On the Hierarchical Information in a Single Contextualised Word Representation (Student Abstract)

Dean L. Slack,[1] **Mariann Hardey,**[2] **Noura Al Moubayed**[1]
[1]Department of Computer Science, Durham University, Durham, UK
[2]Business School, Durham University, Durham, UK
{dean.l.slack, mariann.hardey, noura.al-moubayed}@durham.ac.uk
+44 (0)191 3341736

Abstract

Contextual word embeddings produced by neural language models, such as BERT or ELMo, have seen widespread application and performance gains across many Natural Language Processing tasks, suggesting rich linguistic features encoded in their representations. This work aims to investigate to what extent any linguistic hierarchical information is encoded into a single contextual embedding. Using labelled constituency trees, we train simple linear classifiers on top of single contextualised word representations for ancestor sentiment analysis tasks at multiple constituency levels of a sentence. To assess the presence of hierarchical information throughout the networks, the linear classifiers are trained using representations produced by each intermediate layer of BERT and ELMo variants. We show that with no fine-tuning, a single contextualised representation encodes enough syntactic and semantic sentence-level information to significantly outperform a non-contextual baseline for classifying 5-class sentiment of its ancestor constituents at multiple levels of the constituency tree. Additionally, we show that both LSTM and transformer architectures trained on similarly sized datasets achieve similar levels of performance on these tasks. Future work looks to expand the analysis to a wider range of NLP tasks and contextualisers.

Introduction

Neural language model pretraining has become ubiquitous in Natural Language Processing (NLP), due largely to a recent trend focused on improving the quality of linguistic features contained within word embeddings. The result being *contextual word embeddings*: continuous representations conditioned on the entire input context. Notable examples of this approach are ELMo (Long Short Term Memory-based) (Peters et al. 2018b), and BERT (Transformer-based) (Devlin et al. 2019). Both have led to significant state-of-the-art improvements on downstream tasks, highlighting the potential for improved transfer learning in NLP (Howard and Ruder 2018). However, due to a poor understanding of the linguistic information contained in contextual representations - a thriving new area of research looks to employ novel analysis techniques to improve the interpretability of these models, and how they produce effective, transferable features. Examples of this include probing the behaviour of single neurons to infer which features are encoded (Dalvi et al. 2019), and analysing the internal states and geometry of the pretrained language models in terms of syntactic and semantic sub-spaces (Coenen et al. 2019).

Despite the aforementioned work, little progress has been made to determine the nature of architecture-specific internal representations and how they are mapped to a single contextualised word embedding. The works of (Liu et al. 2019) and (Peters et al. 2018a) explore how the performance of probing classifiers trained on representations produced by language model architectures vary by layer depth across a range of traditional NLP tasks. They show that syntactic and semantic features are encoded by shallower and deeper layers of the model, respectively. Similarly, edge probing tasks are introduced in (Tenney et al. 2019) and (Tenney, Das, and Pavlick 2019) to explore sentence level knowledge of word representations by training classifiers limited to specific spans of the input sequence.

Building on the latter approaches, this work aims to explore the quality of hierarchical information encoded in contextual word representations, through the following contributions: I) Formulation of a range of NLP probing tasks with classification labels at multiple hierarchies of a constituency tree, constraining the input to single word tokens contextualised on the full sentence context. This seeks to identify any hierarchical semantic structure encoded into word representations produced by various prevalent architectures. In this, our work is most similar to that of (Liu et al. 2019). II) Improved understanding of layer-wise performance over different constituent levels of a sentence can be used to make informed decisions regarding the appropriate selection of embedding layer to use, i.e., choice of layer per input token in sequential models where all input tokens are used. III) We show that for sentiment analysis classification, hierarchical information useful for predicting sentiment at multiple constituent levels of a sentence is contained within a single contextual word representation.

Table 1: Results showing accuracy (%) of each classifier on the 5-class ancestor sentiment analysis task. Best performing BERT layers per task are in bold.

Embedder	Root	Leaf	Parent	GParent	GGParent
BERT (base, cased), layer 12	**42.16**	88.33	62.69	53.29	48.15
BERT (base, cased), layer 11	40.97	88.95	62.93	53.34	48.16
BERT (base, cased), layer 10	41.25	89.32	63.12	53.65	**48.42**
BERT (base, cased), layer 9	40.82	89.76	**63.35**	**53.94**	48.36
BERT (base, cased), layer 8	40.37	90.20	63.26	53.48	47.55
BERT (base, cased), layer 7	40.20	90.92	63.10	53.10	47.16
BERT (base, cased), layer 6	39.41	91.31	62.63	52.35	46.45
BERT (base, cased), layer 5	39.09	91.65	62.03	51.30	45.55
BERT (base, cased), layer 4	37.91	91.92	61.80	50.75	44.91
BERT (base, cased), layer 3	38.20	92.15	61.47	50.28	44.29
BERT (base, cased), layer 2	36.98	92.38	61.33	49.97	43.85
BERT (base, cased), layer 1	37.07	**92.84**	60.64	48.90	42.50
BERT (base, cased), layer 0	32.26	92.82	60.10	47.49	39.94
GloVe (840B.300d)	28.81	90.27	60.28	47.53	39.96
State-of-the-art (all input tokens used)	54.70	-	-	-	-

Methodology

We leverage an existing dataset with sentiment classifications for each constituent phrase in a sentence: Stanford Sentiment Treebank (SST) (Socher et al. 2013). SST contains both five class and binary sentiment labels for each constituent phrase of a sentence. Using this hierarchically labelled sentiment data, we formulate a token-wise ancestor sentiment analysis task. For a given token in a sentence, its corresponding contextualised word representation is tasked with predicting the sentiment classification of its parent, grandparent, and great-grandparent constituent phrase. For cases where the token does not have an ancestor phrase, the linear model is tasked to predict a 'None' classification label. As we are probing for semantic features useful for classifying sentiment at varying constituency levels, we perform all tasks using single word representations contextualised on the full root sentence. All classifiers are trained on words contained in full sentences, with no sub-phrases (8544 sentences), for 64 epochs using the Adam optimiser.

Initial Results and Discussion

Table 1 reports the results for each linear classifier layer of the BERT (base, cased) contextualiser, a non-contextual baseline comparison GloVe, and a state-of-the-art comparison (all input tokens used), for ancestor sentiment classification of the root (full sentence), leaf (token), parent, grandparent, and great-grandparent constituents. Results are reported for the test set (2210 sentences). Initial results for ancestor sentiment classification using the BERT contextualiser show that for sentence-level (root) classification, contextual embeddings significantly outperform the non-contextual baseline. This confirms that some notion of global sentiment information is contained within a contextual representation. Furthermore, a single BERT embedding layer is capable of sentiment classification at multiple constituency levels, with lower layers better suited to lower constituent levels.

Conclusion and Future Work

This work shows that a single contextual word representation is capable of encoding information useful for classifying sentiment at multiple constituency levels of the sentence they are contextualised on. Initial probing results for ancestor sentiment analysis using pretrained BERT representations show that deeper layers of the network contain information better suited to the global sentiment of the embedded context, with lower layers containing more local sentiment information. Additionally, we show there is enough global information encoded by the deeper layers of BERT to train a simple linear classifier and achieve 42.16% accuracy on 5-class sentence-level sentiment analysis.

Future work will aim to probe a wider range of NLP classification tasks typically solved by sequential classifiers utilising the full sequence. Single tokens will be then tasked to predict labels at varying constituent levels of the sequence. Additionally, a thorough analysis of the complete set of results including different contextualiser architectures, non-linear probing models, and qualitative assessments of each task will be undertaken.

Acknowledgement

This work is supported by the European Regional Development Fund (ERDF).

References

Coenen, A.; Reif, E.; Yuan, A.; Kim, B.; Pearce, A.; Viégas, F.; and Wattenberg, M. 2019. Visualizing and measuring the geometry of bert. *arXiv preprint arXiv:1906.02715*.

Dalvi, F.; Durrani, N.; Sajjad, H.; Belinkov, Y.; Bau, A.; and Glass, J. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proc. of AAAI*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 4171–4186.

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. In *Proc. of ACL*, 328–339.

Liu, N. F.; Gardner, M.; Belinkov, Y.; Peters, M. E.; and Smith, N. A. 2019. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL*, 1073–1094.

Peters, M.; Neumann, M.; Zettlemoyer, L.; and Yih, W.-t. 2018a. Dissecting contextual word embeddings: Architecture and representation. In *Proc. of EMNLP*, 1499–1509.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018b. Deep contextualized word representations. In *Proc. of NAACL*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*, 1631–1642.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S. R.; Das, D.; and Pavlick, E. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proc. of ICLR*.

Tenney, I.; Das, D.; and Pavlick, E. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.