An ELMo-inspired approach to SemDeep-5's Word-in-Context task

Alan Ansell [1] Felipe Bravo-Marquez [2] Bernhard Pfahringer [1]

[1]University of Waikato

[2]University of Chile

August 12, 2019

Word-in-Context Dataset (Pilehvar and Camacho-Collados, 2019)

- Task: determine whether a word (the "focus word") is being used in the same sense in two different contexts.
- Around 7,500 examples in total.

ELMo Benchmark

- Obtain contextualized embeddings for the focus words in the two contexts by taking hidden states from the ELMo BiLSTM language model (Peters et al., 2018).
- Two methods of making predictions:
  - Calculate cosine similarity between the two embeddings and predict based on a threshold.
  - Feed embeddings into a MLP.
- Best configurations are:
  - ELMo_1 (hidden states of first LSTM layer) + cosine similarity: 57.7%.
  - ELMo_3 (weighted combination of all three LSTM layers) + MLP: 57.2%.

Our System

- Simple system which makes a few improvements over the ELMo baseline.
- Exploits bidirectionality more deeply.
- Better choice of contextual embedding.
- Better similarity measure.

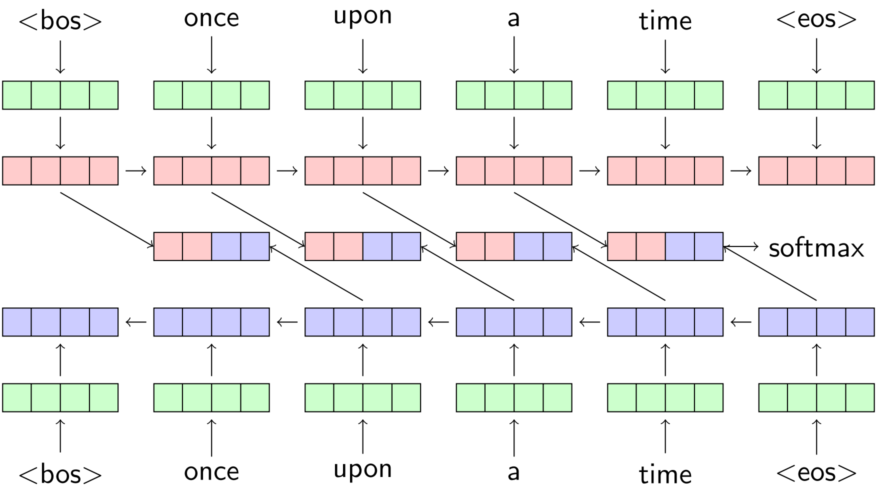
Bidirectionality in ELMo

- ELMo essentially trains forward and backward LSTMs independently.
- In a task like WiC, we want to be able to exploit both sides of the context more effectively.

Bidirectionality in Our Model

- Rather than predicting the next word given a left side context or the previous word given the right side context, predict the missing word given a left and right context.
- Predict the missing word using the concatenation of the forward representation of the left context and the backward representation of the right context.

Architecture



Choice of Contextual Representation

- The hidden states of an LSTM language model for a word of interest contain information which will be useful for predicting future/previous words.
- Some of this information may not be relevant to the sense the word is being used.
- Instead we will use the output from the final layer which is used to *predict* the word of interest, since this representation contains information solely related to this word.

Similarity Measure

- ELMo benchmarks use cosine distance between the contextual representations of the focus words or a MLP.
- Cosine distance is very simple and doesn't exploit the availability of training data.
- MLP seems prone to overfitting on a dataset of this size.

Similarity Measure

We use a weighted dot product

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{w}^\top (\boldsymbol{x}_1 \circ \boldsymbol{x}_2),$$

where $\circ$ denotes element-wise product and $\boldsymbol{w}$ is a weight vector trained on the training set. L2 regularisation is applied with a coefficient tuned on the dev set.

Results

States	Similarity measure	Dev.	Test
Predictor	Weighted dot product	67.4	61.2
Predictor	Unweighted dot product	60.2	59.1
Predictor	Cosine similarity	60.5	59.1
Hidden	Cosine similarity	55.2	54.9
Hidden	Weighted dot product	54.1	53.1

Conclusions

- "Predictor" states provide a better contextualized representation for the purposes of the WiC task than hidden states when forward and backward LSTMs are trained jointly.
- Weighted dot product is a better similarity measure than cosine distance.
- 3.5% improvement over ELMo baseline without looking at the focus word.

References I

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

References II

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal. Association for Computational Linguistics.

References III

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

References IV

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL*, Minneapolis, United States.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.

Parameters

- Embedding size: 256
- LSTM hidden state size: 2048, downprojected to 256 dimensional output.
- Training corpus: Wikipedia 2018
- Vocabulary size: approx. 100k