Probing sentence embeddings for structure-dependent tense

Geoff Bacon
Department of Linguistics
UC Berkeley
bacon@berkeley.edu

Terry Regier
Department of Linguistics and
Cognitive Science Program
UC Berkeley
terry.regier@berkeley.edu

1 Introduction

Learning universal sentence representations which accurately model sentential semantic content is a current goal of natural language processing research (Subramanian et al., 2018; Conneau et al., 2017; Wieting et al., 2016; Kiros et al., 2015). A prominent and successful approach is to train recurrent neural networks (RNNs) to encode sentences into fixed length vectors (Conneau et al., 2018; Nie et al., 2017). Many core linguistic phenomena that one would like to model in universal sentence representations depend on syntactic structure (Chomsky, 1965; Everaert et al., 2015). Despite the fact that RNNs do not have explicit syntactic structural representations, there is some evidence that RNNs can approximate such structure-dependent phenomena under certain conditions (Gulordava et al., 2018; McCoy et al., 2018; Linzen et al., 2016; Bowman et al., 2015), in addition to their widespread success in practical tasks.

In this work, we assess RNNs' ability to learn the structure-dependent phenomenon of main clause tense. To test whether sentence representations derived from RNNs capture main clause tense, we attempt to predict the tense from the representation. This approach is called *probing*, and was introduced by Ettinger et al. (2016) and subsequently used by Adi et al. (2017) and others.

Conneau et al. (2018) probed English sentence representations from various RNN architectures for main clause tense and concluded that these architectures, along with a bag-of-vectors (BoV) baseline, capture tense very well (84-91% accuracy). However, this result was based on a test set in which the tense category (i.e. past or present) to be predicted was the most common tense category in the sentence for 95.2% of sentences. The high performance of the BoV model on this test set is not entirely surprising, given that Köhn (2015, 2016) showed a wide variety of word embedding models capture tense at the word level very well. The high performance of the RNN models is not strong evidence that they are sensitive to the structure-dependence of main clause tense. As suggested by Linzen et al. (2016), these models may be learning a flawed heuristic that only works in grammatically simple examples.

Our goal is to determine whether RNNs learn to perform structure-dependent computation or whether they merely learn practical heuristics. To do this, we extend the experimental setup of Adi et al. (2017), which has a two step nature. First, we train autoencoders for English, Spanish, French and Italian where both the encoder and decoder are either Simple Recurrent Networks (SRNs, Elman, 1990) or Long Short-Term Memory networks (LSTMs, Hochreiter and Schmidhuber, 1997). Second, we use the trained encoder to obtain sentence representations and probe those representations for main clause tense. We investigate whether probing performance is affected by eight potential distractors, one of which is other words in the sentence with tense categories that differ from the tense of the main clause (e.g. *we know who won*). To the extent that the representations are *in*sensitive to structure-dependence, we expect to see probing performance negatively affected by distractors. We compare the RNNs to three BoV baseline models.

In this extended abstract, we report on our work in progress. We have completed data collection and preprocessing, designed our experiments and obtained complete results from our BoV baselines.

2 Data

A guiding principle in our choice of data sources was availability across multiple languages, be-

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 334–336
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

cause we are interested in cross-linguistic generality. To train sentence embedding models (i.e. RNNs and BoV), we extracted one million sentences between 5 and 70 tokens in length from each language's Wikipedia, in line with Adi et al. (2017). This yields between 25 and 29 million tokens per language.

Our labelled probing data are sentences from Universal Dependencies treebanks (UD, Nivre et al., 2016). Because of the way the UD schema annotates tense in multiword verb phrases, extracting main clause tense is not straightforward. Therefore, for each language we developed between five and seven heuristic rules in terms of UD annotations to extract tense. A random sample of 100 sentences for each language shows that our heuristics produce the correct tense in at least 98% of sentences.

To ensure the sentence embedding models see all word types needed for the probing task during training, the embedding vocabulary is set to the union of the 50k most frequent word types in the Wikipedia data and all word types in the probing data. Resulting vocabulary sizes range from 53k to 68k, with OOV rates in the Wikipedia data between 2 and 4% per language. We remove sentences from the probing task that require word types not seen in the Wikipedia data. This results in between 12k and 31k sentences per language in the probing task. We split these into 70% train and 30% test sets, with the constraint that no word form that is responsible for main clause tense in the training set also appears in the test set, following Conneau et al. (2018).

3 Experimental setup

In line with Adi et al. (2017), we trained word embeddings on the Wikipedia data using skipgram (Mikolov et al., 2013), with hierarchical softmax and a window size of five, for five epochs. We trained 50 sets of embeddings per language, with dimension sizes from 20 to 1000 in steps of 20. Our three BoV baselines consist of combining these word embeddings by summing, averaging and using Smooth Inverse Frequency (Arora et al., 2017). Here, we report results from summing, which in contrast to related experiments (Conneau et al., 2018; Arora et al., 2017), consistently and significantly outperforms the other two baselines. For the probing task, we use L1-regularized logistic regression with ten-fold cross validation.

4 Baseline results

Here, we present results for one of our eight distractors. Figure 1 shows the effect on probing performance of the number of words in the sentence with tense categories that differ from the main clause tense. In all four languages, as the number of such conflicting tensed forms in the sentence increases, error rates on the probing task also tend to increase. This is expected given that BoV is not sensitive to syntactic structure, and serves as a baseline for our upcoming work using RNNs.

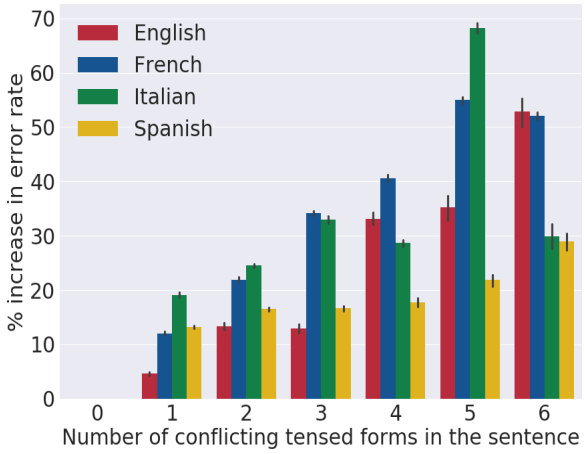

Figure 1: The effect of conflicting tensed words on probing performance for our summed BoV baseline. We measure the absolute percentage increase in error rate over the error rate when no conflicting tensed words are in the sentence. Each bar represents this quantity averaged across all 50 sets of embeddings per language. Error bars are 95% confidence intervals.

Adi et al. (2017) found a negative correlation between performance on one of their probing tasks (content prediction) and sentence length. Surprisingly, we find no correlation between performance of any of our baseline models and sentence length.

5 Remaining work

Our goal is to understand to what extent RNNs show a similar insensitivity to structure-dependence. Our next step is to train SRN- and LSTM-based autoencoders on the Wikipedia data and assess their representations in our probing task. Due to our careful choice of data sources, future work can extend our analysis to any language with i) a sizable Wikipedia, ii) a UD corpus, and iii) tense.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Samuel R Bowman, Christopher D Manning, and Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015 International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches-Volume 1583*, pages 37–42. CEUR-WS. org.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA. MIT Press.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Martin BH Everaert, Marinus AC Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. 2015. Structures, not strings: linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205. Association for Computational Linguistics.

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

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073. Association for Computational Linguistics.

Arne Köhn. 2016. Evaluating embeddings using syntax-based classification tasks as a proxy for parser performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 67–71. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

R Thomas McCoy, Tal Linzen, and Robert Frank. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Allen Nie, Erin D. Bennett, and Noah D. Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *CoRR*, abs/1710.04334.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations*.