Structure-aware Sentence Encoder in BERT-Based Siamese Network

Qiwei Peng David Weir Julie Weeds
University of Sussex
Brighton, UK
{qiwei.peng, d.j.weir, j.e.weeds}@sussex.ac.uk

Abstract

Recently, impressive performance on various natural language understanding tasks has been achieved by explicitly incorporating syntax and semantic information into pre-trained models, such as BERT and RoBERTa. However, this approach depends on problem-specific fine-tuning, and as widely noted, BERT-like models exhibit weak performance, and are inefficient, when applied to unsupervised similarity comparison tasks. Sentence-BERT (SBERT) has been proposed as a general-purpose sentence embedding method, suited to both similarity comparison and downstream tasks. In this work, we show that by incorporating structural information into SBERT, the resulting model outperforms SBERT and previous general sentence encoders on unsupervised semantic textual similarity (STS) datasets and transfer classification tasks.

1 Introduction

Pre-trained models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have demonstrated promising results across a variety of downstream NLP tasks. Though BERT-like models have been shown to capture hidden syntax structures (Clark et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019), recent works have achieved performance improvements on various natural language understanding (NLU) tasks through the use of a graph network that captures syntax and semantics information. Xu and Yang (2019) demonstrate the value of syntax information for pronoun resolution tasks, using Relational Graph Convolutional Networks (RGCNs) (Schlichtkrull et al., 2018) to incorporate syntactic dependency graphs. Wu et al. (2021) argue that semantics has not been brought to the surface of pre-trained models and propose to introduce semantic label information into RoBERTa via RGCNs. Similar ideas have been applied to information extraction (Santosh et al., 2020), sentence-pair classification (Liu et al., 2020) and sentiment analysis (Wang et al., 2020; Yin et al., 2020) tasks. Though problem-specific fine-tuning is required, these improvements suggest that structural supervision is useful, and that RGCNs serve as an effective structure encoder.

BERT can also be used as a general sentence encoder, either by using the CLS token (the first token of BERT output) or applying pooling over its outputs. However, this fails to produce sentence embeddings that can be used effectively for similarity comparison. Furthermore, this method of using BERT for similarity comparison is extremely inefficient, requiring sentence pairs to be concatenated and passed to BERT for every possible comparison. In response, Sentence-BERT (SBERT) has been proposed to alleviate this by fine-tuning BERT on natural language inference (NLI) datasets using a siamese structure (Reimers and Gurevych, 2019). General-purpose sentence embeddings are generated which outperform previous sentence encoders on both similarity comparison and transfer tasks.

In this paper, we show that it is possible to improve the SBERT sentence encoder through the use of explicit syntactic or semantic structure. Inspired by SBERT's success in producing general sentence representations and previous efforts on introducing structural information into pre-trained models, we propose a model that combines the two by training a BERT-RGCN model in a siamese structure. Under specific structural supervision, the proposed model is able to produce structure-aware, general-purpose sentence embeddings. Our empirical results show that it outperforms SBERT and previous sentence encoders on unsupervised similarity comparison and transfer classification tasks. Furthermore, we find that the produced sentence representation generalises better especially

Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 57–63
Bangkok, Thailand (Online), August 6, 2021.

on fine-grained classification tasks.

2 Related Work

Sentence encoders have been studied extensively in years. Skip-Thought (Kiros et al., 2015) has been trained to predict its surrounding sentences by using current sentence in a self-supervised fashion. Hill et al. (2016) proposed a sequential denoising autoencoder (SDAE) to reconstruct given sentence representations. InferSent (Conneau et al., 2017), on the other hand, used labelled NLI datasets to train a general-purpose sentence encoder in a BiLSTM-based siamese structure. Cer et al. (2018) proposed the Universal Sentence Encoder (USE) model based on transformers (Vaswani et al., 2017), and trained it with both unsupervised tasks and supervised NLI tasks. Inspired by InferSent, Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) produces general-purpose sentence embeddings by fine-tuning BERT on NLI datasets in a siamese structure, showing improved performance on a variety of tasks.

Hidden syntax structures in pre-trained models have been well explored. Various probing methods have been used to investigate hidden structures (Clark et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019). The impact of external structures on pre-trained models has also been questioned. Glavaš and Vulić (2021) examined the benefits of incorporating universal dependencies into pre-trained models. Dai et al. (2021) showed that the tree induced from pre-trained models could produce competitive results compared with external trees. However, recent improvements have still been observed on various NLU tasks by incorporating structural information into pre-trained models. Yin et al. (2020) proposed SentiBERT to incorporate constituency tree into BERT for sentiment analysis. Xu and Yang (2019) modelled each sentence as a directed dependency graph by using RGCNs, and achieved large improvements on pronoun resolution. Zhang et al. (2020) proposed a semantics-aware BERT model by further encoding semantic information with BERT using a GRU (Chung et al., 2014). RGCNs have also been used by Wu et al. (2021) to introduce semantic information into RoBERTa, and achieved consistent improvements when fine-tuned on problem-specific datasets. Similar efforts can be seen where researchers try to provide syntax information via self-attention mechanism (Bai et al., 2021; Li et al., 2020).

3 Model

Inspired by Reimers and Gurevych (2019), we train our model in a siamese network to update weights so as to produce similarity-comparable sentence representations. The model we propose consists of two components, as shown in Figure 1.

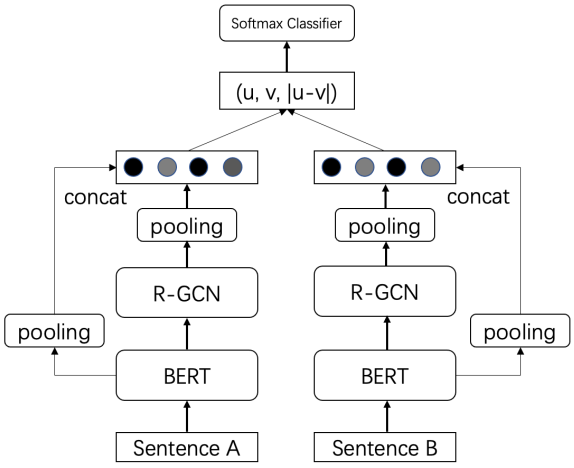

Figure 1: The proposed model in siamese structure

BERT: Each sentence is first fed into the pre-trained BERT-base model to produce both a sentence representation, by applying mean-pooling, and an original contextualised sequence-length token representation, which is used to initialise a RGCN.

Structure Information: We use Spacy dependency parser (Honnibal et al., 2020) with its middle model to obtain dependency parse trees for all input sentences. We also experimented with the use of semantic graphs[1], since Wu et al. (2021) has shown that semantic information benefits pre-trained models. However, we found semantic graphs to be less effective than syntactic dependency trees when evaluated on our development set, and as a result, in the experiments below, we restrict our attention to the use of syntactic dependency graphs.

RGCN: RGCNs, proposed by (Schlichtkrull et al., 2018), can be viewed as a weighted message passing process. At each RGCN layer, each node's representation will be updated by collecting information from its neighbours and applying edge-specific weighting:

$$h_i^{l+1} = ReLU(W_0^l h_i^l + \sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^l h_j^l) \quad (1)$$

[1]For semantic graphs, we use the semantic parser produced by Che et al. (2019).

where N_i^r and W_r^l are the neighbours of node i and the weight of relation $r \in R$, respectively. $c_{i,r}$ is the normalisation constant and normally set to be $|N_i^r|$ which is the number of neighbours under relation r. W_0^l is the self-loop weight. In our case, each sentence is first parsed into a dependency tree, then modelled as a labelled directed graph by an RGCN, where nodes are words and edges are dependency relations. Following Schlichtkrull et al. (2018), we allow information to flow in both directions (from head to dependent and from dependent to head). Following Wu et al. (2021), we pass BERT output through an embedding projection which is made of an affine transformation and ReLU non-linearity, then use the transformed representations to initialise RGCN's node representations. Since BERT and Spacy use different tokenisation strategies, we align them by taking the first subtoken as its word representation from BERT for each word in the RGCN. A structure-aware sentence representation is derived from RGCN's output by applying a mean-pooling over its node representations. During training, rather than using $c_{i,r} = |N_i^r|$, we found it best to apply the normalisation factor across relation types, $c_{i,r} = c_i = \sum_r |N_i^r|$, the number of neighbours. We use a one-layer RGCN, as we find that a deeper network lowers the performance.

Connect BERT and RGCN: The concatenation of BERT and RGCN's sentence representations are then passed through a layer normalisation layer to form the final sentence representation. Sentence embeddings of given sentence-pair are then interacted before passing to the final classifier for training. As for the interaction, we use the concatenation of sentence embedding u, v and the element-wise difference $|u - v|$, which has been found to be the best concatenation mode by Reimers and Gurevych (2019). In this siamese structure, all parameters are shared and will be updated correspondingly. We use cross-entropy loss for optimisation.

4 Experiments

We compare our model with SBERT[2], InferSent[3], USE[4], average GloVe vectors, and also two strategies using pre-trained BERT to produce sentence representations (BERT-CLS and BERT-AVG). For all experiments on these models, we use released pre-trained models and scripts to produce sentence embeddings.

4.1 Training Details

In order to produce general-purpose sentence embeddings, we follow SBERT in training the model on a combination of the SNLI (Bowman et al., 2015) and the MNLI datasets (Williams et al., 2018). They contain $570,000$ and $430,000$ sentence pairs, respectively, which are annotated as contradiction, entailment, or neutral. Our model is trained for one epoch, and we use a batch-size of 16, the Adam optimizer with learning rate $2e{-}5$, and a linear learning rate warm-up over 10% of the training data. For RGCN layer, we use dropout of 0.2 and hidden dimension of 512. Following SBERT, we evaluate our model on the STS benchmark development set in Spearman rank correlation for every $1,000$ steps during training, and save the best model.

4.2 Evaluation - Unsupervised STS

First, we evaluate our model on semantic textual similarity (STS) datasets. Here we use STS12-16 tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016), SICK-Relatedness (SICK-R) (Marelli et al., 2014) test set and STS benchmark (STSb) (Cer et al., 2017) test set. These datasets are labelled from 0 to 5 on semantic relatedness of sentence pairs. We obtain these datasets via SentEval (Conneau and Kiela, 2018). In this evaluation, we test different encoders' performance without using any task-specific training data.

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	AVG
GloVe AVG	52.24	49.91	43.36	55.91	47.67	46.00	55.02	50.02
InferSent	48.42	67.37	61.41	72.87	66.12	64.33	62.95	63.35
USE	63.42	67.50	64.16	76.99	73.23	74.60	**76.67**	70.94
BERT-AVG	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
BERT-CLS	21.54	32.11	21.28	37.89	44.24	20.29	42.42	31.40
SBERT	70.97	76.53	73.19	79.09	74.30	76.98	72.91	74.85
Ours	**72.51**	**77.05**	**74.06**	**80.90**	**76.20**	**78.50**	73.58	**76.11**

Table 1: Results on STS12-16, STSb and SICK-R. Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels is calculated. $\rho \times 100$ is reported

The results are given in Table 1, and show that our model outperforms SBERT on all 7 tasks, obtaining the highest average score, and demonstrating the benefits of including explicit syntax structure during supervision. Both SBERT and

[2]https://github.com/UKPLab/sentence-transformers, we use its BERT-base-nli-mean model

[3]https://github.com/facebookresearch/InferSent

[4]https://tfhub.dev/google/universal-sentence-encoder-large/3

our model perform worse than USE on SICK-R. However, as observed by Reimers and Gurevych (2019), USE is trained on various datasets including question-answering pairs, NLI, online forums and news, which appears to be particularly suitable to the data of SICK-R. Both BERT-AVG and BERT-CLS perform poorly which reflects their weakness as general-purpose sentence encoders.

4.3 Evaluation - Transfer Tasks

While the best results for BERT-like models is achieved with problem-specific fine-tuning, an evaluation on transfer tasks provides a way to test the encoder's generalisation ability and representation quality. Following Reimers and Gurevych (2019), we use SentEval with logistic regression to test different encoders on 8 classification tasks: sentiment analysis, MR (Pang and Lee, 2005); CR (Hu and Liu, 2004); SST-5/SST-2 (Socher et al., 2013); question-type, TREC (Li and Roth, 2002); subjectivity-objectivity, SUBJ (Pang and Lee, 2004); phrase-level opinion polarity, MPQA (Wiebe et al., 2005); and paraphrase detection, MRPC (Dolan et al., 2004). These datasets are provided by SentEval.

As shown in Table 2, the proposed model outperforms previous encoders in general though the difference between SBERT and our model is relatively small. Our model performs significantly worse than USE on TREC, which may be due to the fact that USE is pre-trained on question-answering data, which appears to be beneficial to the TREC question-type classification task. Unlike previous poor performance on STS datasets, BERT-CLS and BERT-AVG produce good results on classification tasks. This shows that the relevant information is encoded in BERT-CLS and BERT-AVG, they just lack the ability to produce similarity-comparable sentence embeddings. Both SBERT and our model perform worse than BERT-AVG and BERT-CLS on SUBJ task, which suggests that, while gaining on sentiment analysis tasks, fine-tuning on NLI datasets leads to information loss on recognising the subjectivity of a sentence.

Extraction Difficulty As we have seen, the difference between SBERT and our model in our previous transfer comparison is small. Our hypothesis is that, since we concatenate the outputs of BERT and RGCN, the representations produced by our model are more complex, and that simple logistic regression lacks the ability to extract useful information from such complex embeddings. To assess this, we replace the logistic regression with a single hidden layer MLP (128 hidden units) which is widely used as a probing classifier. We focus on the comparison between our model and SBERT, re-running these two models with 5 random seeds, and report accuracy in the same fashion, except we adopt a more strict bold strategy to mark the difference (as explained in the caption).

	GloVe AVG	BERT-AVG	BERT-CLS	InferSent	USE	SBERT	Ours
MPQA	87.64±0.11	87.84±0.08	88.17±0.05	**90.32±0.12**	86.52±0.09	89.81±0.06	89.75±0.12
SST-5	44.35±0.11	47.33±0.22	48.03±0.45	44.93±1.14	47.67±0.06	48.57±0.53	**49.19±1.01**
SST-2	80.02±0.24	85.69±0.09	87.21±0.17	84.15±0.33	85.78±0.11	87.8±0.28	**87.99±0.28**
SUBJ	91.26±0.11	95.29±0.05	**95.48±0.1**	92.47±0.1	93.85±0.16	94.03±0.12	93.81±0.16
TREC	80.36±2.13	90.24±0.8	91.36±0.83	87.94±0.56	**92.36±0.32**	86.4±0.83	87.8±0.68
MRPC	72.79±0.21	73.43±0.77	71.68±0.48	**75.33±0.37**	71.2±0.61	74.68±0.75	74.9±0.74
MR	77.26±0.19	81.38±0.08	82.12±0.15	81.71±0.23	79.48±0.1	**82.77±0.22**	82.59±0.13
CR	78.9±0.1	87.12±0.31	87.33±0.23	86.34±0.52	86.03±0.23	88.99±0.16	**89.02±0.13**
AVG	76.57	81.04	81.42	80.40	80.36	81.63	**81.88**

Table 2: Results on SentEval evaluation with logistic regression. For MR, CR, MPQA and SUBJ, we use 10-fold cross validation and report accuracy on test-fold. For remaining tasks, results are reported on test set. We run 5 times with random seeds and report mean with standard deviation.

	SBERT	Ours
MPQA	89.98±0.16	90.11±0.13
SST-5	49.1±0.56	**50.5±0.3**
SST-2	88.51±0.71	88.39±0.39
SUBJ	94.1±0.12	94.05±0.17
TREC	86.96±0.32	**88.4±0.58**
MRPC	74.79±1.28	75.01±0.85
MR	82.7±0.16	82.56±0.14
CR	88.89±0.24	88.94±0.26
AVG	81.88	**82.25**

Table 3: Results on SentEval evaluation with MLP. Cells marked as bold only when the mean minus std is no worse than the mean plus std of the other model

As shown in Table 3, for some tasks, e.g. MR and CR, both models show stable performance cross different classifiers, and their performance remains similar when this more powerful extractor is used. However, for SST-5 (5-way sentiment classification) and TREC (6-way question-type classification), we see that clear improvements are obtained by our model, suggesting that the additional syntax supervision that we bring in through RGCNs is beneficial for fine-grained classification tasks. A similar pattern of results was found when we experimented with a 2 hidden layer MLP.

5 Conclusion

In this work, we show that SBERT can be improved by explicitly incorporating structural information. By using RGCNs to incorporate syntactic structure into supervision, our model is able to produce structure-aware, general-purpose sentence embeddings that achieve improved results on both unsupervised similarity comparison and transfer classification tasks, when compared against previous sentence encoders. By extending probing classifiers, we further show that our syntax-informed supervision method is particularly beneficial for fine-grained tasks such as SST-5 and TREC.

6 Acknowledgement

We thank all anonymous reviewers for their helpful comments, and NVIDIA for the donation of the GPU that supported our work.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Wanxiang Che, Longxu Dou, Yang Xu, Yuxuan Wang, Yijia Liu, and Ting Liu. 2019. HIT-SCIR at MRP 2019: A unified pipeline for meaning representation parsing via efficient training and effective encoding. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 76–85, Hong Kong. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of*

the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1816–1829.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 350–es.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3294–3302.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2020. Improving bert with syntax-aware local attention. *arXiv preprint arXiv:2012.15150*.

Tao Liu, Xin Wang, Chengguo Lv, Ranran Zhen, and Guohong Fu. 2020. Sentence matching with syntax- and semantics-aware bert. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3302–3312.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

T.y.s.s Santosh, Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. SaSAKE: Syntax and semantics aware keyphrase extraction from research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5372–5383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Yinchuan Xu and Junlin Yang. 2019. Look again at the syntax: Relational graph convolutional network for gendered ambiguous pronoun resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 96–101.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.