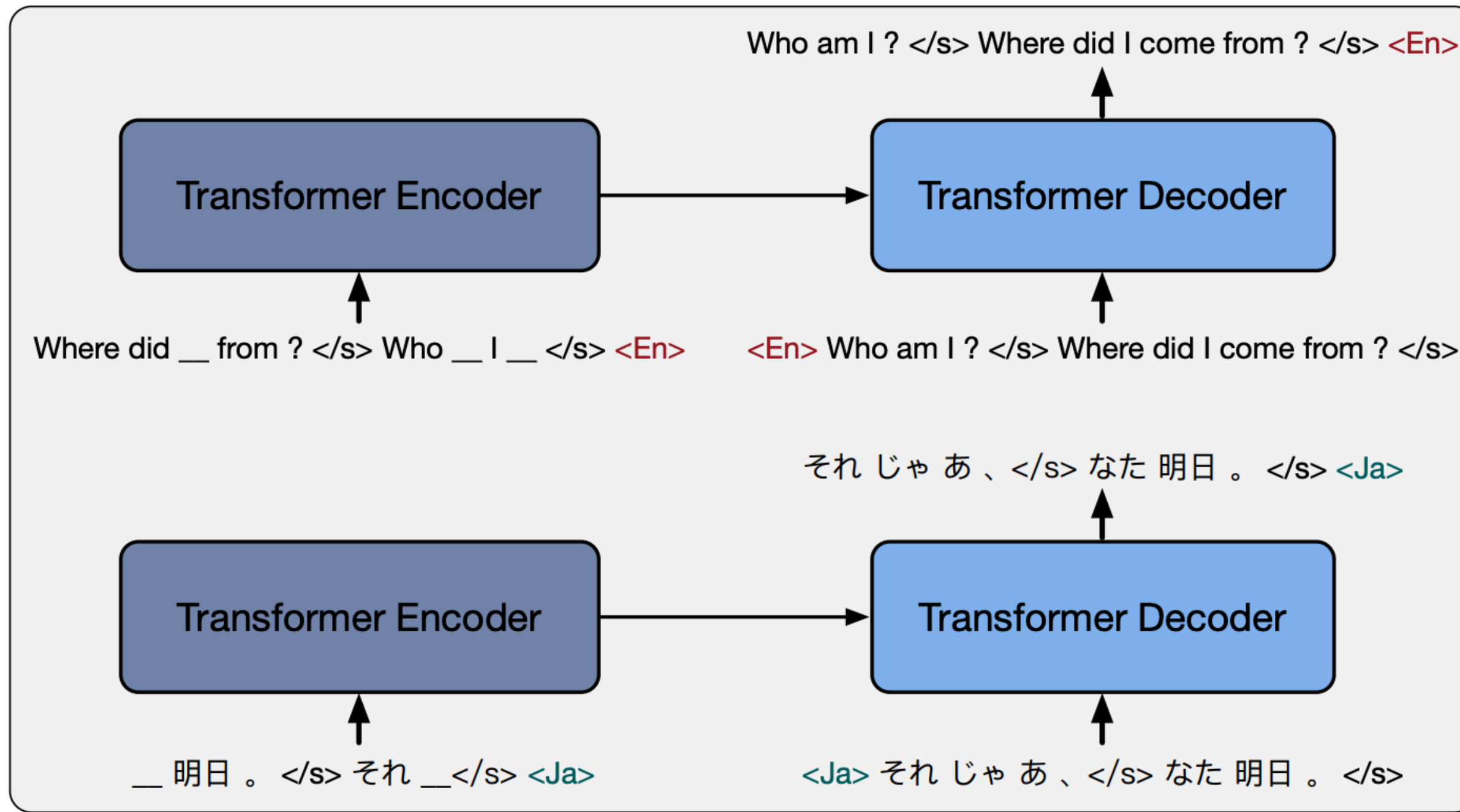
Transferring		BLEU						
Emb.	Inner	De→En parent			En→De parent			avg.
		My→En	Id→En	Tr→En	My→En	Id→En	Tr→En	
Y	Y	17.8	27.4	20.3	17.5	27.5	20.2	21.7
N	Y	13.6	25.3	19.4	10.8	24.9	19.3	18.3
Y	N	3.0	18.2	19.1	3.4	18.8	18.9	13.7
N	N	4.0	20.6	19.0	4.0	20.6	19.0	14.5

Table 2: Transfer learning performance by only transferring parts of the network. Inner layers are the non-embedding layers. N = not-transferred. Y = transferred.

- ▶ Very important to transfer the basic Transformer “skills”, but re-learning the embeddings seems fine in many cases



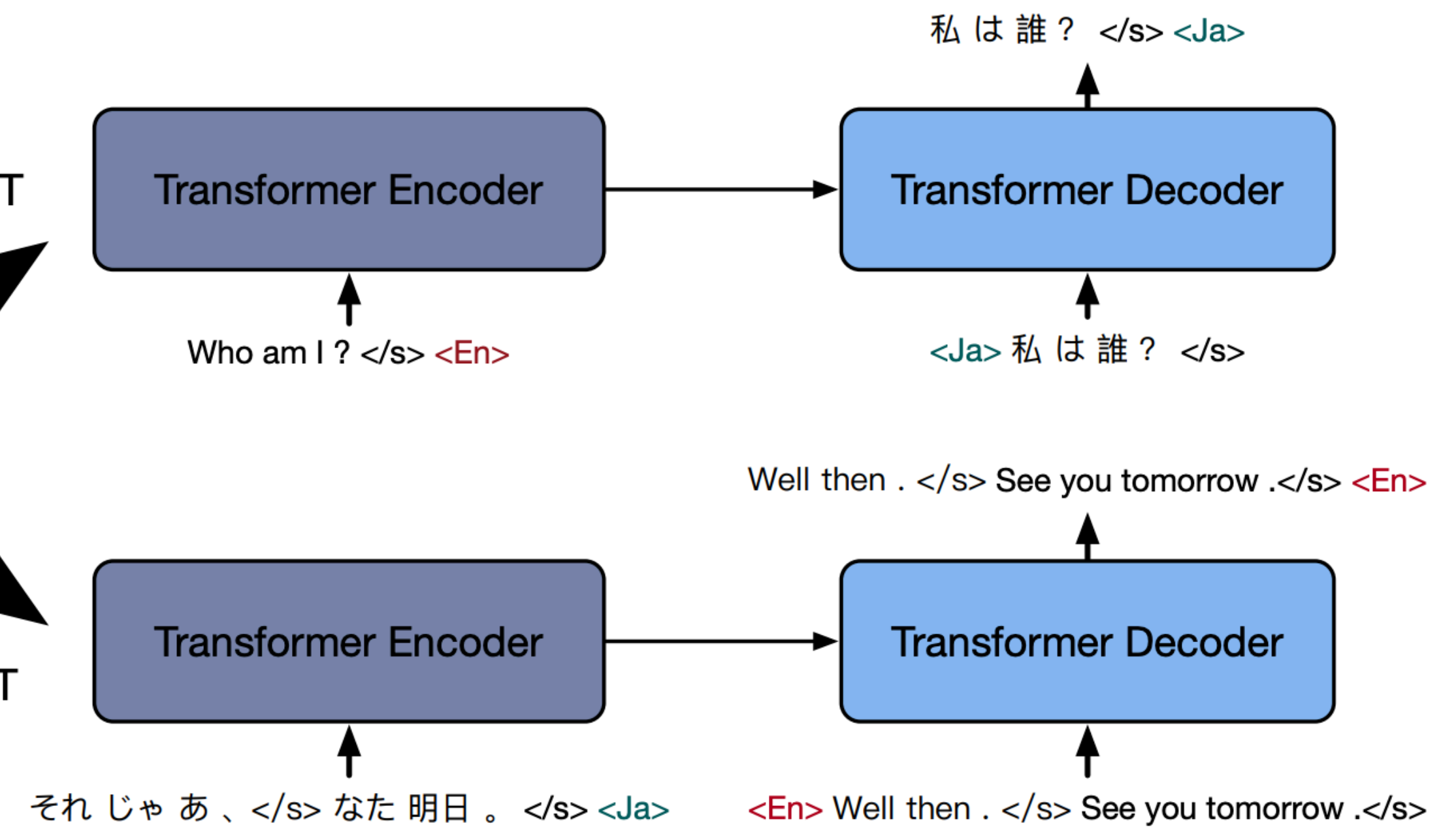
Frontiers in MT: Multilingual Models



Multilingual Denoising **Pre-Training** (mBART)

Sent-MT

Doc-MT



Fine-tuning on Machine Translation



Frontiers in MT: Multilingual Models

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	0.3	0.1	7.4	2.5	36.1	35.4	22.5	17.8	19.1	19.4	24.6	22.6

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	IWSLT17		IWSLT17		IWSLT17		WAT19		FLoRes		WMT16	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	43.3	34.8	37.6	21.6	39.8	34.0	28.3	36.9	14.5	7.4	37.8	37.7

- ▶ Random = random initialization



Frontiers in MT: Multilingual Models

SOURCE
Zh

针对政府的沉默态度,初级医生委员会执行委员会已于今日正式要求英国医学协会理事会召开特别会议批准旨在从九月初开始升级劳工行动的一项长期计划。

TARGET
En

In response to the government's silence, **JDC exec** has today made a formal request for a special meeting of **BMA Council** to authorise a rolling programme of **escalated industrial action** beginning in early September.

mBART25
Ja-En

In response to the government's silence, the **Council of Chief Medical Officers** has formally requested today the **Royal College of Physicians** to hold a special meeting to approve a long-term **workforce action** that starts in September.

mBART25
Ko-En

In response to the government's silence, the **Chief Medical Officers' Council** is calling today for a special session at the **Council of the British Medical Association**, which is a long-term initiative to **upgrade labor** from September.

mBART25
Zh-En

In response to the government's silence, the **Board of Primary Doctors** has today formally asked the **British Medical Association** to hold a special meeting to approve a long-term plan that starts in the beginning of September.



Frontiers in MT: ChatGPT

Table 3: Comparison of different prompts for ChatGPT to perform Chinese-to-English (Zh \Rightarrow En) translation.

System	BLEU \uparrow	ChrF++ \uparrow	TER \downarrow
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
ChatGPT w/ TP1	23.25	53.07	66.03
ChatGPT w/ TP2	24.54	53.05	63.79
ChatGPT w/ TP3	24.73	53.71	62.84

- ▶ Works okay for Chinese-English, but less good at generating into low-resource languages (English \rightarrow Romanian doesn't work well)

Table 5: Performance of ChatGPT with pivot prompting. New results are obtained from the updated ChatGPT version on 2023.01.31. LR: length ratio.

System	De \Rightarrow Zh		Ro \Rightarrow Zh	
	BLEU	LR	BLEU	LR
Google	38.71	0.94	39.05	0.95
DeepL	40.46	0.98	38.95	0.99
ChatGPT (Direct)	34.46	0.97	30.84	0.91
ChatGPT (Direct _{new})	30.76	0.92	27.51	0.93
ChatGPT (Pivot _{new})	34.68	0.95	34.19	0.98

- ▶ Better with “pivoting”

“Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine” Jia et al. (2023)



Frontiers: Evaluation with LLMs

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:
```

Figure 1: The best-performing prompt based on Direct Assessment expecting a score between 0–100. **Template portions in bold face** are used only when a human reference translation is available.

- ▶ Outperforms many learned MT metrics (Transformers trained over (source, target, reference) triples to reproduce human judgments of quality)

Kocmi et al. (2023)



Takeaways

- ▶ Word alignment is a way to learn unsupervised correspondences between words and build phrase tables
- ▶ Phrase-based MT was SOTA for a long time (and until the past couple of years was still best for low-resource settings)
- ▶ Transformers are state-of-the-art for machine translation
- ▶ They work really well on languages where we have a ton of data. When they don't: pre-training can help