Lexicon, Syntax, Semantics IIb: Modeling Meaning

Contextualized word embeddings

Eva Maria Vecchi

Center for Information and Language Processing
LMU Munich

July 16, 2020

bank

The man was accused of robbing a bank.

The man went fishing by the bank of the river.

bank

The man was accused of robbing a bank.

The man went fishing by the bank of the river.

- **word2vec**: same word embedding for the word “bank” in both sentences
  - each word has a fixed representation under Word2Vec regardless of the context within which the word appears

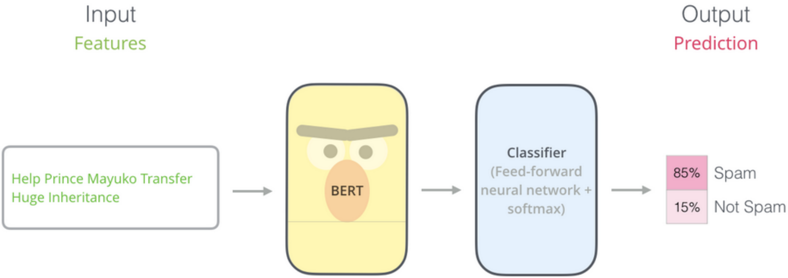
bank

The man was accused of robbing a bank.

The man went fishing by the bank of the river.

- **word2vec**: same word embedding for the word "bank" in both sentences
  - each word has a fixed representation under Word2Vec regardless of the context within which the word appears
- **contextualized embedding** (e.g. BERT): word embedding for "bank" would be different for each sentence
  - produces word representations that are dynamically informed by the words around them

Task Example: Sentence Classification

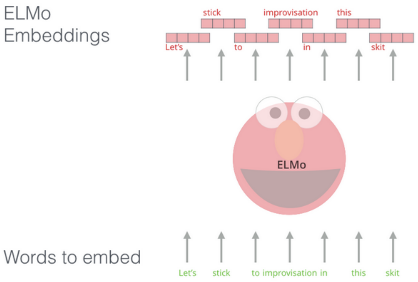


Other examples

- **Sentiment Analysis**
  - Input: Movie/Product review. Output: is the review positive or negative?
  - Example dataset: SST
- **Fact-checking**
  - Input: sentence. Output: “Claim” or “Not Claim”
  - More ambitious/futuristic example:
    - Input: Claim sentence. Output: “True” or “False”
  - Full Fact is an organization building automatic fact-checking tools for the benefit of the public. Part of their pipeline is a classifier that reads news articles and detects claims (classifies text as either “claim” or “not claim”) which can later be fact-checked (by humans now, by with ML later, hopefully).

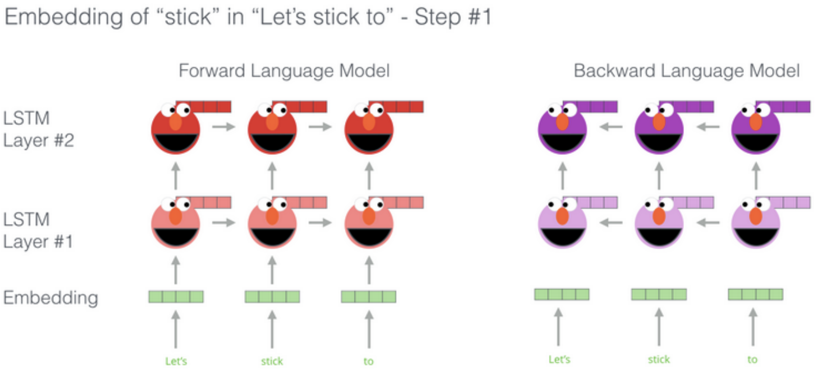
ELMo: Context Matters

ELMo Embeddings

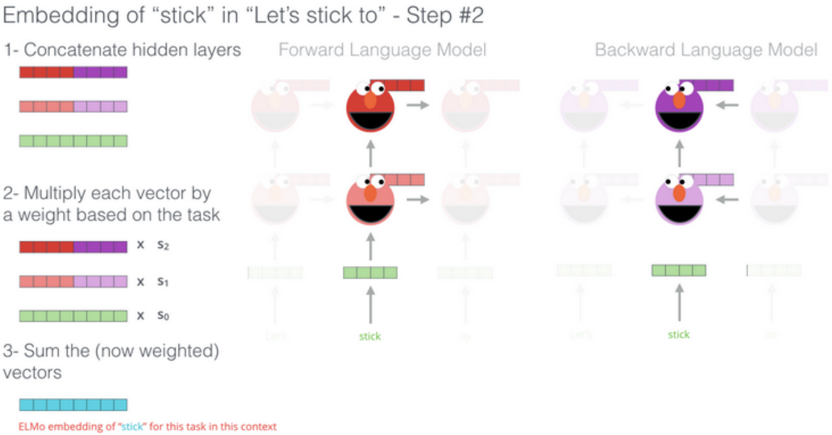


- Instead of using a fixed embedding for each word, ELMo looks at the entire sentence before assigning each word in it an embedding.
- Uses a bi-directional LSTM (e.g. language model) trained on a specific task to be able to create those embeddings

ELMo Embeddings



ELMo Embeddings

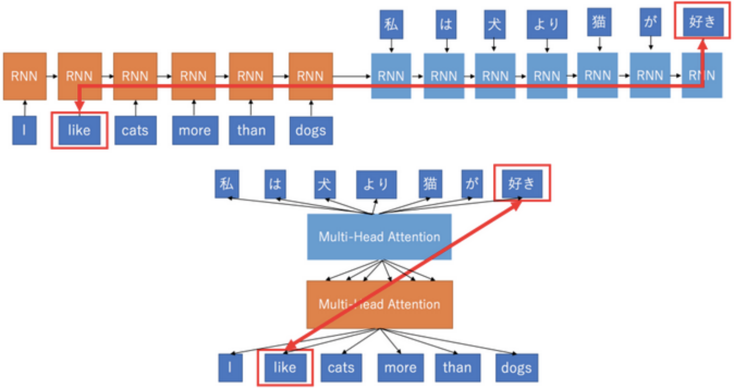


The Transformer: Going beyond LSTMs

- Google AI presented Transformer architecture that showed great benefit over conventional sequence models (LSTM, RNN, GRU)
  (Vaswani et al, 2017)
- Advantages:
  - more effective modeling of long term dependencies among tokens in a temporal sequence
  - more efficient training of the model in general by eliminating the sequential dependency on previous tokens

The Transformer: Going beyond LSTMs

- In a nutshell… a transformer is an **encoder-decoder** architecture model which uses **attention mechanisms** to forward a more complete picture of the whole sequence to the decoder at once rather than sequentially

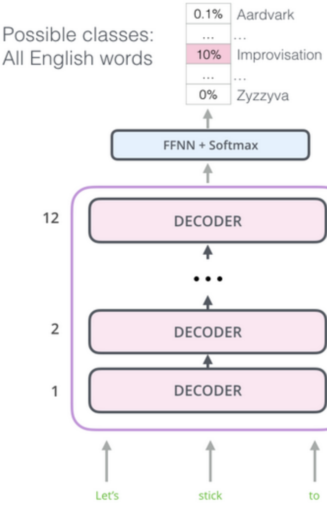


(Source: http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/)

OpenAI Transformer

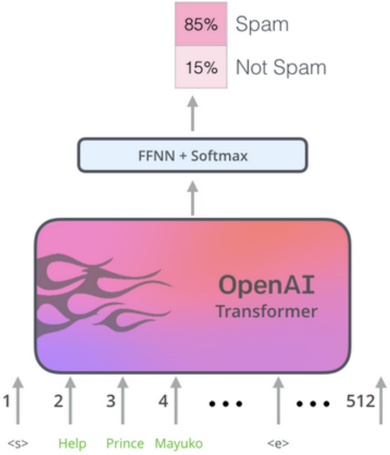
- Turns out, you don't need an entire transformer – all you need is the decoder!
- **Decoder**: a natural choice for language modeling (predicting the next word) as it's built to mask future tokens – a valuable feature when it's generating a translation word by word

OpenAI Transformer



The OpenAI Transformer is now ready to be trained to predict the next word on a dataset made up of 7,000 books.

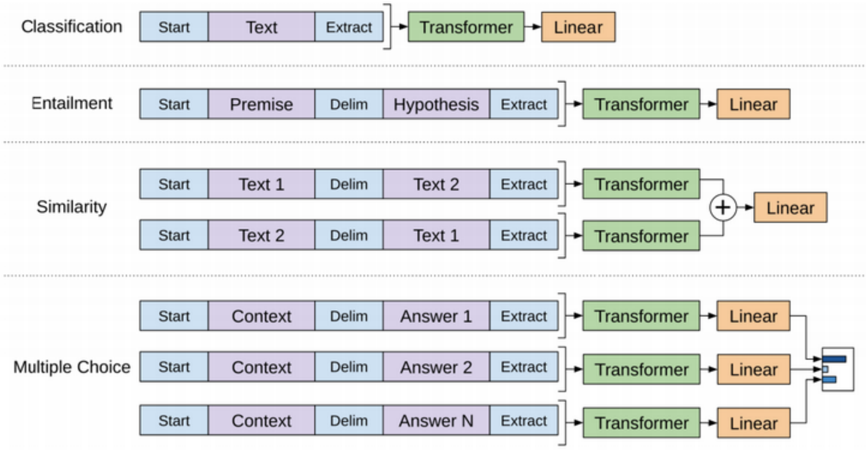
OpenAI Transformer



How to use a pre-trained OpenAI transformer to do sentence clasification

OpenAI Transformer

Various structures of the models and input transformations to carry out different downstream tasks



See Vaswani et al (2017)

Something's missing

- Something went missing in transition from LSTM architecture of ELMo…
- Could we build a transformer-based model whose language model looks both forward and backwards (conditioned on both left and right context)?

Something's missing

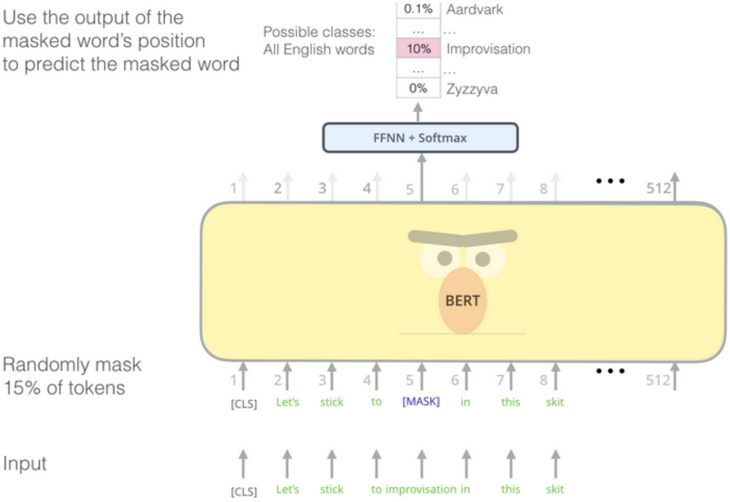
- Something went missing in transition from LSTM architecture of ELMo...
- Could we build a transformer-based model whose language model looks both forward and backwards (conditioned on both left and right context)?
- "Hold my beer", said BERT, aka Bidirectional Encoder Representations from Transformer

Masked Language Model

- "We'll use transformer encoders", said BERT.
- "This is madness", replied Ernie, "Everybody knows bidirectional conditioning would allow each word to indirectly see itself in a multi-layered context."
- "We'll use masks", said BERT confidently.

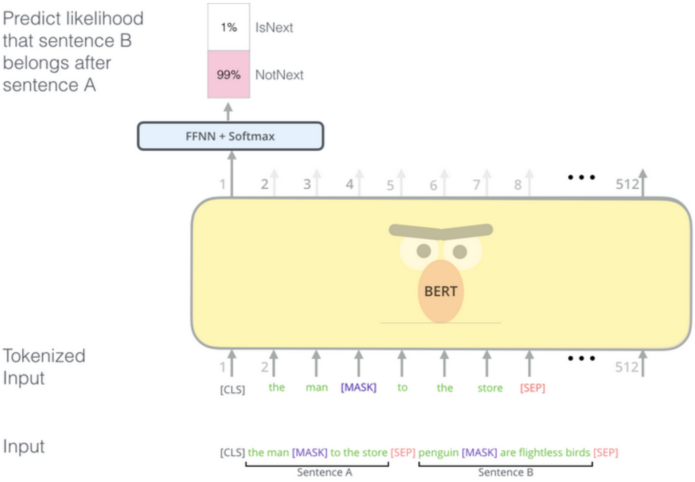
BERT

Delvin et al, 2018 (arXiv)



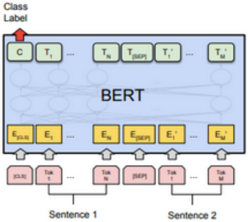
BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Two-sentence Tasks with BERT

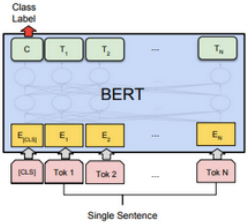


The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

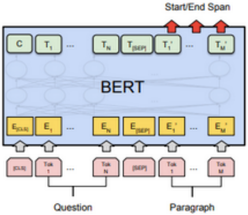
Task Specific Models



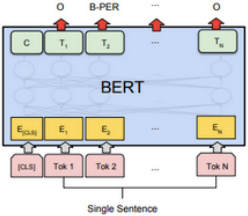
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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

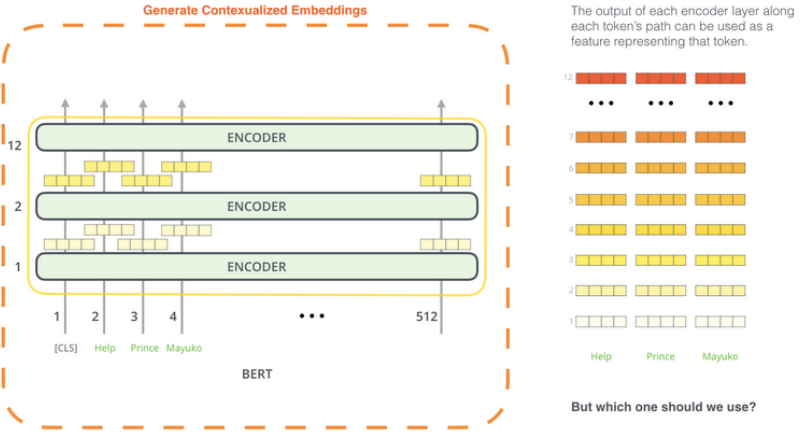


(c) Question Answering Tasks:
SQuAD v1.1



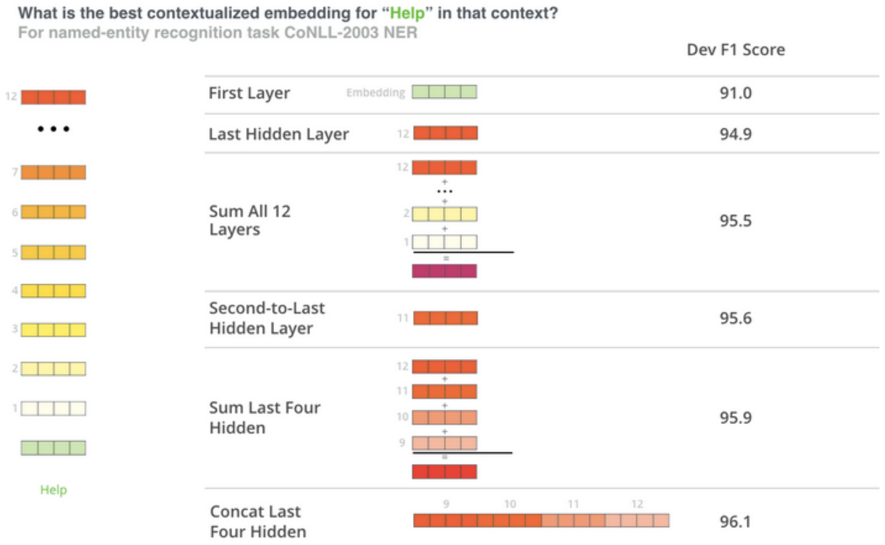
(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

BERT for feature extraction



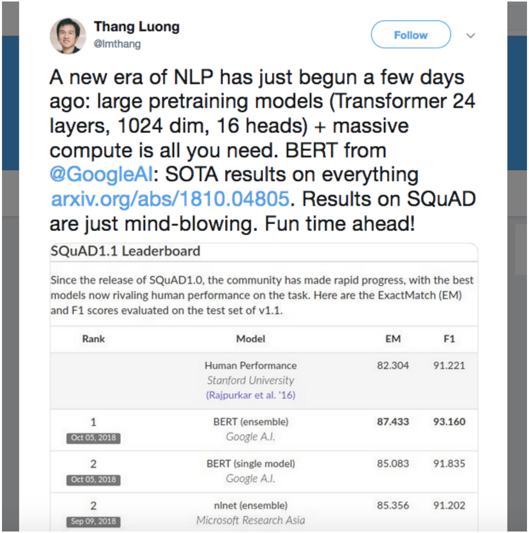
Use the pre-trained BERT to create contextualized word embeddings. Then you can feed these embeddings to your existing model, e.g. named entity recognition.

Which layer to choose?



BERT Unveiled

Delvin et al, 2018 (arXiv)



Thang Luong
@lmthang

Follow

A new era of NLP has just begun a few days ago: large pretraining models (Transformer 24 layers, 1024 dim, 16 heads) + massive compute is all you need. BERT from @GoogleAI: SOTA results on everything arxiv.org/abs/1810.04805. Results on SQuAD are just mind-blowing. Fun time ahead!

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202