ELMo & BERT: Contextualized Word Representations

Presented by Dan Deutsch

Introduction

- Recently, contextual word representations have taken NLP by storm
    - ELMo & BERT
- In this talk:
    - What are contextual word representations
    - How do you train them (ELMo & BERT)
    - How do you use them

Timeline & Stats

- word2vec: 2013
    - 14,900 citations
- ELMo: Feb. 2018
    - Published June 2018
    - 1314 citations
- BERT: Oct. 2018
    - Published June 2019
    - 1545 citations

Outline

- Context-independent word representations
  - Word2vec
  - Updated NLP Pipeline
- Context-dependent word representations
  - Language Models
  - ELMo
  - Transformer
  - BERT
  - Rediscovering the NLP Pipeline

Outline

- **Context-independent word representations**
    - Word2vec
    - Updated NLP Pipeline
- Context-dependent word representations
    - Language Models
    - ELMo
    - Transformer
    - BERT
    - Rediscovering the NLP Pipeline

Context-Independent Word Representations

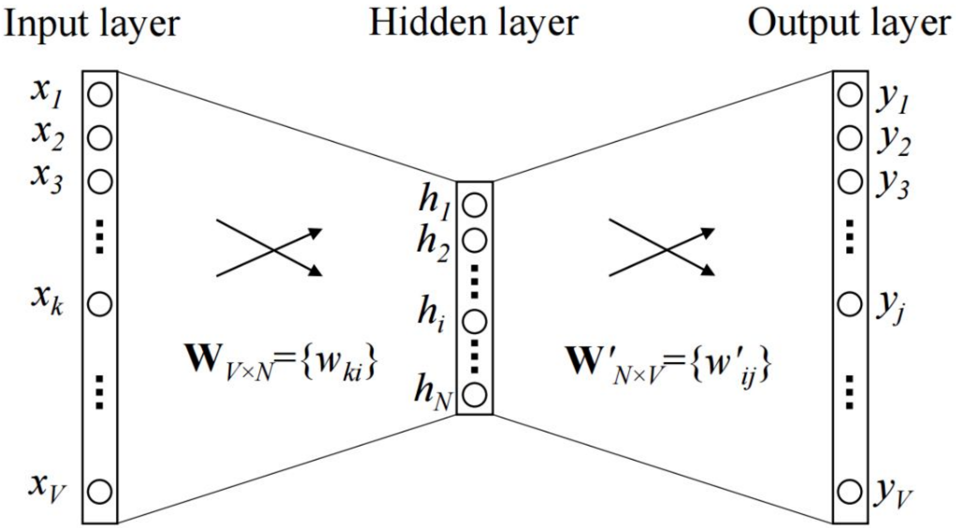
Distributional hypothesis: Use the distribution of a word (the other words it appears around) to create its representation (meaning)

https://medium.com/@keisukeumezawa/word2vec-obtain-word-embeddings-885716a56270

Context-Independent Word Representations

word2vec

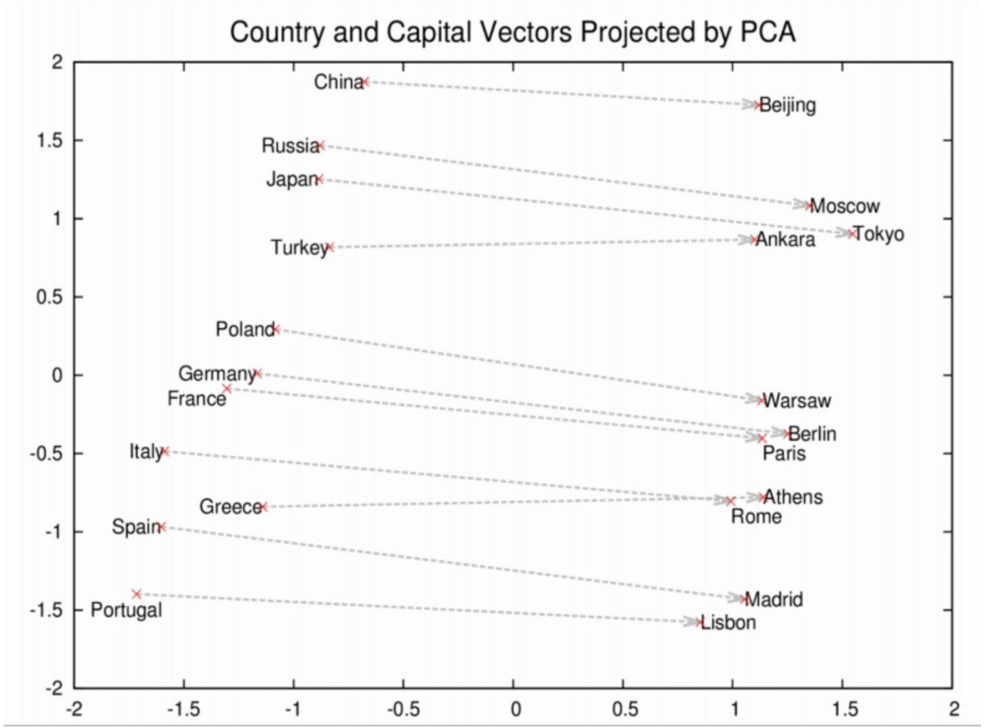
one-hot



predict the context word

https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

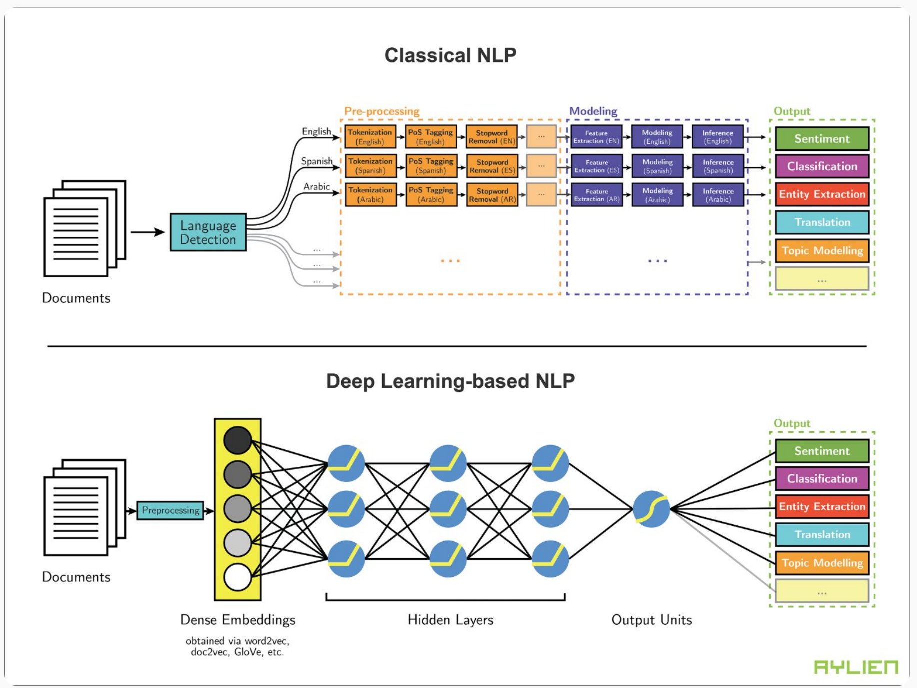
Context-Independent Word Representations



https://skymind.ai/wiki/word2vec

Updated NLP Pipeline

Start with pre-trained word embeddings, end-to-end learning



http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

Context-Independent Word Representations

- General representation of words useful across multiple tasks
- Led to updated NLP Pipeline
- Context-independent

Chico Ruiz made a spectacular play on Alusik's grounder...

Olivia De Havilland signed to do a Broadway play for Garson...

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer

Peters et al. 2018

Problem with Context-Dependent Representations

Intuitively, the representation of a word should change depending on how it's being used

Chico Ruiz made a spectacular <u>play</u> on Alusik's grounder...

Olivia De Havilland signed to do a Broadway <u>play</u> for Garson...

Contextual word embeddings try to address this

Outline

- Context-independent word representations
    - Word2vec
    - Updated NLP Pipeline
- **Context-dependent word representations**
    - Language Models
    - ELMo
    - Transformer
    - BERT
    - Rediscovering the NLP Pipeline

Language Models

Using the previous tokens, predict the next one

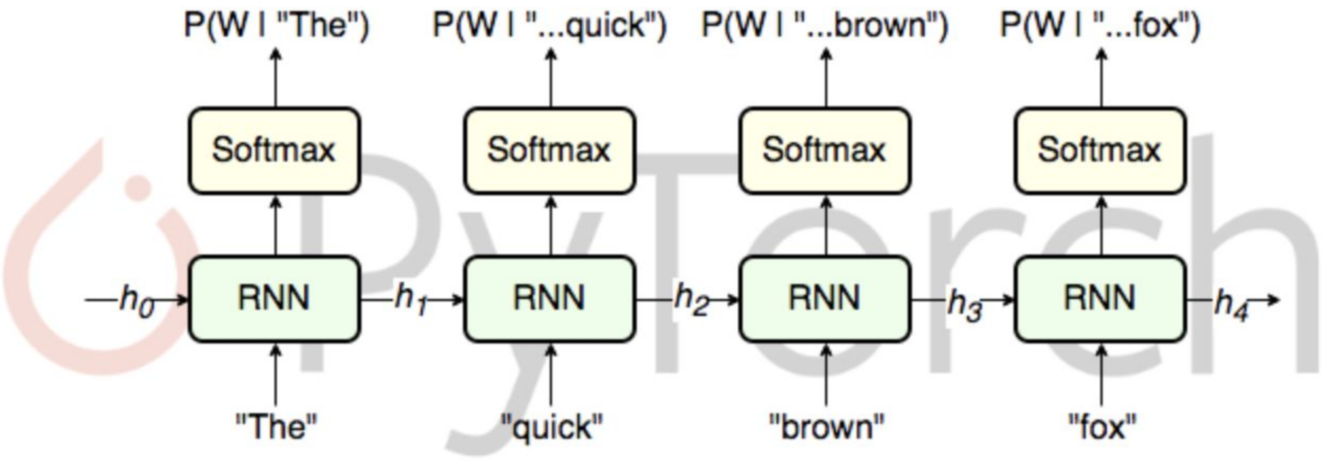
$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

Backwards

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N)$$

Language Models

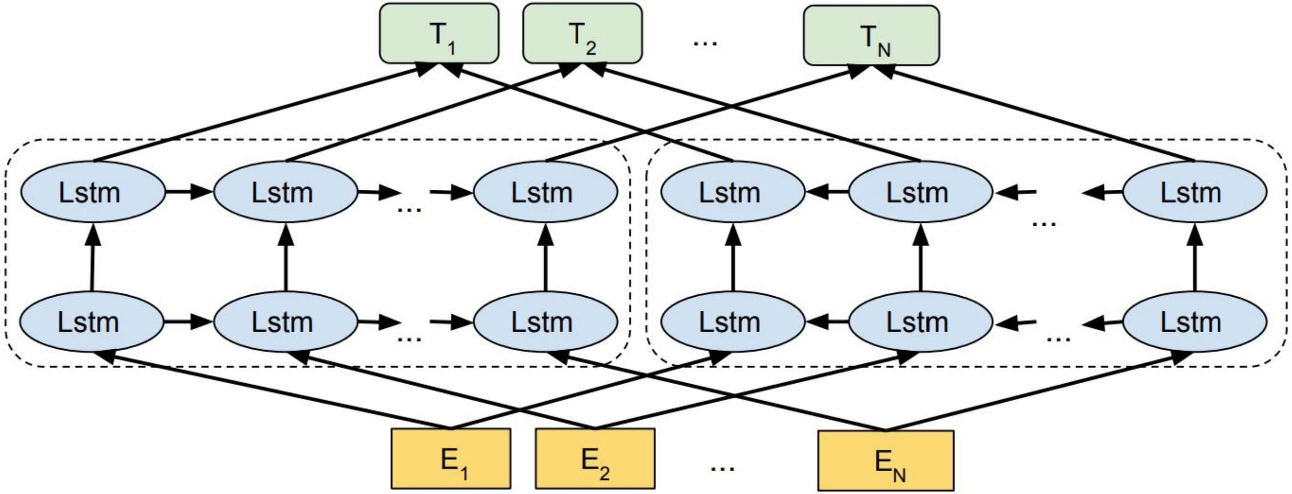
State-of-the-art uses LSTMs



http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

ELMo

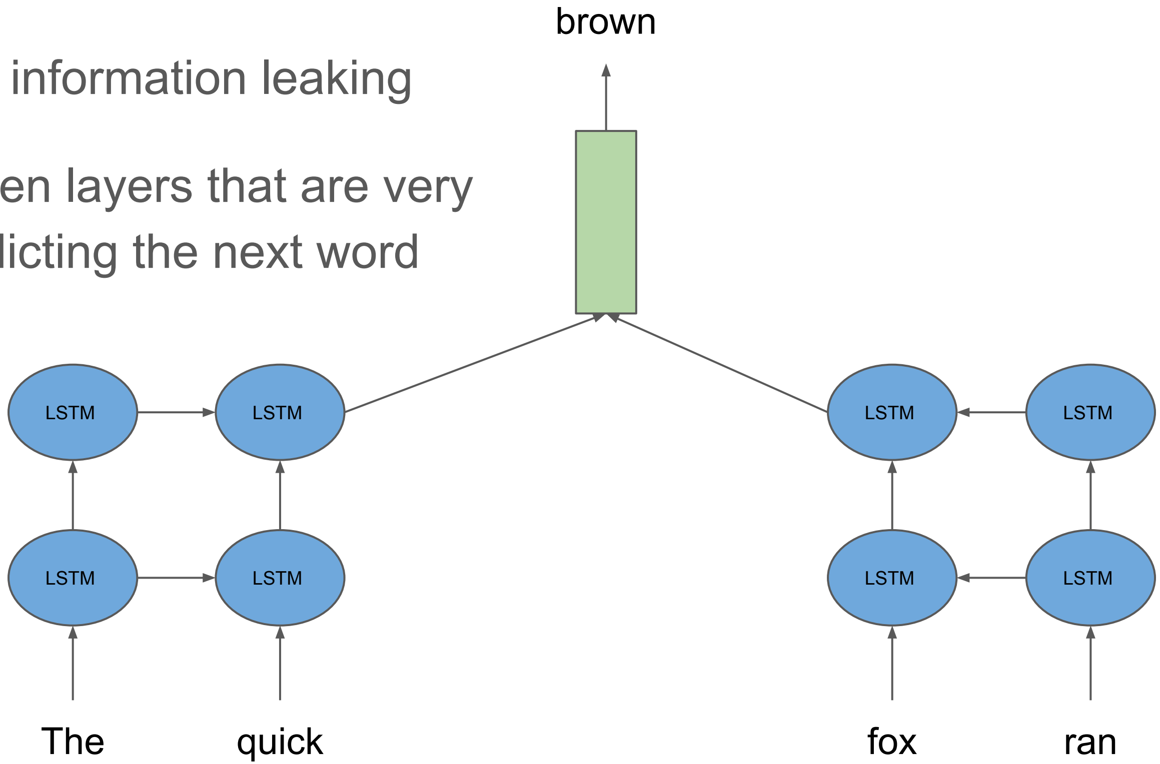
- **E**mbeddings from **L**anguage **Mo**dels
- Jointly train bidirectional LSTM Language Models
- Concatenate representations



ELMo Training

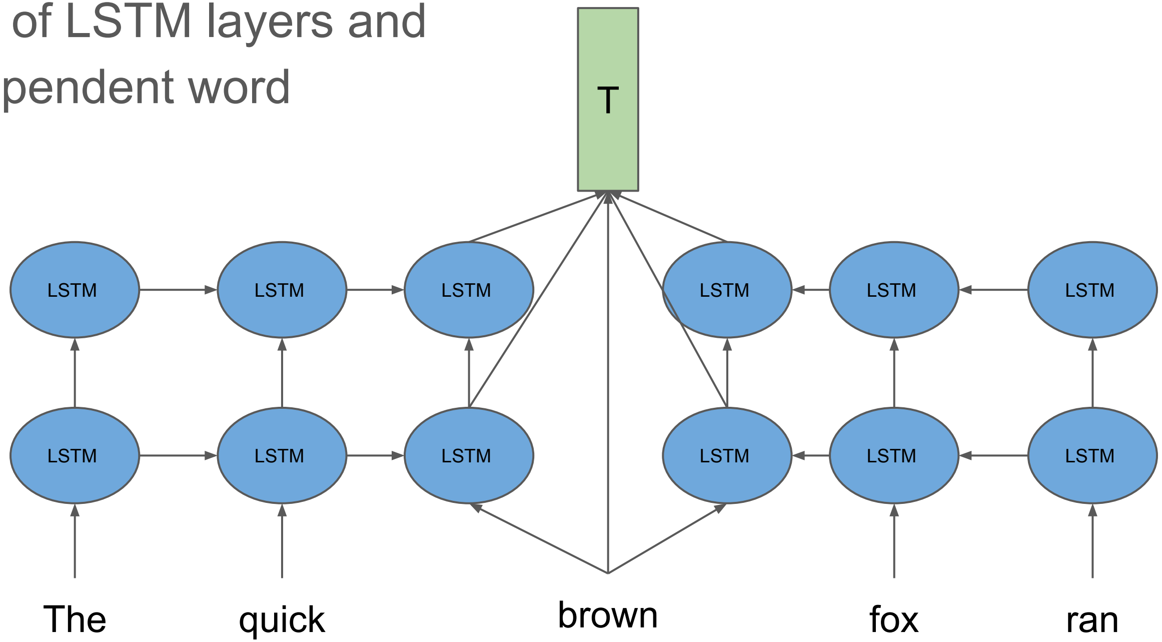
Training: No information leaking

Result: Hidden layers that are very good at predicting the next word



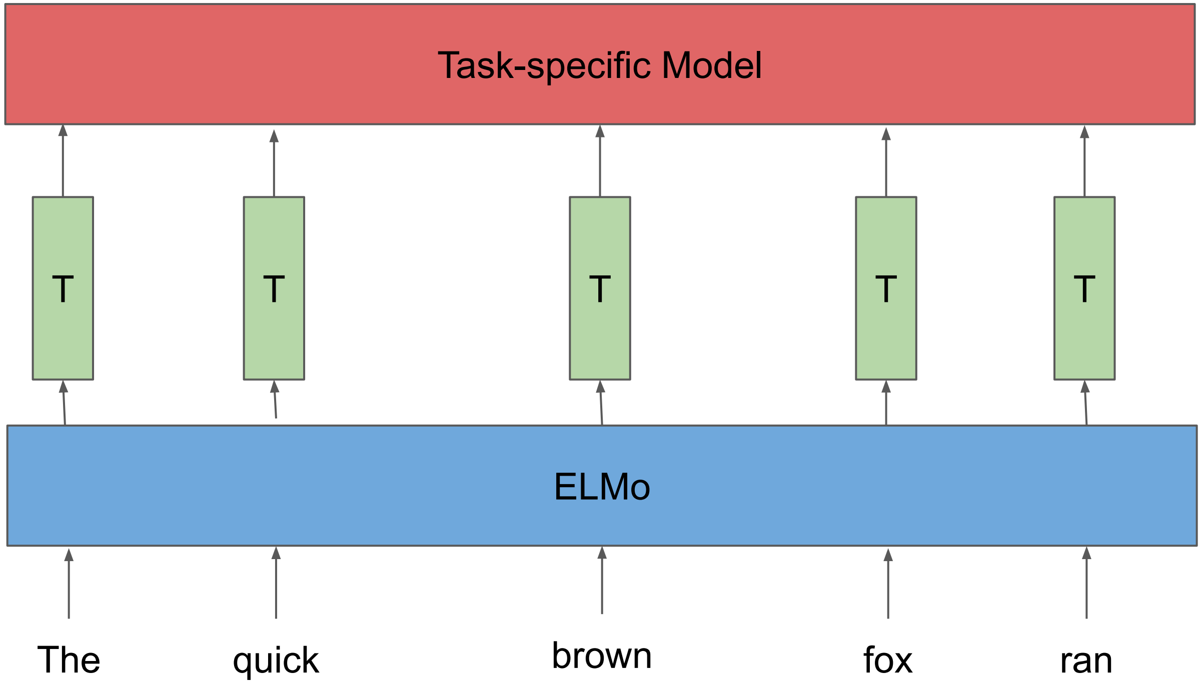
ELMo Embeddings

Contextual representation: linear combination of LSTM layers and context-independent word embedding



Downstream Tasks

Replace word2vec vectors with ELMo vectors



ELMo State-of-the-Art Results

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

SQuAD: Question-Answering
SNLI: Textual entailment
SRL: Semantic role labeling
Coref: Coreference resolution
NER: Named-entity recognition
SST-5: Sentiment analysis

Peters et al. 2018

ELMo: Contextual Word Embeddings

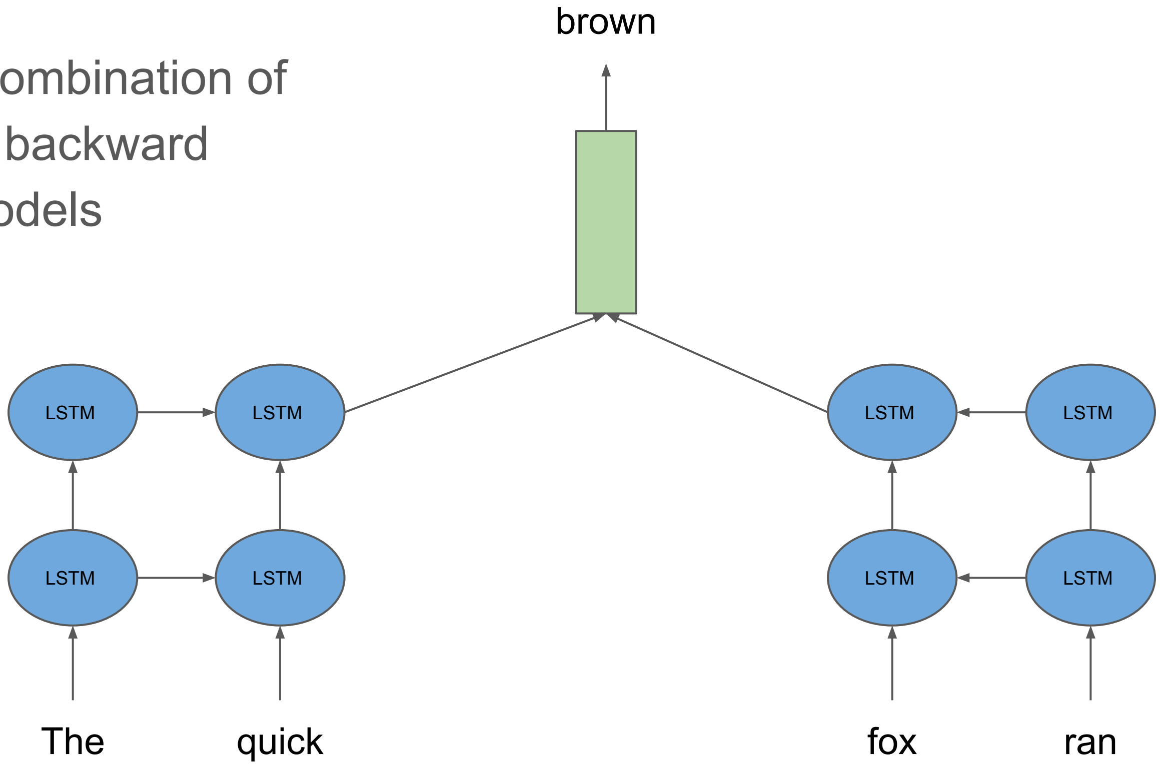
Nearest neighbors in the dataset

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {. . . }	{. . . } they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Peters et al. 2018

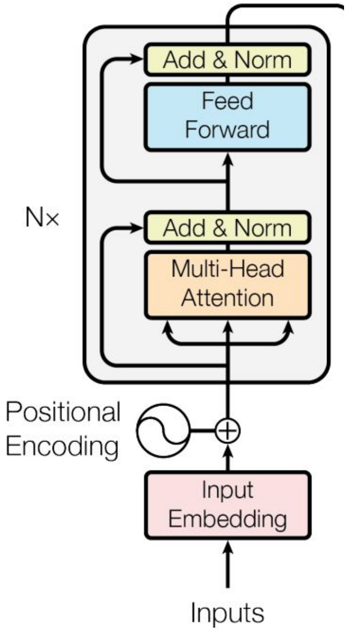
Problems with ELMo

Superficial combination of forward and backward language models



Transformer Encoder

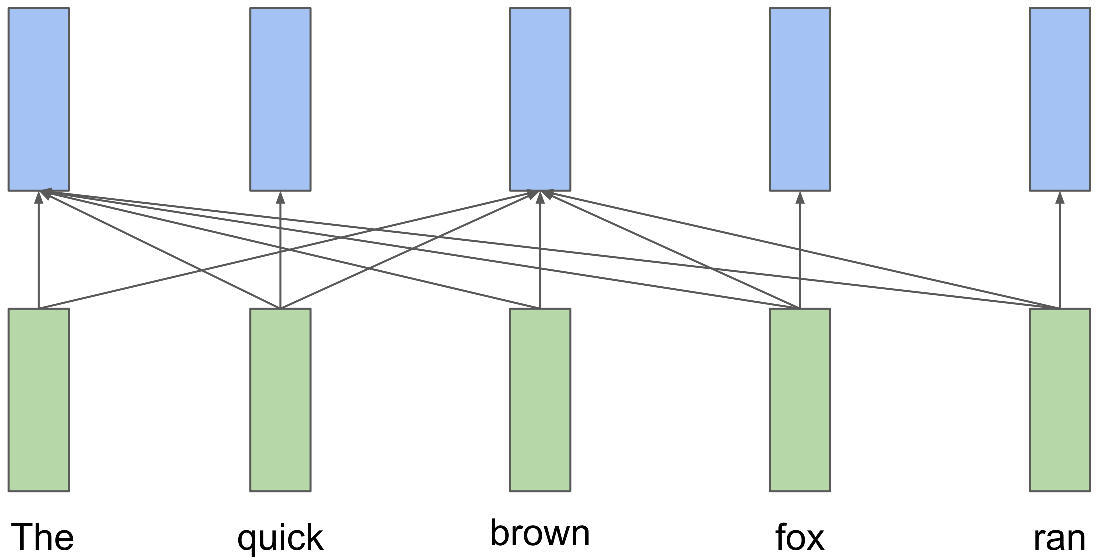
Instead of recurrent encoding,
deep feed forward



Vaswani et al. 2017

Self-Attention

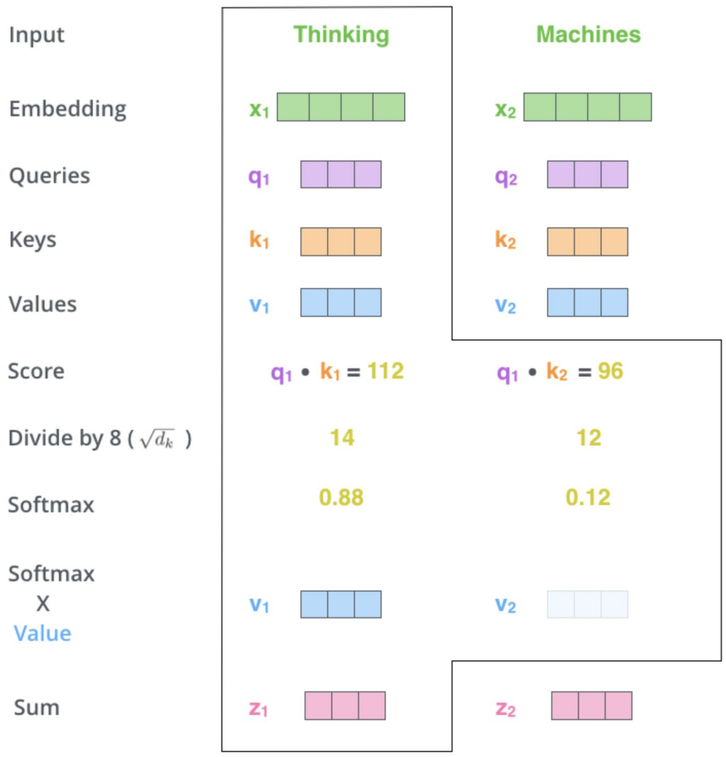
Single head: Compute all pairwise similarities, perform weighted averaging



Self-Attention

Input: Vector for every word

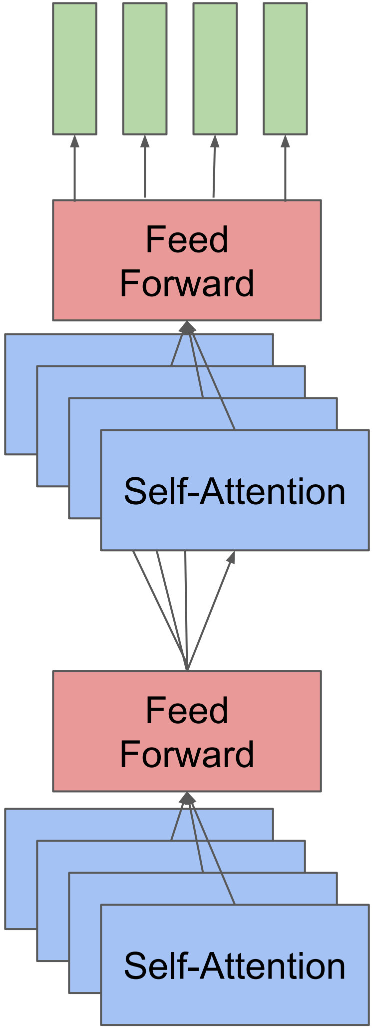
Output: Vector for every word



http://jalammar.github.io/illustrated-transformer/

Encoder

Multiple heads, multiple layers



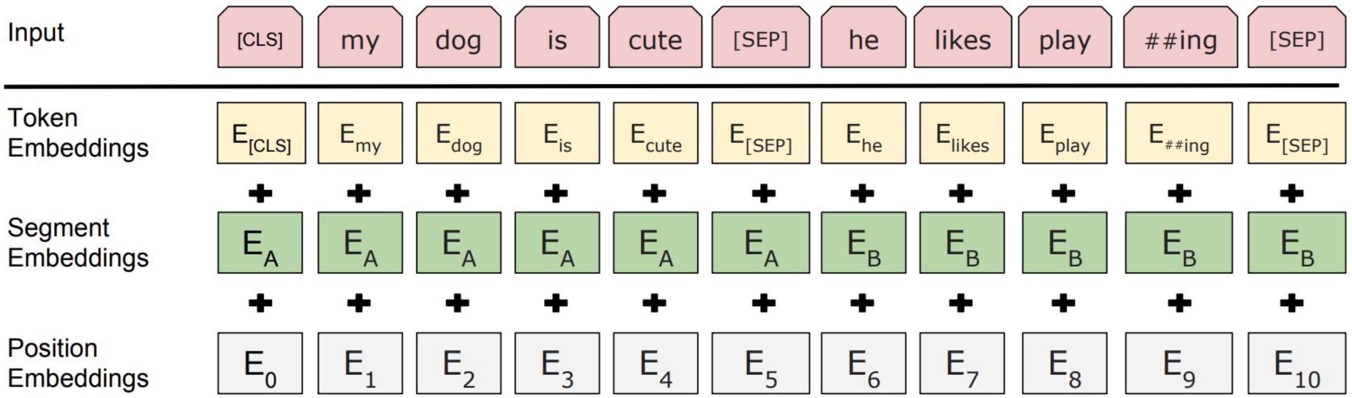
Concatenate vectors for each head (per word), pass through feed forward to get a new vector for each word

BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Encode words with a lot of transformer layers
- Masked word prediction, next sentence classification
    - **Not** traditional language modelling

BERT

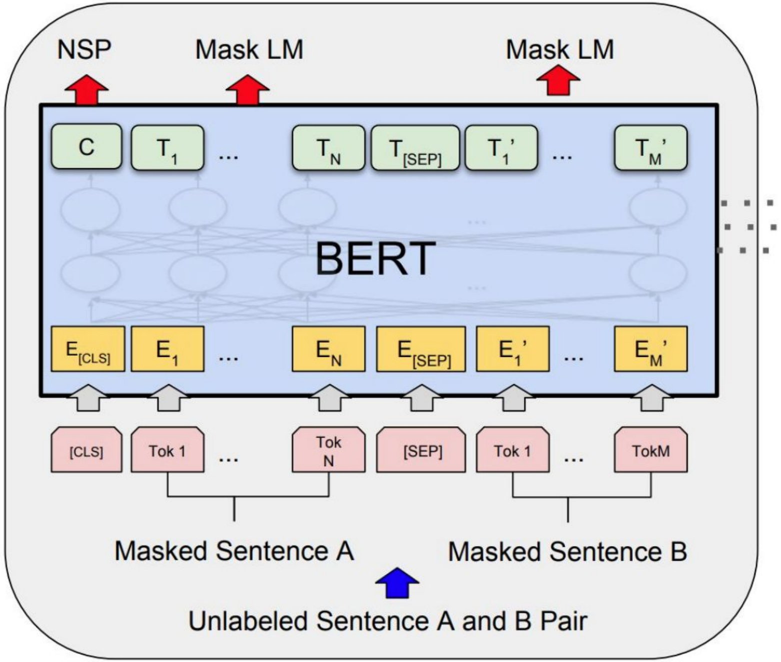
- Two sentences encoded jointly
- Special [CLS] and [SEP] tokens
- Token representation:



Devlin et al. 2019

BERT

- A lot of transformer layers
- Output final representations for each input token, T
- Special “classifier” representation C



Devlin et al. 2019

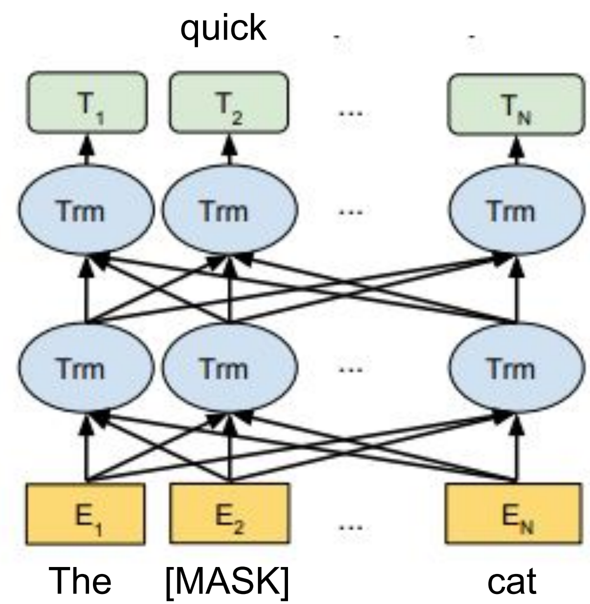
Training Objectives

- Masked word prediction
- Next sentence classification

Masked Word Prediction

- Mask 15% of the words, try to re-predict

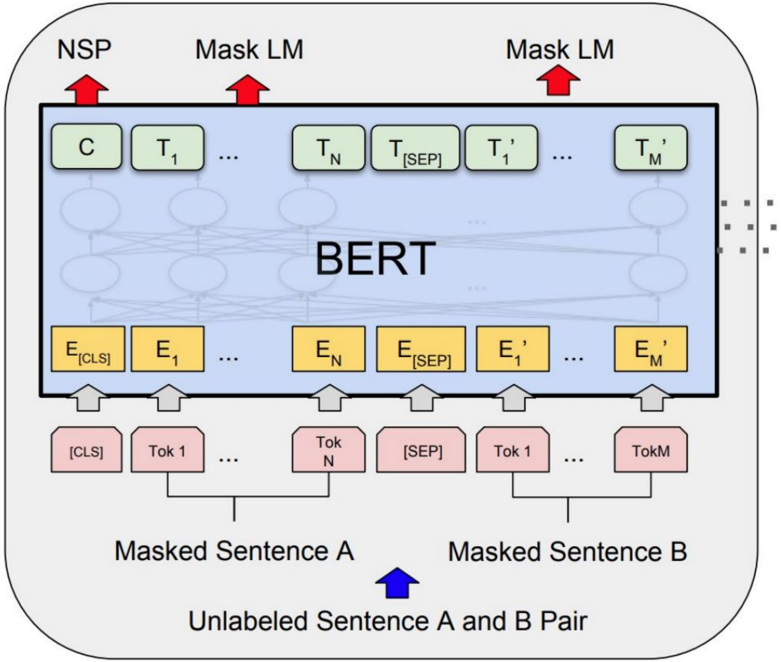
- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.



Devlin et al. 2019

Next Sentence Classification

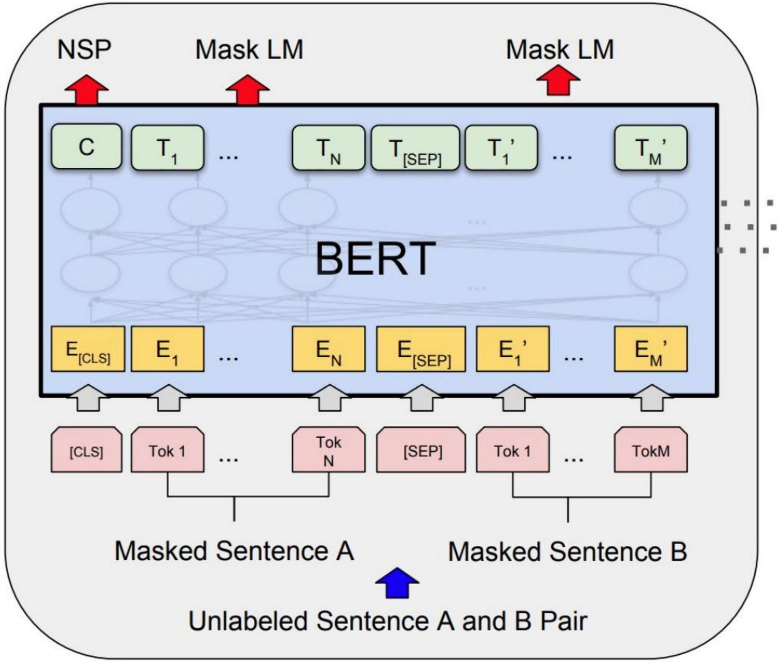
- Given sentences A and B, predict whether B followed A in the original data
- Negative sampling



Devlin et al. 2019

Training Objectives

- Masked word prediction
  - Replace 15% of the input words with [MASK], try to re-predict hidden word
  - 80% of those [MASK], 10% a random word, 10% the real word
- Next sentence prediction
  - Given sentences A and B, predict whether B followed A in the original corpus with C representation
  - Negative sampling

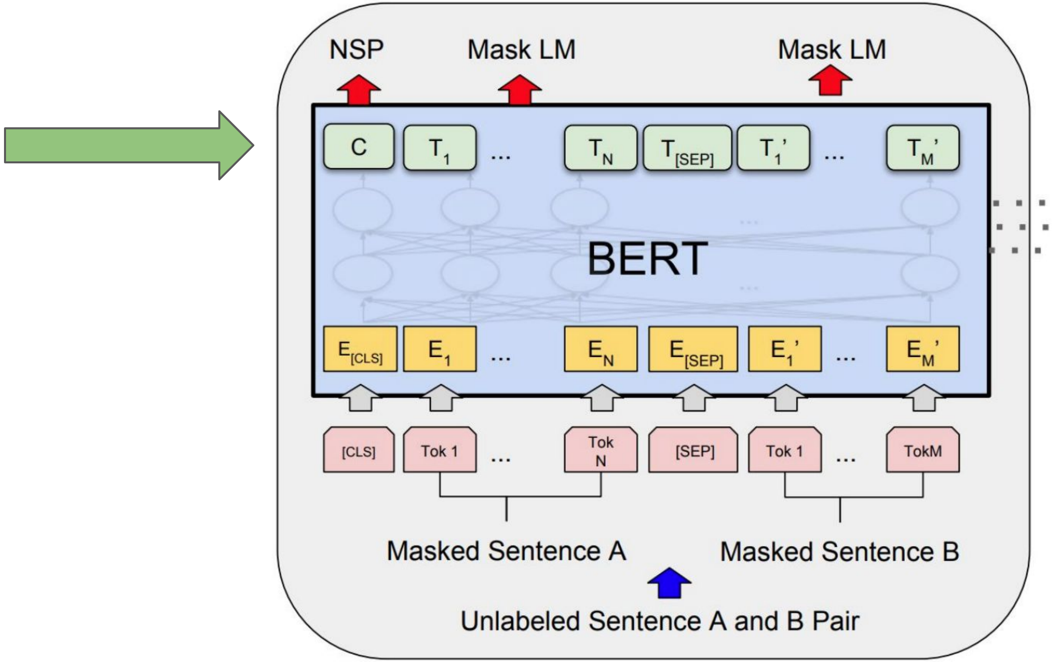


Devlin et al. 2019

Training Details

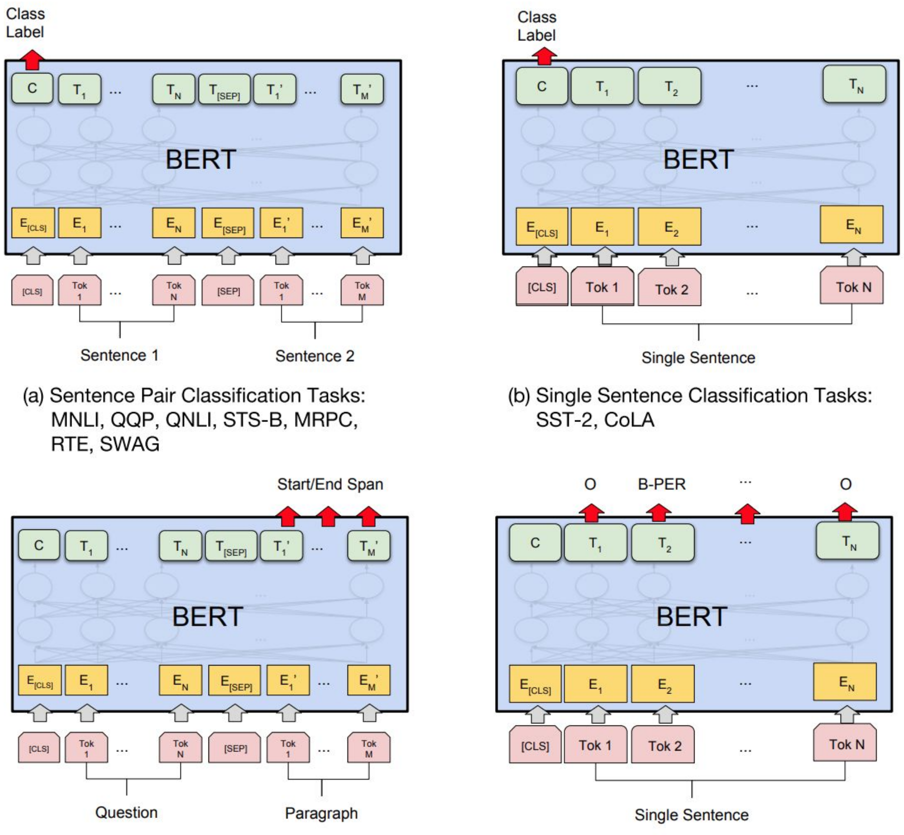
- BookCorpus (800M words) and English Wikipedia (2,500M words)
- Largest model 340M parameters and 24 Transformer layers and 16 self-attention heads
- 16 TPUs over 4 days

BERT Embeddings



Downstream Tasks

- Encode text and null sentence, use like word2vec
- Specialized setup for tasks with text pairs



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin et al. 2019

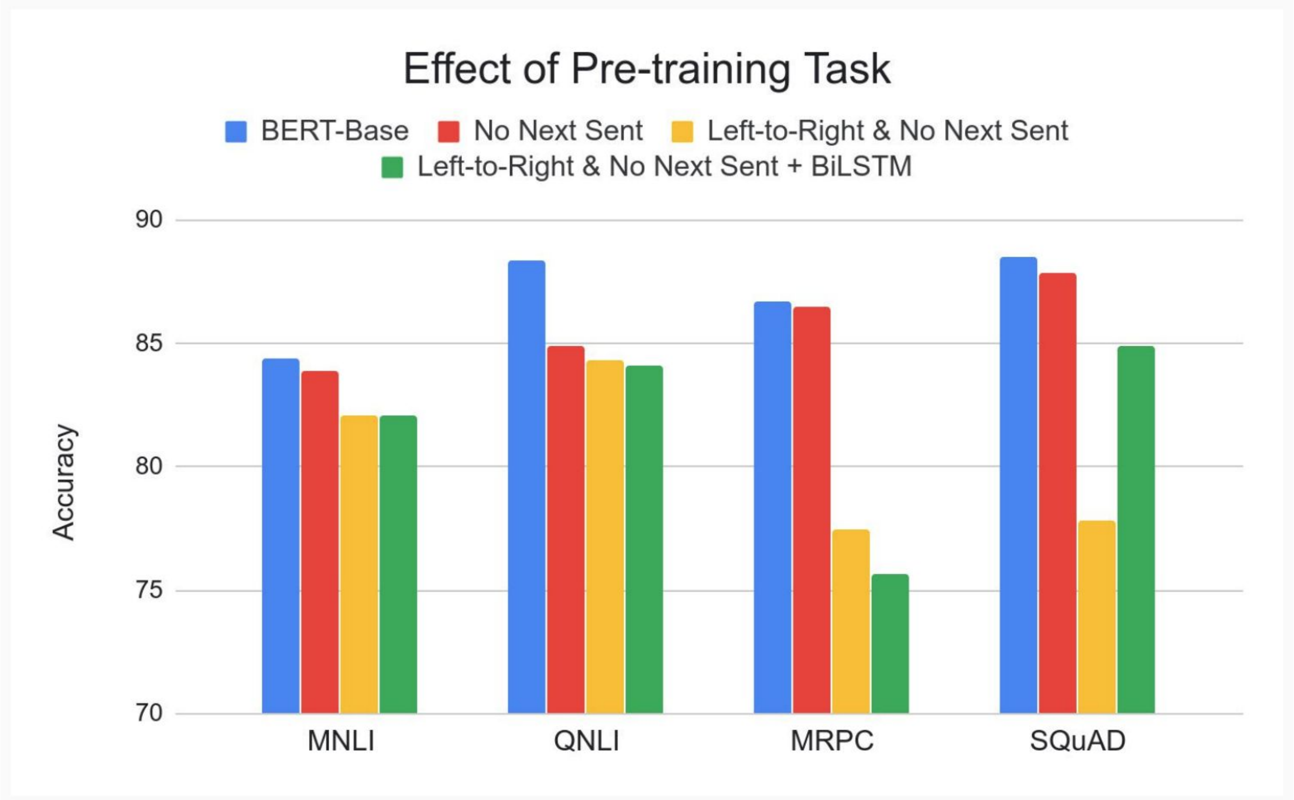
State-of-the-Art Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MNLI: Multi-genre entailment
QQP: Predict if two questions are semantically equivalent
QNLI: Question-Answering as binary classification task
SST-2: Sentiment prediction
CoLA: Prediction acceptability of a sentence
STS-B: Predict how similar two sentences are
MRPC: Predict if two phrases are paraphrases
RTE: Entailment with little training data

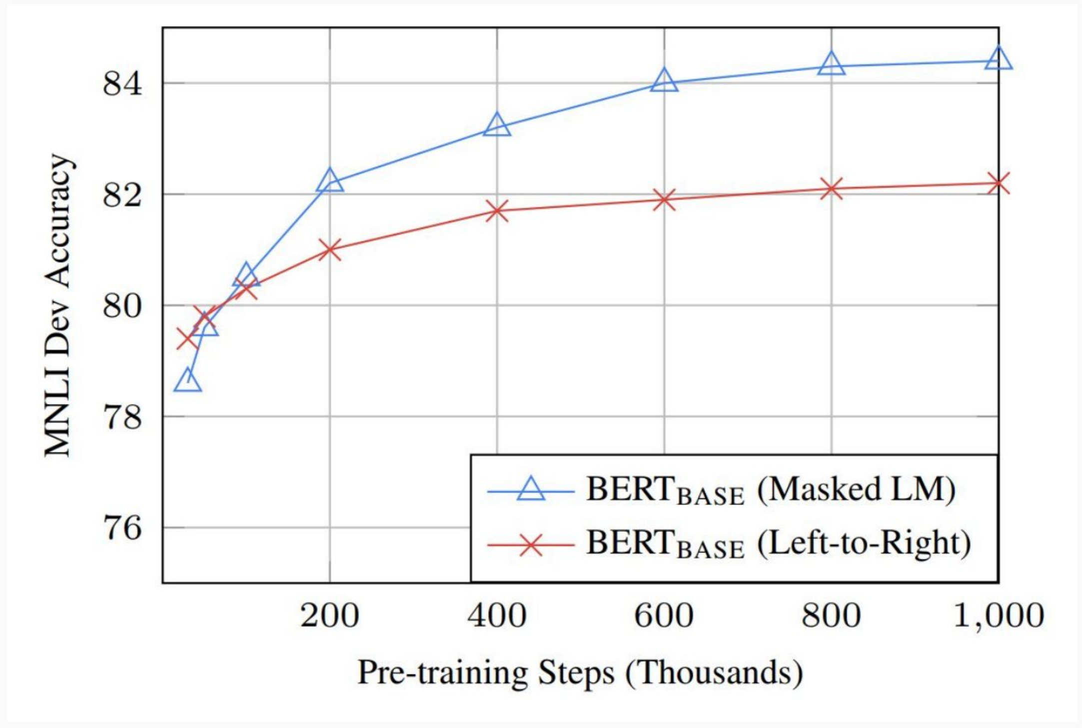
Devlin et al. 2019

Model Ablation



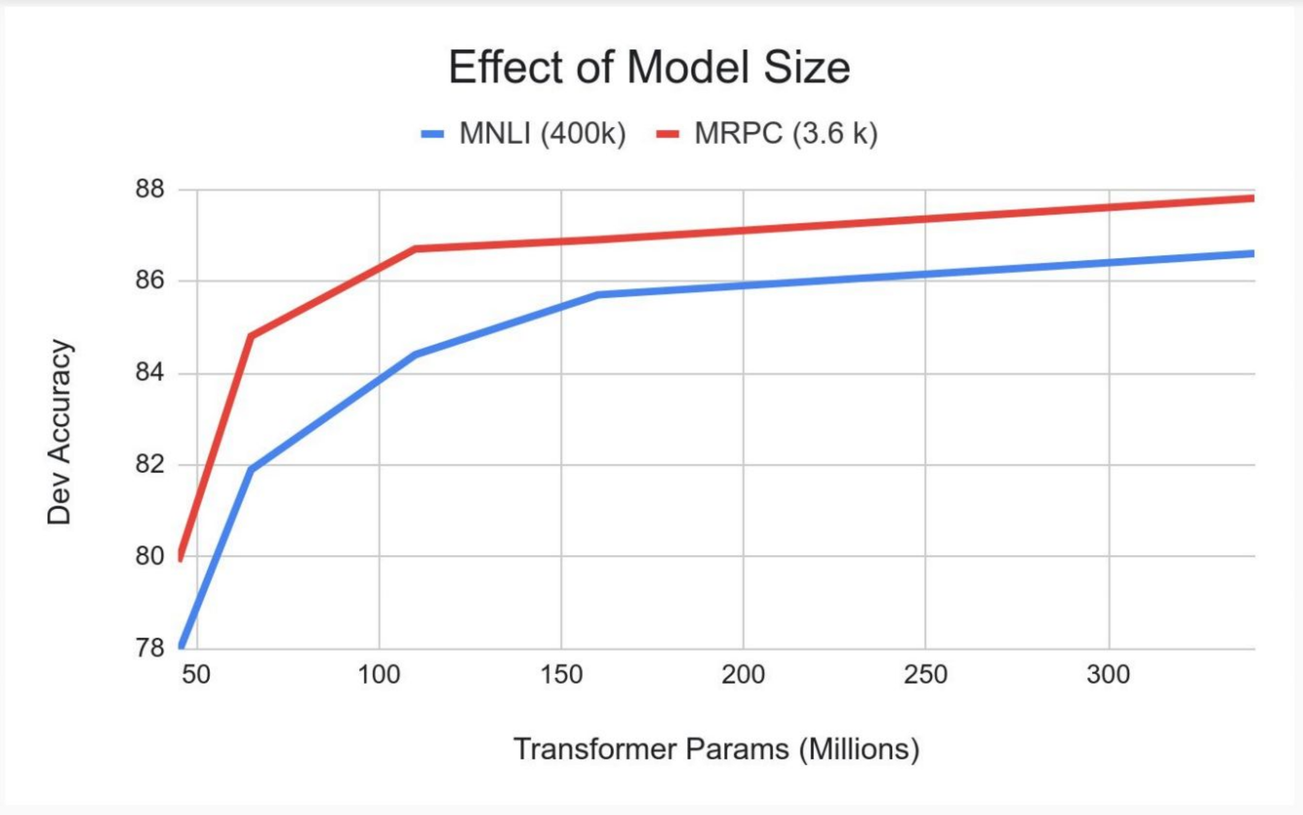
Devlin et al. 2019

Pre-Training Iterations



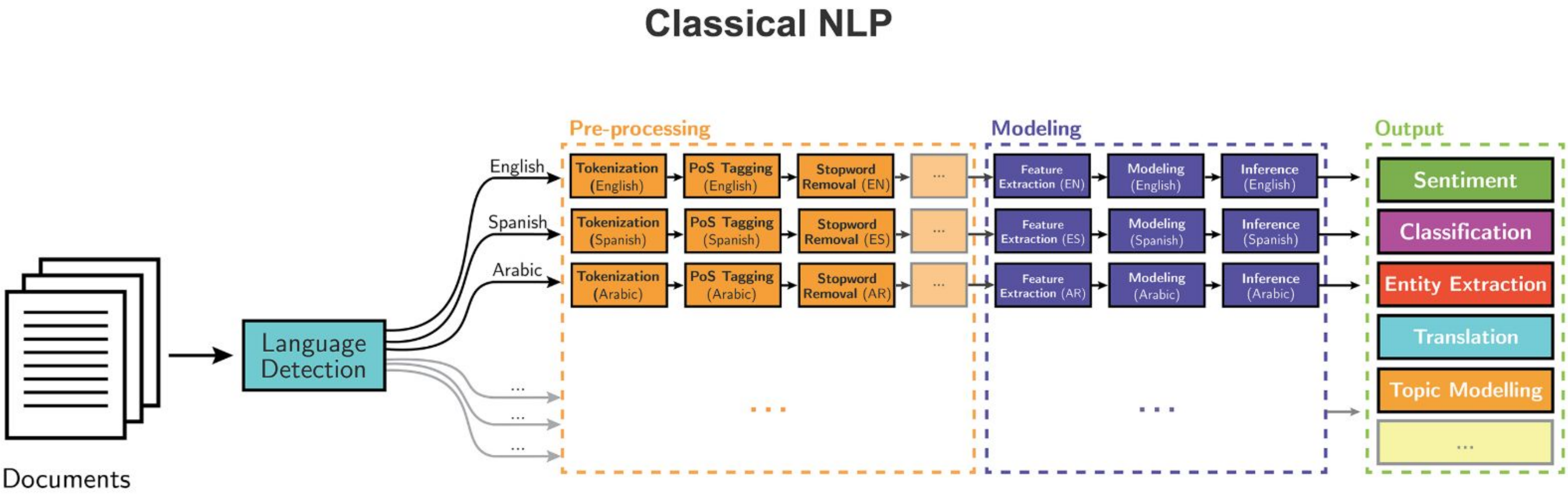
Devlin et al. 2019

Model Size



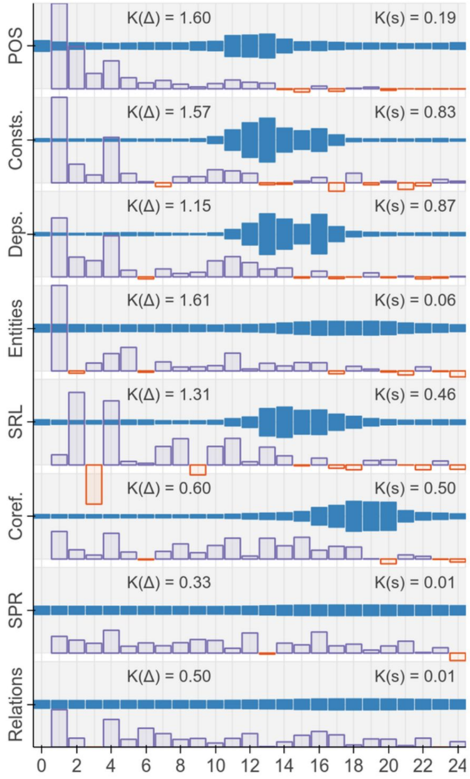
Devlin et al. 2019

Rediscovering the NLP Pipeline



http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

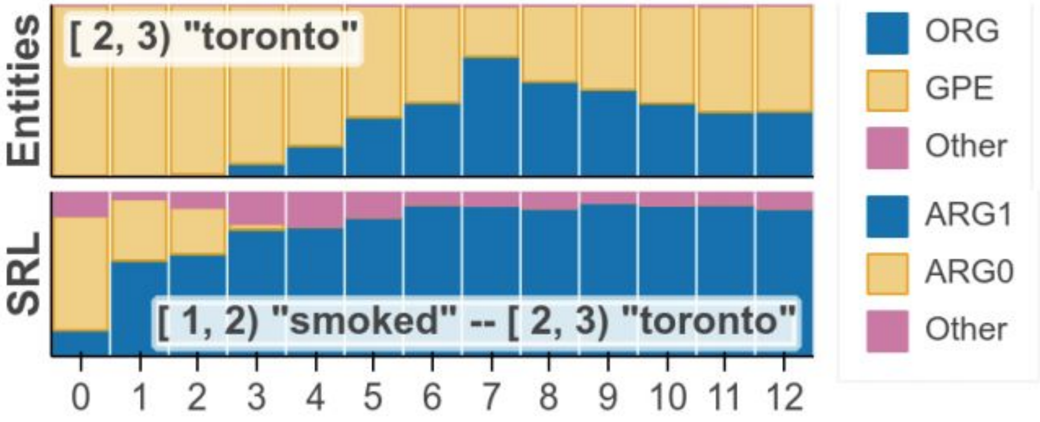
Rediscovering the NLP Pipeline



Tenney et al. 2019

Rediscovering the NLP Pipeline

(a) he smoked **toronto** in the playoffs with six hits, seven walks and eight stolen bases ...



Tenney et al. 2019

References

- “Distributed Representations of Words and Phrases and their Compositionality” Mikolov et al. 2013
- “Attention Is All You Need” Vaswani et al. 2017
- “Deep contextualized word representations” Peters et al. 2018
- “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” Devlin et al. 2019
- “BERT Rediscovers the Classical NLP Pipeline” Tenney et al. 2019