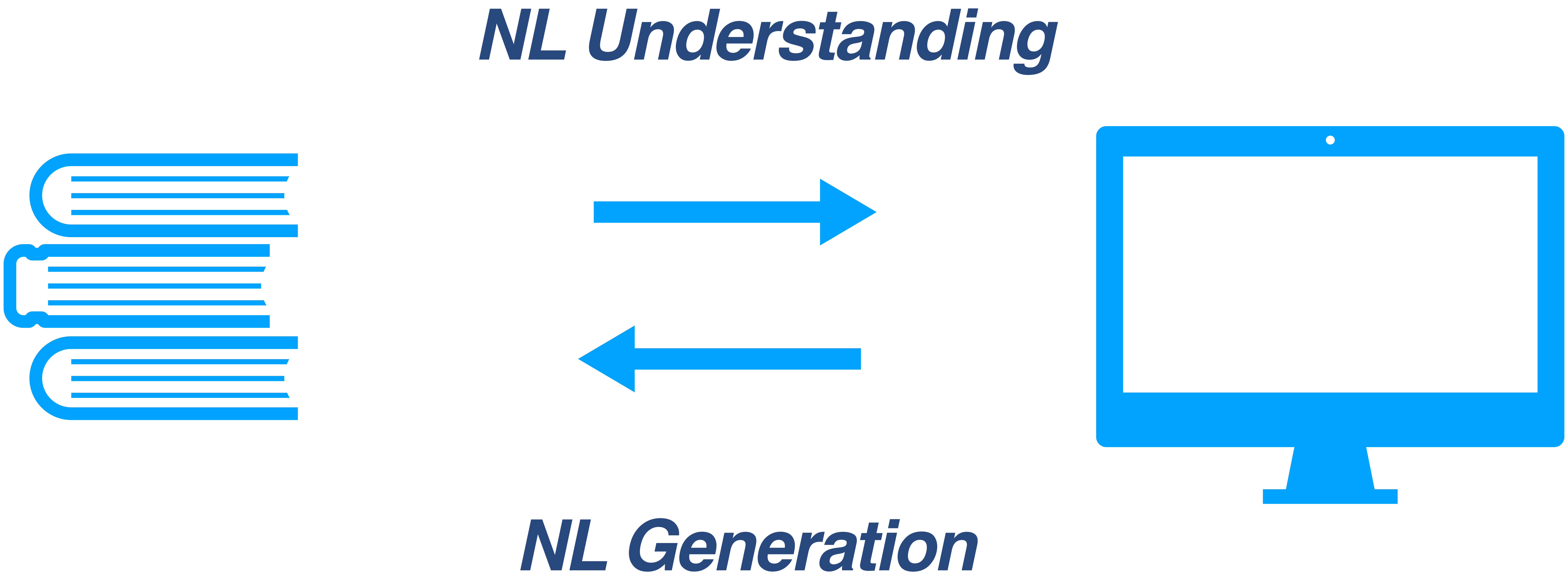
Generating Text

Claire Gardent
Joint work with Angela Fan (Facebook), Antoine Bordes (Facebook) and Chloé Braud (CNRS/IRIT)

cnrs dépasser les frontières

Loria
Laboratoire lorrain de recherche en informatique et ses applications

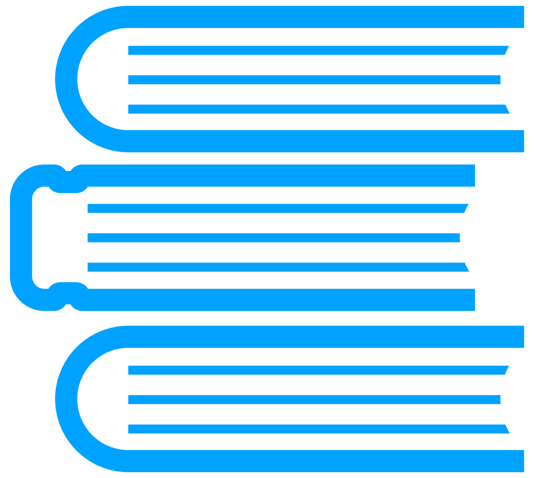
Natural Language Processing



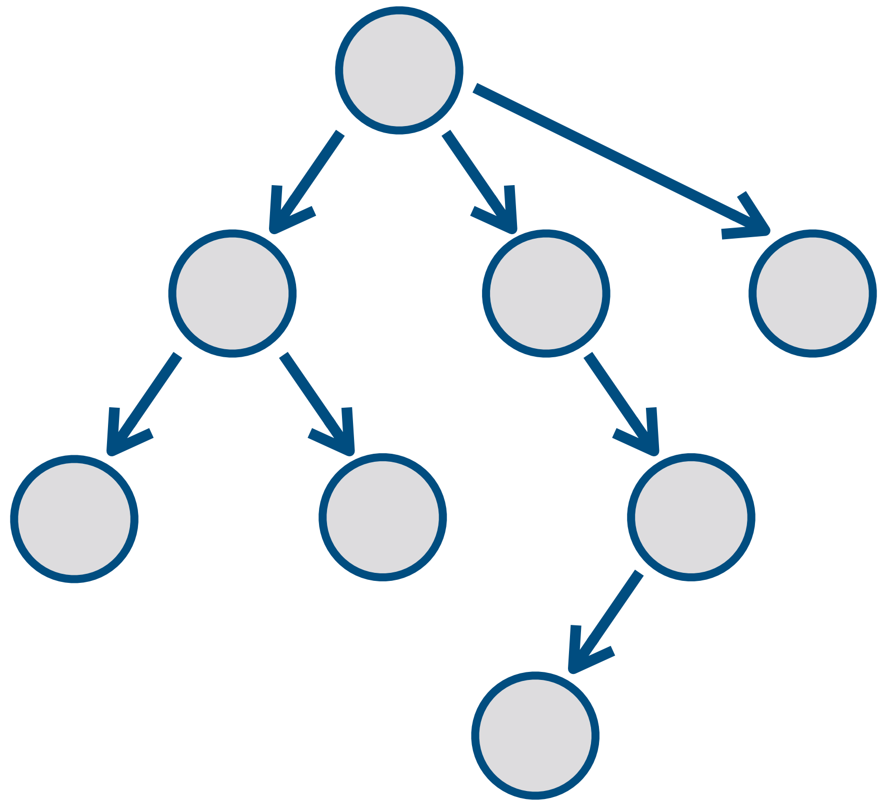
Natural Language Generation

What is NLG useful for ?

Verbalising, Summarising, Querying Knowledge-Bases



Summarising, Simplifying, Paraphrases one or more Text(s)



Converting Graphs into Text

The NLG Hype

ARTIFICIAL INTELLIGENCE

Facebook claims its new chatbot beats Google's as the best in the world

It has also open-sourced the AI system to spur further research.

By Karen Hao

April 29, 2020

Opinion
Artificial intelligence (AI)

This article is more than **3 years old**

AI can write just like me. Brace for the robot apocalypse

Hannah Jane Parkinson

The NLG Hype

Coreference in automatically summarised text

Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... ***Prime Minister Johnson*** was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House.

Generated summary

Boris Johnson has said ***he*** will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

The NLG Hype

Abstract Anaphora in automatically summarised text

Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."

Generated summary

Boris Johnson has said he will raise ***the issue of US diplomat Anne Sacoolas' diplomatic immunity*** with the White House.

The NLG Hype

World Knowledge in automatically summarised text

Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... ***Prime Minister Johnson*** was questioned about the case while speaking to the press at a hospital in Watford. <u>He</u> said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House.

Generated summary

Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

The NLG Hype

Errors in automatically summarised text

Title: Almond-raspberry thumbprint cookies

Deleted

Ingredients: 1 cup butter or margarine softened, 1 cup sugar, 1 can solo almond filling, 2 egg yolks, 1 tsp almond extract, 2 1/2 cups all-purpose flour, 1/2 tsp baking powder, 1/2 tsp salt, 1 can solo raspberry filling

In a large bowl, beat butter and sugar until light and fluffy. Add eggs, one at a time, beating well after each addition. Blend in flour, baking powder, **baking soda**, salt and **baking soda**. Add to creamed mixture and mix well. Stir in nuts and vanilla. Drop dough by rounded table spoonfuls 2 " apart onto ungreased cookie sheets. Bake 8 to 10 minutes or until edges are lightly browned. Cool on wire racks.

Added **Repeated**

The NLG Hype

The factuality and faithfulness issue

A key issue in Neural NLG is correctness. The generated text may

- Contain factually incorrect data
- Include contradictions and repetitions / redundancies
- Omit information present in the input
- Hallucinate information not present in the input

The NLG Hype

KHARI JOHNSON BUSINESS JUN 14, 2022 7:25 PM

LaMDA and the Sentient AI Trap

Arguments over whether Google’s large language model has a soul distract from the real-world problems that plague artificial intelligence.

ARTIFICIAL INTELLIGENCE

OpenAI’s new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

By Will Douglas Heaven

July 20, 2020

Outline

- A short introduction to Neural Generation
- Four Challenges for Neural NLG
  - Long Input
  - Integrating Knowledge
  - Generating into multiple languages
  - Generating long form text

Neural Generation

Neural Generation

- Represents ***(sub)words*** as vectors of real numbers called ***embeddings***
- Generates text by ***predicting the next most probable word*** usin either a Language Model (decoder) or an Encoder-Decoder
- Increasingly uses ***pretrained models***
  - Pretrained word embeddings (Word2Vec, BERT, XLNet ,…)
  - Pretrained decoder (GPT2, LAMDA, …)
  - Pretrained encoder-decoder (T5, BART, …)

Neural Word Representations

Embeddings

Neural Word Representations

- Words with similar contexts have similar meanings

- Words are represented by vectors of real numbers called ***embeddings***

John eats an apple.

Peter eats a pear.

The apple is ripe.

The pear is ripe.

	JOHN	PETER	EAT	APPLE	PEAR	RIPE
JOHN	0	0	1	1	0	0
PETER	0	0	1	0	1	0
EAT	1	1	0	1	1	0
APPLE	1	0	1	0	0	1
PEAR	0	1	1	0	0	1
RIPE	0	0	0	1	1	0

APPLE = <1,0,1,0,0,1>

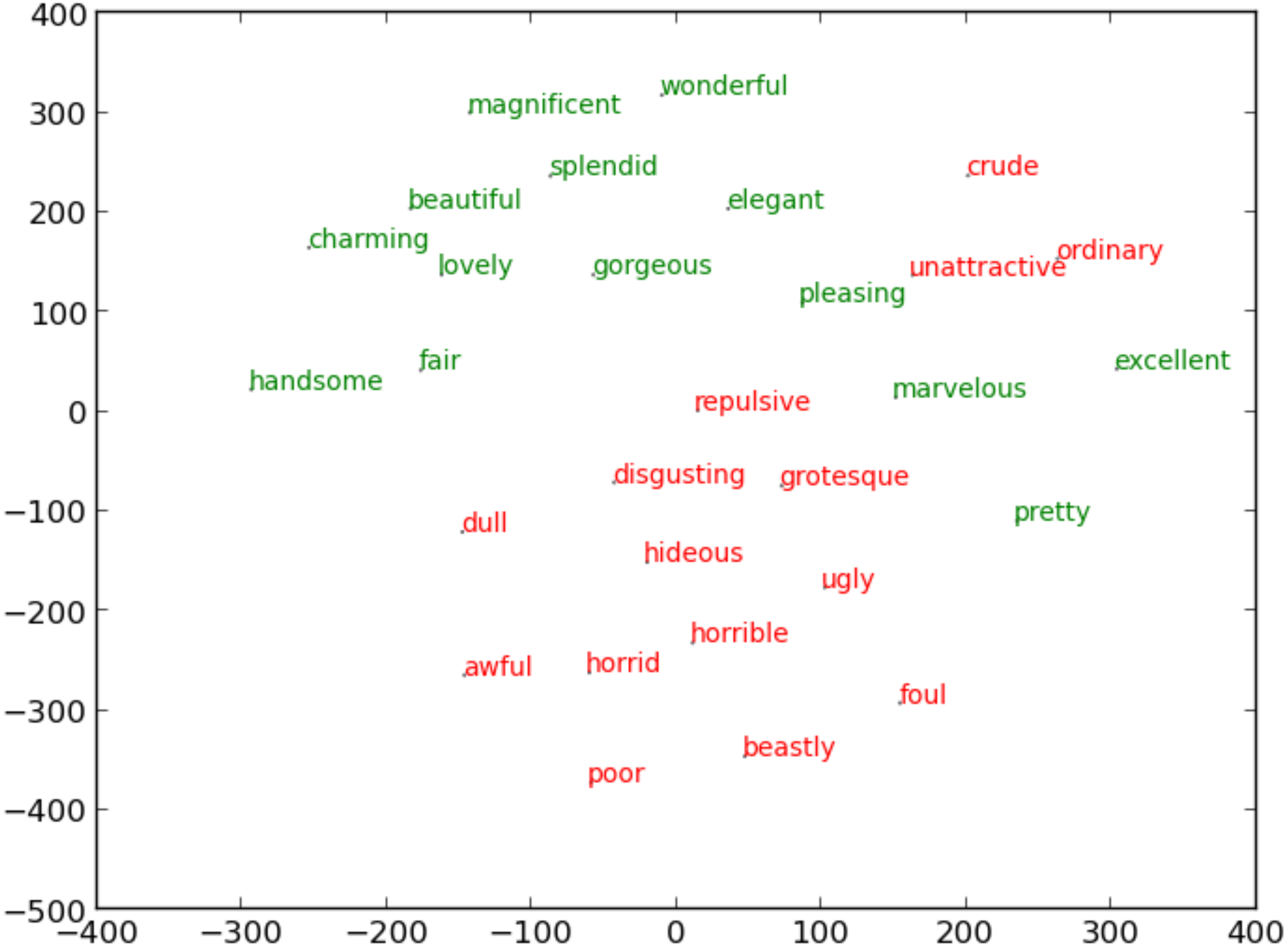
PEAR = <0,1,1,0,0,1>

Embeddings

Neural Word Representations

- Word embeddings are vectors

- The embedding of words with similar meaning are close in the vector space

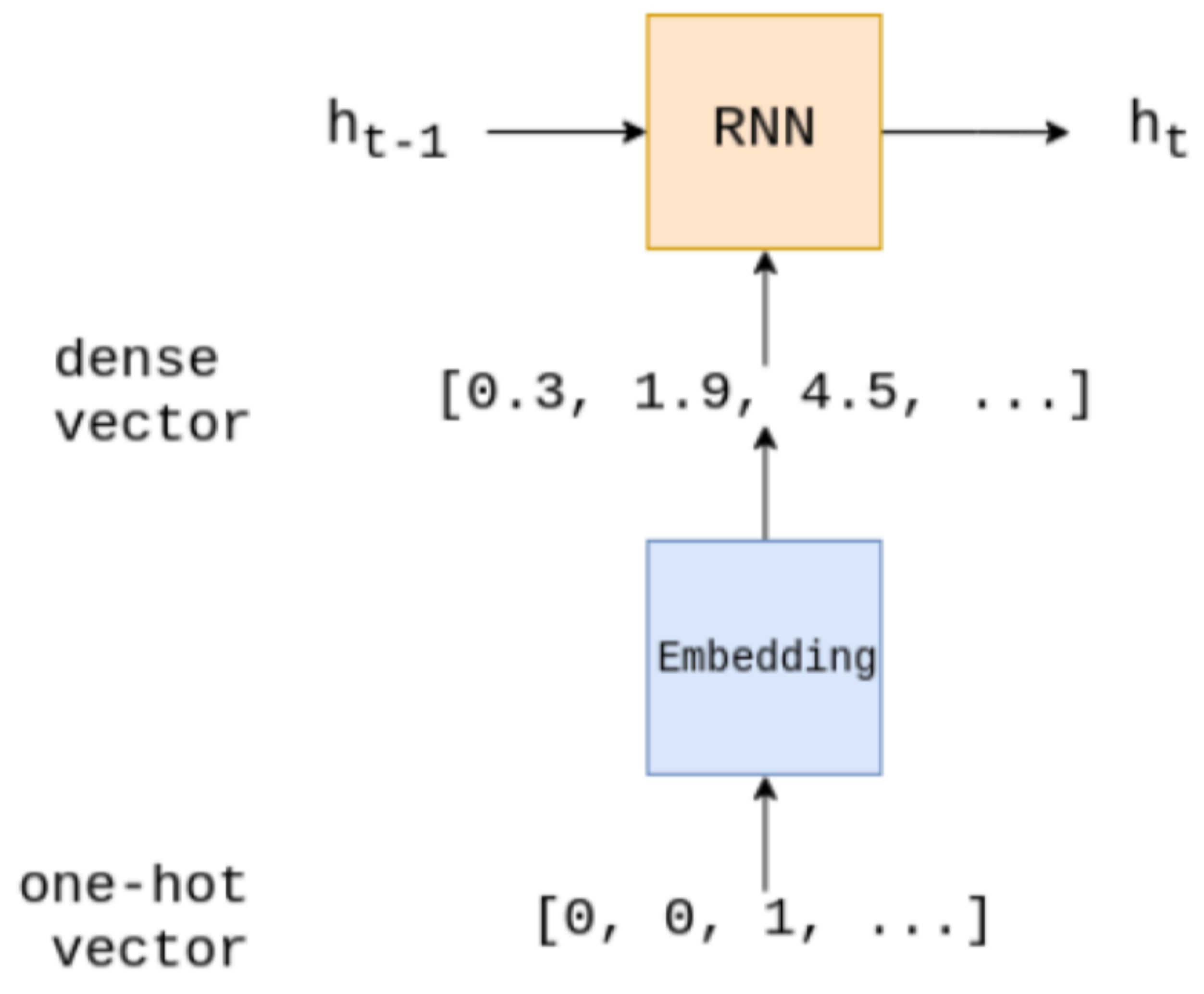


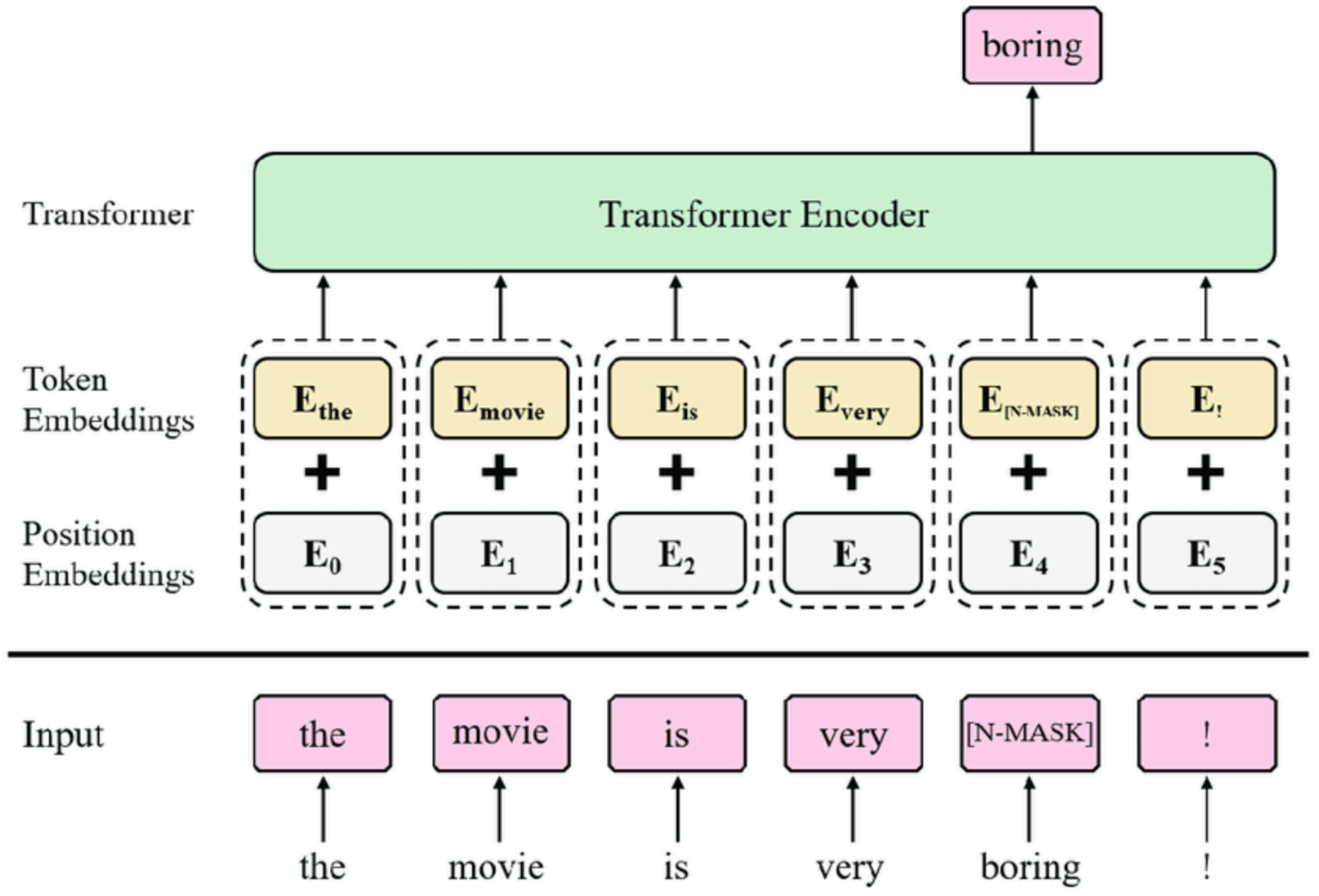
Word embeddings

Can be

- Learned during training

- Pretrained (Word2Vec, Glove, BERT, …)





Neural Language Models

Language Model

How probable is a sequence of words ?

Determines the probability of a sequence of words

$$P(W) = P(w_1, w_2, w_3 \ldots w_n)$$

$\hat{y}_t$

Ex (in english)

$P_1 = P(\text{"a quick brown dog"})$

$P_2 = P(\text{"dog quick a brown"})$

$P_3 = P(\text{"un chien quick brown"})$

$P_4 = P(\text{"un chien brun rapide"})$

- $P_1 > P_2 > P_3 > P_4$

Language Models

Language Models can be used to generate a sentence by auto-regressively predicting the next word given a previous context

France is where I grew up and where I now work. I speak fluent ***??***

p(***French*** | France is where I grew up and where I now work. I speak fluent)

$>$

p(*English* | France … fluent)

$>$

p(*Pizza* | France … fluent)

$>$

p(*the* | France … fluent)

Pre-Neural LM

Example

$P(\text{its water is so transparent}) =$
$P(\text{its}) \times P(\text{ water} \mid \text{its}) \times P(\text{ is} \mid \text{its water })$
$\times P(\text{so} \mid \text{its water is }) \times (\text{ transparent} \mid \text{its water is so})$

The probability of a sequence of words is computed using the chain Rule of Probability and Markov assumption

$$P(w_1, \ldots, w_n) = \prod_i P(w_i \,|\, w_{i-k}, \ldots, w_{i-1})$$

In Pre-neural model, probabilities are estimated on large corpora

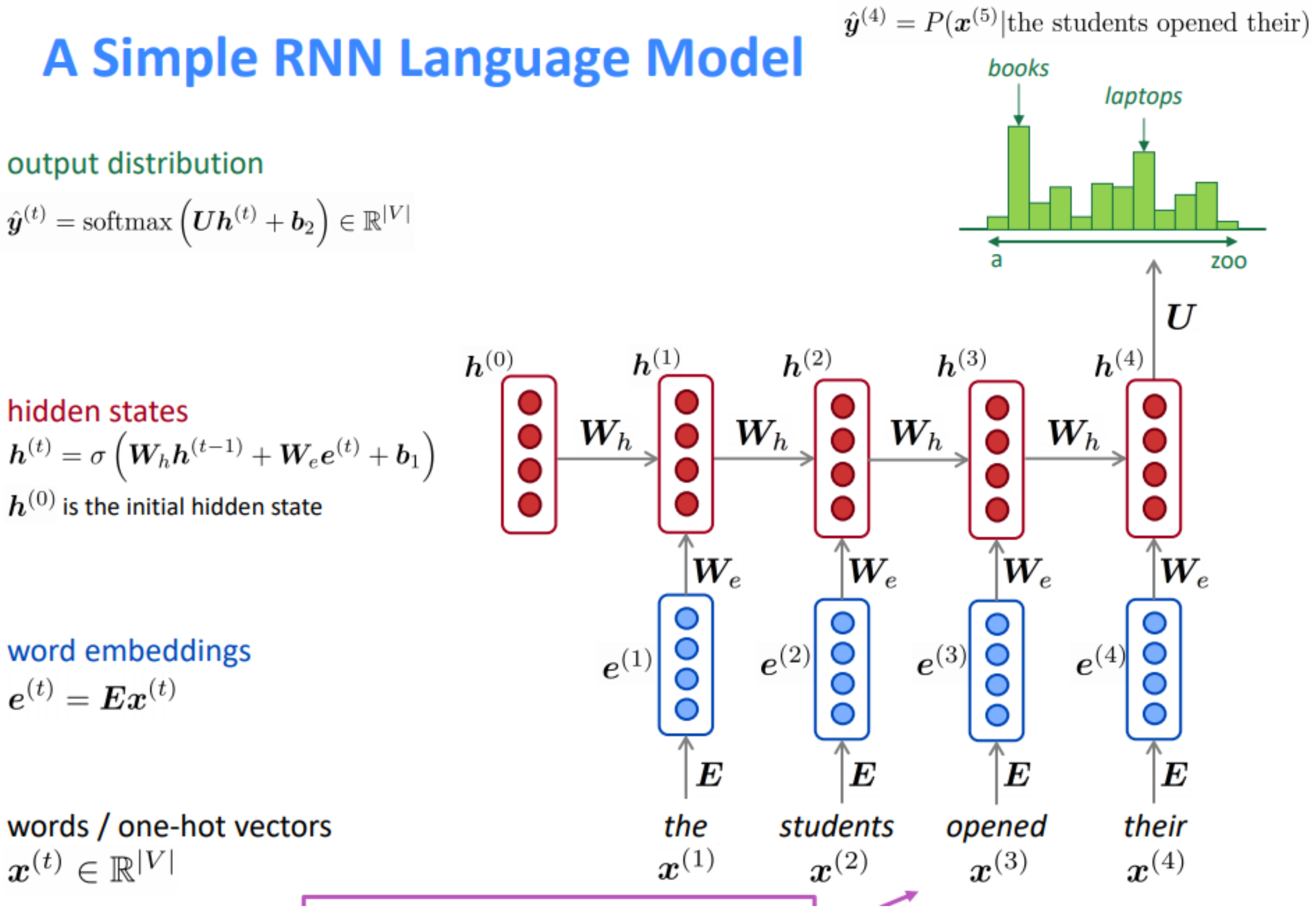
$$P(w_n \mid w_1, \ldots, w_{n-1}) = \frac{count(w_1, w_2, w_3 \ldots w_n)}{count(w_1, w_2, w_3 \ldots w_{n-1})}$$

Neural LMs

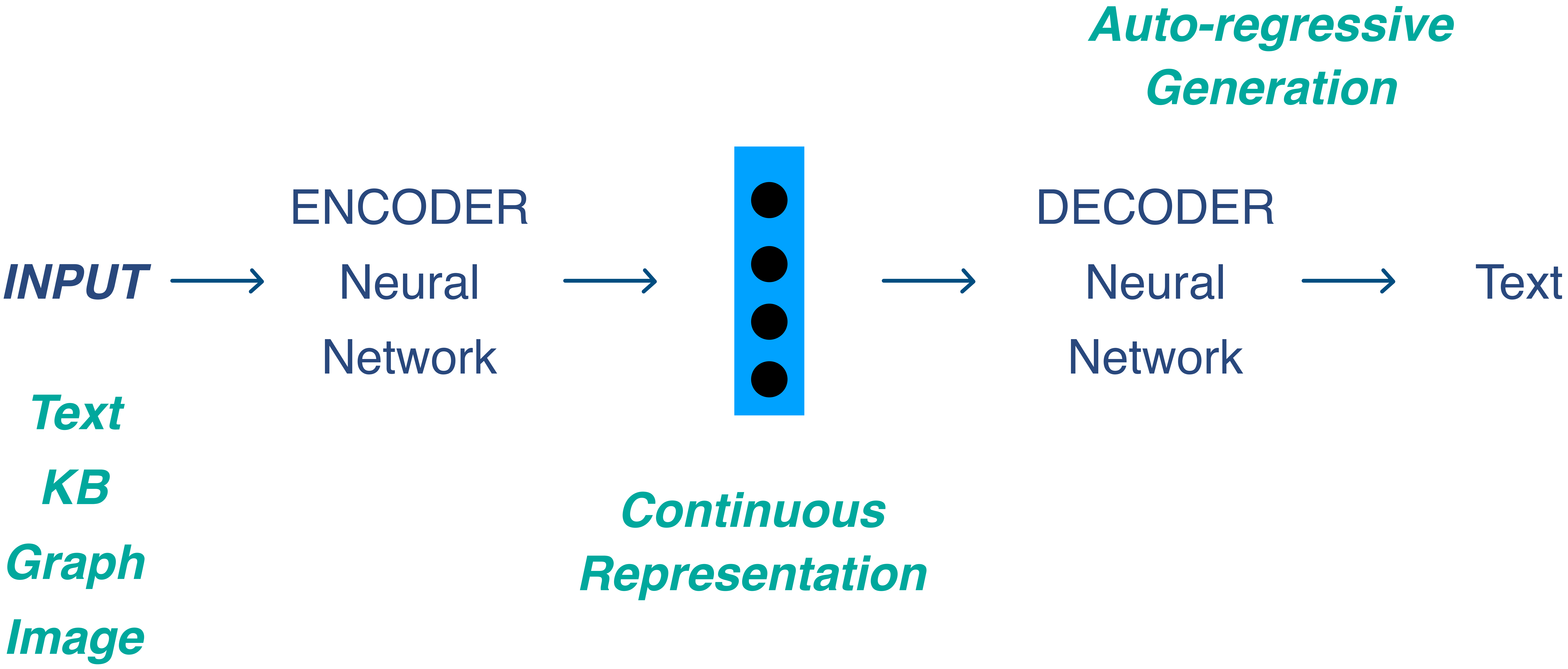
Using an RNN

- At each step, RNNs output a probability distribution over the vocabulary

- The next word is predicted using sampling (or beam search, top-k sampling, nucleus sampling, ranking)



The Encoder-Decoder Model



Encoders

- Recurrent Neural Network (sequences)
- Convolutional Neural Network (Images and Text)
- Graph Encoder (Knowledge Bases, Tabular Data, RDF store)
- Transformer (sequences)

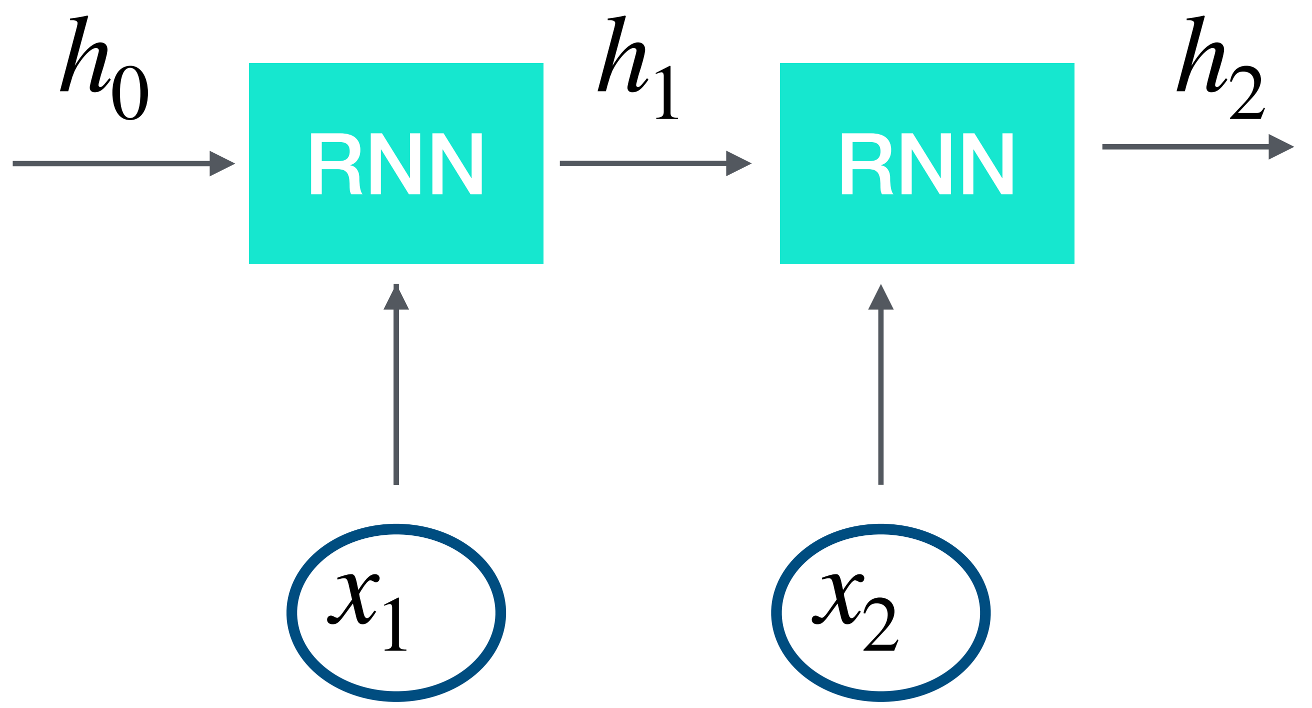
Decoders

- Recurrent Neural Network
- Transformer

Encoding and Decoding with a Recurrent Neural Network (RNN)

Encoding the Input with a Recurrent Neural Network

- Processes sequences from left-to-right



- Recurs over the input

- Outputs a new hidden state at each step

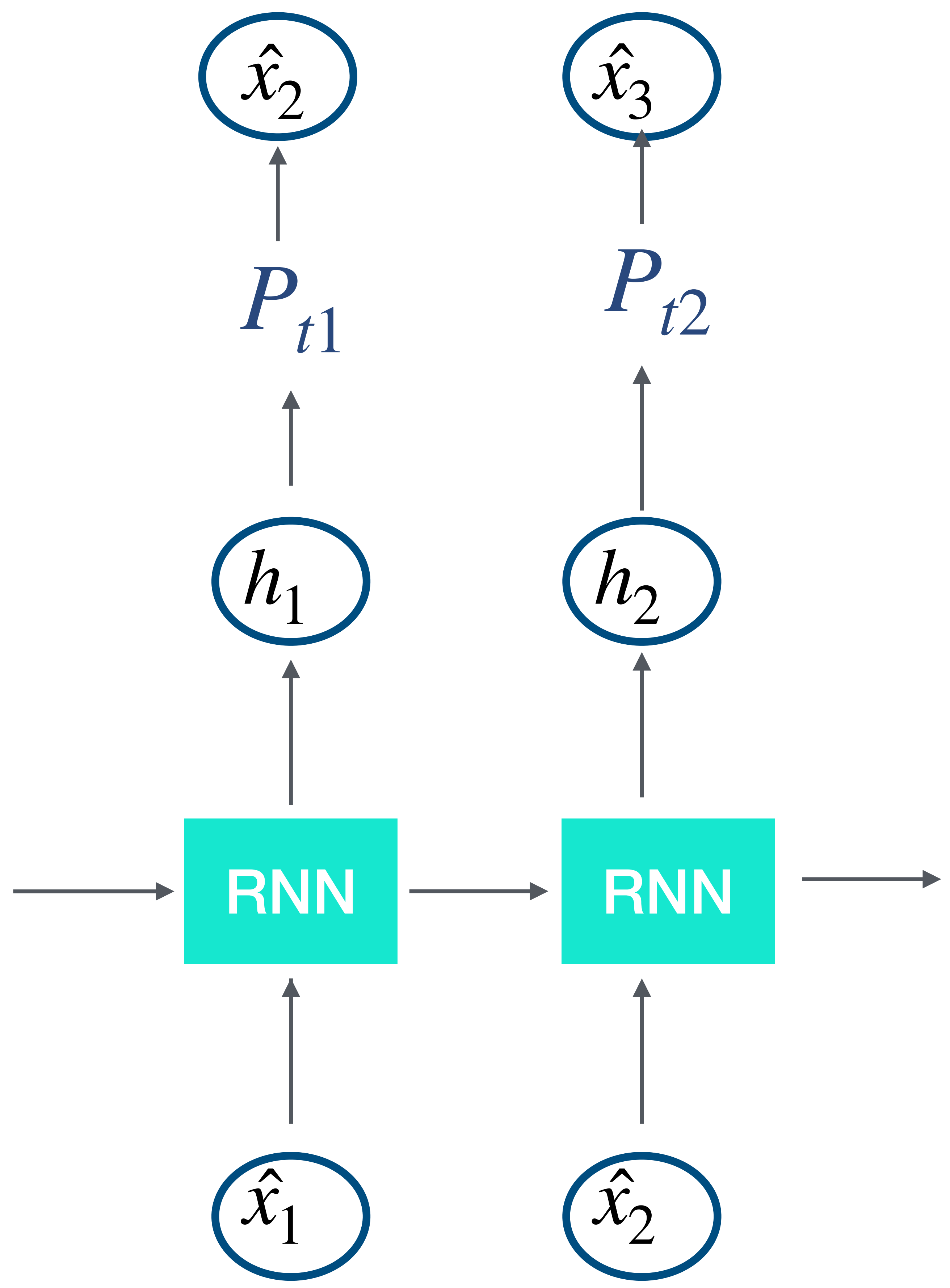
$$h_t = tanh(W_1 h_{t-1} + W_2 x_t)$$

The last hidden state is the input representation

Decoding with a Recurrent NN

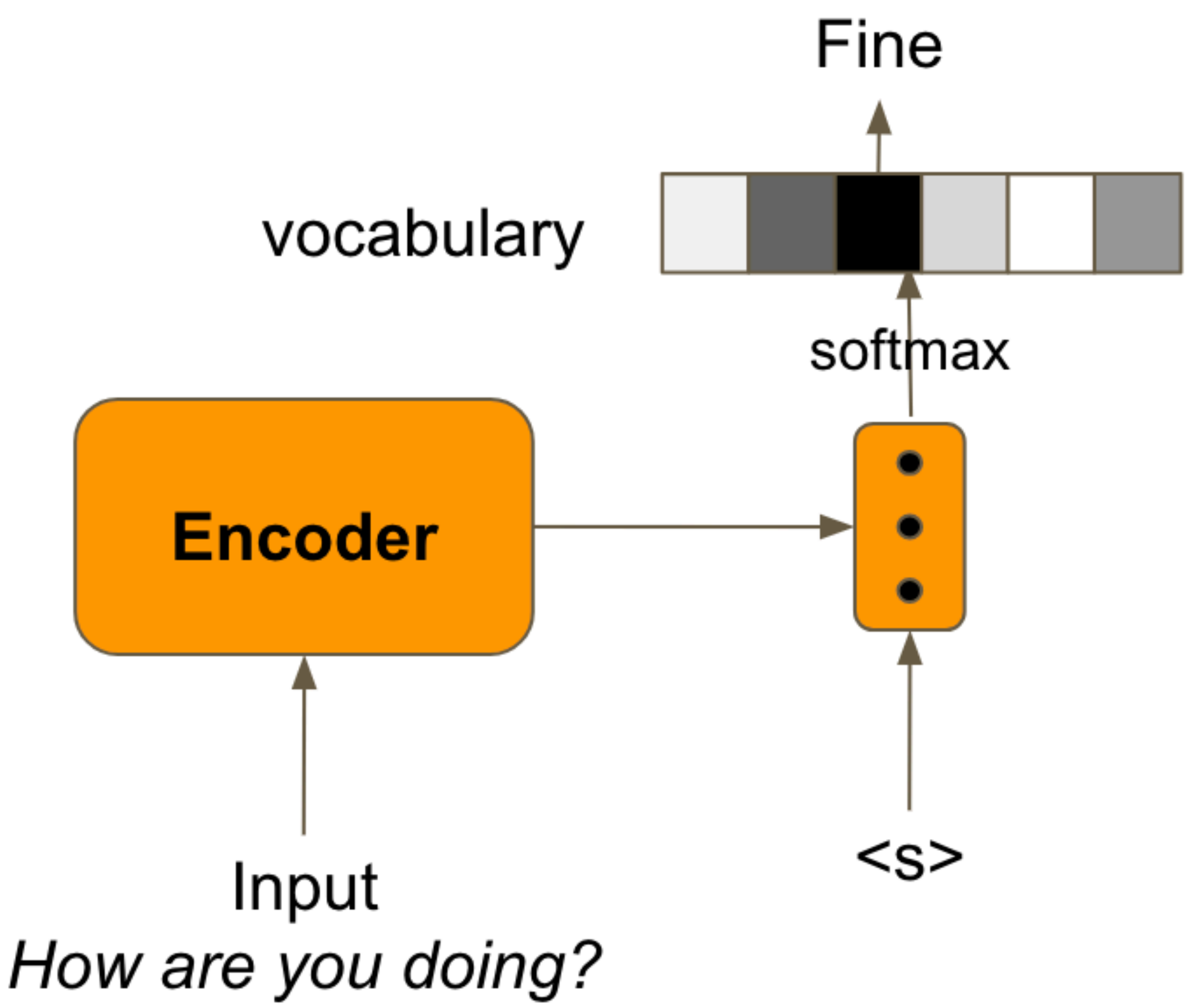
Outputs a word at each step

- Use a softmax layer to compute a ***probability distribution over the output vocabulary***
- Sample a word from this distribution
- The predicted word is the input to the next decoding step



$$P_t = softmax(W_t h_t)$$

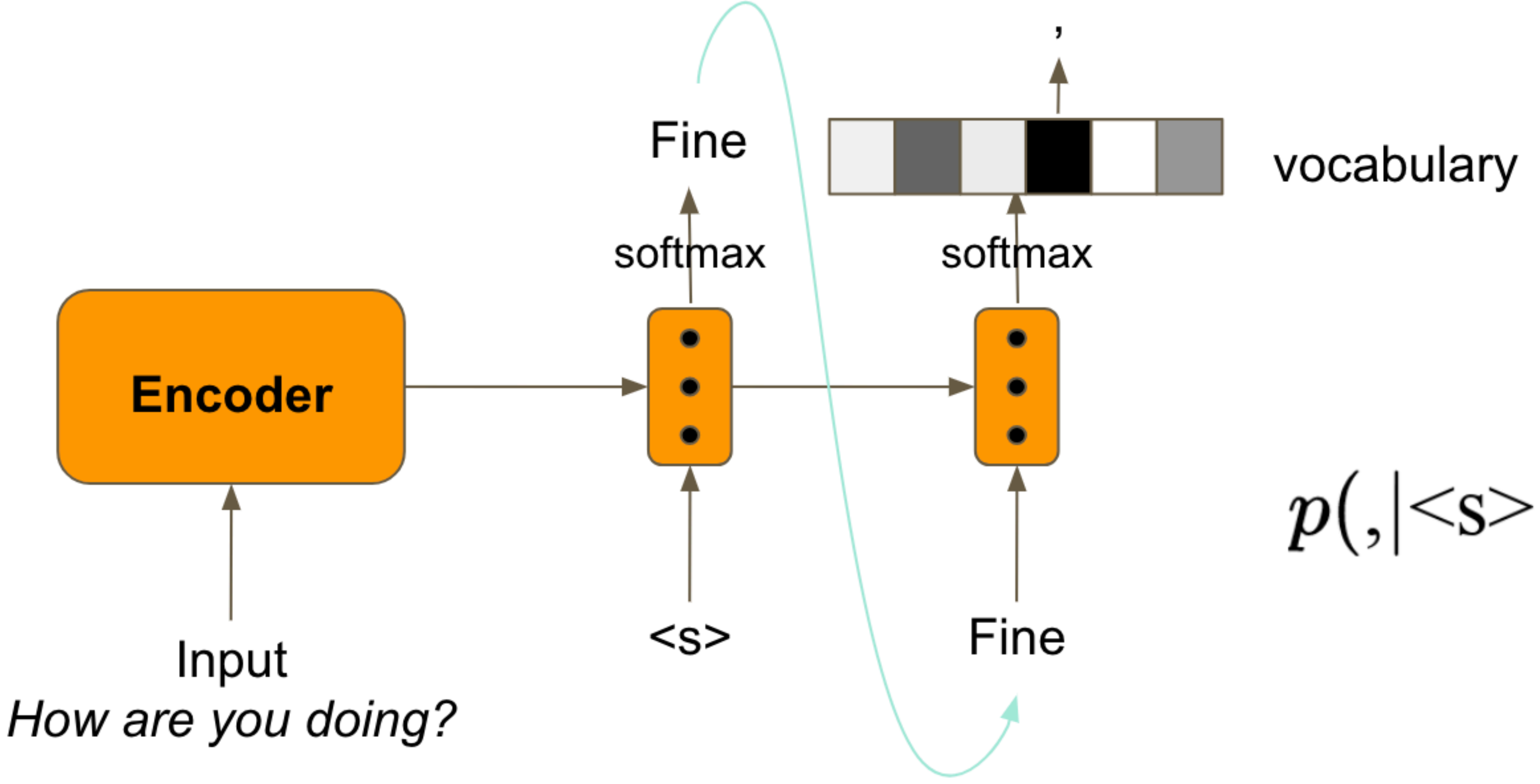
Generating with a Recurrent NN



$$p(\text{Fine}|\text{<s>}, \text{How are you doing?})$$

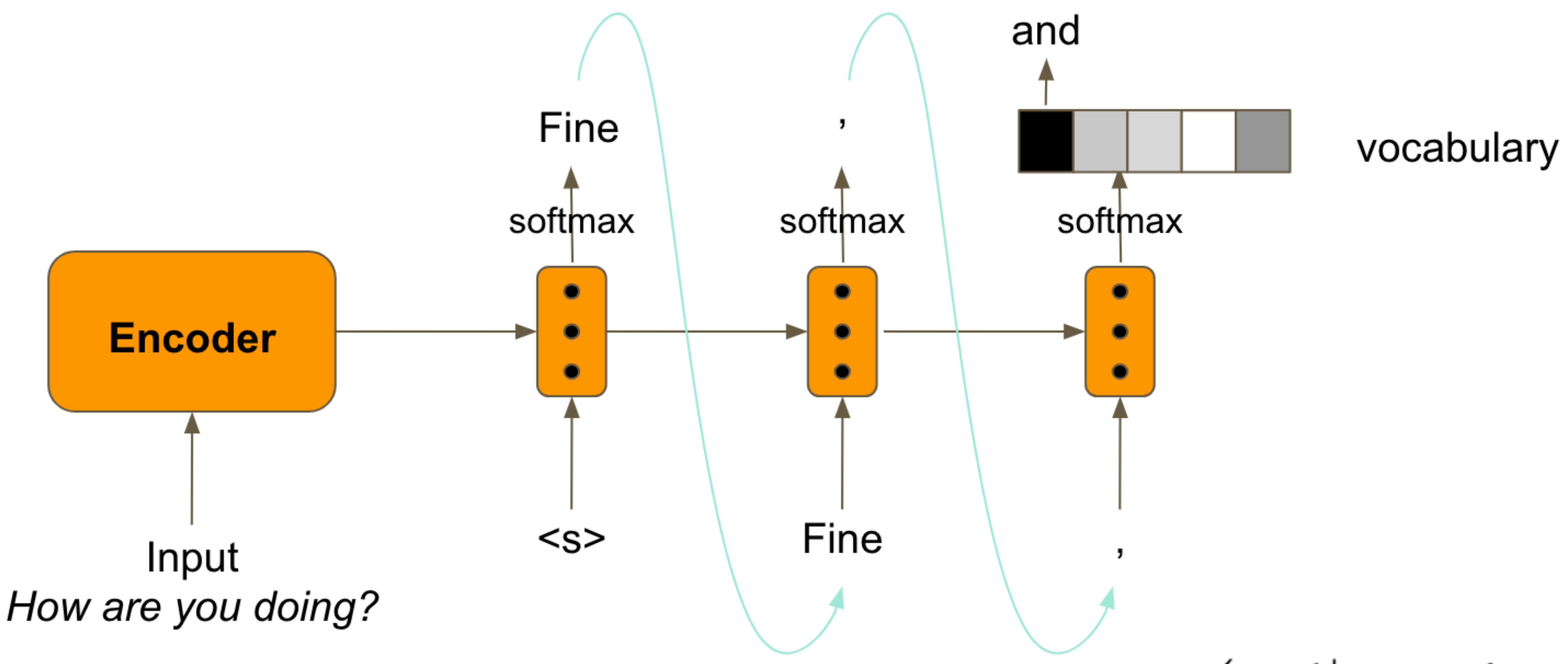
Conditional Generation

Generating with a Recurrent NN



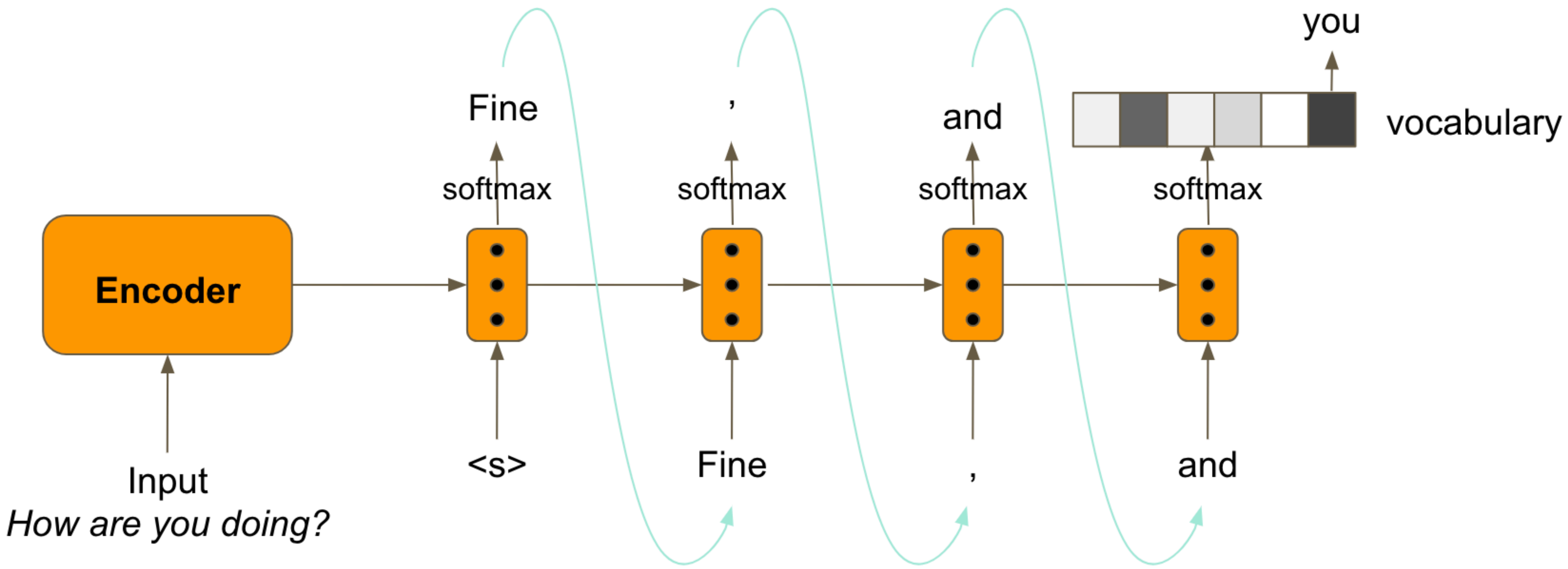
$$p(,|\text{<s> Fine, How are you doing?})$$

Generating with a Recurrent NN

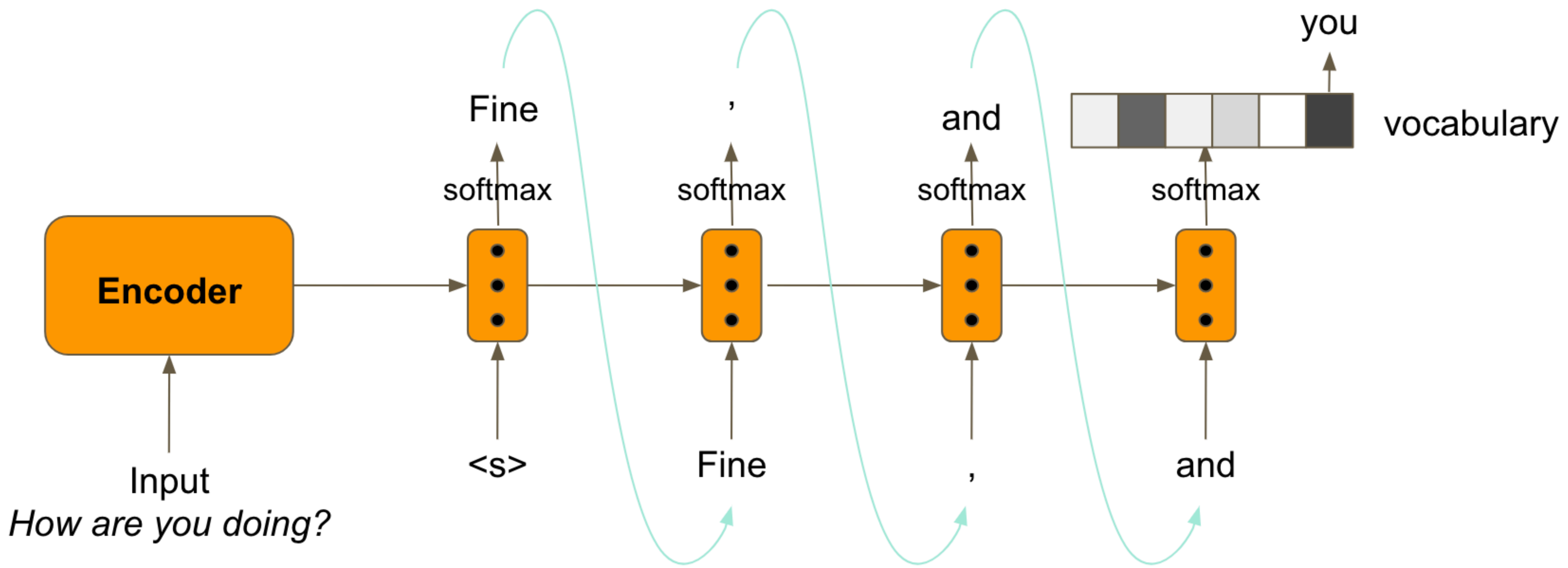


$$p(\text{and} \,|\, \text{<s> Fine,; How are you doing?})$$

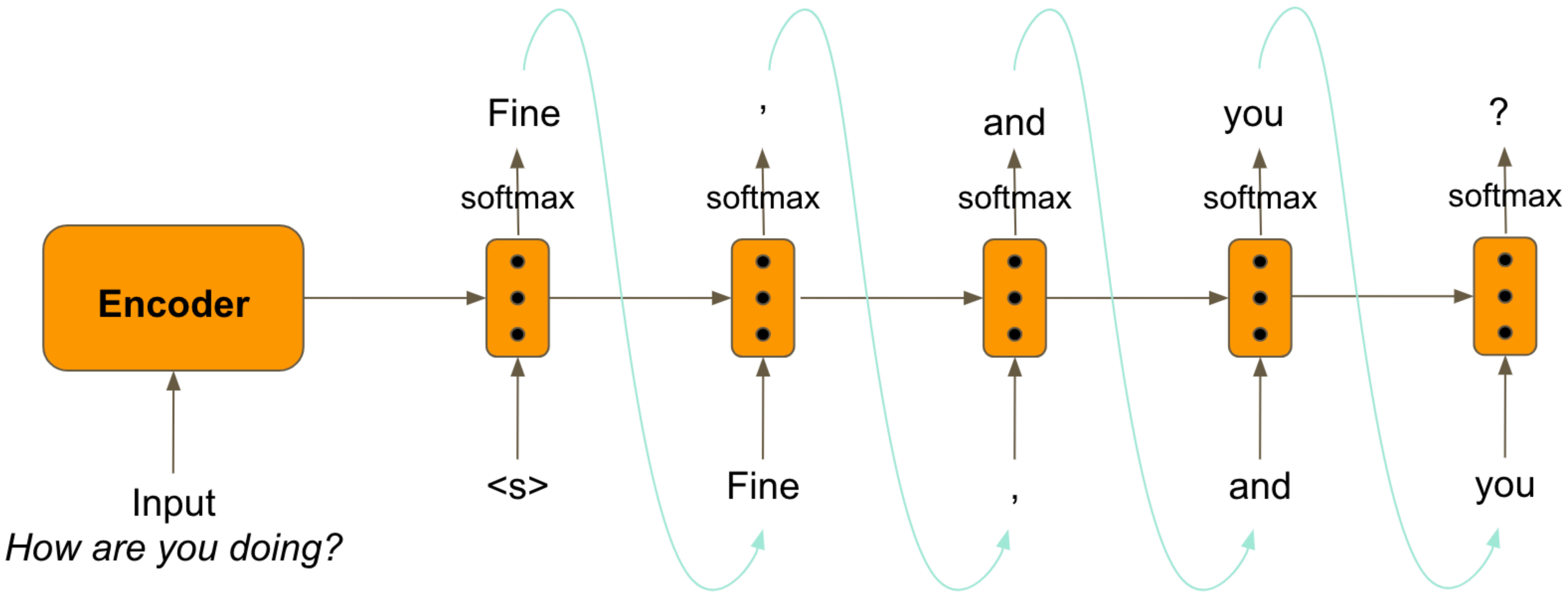
Generating with a Recurrent NN



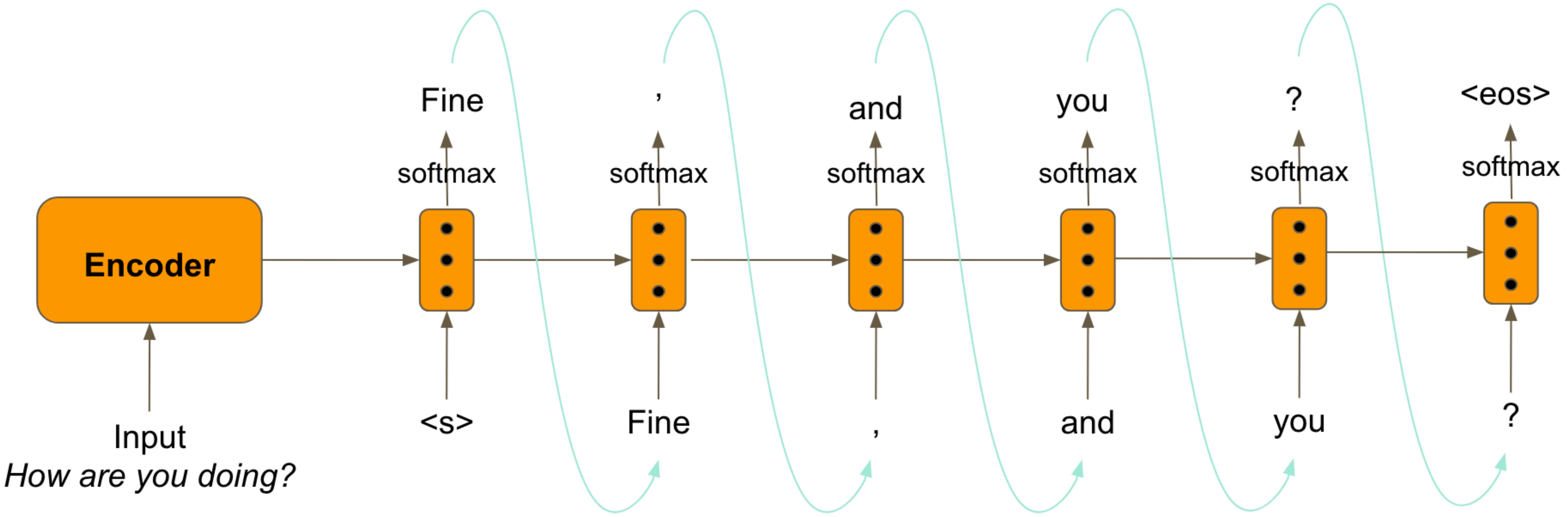
Generating with a Recurrent NN



Generating with a Recurrent NN



Generating with a Recurrent NN



Shortcomings and Solutions

- In practice, RNNs cannot handle long context
  Exploding and Vanishing Gradients
  >> GRU, LSTM

- RNNs only know about the left context
  >> Bi-directional RNN

Shortcomings and Solutions

- The input to decoding is a fixed size vector

>> Attention

permits focusing on the relevant part of the input

- Because they process a sequence one word at a time, RNNs are slow to train

>> Convolutional Neural Network, Transformer

Process each input item in parallel

Shortcomings and Solutions

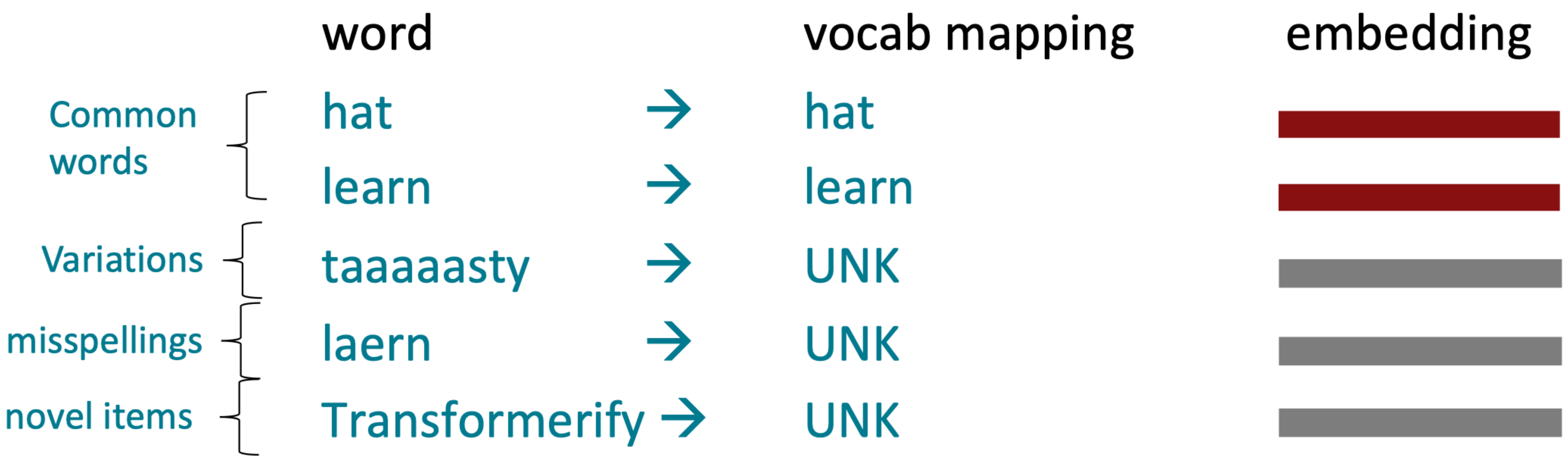
- Word unseen at training time represented as UNK at test time

  >> BPE (Byte Pair Encoding), WordPieces

  Uses subwords rather than words

Words and Subwords

- Neural Language Models assume a finite vocabulary
- All words unseen at training time are mapped to UNK
- For language with a rich morphology (Georgian, Swahili etc.), this is even more problematic



Words and Subwords

- Instead of handling words, subword models learn a vocabulary of subwords
- At training and test time, each word is split into a sequence of subwords
- **Byte-Pair encoding (BPE)** is one way to define a subword vocabulary
  - Start with characters
  - Add most frequent n-grams as subword
  - Iterate until desired vocabulary size is reached

BPE Example

Dictionary

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

Vocabulary

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

Start with all characters in vocab

Dictionary

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

Vocabulary

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

Add a pair (e, s) with freq 9

BPE Example

Dictionary

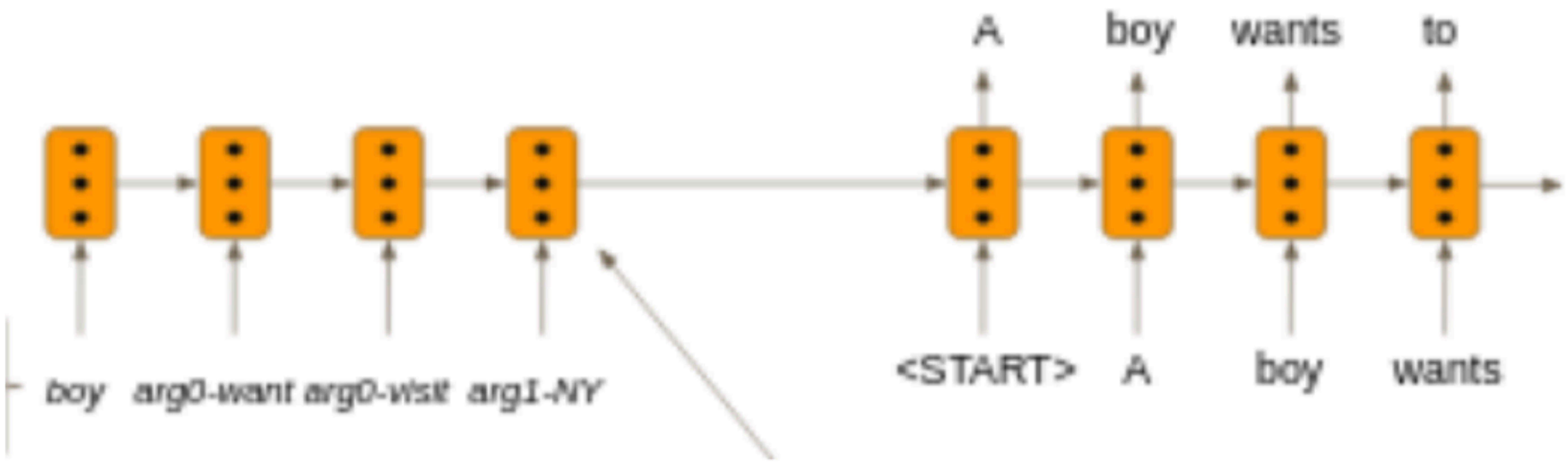
5 l o w
2 l o w e r
6 n e w **est**
3 w i d **est**

Vocabulary

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

Add a pair (es, t) with freq 9

Attention



- The input is compressed into a fixed-length vector
- Performance decreases with the length of the input

Attention

- Computes a score between each input token encoder state and the current state

$$a_{t,j} = score(s_t, h_j)$$

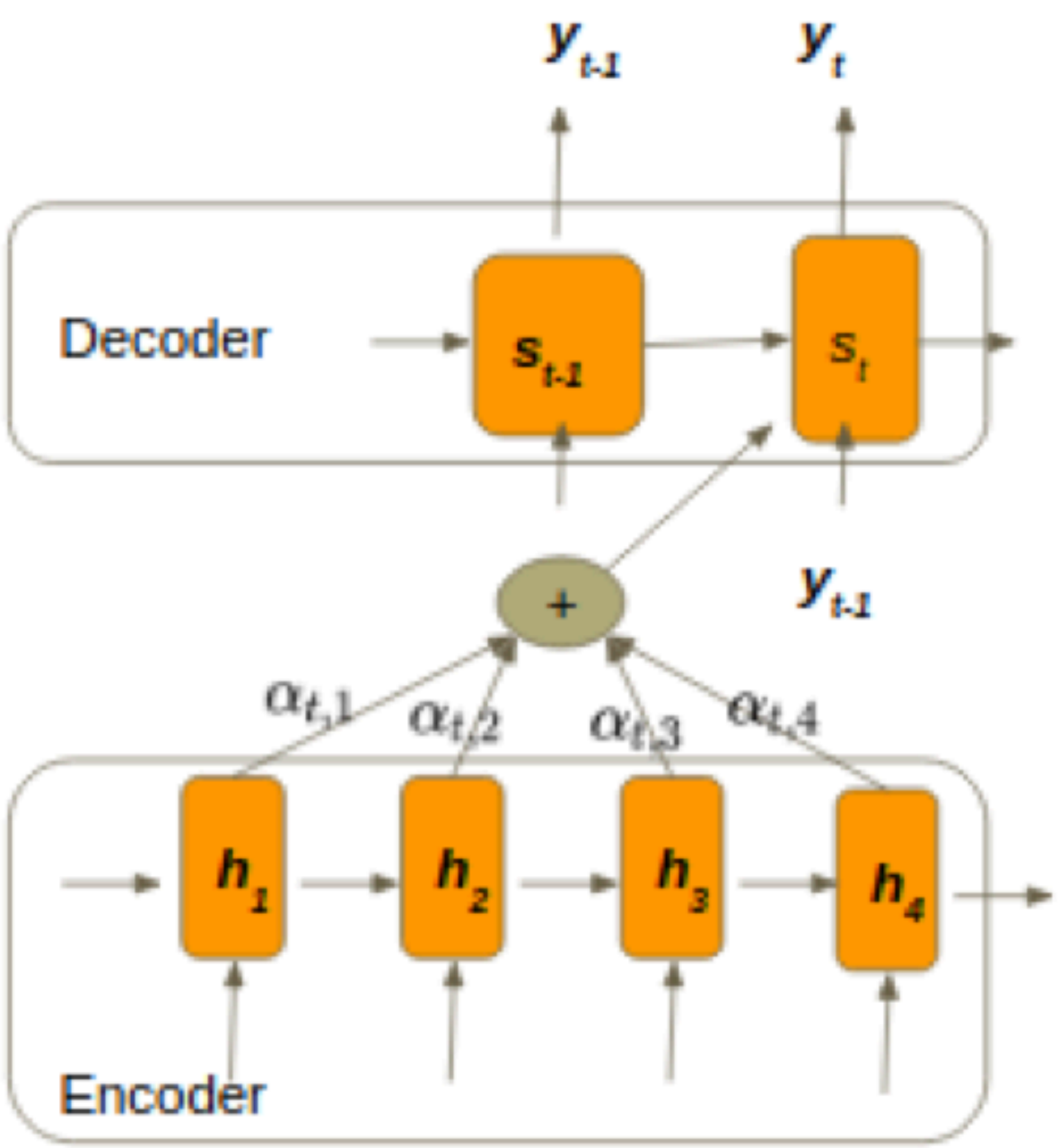
$$\alpha_t = softmax(a_t)$$

- The context vector is the weighted sum of the encoder states.

$$c_t = softmax(\sum_j \alpha_{t,j} \, . \, h_j)$$

- The new state is computed taking into account this context vector.

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$



$$\alpha_t = softmax(a_t)$$

Attention

Attention is a way to obtain a fixed-size representation of an arbitrary set of representations (the values) dependent on some other representation (the query)

Encoder-Decoder Cross-Attention

- Query = decoder state
- Values = encoder hidden states

Transformer Self-Attention

- Query = token embedding
- Values = surrounding tokens embeddings

Pretraining *(Self-Supervised Learning)*

Word Embedding, Encoder

- Word2Vec, Glove, BERT …

Language Models

- GPT2, DialoGPT, LAMDA …

Encoder-Decoder

- T5, BART ….

BERT

Transformer encoder pretrained on BooksCorpus (800M words) and English Wikipedia (2,500M words)

Two loss functions

- Predict masked tokens (Masked Language Modelling)
- Next sentence prediction classification (true if next sentence is the correct continuation)

The training loss is the sum of the mean masked LM likelihood and mean next sentence prediction likelihood

BERT Masked Language Modeling

the   of

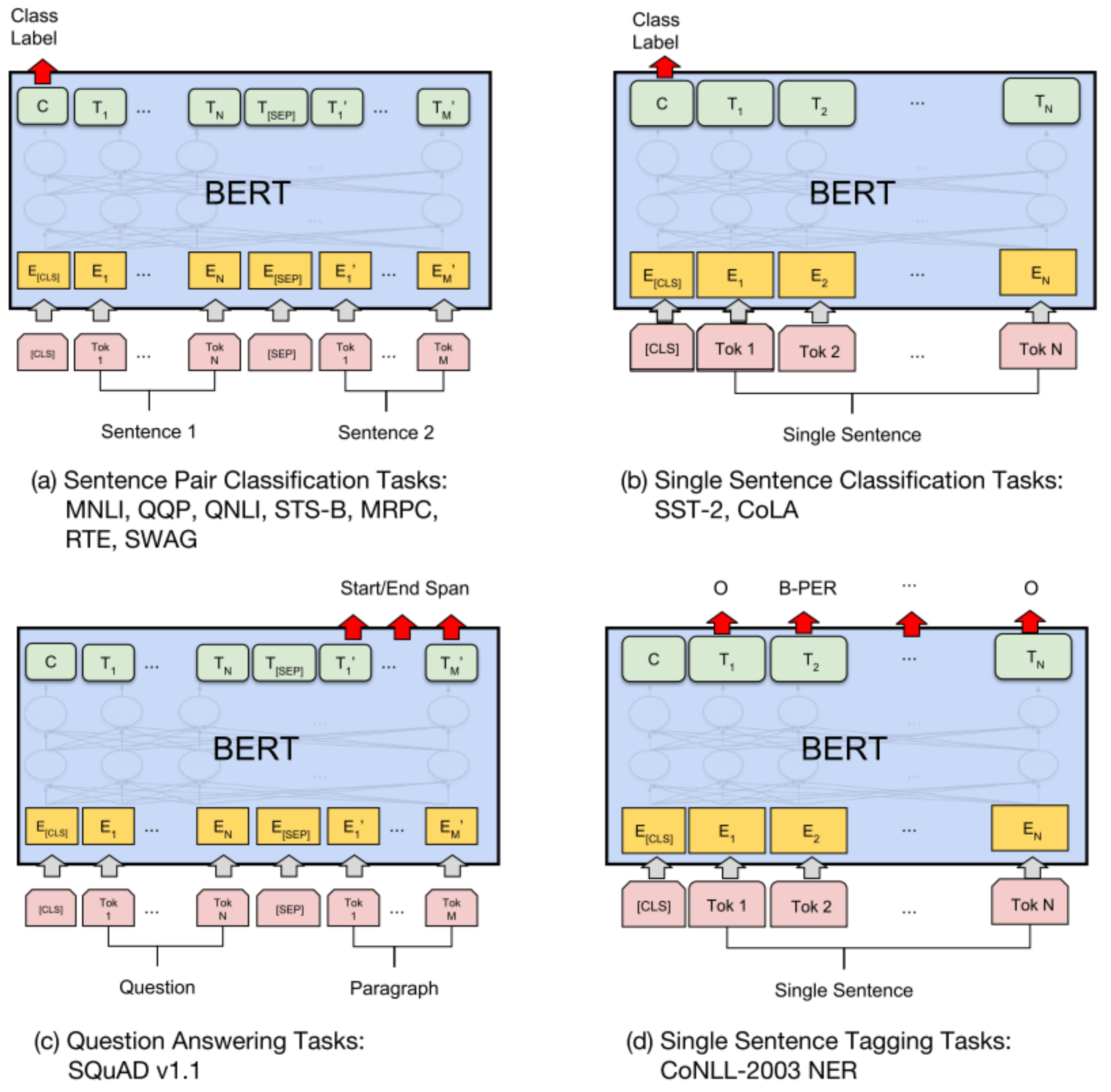
Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

Mask 1 word in 7

- Too litle masking: too expensive to train
- Too much masking: not enough context

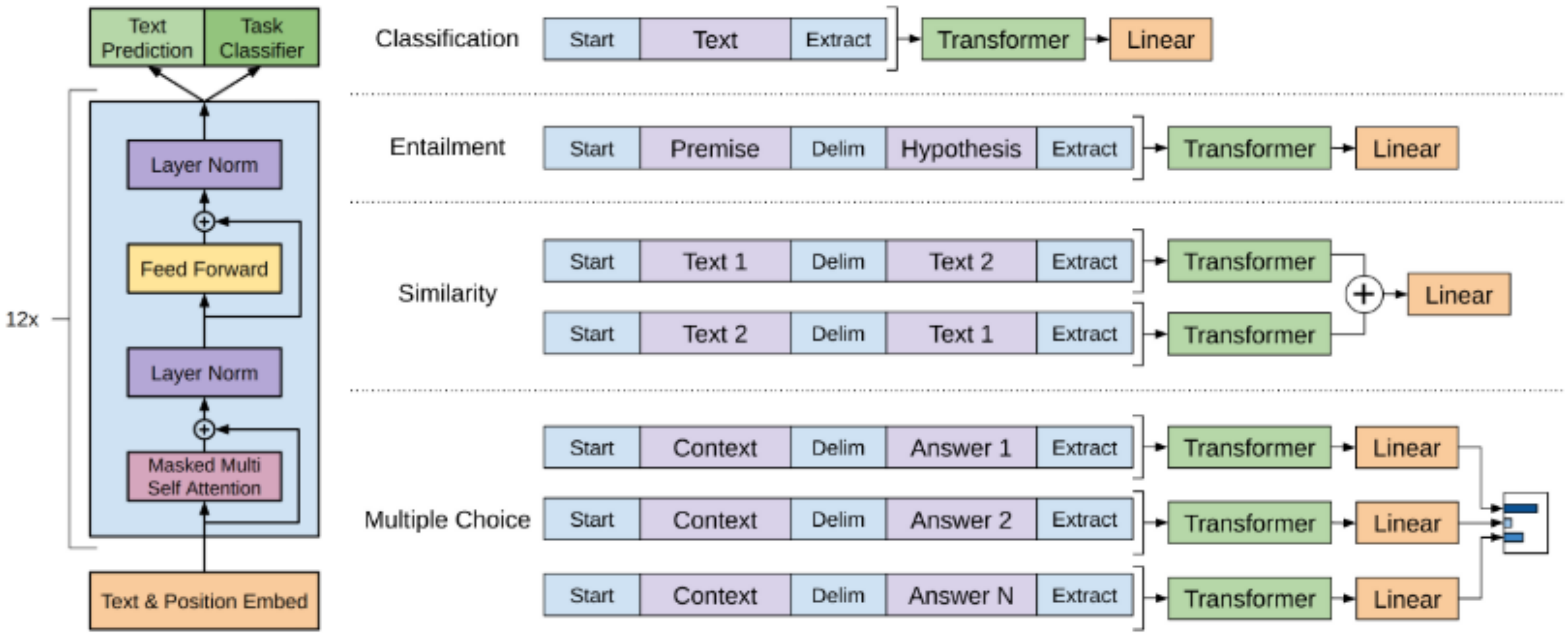
BERT Fine Tuning

- Sentence representation = final hidden state output by the Transformer (= [CLS] word embedding)
- Add a classification layer
- Fine tune all BERT parameters and the classification layer jointly to maximize the log-probability of the correct label

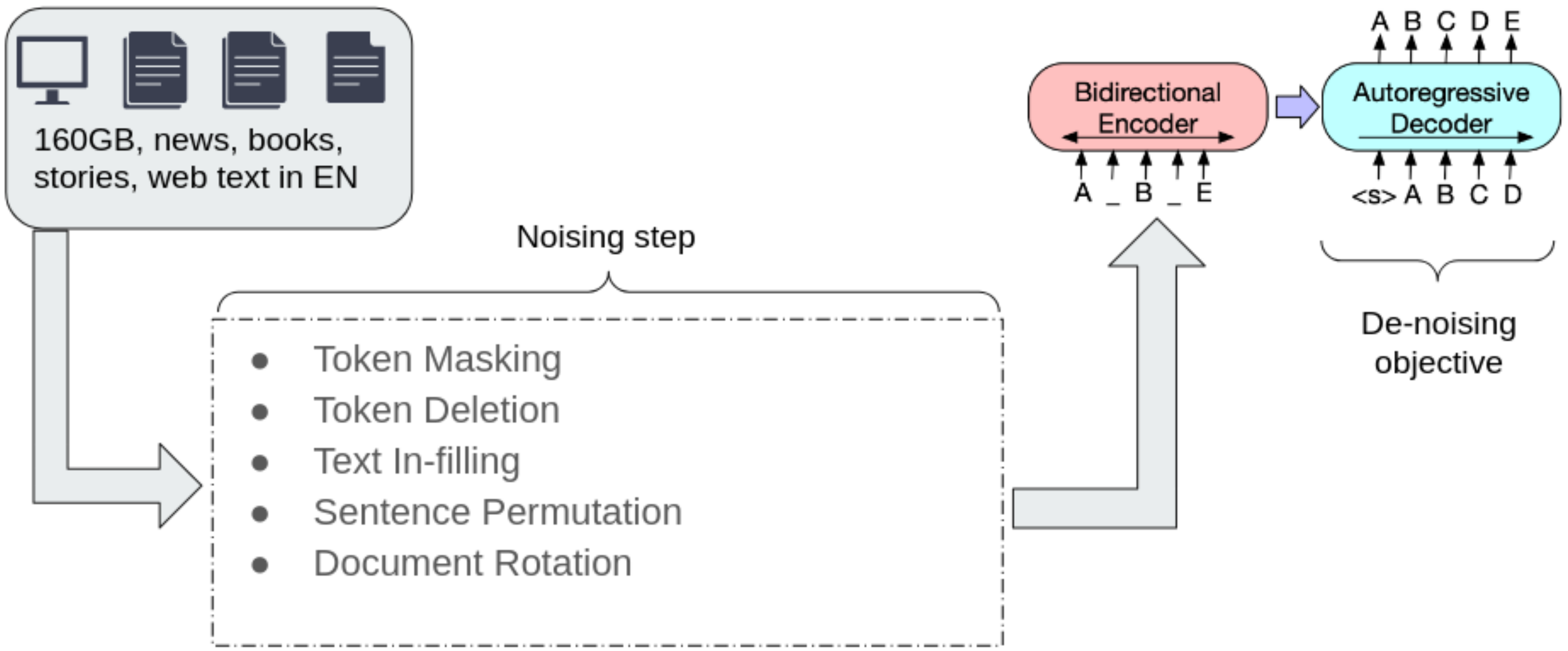


GPT2

- Unsupervised Pre-training
  Train a LM on a large corpus of text (BookCorpus 7K books)
- Supervised Fine-tuning
  - Input passed through pre-trained LM
  - Feed final LM activation to added linear + softmax output layers to predict output
  - Task -aware input transformations
- Significantly improves upon the SOTA in 9 out of 12 NLU tasks



BART



- Transformer Encoder-Decoder
- Denoising Auto-Encoder
  - Corrupt text with a noising function
  - Model learns to reconstruct original text
- Experiment with different noising functions

BART

Achieves new state-of-the-art results on a number of text generation tasks

Text infilling (reconstruct spans) demonstrates the most consistently strong performance

Token masking (reconstruct missing tokens) is crucial

Document Rotation and Sentence Shuffling perform poorly in isolation

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	**89.0**/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

Table 2: Results for large models on SQuAD and GLUE tasks. BART performs comparably to RoBERTa and XLNet, suggesting that BART's uni-directional decoder layers do not reduce performance on discriminative tasks.

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

Table 3: Results on two standard summarization datasets. BART outperforms previous work on summarization on two tasks and all metrics, with gains of roughly 6 points on the more abstractive dataset.

Four Challenges for Neural Generation

Challenges for Neural NLG

- Generating from long Input

Challenges for Neural NLG

- Generating from Dealing long Input
- Retrieving and Integrating Relevant Knowledge

Challenges for Neural NLG

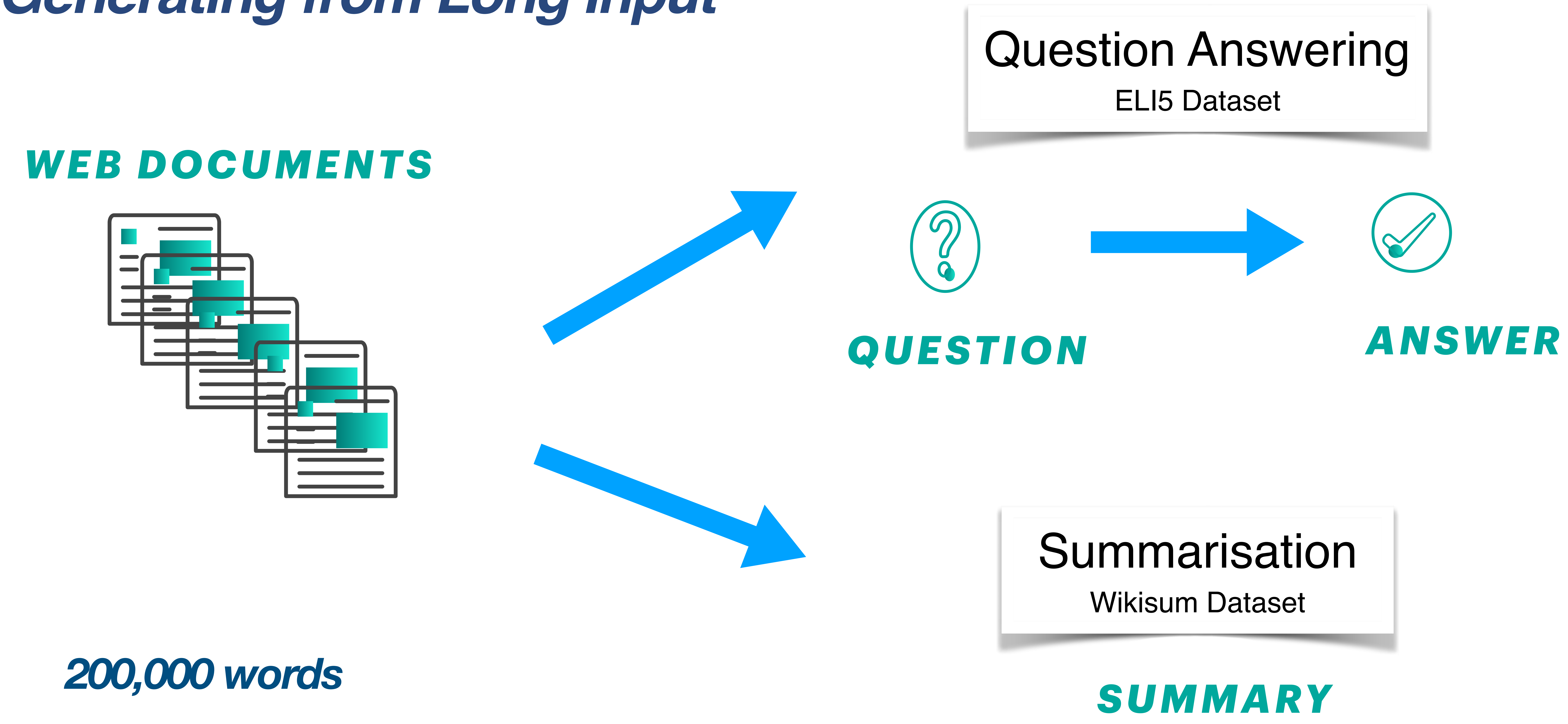
- Generating from long Input
- Retrieving and Integrating Relevant Knowledge
- Generating into Languages other than English

Challenges for Neural NLG

- Generating from Long Input
- Retrieving and Integrating Relevant Knowledge
- Generating into Languages other than English
- Generating Long Form Text

Handling Long Input

Generating from Long Input

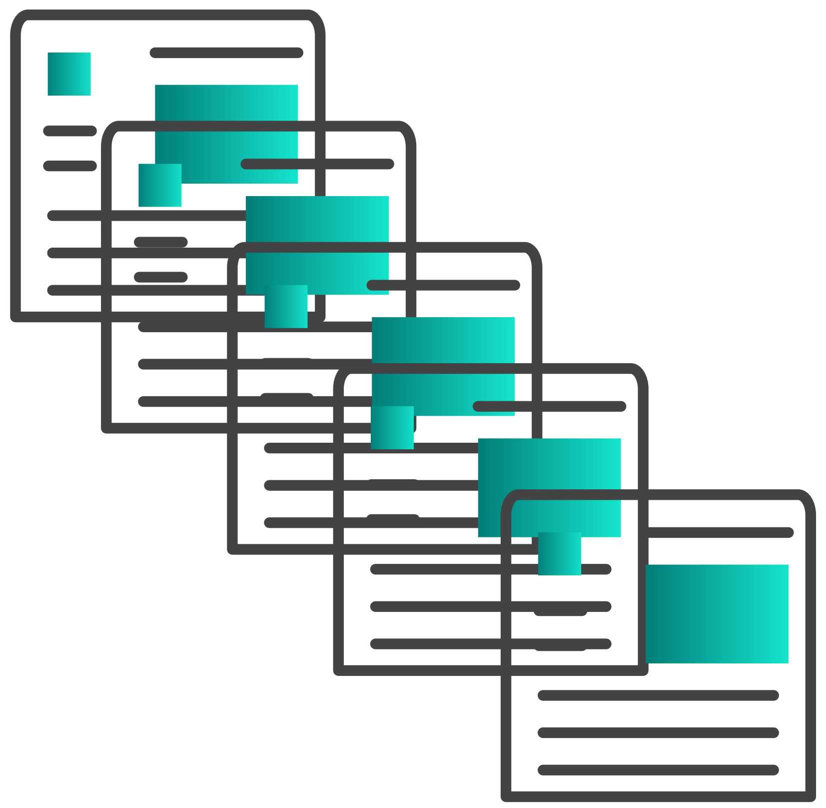


Question Answering

Explain Like I'm Five Dataset

270,000 TRAINING INSTANCES

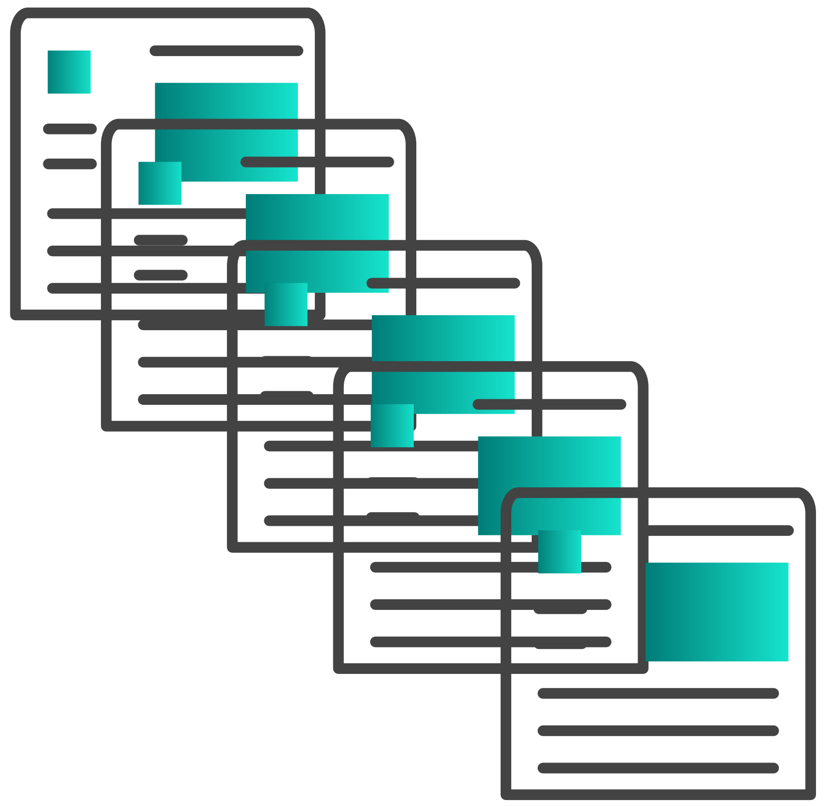
QUESTION + **WEB DOCUMENTS** + **ANSWER**



200,000 words

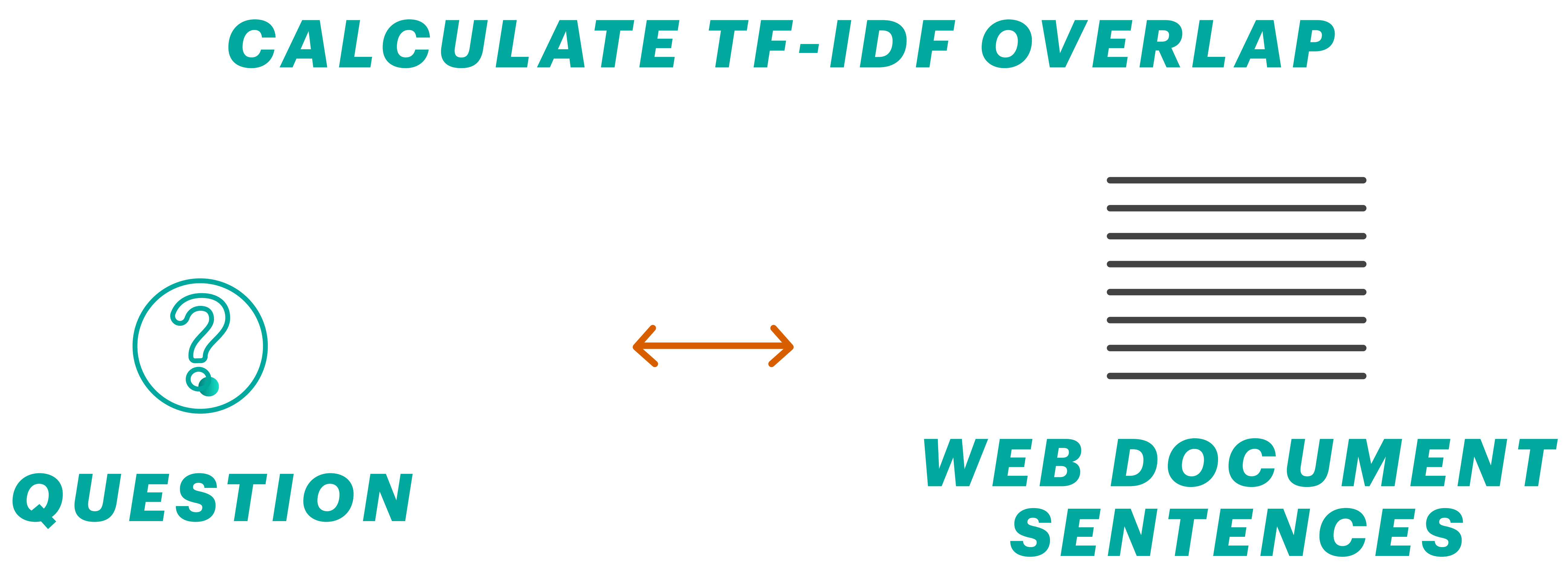
Dealing with Long Web Input

WEB DOCUMENTS

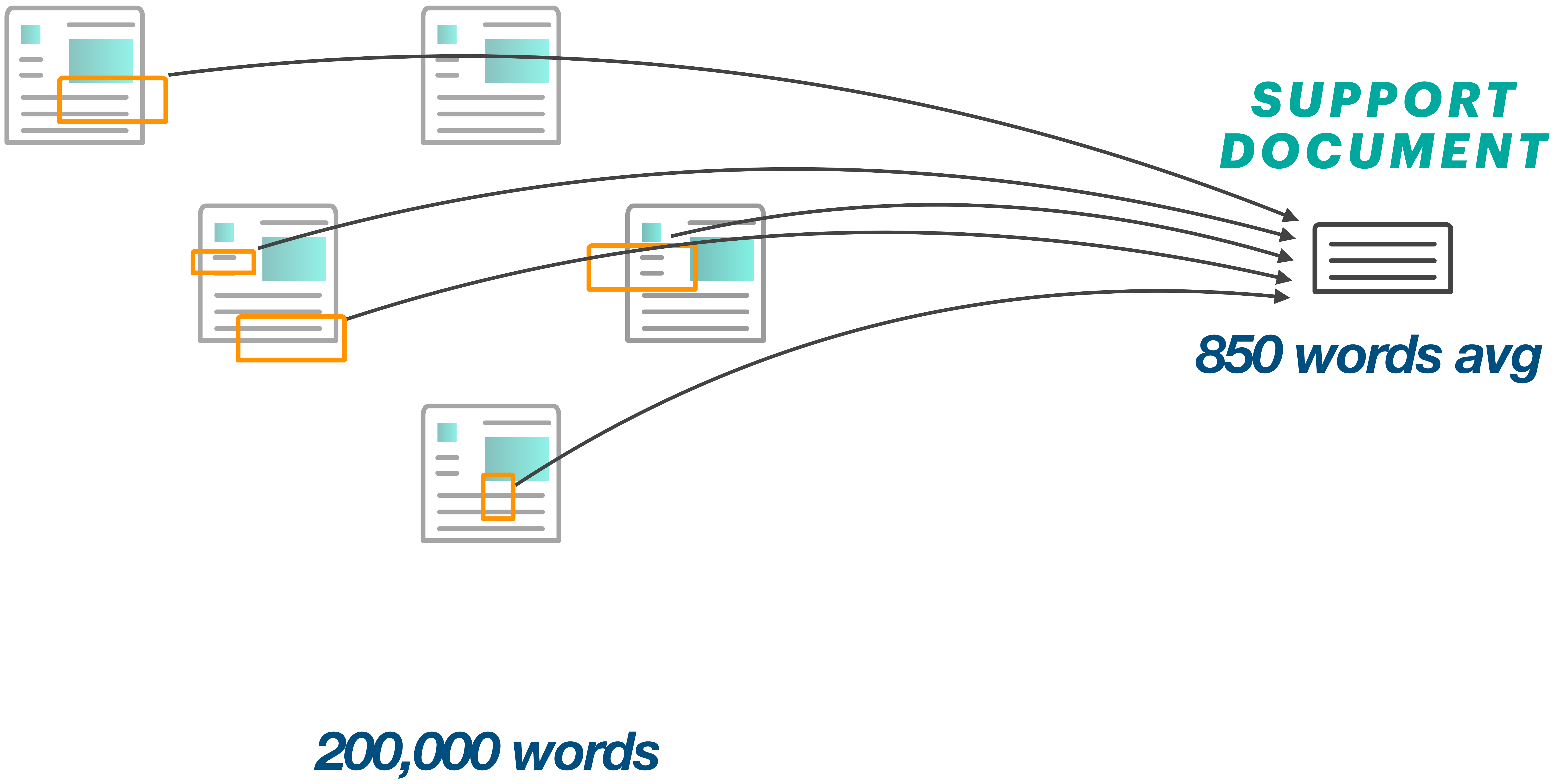


Over 200,000 words long

Creating a Shorter Support Document



Creating a Shorter Support Document



Downsides of Short Support Document

SUPPORT DOCUMENT

850 words avg

40% of the Answer Tokens are Missing

Downsides of Short Support Document

SUPPORT DOCUMENT

850 words avg

40% of the Answer Tokens are Missing

Information selected is Redundant

Downsides of Extractive Support Document

SUPPORT DOCUMENT

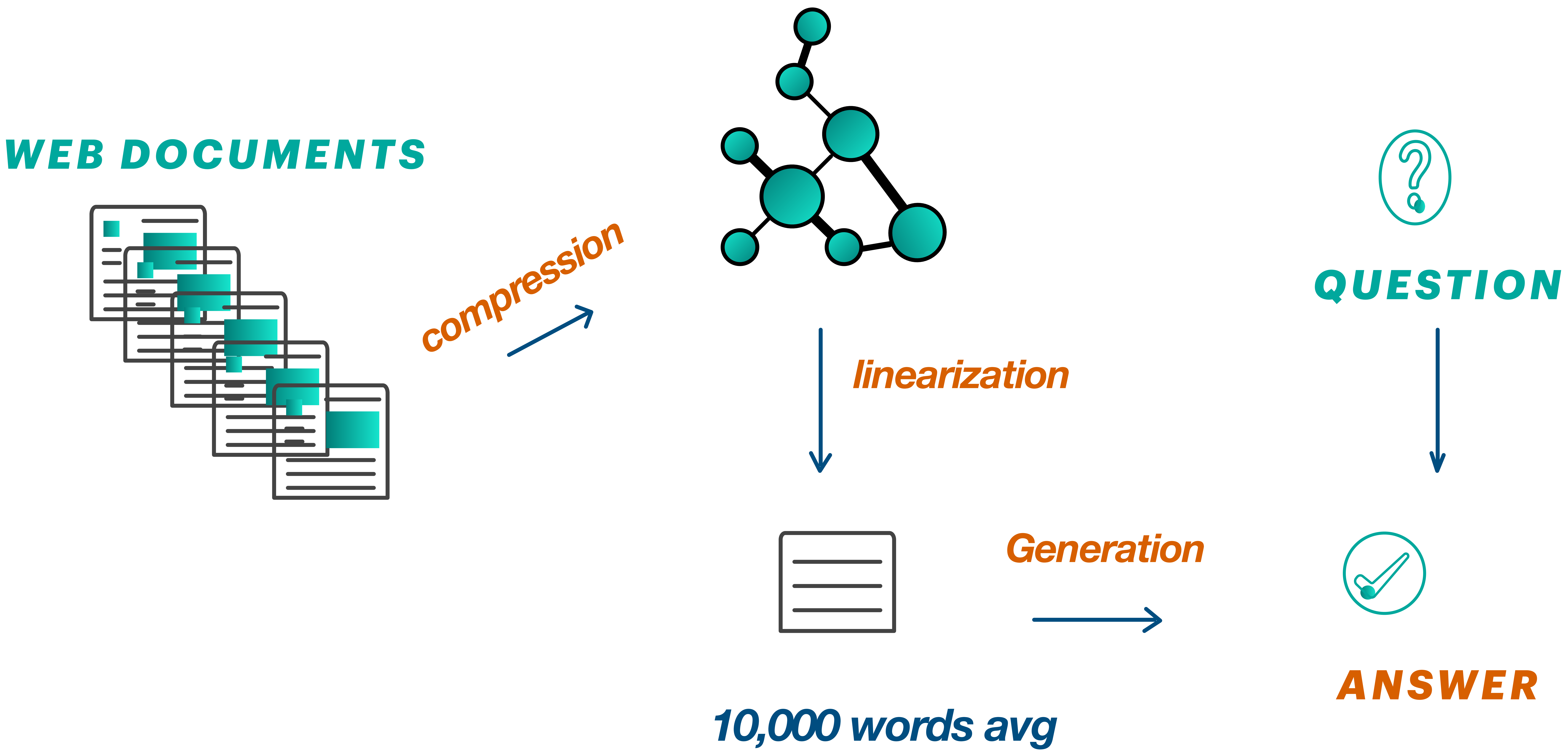
850 words avg

40% of the Answer Tokens are Missing

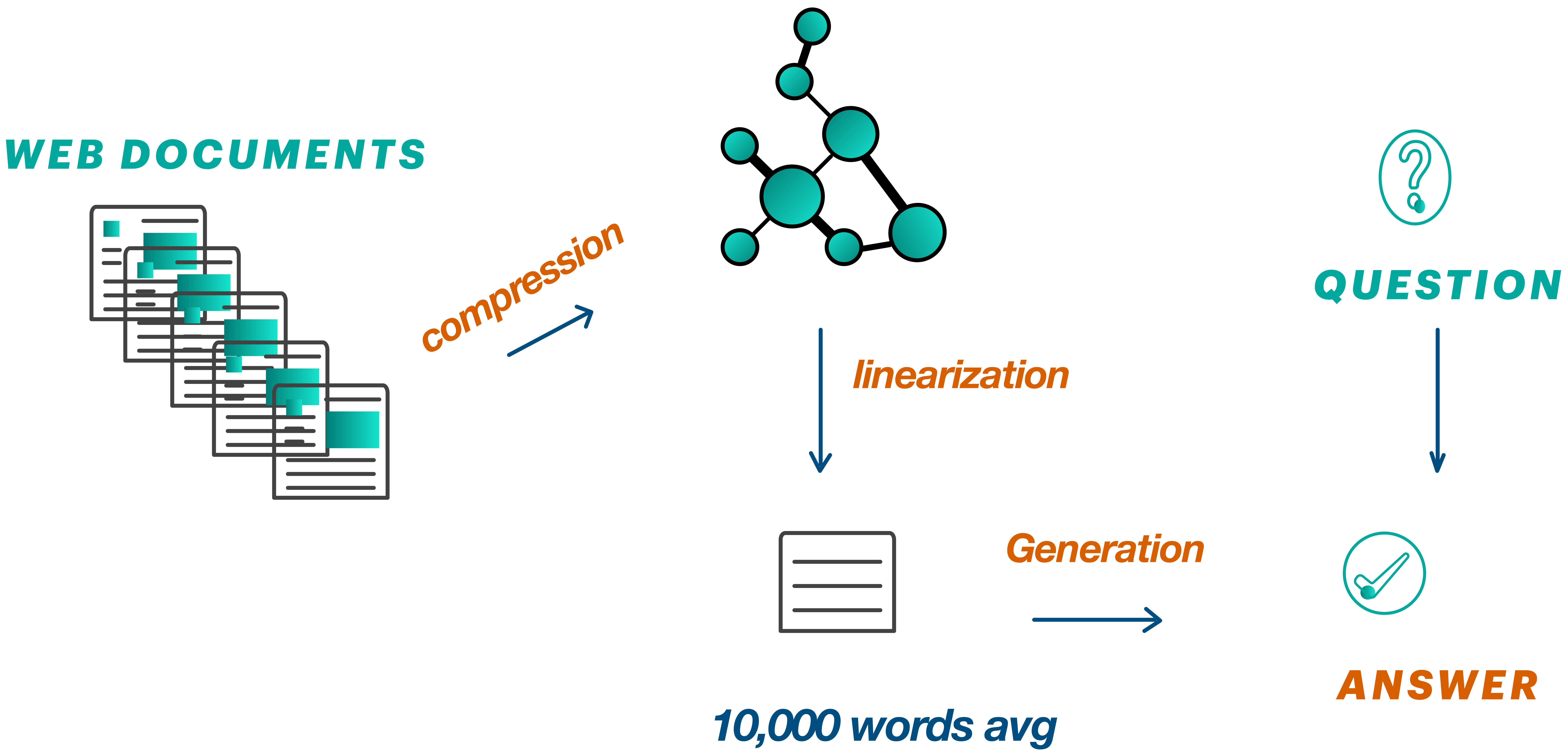
Information selected is Redundant

Web Input is Noisy, Selection is Hard

Knowledge Graph Construction

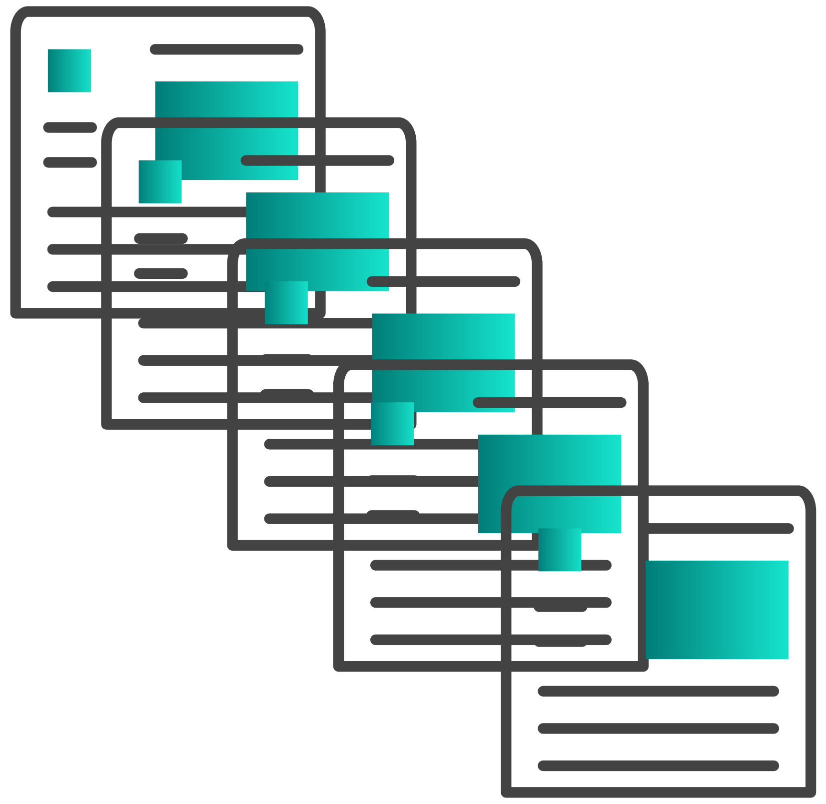


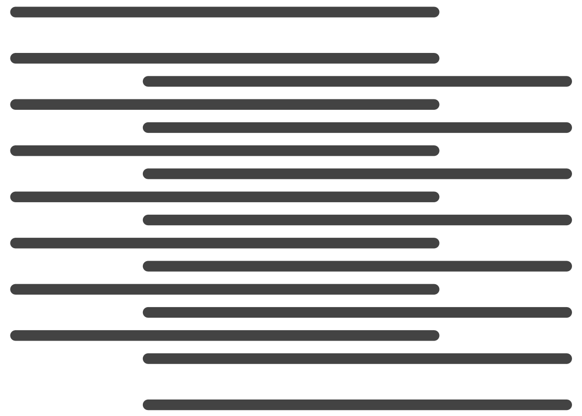
Knowledge Graph Construction



Converting a Text to a Graph

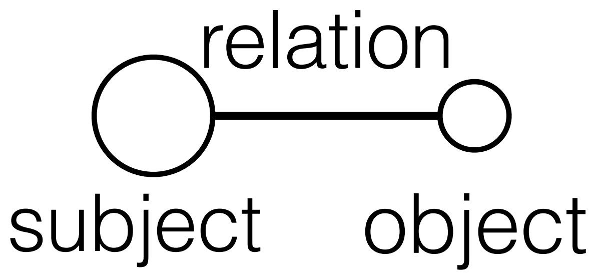
WEB DOCUMENTS





WEB DOCUMENT SENTENCES

open information extraction



coreference Resolution

Merge nodes
Increment
Nodes Weight

Tf-idf filtering

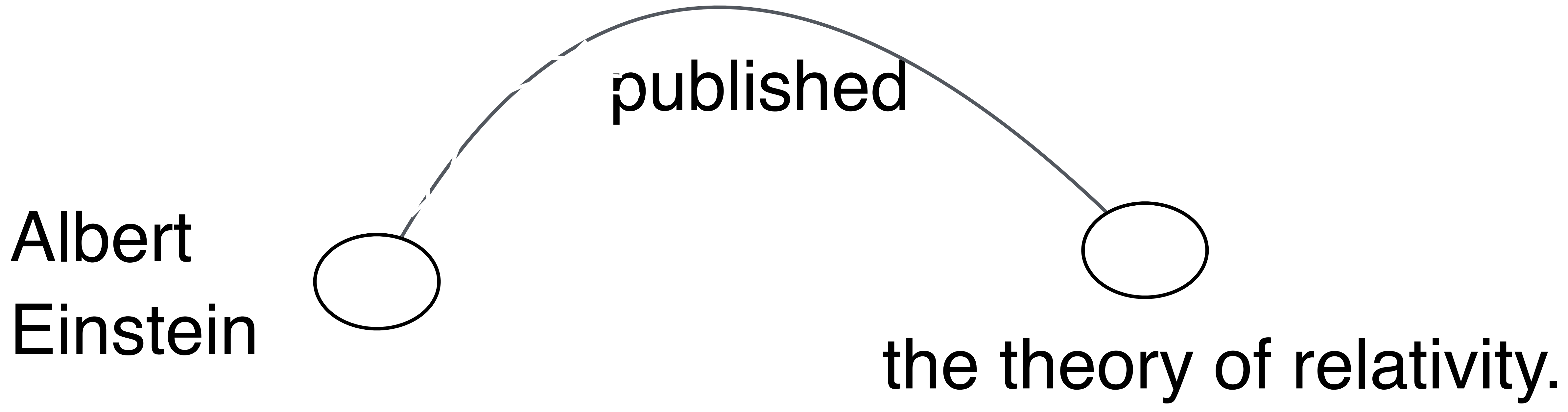
Filter Irrelevant
Input

Open Information Extraction

Converting text to edges

Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published the theory of relativity.



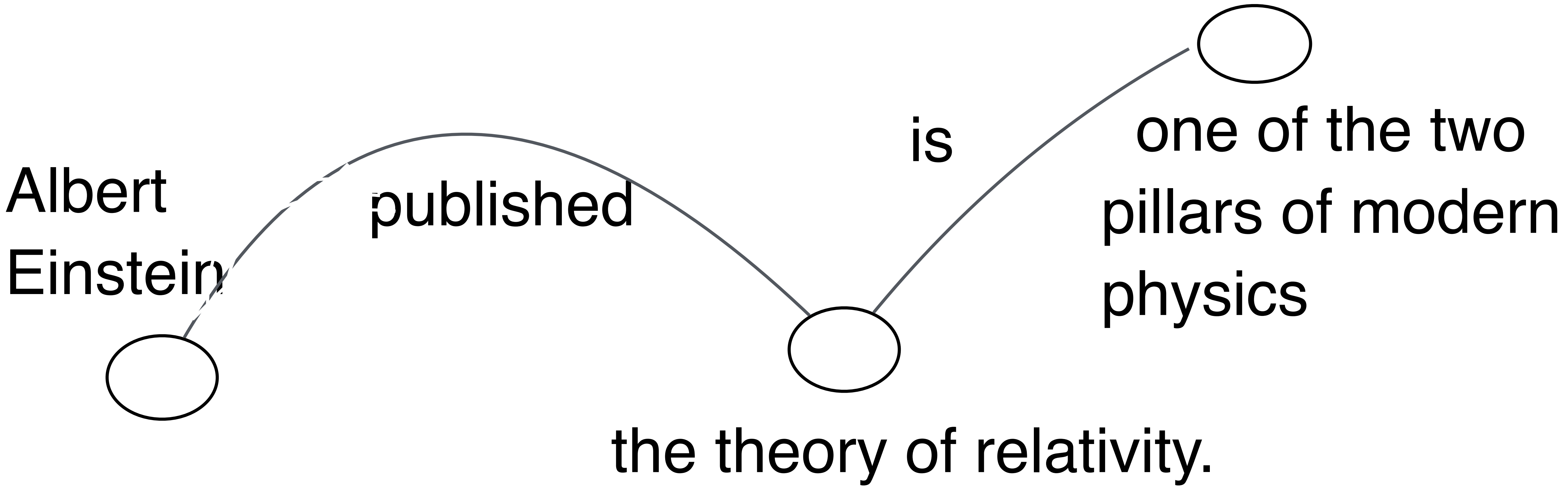
Coreference

Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published the theory of relativity.

The theory of relativity is one of the two pillars of modern physics

node weight +1

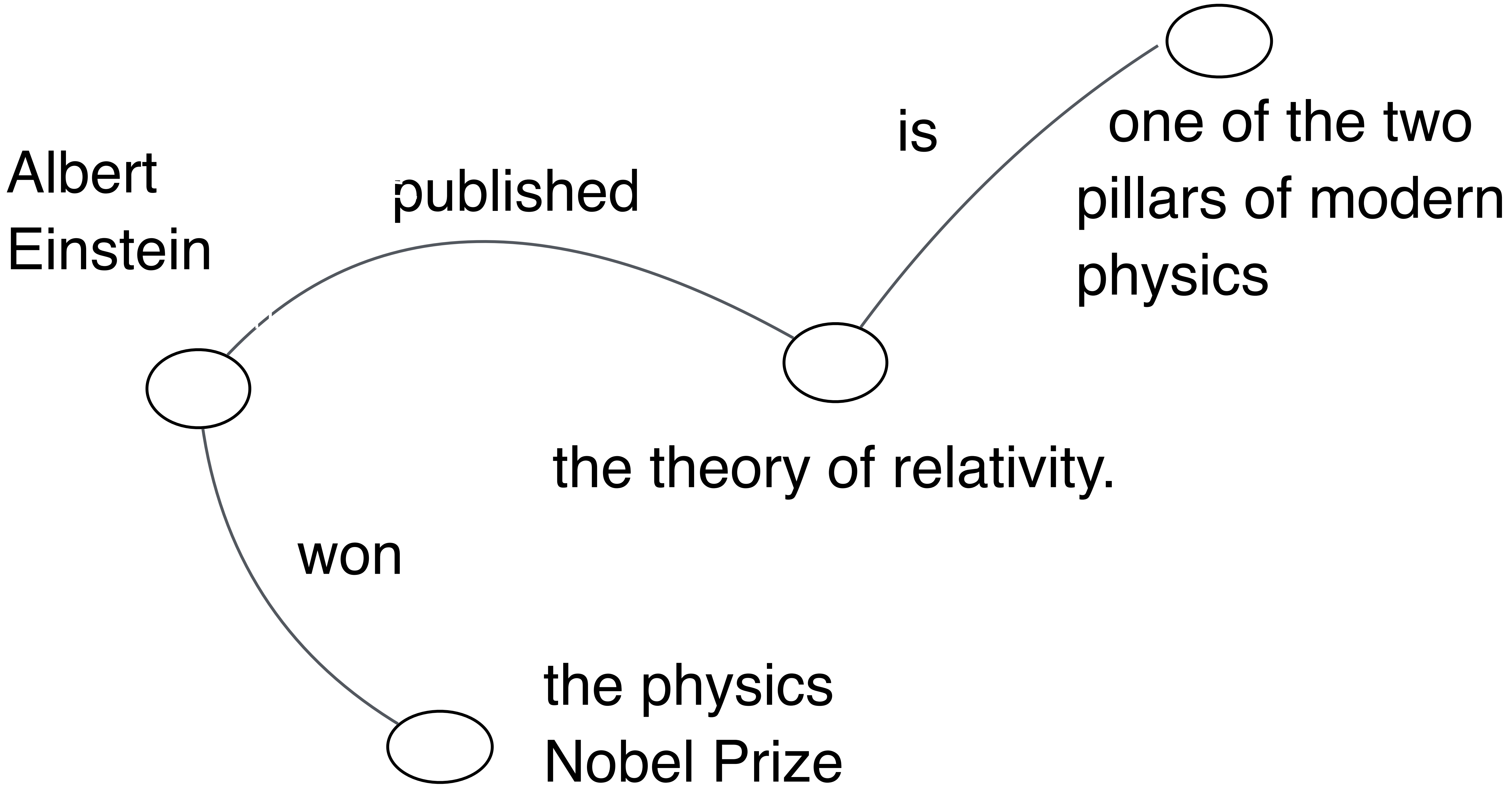


Coreference

Albert Einstein, a German theoretical physicist , published the theory of relativity.

The theory of relativity is one of the two pillars of modern physics.

He won the physics Nobel Prize
node weight +1



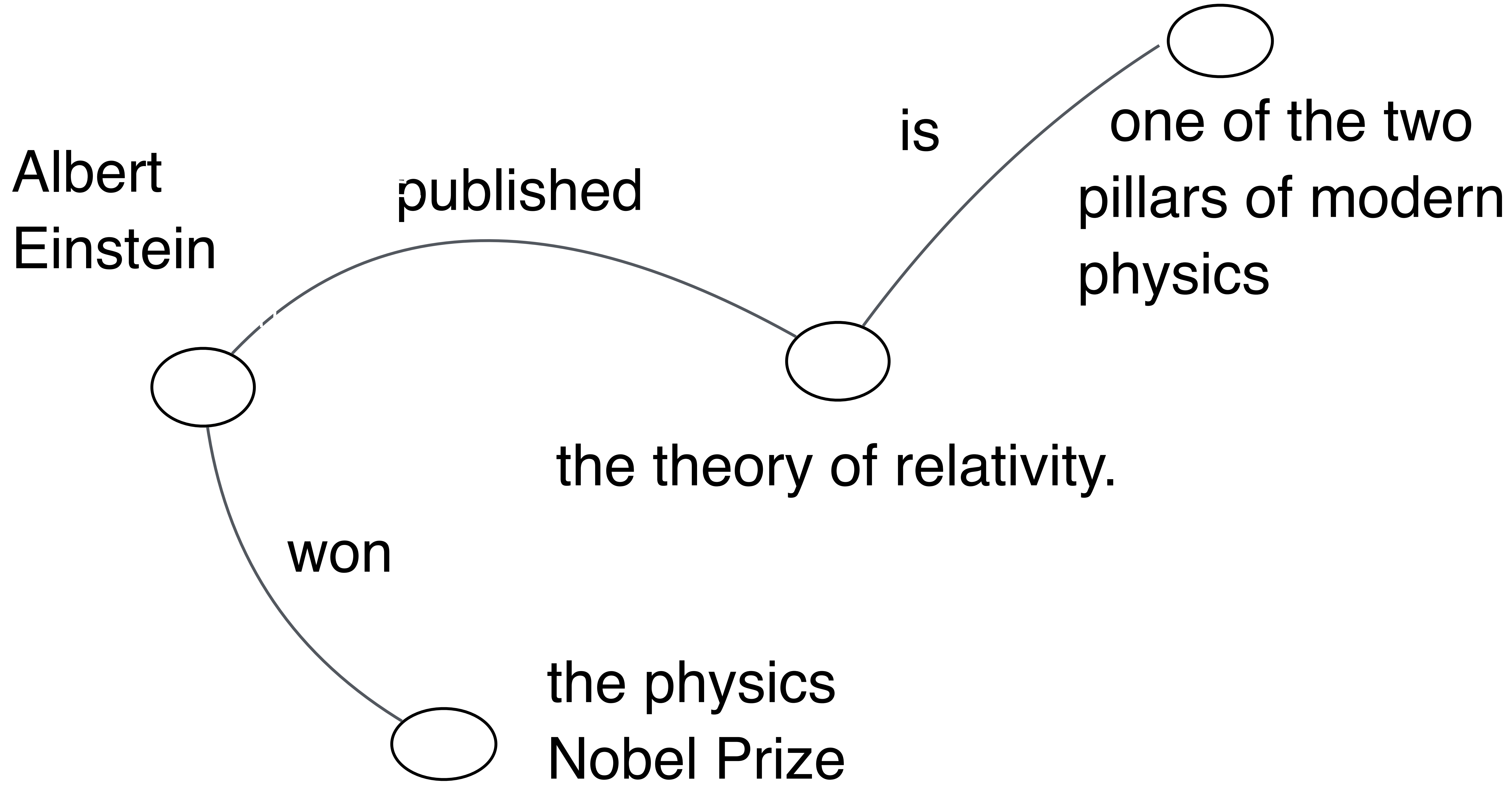
Relevance Filtering

Albert Einstein, a German theoretical physicist , published the theory of relativity.

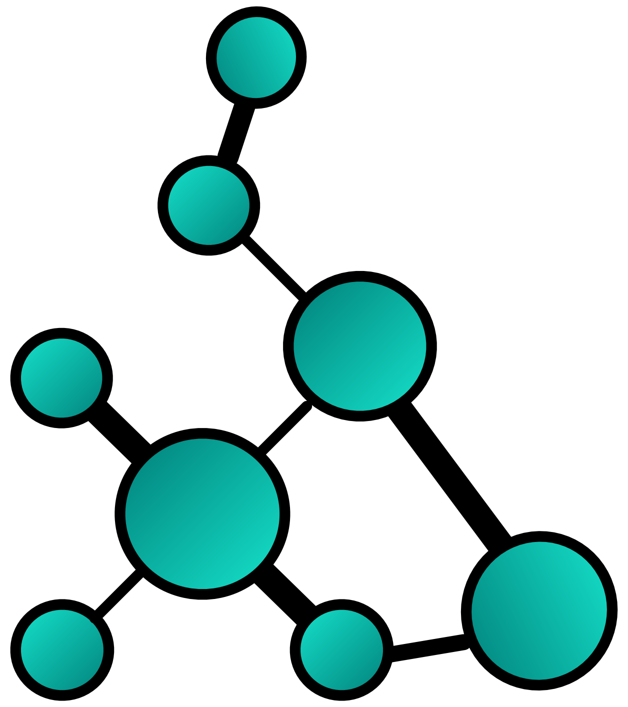
The theory of relativity is one of the two pillars of modern physics.

He won the physics Nobel Prize

Puppies are very cute.
Low TF-IDF overlap with query
Not added



Knowledge Graph Construction



Compresses the input by

- Merging redundant information
- Dropping words
- Filtering out irrelevant triples

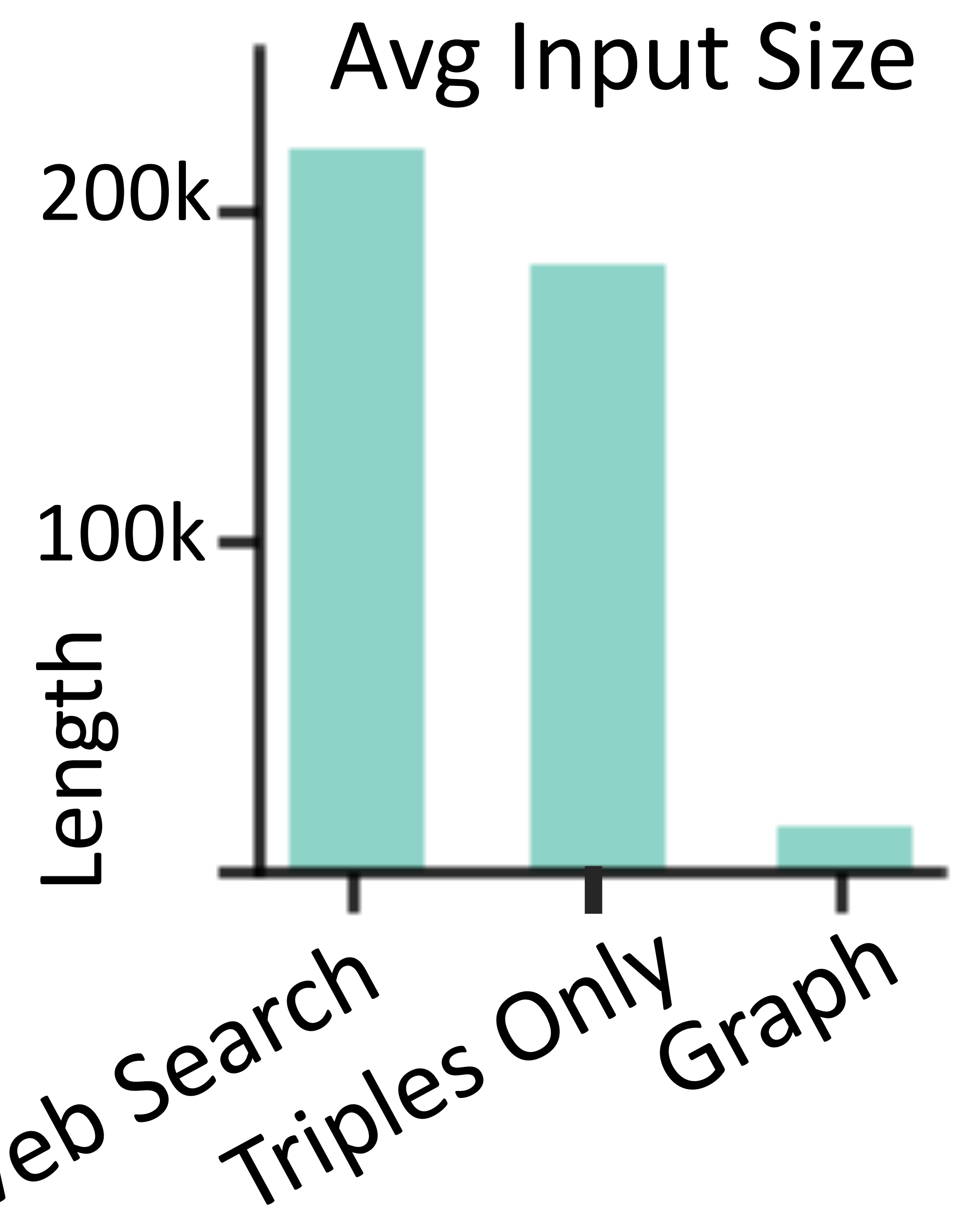
Reduces redundancy

- Merging nodes, edges and redundant triples

Filters out irrelevant content

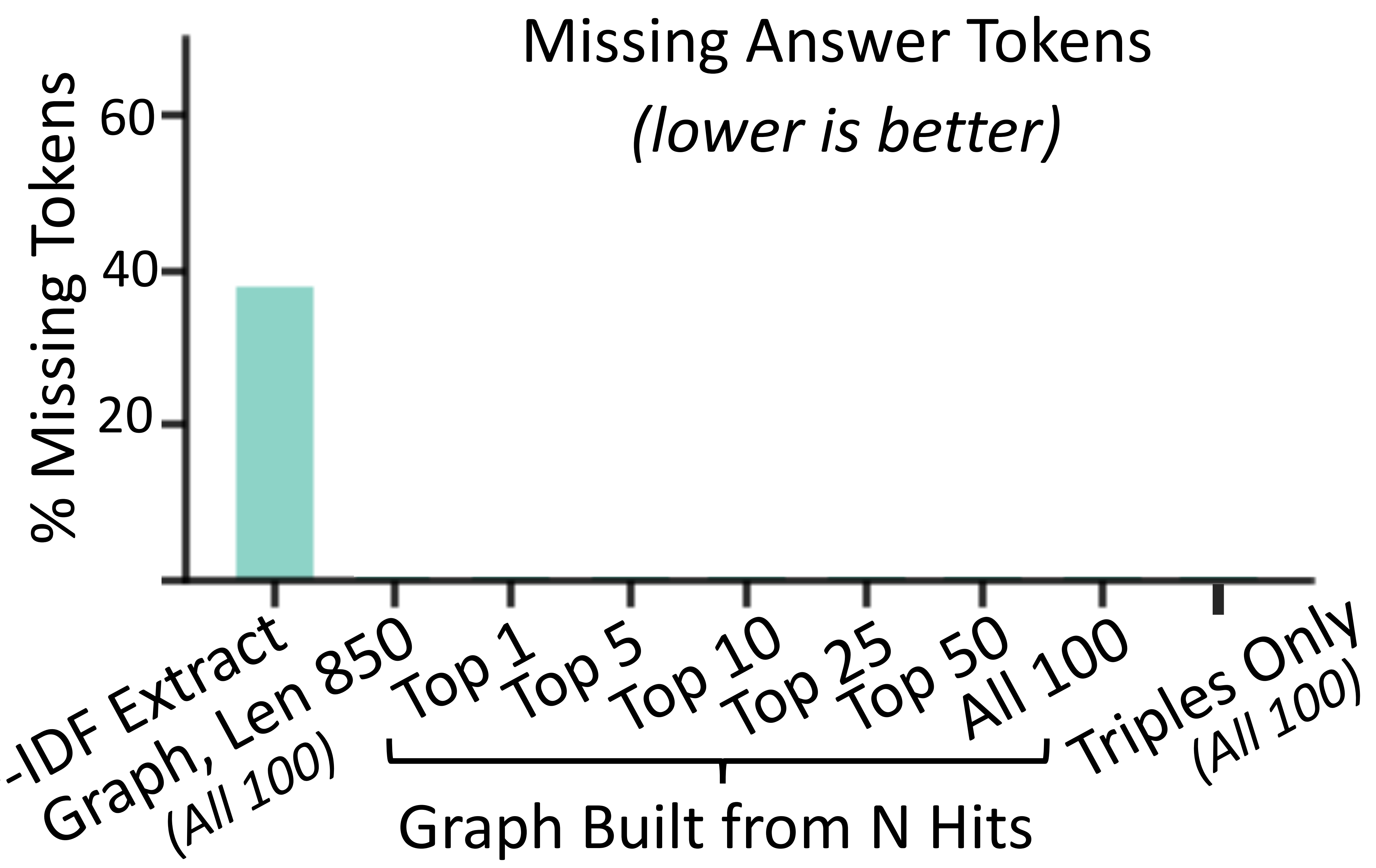
- Tf-idf overlap (Question,Triple)

How much does the graph manage to compress the input ?



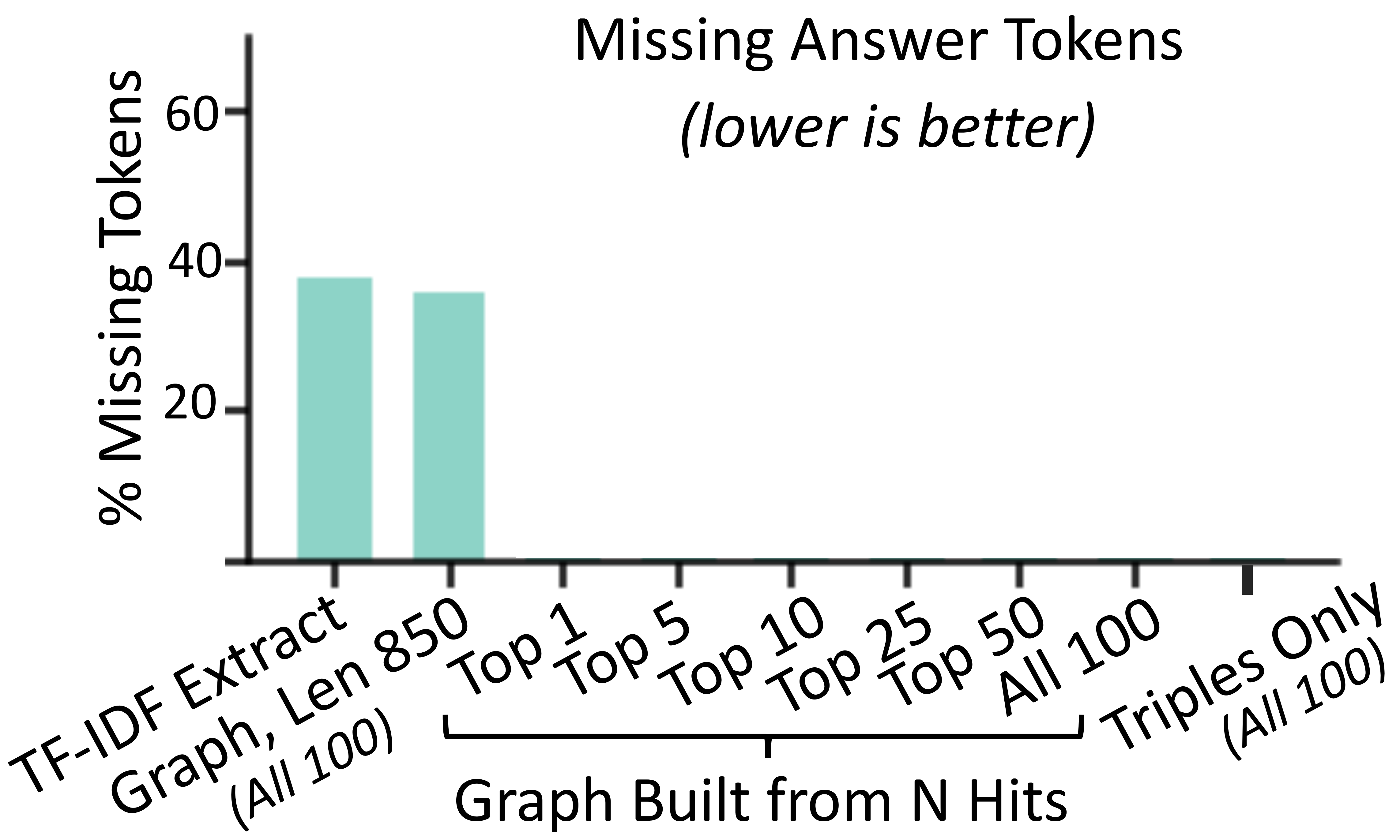
Knowledge Graph Construction drastically reduces the input size

How much does the graph preserve relevant information ?



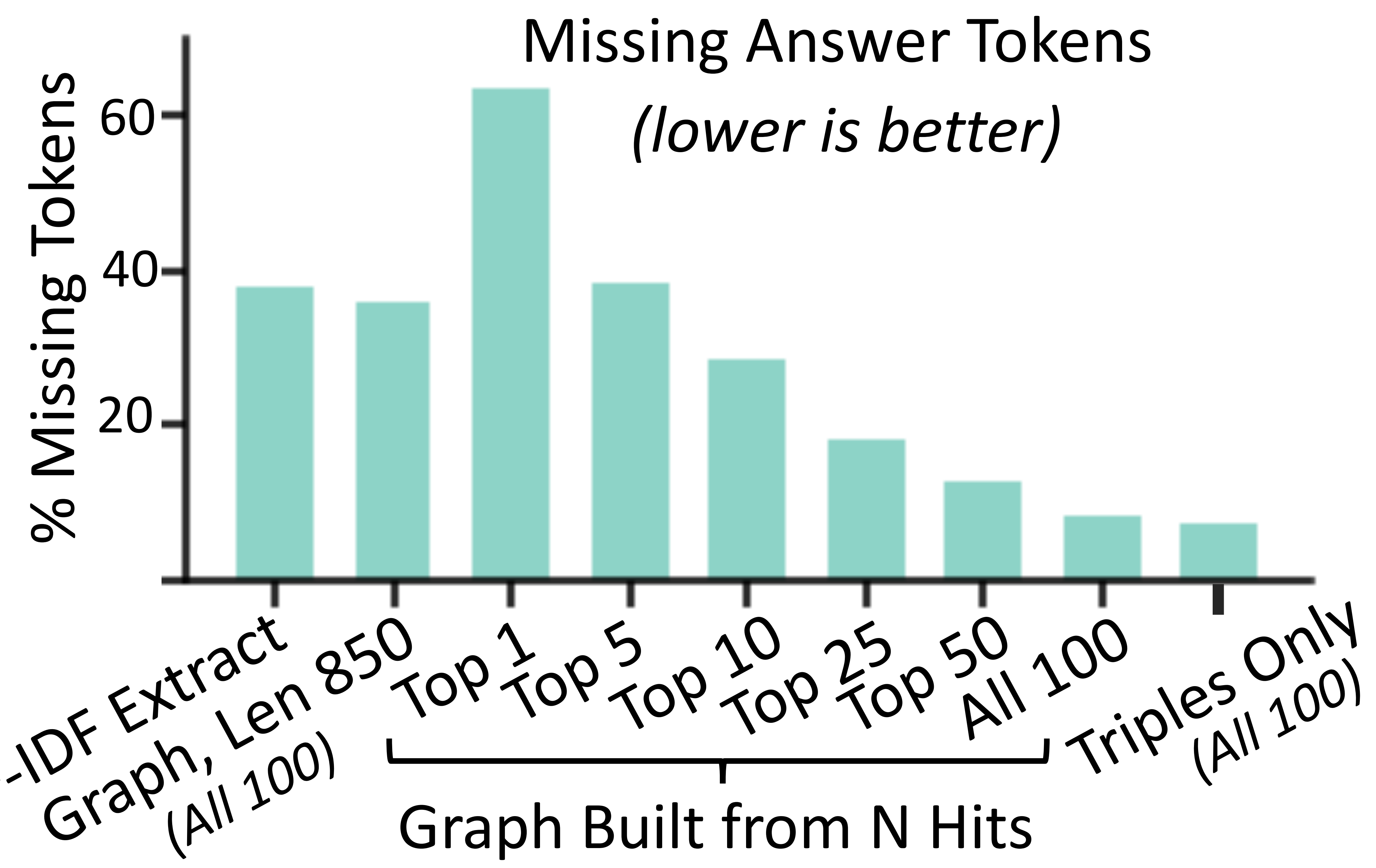
TF-IDF extraction is missing 38% of the answer tokens

Knowledge Graph Construction contains More Answer



The graph extracted for 850 tokens is missing 35% of the answer tokens

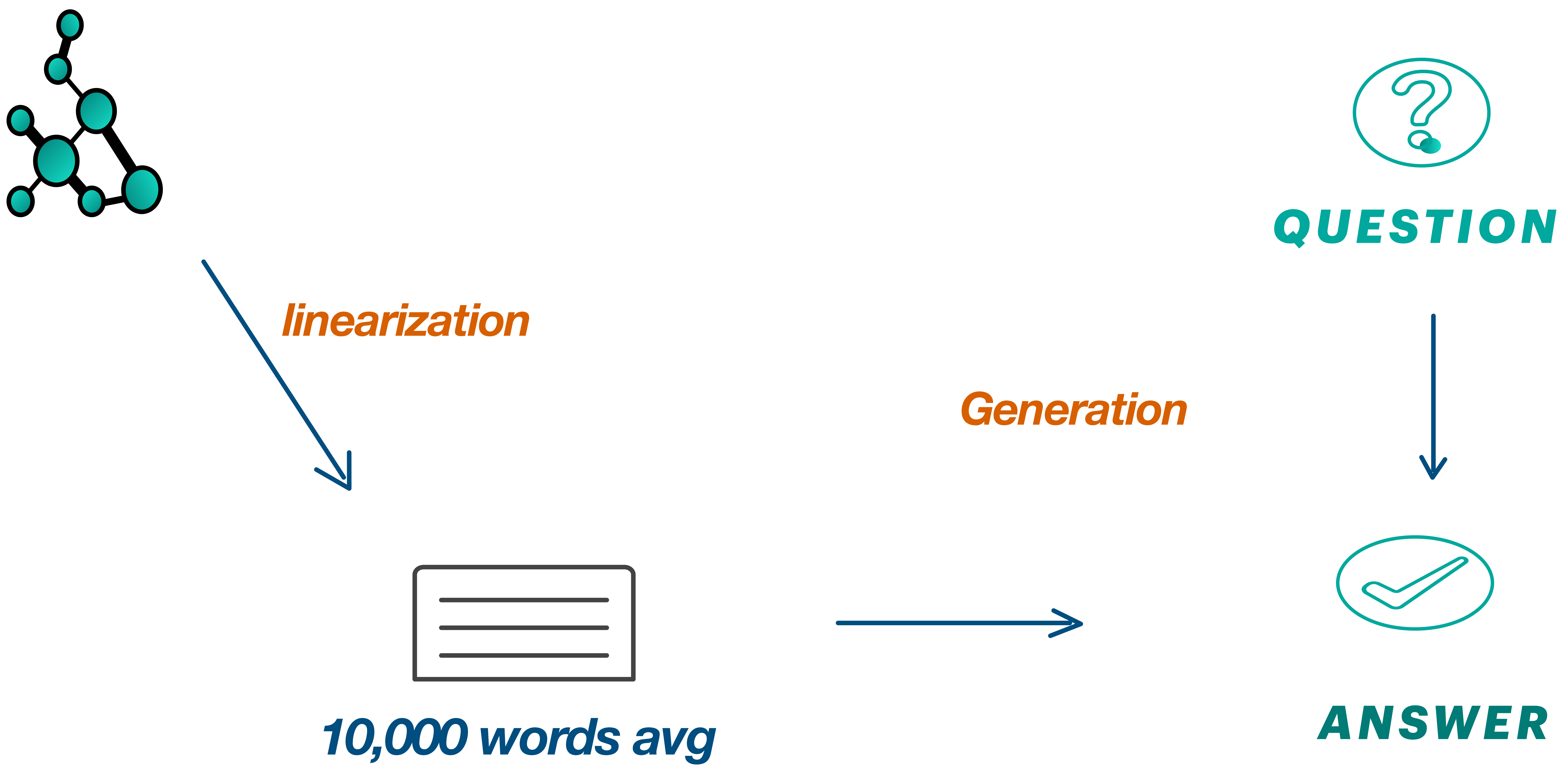
Knowledge Graph Construction contains More Answer



The graph for the full Input is missing only 8.7% of the answer tokens

Model

Generation Model



Encoding Graph Structure in a Seq2Seq Model

WORD EMBEDDING <sub> Albert Einstein <obj> the theory of relativity <pred> published <s> developed <obj> the Physics Nobel Prize <s> won

POSITION EMBEDDING 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Encoding Graph Structure in a Seq2Seq Model

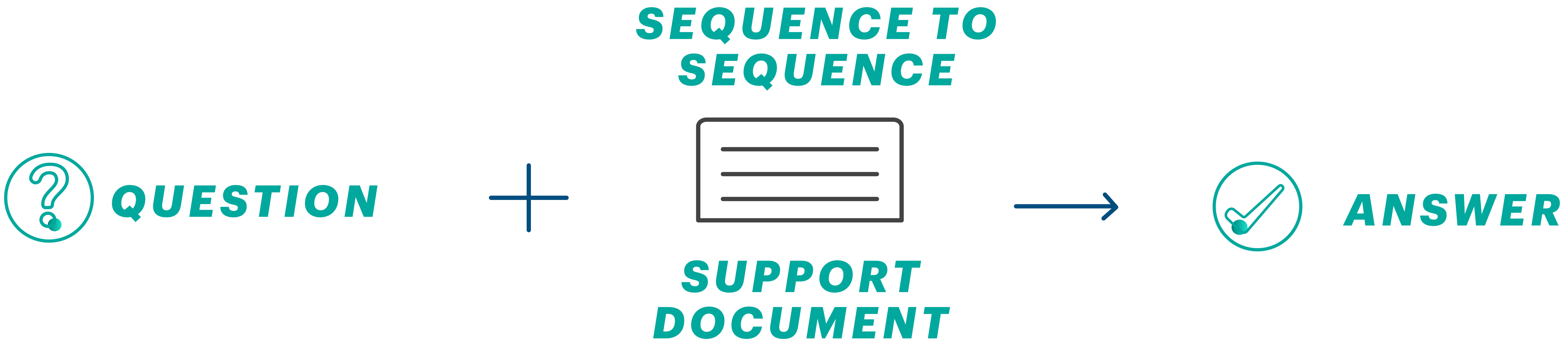
WORD EMBEDDING	<sub>	Albert	Einstein	<obj>	the	theory	of	relativity	<pred>	published	<s>	developed	<obj>	the	Physics	Nobel	Prize	<s>	won
POSITION EMBEDDING	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
GRAPH WEIGHT EMBEDDING	0	4	4	0	2	2	2	2	0	1	0	1	0	3	3	3	3	0	2

Encoding Graph Structure in a Seq2Seq Model

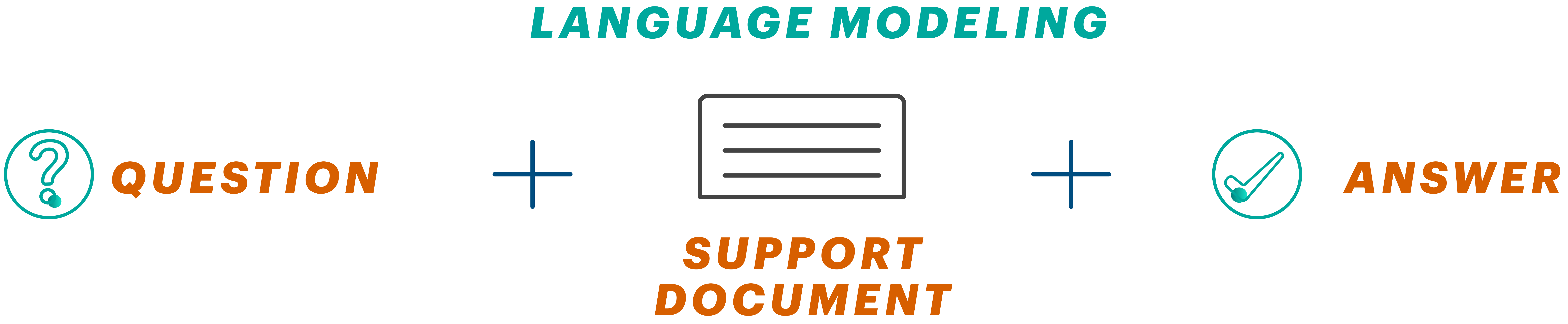
WORD EMBEDDING	<sub>	Albert	Einstein	<obj>	the	theory	of	relativity	<pred>	published	<s>	developed	<obj>	the	Physics	Nobel	Prize	<s>	won
POSITION EMBEDDING	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
GRAPH WEIGHT EMBEDDING	0	4	4	0	2	2	2	2	0	1	0	1	0	3	3	3	3	0	2
QUERY RELEVANCE EMBEDDING	0	1	1	0	1	1	1	1	0	1	0	2	0	1	1	1	1	0	1

Sequence-to-Sequence Model

Generate each word of the answer



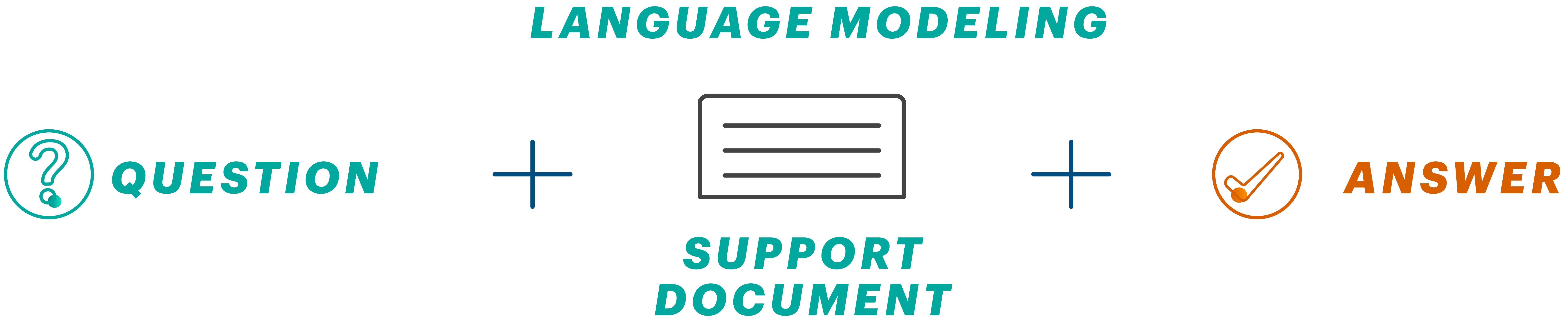
Language Model



Language Modeling Model

Inference time: provide true question and support document

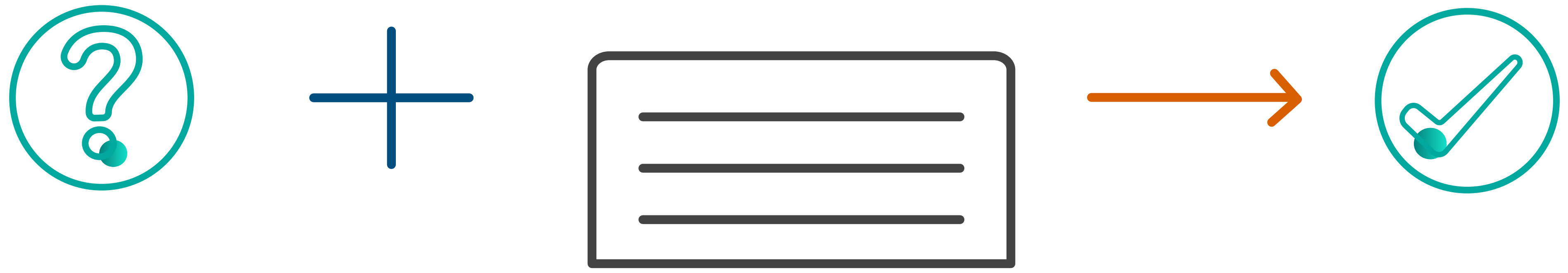
evaluate answer



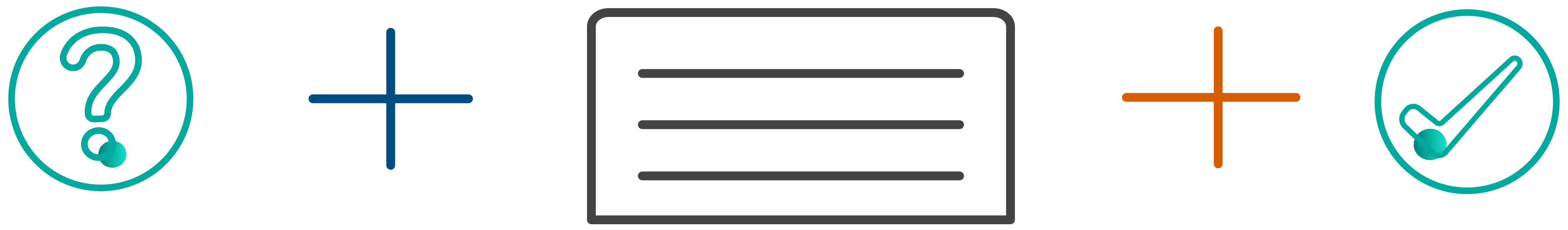
MULTITASK LEARNING

training time: train on many tasks

SEQUENCE TO SEQUENCE

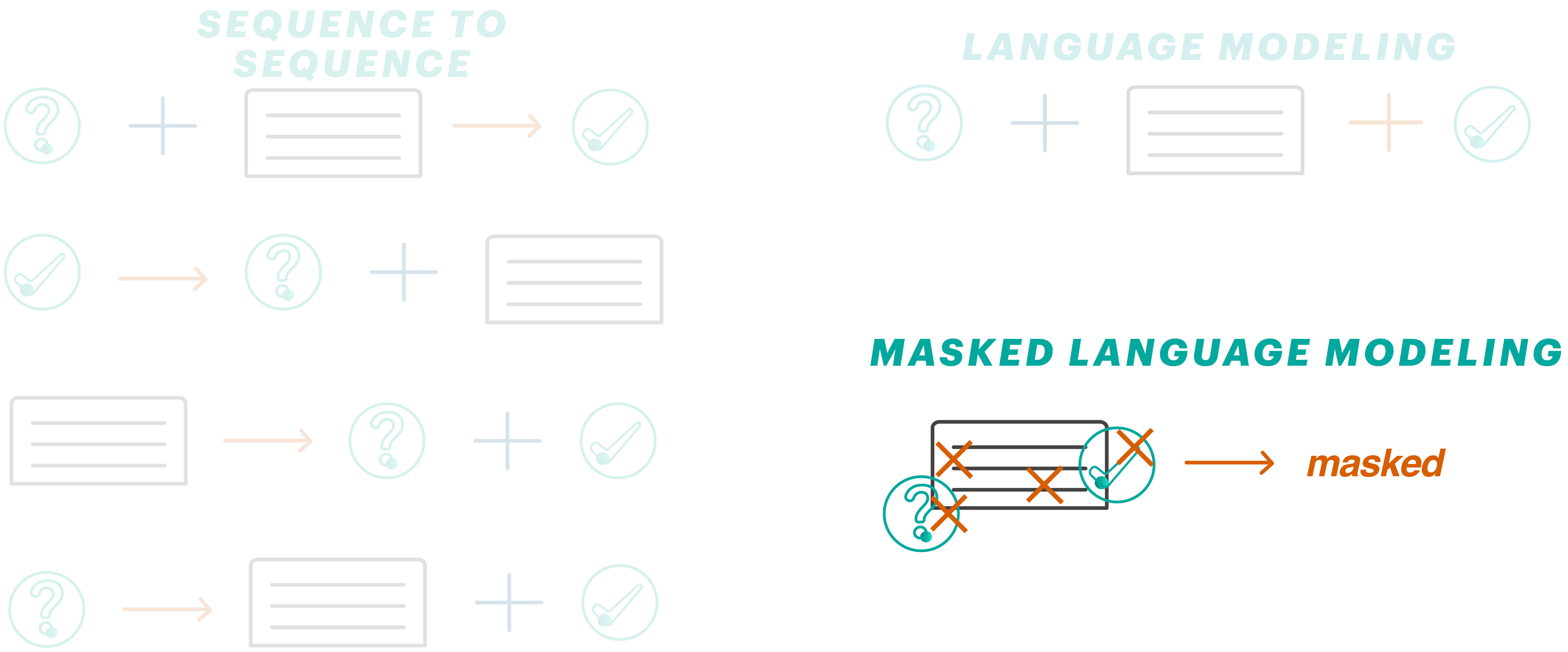


LANGUAGE MODELING

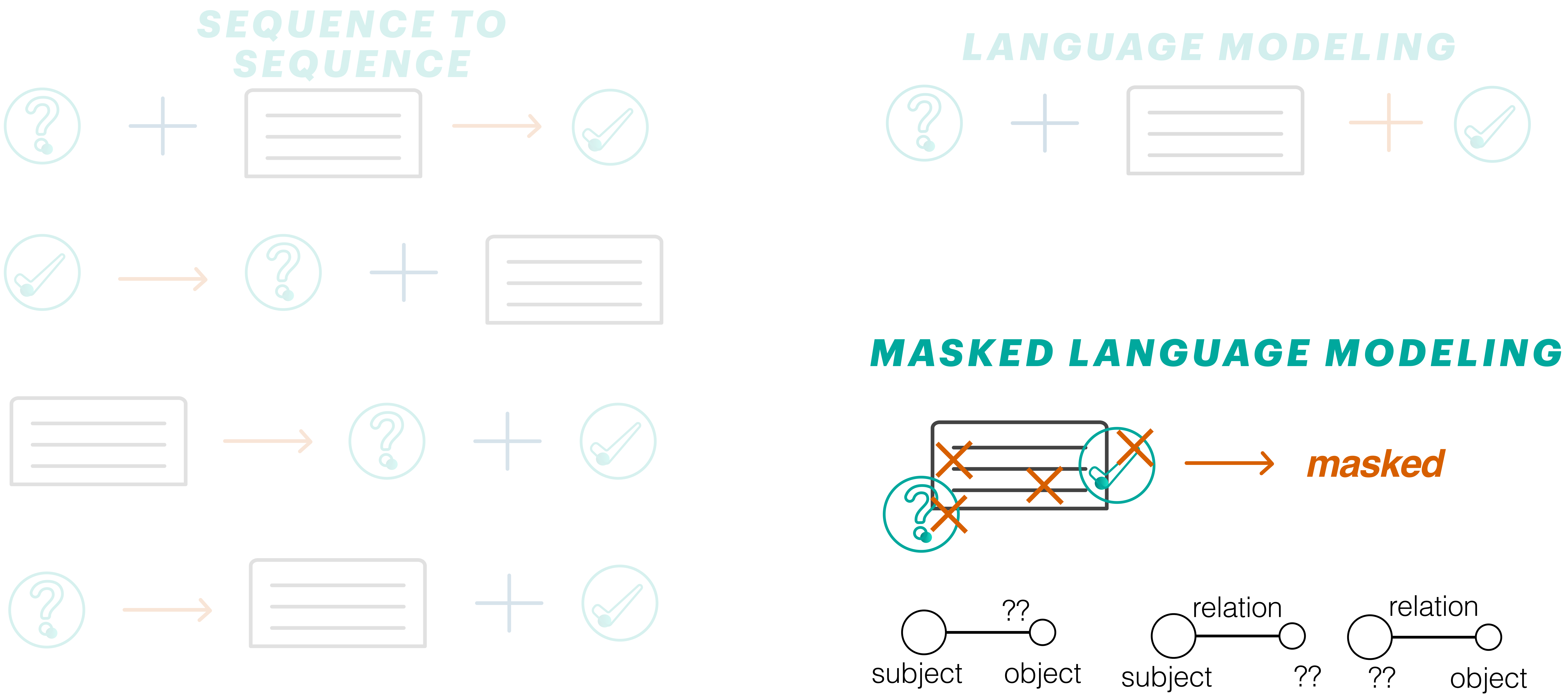


training time: train on many tasks

training time: train on many tasks

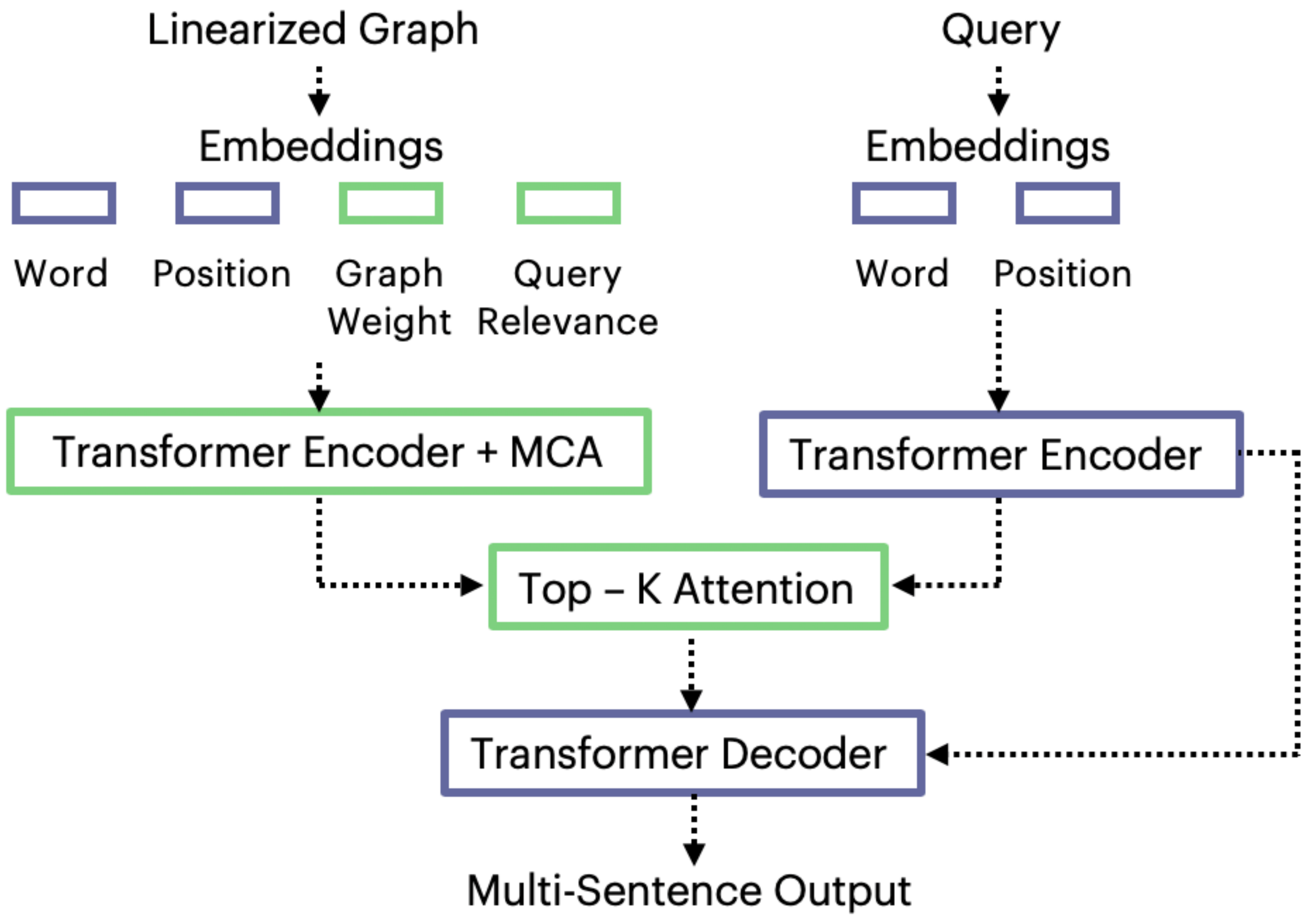


training time: train on many tasks



Handling Long Input

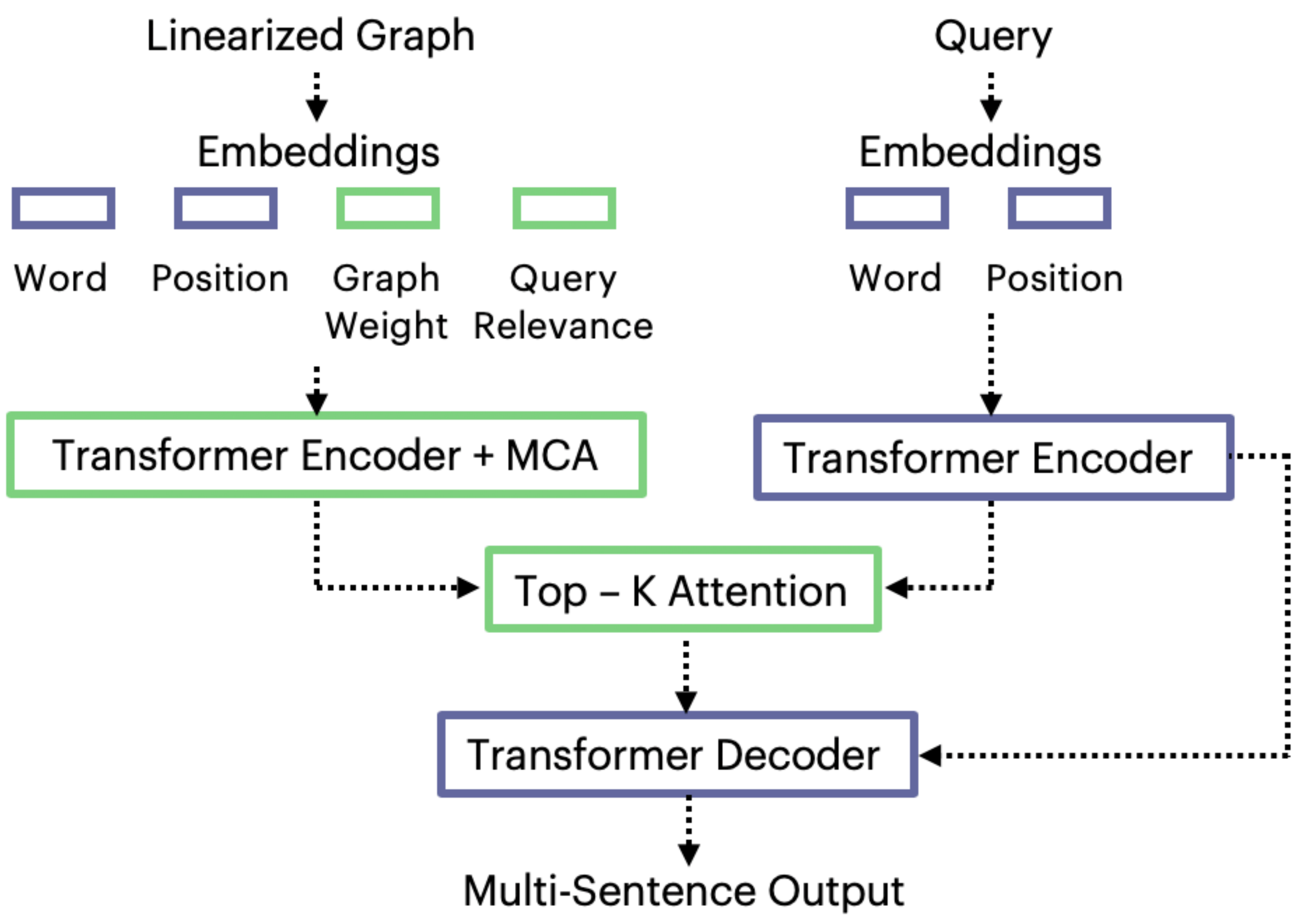
How do we encode 10K tokens in a Transformer ?



- *MCA in Encoder*

  Memory Compressed Attention

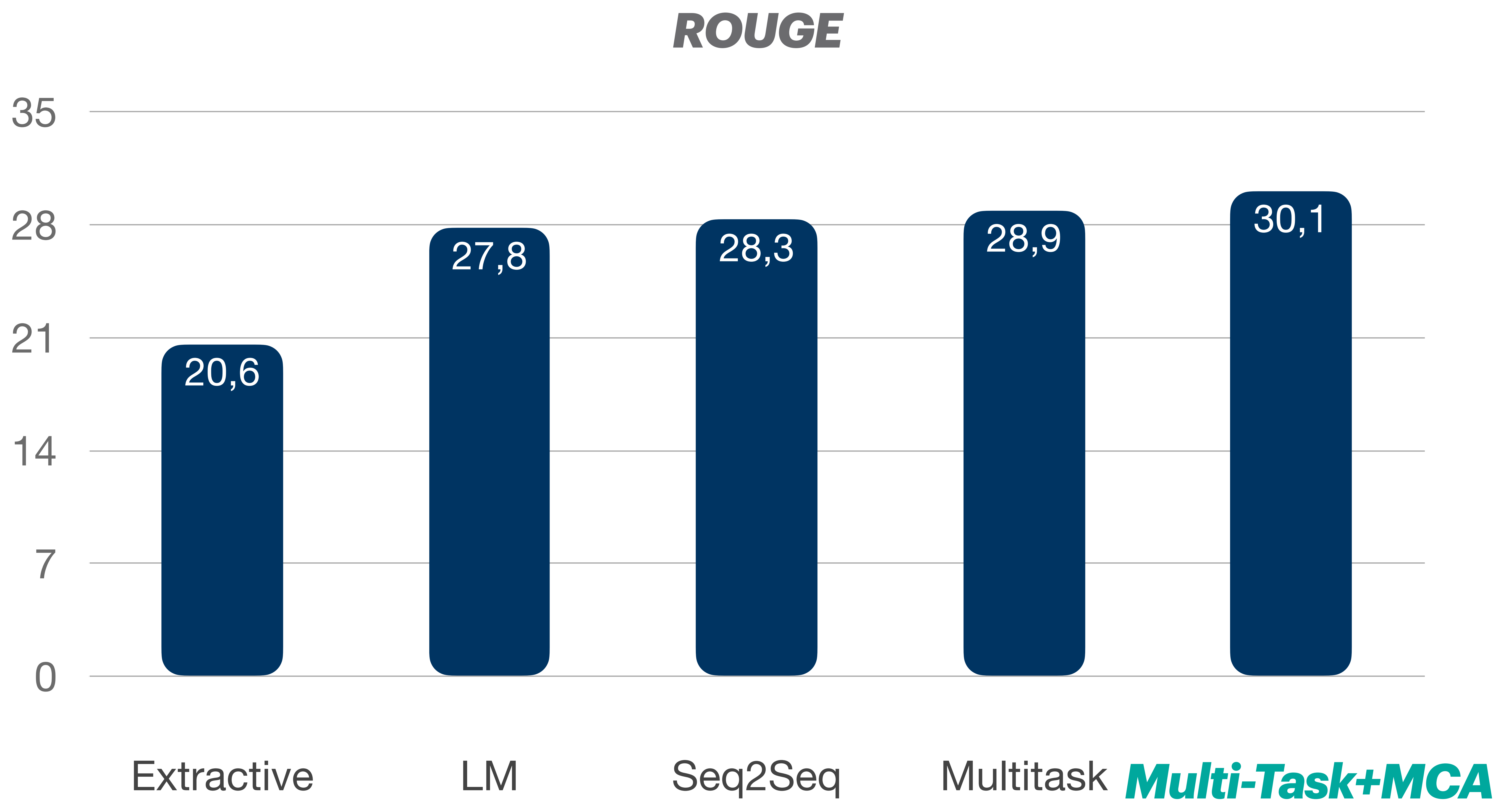
Handling Long Input



- ***MCA in Encoder***
  Memory Compressed Attention
- ***Hierarchical Top-k Attention***

Evaluation

Automatic Evaluation



Human Evaluation: Preference

Multi-task

Multi-Task+MCA

Generalising to other Datasets: WikiSum

Title: dwight h. perkins (economist)

Beginning of Web Search: memorial minute adopted by the faculty of arts and sciences, harvard university: abram bergson, john kenneth galbraith, carl kaysen, raymond vernon, dwight h. perkins, chairman. ed was a generous and understanding man who could see the good qualities in people and bring them out. he was always determined and persistent, which meant that he also saw to completion what he started . the list of projects, large and small, that he led is long. in 1946 he was one of the authors of the speech of secretary of state james byrnes in which the secretary announced the return of responsibility for the german economy to the germans. in 1956, he, together with ray vernon, did a pioneering study of the new york metropolitan region, a study that tried to identify the economic, political and social forces that were shaping that vast urban area. at the time the problems of urban areas were mainly the concern of architects and philosophers; almost no economist and only a few political scientists knew much or cared greatly about such issues. the studies that resulted provided rich fare for a generation of urban planners to follow. mason, edward s.(1899-1992) — harvard square library home biographies mason, edward s.(1899-1992) mason, edward s.(1899-1992) edward sagendorph mason was a member of the first parish unitarian church in harvard square who exemplified liberal religion in both thought and action. in addition to his notable contributions to the science of government, he served as chairman of the sloan commission on cable communication which issued recommendations for the future, on the cable. [...]

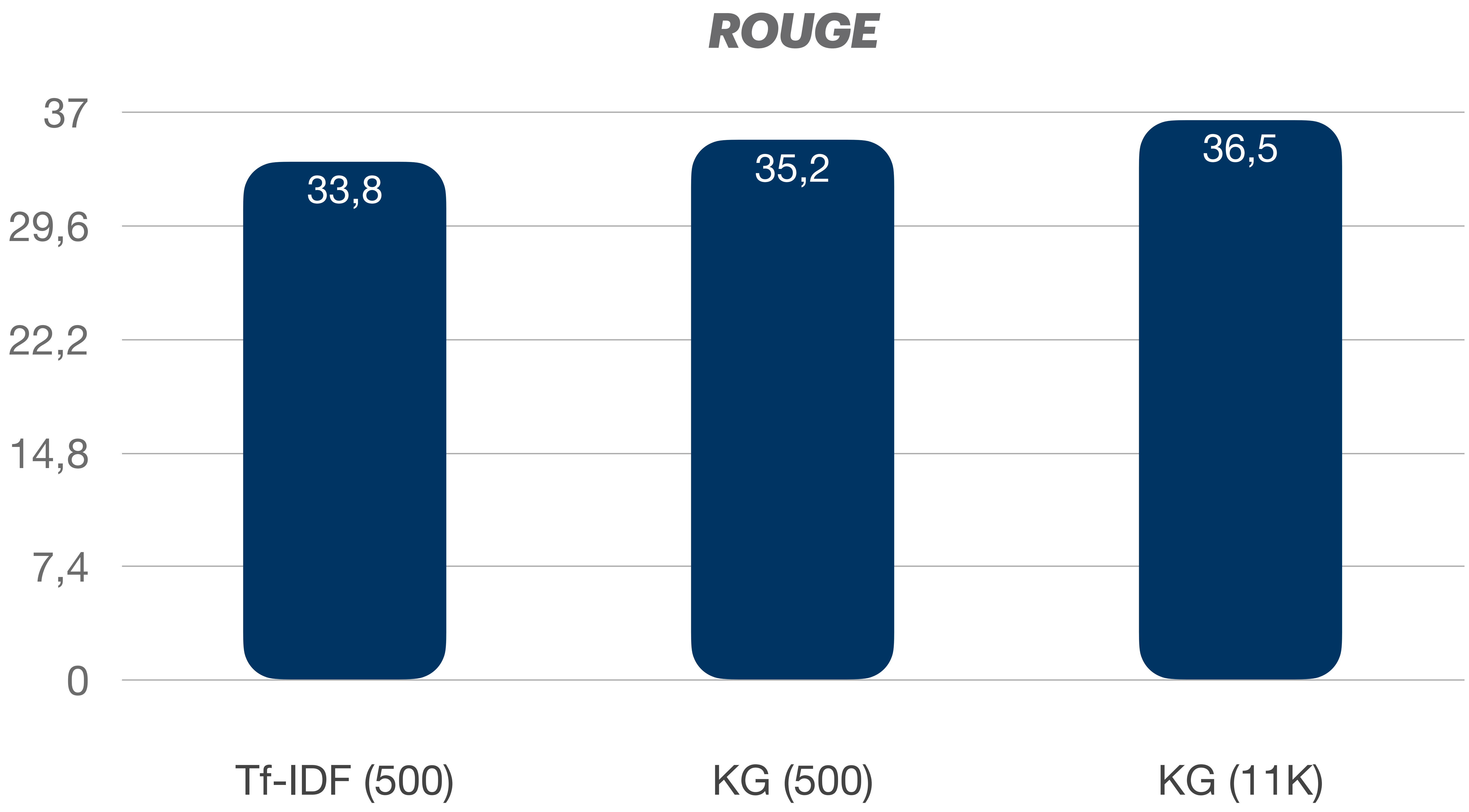
Target Lead Paragraph: dwight heald perkins (born in chicago, illinois in 1934) is an american academic, economist, sinologist and professor at harvard university. he is the son of lawrence bradford perkins, architect, and margery blair perkins and the grandson of dwight heald perkins, the architect, for whom he was named. he married julie rate perkins in 1957 and they have three adult children. == early life == perkins earned an undergraduate degree at cornell university in 1956. after two years military service in the us navy, perkins resumed his studies at harvard. he earned a ma in economics in 1961 and a ph.d. in economics in 1964. his doctoral thesis was "price formation in communist china". == academic career == perkins' teaching career at harvard began when he was still a graduate student, and continued uninterrupted through 2006 when he became a research and emeritus professor. he was a member of both the department of economics of the faculty of arts and sciences at harvard and of the harvard kennedy school. [...]

Title
= query for web search

Multi-document input
= Result from web search

Target
= Wikipedia Article lead paragraph

Automatic Evaluation



Example of Generated Text (ELI5)

Question: Why is touching microfiber towels such an uncomfortable feeling?

True Answer: Do you mean the kind of cloths used to clean glasses and lenses? I've never noticed any uncomfortable feeling myself, but I do find touching certain cleaning cloths can be quite uncomfortable. There's a brand called "e - cloth" which market themselves as not needing any cleaning supplies. Basically the threads have tiny little hooks and barbs in them, so dragging them over skin makes you feel like it 's catching against you.

Generation: The microfiber is made up of a bunch of tiny fibers that are attached to them. When you touch them, the fibers that make up the microfiber are attracted to each other. When they touch, they're attracted to the other end of the fiber, which is what makes them feel uncomfortable.

Fluent answer, factually correct

Some repetition ("attracted to each other")

Example of Generated Text (ELI5)

Question: Why do we have nightmares when stressed out?

True Answer: I can tell you that your mind, and in particular your dreams, are subject to the input that it the journey of your life, as you have interpreted it. Yours dreams tend to be your subconscious representing issues you have within, whether it be past or present, with things and people that are sort of "preinterpreted symbology" from your deep and ever power subconsciousness. All your nightmares are just dreams, and your inner mind knows exactly can mess with you the most, and sometimes it brings that to light, it usually being an issue you may be stressed about.

Generation: When you're stressed, your brain releases a hormone called cortisol. This hormone is released by your adrenal glands, and is released when you are stressed. This is why it's called "fight-or-flight" mode. When your body is stressed, it releases hormones that make you feel more stressed. The reason why you don't feel stressed is because your brain is trying to get rid of the stress hormone, which causes you to feel stressed, and you're less likely to get stressed. If you're really stressed, you're probably going to be more stressed than if you weren't stressed.

Repetitions ("when you are stressed")
Non sense fragments ("The reason …get stressed")

Example of Generated Text (WikiSum)

Title: wylie rogers

True Paragraph: wylie rogers (born march 16, 1985) is an american ice hockey goaltender. he played with the rio grande valley killer bees of the central hockey league. prior to turning professional, rogers attended the university of alaska fairbanks where he played four seasons of ncaa men's division i ice hockey with the alaska nanooks men's ice hockey team. rogers played the 2010-11 season in the netherlands with the heerenveen flyers.

Generation: wylie rogers is an american football coach and former player. he is currently the head football coach at the university of north carolina at chapel hill. rogers played college football at north carolina state university. he was a member of the north carolina tar heels men's football team that won the ncaa division i football championship in 2012. he also played for the university's men's basketball team. after his playing career ended, rogers was hired as the head coach at unc chapel hill in 2013. he led the tar heels to their first ncaa division ii championship in 2015.

Fluent text with nice, elaborate discourse structure

Incorrect facts (Wylie Rogers is an ice hockey player not a football coach. University of Alaska not North Carolina. etc;)

Retrieving and Integrating Relevant Knowledge

Human-Machine Dialog

Retrieving Knowledge to generate good Responses

Human: I love watching Disney movies! Do you like Disney ?

Model: That’s nice !

Human-Machine Dialog

Retrieving Knowledge to generate good Responses

Human: I love watching Disney movies! Do you like Disney ?

Model: That’s nice !

Model: I love Disney, I love watching Disney movies and different animations and characters that makes them unique

Human-Machine Dialog

Retrieving Knowledge to generate good Responses

Human: I love watching Disney movies! Do you like Disney ?

Model: I love Disney, I love watching Disney movies and different animations and characters that makes them unique

Human: What's your favorite Disney movie ?

Model: I love the incredibles, they are my favorite Disney movie.

Human: omg, me too! The second one just came out. I don't watch a lot of sequels though, I think the last one I saw was Toy Story

Model: That's cool. I remember when Toy Story was first released in 1995

Retrieval Augmented Generation

Two retrieval sources

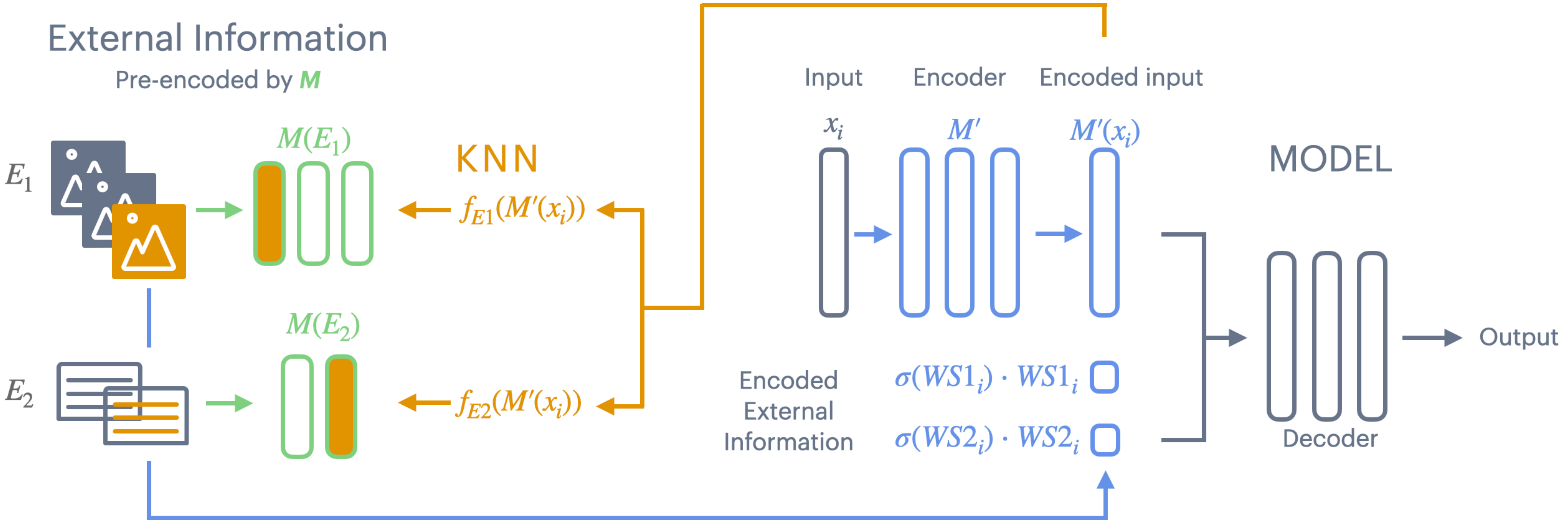
Human: What's your favorite Disney movie ?

Model: I love the incredibles, they are my favorite Disney movie.

Knowledge: Disney announced intentions to develop additional superhero films after the success by the incredibles

Template: I love kiteboarding, it is one of my favorite activities on the water.

Extending Human-Machine Dialog with External Retrieval using K-Nearest Neighbour Search



Wizard of Wikipedia

Dialog about a topic

Retrieval Corpus for KL

- WKP passages
- 34 sentences per topic

Retrieval Corpus for Template

- Dialog turns
- 170K dialog turns

Image Chat

Dialog about an image

Retrieval Corpus for KL

- Image + dialog
- 184K images

Retrieval Corpus for Template

- Dialog turns
- 350K dialog turns

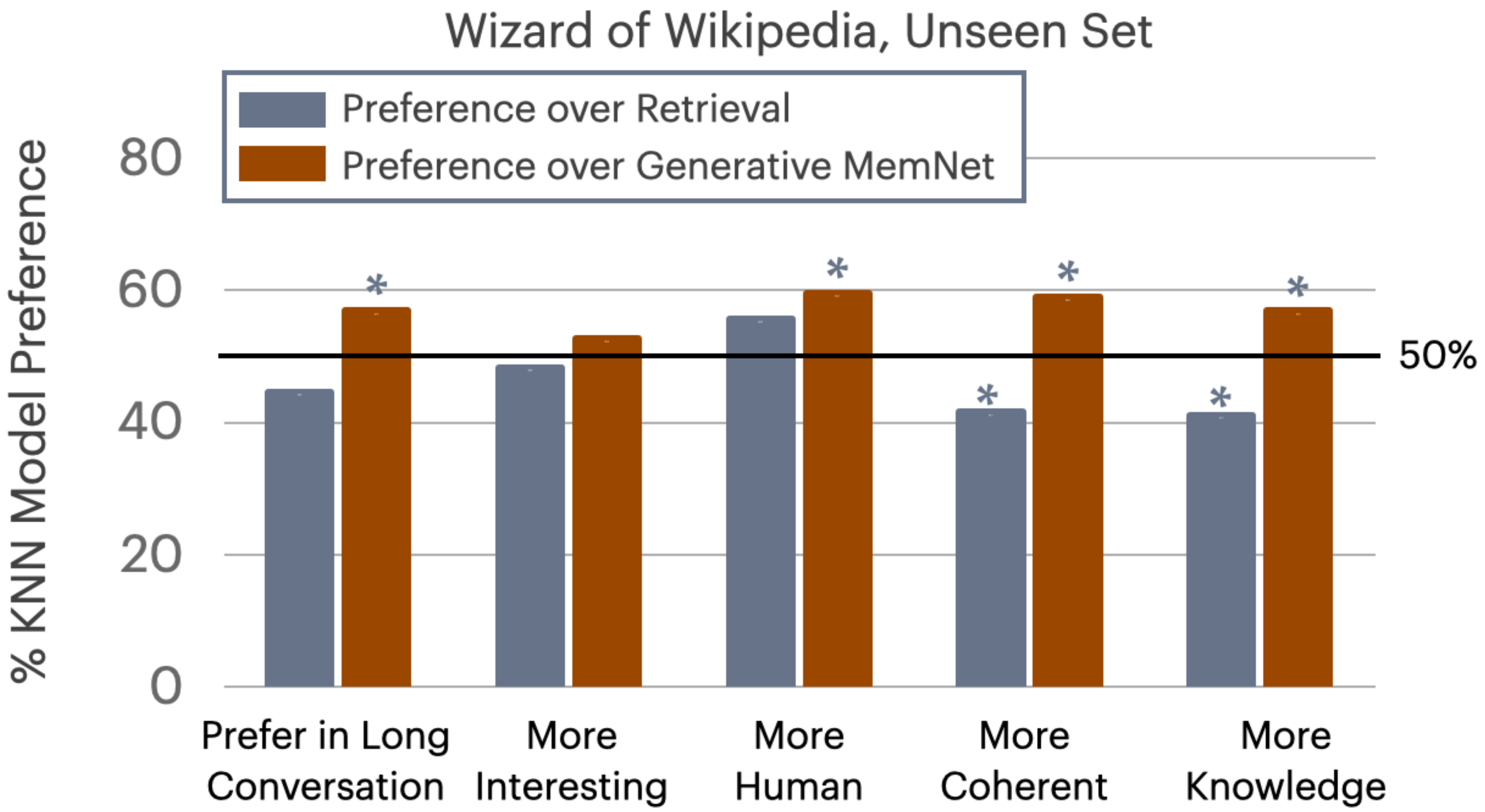
Effect of Fetched Text on Generation

Knowledge	Training Utterance	Generation
buzz lightyear's name is in honor of astronaut edwin 'buzz' aldrin	my favorite character in that book series is hermione granger	cool! my favorite character in that movie is **buzz lightyear**
mr potato head is based on the **real-life mr. potato head** toy	my favorite character in that book series is hermione granger	my favorite character in that movie is **real-life mr potato head**
slinky dog is a toy dachschund with a metal slinky for a body	my favorite character in that book series is hermione granger	cool! my favorite character is the **slinky dog**
slinky dog is a toy dachschund with a metal slinky for a body	**i really like** the character hermione granger	cool! **i really like** slinky dog
slinky dog is a toy dachschund with a metal slinky for a body	**my favorite character** of all time **has to be** hermione granger	i love that movie, **my favorite character has to be** slinky dog the dachshund
slinky dog is a toy dachschund with a metal slinky for a body	i agree with you! that's **my favorite** character as well	i think so too! **my favorite** is slinky

Keeping the template fixed

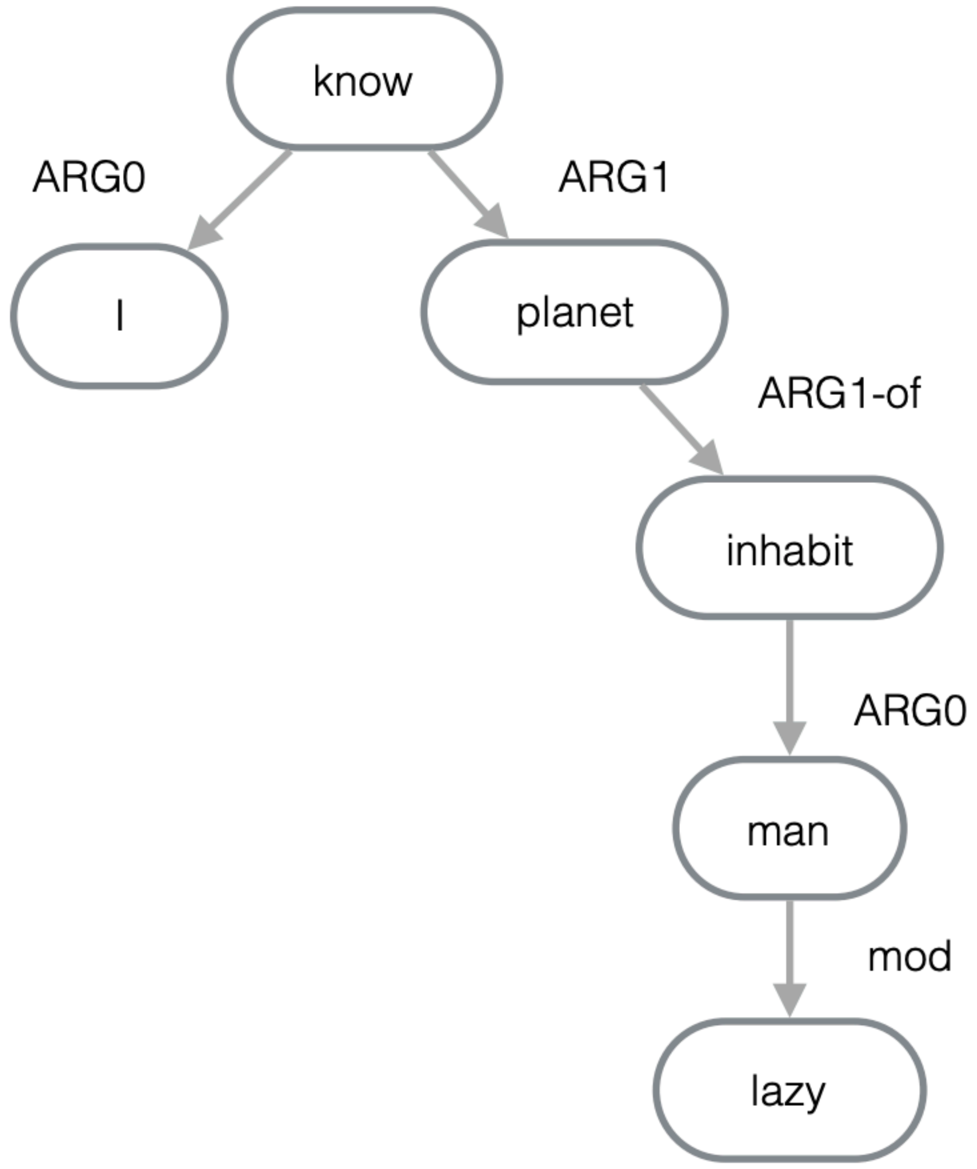
Keeping the KL fixed

Human Evaluation



Generating from AMR Graphs into Multiple Languages

AMR Graph



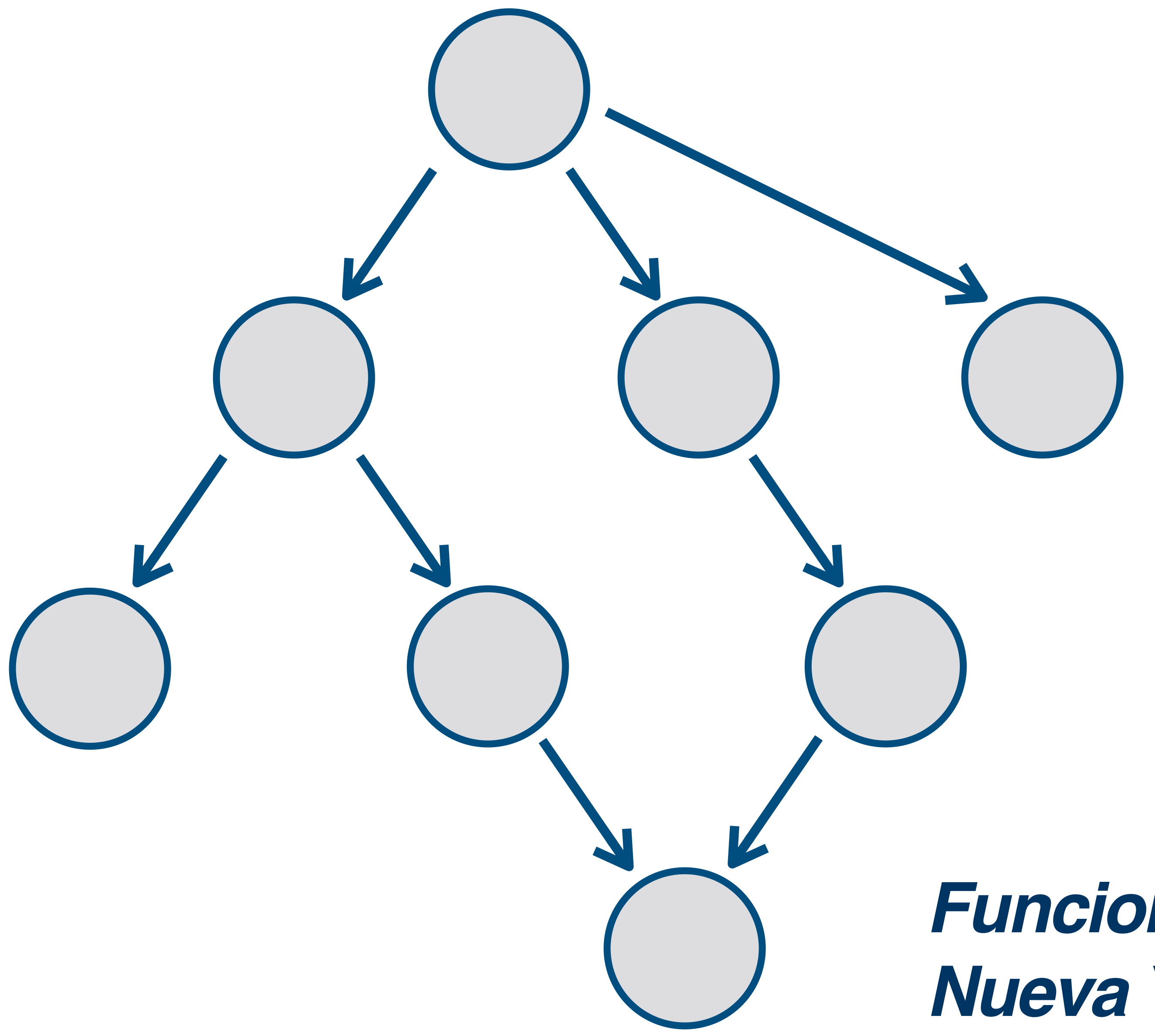
I have known a planet that was inhabited by a lazy man

- Rooted Directed Acyclic Graph
- Nodes: concepts (nouns, verbs, NE, etc.)
- Edges: Semantic Roles

Graphs are frequent data structures

- Knowledge Graphs
- RDF stores
- Tabular data
- Meaning Representations

Graph —> 21 Languages



Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

US officials held an expert group meeting in January 2002 in New York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à Ne York.

Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Romance, Germanic, Slavic, Uralic

Challenges

- Structured Input has a different surface form

Challenges

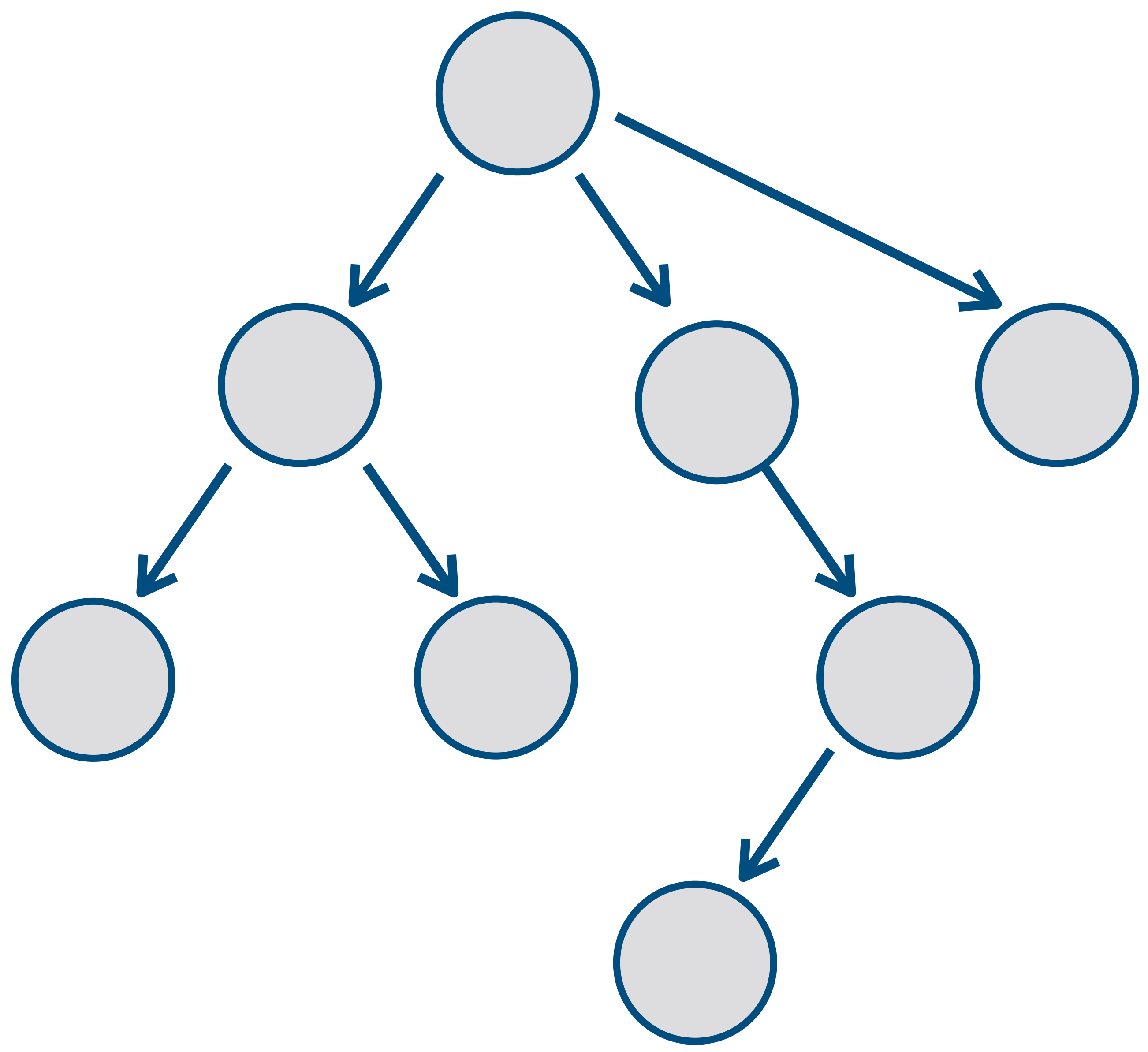
- Structured Input has a different surface form
- Structured Input is underspecified

Challenges

- Structured Input has a different surface form
- Structured Input is often very underspecified
- Multilingual: decoding into languages with varied morphology and word order

Encoder-Decoder MODEL

Graph



hold

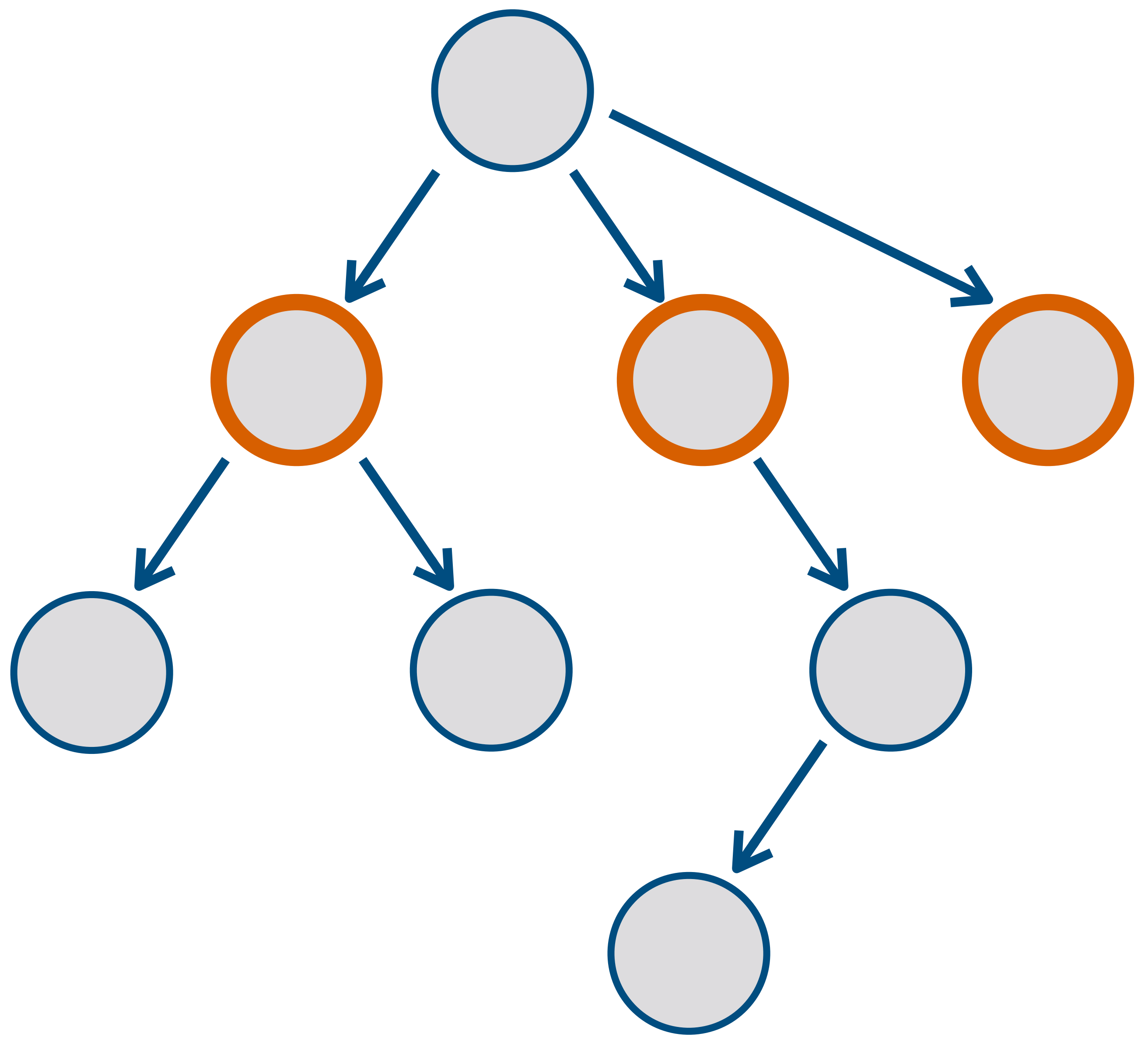
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official

:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group

:time date-entity :year 2002 :month 1

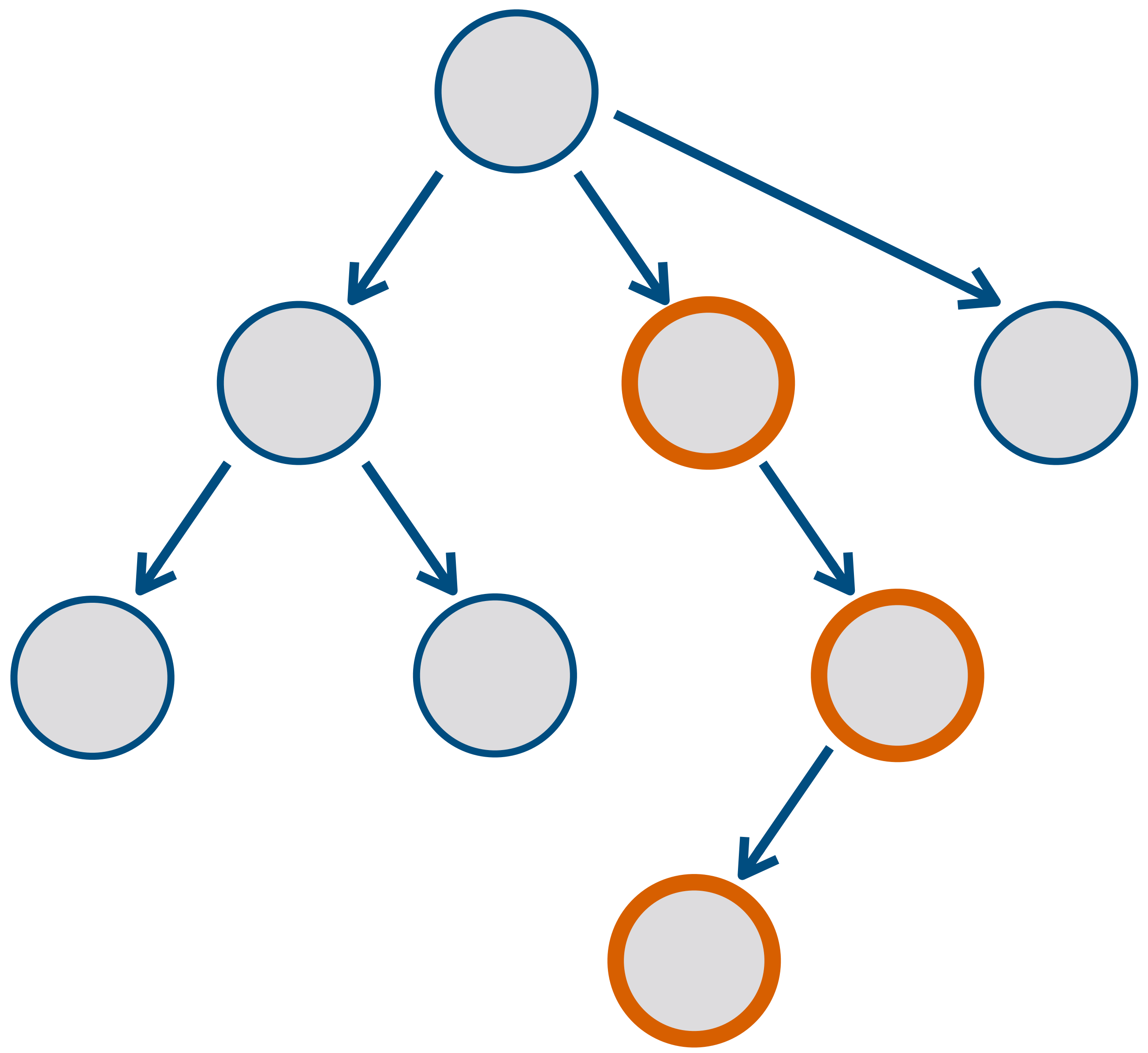
:location city :op1 New :op2 York

Graph Encoding



hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Graph Encoding



hold

:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official

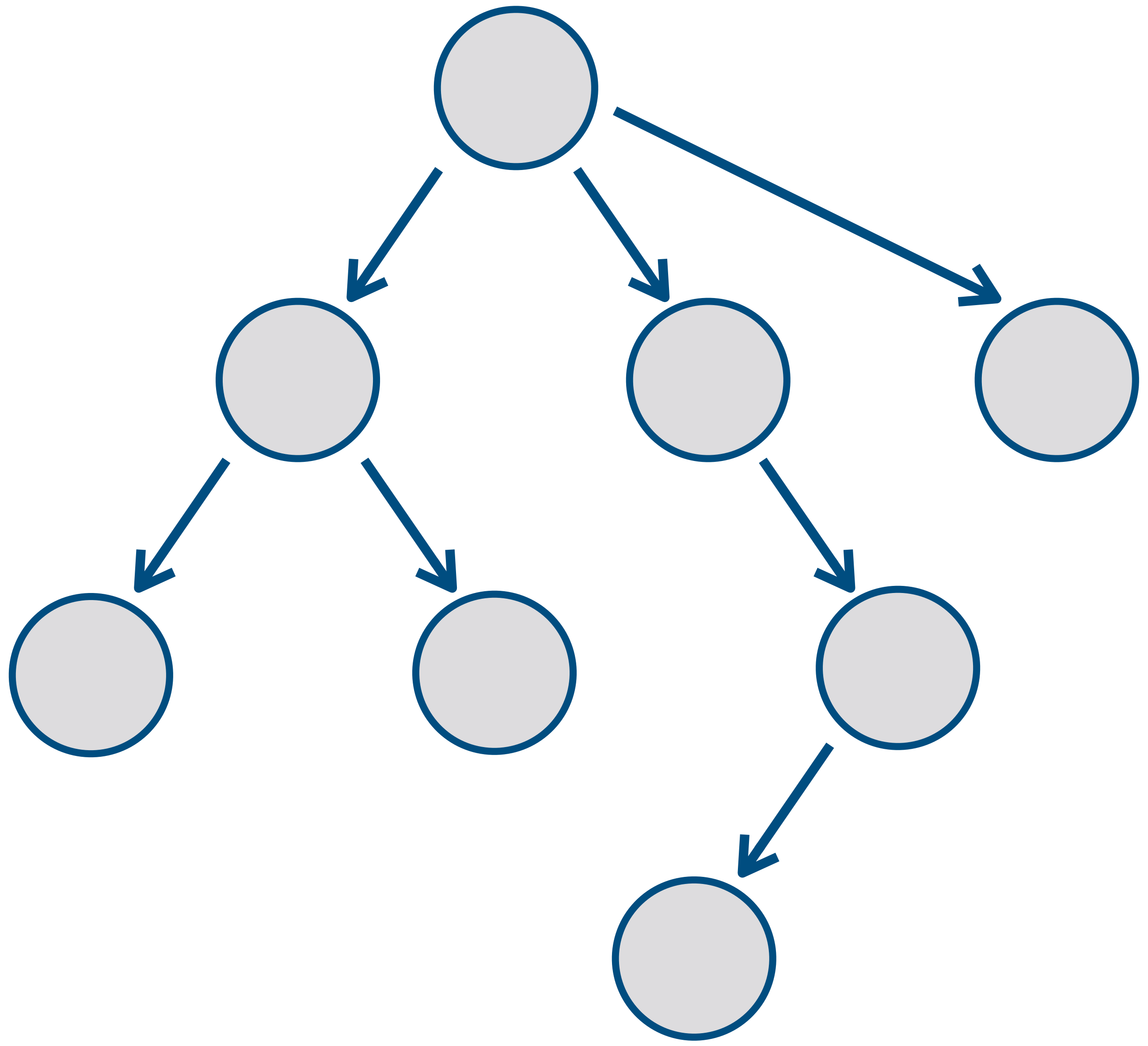
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group

:time date-entity :year 2002 :month 1

:location city :op1 New :op2 York

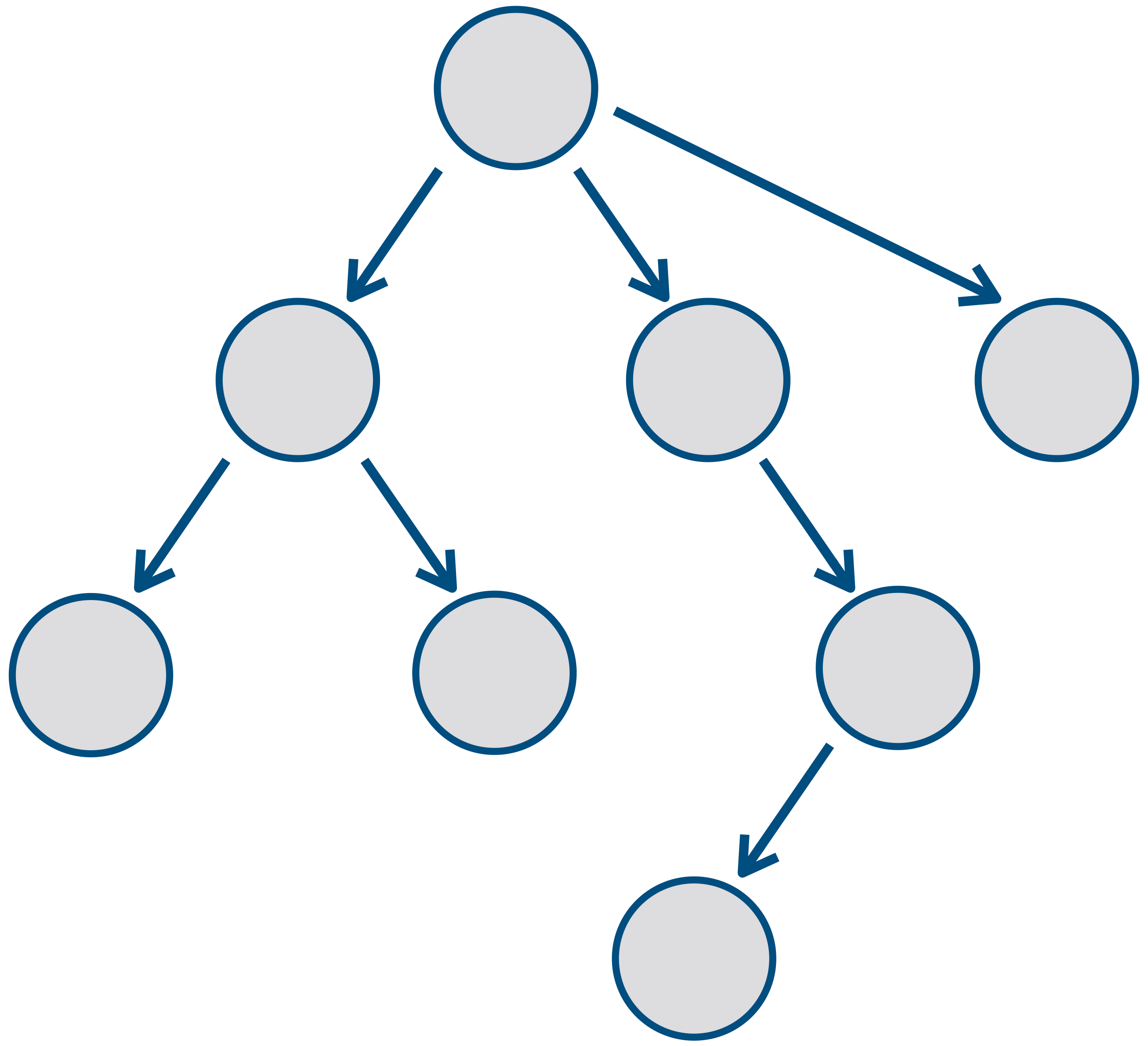
Preprocessing

- Remove variable names and instance-of relation
- No anonymisation
- Sentence piece model with 32K operations



Pretraining

- Pretraining on silver AMRs
  - 30M sentences from CCNET
  - Using JAMR



Decoding into multiple Languages

- XLM cross-lingual embeddings and vocabulary (32K sentence piece subwords)
- Language Model pretraining on 30M sentences
- Multilingual Encoder-Decoder

French

Des responsables américains

….

Spanish

Funcionarios estadounidenses

….

Slovak

Americkí predstavitelia

….

Bulgarian

Американските служители

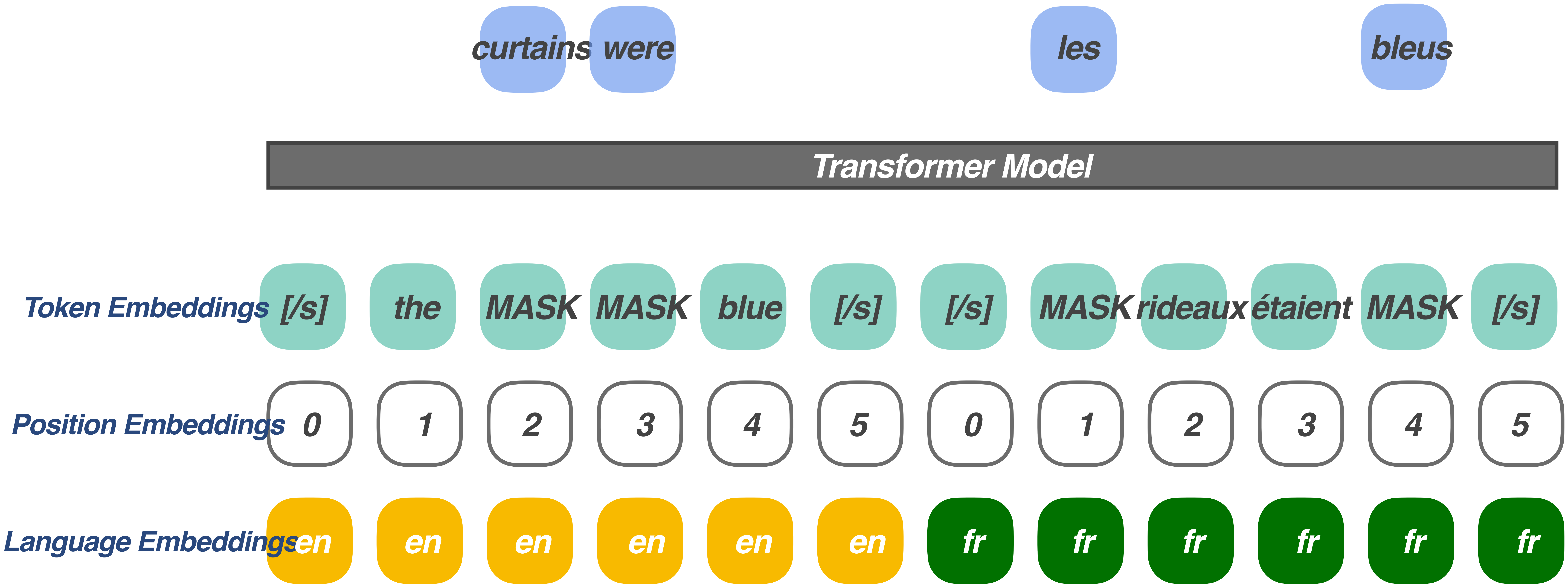
….

Swedish

Amerikanska tjänstemän

…

XLM Cross-lingual embeddings



Cross-lingual Language Model Pretraining
Guillaume Lample, Alexis Conneau

Multilingual Encoder-Decoder

Decoding into Slovak

sv

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

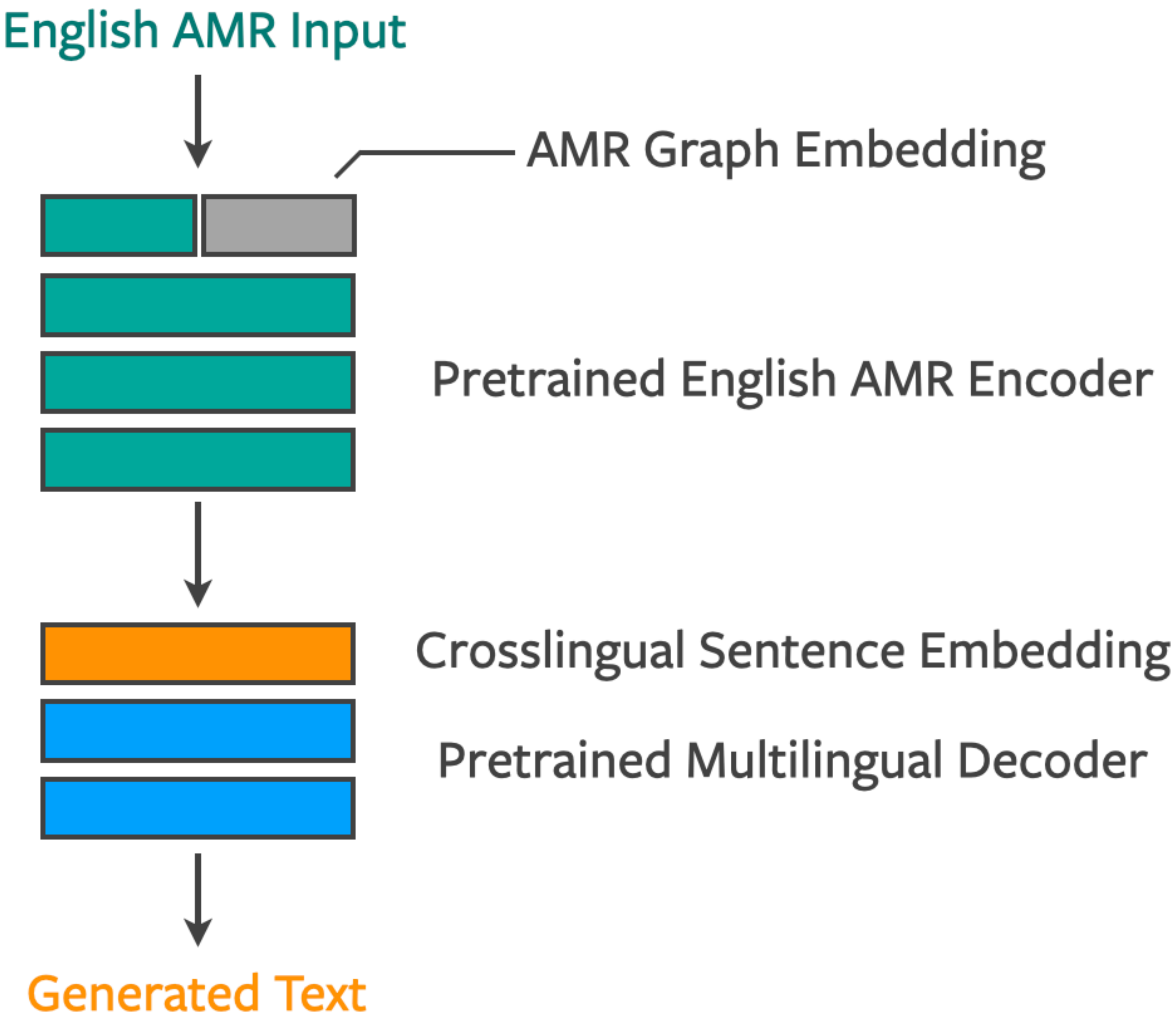
Decoding into French

fr

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Multilingual AMR-to-NL Model



- Encoder: pretraining on Silver AMRs
- Decoder: language model pretraining, multilingual model

DATA

Training Data

- Europarl: 21 Languages
- Construct AMR: create AMR structure with JAMR parser

https://github.com/jflanigan/jamr

Training Data

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

French

Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Spanish

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

Slovak

Американските служители проведоха среща на експертна група през януари 2002 г. в Ню Йорк.

Bulgarian

Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

Swedish

EVALUATION

Evaluation

Automatic (BLEU)

- Ablation
- Comparison with two strong baselines
- Impact of training data (which languages ?)
- Correlation I/O (sub)word overlap and BLEU

Human-Based

- Word-Oder, Morphology, Semantic adequacy, Paraphrasing

Ablation Study

Base Model (English)	32.5
+ Graph embeddings	32.9
+ Crosslingual embeddings.	33.0
+ Encoder pretraining	33.4
+ Decoder pretraining	33.8

https://github.com/jflanigan/jamr

Comparison: Monolingual v. Multilingual

Monolingual Baseline

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparison: Monolingual v. Multilingual

Monolingual Baseline

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

fr

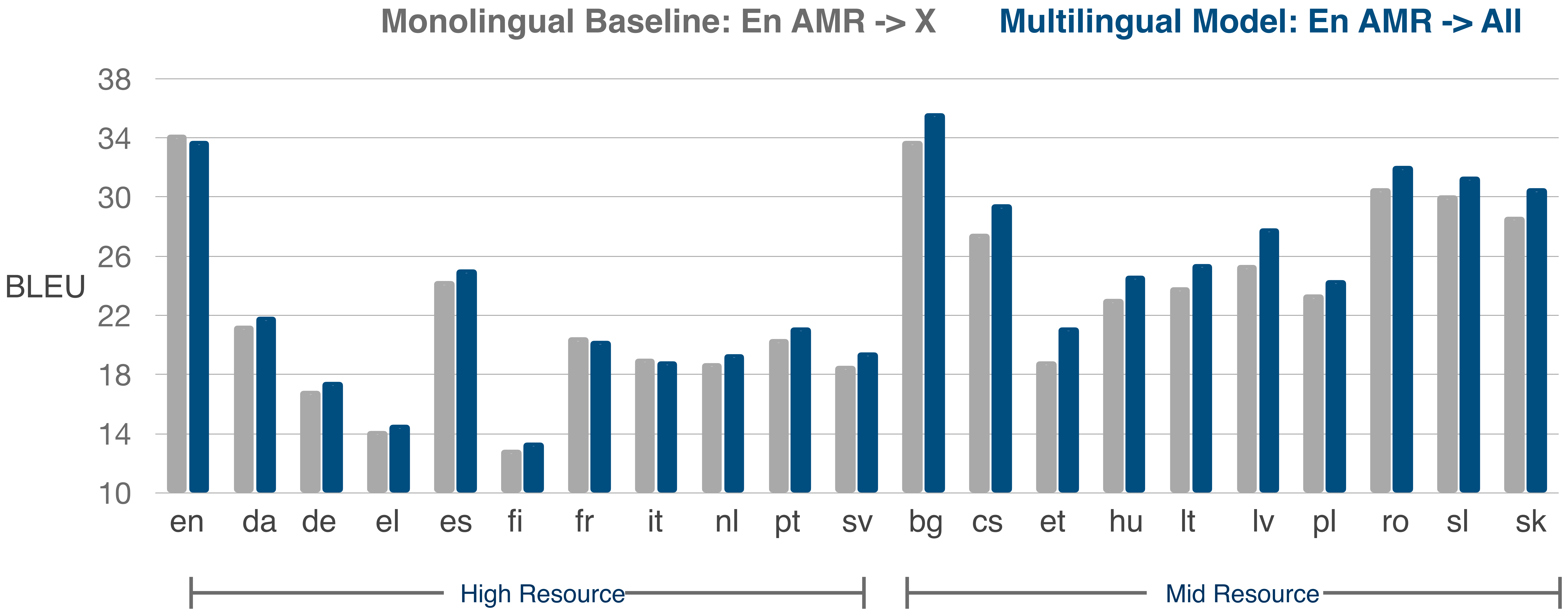
Multilingual Model

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

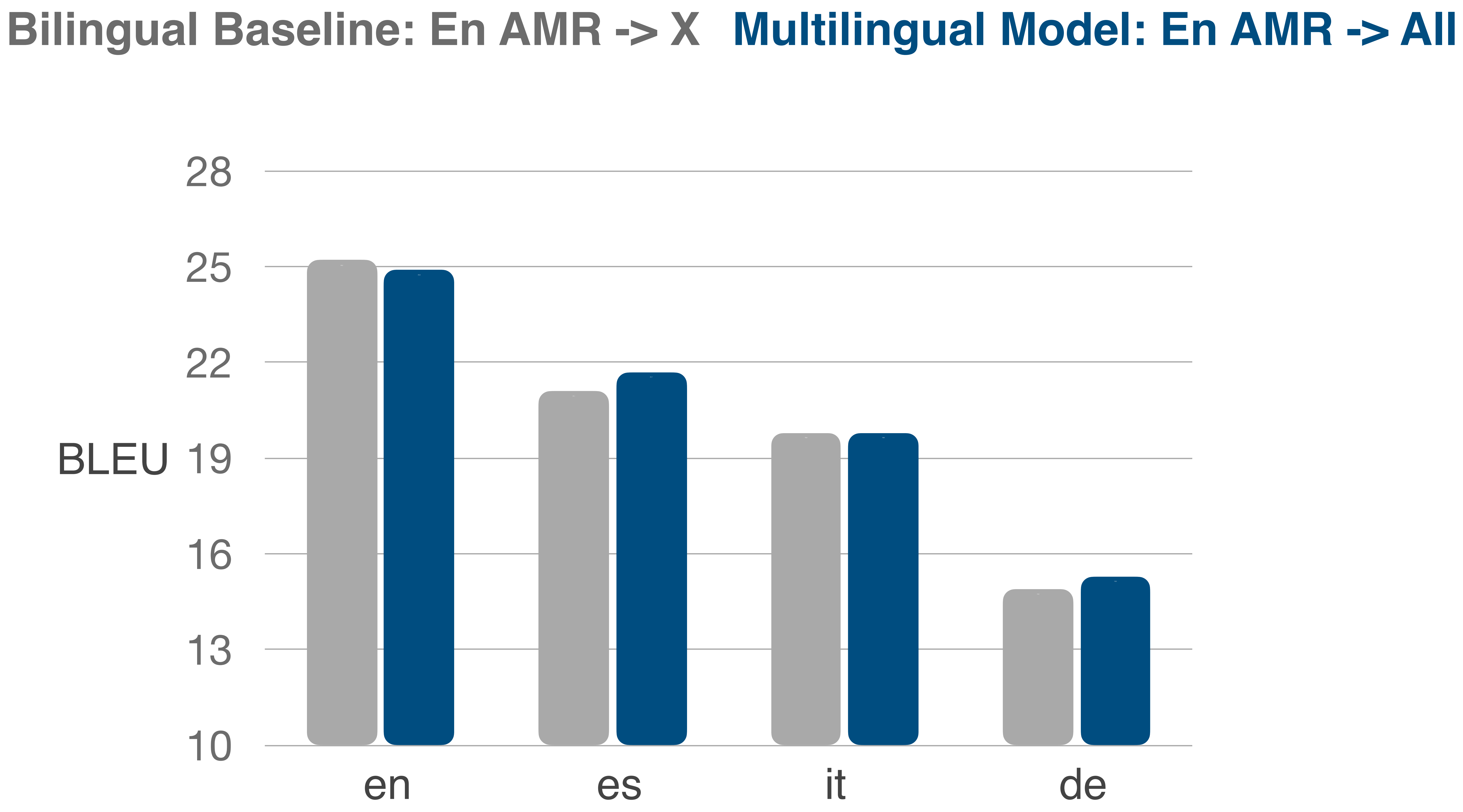
↓

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Results: Europarl



Results: Gold AMR



Comparison: Hybrid Translation v.

Hybrid Translation Model

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

AMR to English

US officials held an expert group meeting in January 2002 in New York.

Translation Model

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Multilingual Model

fr

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparison: Hybrid Translation v. Multilingual

Hybrid Translation Model

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

AMR to English

US officials held an expert group meeting in January 2002 in New York.

Translation Model

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Multilingual Model

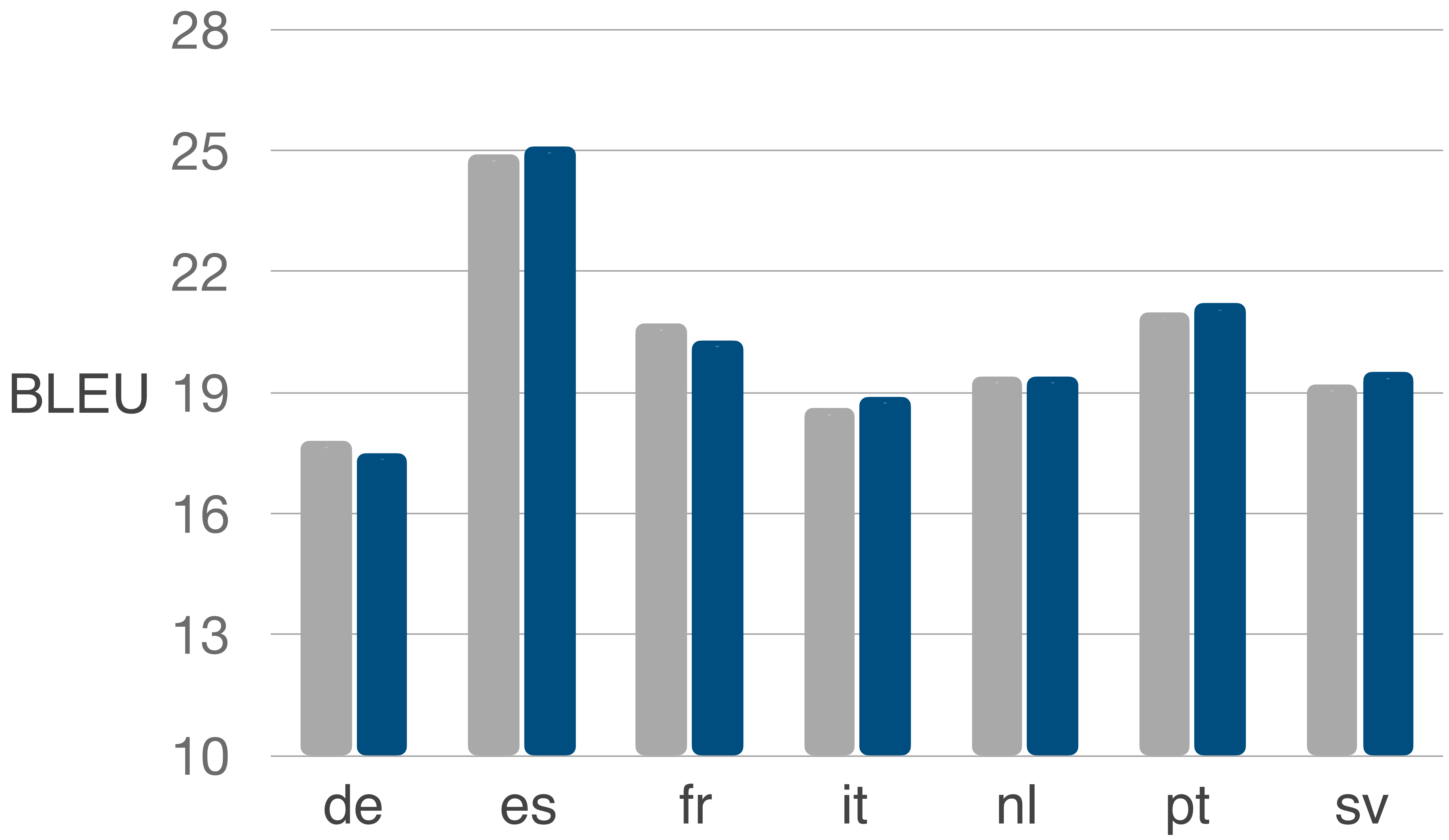
fr

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparison to Hybrid Translation Baseline

Hybrid Translation: En AMR -> En -> Translate to X **Multilingual Model: En AMR -> All**

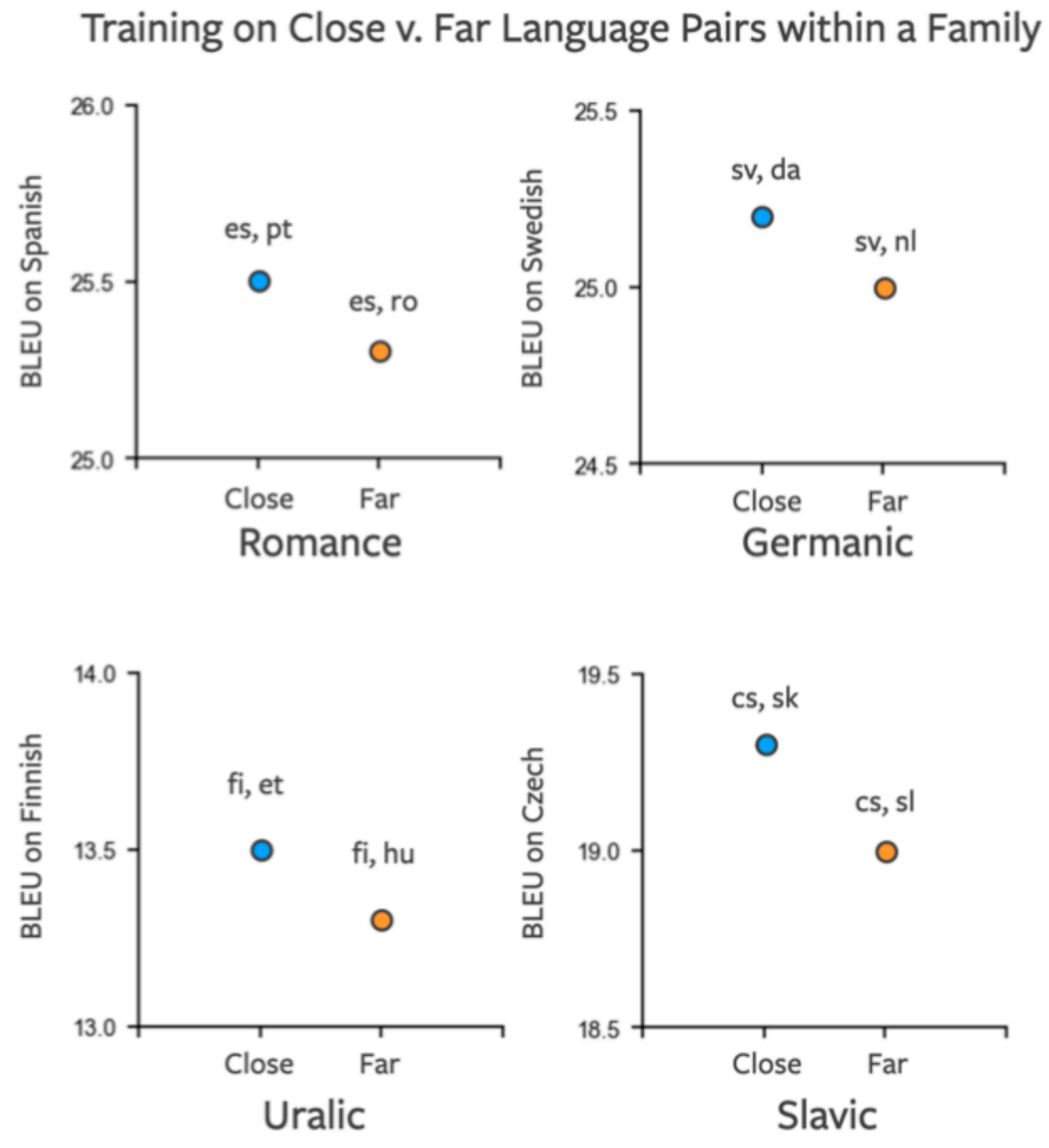


Training on languages from the same family

	Da	De	Nl	Sv
One Language	21.3	17.0	18.5	18.7
Germanic Family	21.8	21.9	19.6	19.3
All Languages	21.9	17.5	19.4	19.5

Training on the closest language

- Multilingual models trained on language pairs
- Within a family, the most closely related pairs get best results
  - Romance: Spanish/Portuguese
  - Germanic: Swedish/Danish
  - Uralic: Finnish/Estonian
  - Slavic: Czech/Slovak



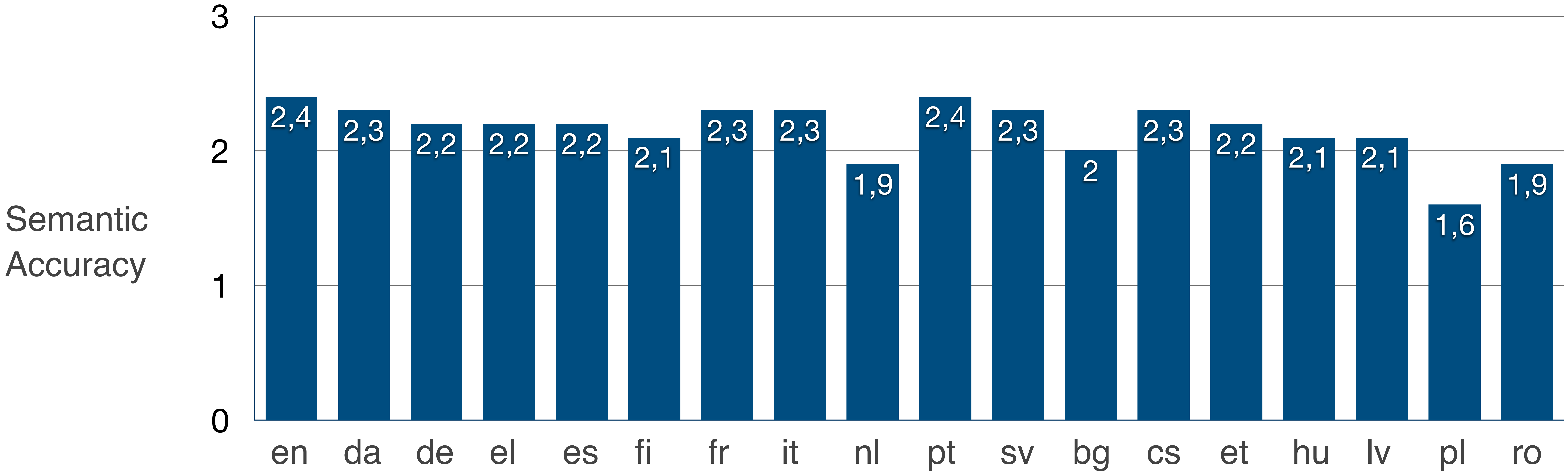
Human Evaluation

- **Semantic Accuracy:**
  Does the hypothesis correctly paraphrase the reference?

- **Morphology:**
  Is the morphology correct? Are agreement constraints e.g., verb/subject, noun/adjective respected?

- **Word Order:**
  Is the word order natural sounding?

Human Evaluation: Semantic Accuracy



Human Evaluation

The scores are uniformly high across languages for both Morphology and Word Order

A Multilingual model generalises well across languages

Example Paraphrases

This point will certainly be the subject of subsequent further debates in the council

This is a point that will undoubtedly be discussed later in the council.

Je ne suis pas favorable à des exceptions à cette règle.

A mon avis, il n’est pas bon de faire des exceptions à cette règle .

Human evaluation demonstrates multilingual techniques generalize across languages

Human evaluation demonstrates multilingual techniques generalize across languages

Multilingual benefits from increased training data and performs better than monilingual

Human evaluation demonstrates multilingual techniques generalize across languages

Multilingual benefits from increased training data and performs better than monilingual

Using English-Centric AMR, we can decode into many different target-side languages

Retrieval-Based Generation of Long Form Text

Generating Woman Biographies

Generating Wikipedia Biographies from Web Retrieval

PERSON NAME

WIKIPEDIA

Joan Paton

Joan Burton Paton AM née Cleland (1916–April 2000) was an Australian teacher, naturalist, environmentalist and ornithologist. One of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected President 1991–1993, she also served as president of the South Australian Ornithological Association (1979-1982). Her father was Professor Sir John Burton Cleland, a notable microbiologist and pathologist who strongly encouraged her early interest in natural history.

Contents

Early life and education

Joan Burton Paton was born in Sydney, New South Wales, the daughter of John Burton Cleland (1878–1971) and his wife, Dora Isabel Paton (1880–1955).[1] She had three sisters, Dr Margaret Burton Cleland, Elizabeth Robson Cleland and Barbara Burton Cleland; and a brother, William Paton 'Bill' Cleland, who became a surgeon. The father encouraged his children's interest in science. Joan Paton was educated at the University of Adelaide, where she majored in organic chemistry and biochemistry. In 1951 she married Erskine Norman Paton (1922–1985), son of Adolph Ernest Paton and Ida Marie Poynton. Their son is Prof David Cleland Paton.[2]

Career

In 1967 Paton became a lecturer on ornithology in South Australia's Workers' Educational Association.[3][4] Among those she inspired to work in ornithology and environmental conservation was Margaret Cameron, who became the President of the Royal Australasian Ornithologists Union (RAOU).[5]

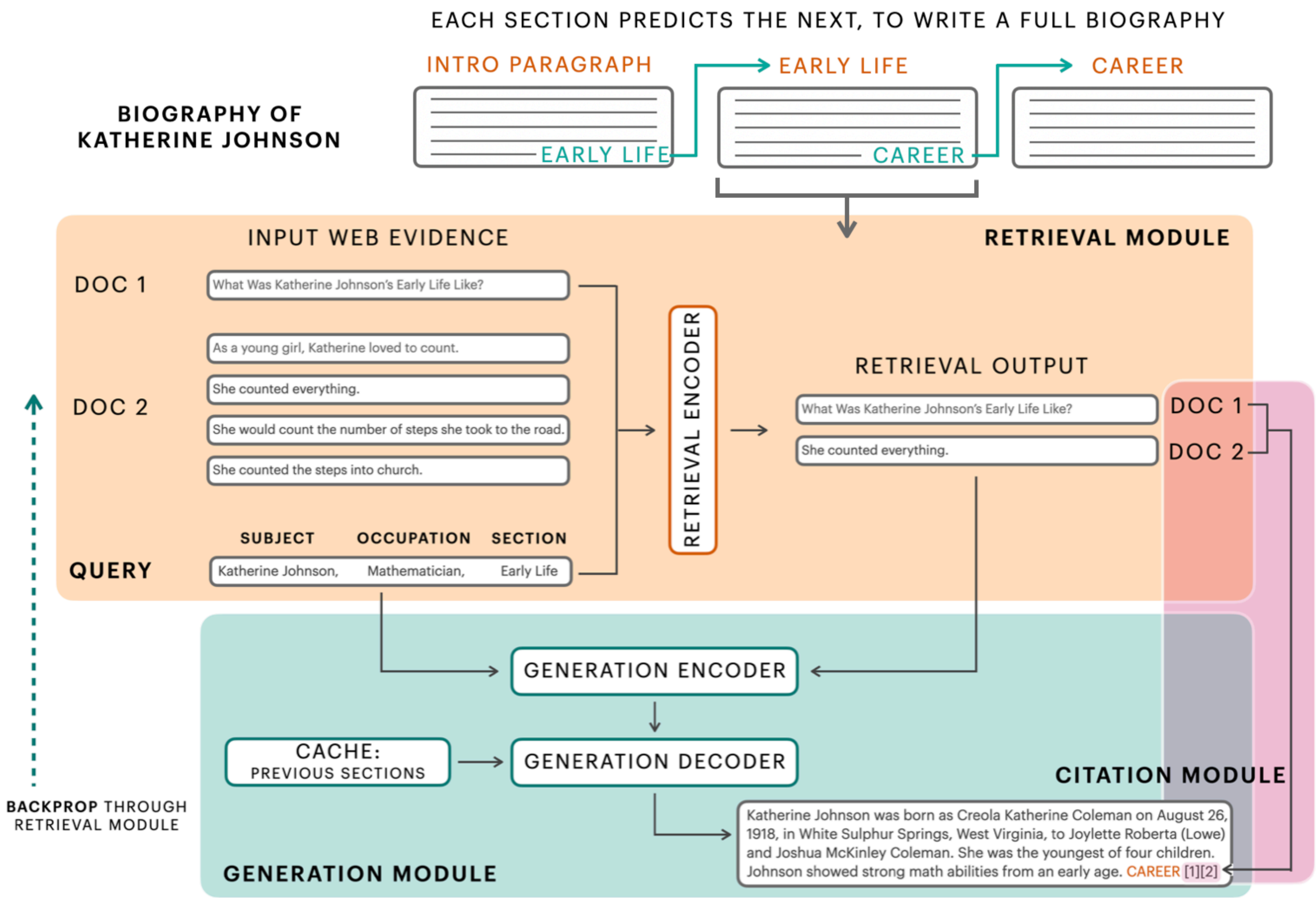
Paton was active in the RAOU, as well as in the South Australian Ornithological Association (SAOA), of which she was elected Vice-President 1974–1979, and President 1979–1982. She was one of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected president (1991-1993).[6]

Legacy and honours

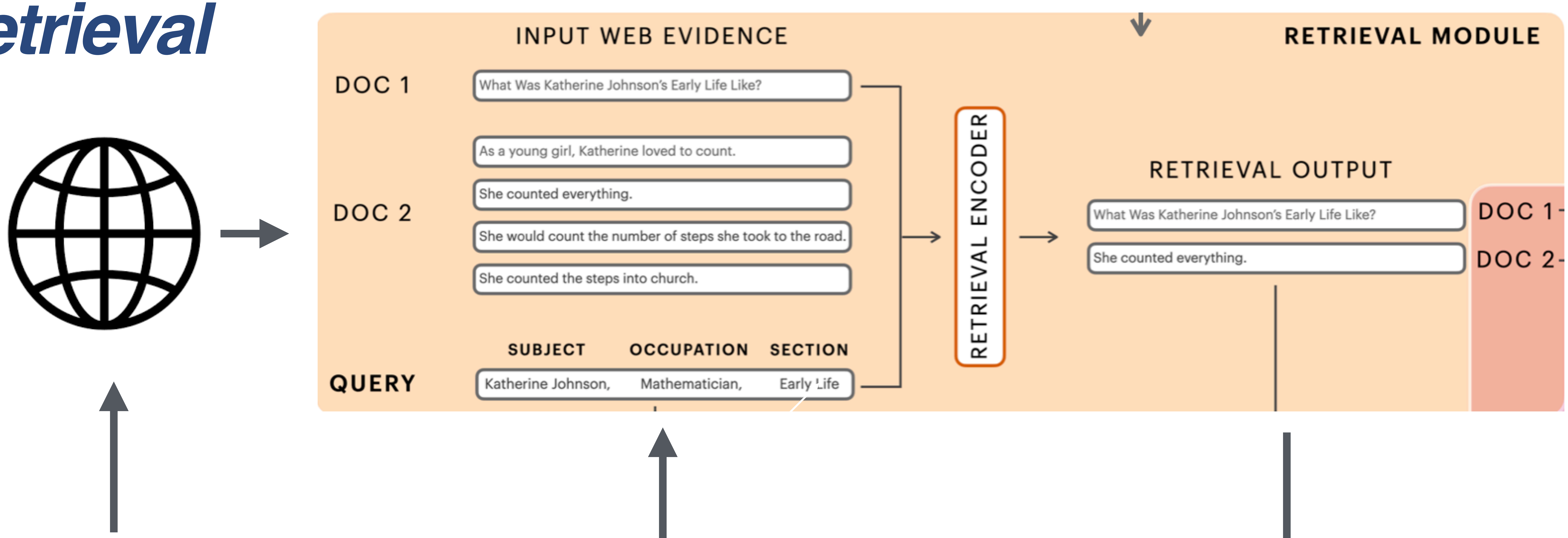
- 1990, she was made an Honorary Member of the SAOA.
- 1996, she was made an Honorary Member of the Adelaide Ornithologists Club.

Challenges

- Gather relevant evidence
- Generate a structured text
- Ensure factuality


EACH SECTION PREDICTS THE NEXT, TO WRITE A FULL BIOGRAPHY
INTRO PARAGRAPH
EARLY LIFE
CAREER
BIOGRAPHY OF KATHERINE JOHNSON
EARLY LIFE
CAREER
INPUT WEB EVIDENCE
RETRIEVAL MODULE
DOC 1
What Was Katherine Johnson's Early Life Like?
DOC 2
As a young girl, Katherine loved to count.
She counted everything.
She would count the number of steps she took to the road.
She counted the steps into church.
RETRIEVAL ENCODER
RETRIEVAL OUTPUT
What Was Katherine Johnson's Early Life Like?
She counted everything.
DOC 1
DOC 2
SUBJECT
OCCUPATION
SECTION
QUERY
Katherine Johnson,
Mathematician,
Early Life
GENERATION ENCODER
CACHE:
PREVIOUS SECTIONS
GENERATION DECODER
CITATION MODULE
BACKPROP THROUGH RETRIEVAL MODULE
Katherine Johnson was born as Creola Katherine Coleman on August 26, 1918, in White Sulphur Springs, West Virginia, to Joylette Roberta (Lowe) and Joshua McKinley Coleman. She was the youngest of four children. Johnson showed strong math abilities from an early age. CAREER [1][2]
GENERATION MODULE

Retrieval



QUERY

Katherine Johnson

Mathematician

Early Life

SEARCH OUTPUT

Top 20 search results segmented into sentences

OUTPUT

40 sentences most similar with the query

(1,000 words)

Generation

QUERY

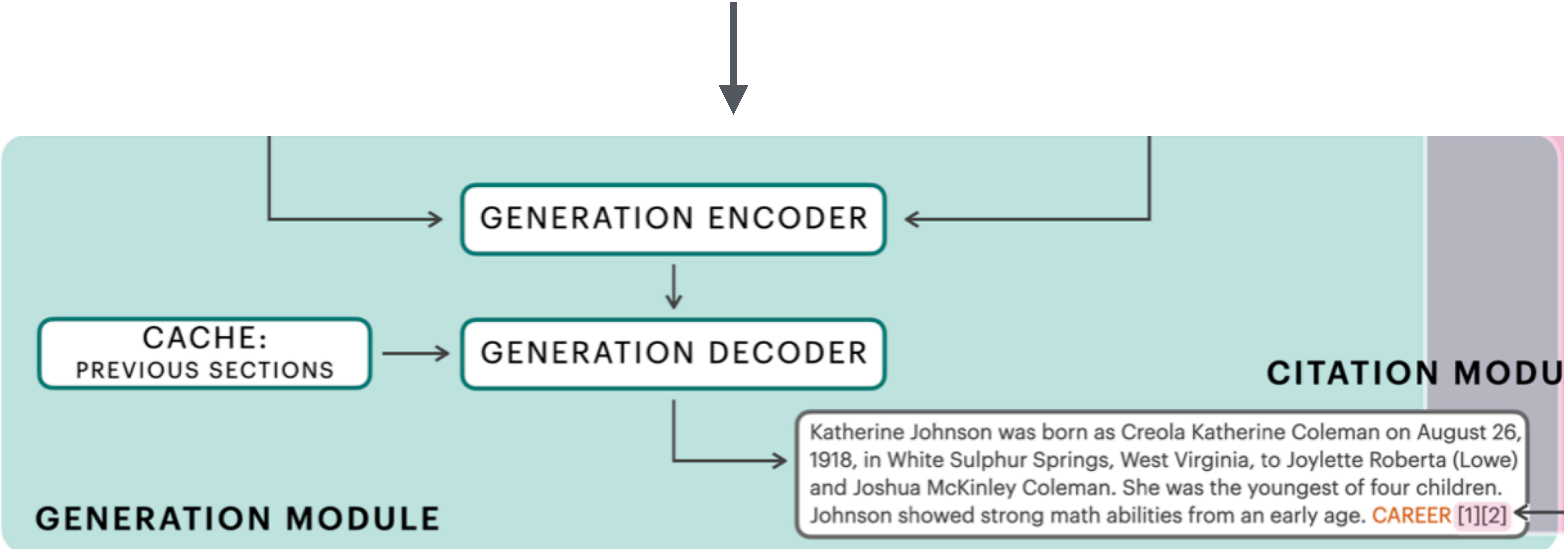
Katherine Johnson

Mathematician

Early Life

RETRIEVED EVIDENCE

1,000 words



Transformer-XL Cache Mechanism

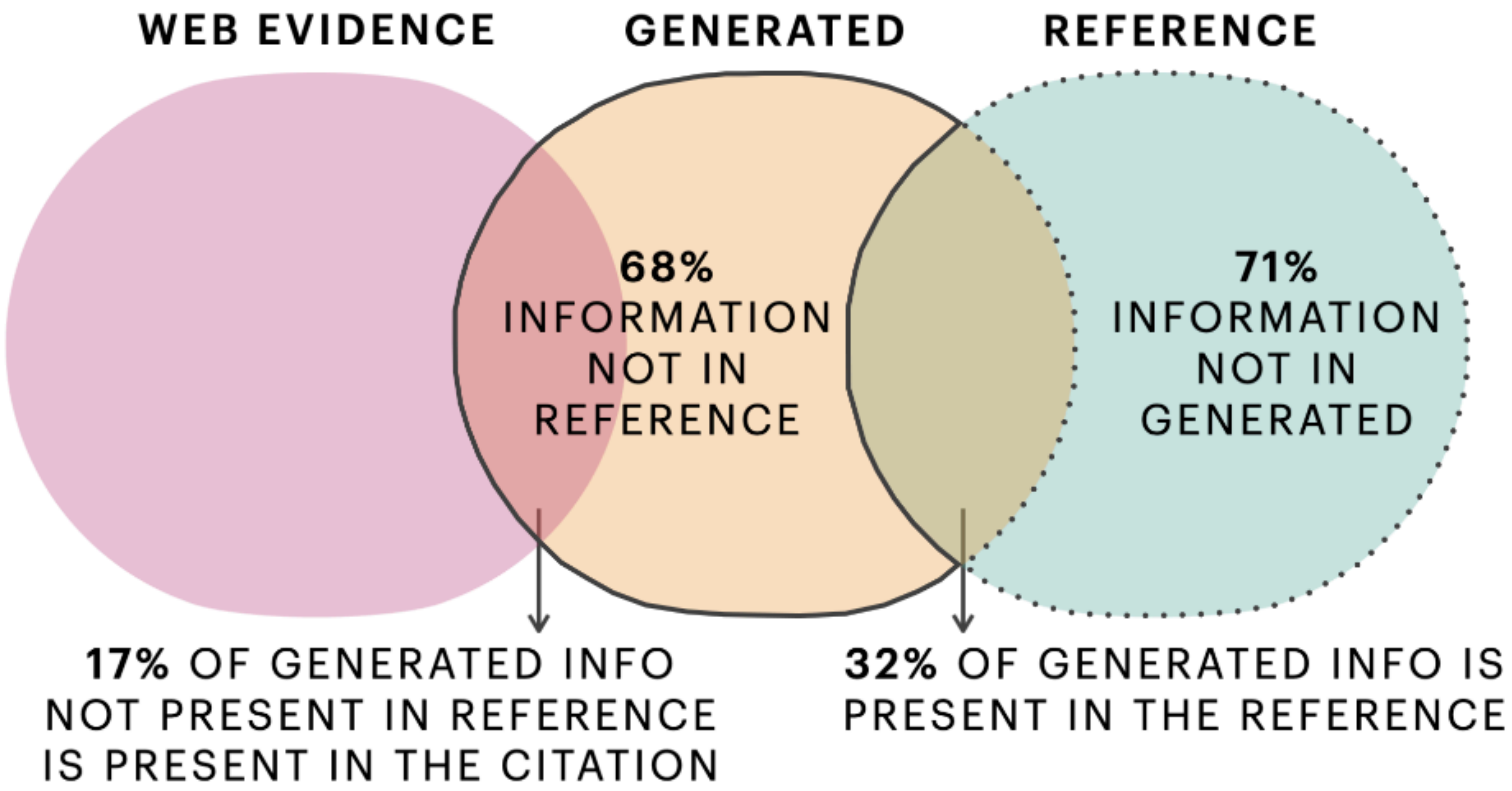
- Caches the previous section’s hidden states at every later
- Usd as a memory to generate the current section

Ablation

Model	ROUGE-L	Entailment	Named Entity Coverage
BART Pretraining + Finetuning	17.4	15.8	21.9
+ Retrieval Module	18.8	17.2	23.1
+ Caching Mechanism	19.3	17.9	23.4

The retrieval and the cache module statistically significantly improve results

Human Evaluation of Factuality



The Evidence Gap

Data

(person name, web evidence, Wikipedia biography)

- Wikisum: Wikipedia biographies
- Our dataset: Women biographies

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Less Web Evidence, Less Good Texts

Model	**WikiSum Test**	**Women**	**Scientists**	**Women in Asia**	**Women in Africa**
BART Pretraining	19.0	17.4	18.2	16.7	16.4
+ Retrieval	21.4	18.8	19.3	17.9	17.1
+ Caching	21.8	19.3	19.7	18.4	17.3

Example output

toplevel ashley mckenzie is a canadian film director, screenwriter and producer. she is the winner of the stella artois jay scott prize for emerging talent at the 2016 toronto international film festival. her first feature film, werewolf, premiered at the toronto film festival in 2016. she has also directed short films for the national film board of canada and the canadian screen actors guild. she was born in montreal, quebec, canada, and grew up in ottawa, ontario. [1,3,11,13,14]

=personal life= mckenzie was born in london, england. she is the daughter of alexander mckenzie, who was a member of the british rock band the beatles. she has a younger sister, jessica, who is also a singer. she was educated at st mary 's college, oxford, where she graduated with a bachelor of arts degree in english literature. she also studied at the university of london. she married fellow x factor contestant andrew davies in september 2006. they have two children, a son and a daughter. [3,4,7,8,10,11]

=career= mckenzie was a contestant on the third series of the x - factor in 2006. she was eliminated in the first week of the competition. in 2007, mckenzie released her debut single "don 't pretend you hadn' t, now..." which peaked at no .160; 2 on the uk singles chart. she also released a second single ," i 'm not afraid ", in 2008. in 2009, she released her third single ," don' t pretend you haven 't, now ". in 2010, she was a judge on the x factor uk. [2]

Conclusion

Open Challenges

- ***Factuality, faithfulness to the input*** (evaluation and implementation)
- Multilingual NLG: generating from and into multiple languages Document level Simplification
- Multi-document, multi-format, summarisation
- Domain adaptation, style transfer

Thank you !