IN5550: Neural Methods in Natural Language Processing
Sub-lecture 11.2
Contextualized embeddings: ELMo

Andrey Kutuzov

University of Oslo

31 March 2022



Contents



1. Solution: contextualized embeddings
   - We need to talk about ELMo

Solution: contextualized embeddings

- Idea: at inference time, assign each word a vector which is a function of the whole input phrase! [Melamud et al., 2016, McCann et al., 2017]

Solution: contextualized embeddings

- Idea: at inference time, assign each word a vector which is a function of the whole input phrase! [Melamud et al., 2016, McCann et al., 2017]
- Now our word representations are context-dependent: one and the same word has different vectors in different contexts.

Solution: contextualized embeddings

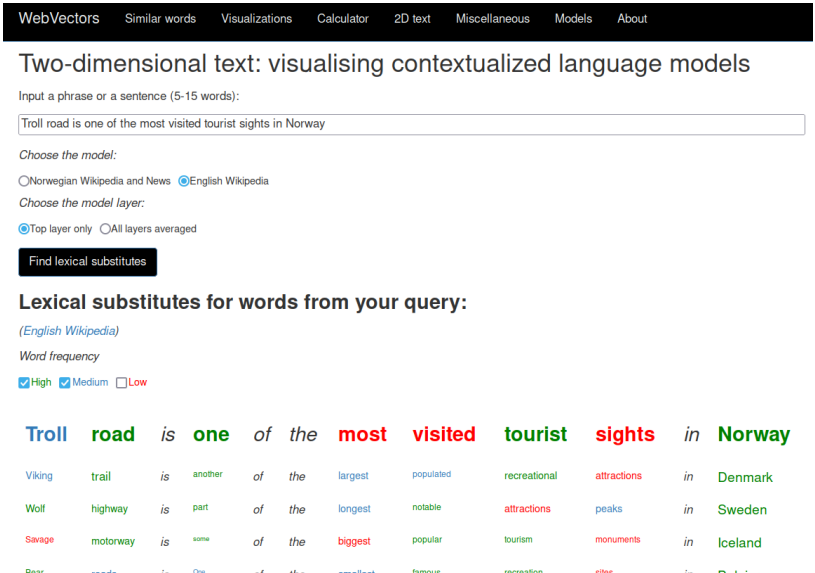
- Idea: at inference time, assign each word a vector which is a function of the whole input phrase! [Melamud et al., 2016, McCann et al., 2017]
- Now our word representations are context-dependent: one and the same word has different vectors in different contexts.
- Token embeddings VS type embeddings.

Solution: contextualized embeddings

- Idea: at inference time, assign each word a vector which is a function of the whole input phrase! [Melamud et al., 2016, McCann et al., 2017]
- Now our word representations are context-dependent: one and the same word has different vectors in different contexts.
- Token embeddings VS type embeddings.
- As an input, our model takes not an isolated word, but a phrase, sentence or text.

Solution: contextualized embeddings

http://vectors.nlpl.eu/explore/embeddings/en/contextual/



Solution: contextualized embeddings

- There is no one-to-one correspondence between a word and its embedding any more.
- Word vectors are not fixed: they are learned functions of the internal states of a deep neural network.
- The architecture itself is no more a simple lookup table:
- it is a full-fledged language model.

Solution: contextualized embeddings

- There is no one-to-one correspondence between a word and its embedding any more.
- Word vectors are not fixed: they are learned functions of the internal states of a deep neural network.
- The architecture itself is no more a simple lookup table:
- it is a full-fledged language model.
- A somewhat awkward term 'pre-trained model' is widely used now.

Google Brain unveils trillion-parameter AI language model, the largest yet

by Louis Stone 1/15/2021



Currently more of a research project than a commercial product

Google Brain has developed an artificial intelligence language model with some 1.6 trillion parameters.

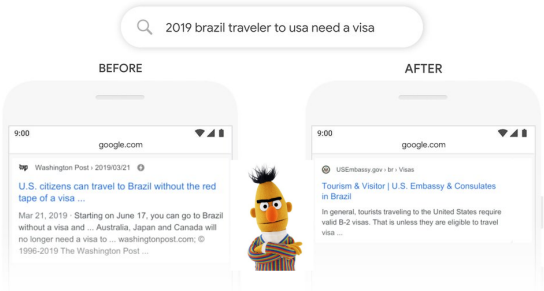
That puts it at nine times the size of the OpenAI's 175 billion parameter GPT-3, previously considered to be the world's largest language model.

While it gives some indication of the project's scale, the models are too different in architecture for a meaningful 'apples-to-apples' comparison.

I like big parameters and I cannot lie

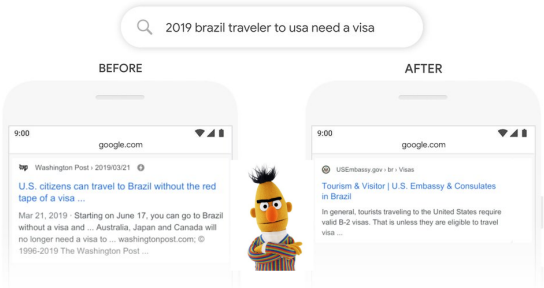
Solution: contextualized embeddings

BERT was created and popularized by Google in 2018-2019:



Solution: contextualized embeddings

BERT was created and popularized by Google in 2018-2019:



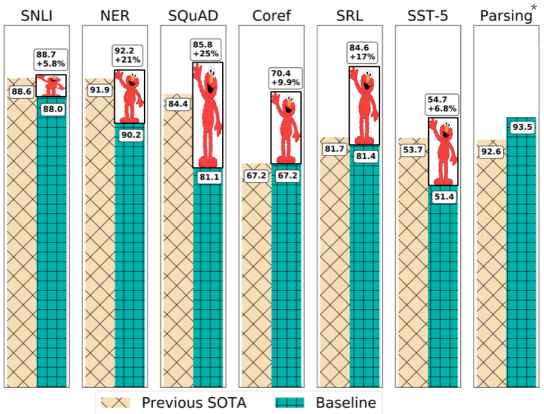
...but BERT is not alone.

Solution: contextualized embeddings

- 2018: ELMo (Embeddings from Language MOdels) conquered almost all NLP tasks
  [Peters et al., 2018]

Solution: contextualized embeddings

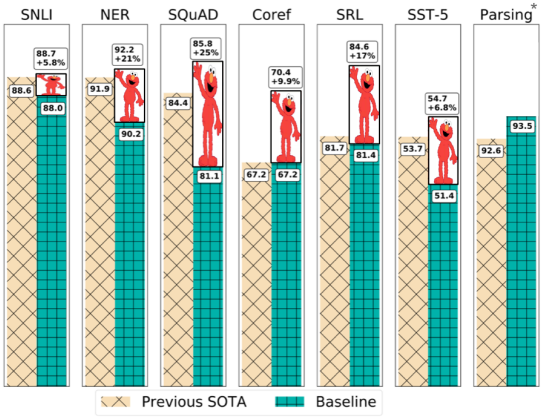
- 2018: ELMo (Embeddings from Language MOdels) conquered almost all NLP tasks
  [Peters et al., 2018]



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

Solution: contextualized embeddings

- 2018: ELMo (Embeddings from Language MOdels) conquered almost all NLP tasks
  [Peters et al., 2018]



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

- 2019: BERT (Bidirectional Encoder Representations from Transformer) did the same
  [Devlin et al., 2019]

Solution: contextualized embeddings

Both architectures use deep neural networks

- ELMo employs bidirectional LSTMs.
- BERT employs transformers with self-attention.
- '*ImageNet for NLP*' (Sebastian Ruder)
- Many other Sesame street characters made it into NLP since then!

Deep Neural Networks Are Our Friends



Wang Ling Google DeepMind

We need to talk about ELMo

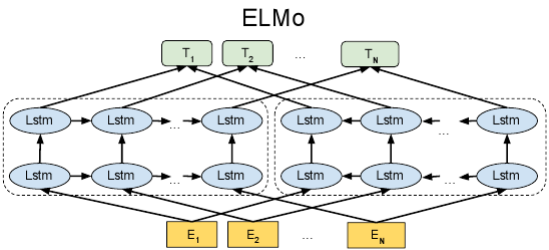
Embeddings from Language MOdels

- Contextualized ELMo models are trained on raw text, optimizing for the language modeling task (next word prediction).

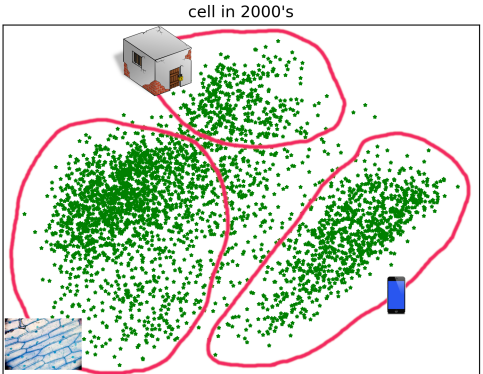
We need to talk about ELMo

Embeddings from Language MOdels

- Contextualized ELMo models are trained on raw text, optimizing for the language modeling task (next word prediction).
- Two BiLSTM layers over one layer of character-based CNN.
- Takes sequences of characters as an input.
  - ...actually, they are UTF-8 code units (bytes), not characters *per se*.
- LSTM output states for each word are the ELMo embeddings.



We need to talk about ELMo



- 2-dimensional PCA projections of ELMo embeddings for each occurrence of *'cell'* in the Corpus of Historical American English (2000-2010).
- Left clusters: biological and prison senses.
- Large cluster to the right: 'mobile phone' sense.

NB: each dot represents an embedding of *'cell'* in a particular sentence.

We need to talk about ELMo

Norwegian ELMo in word sense disambiguation (PCA projection of '*rett*' embeddings):



References I

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017).
Learned in translation: Contextualized word vectors.
In Advances in Neural Information Processing Systems, pages 6294–6305.

Melamud, O., Goldberger, J., and Dagan, I. (2016).
context2vec: Learning generic context embedding with bidirectional LSTM.
In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

References II



Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237. Association for Computational Linguistics.