A guided tour of contextual word representations for language understanding

Matthew Peters

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

About me

Allen Institute for AI

AI for the Common Good.

Our mission is to contribute to humanity through high-impact AI research and engineering.

AllenNLP

Open Source NLP Platform

Design, evaluate, and contribute new models on our open-source PyTorch-backed NLP platfom, where you can also find state-of-the-art implementations of several important NLP models and tools.

Learn More →

Aristo

Systems That Read and Reason

The Aristo Project aims to build systems that demonstrate a deep understanding of the world, integrating technologies for reading, learning, reasoning, and explanation.

Learn More →

Mosaic

Common Sense for AI

The Mosaic team seeks to define, develop, and improve common sense for AI — an important, fundamental skill required to go beyond the narrow and brittle AI applications we have today.

Learn More →

PRIOR

Computer Vision

The Perceptual Reasoning and Interaction Research (PRIOR) team seeks to create AI systems that can see, explore, learn, and reason about the world.

Learn More →

Semantic Scholar

Semantic Literature Search

Combining NLP, data mining, and computer vision to create a rich academic search experience that helps scientists discover and understand research papers more efficiently than ever.

Learn More →

AI & Fairness

Applying AI for Good

Building on AI2's expertise in NLP, computer vision, and engineering, we seek to deliver a tangible positive impact on fairness.

Learn More →

Incubator

Launching AI-Powered Startups

The incubator combines AI2's world class engineering and research organization with proven business leaders to bring innovative, AI-powered ideas to life.

Learn More →

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Why Natural Language Processing?

Natural language is one of the primary ways humans communicate, and is a cornerstone of human intelligence.

Why Natural Language Processing?

Natural language is one of the primary ways humans communicate, and is a cornerstone of human intelligence.

There exists a huge amount of digitized text in many languages → easy to process with computer.

Why Natural Language Processing?

Natural language is one of the primary ways humans communicate, and is a cornerstone of human intelligence.

There exists a huge amount of digitized text in many languages → easy to process with computer.

Natural Language Processing (NLP) is an interdisciplinary field combining computational linguistics, computer science, machine learning, AI, mathematics, statistics, and others.

Why Natural Language Processing?

Natural language is one of the primary ways humans communicate, and is a cornerstone of human intelligence.

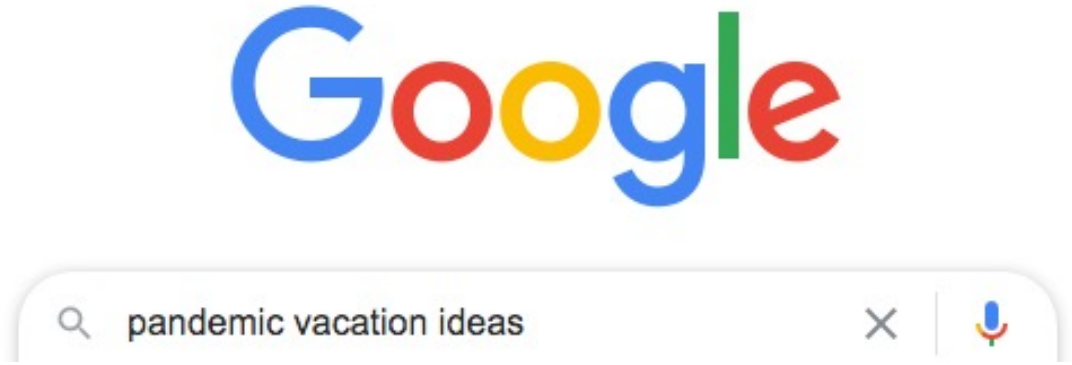
There exists a huge amount of digitized text in many languages → easy to process with computer.

Natural Language Processing (NLP) is an interdisciplinary field combining computational linguistics, computer science, machine learning, AI, mathematics, statistics, and others.

Long term goal: build a language understanding agent that can understand and communicate with humans in the same way humans communicate with each other.

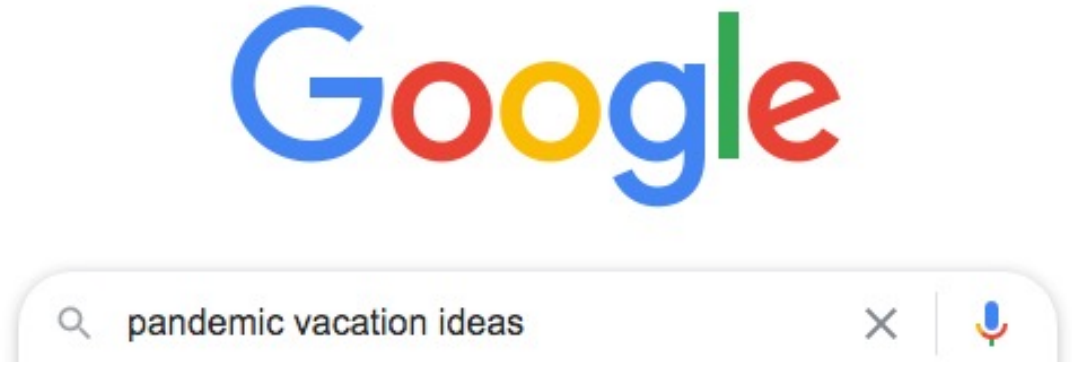
NLP Applications

NLP Applications



Search & Ad ranking

NLP Applications



Search & Ad ranking

---

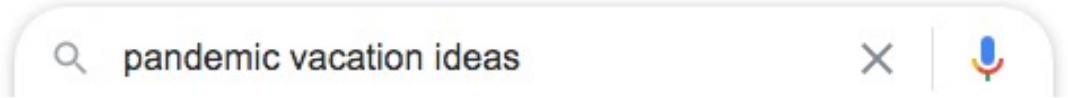
salesforce

"I like the clothes but they are too expensive." 👎

Sentiment classification

NLP Applications

Google



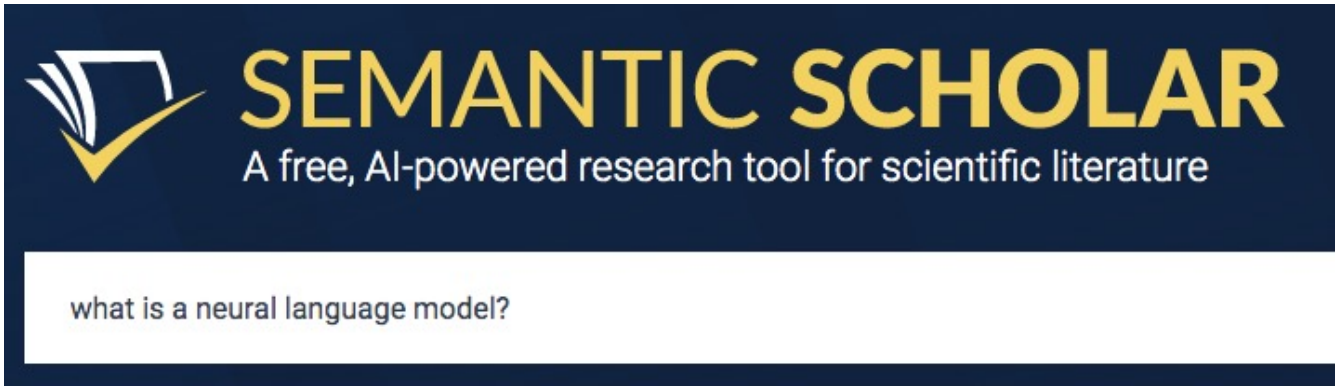
Search & Ad ranking

---

salesforce

"I like the clothes but they are too expensive." 👎

Sentiment classification



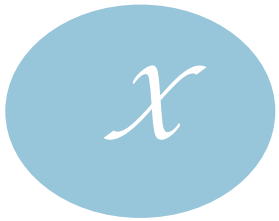
Question answering

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Supervised learning

Supervised learning

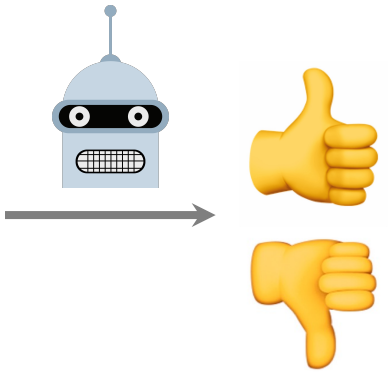
Turns potentially forgettable formula into something strangely diverting. [1]



[1] an example from Stanford Sentiment Treebank (SST, Socher et al 2013)

Supervised learning

Turns potentially forgettable formula into something strangely diverting. [1]



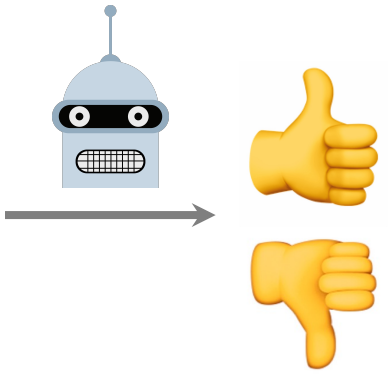
x

y

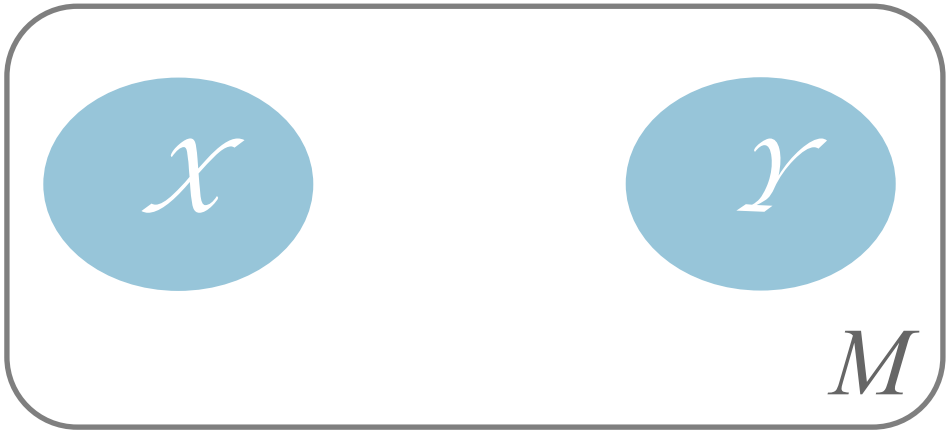
[1] an example from Stanford Sentiment Treebank (SST, Socher et al 2013)

Supervised learning

Turns potentially forgettable formula into something strangely diverting. [1]



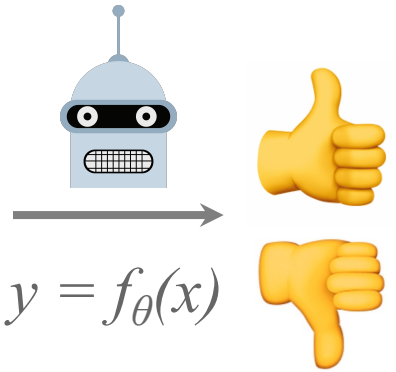
$$D = \{(x_i, y_i),\ i = 1,\ \ldots,\ M\}$$



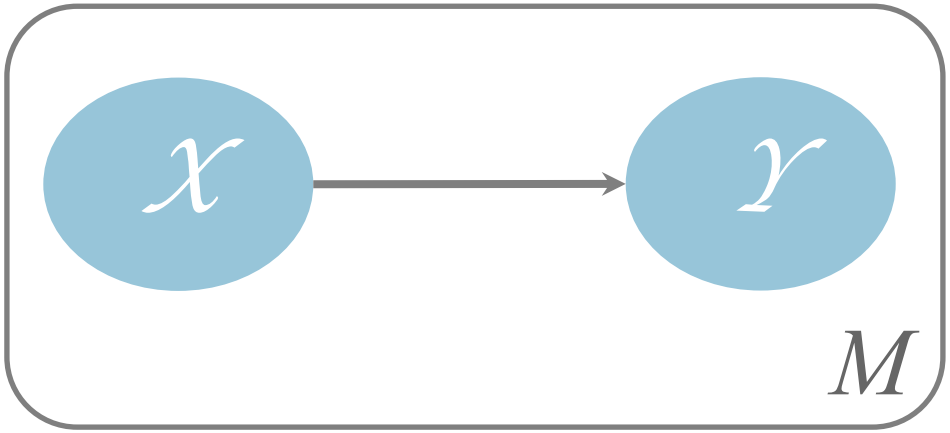
[1] an example from Stanford Sentiment Treebank (SST, Socher et al 2013)

Supervised learning

Turns potentially forgettable formula into something strangely diverting. [1]



$$y = f_\theta(x)$$



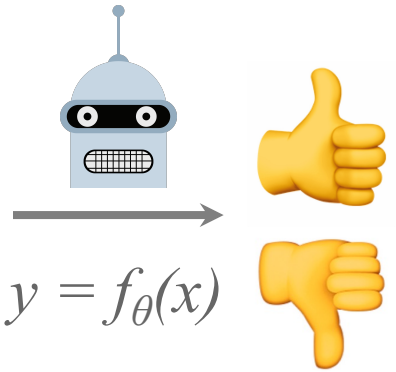
$$D = \{(x_i, y_i),\ i = 1,\ \ldots,\ M\}$$

$$p(y \mid x) = f_\theta(x)$$

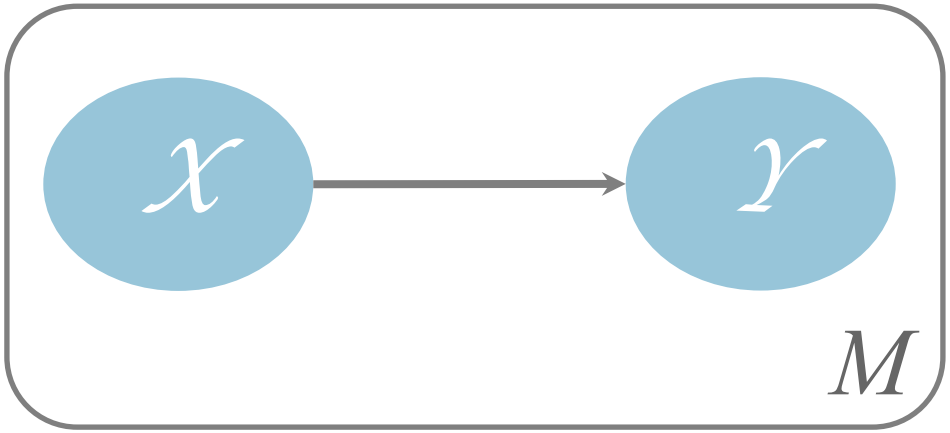
[1] an example from Stanford Sentiment Treebank (SST, Socher et al 2013)

Supervised learning

Turns potentially forgettable formula into something strangely diverting. [1]



$y = f_\theta(x)$



$$D = \{(x_i, y_i),\ i = 1,\ \ldots,\ M\}$$

$$p(y \mid x) = f_\theta(x)$$

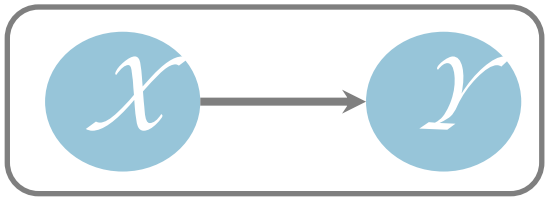
$$\theta = \text{argmax}\ L(D, \theta)$$

$$L(D, \theta) = \Sigma_i\ (y_i \log f_\theta(x) - (1 - y_i) \log (1 - f_\theta(x)))$$

[1] an example from Stanford Sentiment Treebank (SST, Socher et al 2013)

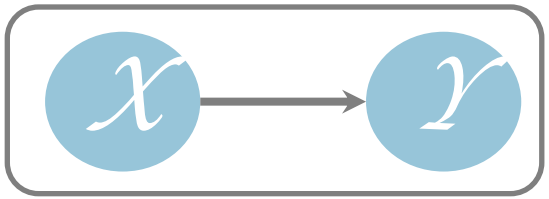
Word representations

One of the central challenges in NLP is converting text to a form that is usable in a statistical learning algorithm.



Word representations

One of the central challenges in NLP is converting text to a form that is usable in a statistical learning algorithm.

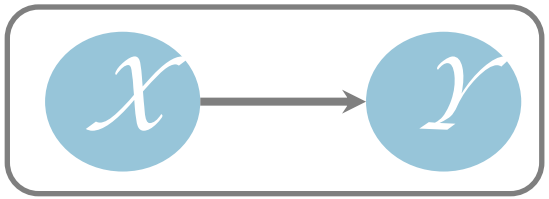


For classification over C classes, we can decompose $f_\theta(x)$ into two functions:

- embedder: represents text as a vector, $e(text) = \boldsymbol{x} \in \mathbb{R}^N$
- predictor: $g(\boldsymbol{x}) = p(y \mid x) \in \mathbb{R}^C$

Word representations

One of the central challenges in NLP is converting text to a form that is usable in a statistical learning algorithm.



For classification over *C* classes, we can decompose $f_\theta(x)$ into two functions:

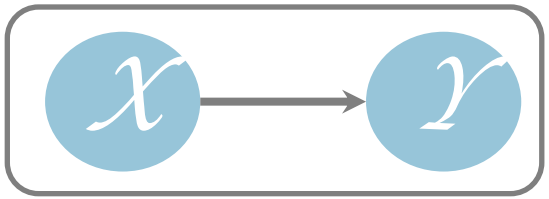
- embedder: represents text as a vector, $e(text) = \boldsymbol{x} \in \mathbb{R}^N$
- predictor: $g(\boldsymbol{x}) = p(y \mid x) \in \mathbb{R}^C$

Desired properties of embedder / encoder:

- Syntax, e.g. “Bob saw Alice” vs. “Alice saw Bob”
- How context influences word meaning, e.g. “play a game” vs “attend a play”
- Commonsense / World knowledge
- and many more

Word representations

One of the central challenges in NLP is converting text to a form that is usable in a statistical learning algorithm.



For classification over *C* classes, we can decompose $f_\theta(x)$ into two functions:

- embedder: represents text as a vector, $e(text) = \boldsymbol{x} \in \mathbb{R}^N$
- predictor: $g(\boldsymbol{x}) = p(y \mid x) \in \mathbb{R}^C$

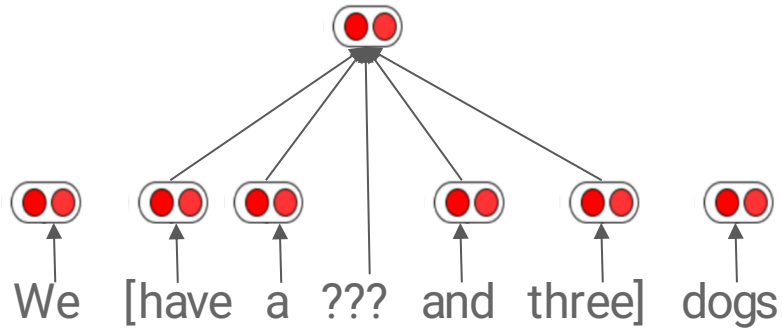
Desired properties of embedder / encoder:

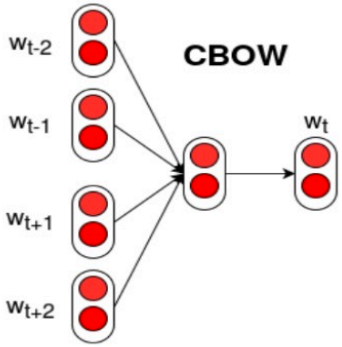
- Syntax, e.g. “Bob saw Alice” vs. “Alice saw Bob”
- How context influences word meaning, e.g. “play a game” vs “attend a play”
- Commonsense / World knowledge
- and many more

Key idea: *learn the embedder* by optimizing one objective on unlabeled text, and transfer / reuse it for a different end task.

Word vectors

Word vector approaches like word2vec define a fixed vocabulary of words, and learn one vector $w \in \mathbb{R}^N$ for each word.



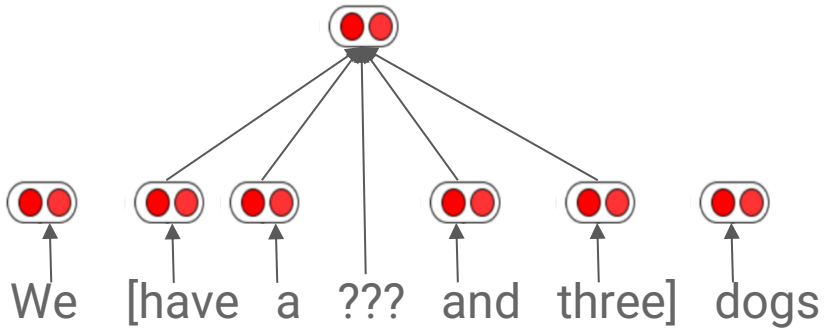


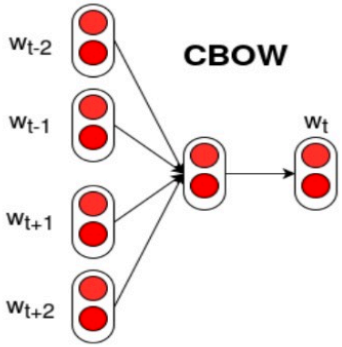
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_t|w_{t+j})$$

word2vec: Mikolov et al (2013)

Word vectors

Word vector approaches like word2vec define a fixed vocabulary of words, and learn one vector $w \in \mathbb{R}^N$ for each word.





$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_t|w_{t+j})$$

MIT Technology Review

Artificial intelligence

King – Man + Woman = Queen: The Marvelous Mathematics of Computational Linguistics

word2vec: Mikolov et al (2013)

Outline for rest of talk

- ELMo: first embedding approach to learn contextual word vectors that satisfies many of these properties. Maximizes log likelihood *p(x)* with a recurrent neural network.
- Analyzing and understanding aspects of language meaning encoded in word representations
- Knowledge enhanced contextual word representations: adding sparse, structured human curated knowledge
- Longformer: extending to long sequences
- Beyond supervised learning

ELMo

Contextual word vectors

Language understanding requires context, but word vectors must compress all contexts into a *single vector.*

Nearest neighbor vectors to “play”:

VERB	NOUN	ADJ	??
playing	game	multiplayer	plays
played	games		Play
	players		
	football		

Using 840B.300d GloVe vectors, Pennington et al (2014).

Contextual word vectors

Compute contextual vector:

$$\boldsymbol{c}_k = f(w_k \mid w_1, \ldots, w_n) \in \mathbb{R}^N$$

f(play | Elmo and Cookie Monster play a game .)

$$\neq$$

f(play | The Broadway play premiered yesterday .)

ELMo

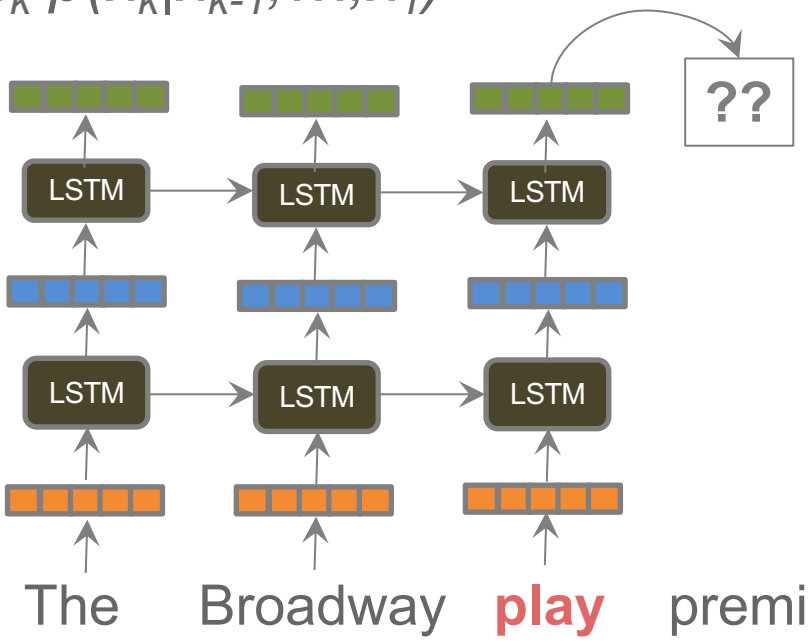
The Broadway **play** premiered yesterday .

Peters et al (2018): Deep contextualized word representations

ELMo

Neural language model

$p(\boldsymbol{x}) = \Pi_k \, p(x_k | x_{k-1}, \ldots, x_1)$

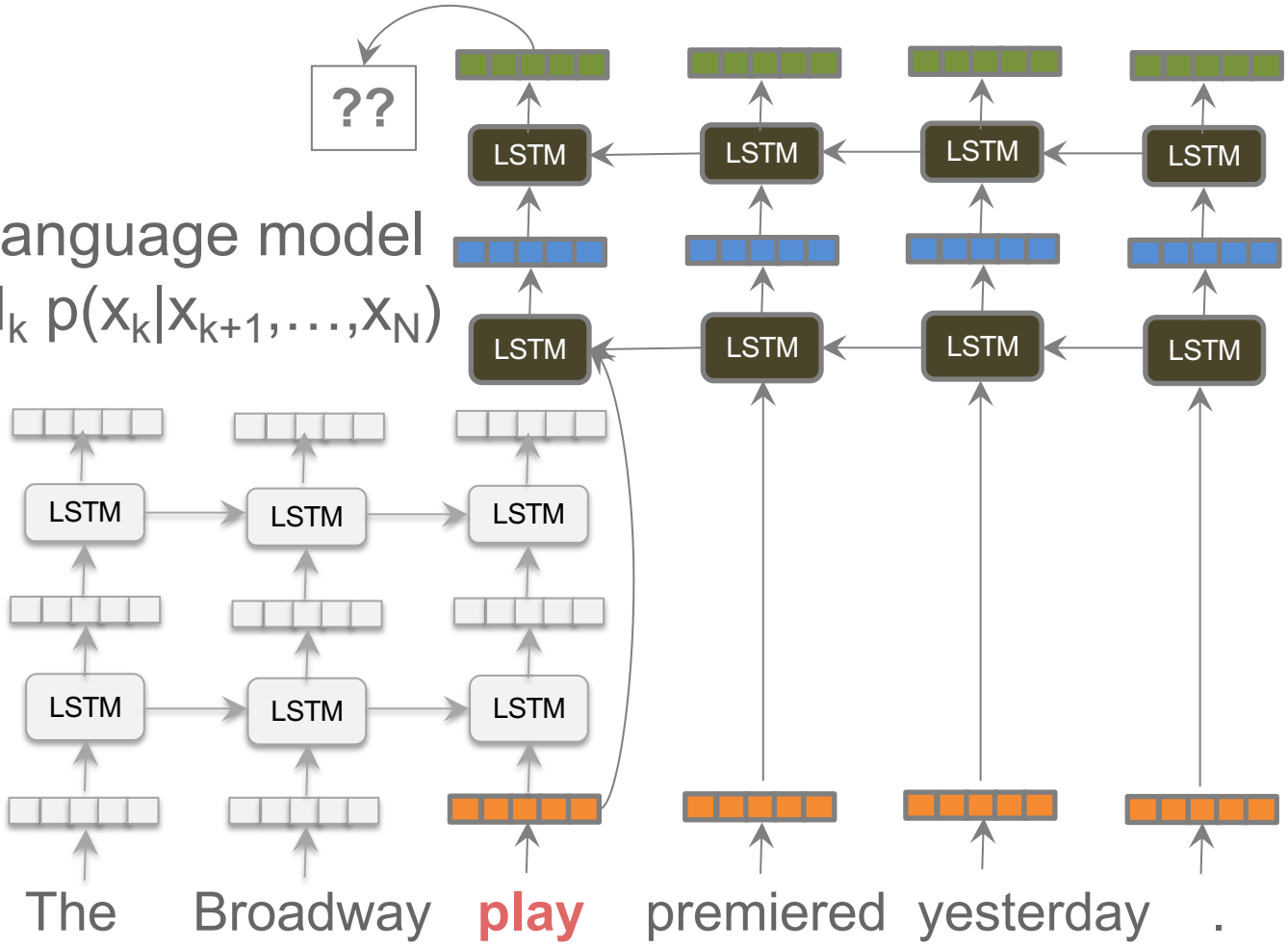


Peters et al (2018): Deep contextualized word representations

ELMo

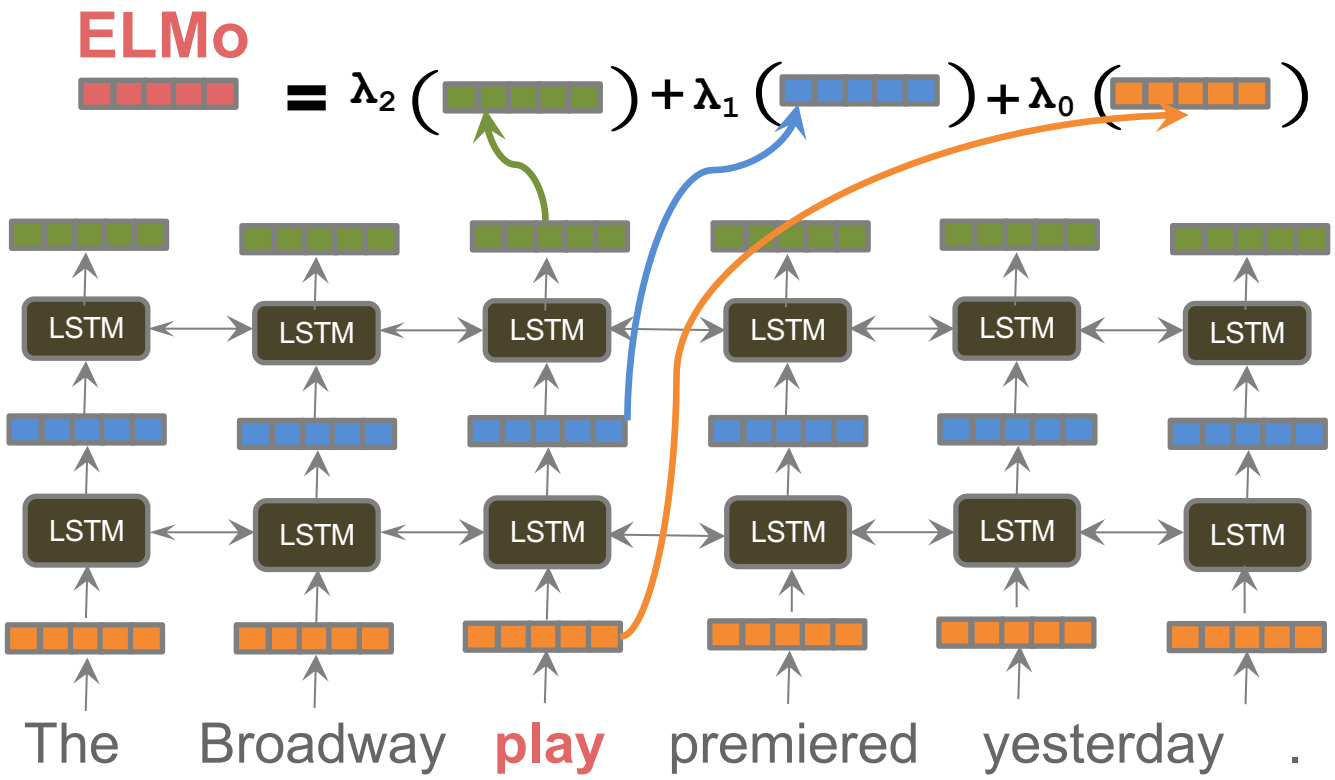
Neural language model

$p(x) = \Pi_k \, p(x_k | x_{k+1}, \ldots, x_N)$



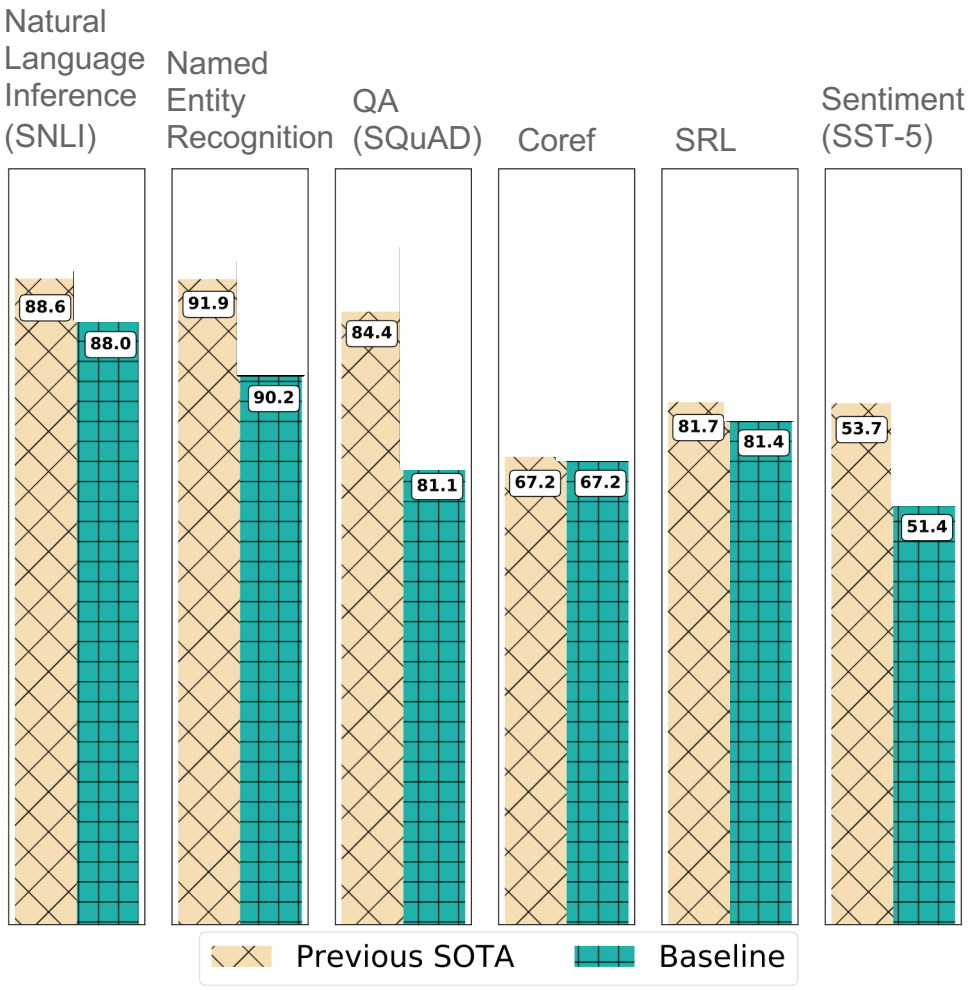
Peters et al (2018): Deep contextualized word representations

ELMo



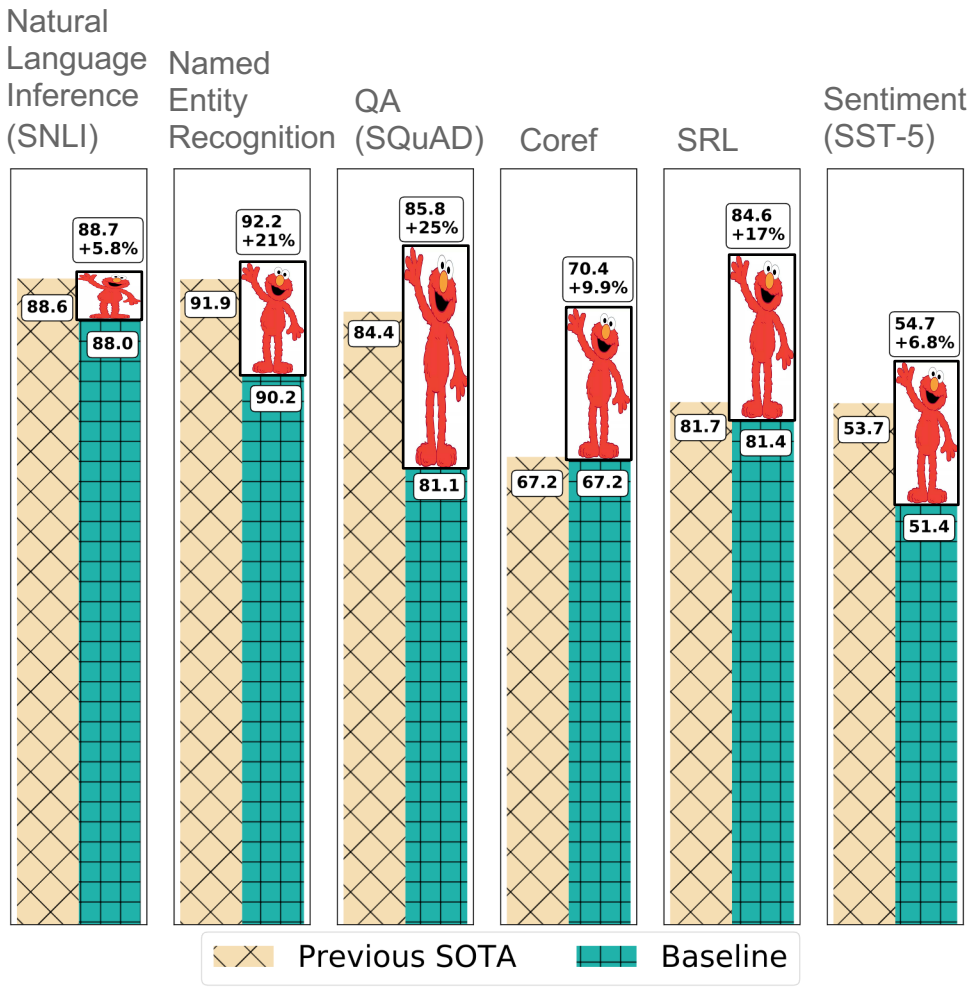
Peters et al (2018): Deep contextualized word representations

ELMo: Empirical results



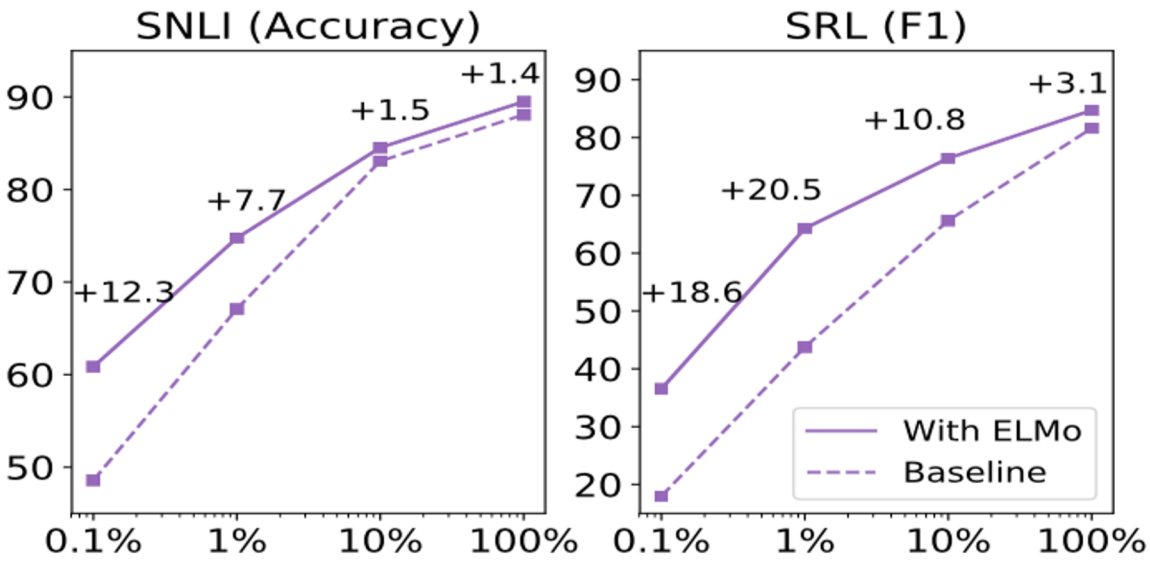
Peters et al (2018): Deep contextualized word representations

ELMo: Empirical results



Peters et al (2018): Deep contextualized word representations

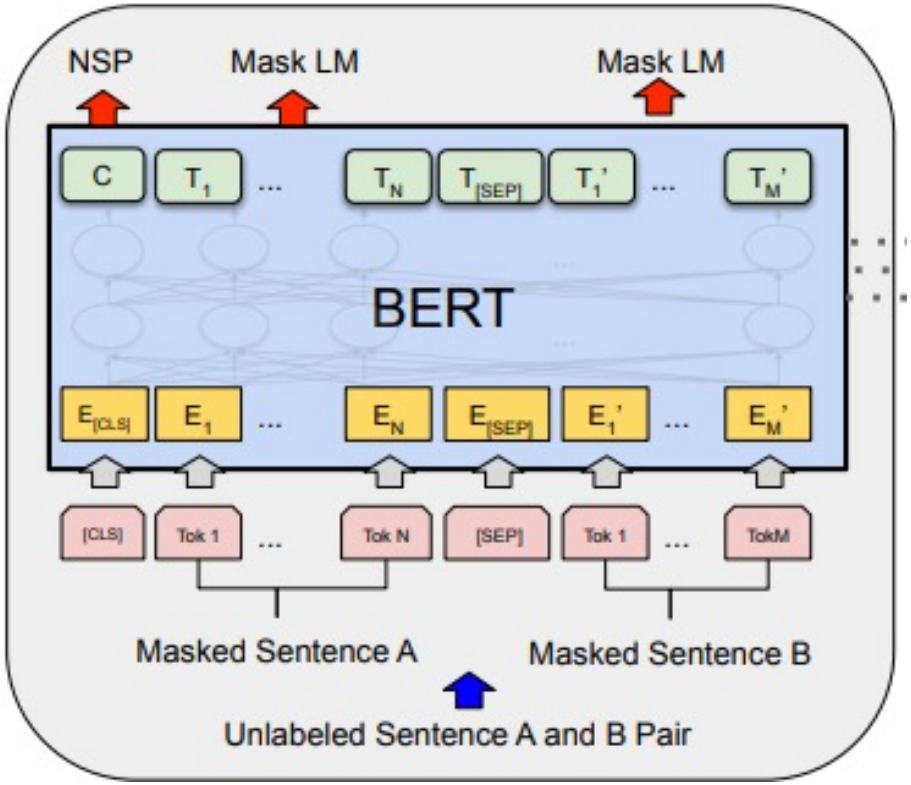
ELMo: Empirical results



Percent of training data

Peters et al (2018): Deep contextualized word representations

BERT – Devlin et al (2019)



Differences vs. ELMo:

- Transformer vs LSTM
- Masked language modeling vs autoregressive
- Scale! 60X more compute

Sesame Street LMs

Sesame Street LMs

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Sesame Street LMs

THE VERGE

TECH | SCIENCE | MORE

REPORT \ TECH \ ARTIFICIAL INTELLIGENCE

Why are so many AI systems named after Muppets?

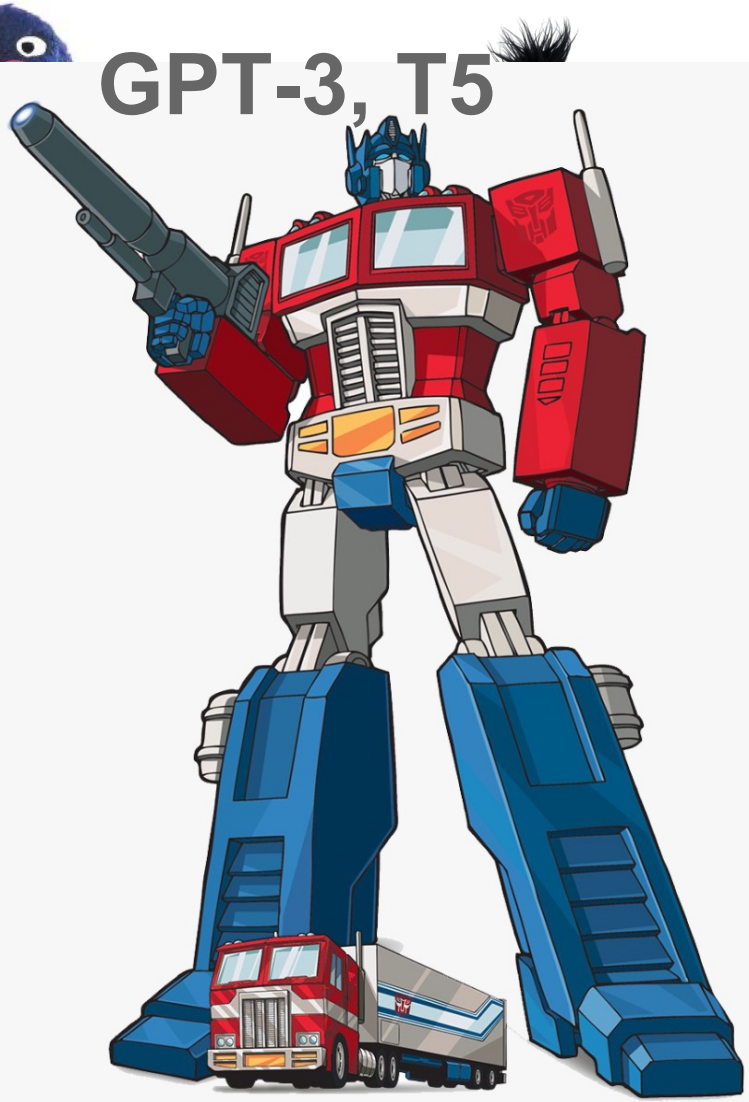
An inside joke that says a lot about AI development

By James Vincent | Dec 11, 2019, 8:26am EST

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Sesame Street LMs

GPT-3, T5



Google

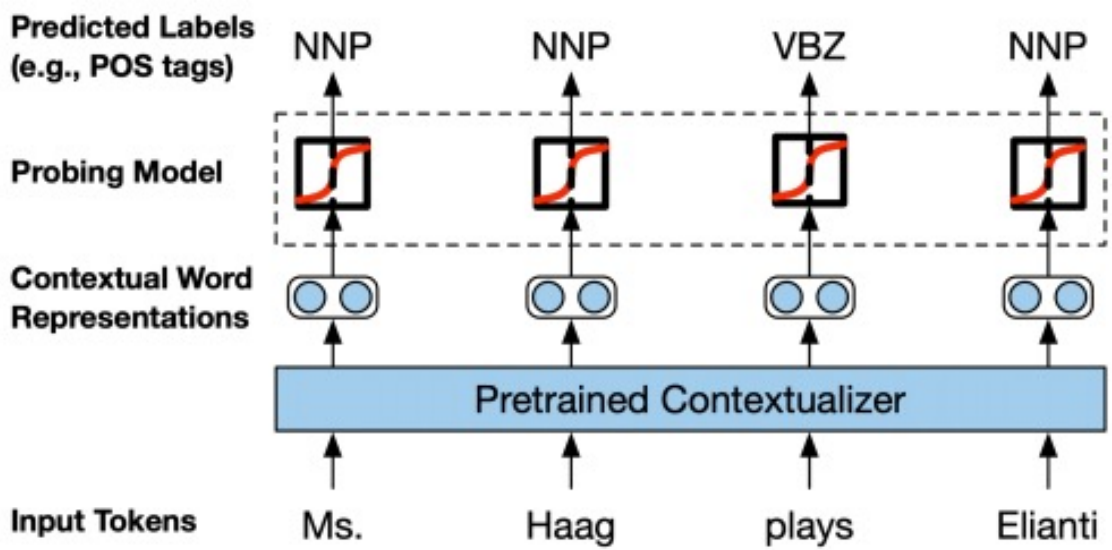
MORE

Bai du 百

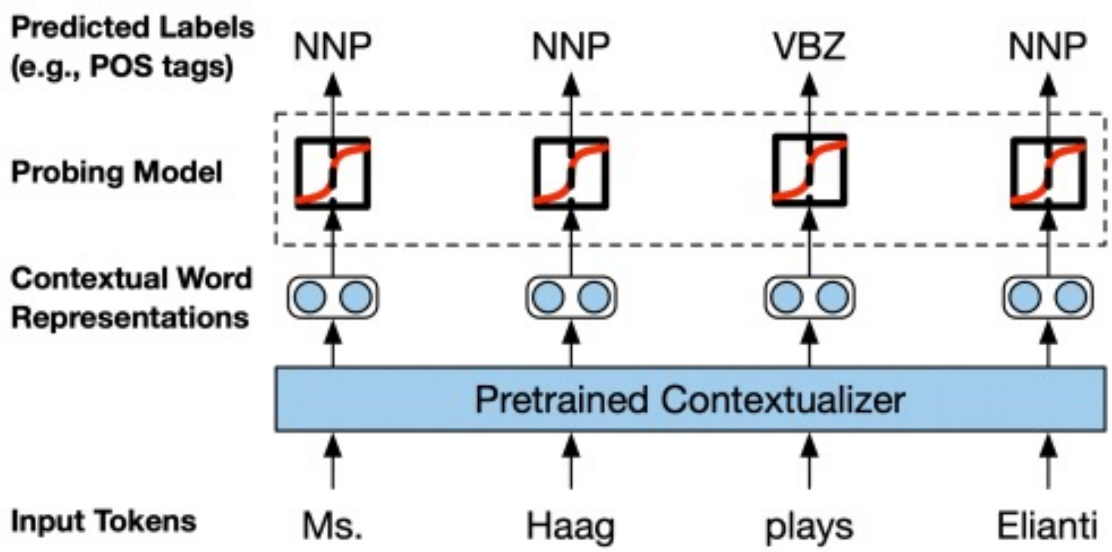
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

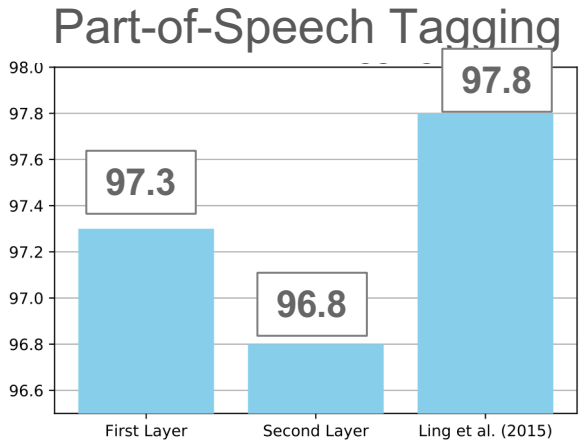
Analyzing contextual representations

Why do these pretraining methods work?

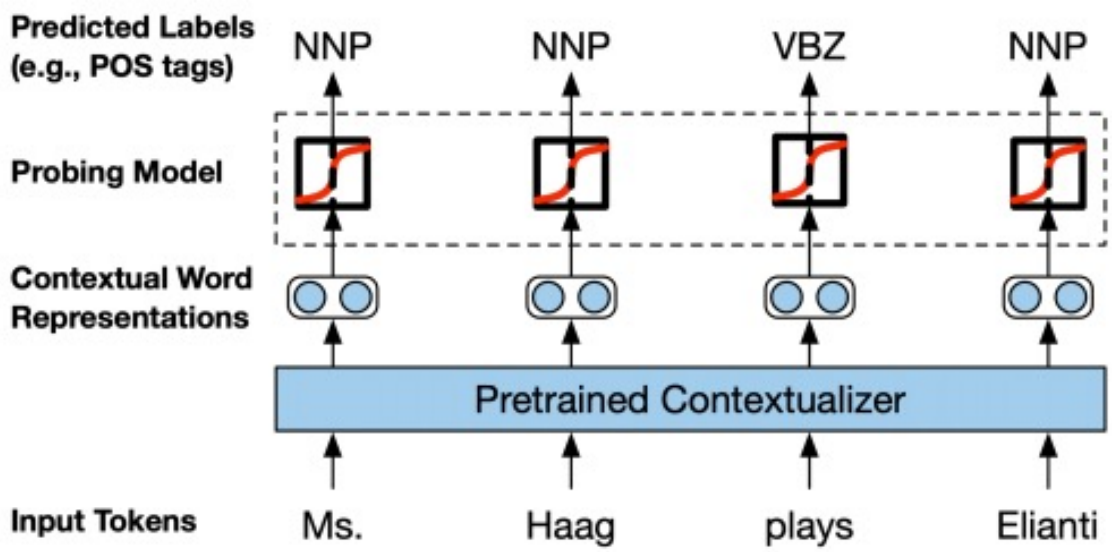


Why do these pretraining methods work?

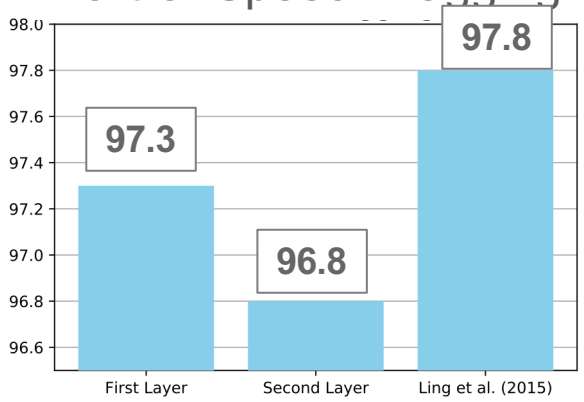




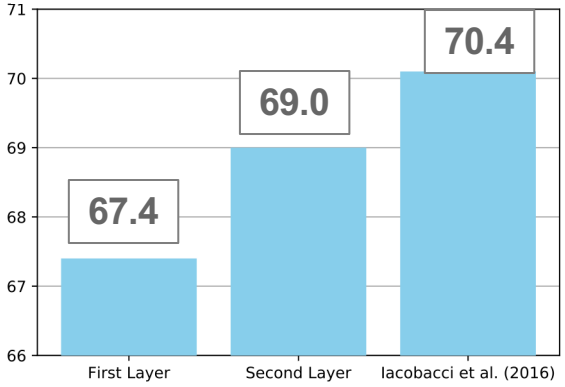
Why do these pretraining methods work?



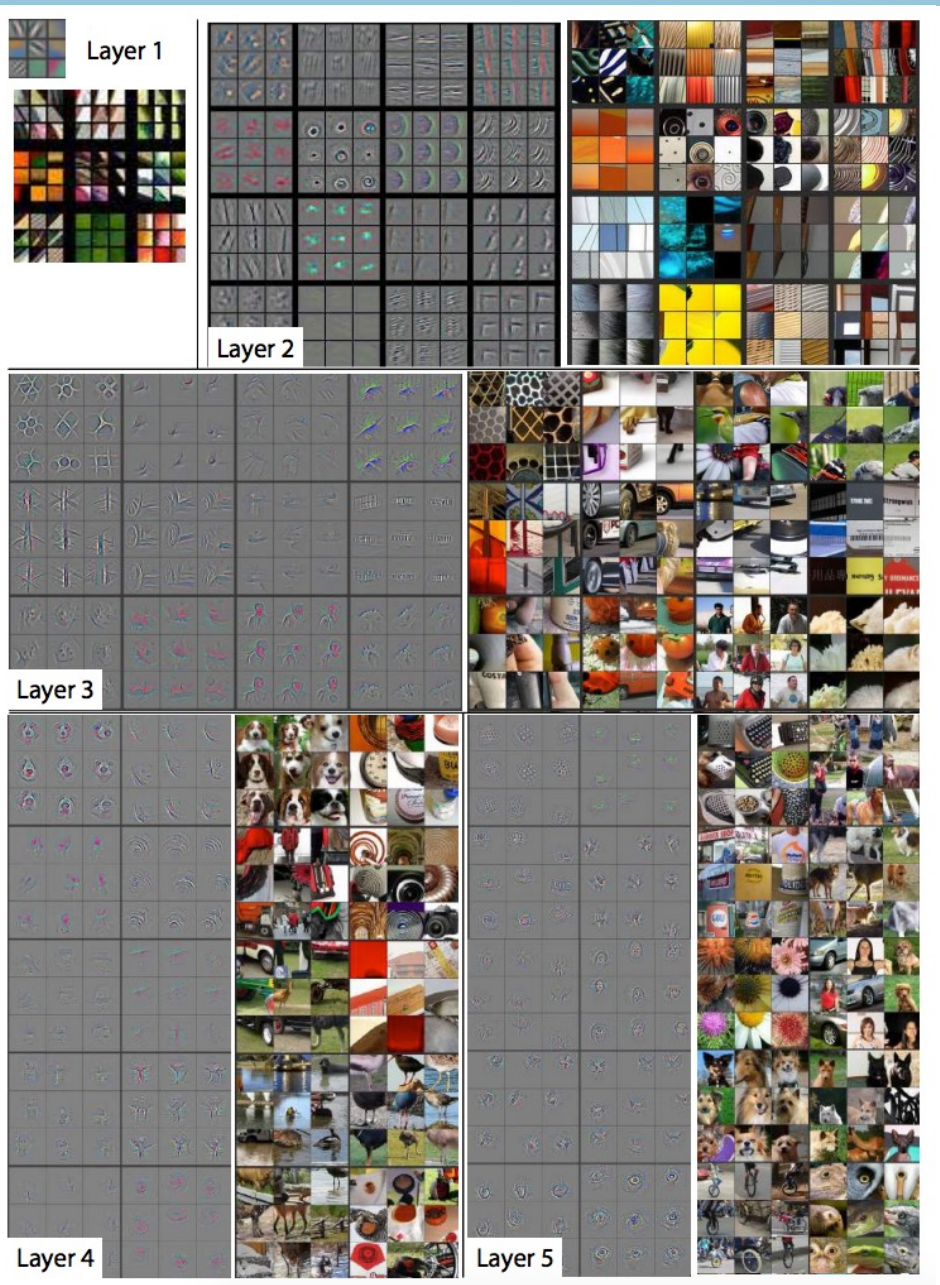
Part-of-Speech Tagging



Word sense disambiguation



Emergence of hierarchical features



Zeiler and Fergus, 2014:
Visualizing and Understanding Convolutional Networks

Emergence of hierarchical features

Hierarchical features naturally arise in computer vision models.

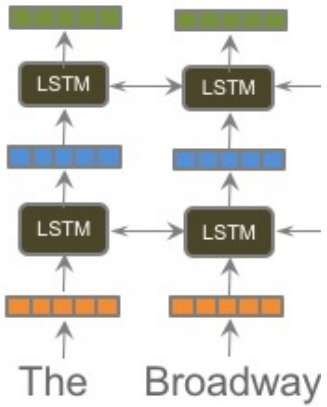
Can we find something similar in these word representations?
→ compare 3 models: 4-layer LSTM, 6-layer transformer, 16-layer CNN

Emergence of hierarchical features

Hierarchical features naturally arise in computer vision models.

Can we find something similar in these word representations?
→ compare 3 models: 4-layer LSTM, 6-layer transformer, 16-layer CNN

Vector analogy results for *uncontextualized* word vectors



Emergence of hierarchical features

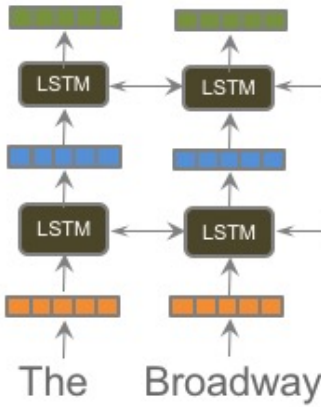
Hierarchical features naturally arise in computer vision models.

Can we find something similar in these word representations?
→ compare 3 models: 4-layer LSTM, 6-layer transformer, 16-layer CNN

cat : dog :: cats : dogs

king : man :: queen : woman

Representation	Syntactic	Semantic
GloVe	77.9	**79.2**
n-gram hash	72.3	0.5
LSTM 4-layer		
Transformer		
Gated CNN		



Vector analogy results for *uncontextualized* word vectors

Emergence of hierarchical features

Hierarchical features naturally arise in computer vision models.

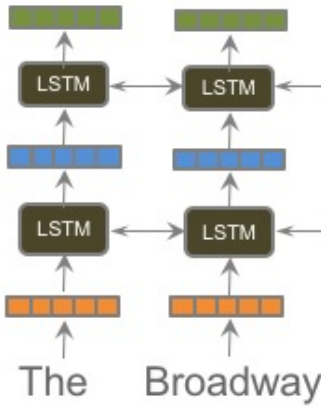
Can we find something similar in these word representations?
→ compare 3 models: 4-layer LSTM, 6-layer transformer, 16-layer CNN

cat : dog :: cats : dogs

king : man :: queen : woman

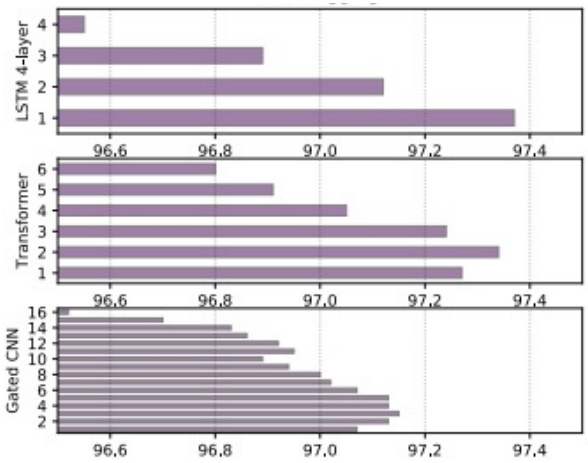
Representation	Syntactic	Semantic
GloVe	77.9	**79.2**
n-gram hash	72.3	0.5
LSTM 4-layer	74.2	11.5
Transformer	**87.1**	48.8
Gated CNN	83.6	26.3

Vector analogy results for *uncontextualized* word vectors

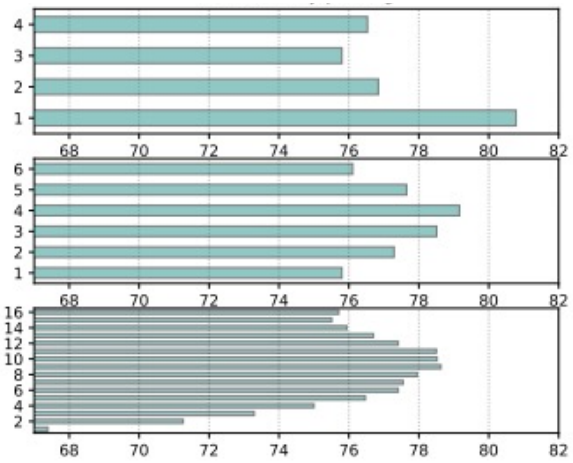


Emergence of hierarchical features

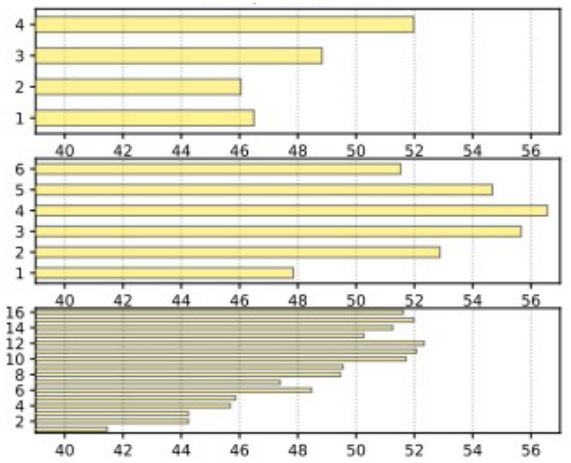
Part-of-speech tagging



Syntactic parsing



Unsupervised coreference



Peters et al (2018), Dissecting contextual word representations: Architecture and Representation

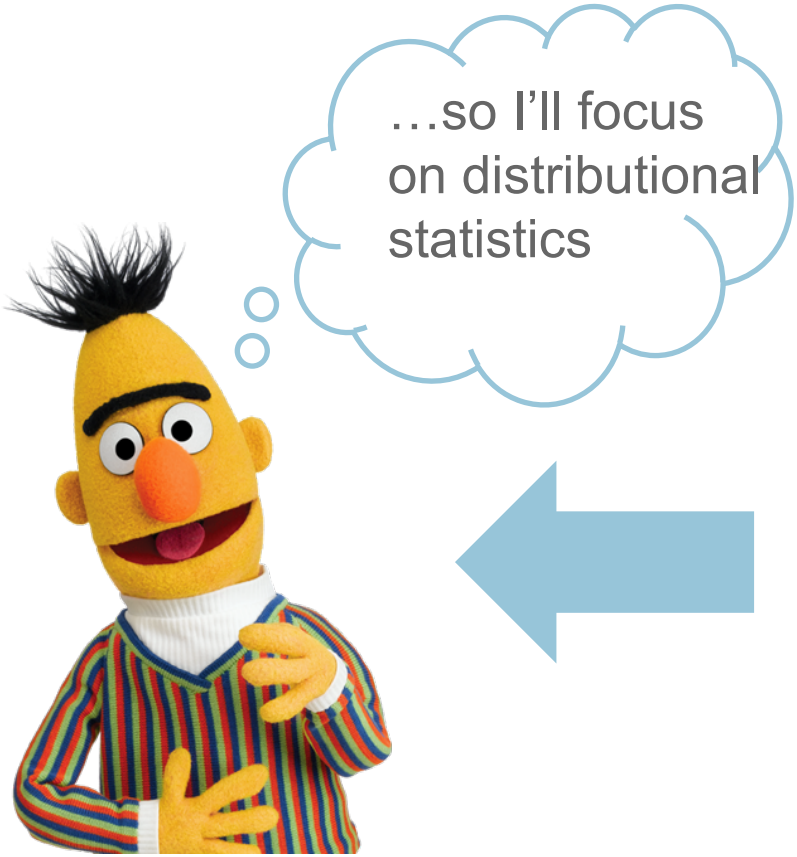
See also:

- Tenney et al (2019): BERT Rediscovers the Classical NLP Pipeline
- Liu et al (2019): Linguistic Knowledge and Transferability of Contextual Word Representations
- Rogers et al (2020): A Primer in BERTology: What we know about how BERT works

Knowledge enhanced contextual word representations

~1 GB

~10-100 GBs



~1 GB

~10-100 GBs

Paris is located in [MASK].

Parc national de la Mauricie is located in [MASK].

Parc national de la Mauricie is located in [MASK].

La Mauricie National Park

From Wikipedia, the free encyclopedia

La Mauricie National Park (French: *Parc national de la Mauricie*) is located near Shawinigan in the Laurentian mountains, in the Mauricie region of Quebec, Canada. It covers 536 km^2 (207 sq mi) in the southern Canadian Shield region bordering the Saint Lawrence lowlands. The park contains 150 lakes and many ponds.

Parc national de la Mauricie is located in [MASK].

La Mauricie National Park

From Wikipedia, the free encyclopedia

La Mauricie National Park (French: *Parc national de la Mauricie*) is located near Shawinigan in the Laurentian mountains, in the Mauricie region of Quebec, Canada. It covers 536 km^2 (207 sq mi) in the southern Canadian Shield region bordering the Saint Lawrence lowlands. The park contains 150 lakes and many ponds.

Human curated, compressed KB

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Let me look it up…

Parc national de la Mauricie is located in [MASK].

La Mauricie National Park

From Wikipedia, the free encyclopedia

La Mauricie National Park (French: *Parc national de la Mauricie*) is located near Shawinigan in the Laurentian mountains, in the Mauricie region of Quebec, Canada. It covers 536 km^2 (207 sq mi) in the southern Canadian Shield region bordering the Saint Lawrence lowlands. The park contains 150 lakes and many ponds.

Human curated, compressed KB

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Quebec

Parc national de la Mauricie is located in [MASK].

La Mauricie National Park

From Wikipedia, the free encyclopedia

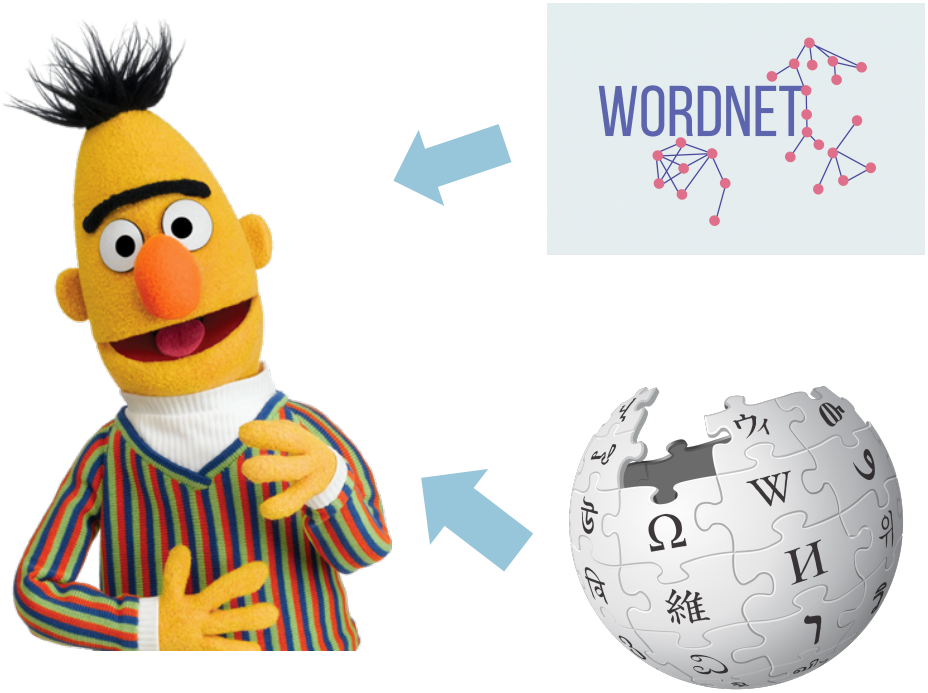
La Mauricie National Park (French: *Parc national de la Mauricie*) is located near Shawinigan in the Laurentian mountains, in the Mauricie region of Quebec, Canada. It covers 536 km^2 (207 sq mi) in the southern Canadian Shield region bordering the Saint Lawrence lowlands. The park contains 150 lakes and many ponds.

Human curated, compressed KB

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Introducing KnowBert

KnowBert (Peters et al 2019) learns general purpose knowledge enhanced representations by retrieving relevant information from multiple knowledge bases with a sparse O(1) lookup.



Task agnostic

Outperforms BERT

Related work:

- Retrieval: REALM (Guu et al 2020), RAG (Lewis et al 2020)
- Sparsity: Switch Transformer (Fedus et al 2021)

Knowledge bases

- Adopt very general definition of knowledge base, so easy to incorporate heterogenous knowledge sources: (subject, relation, object) tuples, entity metadata, etc.
- Easy to add large number of parameters via entity embeddings
- First compute initial contextual representations without KB for disambiguated entity linking, then re-contextualize to incorporate knowledge.

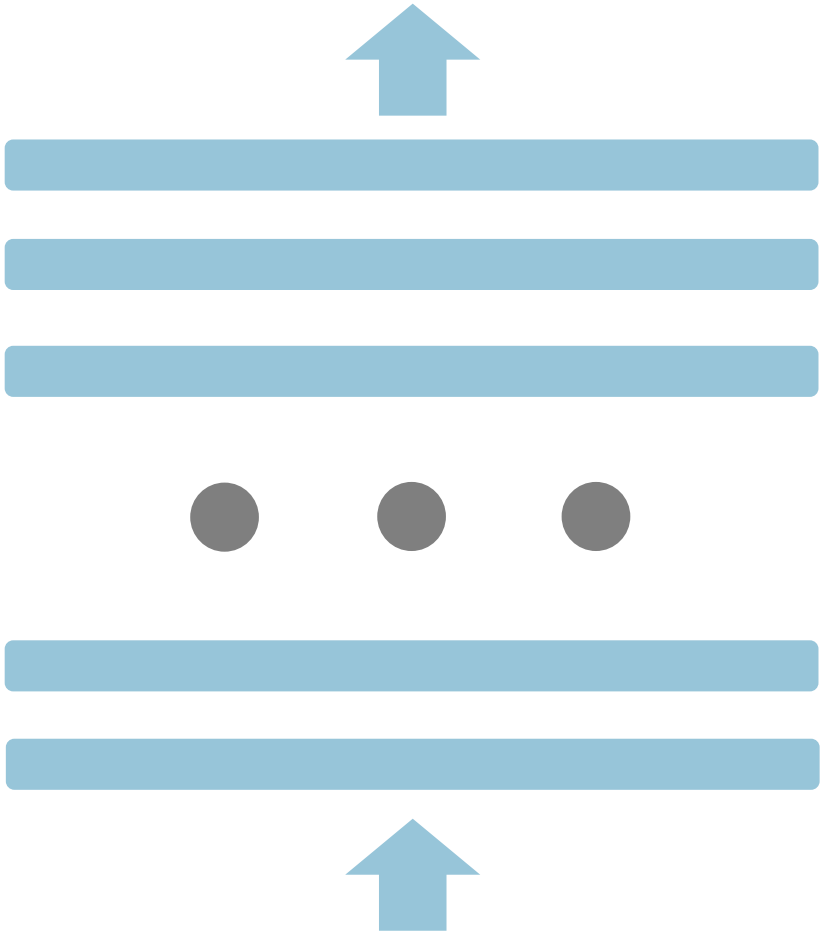
KnowBert

Loss: masked language modeling

La Mauricie National Park is …

KnowBert

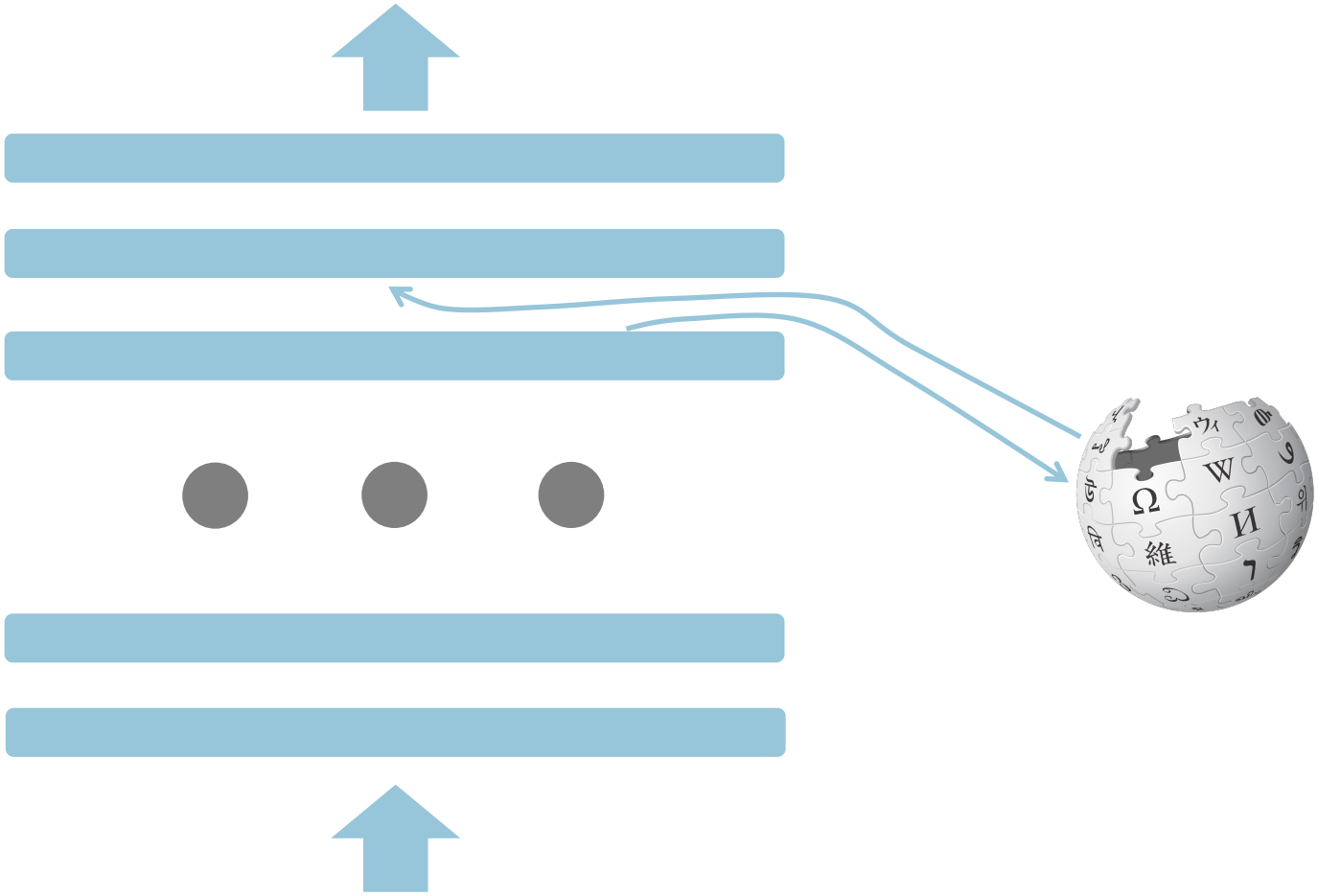
Loss: masked language modeling



La Mauricie National Park is …

KnowBert

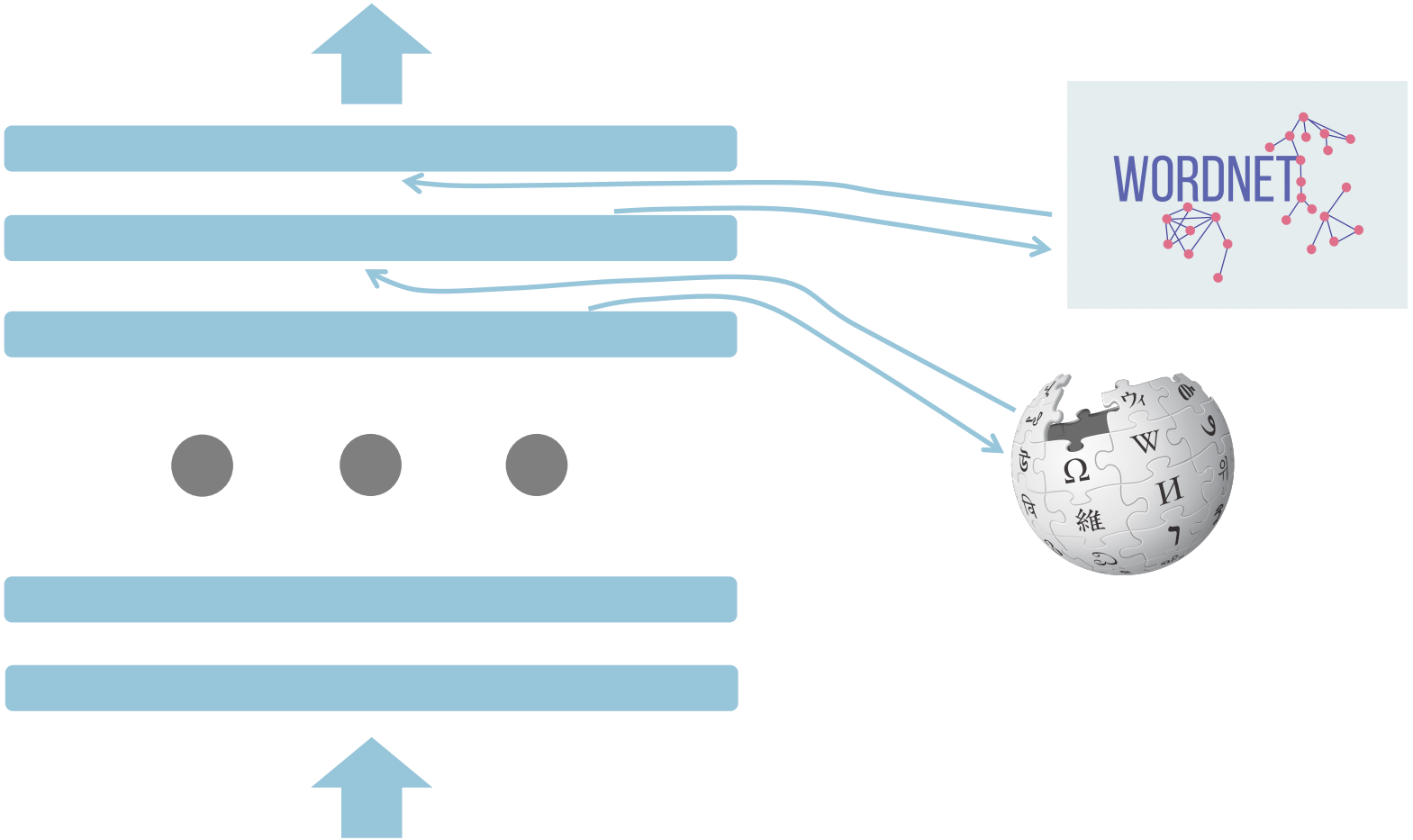
Loss: masked language modeling **+ entity linking**



La Mauricie National Park is …

KnowBert

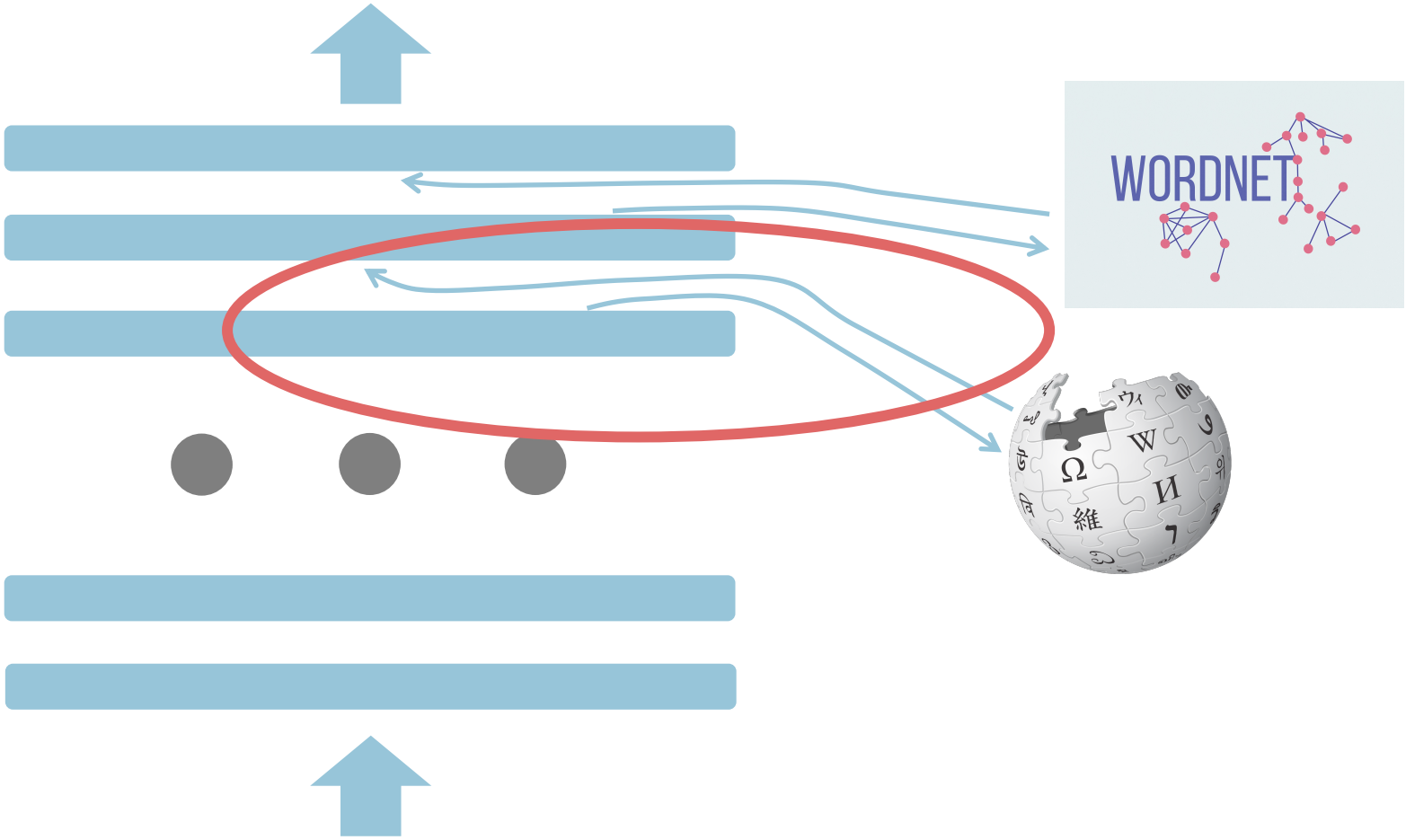
Loss: masked language modeling **+ entity linking**



La Mauricie National Park is …

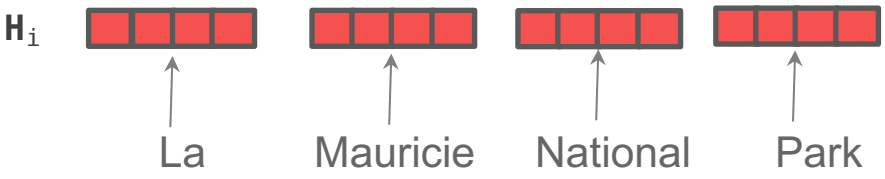
Knowledge attention and recontextualization

Loss: masked language modeling **+ entity linking**

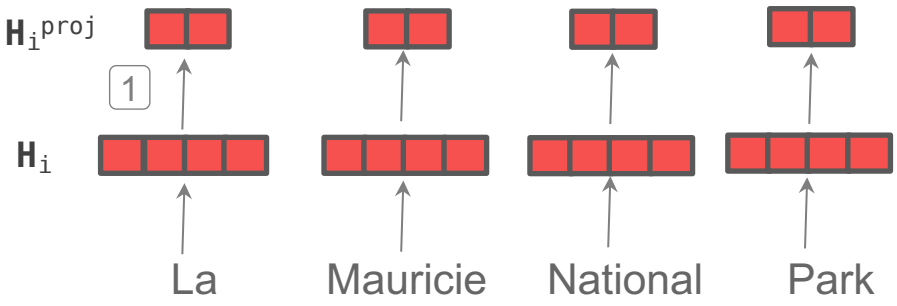


La Mauricie National Park is …

Knowledge attention and recontextualization

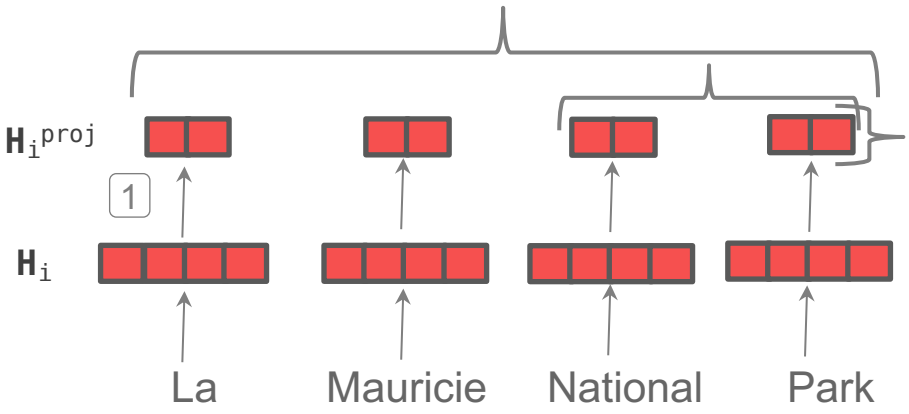


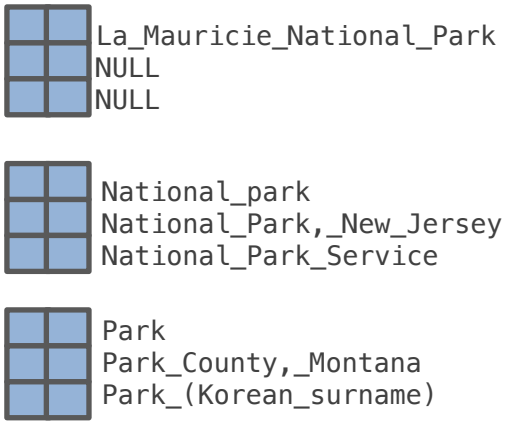
Knowledge attention and recontextualization



1 Projected contextual word embeddings

Knowledge attention and recontextualization

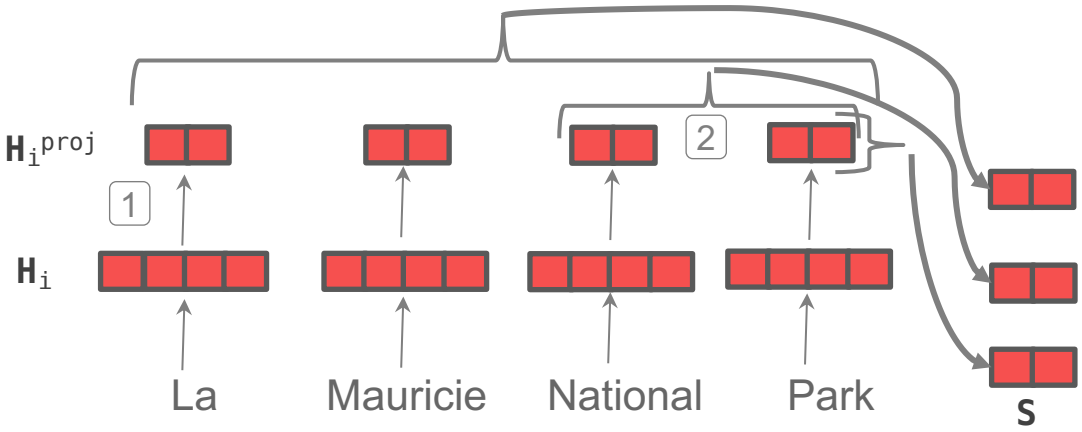


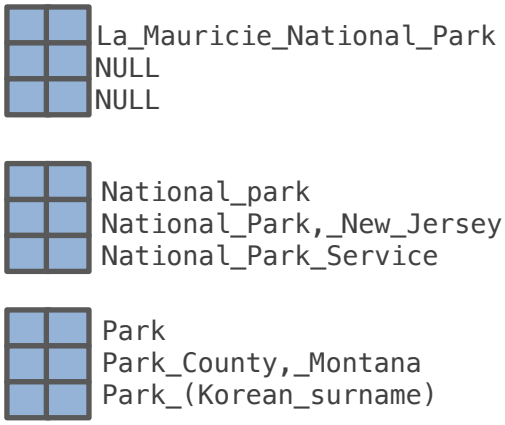


1 Projected contextual word embeddings

Knowledge attention and recontextualization

Span representation for all candidates

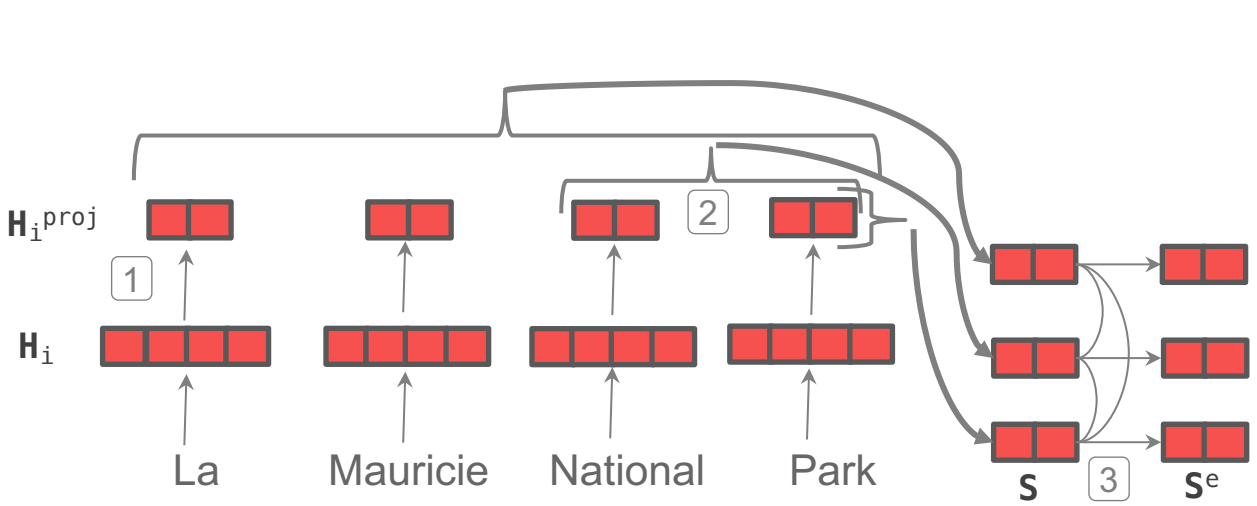


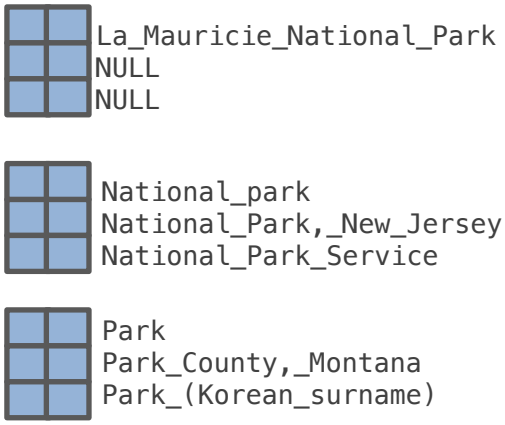


1 Projected contextual word embeddings
2 Pooled entity mention-span representations

Knowledge attention and recontextualization

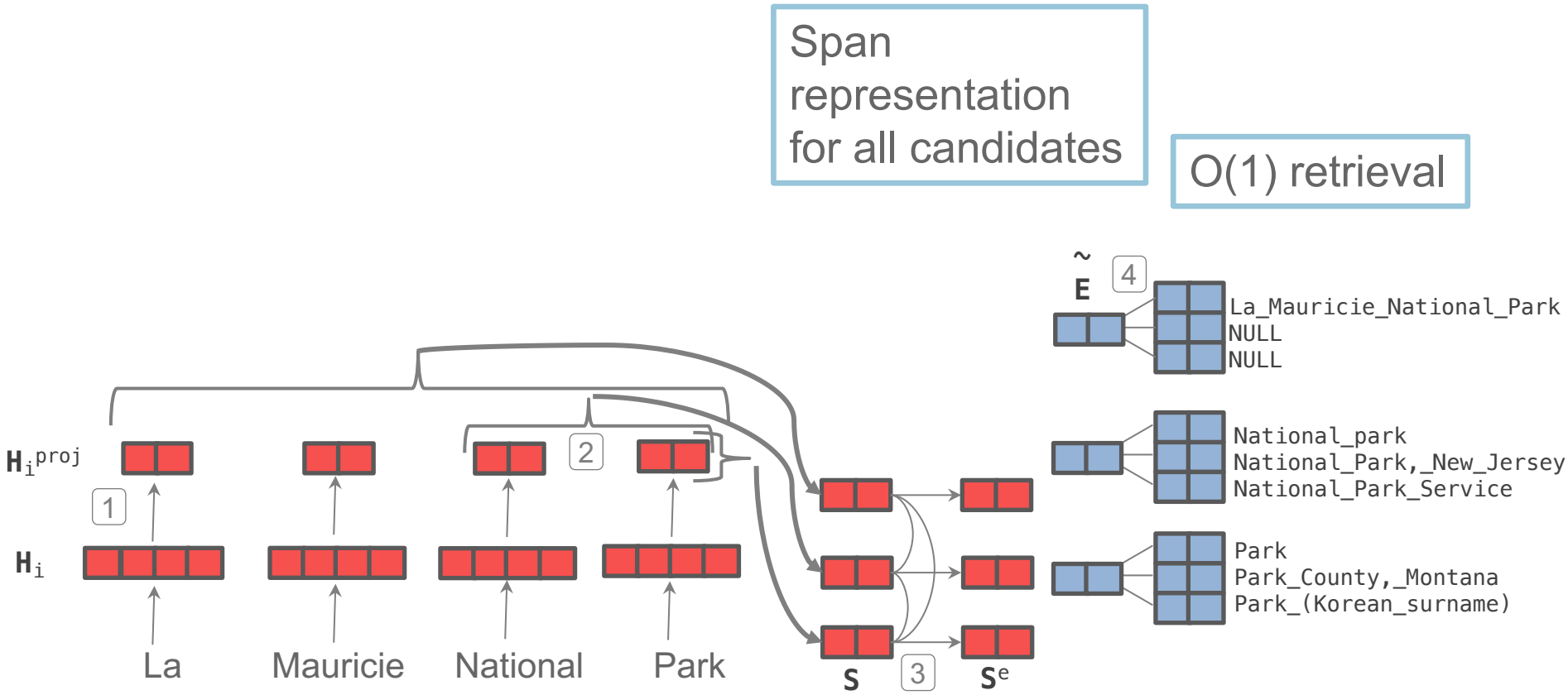
Span representation for all candidates





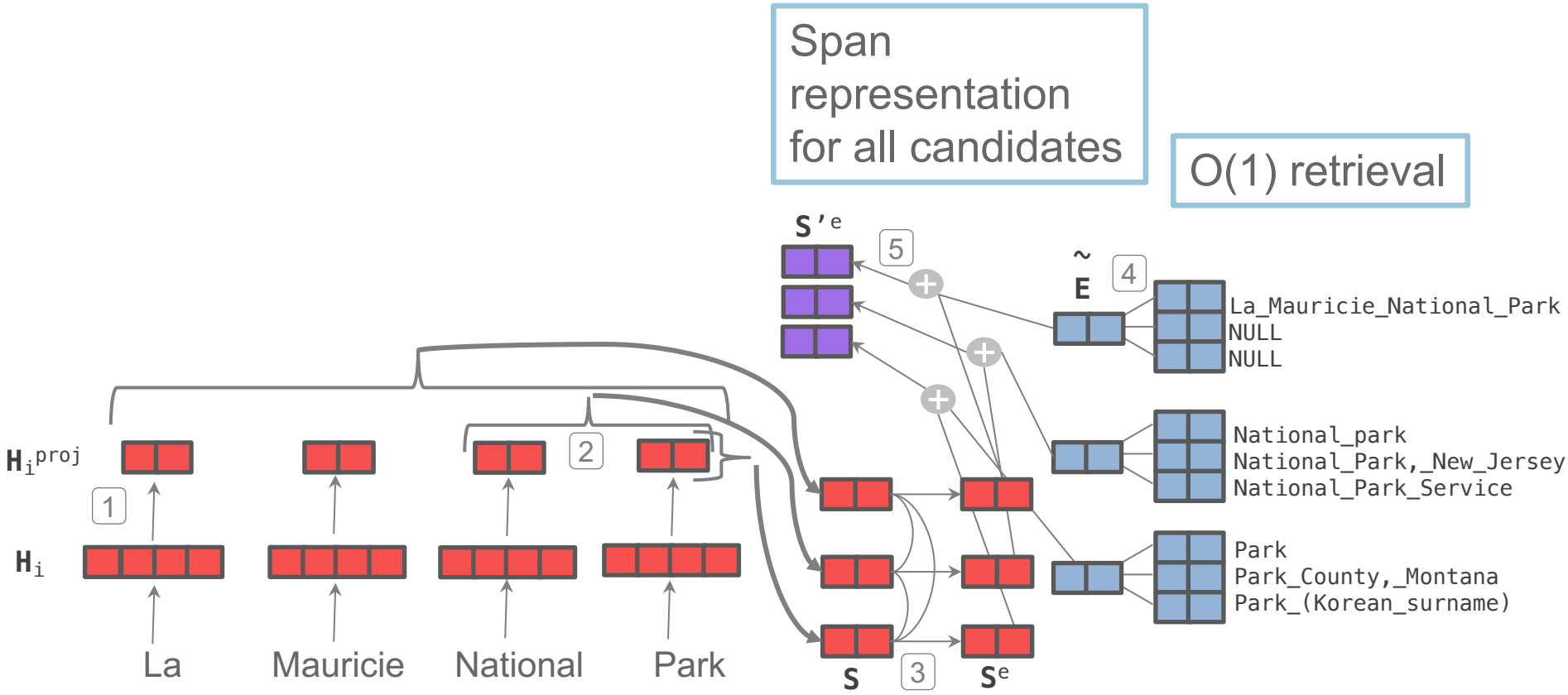
1 Projected contextual word embeddings
2 Pooled entity mention-span representations
3 Span-span self-attention

Knowledge attention and recontextualization



1 Projected contextual word embeddings
2 Pooled entity mention-span representations
3 Span-span self-attention
4 Weighted entity vectors

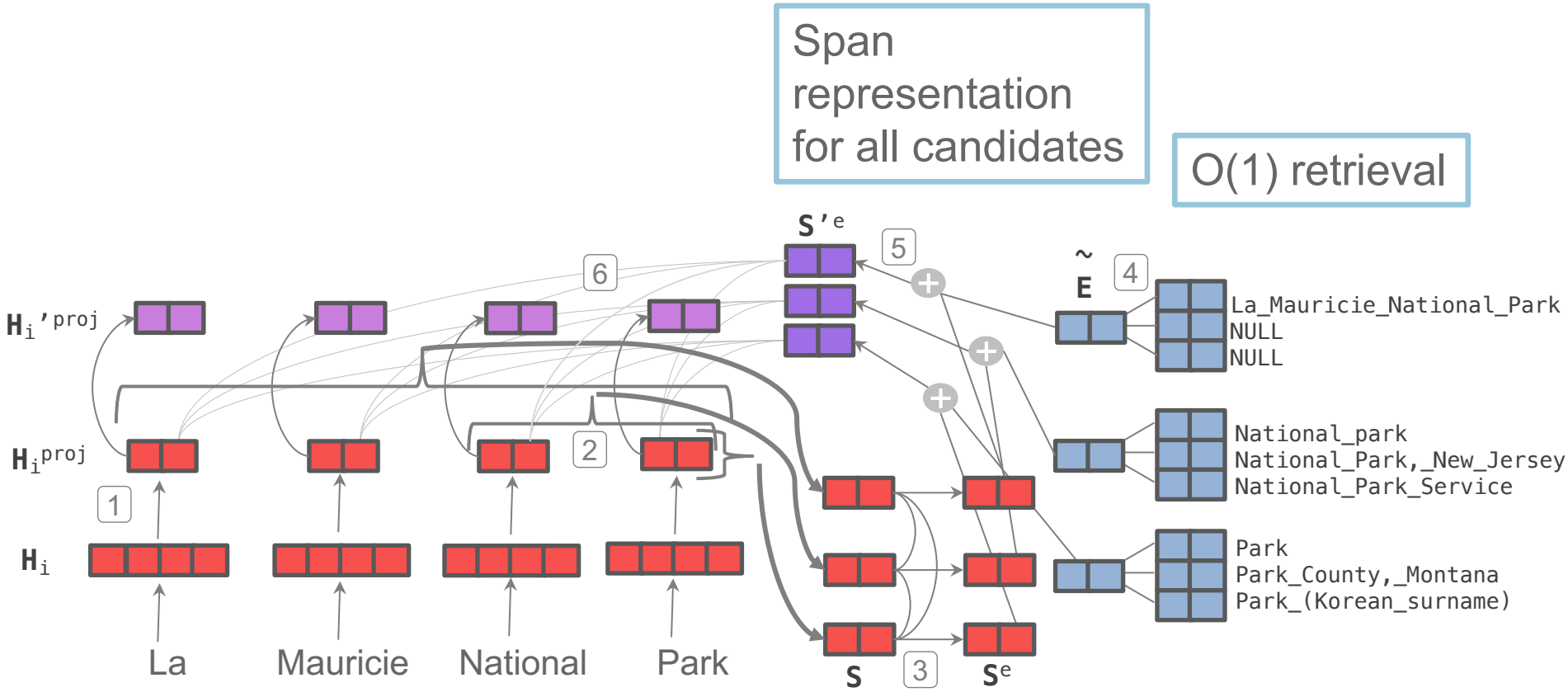
Knowledge attention and recontextualization



1 Projected contextual word embeddings
2 Pooled entity mention-span representations
3 Span-span self-attention
4 Weighted entity vectors

5 Knowledge enhanced entity-spans

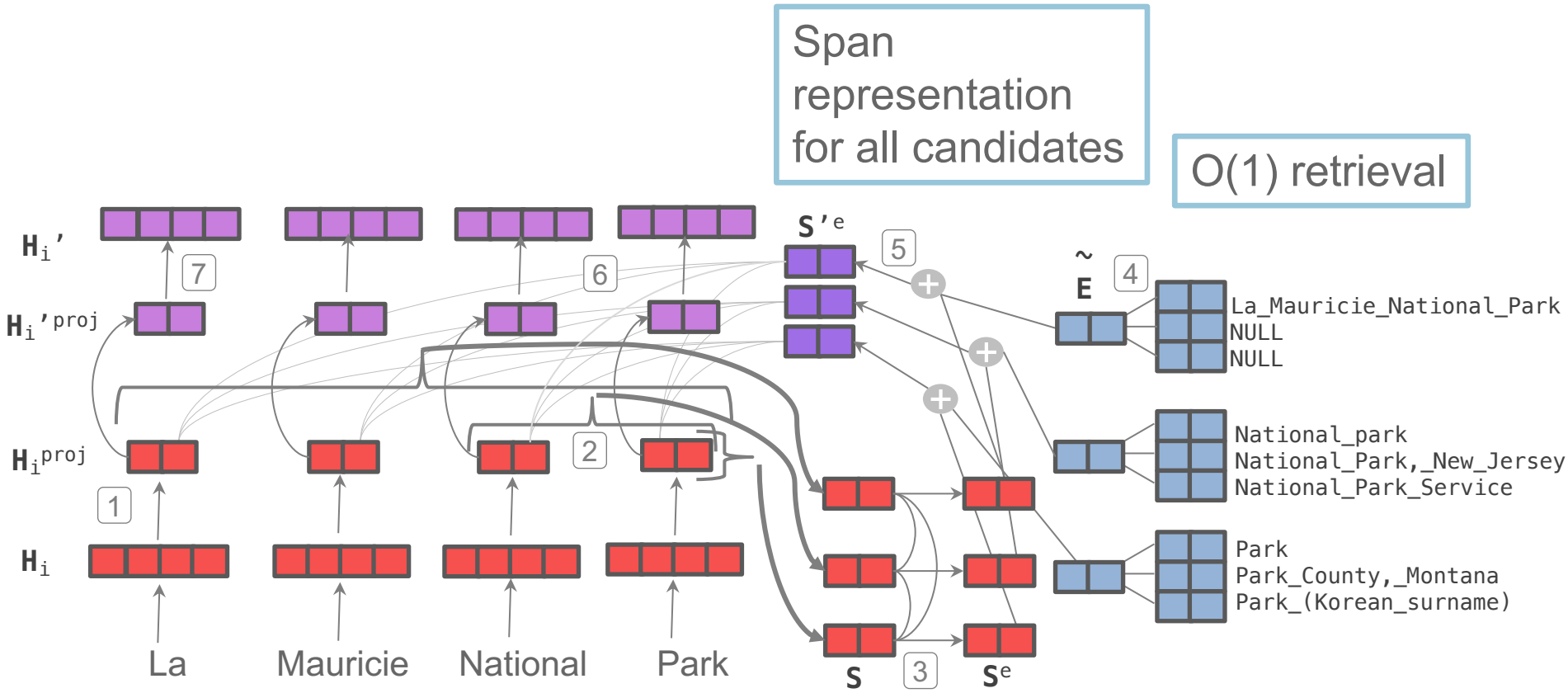
Knowledge attention and recontextualization



1 Projected contextual word embeddings
2 Pooled entity mention-span representations
3 Span-span self-attention
4 Weighted entity vectors

5 Knowledge enhanced entity-spans
6 Recontextualized knowledge enhanced representations

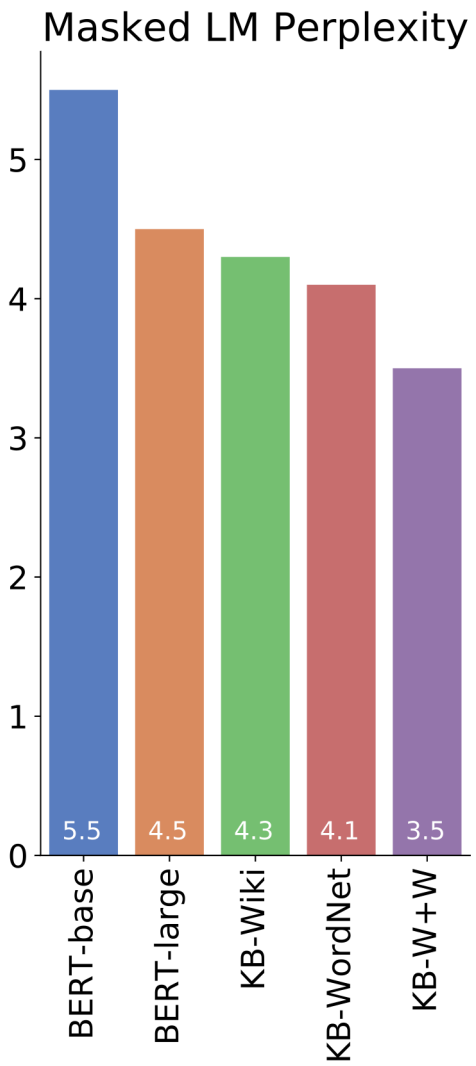
Knowledge attention and recontextualization



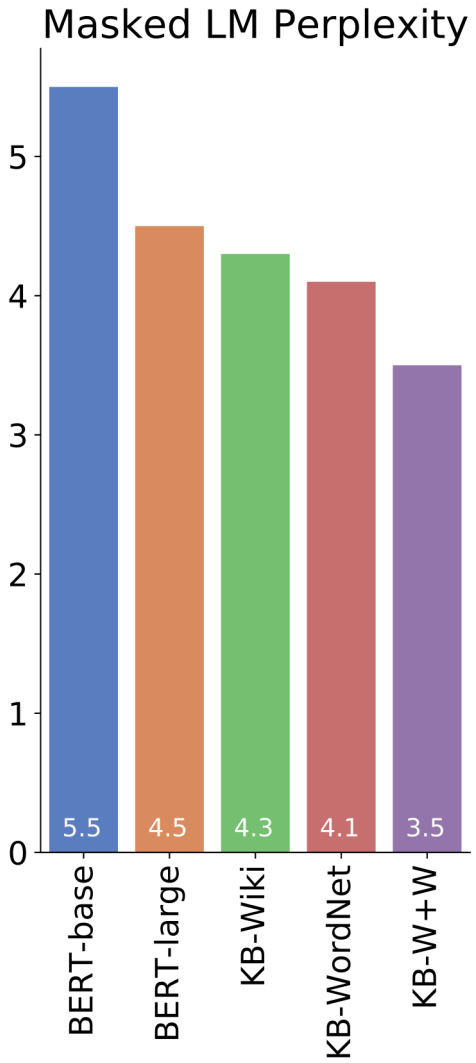
1 Projected contextual word embeddings
2 Pooled entity mention-span representations
3 Span-span self-attention
4 Weighted entity vectors

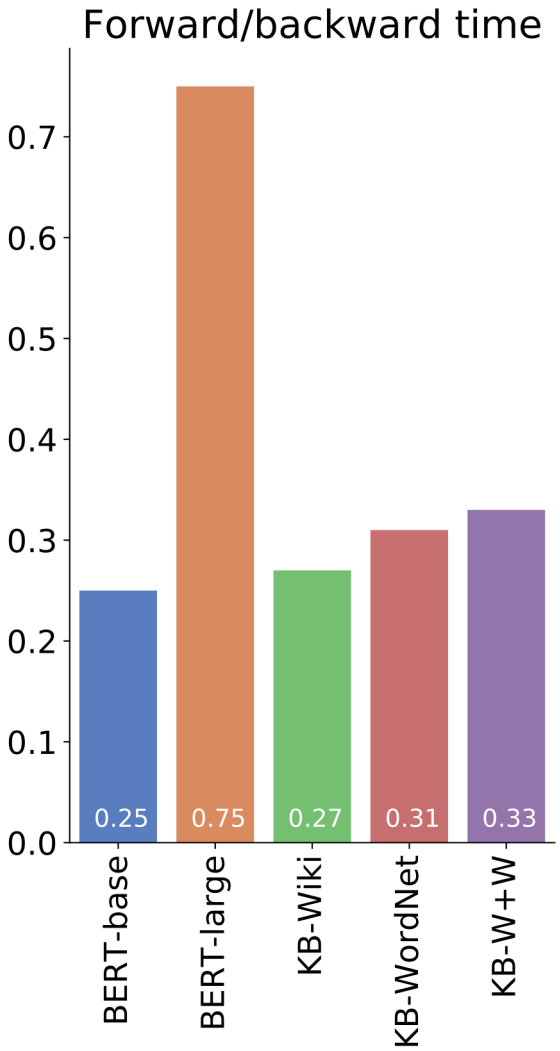
5 Knowledge enhanced entity-spans
6 Recontextualized knowledge enhanced representations
7 Projected knowledge enhanced representations

Intrinsic evaluation



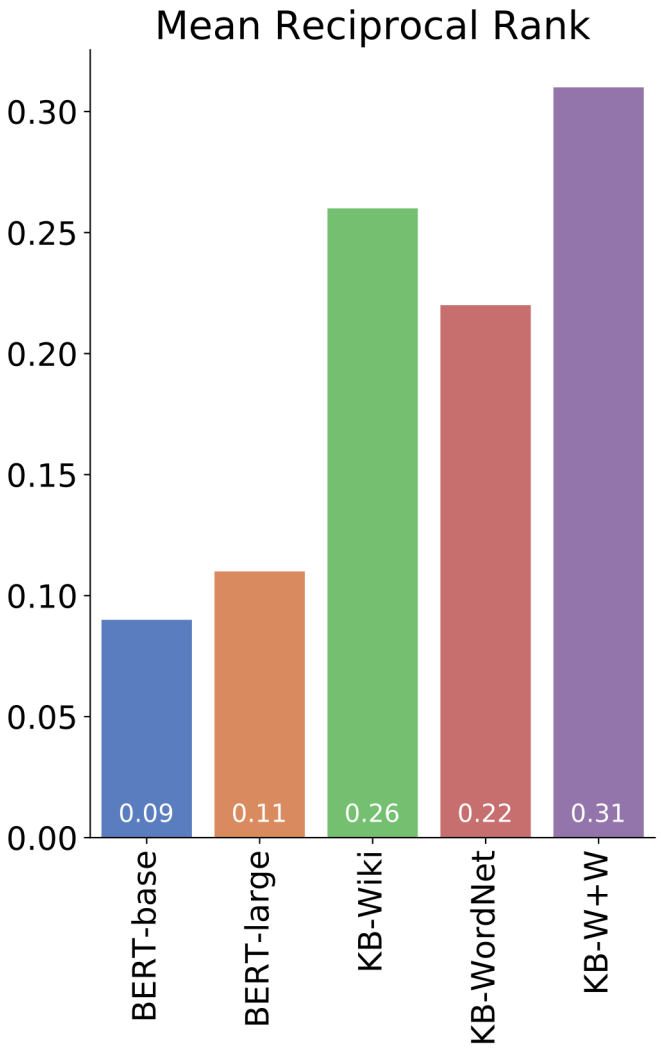
Intrinsic evaluation





ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

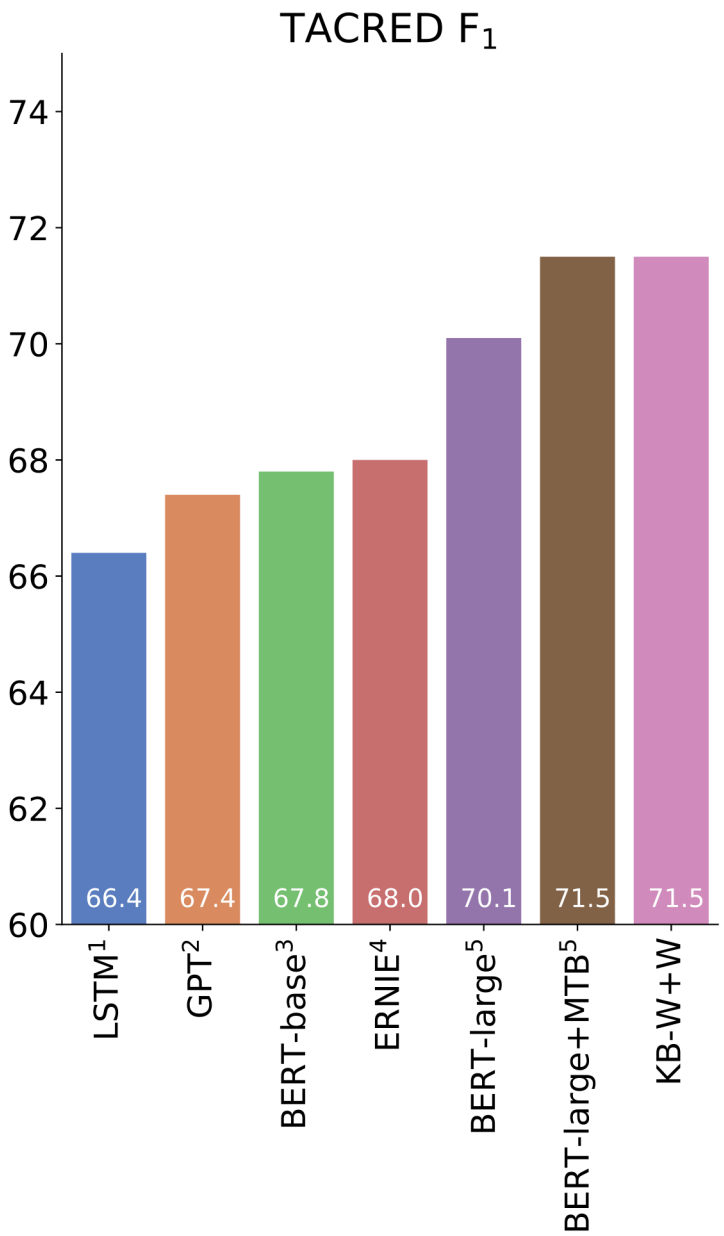
Knowledge probe



The song Eleanor Rigby
is performed by [MASK] [MASK].

180K tuples, 17 relation types
collected from Wikidata.

TACRED relationship extraction



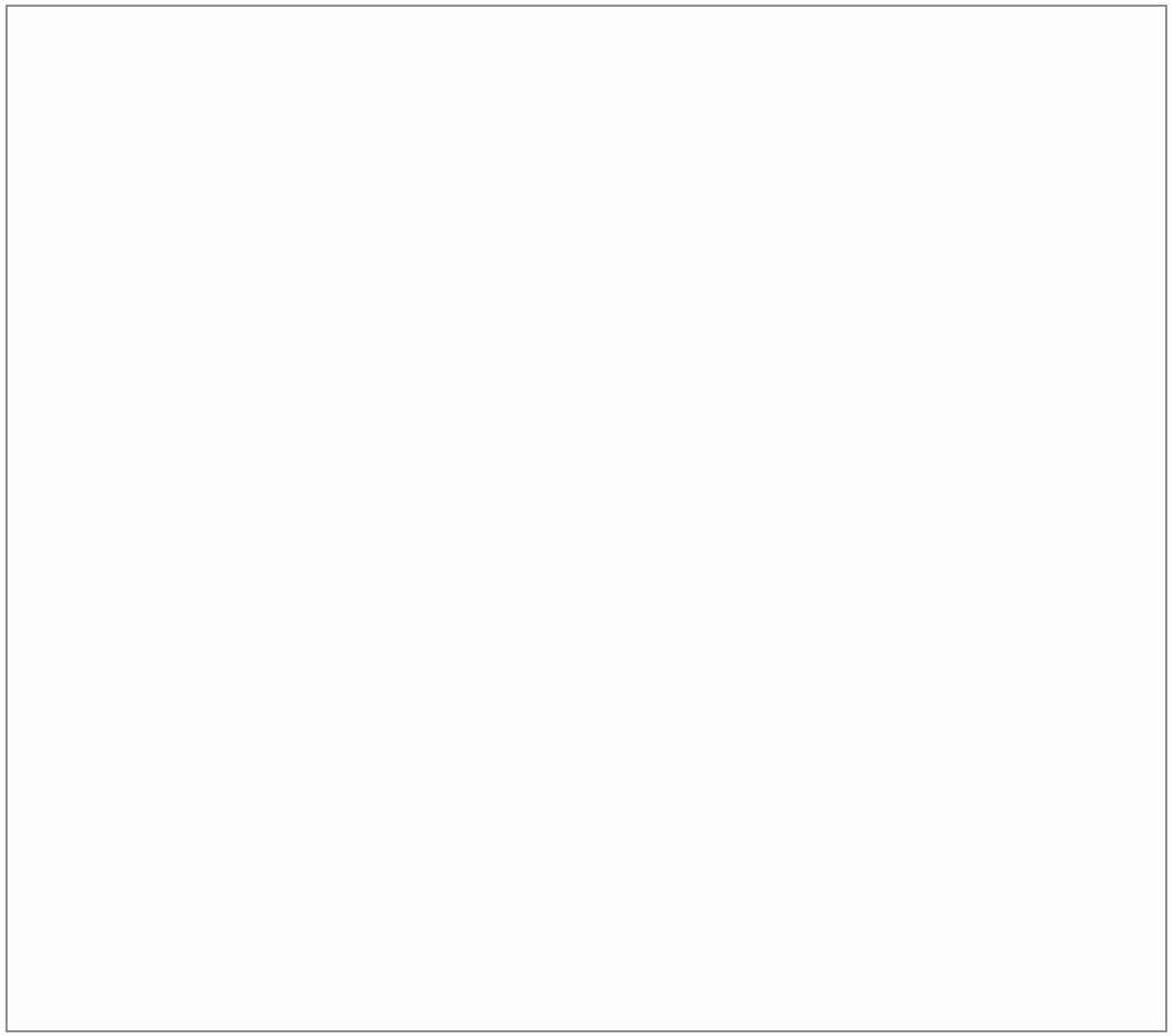
TACRED relationship extraction (Zhang et al 2017).

42 relations

[1] Zhang et al. (2018)
[2] Alt et al. (2019)
[3] Shi and Lin (2019)
[4] Zhang et al. (2019)
[5] Soares et al. (2019)

Longformer:
Long Document Transformer

Transformer attention



Encoder-Decoder
Machine translation

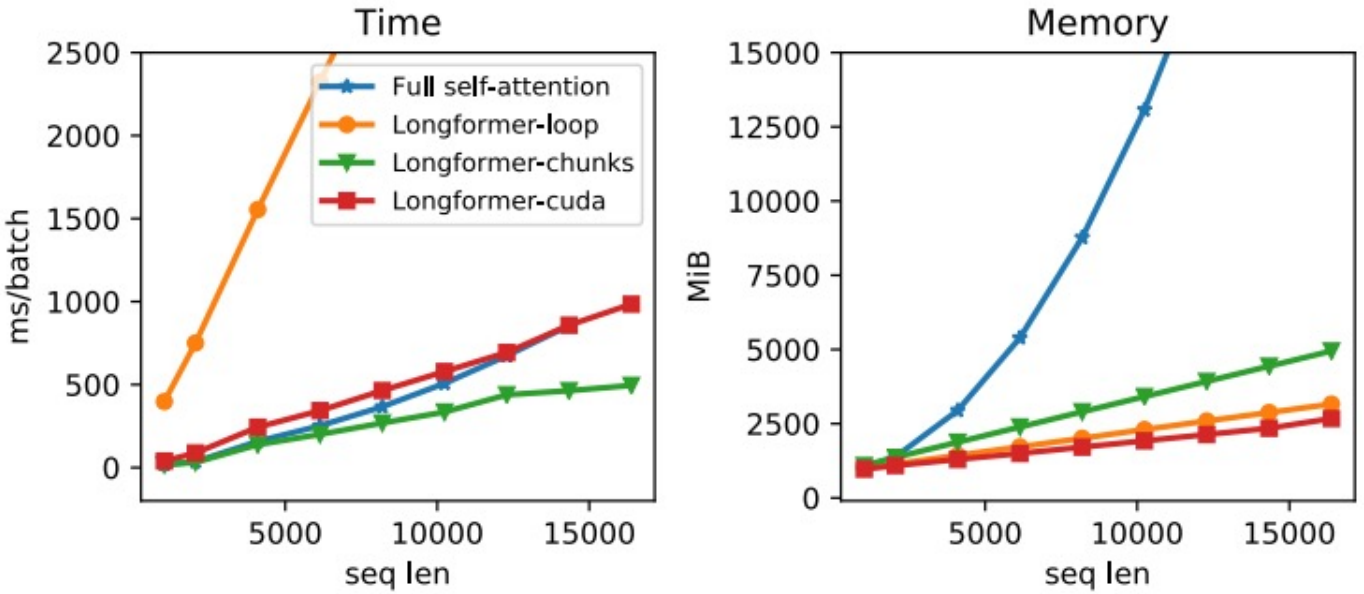
h = Encoder(English source)
French translation =
Decoder(***h***)

L = Enc. sequence length
D = Dec. sequence length

Enc. Attention = $O(L^2)$
Dec. Attention = $O(L*D + D^2)$

Vaswani et al 2017, “Attention is all you need”

Long sequence transformers



Longformer

Modification of attention mechanism in transformer
→ Reduces $O(L^2)$ to $O(L*(w + c))$ where c, w << L

Beltagy et al (2020), Longformer: The Long-Document Transformer

Longformer

Modification of attention mechanism in transformer
→ Reduces $O(L^2)$ to $O(L*(w + c))$ where $c, w << L$

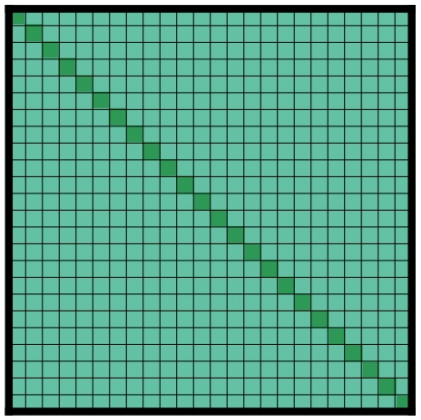
Combination of a local “sliding window” attention and a “global” attention
→ Local context is more important then long range context

Beltagy et al (2020), Longformer: The Long-Document Transformer

Longformer

Modification of attention mechanism in transformer
→ Reduces $O(L^2)$ to $O(L*(w + c))$ where c, w << L

Combination of a local “sliding window” attention and a “global” attention
→ Local context is more important then long range context



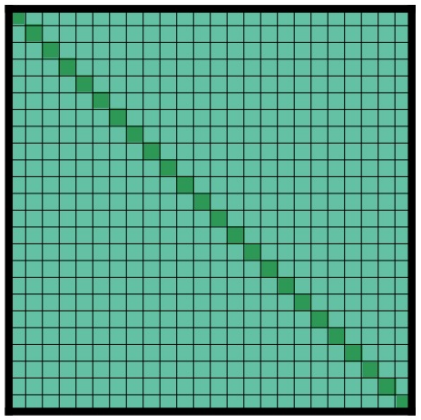
$O(L^2)$ Full attention

Beltagy et al (2020), Longformer: The Long-Document Transformer

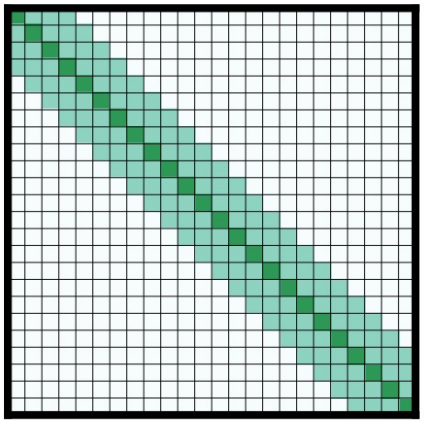
Longformer

Modification of attention mechanism in transformer
→ Reduces $O(L^2)$ to $O(L*(w + c))$ where c, w << L

Combination of a local “sliding window” attention and a “global” attention
→ Local context is more important then long range context



$O(L^2)$ Full attention



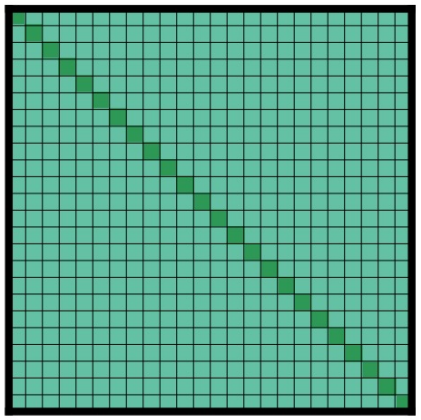
O(L*w) Sliding window

Beltagy et al (2020), Longformer: The Long-Document Transformer

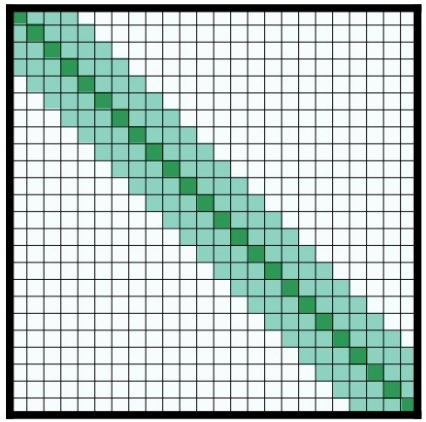
Longformer

Modification of attention mechanism in transformer
→ Reduces $O(L^2)$ to $O(L*(w + c))$ where c, w << L

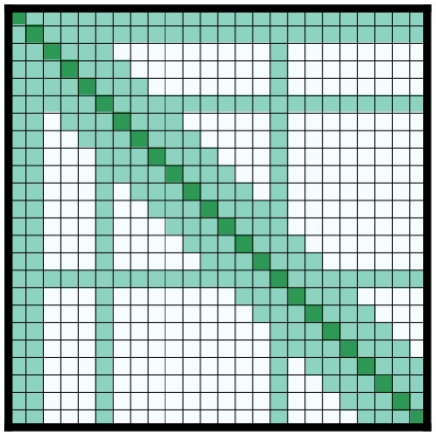
Combination of a local “sliding window” attention and a “global” attention
→ Local context is more important then long range context



$O(L^2)$ Full attention



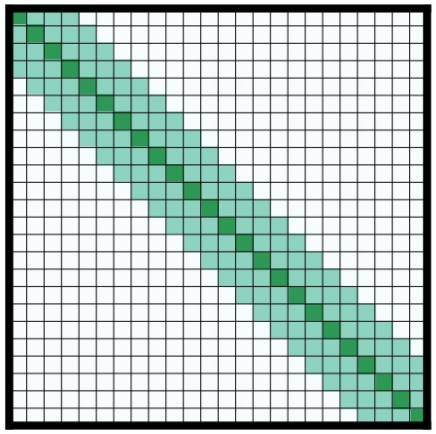
O(L*w) Sliding window

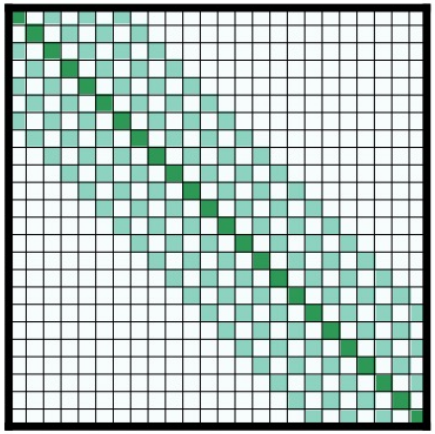


O(L*(w+c))
Window + Global

Beltagy et al (2020), Longformer: The Long-Document Transformer

Longformer - results





Character level, autoregressive, left-to-right generative language modeling

Used combination of sliding window and dilated windows

Sequence lengths of 32K characters

Model	#Param	Dev	Test
Dataset `text8`			
T12 (Al-Rfou et al., 2018)	44M	-	1.18
Adaptive (Sukhbaatar et al., 2019)	38M	1.05	1.11
BP-Transformer (Ye et al., 2019)	39M	-	1.11
Our Longformer	41M	1.04	**1.10**
Dataset `enwik8`			
T12 (Al-Rfou et al., 2018)	44M	-	1.11
Transformer-XL (Dai et al., 2019)	41M	-	1.06
Reformer (Kitaev et al., 2020)	-	-	1.05
Adaptive (Sukhbaatar et al., 2019)	39M	1.04	1.02
BP-Transformer (Ye et al., 2019)	38M	-	1.02
Our Longformer	41M	1.02	**1.00**

Longformer - results

Model	base	large
RoBERTa (seqlen: 512)	1.846	1.496
Longformer (seqlen: 4,096)	10.299	8.738
+ copy position embeddings	1.957	1.597
+ 2K gradient updates	1.753	1.414
+ 65K gradient updates	1.705	1.358

Masked language modeling BPC improves (decreases) when using longer sequences, and careful initialization allows re-use of existing models.

Longformer - results

Downstream task performance improves across the board for long sequence tasks.

	QA			Coref.	Classification	
Model	WikiHop	TriviaQA	HotpotQA	OntoNotes	IMDB	Hyperpartisan
RoBERTa-base	72.4	74.3	63.5	78.4	95.3	87.4
Longformer-base	**75.0**	**75.2**	**64.4**	**78.6**	**95.7**	**94.8**

Sequence lengths

Wordpieces	WH	TQA	HQA	ON	IMDB	HY
avg.	1,535	6,589	1,316	506	300	705
95th pctl.	3,627	17,126	1,889	1,147	705	1,975

Beyond supervised learning

Big idea

We'd like to build a general purpose language understanding agent.

Supervised learning learns a narrowly specified model to perform one specific task. Learning new tasks requires gathering new data and training another model.

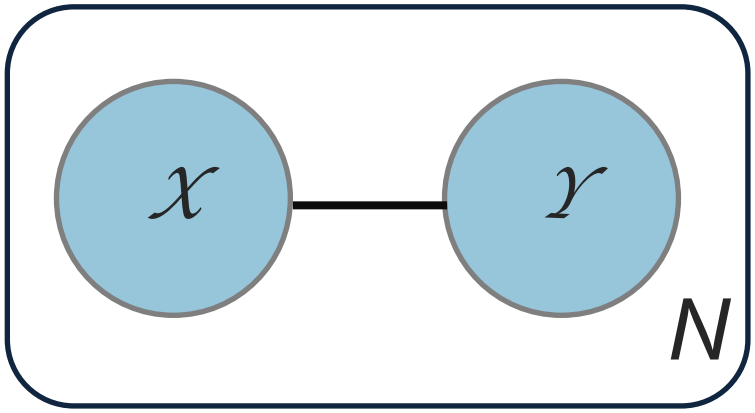
Task is learned with many labeled examples but without any explicit instructions about the underlying task: "learning from examples"

A more natural approach is "learning from task descriptions" → allows generalization to unseen tasks without labeled data by providing task description.

Shifts learning problem from fitting probability distribution to understanding semantics of description.

Learning from task descriptions

Learning from examples

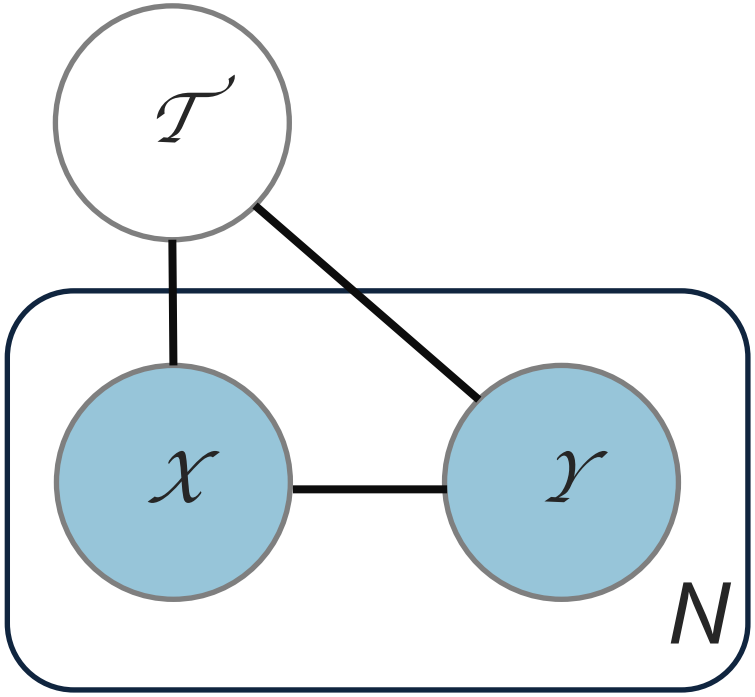


$$D = \{(x_i, y_i),\ i = 1, \ldots, N\}$$

Weller et al (2020): Learning from task descriptions

Learning from task descriptions

Learning from examples

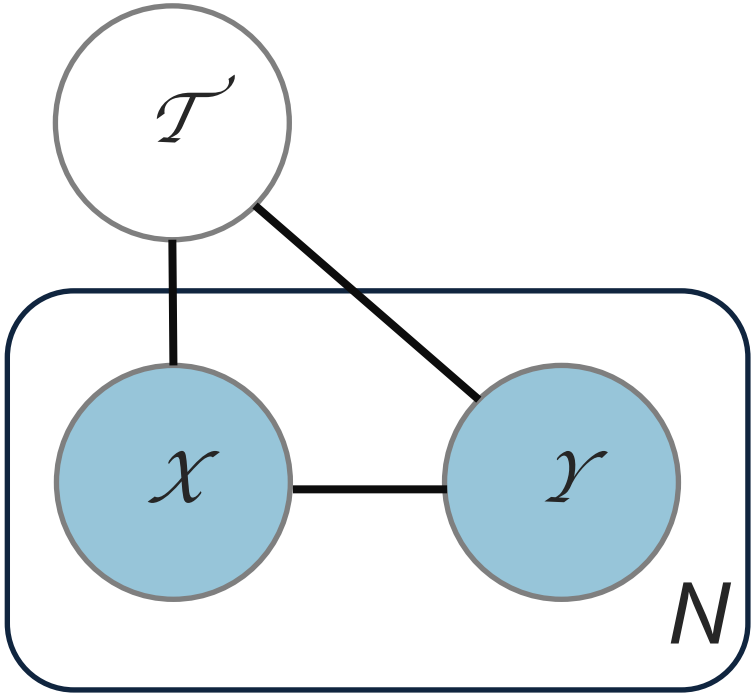


$$D = \{(x_i, y_i),\ i = 1, \ldots, N\}$$

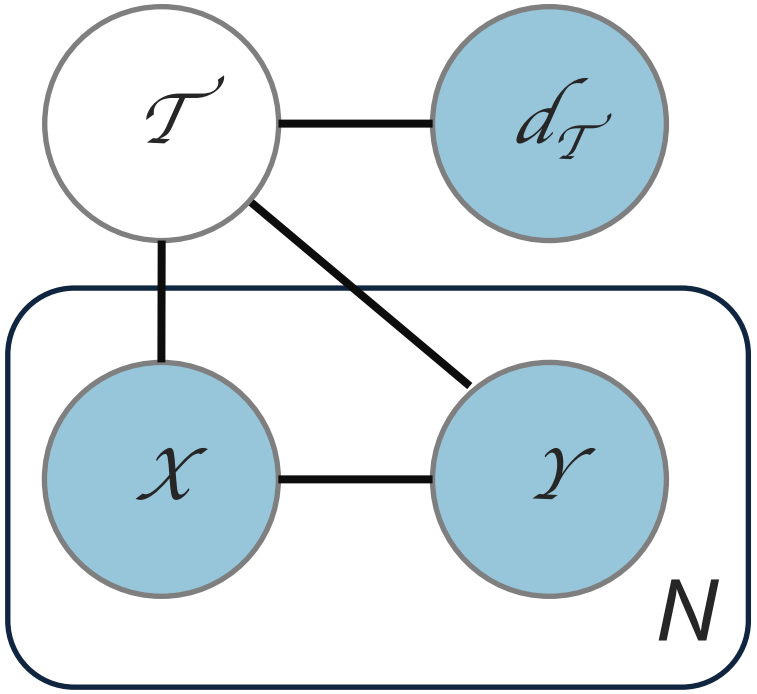
Weller et al (2020): Learning from task descriptions

Learning from task descriptions

Learning from examples



Learning from task descriptions

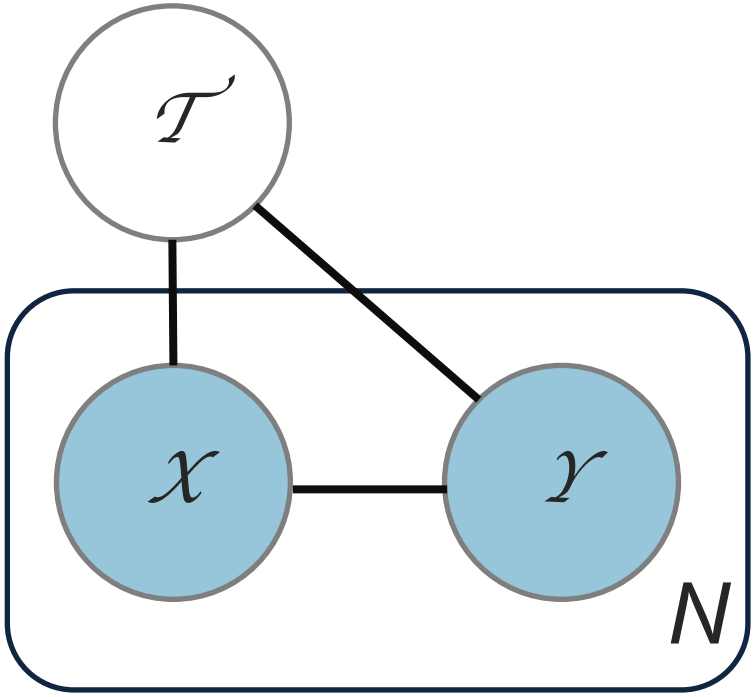


$$D = \{(x_i, y_i),\ i = 1, \ldots, N\}$$

Weller et al (2020): Learning from task descriptions

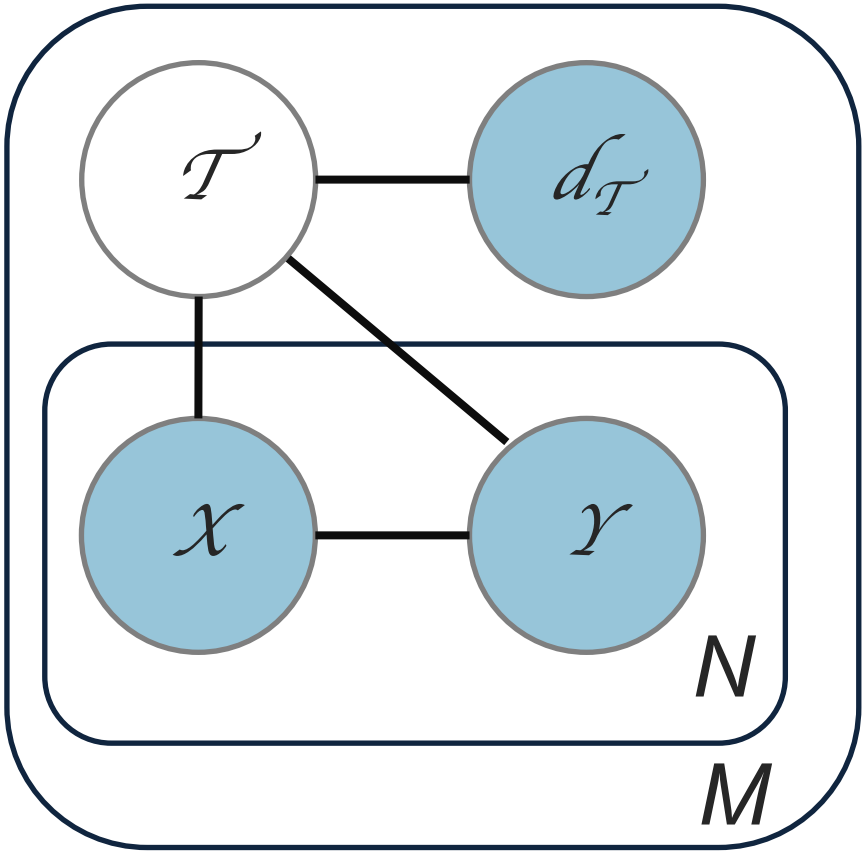
Learning from task descriptions

Learning from examples



$$D = \{(x_i, y_i),\ i = 1,\ldots,N\}$$

Learning from task descriptions



$$D_j = \{(x_i, y_i),\ i = 1,\ldots,N_j\}$$
$$D = \{(d_{\mathcal{T}}, D_j),\ j = 1,\ldots,M\}$$

Weller et al (2020): Learning from task descriptions

ZEST

Benchmark dataset for “zero-shot” generalization to unseen tasks, given natural language descriptions.

Provides >1000 tasks, each paired with 20 input/output annotations and a description.

Introduce task Competence as evaluation metric: C@75 = 30 → model can solve 30% of unseen tasks at >=75% accuracy.

Classification, entity extraction, relationship extraction.

Weller et al (2020): Learning from task descriptions

ZEST

Split all tasks into meta-train/meta-development/meta-test sets.

Use unified text-to-text approach (T5, Raffel et al, 2019) that takes task description and context as input and predicts answer.

Weller et al (2020): Learning from task descriptions

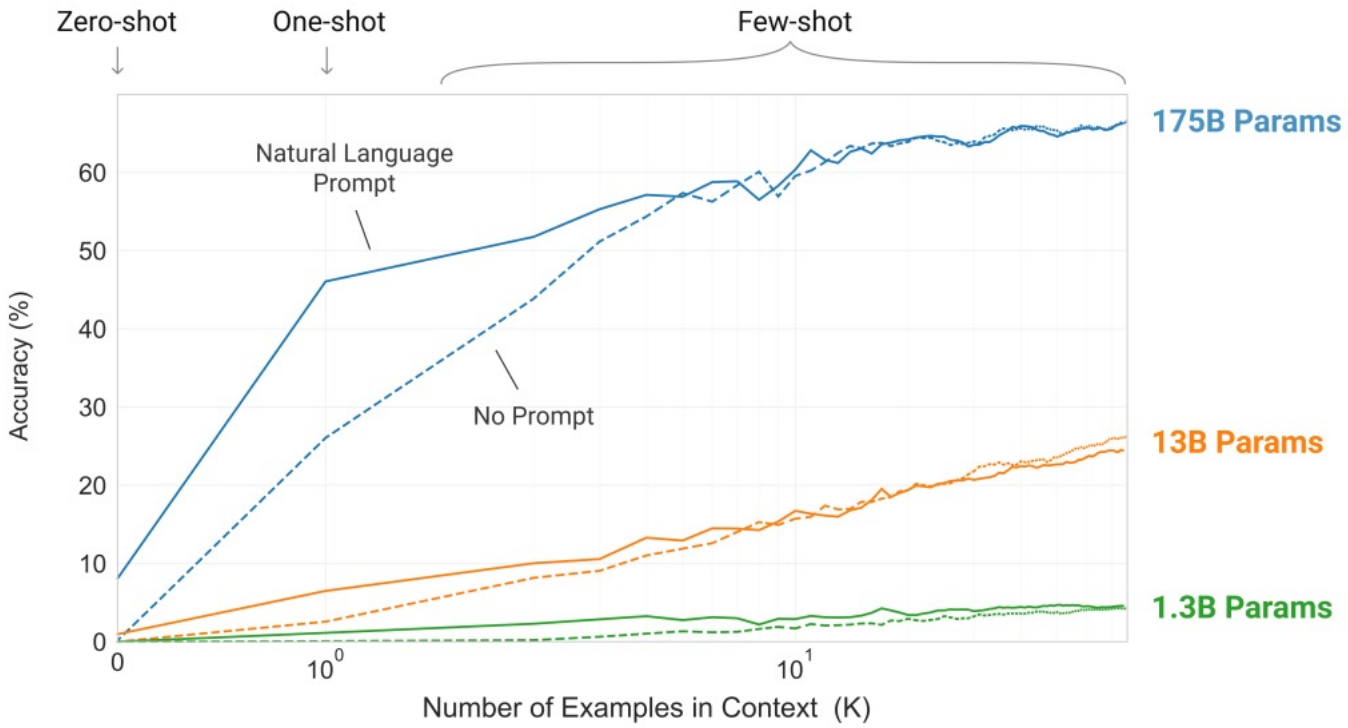
ZEST Results

	# params	Development		Test	
		Mean	C@75	**Mean**	C@75
T5-11B ZEST with multi-task learning	11B	56	35	56	28
Human estimate		-	-	74	61

State-of-the-art models have some skill but still a long way from humans.

Weller et al (2020): Learning from task descriptions

GPT-3



GPT-3 is a 175 Billion parameter neural language model trained by OpenAI
~$4.6 million of compute time to train
Non-trivial zero-shot performance with “prompt” (=task description)

Brown et al (2020): Language models are few-shot learners

ZEST Results

	# params	Development		Test	
		Mean	C@75	**Mean**	C@75
T5-11B ZEST with multi-task learning	11B	56	35	56	28
Human estimate		-	-	74	61

State-of-the-art models have some skill but still a long way from humans.

ZEST Results

	# params	Development		Test	
		Mean	C@75	**Mean**	C@75
T5-11B ZEST with multi-task learning	11B	56	35	56	28
GPT-3	175B	22	2	-	-
Human estimate		-	-	74	61

State-of-the-art models have some skill but still a long way from humans.

Conclusion

Significant recent advances in NLP to near human performance in some cases by learning contextual word representations with some variant of neural language modeling on unlabeled data then transferring to end task.

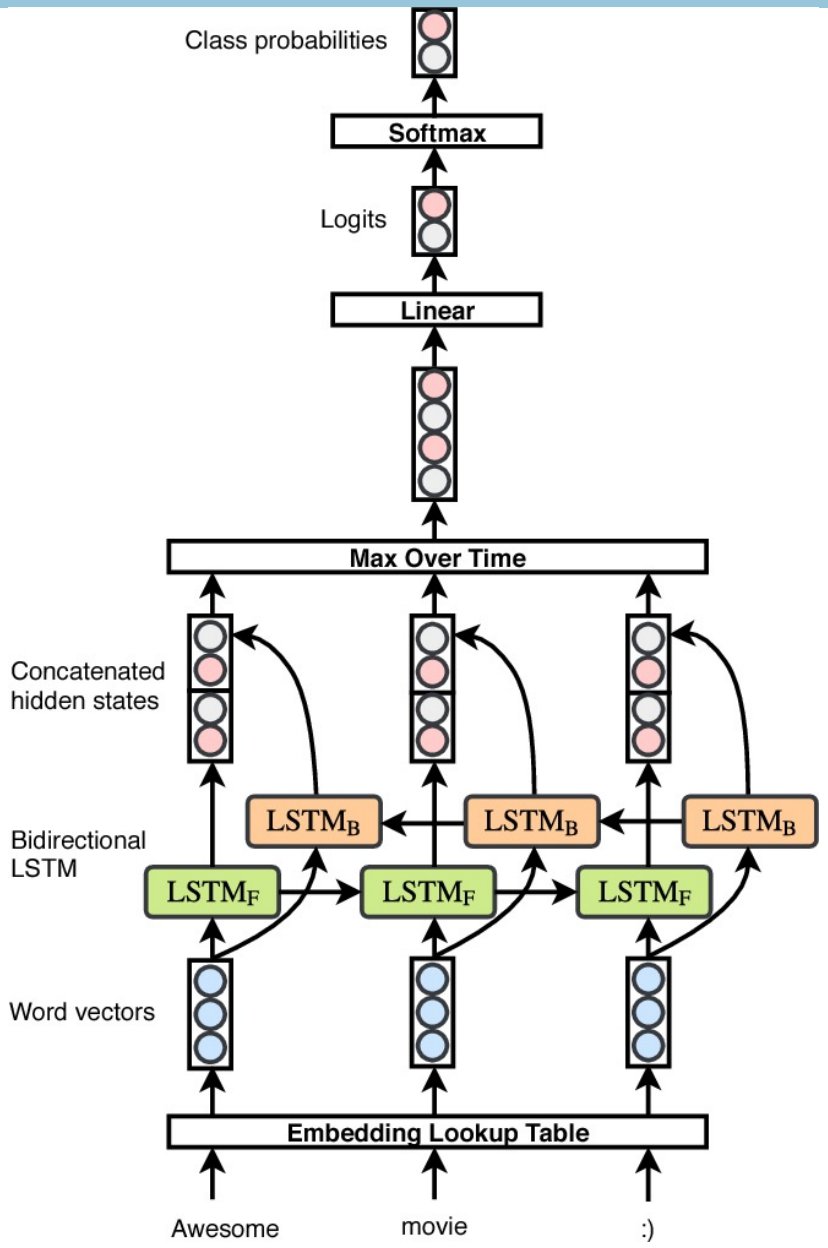
These language models learn properties of language meaning – syntax, semantics, commonsense and others.

Model performance is directly related to scale – size of model, amount of data, number of FLOPs used for training

Very large models can in some cases generalize to unseen tasks with a description, but fail in most cases.

Thank you!

Recurrent neural networks



Word vectors combined with recurrent neural networks (usually LSTM) were state-of-the-art for most NLP tasks in 2016/2017.