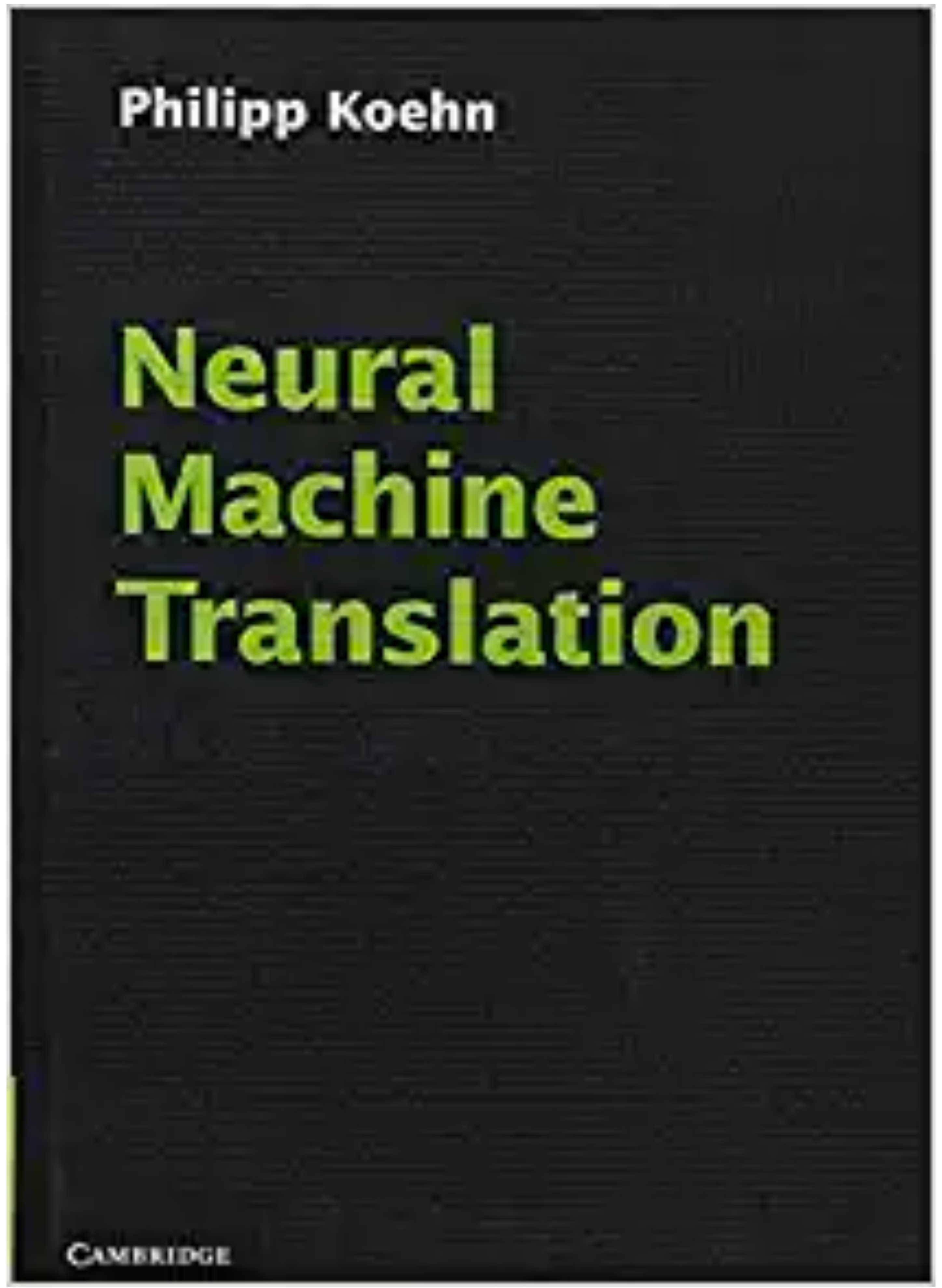
Encoder-Decoder (aka Seq2Seq)

Wei Xu

(many slides from Greg Durrett)

This Lecture

- Machine Translation
- Sequence-to-Sequence Model



- Reading — Eisenstein 18.3-18.5

MT Basics

MT Basics

People’s Daily, August 30, 2017

Trump Pope family watch a hundred years a year in the White House balcony

MT Basics

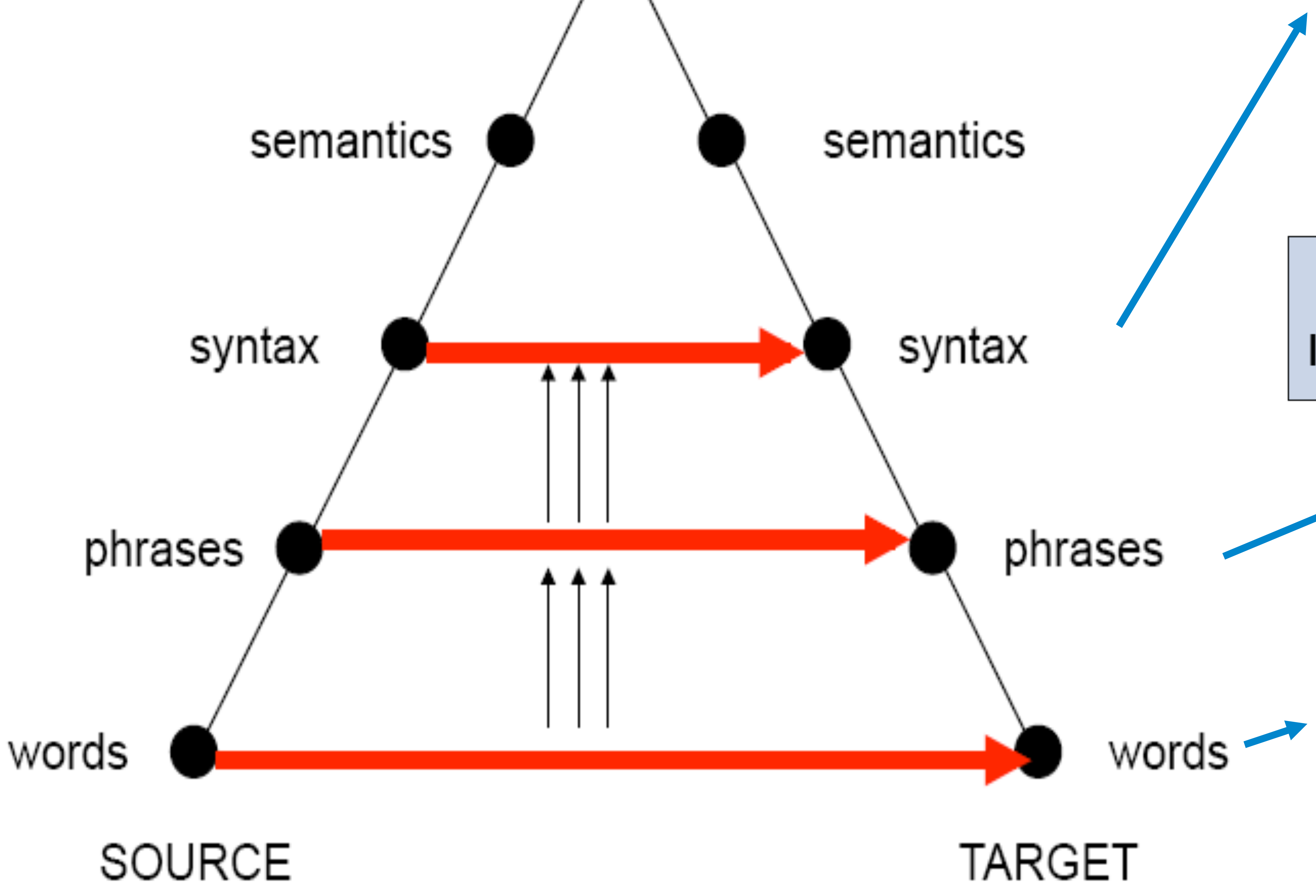
People’s Daily, August 30, 2017

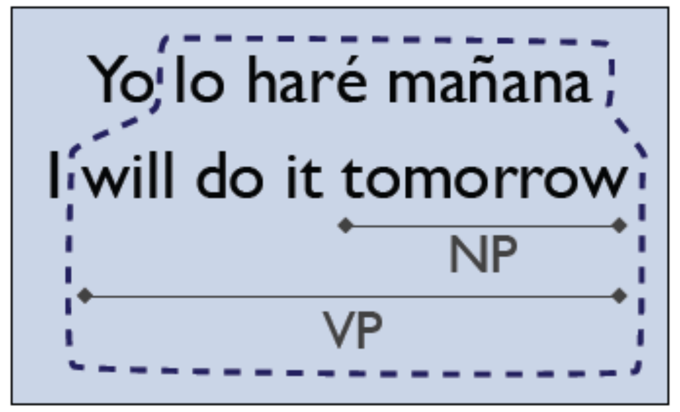
Trump and his family watch the once-in-a-century total solar eclipse from the White House balcony

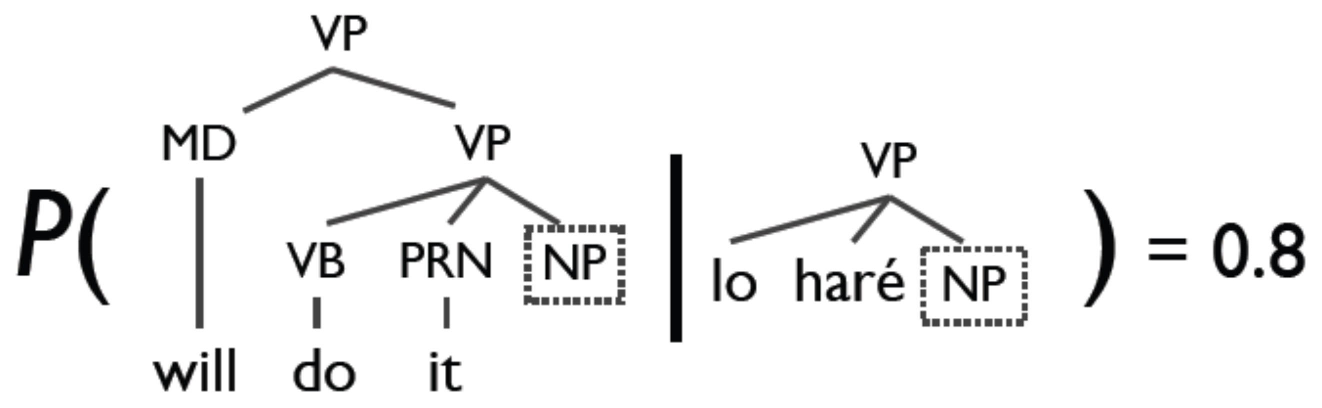
MT Ideally

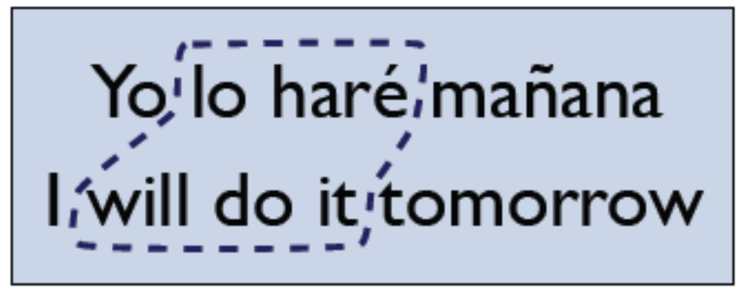
- I have a friend => $\exists x$ `friend(x,self)` => J'ai un ami
  J'ai une amie
  - May need information you didn't think about in your representation
  - Hard for semantic representations to cover everything
- Everyone has a friend => $\exists x \forall y$ `friend(x,y)` => Tout le monde a un ami
  $\forall x \exists y$ `friend(x,y)`
  - Can often get away without doing all disambiguation — same ambiguities may exist in both languages

Levels of Transfer: Vauquois Triangle (1968)

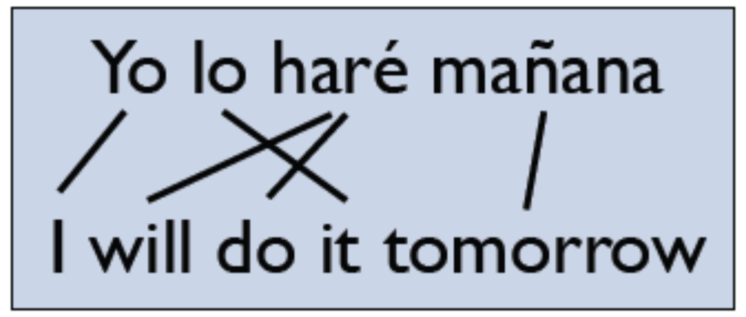








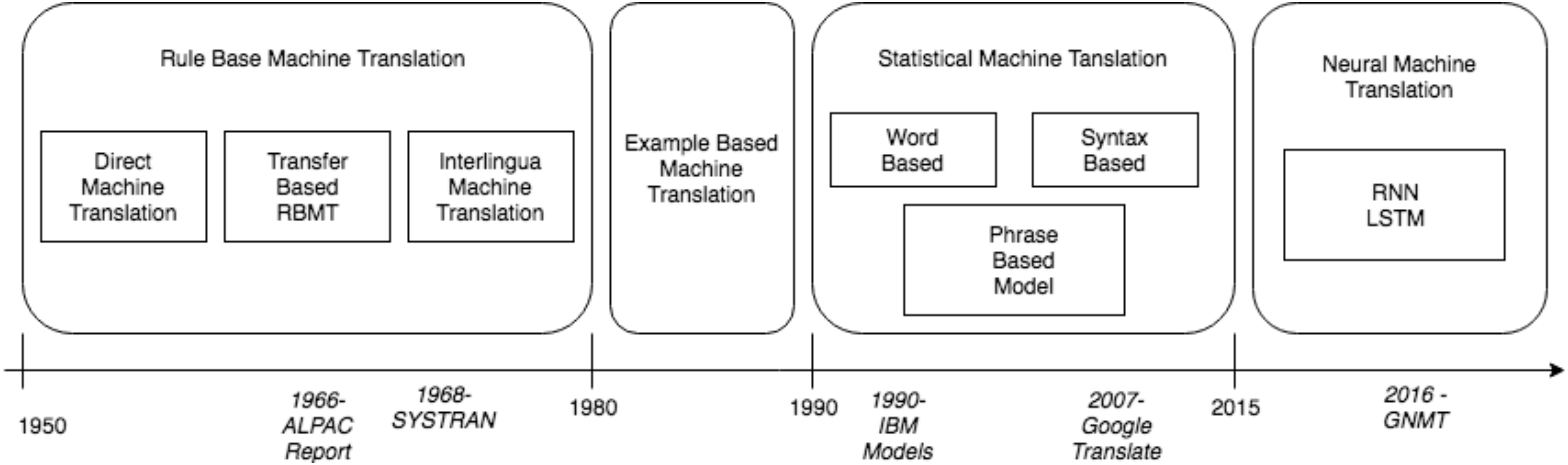
English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

Slide credit: Dan Klein

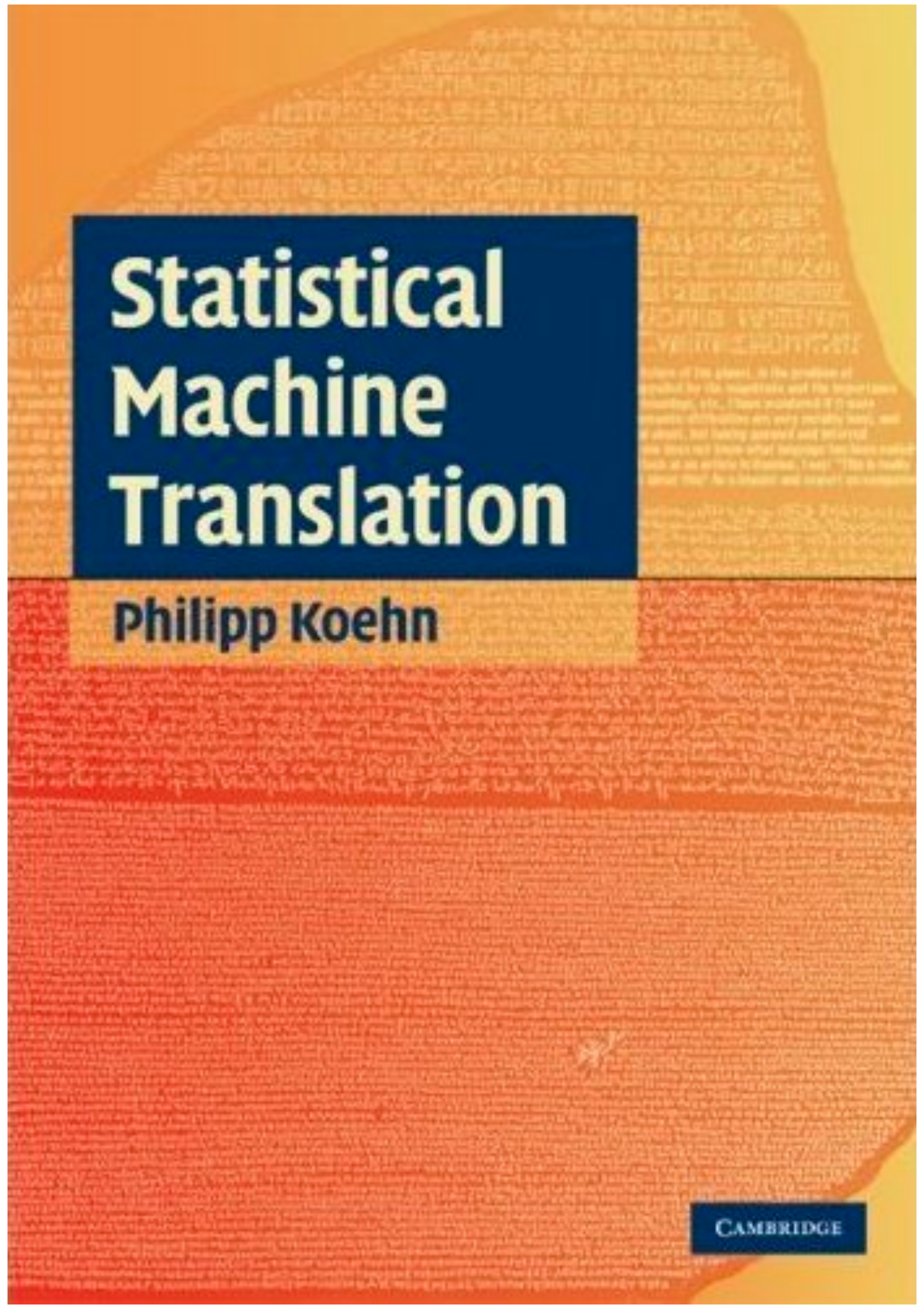
History of MT



Parallel Training Corpus

		facing with the swelling flow of through traffic zooming past their doors .		que pasa por delante_de sus casas , que aumenta a_diario .
5	#77501757	Weekend traffic bans and traffic jams are a curse to road transport .	#74765580	Las prohibiciones de conducir los fines de semana y los embotellamientos asolan el transporte por carretera .
6	#79500725	Some people also want to recoup the cost of traffic jams from those who get stuck in them , according to the ' polluter pays ' principle .	#76764676	Algunos son partidarios de que incluso los costes ocasionados por los atascos se carguen a el ciudadano que se encuentra atrapado en ellos , de conformidad con el principio de que " quien contamina paga " .
7	#79500765	I think this is an excellent principle and I would like to see it applied in full , but not to traffic jams .	#76764713	Me parece un principio acertado y estoy dispuesta a aplicarlo íntegramente , pero no sobre los atascos , ya_que éstos son un claro indicio de el fracaso de la política gubernamental en_materia_de infraestructuras .
8	#79500768	Traffic jams are indicative of failed government policy on the infrastructure front , which is why the government itself , certainly in the Netherlands , must be regarded as the polluter .	#76764747	Por eso es preciso subrayar que en estos casos quien contamina es el propio Gobierno , a el menos en los Países_Bajos .
9	#81309716	This would increase traffic jams , weaken road safety and increase costs .	#78586130	Esto aumentaría los atascos , mermaría la seguridad vial e incrementaría los costes .
10	#81997391	In the previous legislature , Parliament gave its opinion on the Commission ' s proposals on the simplification of vertical directives on sugar , honey , fruit juices , milk and jams .	#79281114	En efecto , durante la precedente legislatura , el Parlamento se manifestó sobre las propuestas de la Comisión relativas a la simplificación de directivas verticales sobre el azúcar , la miel , los zumos de frutas , la leche y las confituras .
11	#81998167	For jams , I personally reintroduced an amendment that was not accepted by the Committee on the Environment , Public Health and Consumer Policy , but which I hold to .	#79281936	Para las confituras , yo personalmente volví a introducir una enmienda que no fue aceptada por la Comisión_de_Medio_Ambiente , Salud_Pública y Política_de_el_Consumidor , pero que es importante para mí .
12	#81998209	It concerns not accepting the general use of a chemical flavouring in jams and marmalades , that is vanillin .	#79281966	Se trata de no aceptar la utilización generalizada de un aroma químico en las confituras y " marmalades " , a saber , la vainillina .
13	#82800065	This is highlighted particularly in towns where it is necessary to find ways of solving environmental problems and the difficulties caused by traffic jams .	#80085988	Esto se pone_de_relieve aún más en las ciudades , en las que hay que encontrar medios para eliminar los inconvenientes derivados de los problemas medioambientales y de la congestión de el tráfico .

Phrase-based MT (very briefly)



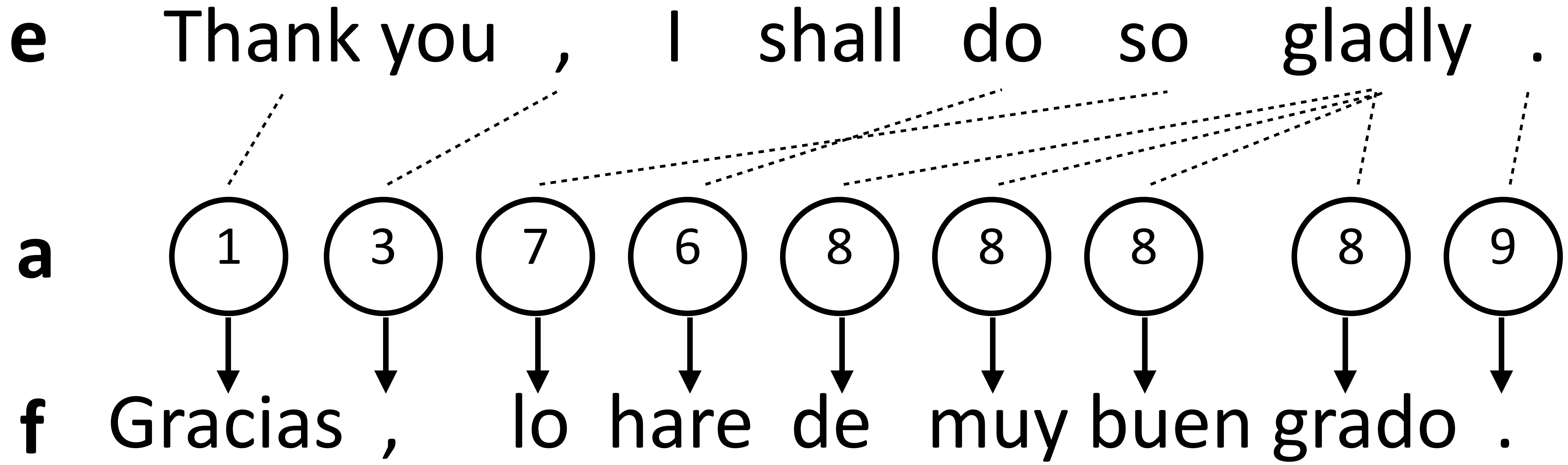
Phrase-Based MT

- Key idea: translation words 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)

Word Alignment: IBM Model 1

- Each "Foreign" word is aligned to *at most* one English word

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=1}^{n} P(f_i|e_{a_i})P(a_i)$$



- Set P(a) uniformly (no prior over good alignments) = 1 / (#words in e + 1)

- $P(f_i|e_{a_i})$: word translation probability. Learn with EM (Eisenstein ch 18.2.2)

Brown et al. (1993)

Word Alignment

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

de assister à la runion et ||| to attend the meeting and

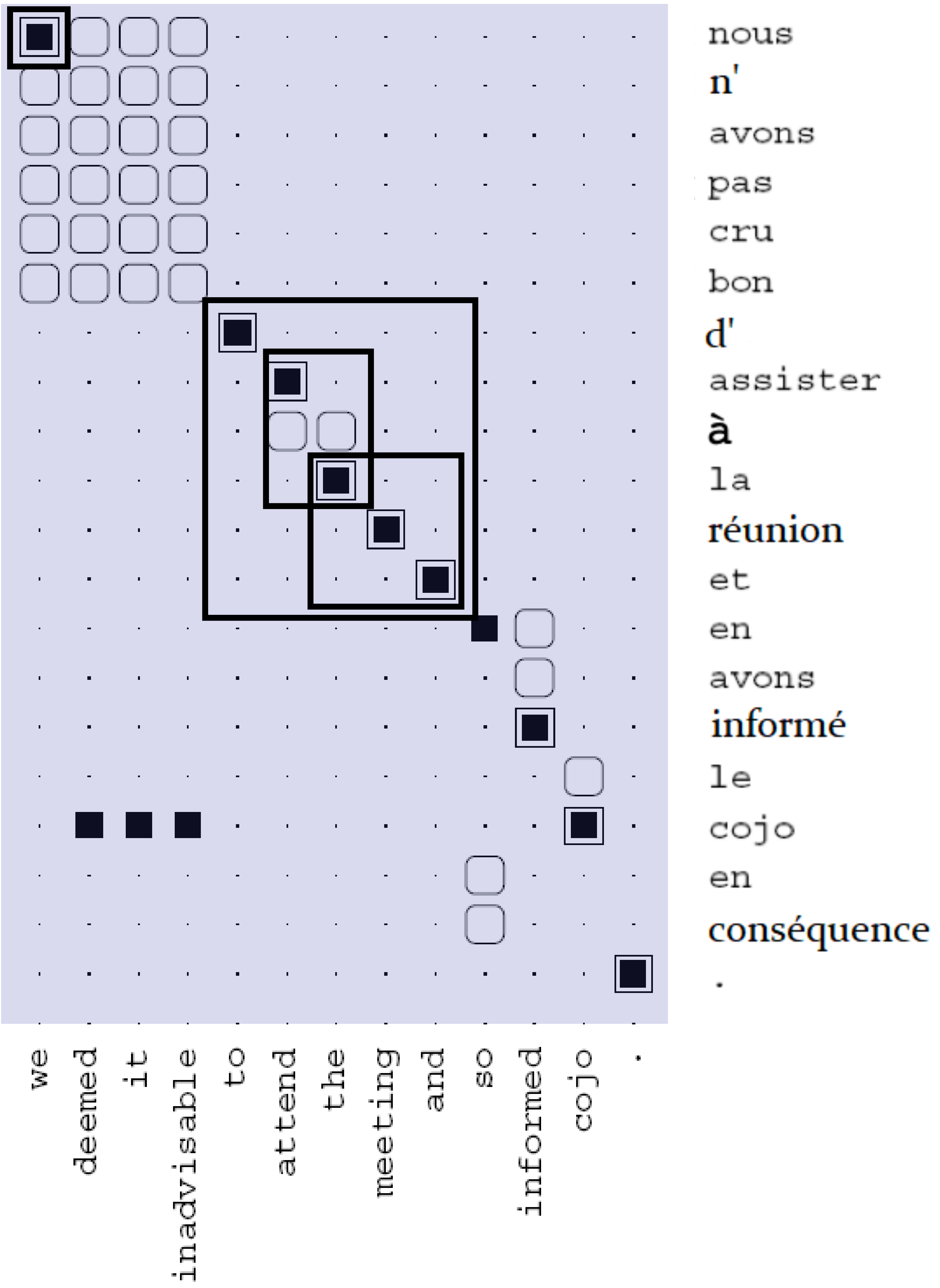
assister à la runion ||| attend the meeting

la runion and ||| the meeting and

nous ||| we

…

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

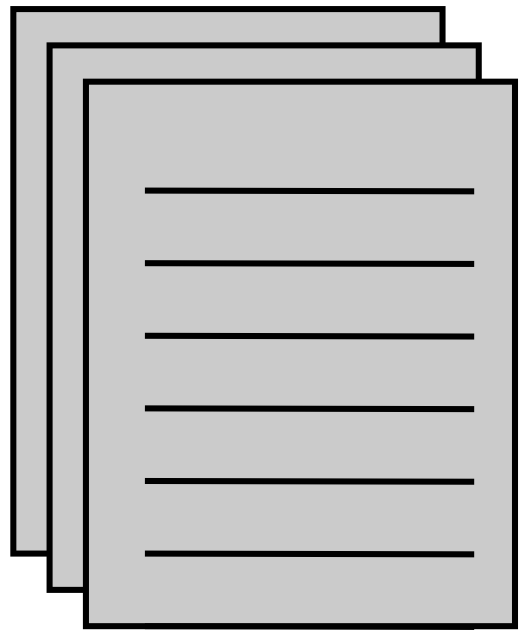


Phrase-Based MT

- Goal: translate from Foreign language to English

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 $P(e)$

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

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

“Translate faithfully but make fluent English”

MT Evaluation

Mean (Math Review)

- Arithmetic Mean = (P + R) / 2
- Geometric Mean = $\sqrt{P \times R}$
- Harmonic Mean = 2 x P x R / (P + R)

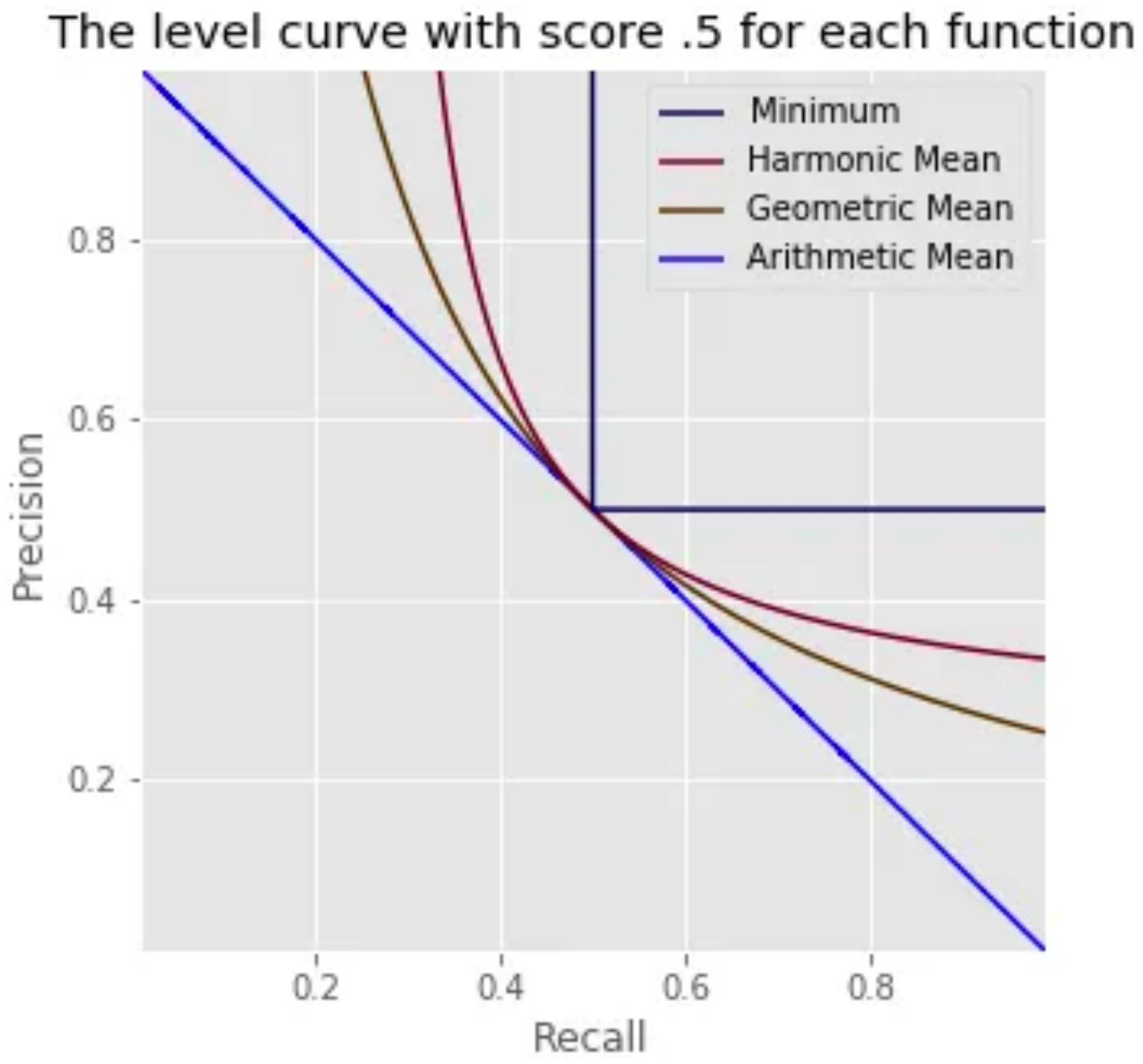


Image credit: Greg Gandenberger

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1
reference 1	I am tired			
reference 2	I am ready to sleep now and so exhausted			

Papineni et al. (2002)

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

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

- Typically $N = 4$, $w_i = 1/4$

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

- r = length of reference
  c = length of system output

- Does this capture fluency and adequacy?

Papineni et al. (2002)

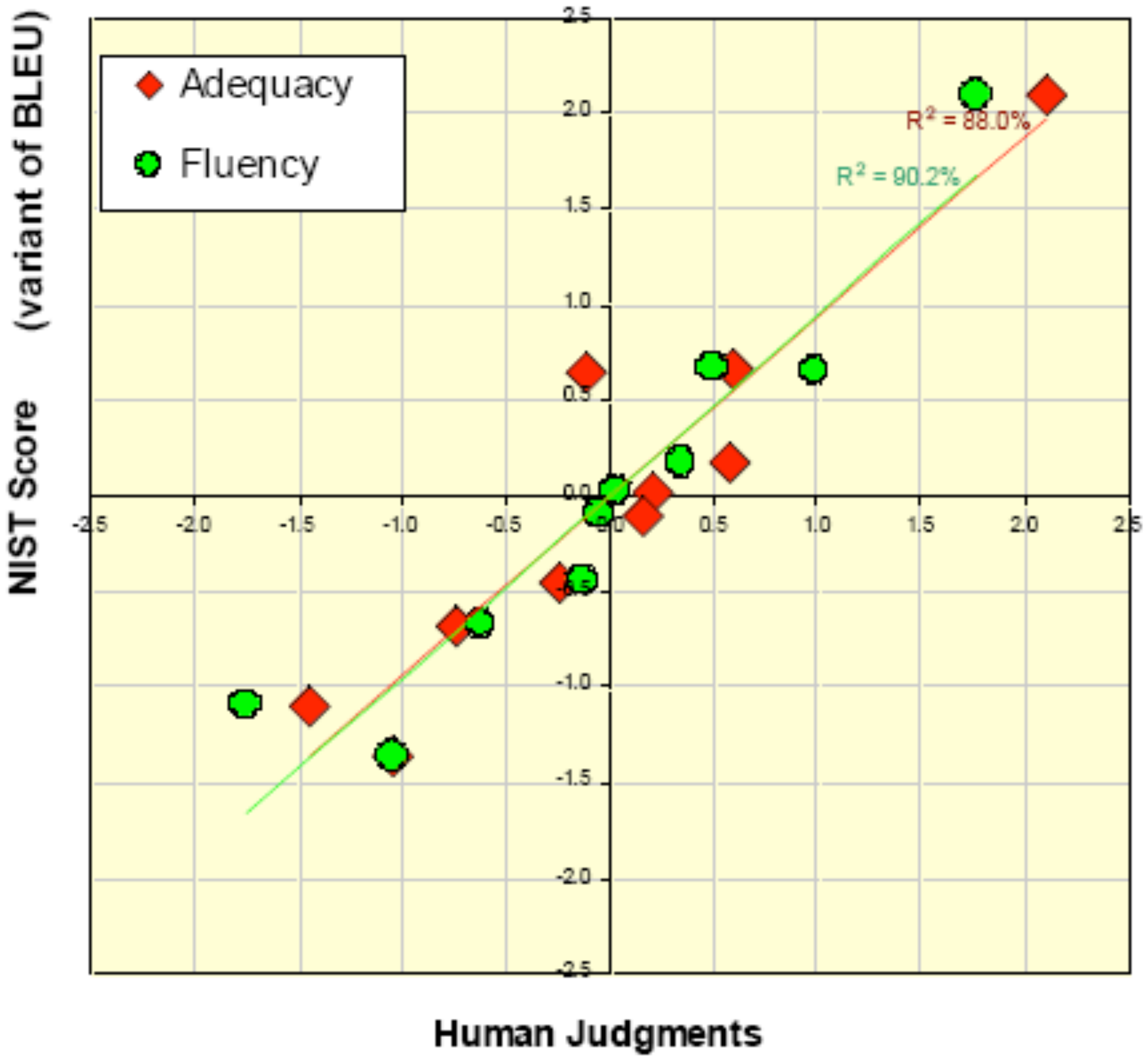
https://github.com/mjpost/sacrebleu

BLEU Score

$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} {p_n}^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

BLEU Score

- Better methods with human-in-the-loop
- HTER: human-assisted translation error rate
- If you're building real MT systems, you do user studies. In academia, you mostly use BLEU, COMET, etc.



slide from G. Doddington (NIST)

Appraise - Human Evaluation Interface

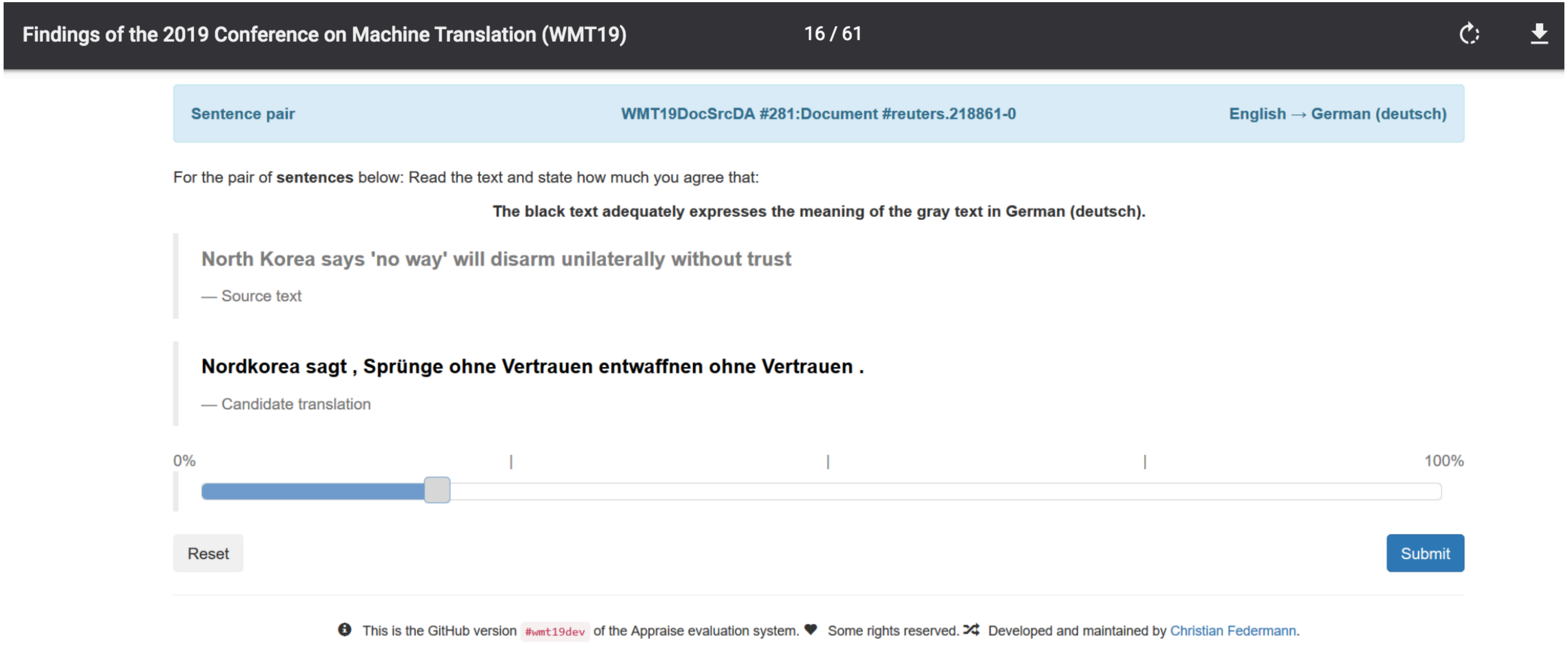
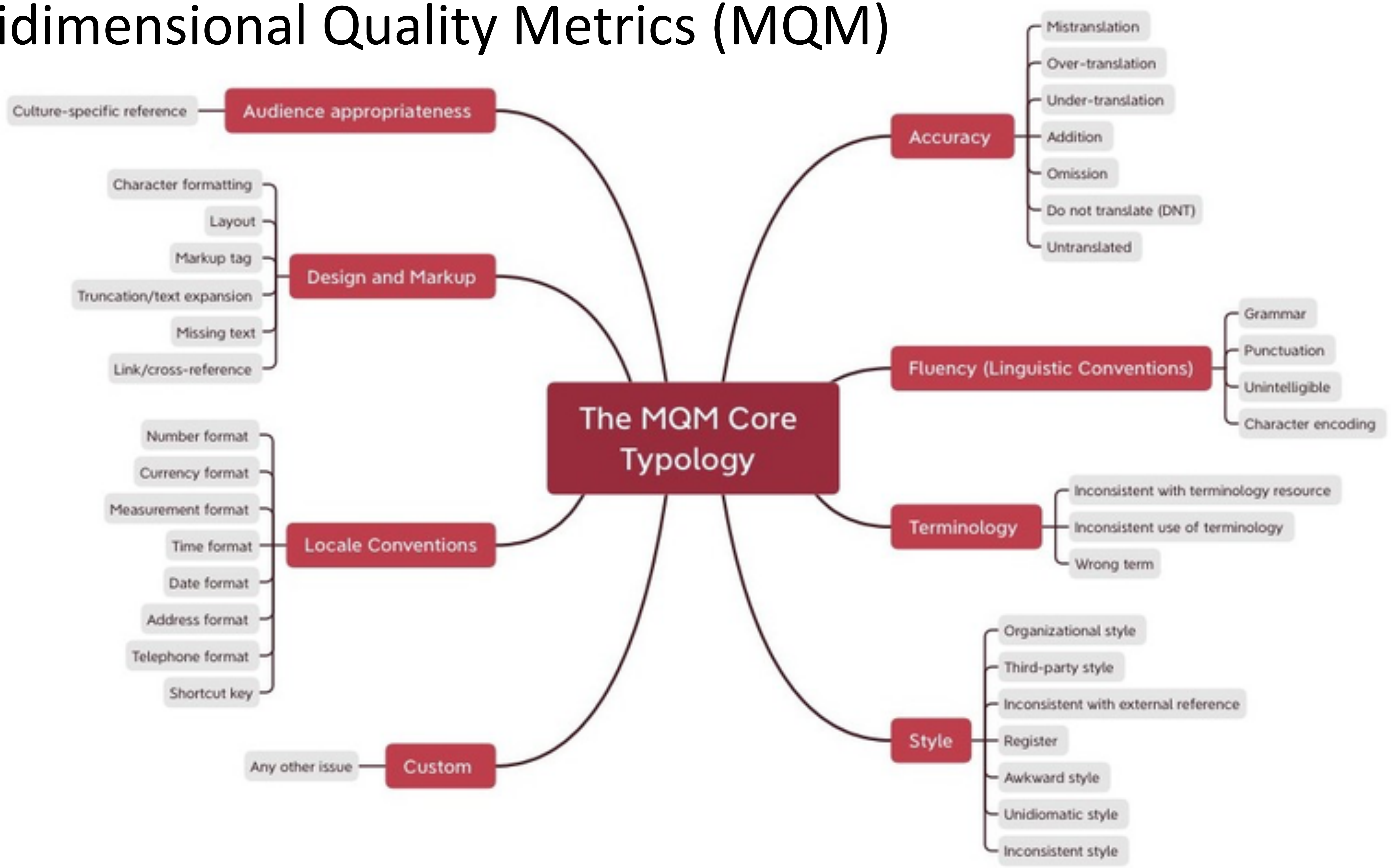


Figure 3: Screen shot of segment-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output segment randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

Federmann (2010)

MQM - fine-grained Human Eval

- Multidimensional Quality Metrics (MQM)



Lommel et al. (2014)

Thresh 🌾 - fine-grained Human Eval

https://github.com/davidheineman/thresh

thresh.tools

A Unified, Customizable and Deployable Platform
for Fine-Grained Text Evaluation

Prompt (human-written):

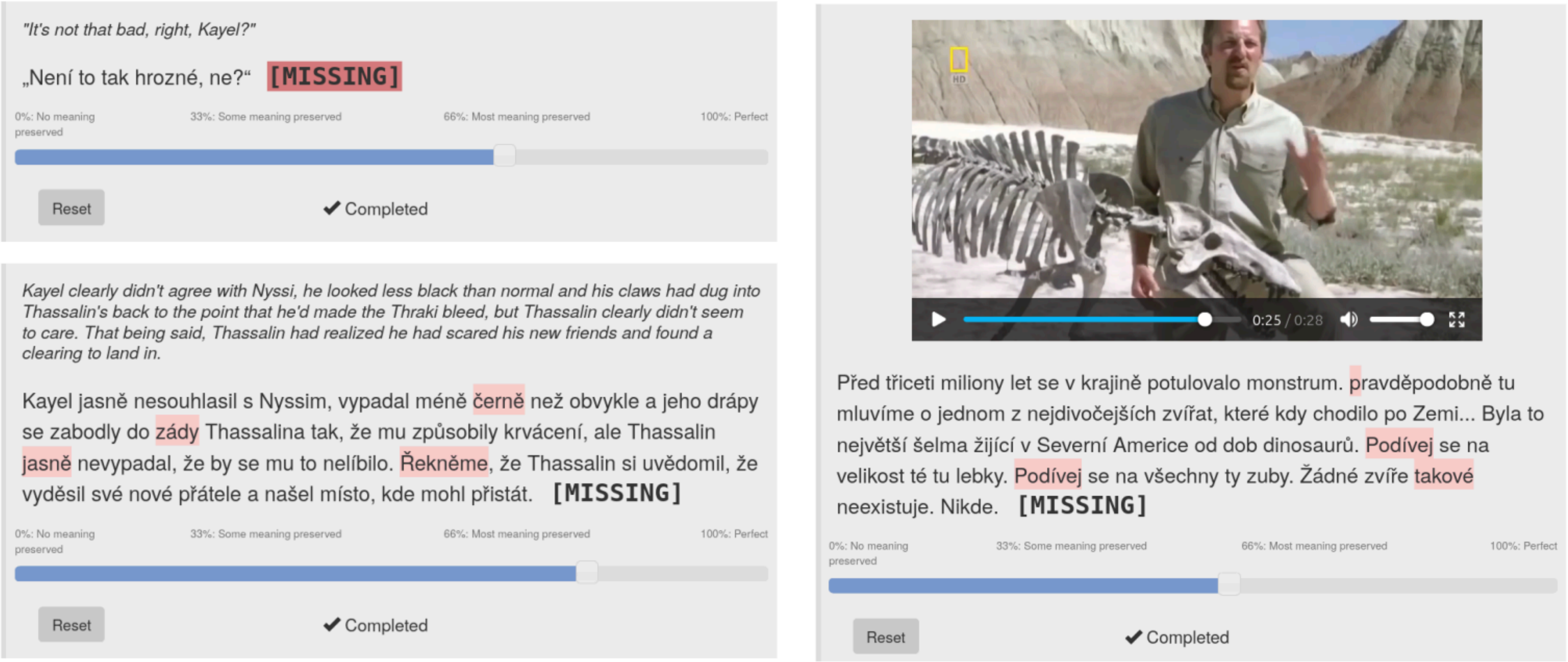
France's former President Nicolas Sarkozy was found guilty of corruption on Monday and sentenced to three years in prison, a stunning fall from grace for a man who led his country and bestrode the world stage for five years.

Continuation:

Given the string of accusations against Mr Sarkozy and his hardline France, it could be argued that the 6/7 AIPAC scandal and the background of the president's class political career were alarming indeed. Some of Mr Sarkozy's closest aides had openly suggested that his remarks — in which he dismissed other American Democrats as "crazies" who used unsavoury language — sounded like he had been flirting with suicide. Nevertheless, - or perhaps because he did not believe in lost his job — he himself lost his job rather than resign from office in 2011 and became the president without having to resign. In 2008, however, he became a totally shut-door candidate for the presidential election.

David Heineman, Yao Dou, Wei Xu. "Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation" (EMNLP 2023 demo)
David Heineman, Yao Dou, Mounica Maddela, Wei Xu. "Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA" (EMNLP 2023)

WMT 2024



(a) Excerpt of two segments from a larger document. In the first segment, the name *"Kayel"* is omitted which is a major error. In the second segment, there are many minor errors.

(b) Example of a video to text translation with several minor errors. The annotator can control the video player.

Figure 1: Two screenshots of ESA (Kocmi et al., 2024b) and the annotator instructions. ESA shows multiple segments within a document at once as well as video sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100. The tool is implemented in Appraise (Federmann, 2018).

Kocmi et al. (2024)

WMT 2024

Czech→Ukrainian

Rank	System	Human	AutoRank
1-2	Claude-3.5 §	93.0	1.7
2-2	HUMAN-A	92.7	-
3-3	Gemini-1.5-Pro	92.6	2.0
3-4	Unbabel-Tower70B	92.2	1.0
5-5	IOL-Research	90.2	1.9
6-7	CommandR-plus §	89.7	1.9
6-8	ONLINE-W	88.7	2.3
7-9	GPT-4 §	88.6	2.0
8-9	IKUN	87.1	2.3
10-10	Aya23	86.6	2.5
11-11	CUNI-Transformer	85.3	3.0
12-12	IKUN-C	82.6	3.0

English→Spanish

Rank	System	Human	AutoRank
1-1	HUMAN-A	95.3	-
2-2	Dubformer	93.4	2.0
3-4	GPT-4	91.9	1.9
4-7	IOL-Research	91.4	2.3
5-8	Mistral-Large	89.3	2.2
5-9	Unbabel-Tower70B	88.9	1.0
3-8	Claude-3.5	88.8	2.1
5-8	Gemini-1.5-Pro	88.8	2.4
7-9	CommandR-plus	88.3	2.1
9-10	Llama3-70B §	87.2	2.6
11-11	ONLINE-B	85.6	2.7
12-13	IKUN	84.7	2.8
12-13	IKUN-C	80.4	3.4
14-14	MSLC	63.9	7.4

English→Czech

Rank	System	Human	AutoRank
1-2	HUMAN-A	92.9	-
2-2	Unbabel-Tower70B	91.6	1.0
2-3	Claude-3.5 §	91.2	2.1
4-5	ONLINE-W	89.0	2.8
4-6	CUNI-MH	88.4	2.1
6-6	Gemini-1.5-Pro	88.2	2.6
6-8	GPT-4 §	87.7	2.6
8-8	CommandR-plus §	86.9	2.9
8-9	IOL-Research	86.5	2.8
10-11	SCIR-MT	85.4	3.2
10-11	CUNI-DocTransformer	84.3	4.4
12-12	Aya23	84.2	4.3
13-13	CUNI-GA	82.1	2.3
14-14	IKUN	81.7	3.9
15-15	Llama3-70B §	77.4	4.1
16-16	IKUN-C	75.4	4.7

English→Hindi

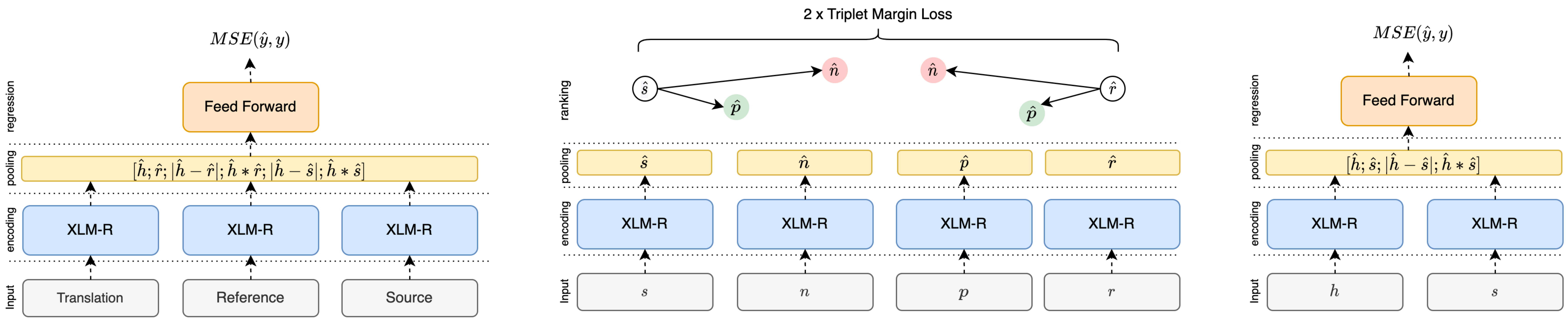
Rank	System	Human	AutoRank
1-3	TranssionMT	91.3	1.3
1-4	Unbabel-Tower70B	90.5	1.0
3-3	Claude-3.5 §	90.2	1.2
3-4	ONLINE-B	90.1	1.4
3-5	Gemini-1.5-Pro §	90.0	1.6
6-6	GPT-4 §	88.5	2.1
7-8	HUMAN-A	88.5	-
8-8	IOL-Research	87.2	2.1
8-9	Llama3-70B §	86.7	2.1
10-10	Aya23	84.7	3.2
11-11	IKUN-C	70.7	5.5

Kocmi et al. (2024)

Other MT Evaluation Metrics

- BLEU (2002): n-gram overlap
- METEOR (2005): also take into consideration of synonyms
- HTER (2009): human-assisted translation error rate
- BERTScore (2019): embedding-based
- BLEURT (2020) and COMET (2020): trained neural network model using human evaluation data
- and many more … e.g., CometKiwi-DA-XL (2023), MetricX-23-XL (2023)

COMET - Learnt Metric



- Regression Metric (left): trained on a regression task using source, MT and reference; Ranking Metric (middle): optimize to encode good translations closer to the anchors (source, reference) while pushing bad translations away; Reference-less Metric (right): does not use the reference translation.

Rei et al. (2020)

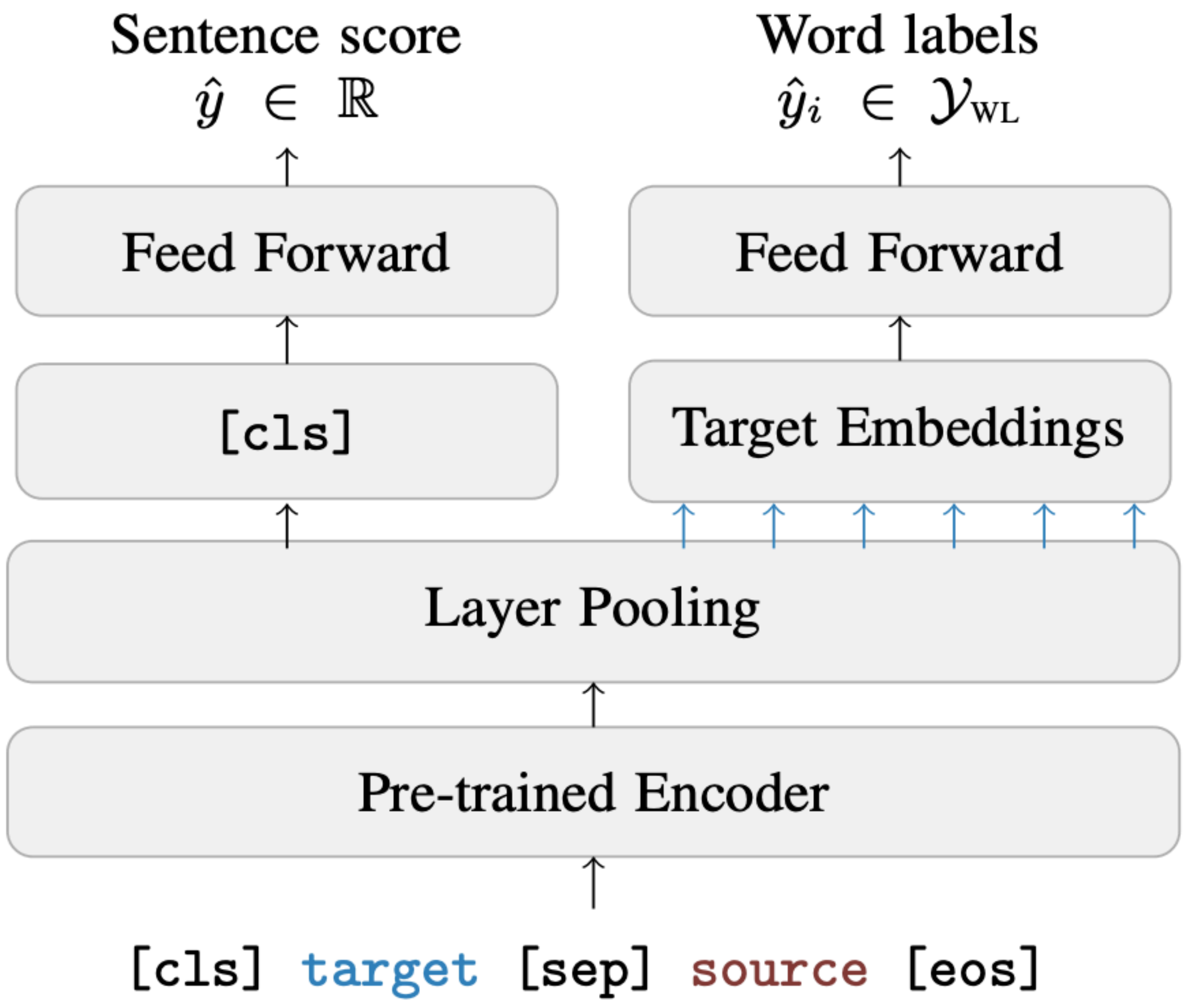
COMETKIWI - Learnt Metric

- Learn from both sentence-level and word-level quality estimations
- Use a (trainable) weighted sum of the hidden states of each layer of the encoder

$$\mathcal{L}_{\mathrm{SL}}(\theta) = \frac{1}{2}(y - \hat{y}(\theta))^2$$

$$\mathcal{L}_{\mathrm{WL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{y_i} \log p_\theta(y_i)$$

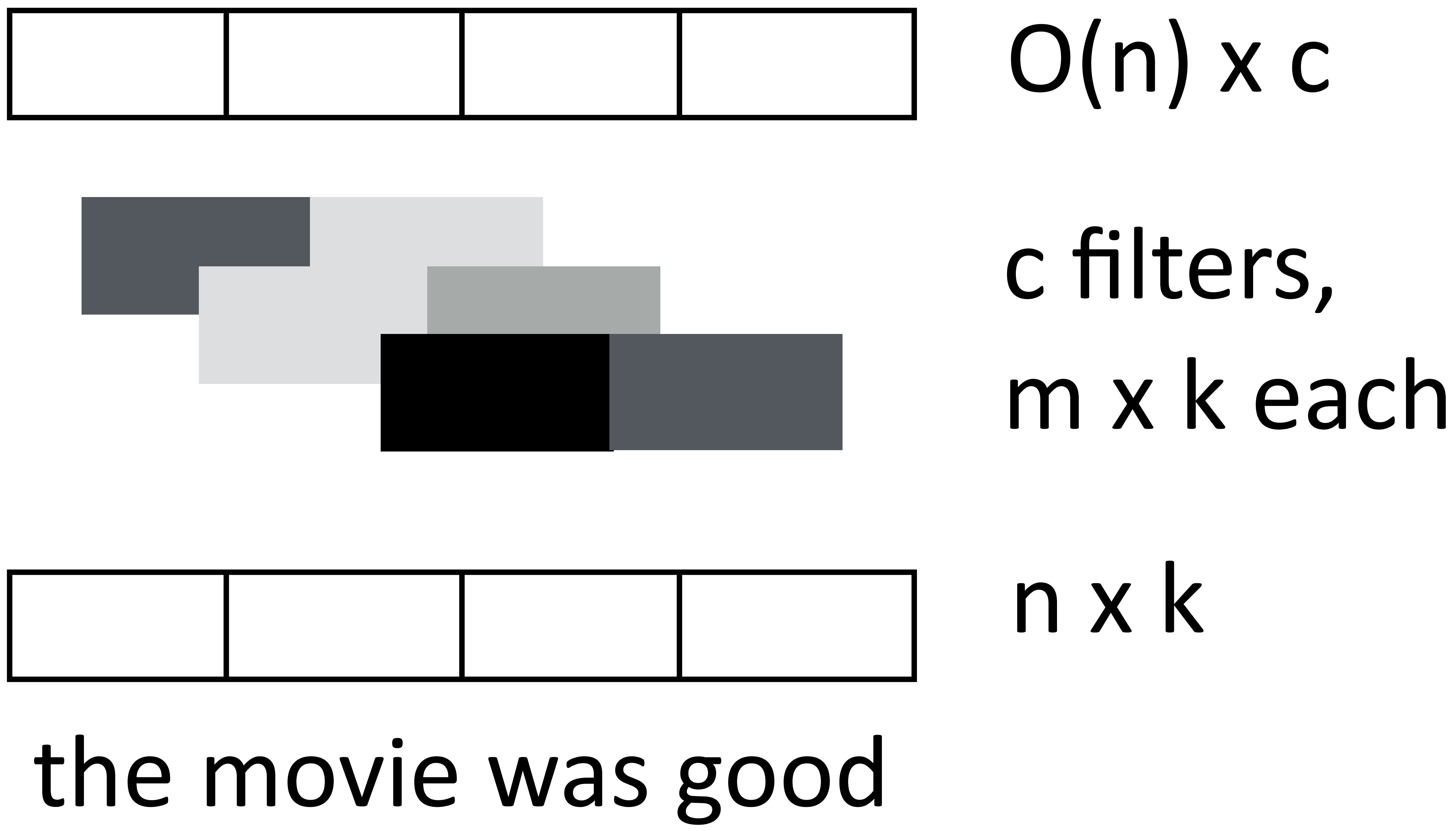
$$\mathcal{L}(\theta) = \lambda_{\mathrm{SL}}\mathcal{L}_{\mathrm{SL}}(\theta) + \lambda_{\mathrm{WL}}\mathcal{L}_{\mathrm{WL}}(\theta),$$

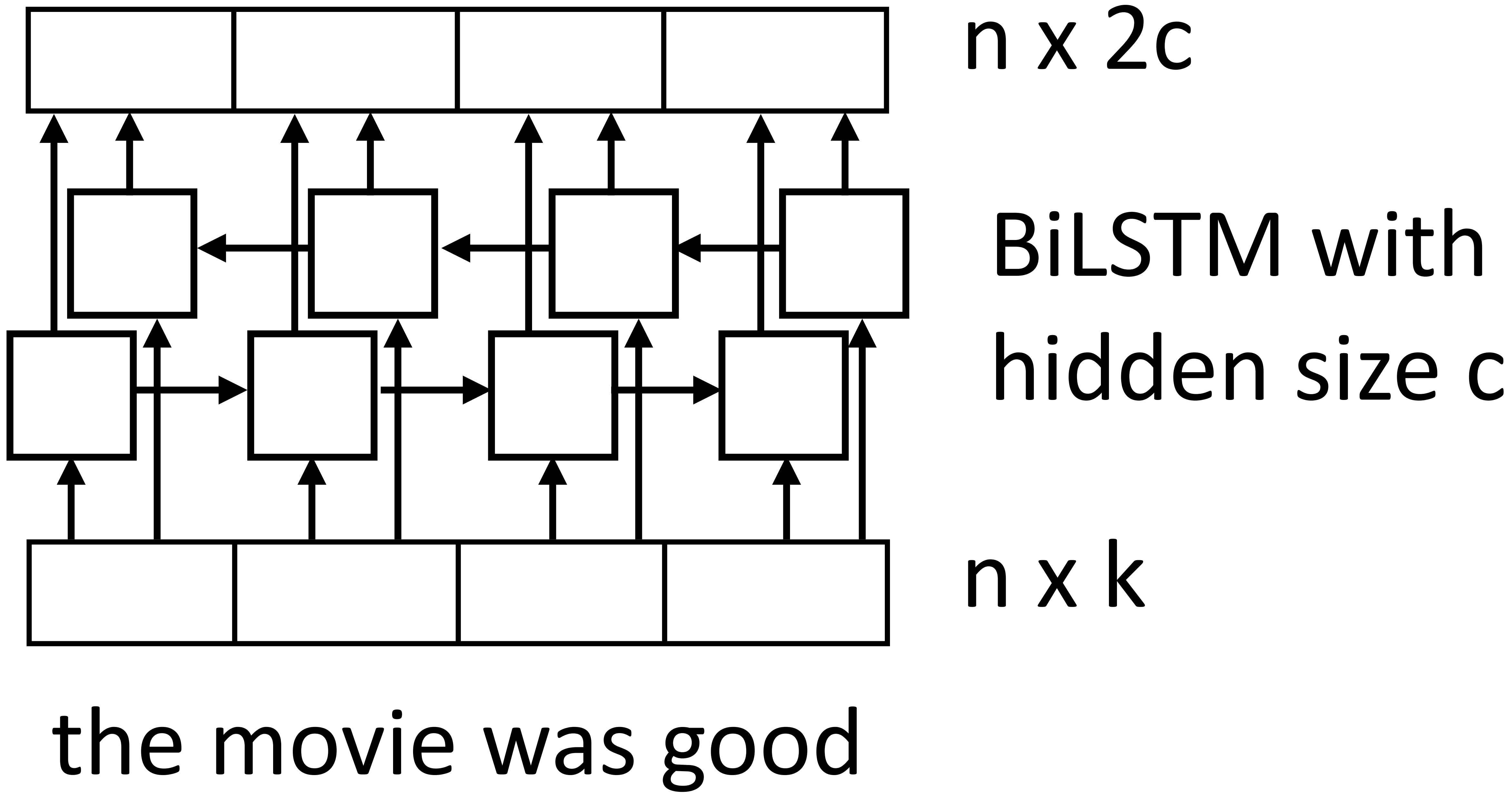


Rei et al. (2022)

Seq2Seq Models

Recall: CNNs vs. LSTMs

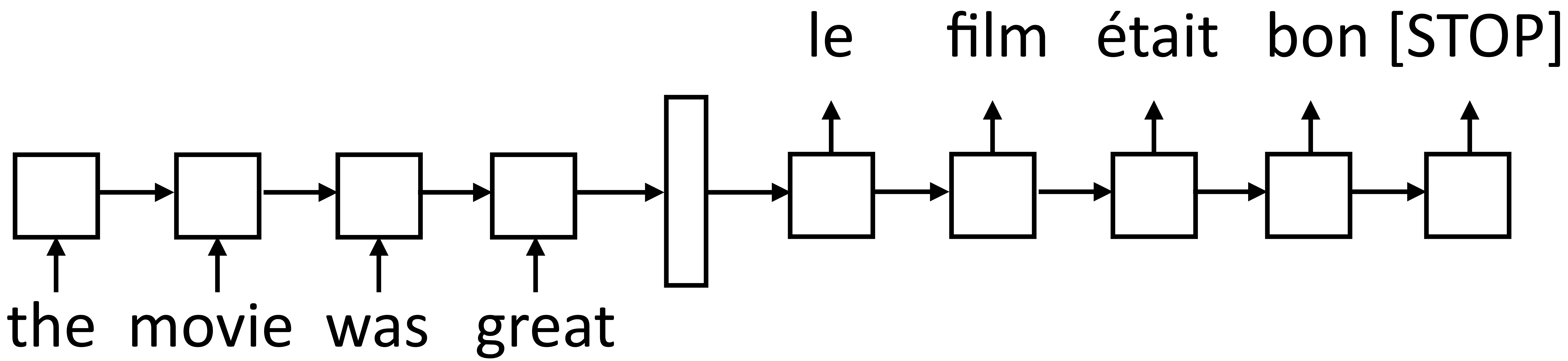




- Both LSTMs and convolutional layers transform the input using context
- LSTM: “globally” looks at the entire sentence (but local for many problems)
- CNN: local depending on filter width + number of layers

Encoder-Decoder

- Encode a sequence into a fixed-sized vector



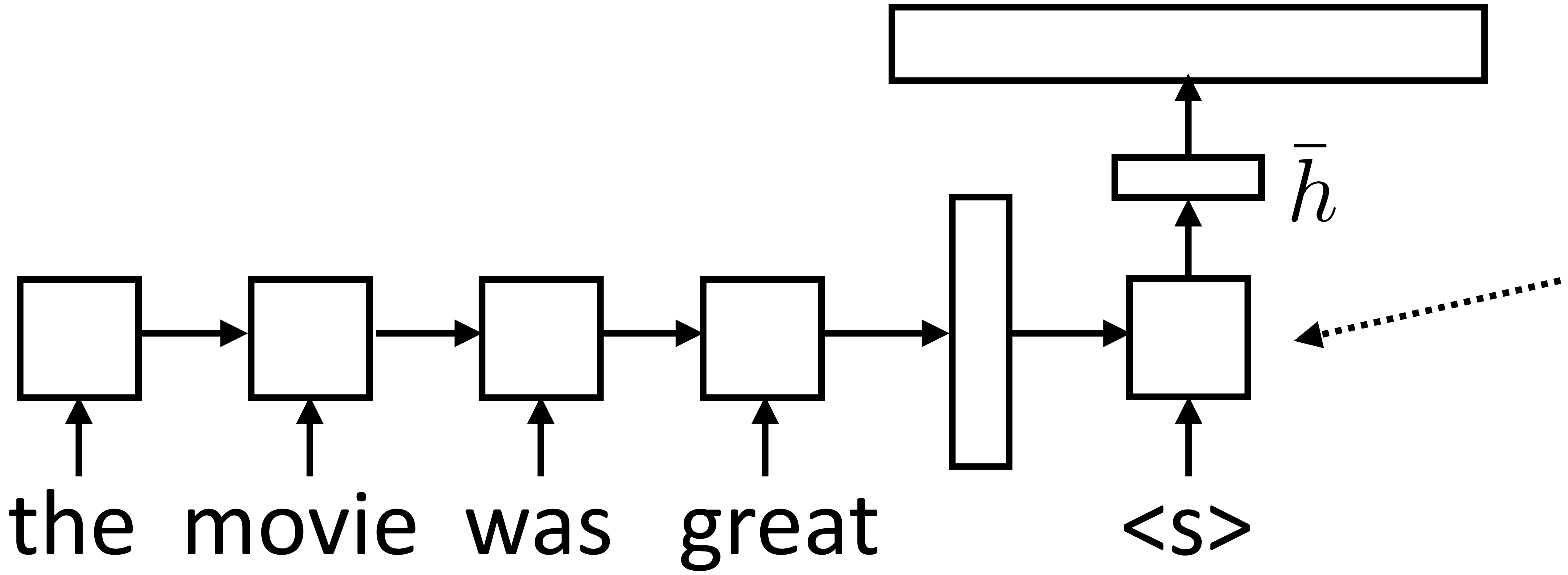
- Now use that vector to produce a series of tokens as output from a separate LSTM *decoder*
- Machine translation, NLG, summarization, dialog, and many other tasks (e.g., semantic parsing, syntactic parsing) can be done using this framework.

Sutskever et al. (2014)

Model

- Generate next word conditioned on previous word as well as hidden state

- W size is |vocab| x |hidden state|, softmax over entire vocabulary



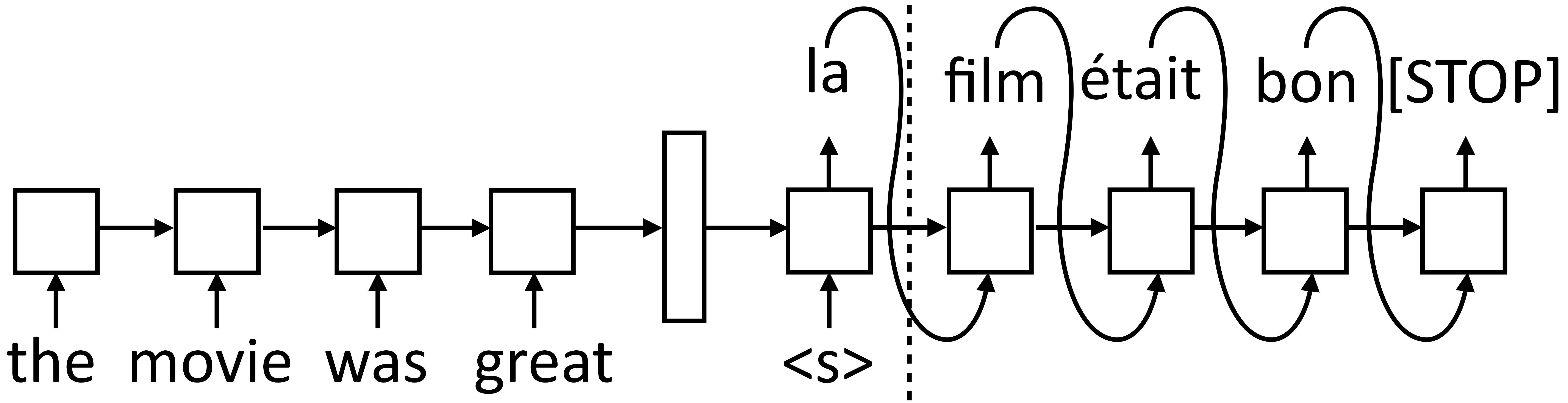
$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Decoder has separate parameters from encoder, so this can learn to be a language model (produce a plausible next word given current one)

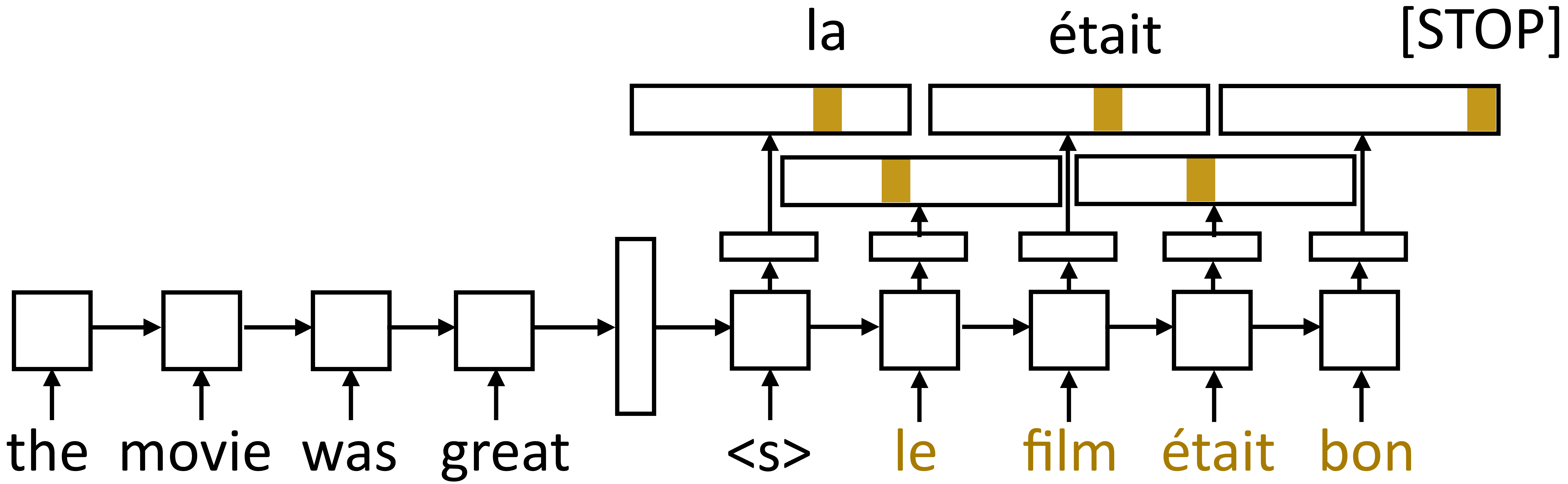
Inference

- Generate next word conditioned on previous word as well as hidden state



- During inference: need to compute the argmax over the word predictions and then feed that to the next RNN state
- Decoder is advanced one state at a time until [STOP] is reached

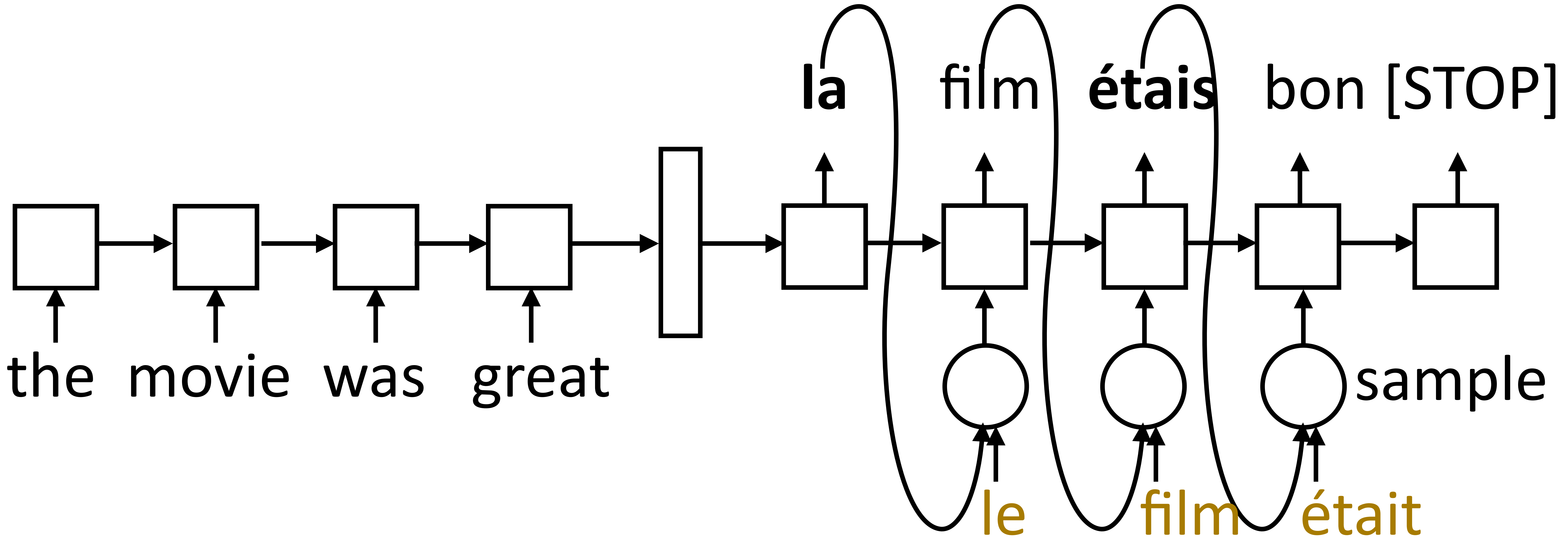
Training



- Objective: maximize $\sum_{(\mathbf{x},\mathbf{y})} \sum_{i=1}^{n} \log P(y_i^* | \mathbf{x}, y_1^*, \ldots, y_{i-1}^*)$
- One loss term for each target-sentence word, feed the correct word regardless of model's prediction (called "teacher forcing")

Training: Scheduled Sampling

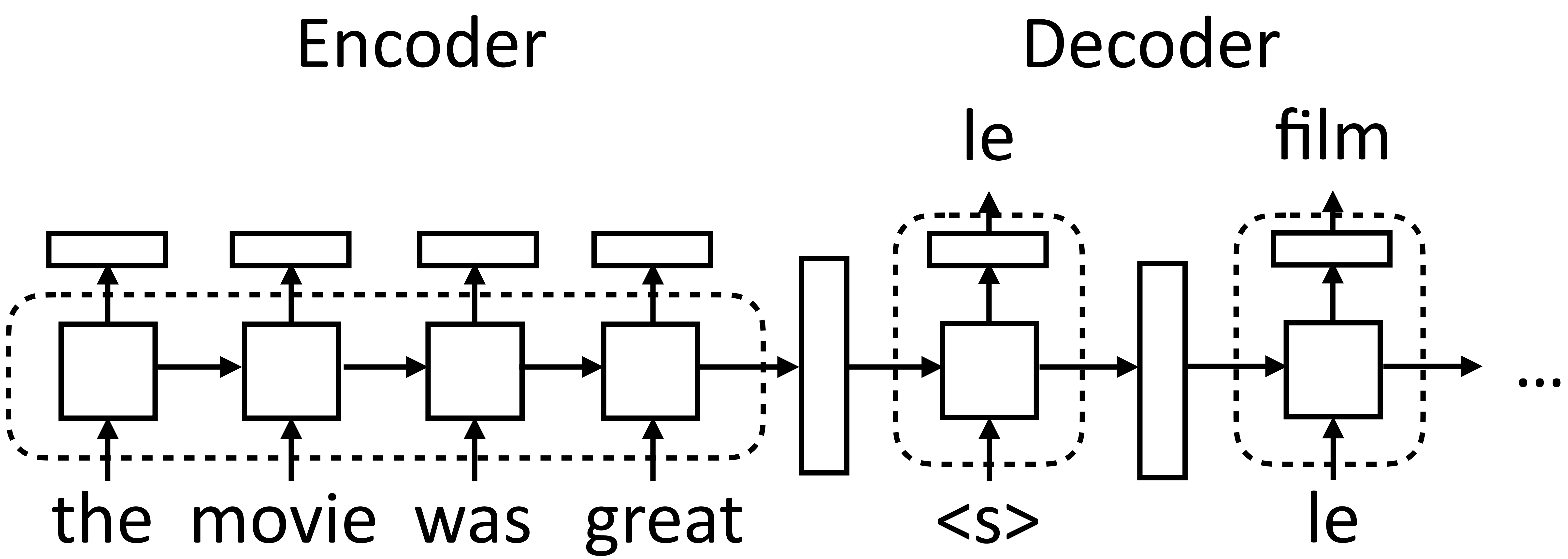
- Model needs to do the right thing even with its own predictions



- Scheduled sampling: with probability *p*, take the gold (human) translation as input, else take the model's prediction
- Starting with *p* = 1 and decaying it works best

Bengio et al. (2015)

Implementing seq2seq Models



- Encoder: consumes sequence of tokens, produces a vector. Analogous to encoders for classification/tagging tasks
- Decoder: separate module, single cell. Takes two inputs: hidden state (vector h or tuple (h, c)) and previous token. Outputs token + new state

Implementation Details

- Sentence lengths vary for both encoder and decoder:
  - Typically pad everything to the right length
- Encoder: Can be a CNN/LSTM/Transformer…
- Batching is a bit tricky:
  - encoder should use pack_padded_sequence to handle different lengths.
  - The decoder should pad everything to the same length and use a mask to only accumulate "valid" loss terms
  - Label vectors may look like [num timesteps x batch size x num labels]

Implementation Details (cont')

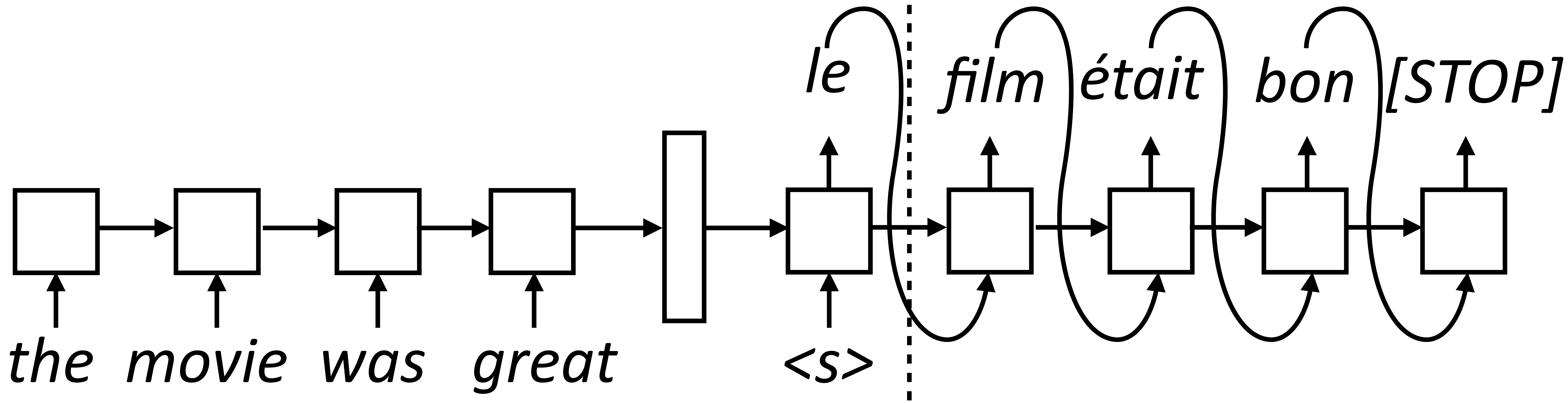
- Decoder: execute one step of computation at a time, so computation graph is formulated as taking one input + hidden state.
  - Test time: do this until you generate the [STOP] token
  - Training time: do this until you reach the gold stopping point
- Beam search: can help with lookahead. Finds the (approximate) highest scoring sequence:

$$\mathrm{argmax}_{\mathbf{y}} \prod_{i=1}^{n} P(y_i | \mathbf{x}, y_1, \ldots, y_{i-1})$$

Decoding Strategies

Greedy Decoding

- Generate next word conditioned on previous word as well as hidden state



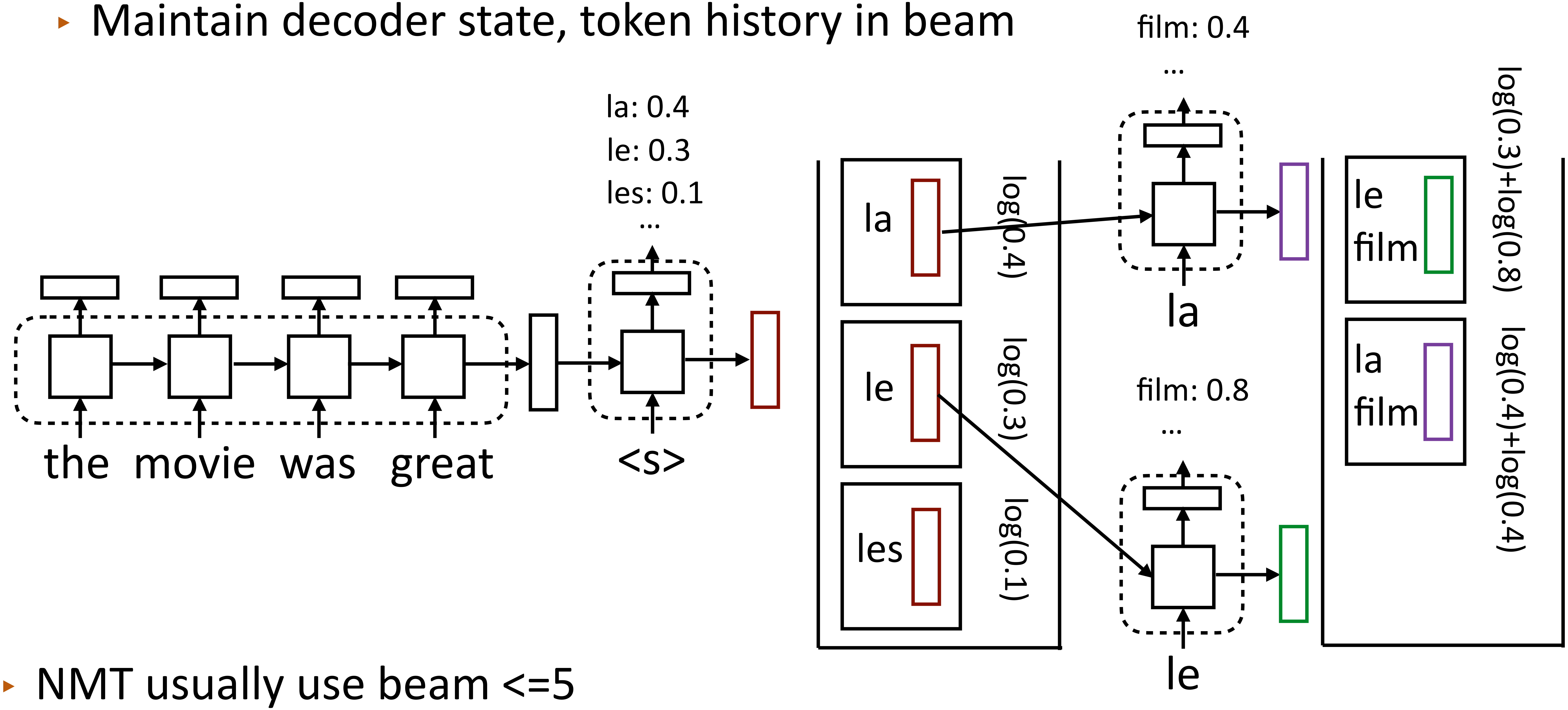
- During inference: need to compute the argmax over the word predictions and then feed that to the next RNN state. This is **greedy decoding**

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$y_{\mathrm{pred}} = \mathrm{argmax}_y P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Beam Search

- Maintain decoder state, token history in beam



- NMT usually use beam <=5
- Keep **both** *film* states! Hidden state vectors are different

Problems with Greedy Decoding

- Only returns one solution, and it may not be optimal
- Can address this with **beam search**, which usually works better…but even beam search may not find the correct answer! (max probability sequence)

Model	Beam-10	
	BLEU	#Search err.
LSTM*	28.6	58.4%
SliceNet*	28.8	46.0%
Transformer-Base	30.3	57.7%
Transformer-Big*	31.7	32.1%

A sentence is classified as search error if the decoder does not find the global best model score.

Stahlberg and Byrne (2019)

“Problems” with Beam Decoding

- For machine translation, the highest probability sequence is often the empty string, i.e.. a single </s> token!   (>50% of the time)

Search	BLEU	Ratio	#Search errors	#Empty
Greedy	29.3	1.02	73.6%	0.0%
Beam-10	30.3	1.00	57.7%	0.0%
Exact	2.1	0.06	0.0%	51.8%

- Beam search results in *fortuitous search errors* that avoid these bad solutions. NMT usually use beam <=5.
- Exact inference uses depth-first search, but cut off branches that fall below a lower bound.

Stahlberg and Byrne (2019)

Sampling

- Beam search may give many similar sequences, and these actually may be *too close* to the optimal. Can sample instead:

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$y_{\mathrm{sampled}} \sim P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

- Greedy solution can be uninteresting / vacuous for various reasons (so called text *degeneration)*. Sampling can help - especially for some text generation tasks.

Beam Search vs. Sampling

- These are samples from an unconditioned language model GPT-2 (not seq2seq model)

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

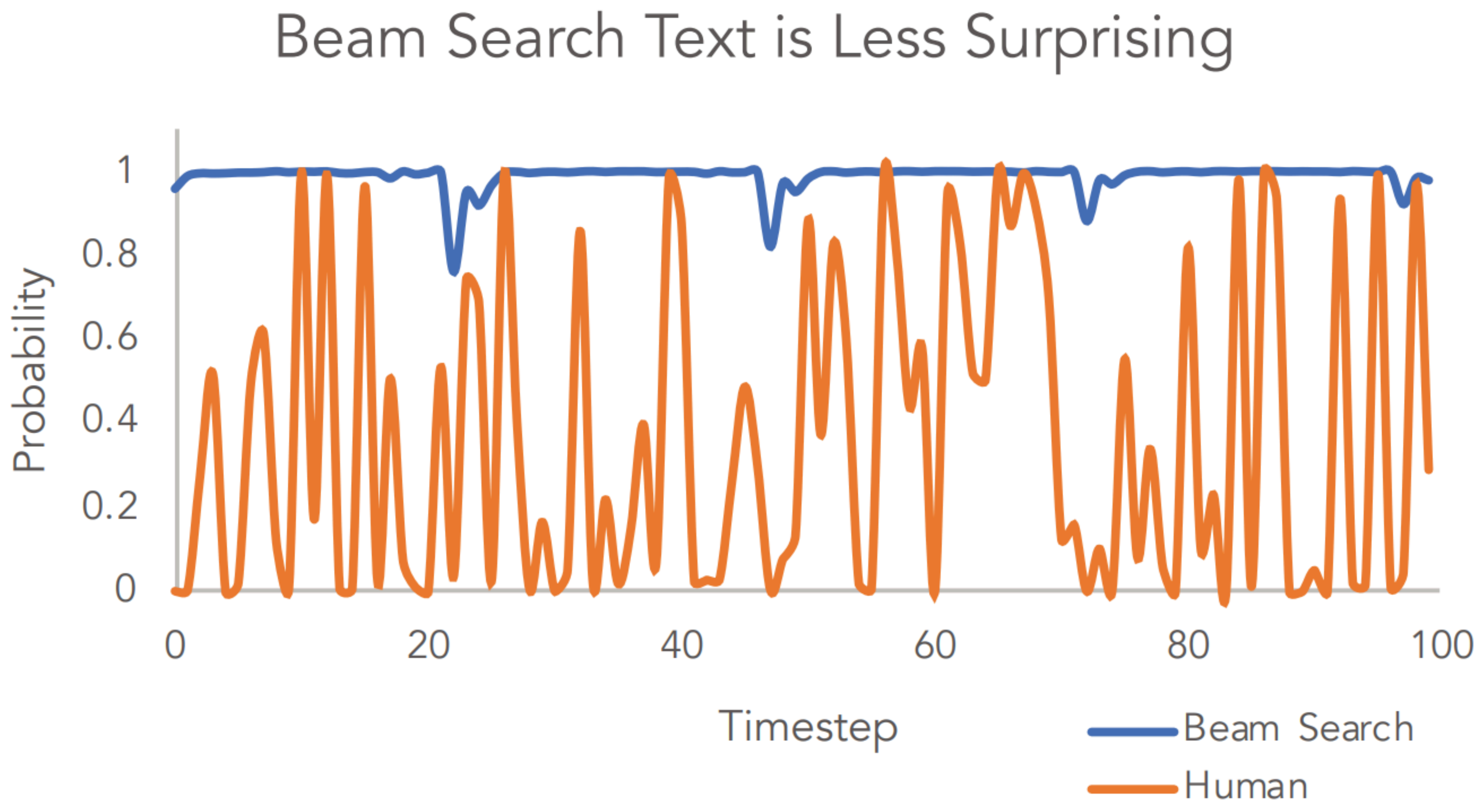
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

- Sampling is better but sometimes draws too far from the tail of the distribution (relatively low prob. over thousands of candidate tokens).

Holtzman et al. (2019)

Beam Search vs. Sampling



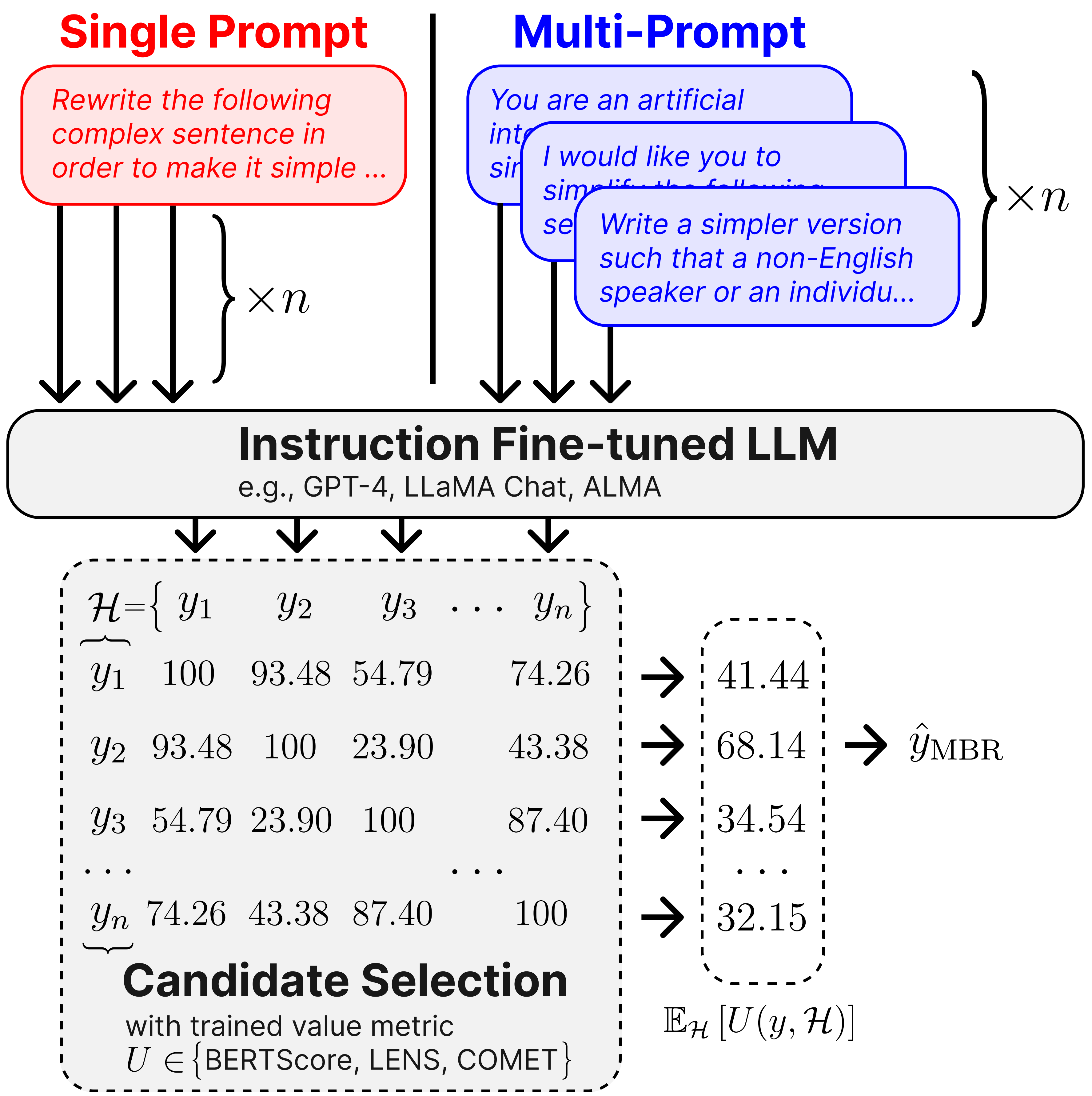
Holtzman et al. (2019)

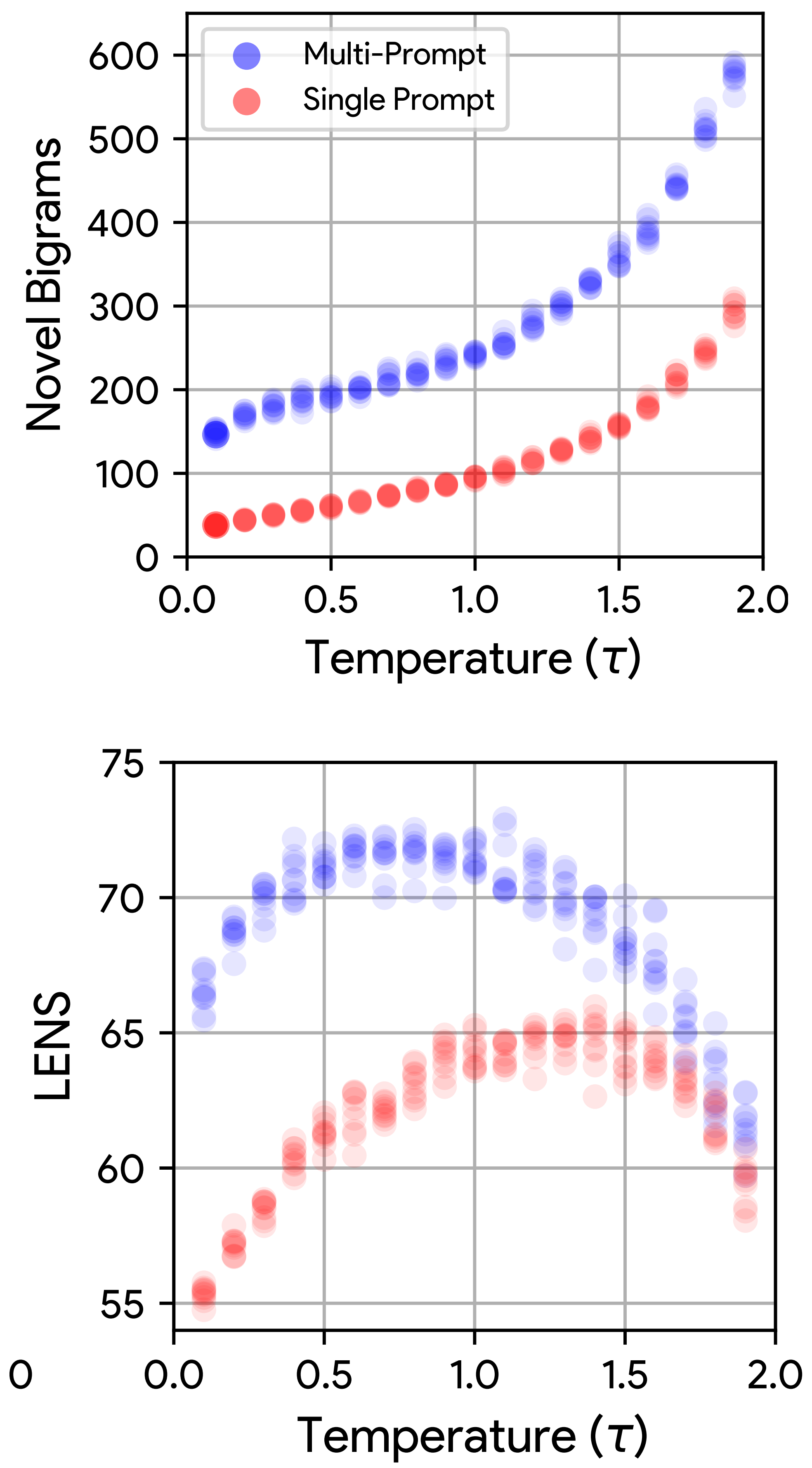
Decoding Strategies

- Greedy
- Beam search
- Sampling (e.g., top-k or Nucleus sampling)
  - Top-k: take the top k most likely words (k=5), sample from those
  - Nucleus: take the top p% (95%) of the distribution, sample from within that

Minimal Bayes Risk (MBR) Decoding

- MBR aims to find an output that maximizes the expected utility, i.e., metrics like COMET (for translation) or LENS (for simplification).



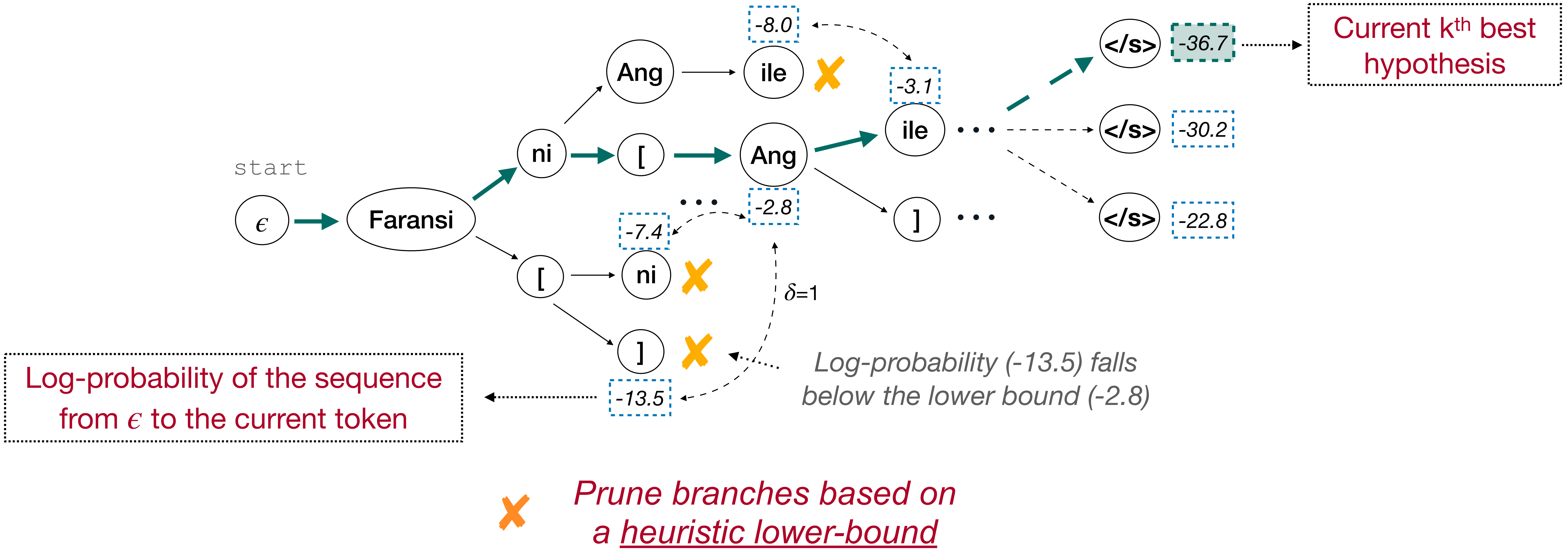


Improving Minimum Bayes Risk Decoding with Multi-Prompt. David Heineman, Yao Dou, Wei Xu (EMNLP 2024)

CODEC - Constrained Decoding

- A branch-and-bound search algorithm with a heuristic lower bound

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ ”Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ ”Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



Constrained Decoding for Cross-lingual Label Projection. Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu (ICLR 2024)

Other Applications of Seq2Seq

Generation Tasks

- There are a range of seq2seq modeling tasks we will address
- For more constrained problems: greedy/beam decoding are usually best
- For less constrained problems: nucleus sampling introduces favorable variation in the output

Less constrained ⟷ More constrained

Unconditioned sampling/ e.g., story generation

Dialogue

Summarization

Data-to-text

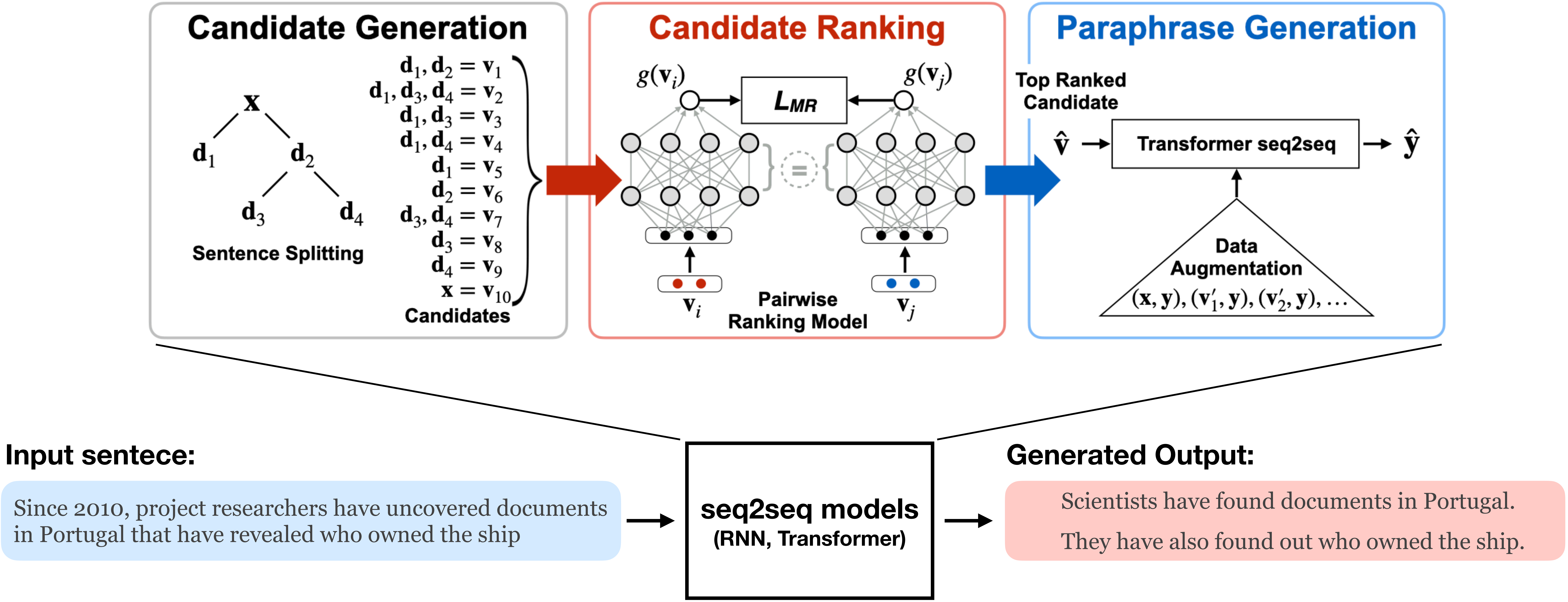
Translation

Text-to-text

Text-to-code

Text-to-Text Generation

- Text Simplification (with readability constraints)

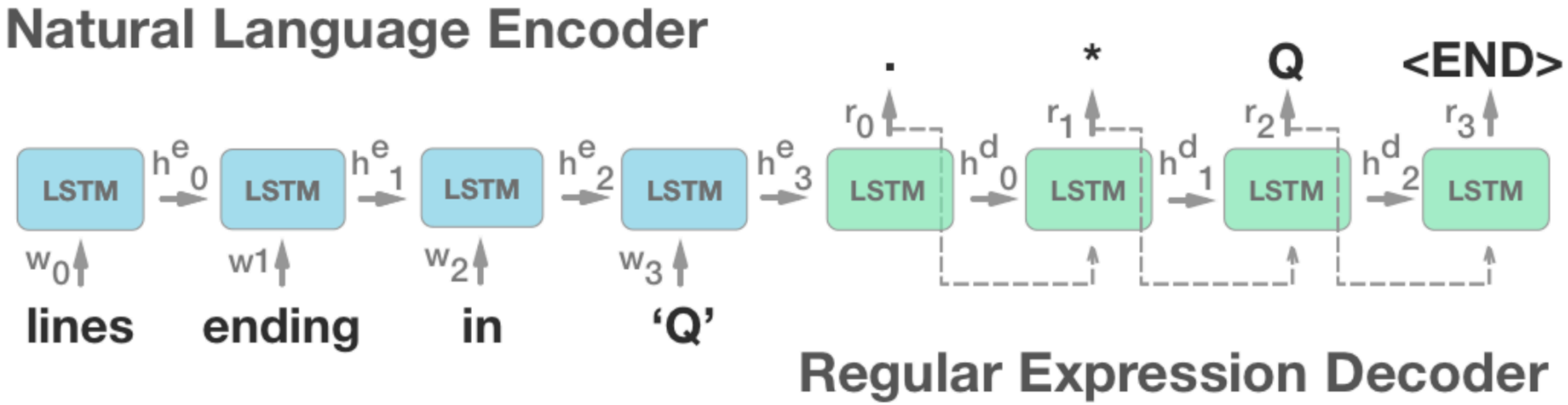


Mounica Maddela, Fernando Alva-Manchego, Wei Xu. “Controllable Text Simplification with Explicit Paraphrasing” (NAACL 2021)

Regex Prediction

- Seq2seq models can be used for many other tasks!

- Predict regex from text



- Problem: requires a lot of data: 10,000 examples needed to get ~60% accuracy on pretty simple regexes

Locascio et al. (2016)

Semantic Parsing as Translation

"what states border Texas"

$$\downarrow$$

$$\lambda\ x\ \texttt{state(}\ x\ \texttt{)}\ \land\ \texttt{borders(}\ x\ ,\ \texttt{e89}\ \texttt{)}$$

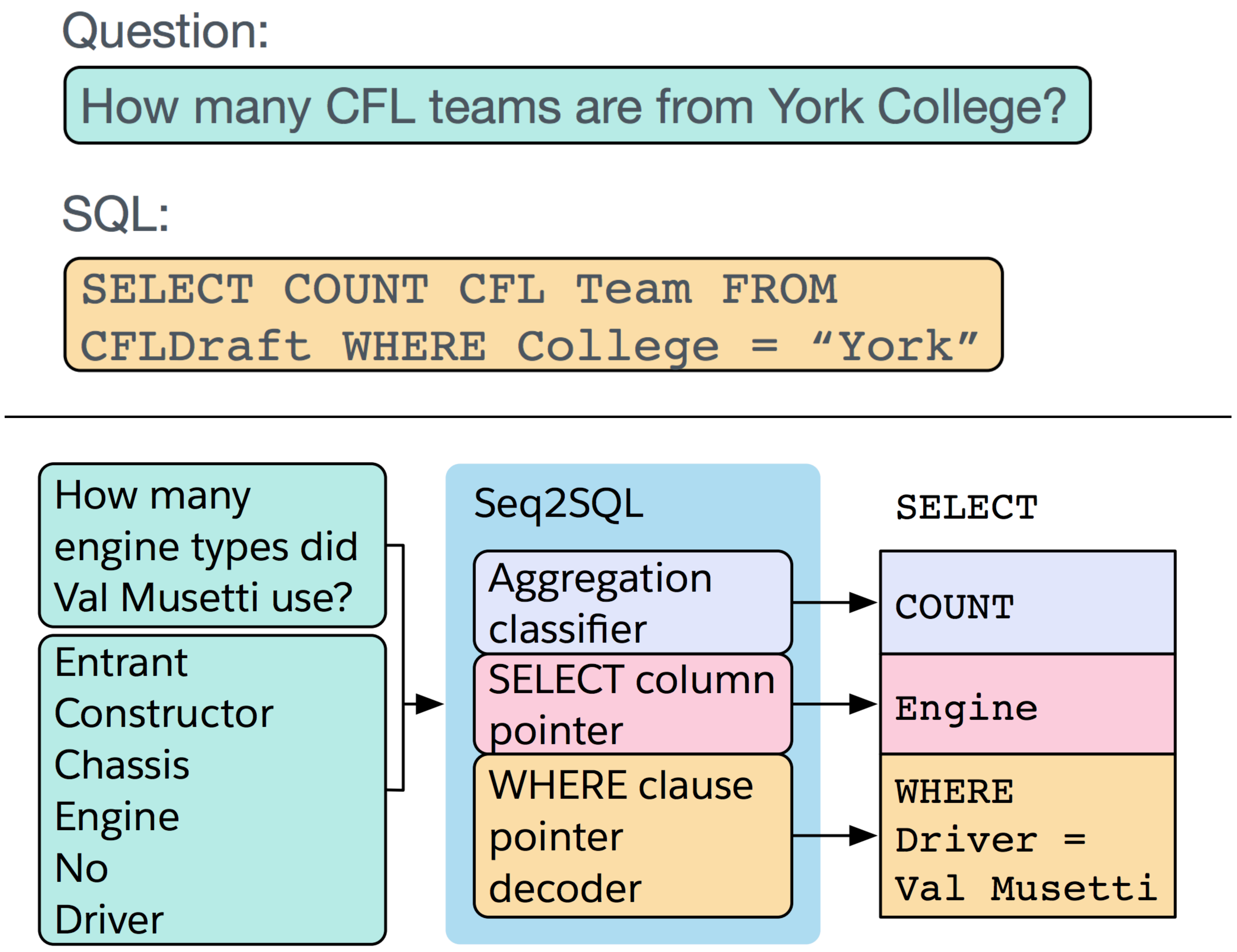
- Write down a linearized form of the semantic parse, train seq2seq models to directly translate into this representation; using copy mechanism
- No need to have an explicit grammar, simplifies algorithms
- Might not produce well-formed logical forms, might require lots of data

Semantic Parsing/Lambda Calculus: https://www.youtube.com/watch?v=OocGXG-BY6k&t=200s

Jia and Liang (2016)

SQL Generation

- Convert natural language description into a SQL query against some DB
- How to ensure that well-formed SQL is generated?
  - Three components
- How to capture column names + constants?
  - Pointer mechanisms



Zhong et al. (2017)

Constrained Decoding for Cross-lingual Label Projection (CODEC)

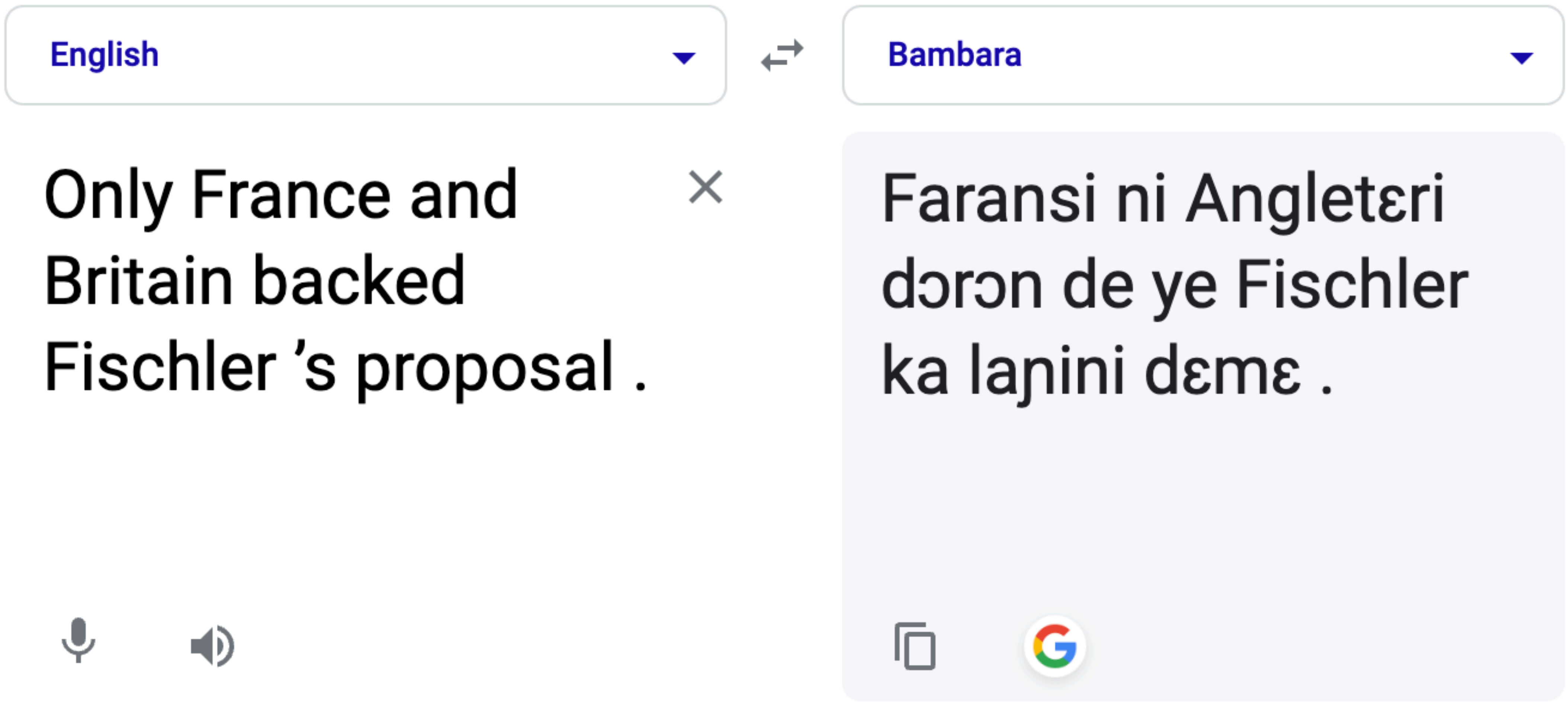



Duong Minh Le Yang Chen Alan Ritter Wei Xu

A better technical solution for marker-based label projection

Key Idea

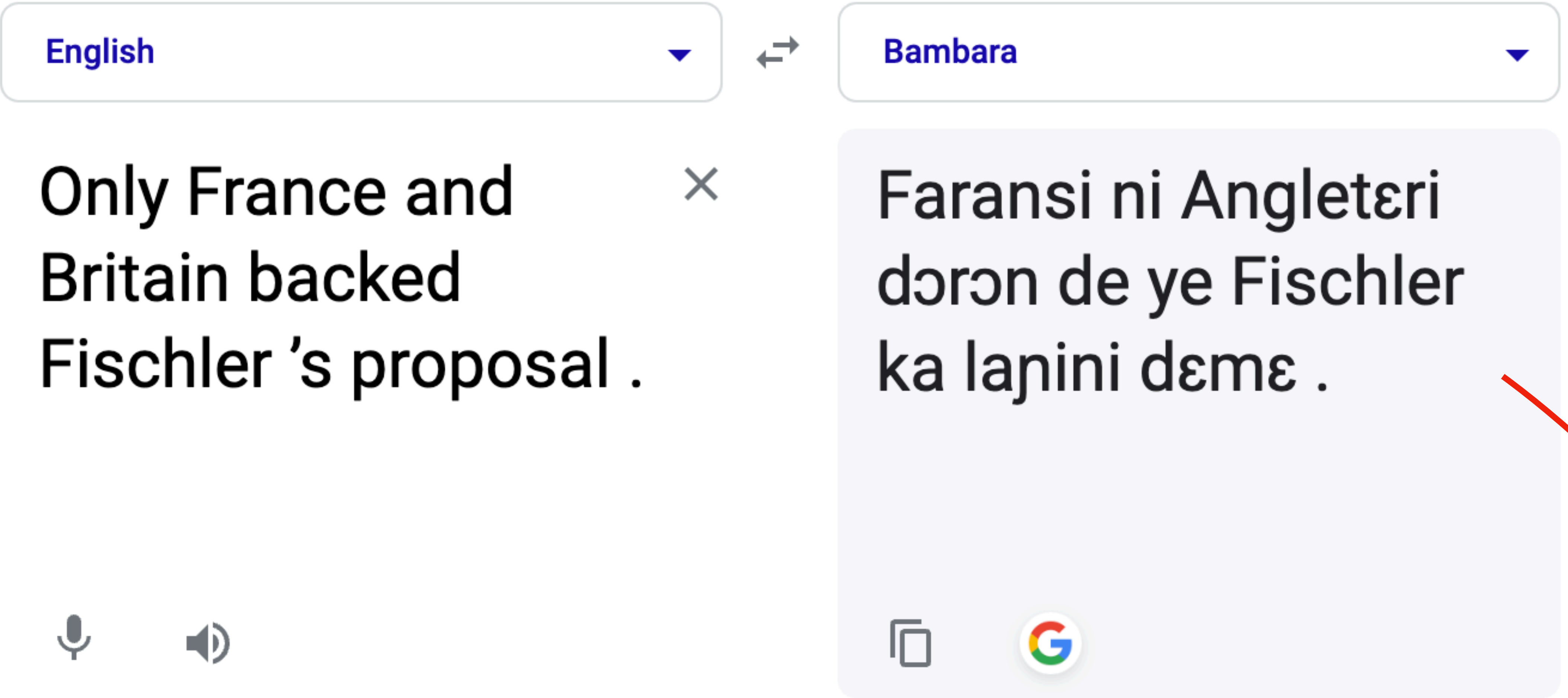
Step 1. Translate the original sentence as usual without markers.



Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Key Idea

Step 1. Translate the original sentence as usual without markers.



Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [ ] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

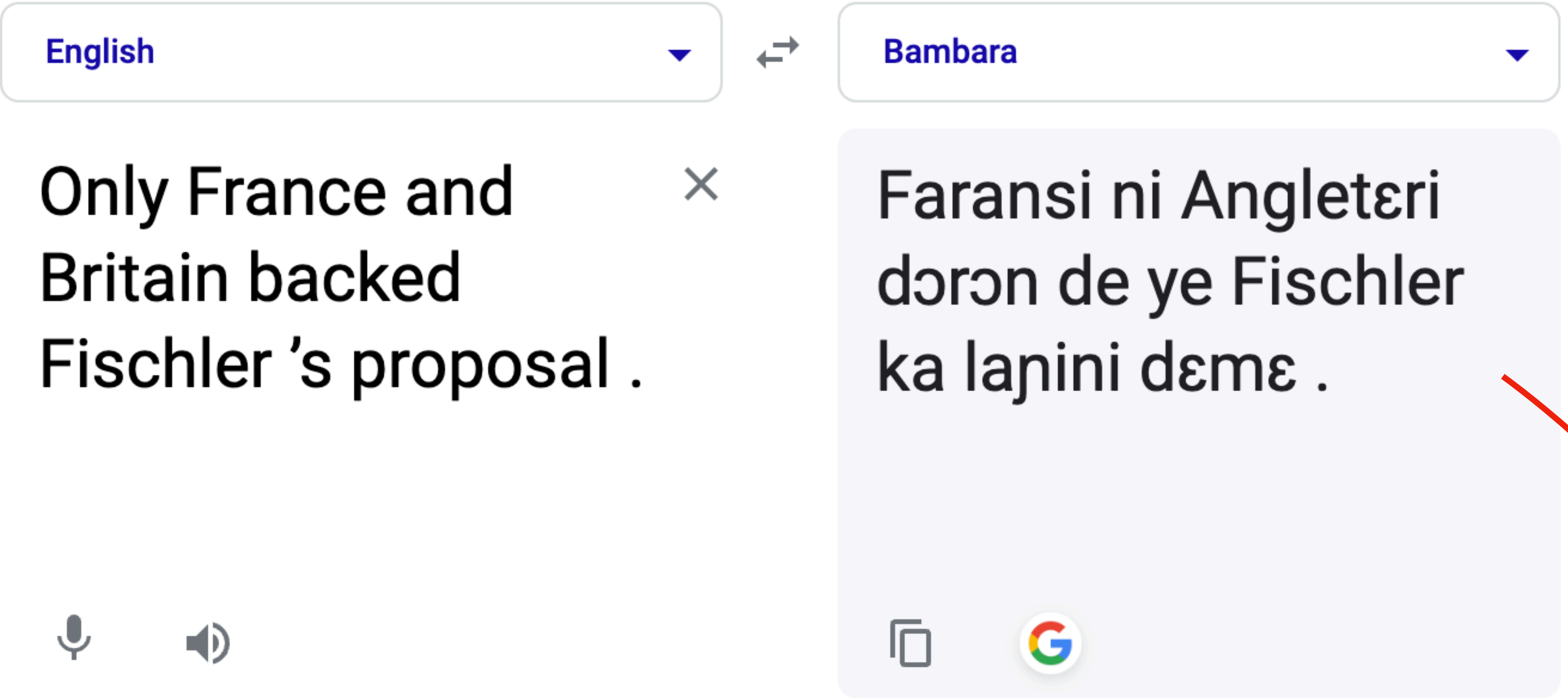
Input sentece:

Only [France] and [Britain] backed [Fischler]'s proposal. → **LLMs (Translation Models)** →

Translated Output:

Key Idea

Step 1. Translate the original sentence as usual without markers.



Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [ ] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Input sentece:

Only [France] and [Britain] backed [Fischler]'s proposal. → **LLMs (Translation Models)** → **Translated Output:**

[Faransi] ni [Angiletɛri] dɔrɔn de ye [Fischler] ka laɲini dɛmɛ .

Key Idea — more formally

Step 1. Translate the original sentence as usual without markers.

$$y^{tmpl} = \arg\max_{y} \log P_{\tau}(y|x)$$

Step 2. Run translation model another time to insert m marker pairs [ ] into y^{tmpl} .

$$y^* = \arg\max_{y \in \mathcal{Y}} \log P_{\tau}(y|x^{mark}; y^{tmpl})$$

$$\boxed{O(n^{2m})}$$

An Efficient Constrained Decoding Algorithm

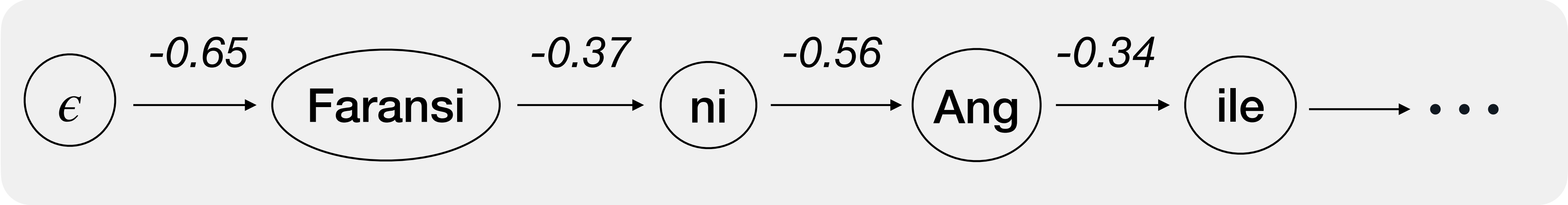
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laŋini dɛmɛ ."

$p_1^i = \log P(y_i^{tmpl} \mid y_{<i}^{tmpl}, x)$ (Conditioned on source text)



$p_2^i = \log P(y_i^{tmpl} \mid y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)

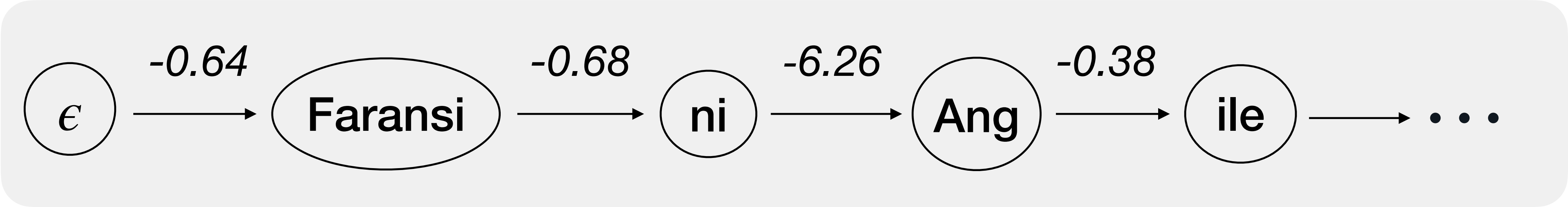
An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ “Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”

$p_1^i = \log P(y_i^{tmpl} \mid y_{<i}^{tmpl}, x)$ (Conditioned on source text)

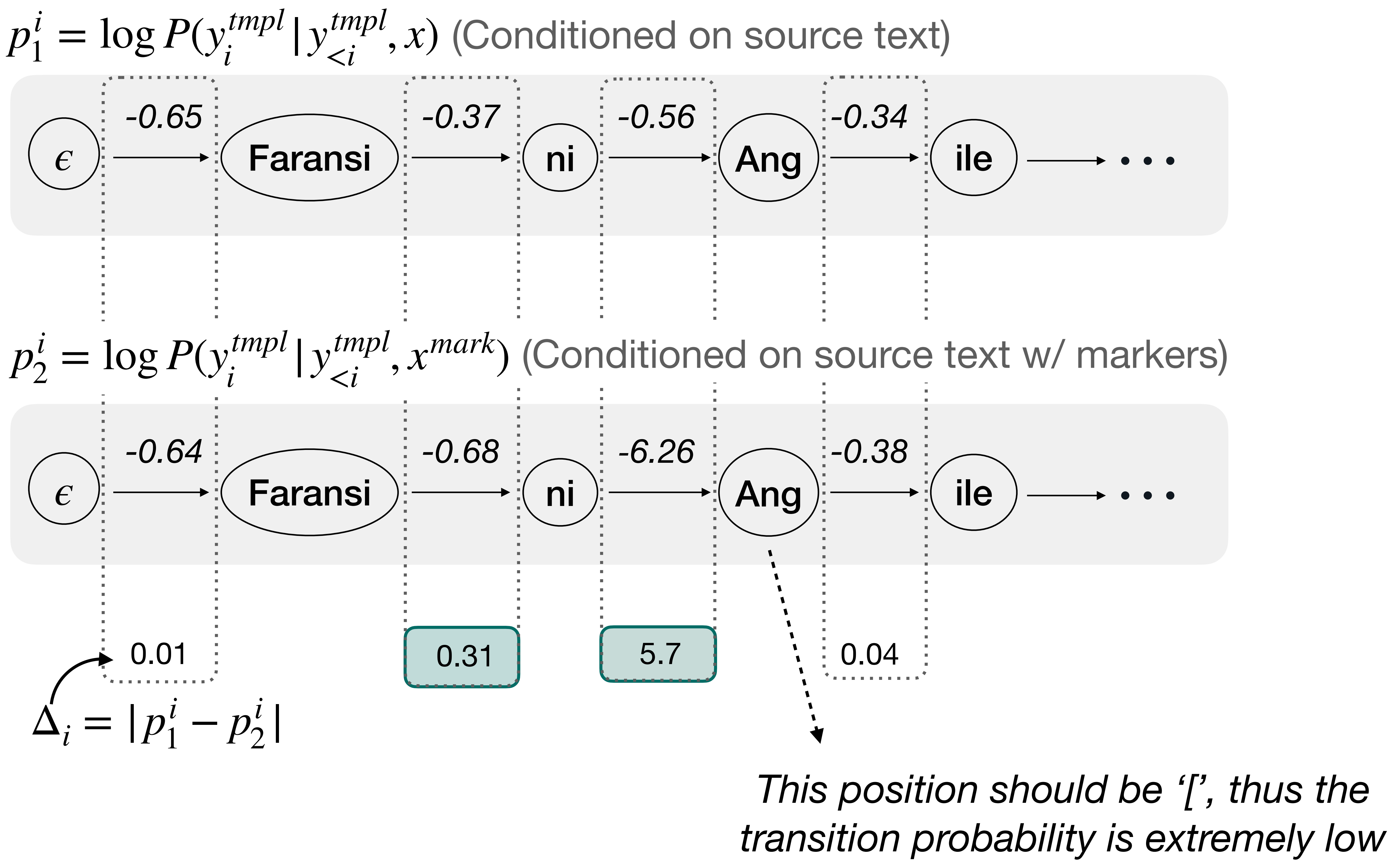
$p_2^i = \log P(y_i^{tmpl} \mid y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)



An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

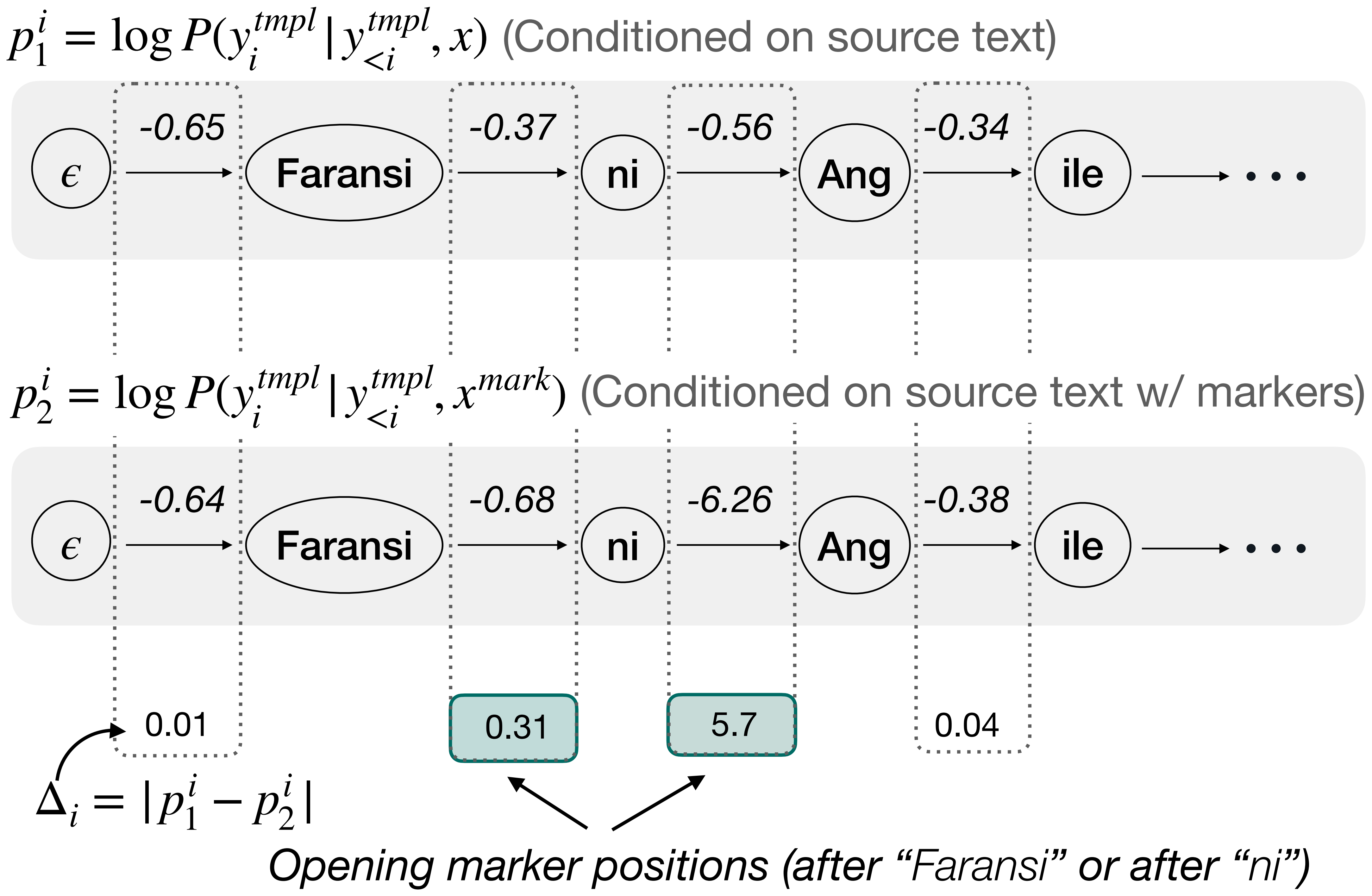
Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

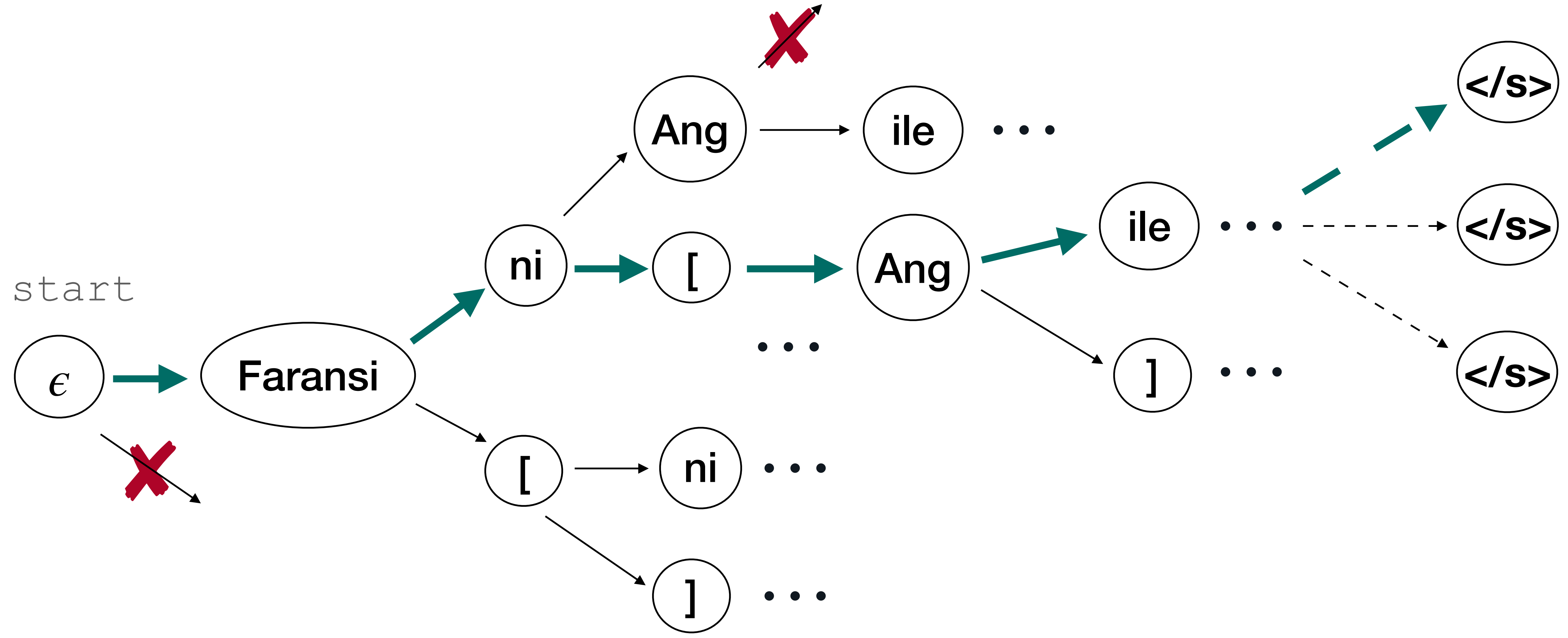
$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



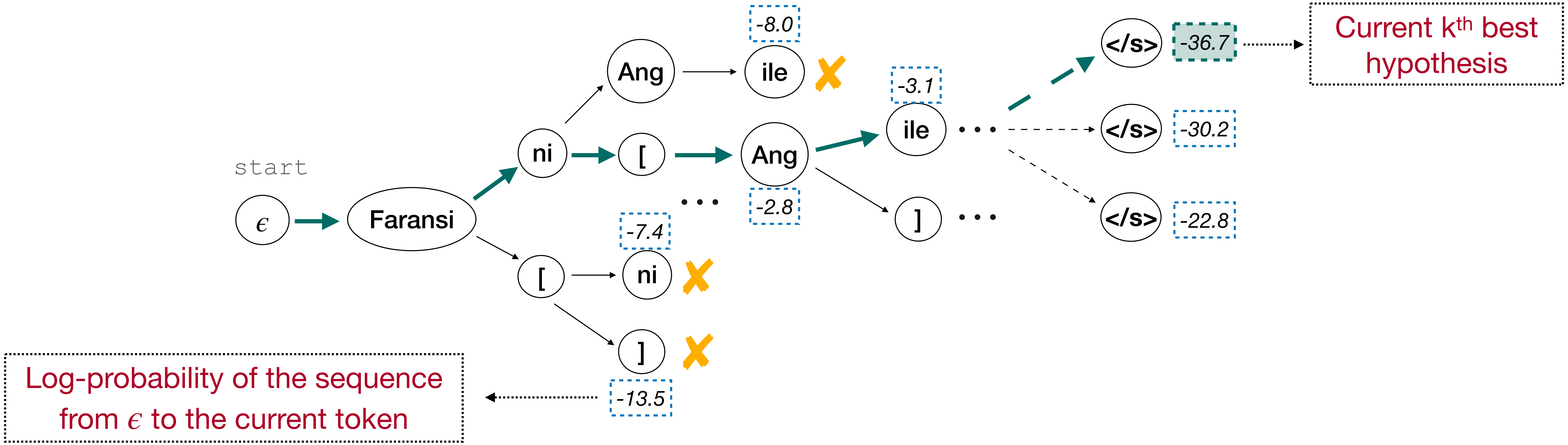
Prune opening-marker positions

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angilətɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."

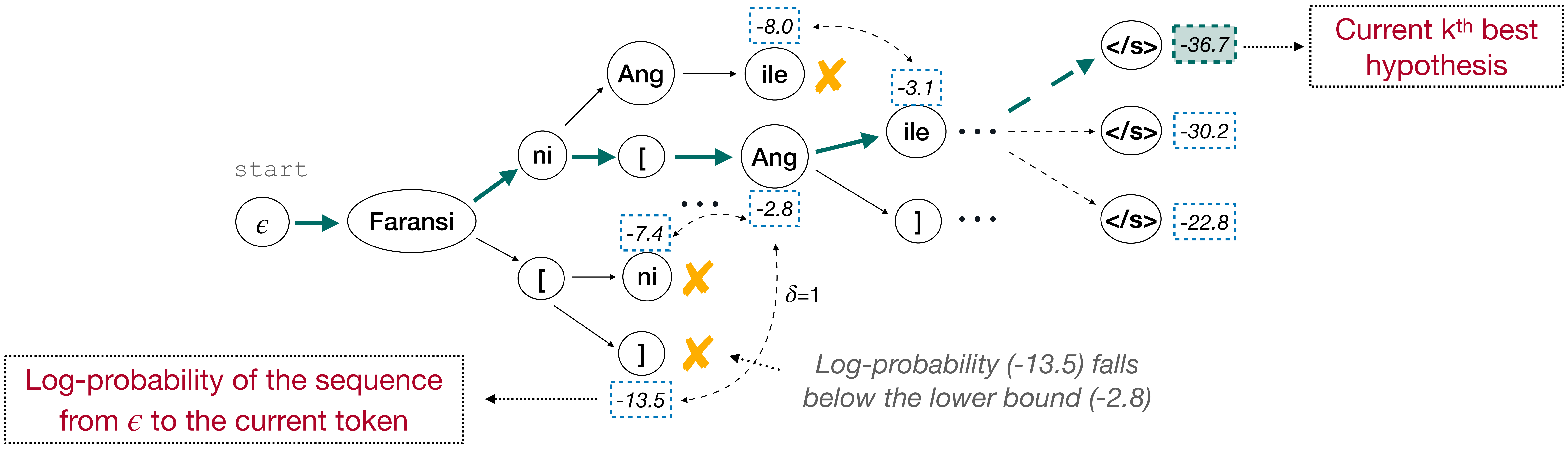


An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ ”Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ ”Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



Prune branches based on a heuristic lower-bound

An Efficient Constrained Decoding Algorithm

Algorithm 1 Constrained_DFS: Searching for top-k best hypotheses

Input x^{mark}: Source sentence with marker, y: translation prefix (default: ϵ), y^{tmpl}: translation template,
L: $[\log P(y_1|x), \log P(y_{1:2}|x), \ldots, \log P(y|x)]$ (default=[0.0]), $\mathcal{M}$: opening marker positions
H: min heap to record the results, k: number of hypotheses, δ: lower bound hyperparameter

$flag \leftarrow$ {check if all markers are generated}
if $y_{|y|} = \texttt{</s>}$ and $flag = \texttt{TRUE}$: **then**
  $H.\text{push}((L_{|y|}, L, y))$ ▷ H sorts by the first element
  if $\text{len}(H) > k$ **then**
    $H.\text{pop}()$
else
  $\mathcal{T} \leftarrow []$
  $w_1 \leftarrow$ {get the next token in y^{tmpl}}
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_1, \log P(w_1|y, x^{mark}))\}$
  $j \leftarrow |y| + 1$ ▷ position of the token to be generated next
  $w_2 \leftarrow$ {get the next marker}
  if $\exists\, w_2$ and not ($w_2 =$ '[' land $j \notin \mathcal{M}$) **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_2, \log P(w_2|y, x^{mark}))\}$
  $\mathcal{T} \leftarrow$ {sort $\mathcal{T}$ by the second element in decreasing order}
  for $(w, p) \in \mathcal{T}$ **do**
    $logp \leftarrow L_{|y|} + p$
    $\gamma \leftarrow$ {compute lower bound following Eq 7}
    if $logp > \gamma$ **then**
      Constrained_DFS($x^{mark}, y \cdot w, y^{tmpl}, L \cup \{logp\}, \mathcal{M}, H, k, \delta$)
return H

Experiment Results

CODEC outperforms GPT-4, EasyProject and Awesome-align for NER and Event Extraction tasks.

- **Label Projection baselines:**
  - Alignment-based (***Awes-align***): Utilize a word-alignment system (*Awesome-align*[1]) to perform label projection
  - Marker-based (***EasyProject***): insert markers into the source sentence then translate

- **Zero-shot Cross-lingual transfer (*FT_{En}*)**

  The multilingual model is fine-tuned only on the English data

[1] *Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora.* In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2112–2128, Online, April 2021

Experiment Results

More importantly, CODEC shines on low-resource languages, such as MasakhaNER 2.0 dataset.

Lang.	GPT-4†	FT_{En}	Translate-train		
			Awes-align	EasyProject	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)
Ewe	75.5	75.3	78.3	78.5	**79.1** (+3.8)
Fon	19.4	49.6	59.3	61.4	**65.5** (+15.9)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)
Luo	**55.2**	35.8	46.5	50.2	49.6 (+13.8)
Mossi	44.2	45.0	52.2	53.1	**55.6** (+10.6)
Chichewa	75.8	**79.5**	75.1	75.3	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)
Kiswahili	82.6	**87.7**	82.4	83.6	83.1 (-4.6)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)
isiXhosa	69.5	24.0	61.7	**71.1**	70.4 (+46.4)
Yoruba	**58.2**	36.0	38.1	36.8	41.4 (+5.4)
isiZulu	60.2	43.9	68.9	73.0	**74.8** (+30.9)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)

- NER: mDeBERTa-v3
- MT: NLLB

prior marker-based approach cannot do this

Experiment Results

"Translate-test" - CODEC can also translate test data in source language into a high-resource language to run inference on, then project predicted span labels back to the test data.

Lang.	GPT-4†	FT_{En}	**Translate-train**			**Translate-test**	
			Awes-align	EasyProject	CODEC (Δ_{FT})	Awes-align	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	**55.6** (+18.5)
Ewe	75.5	75.3	78.3	78.5	**79.1** (+3.8)	72.5	**79.1** (+3.8)
Fon	19.4	49.6	59.3	61.4	**65.5** (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	**73.7** (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	**77.2**	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	**78.0** (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	**82.4**	82.3 (+7.0)
Luo	**55.2**	35.8	46.5	50.2	49.6 (+13.8)	52.6	52.9 (+17.1)
Mossi	44.2	45.0	52.2	53.1	**55.6** (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	**79.5**	75.1	75.3	76.8 (-2.7)	78.0	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)	67.0	**78.4** (+43.2)
Kiswahili	82.6	**87.7**	82.4	83.6	83.1 (-4.6)	80.2	81.5 (-6.2)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)	**81.4**	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	**73.5** (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	**67.2** (+23.0)
isiXhosa	69.5	24.0	61.7	**71.1**	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	**58.2**	36.0	38.1	36.8	41.4 (+5.4)	49.1	58.0 (+22.0)
isiZulu	60.2	43.9	68.9	73.0	**74.8** (+30.9)	64.1	**76.9** (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	**70.4** (+16.0)

A Lot More Experiments in the Paper

- Using different MT systems:

  NLLB (600m, 1.3b, 3b) , M2M, mBART50 many-to-many, Google Translate

- Using different encoder LLMs for Word Alignment, NER. Event Extraction:

  mBERT, mDebertaV3, AfroXLMR, Glot500 — specialized for African languages

- Compare to a modified version of beam search with the constrained search space
- And more ….

Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu. "Constrained Decoding for Cross-lingual Label Projection" (ICLR 2024)

Recent Work and more are ongoing …

EMNLP 2024 papers: (1) decoding; (2) multilingual multi-domain; (3) specialized domain, such as medicine.

Improving Minimum Bayes Risk Decoding with Multi-Prompt

David Heineman, Yao Dou, Wei Xu
School of Interactive Computing, Georgia Institute of Technology
{david.heineman, douy}@gatech.edu; wei.xu@cc.gatech.edu

Abstract

While instruction fine-tuned LLMs are effective text generators, sensitivity to prompt construction makes performance unstable and sub-optimal in practice. Relying on a single 'best' prompt cannot capture all differing approaches to a generation problem. Using this observation, we propose *multi-prompt decoding*, where many candidate generations are decoded from a prompt bank at inference-time. To ensemble candidates, we use Minimum Bayes Risk (MBR) decoding, which selects a final output using a trained value metric. We show multi-prompt improves MBR across a comprehensive set of conditional generation tasks (Figure 1), and show this is a result of estimating a more diverse and higher quality candidate space than that of a single prompt. Further experiments confirm multi-prompt improves generation across tasks, models and metrics.[1]

1 Introduction

Minimum Bayes Risk (MBR) decoding (Bickel and Doksum, 1977) improves the generation quality of large language models (LLMs) over standard, single-output decoding methods, such as beam search and sampling. MBR generates a set of candi-

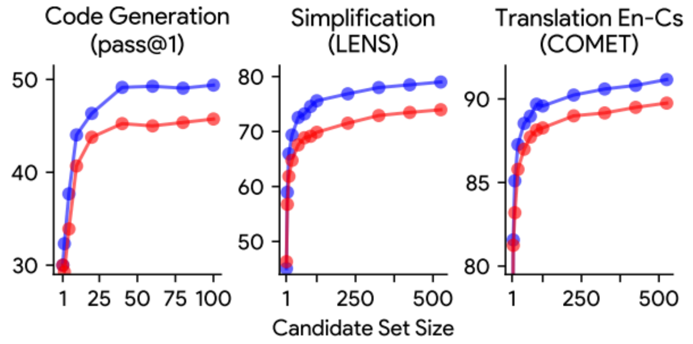
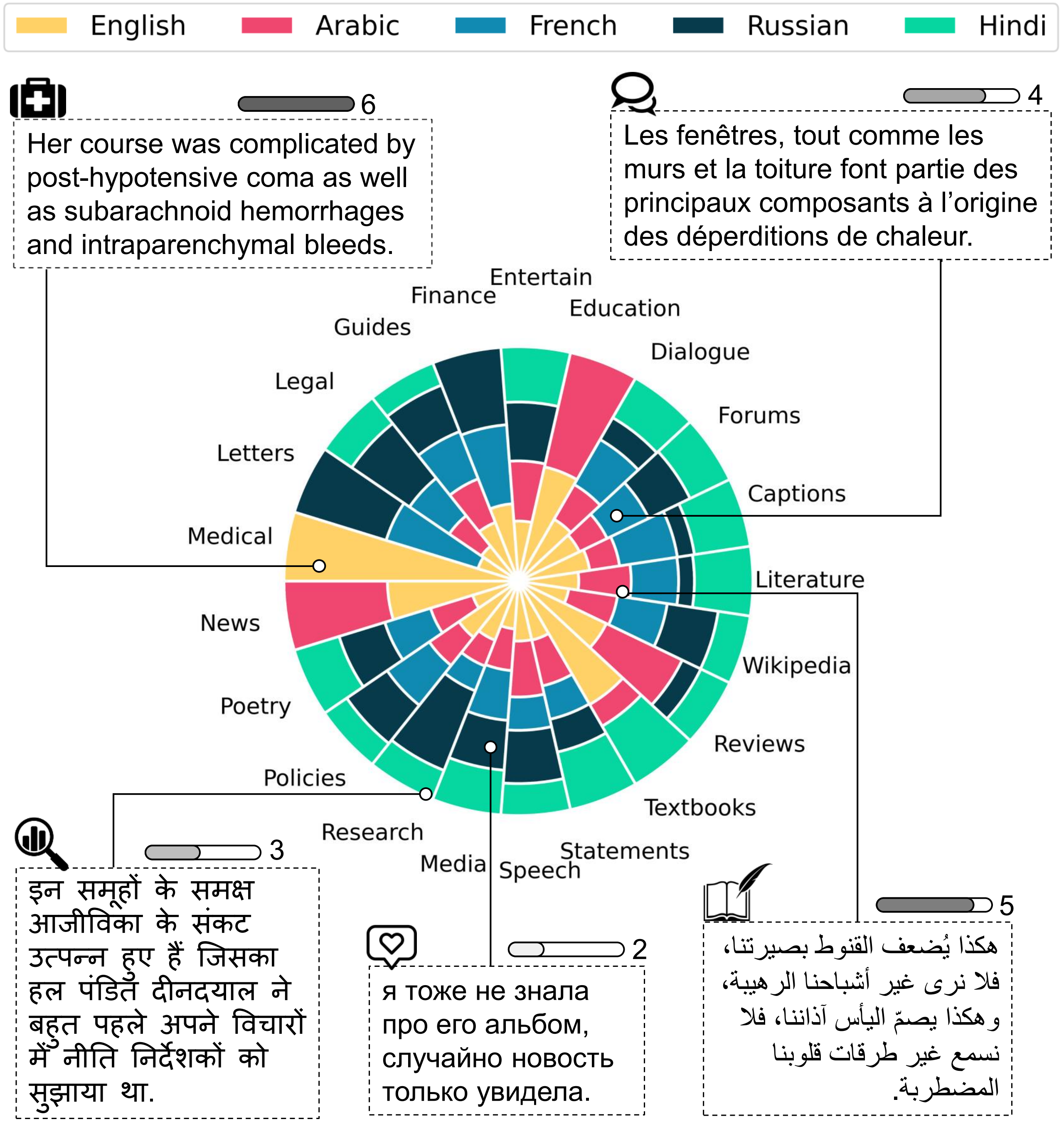


Figure 1: Multi-prompt and single prompt MBR results for code generation on HUMANEVAL, text simplification on SIMPEVAL, and translation on WMT '22 EN-CS generated with open-source 7B LLMs (details in §4).

set. Prior work has found success using sampling-based decoding to generate diverse hypotheses (Eikema and Aziz, 2020; Freitag et al., 2022a, 2023a). However, naively increasing the sampling temperature eventually degrades the quality of the candidates. Recently, instruction fine-tuned LLMs (Ouyang et al., 2022; Chung et al., 2022) have opened up the possibility of writing *prompts* in various formats to elicit higher diversity generations. As these models are observed to be sensitive to prompt design, a slight change in phrasing or the inclusion of more relevant example can signif-

README++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu
College of Computing
Georgia Institute of Technology
{tareknaous, michaeljryan, antonlavrouk, mchandra9}@gatech.edu; wei.xu@cc.gatech.edu



MEDREADME: A Systematic Study for Fine-grained Sentence Readability in Medical Domain

Chao Jiang
College of Computing
Georgia Institute of Technology
chaojiang@gatech.edu

Wei Xu
College of Computing
Georgia Institute of Technology
wei.xu@cc.gatech.edu

Abstract

Medical texts are notoriously challenging to read. Properly measuring their readability is the first step towards making them more accessible. In this paper, we present a systematic study on fine-grained readability measurements in the medical domain at both sentence-level and span-level. We introduce a new dataset MEDREADME, which consists of manually annotated readability ratings and fine-grained complex span annotation for 4,520 sentences, featuring two novel "Google-Easy" and "Google-Hard" categories. It supports our quantitative analysis, which covers 650 linguistic features and automatic complex word and jargon identification. Enabled by our high-quality annotation, we benchmark and improve several state-of-the-art sentence-level readability metrics for the medical domain specifically, which include unsupervised, supervised, and prompting-based methods using recently developed large language models (LLMs). Informed by our fine-grained complex span annotation, we find that adding a single feature, capturing the number of jargon spans, into existing readability formulas can significantly improve their correlation with human judgments. We will publicly release the dataset and code.

1 Introduction

If you can't measure it, you can't improve it.
– Peter Drucker

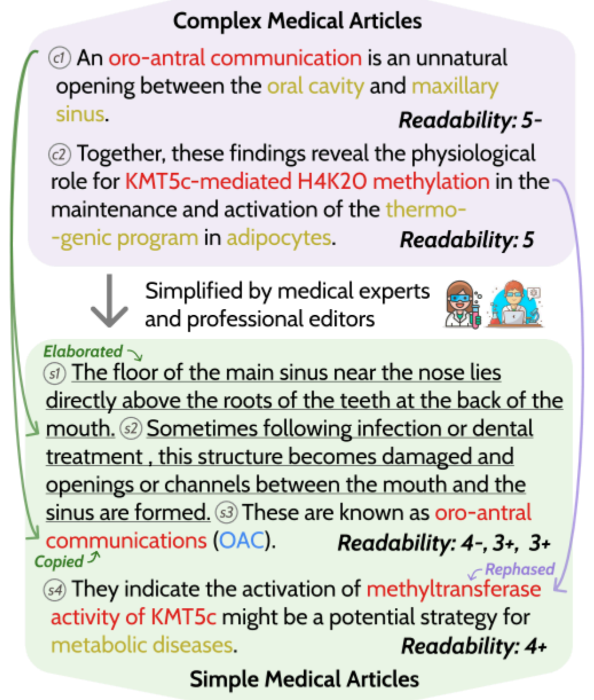


Figure 1: An illustration of our dataset, with sentence readability ratings and fine-grained complex span annotation on 4,520 sentences, including "Google-Hard" and "Google-Easy", abbreviations, and general complex terms, etc. We also analyze how medical jargon are being handled during simplification. e.g., a Google-Hard "oro-antral communication" is copied and elaborated. Some jargon are ignored for clarity.

them more accessible, properly measuring the readability of medical texts is crucial (Rooney et al., 2021; Echuri et al., 2022). However, a high-quality