Neural Multi-Task Learning for Stance Prediction

Wei Fang, Moin Nadeem, Mitra Mohtarami, James Glass
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA
{weifang,mnadeem,mitram,glass}@mit.edu

Abstract

We present a multi-task learning model that leverages large amount of textual information from existing datasets to improve stance prediction. In particular, we utilize multiple NLP tasks under both unsupervised and supervised settings for the stance prediction task. Our model obtains state-of-the-art performance on a public benchmark dataset, Fake News Challenge, outperforming current approaches by a wide margin.

1 Introduction

For journalists and news agencies, fact checking is the task of assessing the veracity of information and claims. Due to the large volume of claims, automating this process is of great interest to the journalism and NLP communities. A main component of automated fact-checking is stance detection which aims to automatically determine the perspective (stance) of given documents with respect to given claims as *agree*, *disagree*, *discuss*, or *unrelated*.

Previous work (Riedel et al., 2017; Hanselowski et al., 2018; Baird et al., 2017; Chopra et al., 2017; Mohtarami et al., 2018; Xu et al., 2018) presented various neural models for stance prediction. One of the challenges for these models is the limited size of human-labeled data, which can adversely affect the resulting performance for this task. To overcome this limitation, we propose to supplement data from other similar Natural Language Processing (NLP) tasks. However, this is not a straightforward process due to differences between NLP tasks and data sources. We address this problem using an effective multi-task learning approach which shows sizable improvement for the task of stance prediction on the Fake News Challenge benchmark dataset. The contributions of this work are as follows:

- To the best of our knowledge, we are the first to apply multi-task learning to the problem of stance prediction across different NLP tasks and data sources.
- We present an effective multi-task learning model, and investigate the effectiveness of different NLP tasks for stance prediction.
- Our model outperforms the state-of-the-art baselines on a publicly-available benchmark dataset with a substantial improvement.

2 Multi-task Learning Framework

We propose a multi-task learning framework which utilizes the commonalities and differences across existing NLP datasets and tasks to improve stance prediction performance. More specifically, we use both unsupervised and supervised pre-training on multiple tasks, and then fine-tune the resulting model on our target stance prediction task.

2.1 Model Architecture

The architecture of our model is shown in Figure 1. We use a transformer encoder (Vaswani et al., 2017) that is shared across different tasks to encode the inputs before feeding the contextualized embeddings into task-specific output layers. In what follows, we explain different components of our model.

Input Representation The input sequence $x = \{x_1, \ldots, x_l\}$ of length l is either a single sentence or multiple texts packed together. The input is first converted to word piece sequences (Wu et al., 2016) and, in the case of multiple texts, a special token `[SEP]` is inserted between the tokenized sequences. Another special token `[CLS]` is inserted at the beginning of the sequence, which corresponds to the representation of the entire sequence.

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 13–19
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics

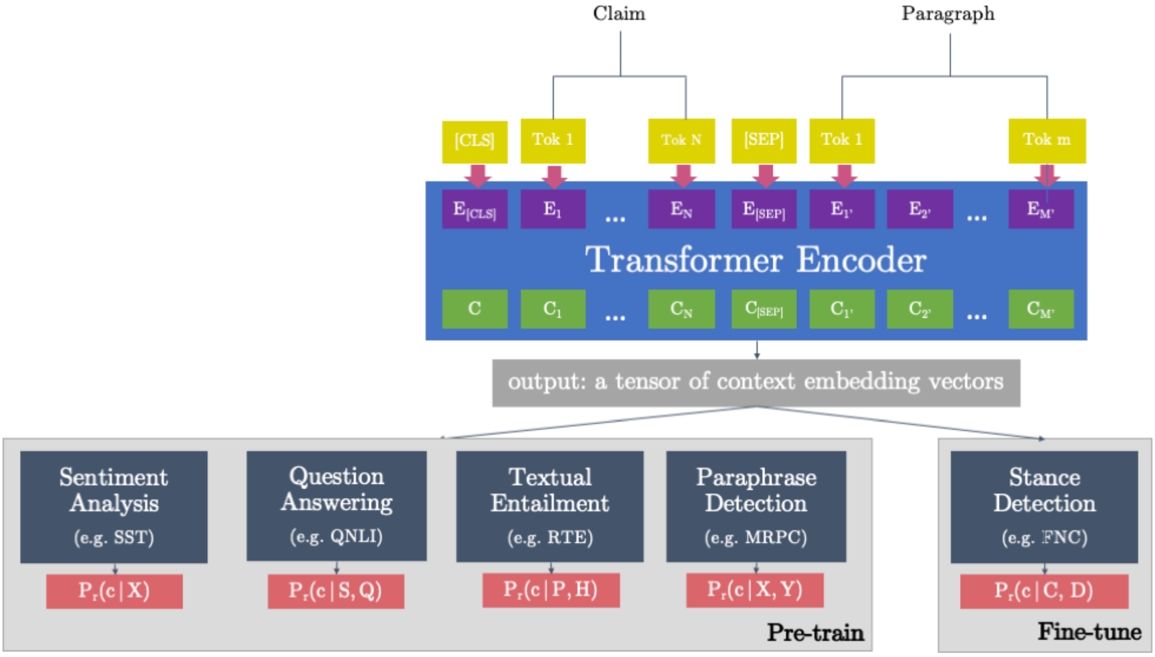

Figure 1: The architecture of our multi-task learning model for stance prediction.

Transformer Encoder We use a bidirectional Transformer encoder that takes x as input and produces contextual embedding vectors $\mathbf{C} \in \mathbb{R}^{d \times l}$ via multiple layers of self-attention (Devlin et al., 2019).

Task-specific Output Layers For single-sentence classification tasks, we take the vector from the first column in $\mathbf{C}$, corresponding to the special token `[CLS]`, as the semantic representation of the input sentence x. We then feed this vector through a linear layer followed by `softmax` to obtain the prediction probabilities.

For pairwise classification tasks, we use the answer module from the stochastic answer network (SAN) (Liu et al., 2018) as the output classifier. It performs K-step reasoning over the two pieces of text with bi-linear attention and a recurrent mechanism, producing output predictions at each step and iteratively refining its predictions. At training time, some predictions are randomly discarded (stochastic dropout) before averaging, and during inference all output probabilities are utilized.

2.2 Unsupervised Pre-training

To utilize large amounts of text data, we use the BERT model which pre-trains the transformer encoder parameters with two unsupervised learning tasks: masked language modeling, for which the model has to predict a randomly masked out word in the sequence, and next sentence prediction, where two sentences are packed and fed into the encoder and the embedding corresponding to the `[CLS]` token is used to predict whether they are adjacent sentences (Devlin et al., 2019).

2.3 Multi-task Supervised Pre-training

In addition to learning contextual representations under an unsupervised setting with large data, we investigate whether existing NLP tasks that are conceptually similar to stance prediction can improve performance. We introduce four types of such tasks for pre-training:

Textual Entailment: Given two sentences, a premise and an hypothesis, the model determines whether the hypothesis is an *entailment*, *contradiction*, or *neutral* with respect to the premise. Since stance prediction could be cast as a textual entailment task, we investigate if the addition of this task will benefit our model.

Paraphrase Detection: Given a pair of sentences, the model should predict whether they are semantically equivalent. This task is considered because we may be able to benefit from detecting document sentences that are equivalent to claims.

Question Answering: Question answering is similar to the stance prediction task in that the model has to make a prediction given a question and a passage containing several sentences.

Sentiment Analysis: Fake claims or articles may exhibit stronger sentiment, thus we explore if pre-training on this task would be beneficial.

2.4 Training Procedure and Details

There are two stages in our training procedure: multi-task supervised pre-training, and fine-tuning on stance prediction. Before the training stages, the transformer encoder is initialized with pre-trained parameters to take advantage of knowledge learned from unlabeled data[1].

During multi-task pre-training, we randomly pick an ordering on tasks between each epoch, and train on 10% of a task's training data for each task in that order. This process is repeated 10 times in each epoch so that all the training examples are trained once. The shared encoder is learned over all tasks while each task-specific output layer is learned only for its corresponding task.

For fine-tuning, the task-specific output layers for pre-training are discarded, and a randomly initialized output layer is added for stance prediction. Then the entire model is fine-tuned over the training set for stance prediction.

For both multi-task pre-training and fine-tuning, we train with cross-entropy loss at each output layer. We use the Adam optimizer (Kingma and Ba, 2014) with learning rate of $3e$-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and mini-batch size of 16 for 10 epochs. For the SAN answer module we set $K = 5$ and use stochastic dropout rate of 0.1.

3 Experiments

3.1 Data

The BERT model was pre-trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia. For multi-task pre-training, we use the following datasets:

SNLI Stanford Natural Language Inference is the standard entailment classification task that contains 549K training sentence pairs after removing examples with no gold labels (Bowman et al., 2015). The relation labels are *entailment*, *contradiction*, and *neutral*.

MNLI Multi-genre Natural Language Inference is a large-scale entailment classification task from a diverse set of sources with the same relation classes as SNLI (Williams et al., 2018). We use its training set that contains 393K pairs of sentences.

RTE Recognizing Textual Entailment is a binary entailment task with 2.5K training examples (Wang et al., 2019).

QQP Quora Question Pairs[2] is a QA dataset for binary classification where the goal is to predict whether two questions are semantically equivalent. We use its 364K training examples for pre-training.

MRPC Microsoft Research Paraphrase Corpus consists of automatically extracted sentence pairs from new sources, with human annotations for whether the pairs are semantically equivalent (Dolan and Brockett, 2005). The training set used for pre-training contains 3.7K sentence pairs.

QNLI Question Natural Language Inference (Wang et al., 2019) is a QA dataset which is derived from the Stanford Question Answering Dataset (Rajpurkar et al., 2016) and used for binary classification. For a given question-sentence pair, the task is to predict whether the sentence contains the answer to the question. QNLI contains 108K training pairs.

SST-2 Stanford Sentiment Treebank is used for binary classification for sentences extracted from movie reviews (Socher et al., 2013). We use the GLUE version that contains 67K training sentences (Wang et al., 2019).

IMDB The Large Movie Review Dataset contains 50K movie reviews which are categorized as either *positive* or *negative* in terms of sentiment orientation (Maas et al., 2011).

For fine-tuning on stance prediction, we use the dataset provided by the Fake News Challenge Stage 1 (**FNC-1**)[3], consisting of a total of 75K claim-document pairs collected from a variety of sources such as rumor sites and social media. The claim-document relation classes are: *agree*, *disagree*, *discuss*, and *unrelated*. The FNC-1 dataset has an imbalanced distribution over stance labels, especially lacking data for *agree* (7.3%), and *disagree* (1.7%) classes.

3.2 Evaluation Metrics

For evaluation, the standard measures of **accuracy** and **macro-F1** are used. Additionally, as per previous work, **weighted accuracy** is also reported, which is a two-level scoring scheme that gives 0.25 weight to predicting examples as *related* v.s. *unrelated* correctly, and an additional 0.75 weight to classifying related examples as *agree*, *disagree*, and *discuss* correctly.

[1] In this work we use the pre-trained BERT weights released by the authors.

[2] `https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs`

[3] `http://www.fakenewschallenge.org`

	Model	Auxiliary Data	Weigh. Acc.	Acc.	Macro-F1
1	Gradient Boosting	-	75.2	86.3	46.1
2	TALOS	-	82.0	89.1	57.8
3	UCL	-	81.7	88.5	57.9
4	Memory Network	-	81.2	88.6	56.9
5	Adversarial Adaptation	FEVER	80.3	88.2	60.0
6	TransLinear	-	84.9	89.3	66.3
7	TransSAN	-	85.1	90.3	67.9
		Textual Entailment			
8	MTransSAN	SNLI	86.7	91.9	72.3
9	MTransSAN	MNLI	86.4	90.8	71.0
10	MTransSAN	RTE	85.6	90.7	69.3
11	MTransSAN	SNLI, MNLI, RTE	86.1	91.3	71.6
		Paraphrase Detection			
12	MTransSAN	QQP	87.6	92.1	74.1
13	MTransSAN	MRPC	87.0	92.0	73.5
14	MTransSAN	QQP, MRPC	**88.0**	**92.3**	**74.4**
		Question Answering			
15	MTransSAN	QNLI	86.5	91.2	71.9
		Sentiment Analysis			
16	MTransSAN	SST	86.7	91.8	70.0
17	MTransSAN	IMDB	85.6	91.2	70.4
18	MTransSAN	SST, IMDB	86.5	91.7	71.1
		Joint			
19	MTransSAN	SNLI, MNLI, QNLI	84.7	90.6	70.1
20	MTransSAN	MNLI, RTE, QQP, MRPC, QNLI, SST	87.0	91.6	71.8
21	MTransSAN	SNLI, MNLI, RTE, QQP, MRPC, QNLI, SST, IMDB	86.5	91.6	72.1

Table 1: Results on the FNC test data. TransLinear, TransSAN and MTransSAN show our model where the first two are based on a transformer followed by a MLP or neural model, and the later further uses multi-task learning.

3.3 Baselines

We compare our model with existing state-of-the-art stance prediction models including the top-ranked models from FNC-1 and neural models:

Gradient Boosting This baseline[4] uses a gradient-boosting classifier with hand-crafted features including n-gram features, and indicator features for polarity and refutation.

TALOS (Baird et al., 2017) An ensemble of gradient-boosted decision trees and a convolutional neural network.

UCL (Riedel et al., 2017) A Multi-Layer Perceptron (MLP) with Bag-of-Words and similarity features extracted from claims and documents.

Memory Network (Mohtarami et al., 2018) A feature-light end-to-end memory network that attends over convolutional and recurrent encoders.

Adversarial Domain Adaptation (Xu et al., 2018) This baseline uses a domain classifier with gradient reversal on top of a convolutional network and TF-IDF features to perform adversarial domain adaptation from another fact-checking dataset (Thorne et al., 2018) to FNC.

3.4 Results and Discussion

The performance of the existing models are shown in Table 1 from rows 1–5, and our models (MTransSAN) are in rows 8–21. All variants of MTransSAN consistently outperform existing models on all three metrics by a considerable margin. In particular, our best MTransSAN (row 14) **achieves 6.0 and 14.4 points of absolute improvement** in terms of weighted accuracy and macro-F1, respectively, over existing state-of-the-art results.

We also compare MTransSAN versus a model with the same architecture but without pre-training on the NLP tasks (TransSAN), shown in row 7, and another version of that model with a linear layer instead of the SAN answer module (TransLinear), shown in row 6. Using the SAN answer module improves over a linear layer for all three metrics, and generally most MTransSAN models outperform the TransSAN model. Our best MTransSAN model exceeds TransSAN by 3.1 and 6.5 points in weighted accuracy and macro-F1, respectively, justifying the effectiveness of model pre-training with NLU tasks. Note that even the TransLinear model outperforms previously state-of-the-art models by a wide margin, suggesting that a neural model pre-trained on large amounts

[4] `https://github.com/FakeNewsChallenge/fnc-1-baseline`

of unlabeled data and fine-tuned on stance prediction is superior to models that require hand-crafted features.

Additionally, we conduct experiments where we use different combinations of language understanding tasks for pre-training. We pre-train with single tasks, multiple tasks with the same task type, and joint learning across multiple task types. For textual entailment (rows 8–11), we see that pre-training on SNLI gives us best improvement, and that pre-training across all three entailment tasks did not improve compared to just training on SNLI. However, for paraphrase detection (rows 12–14) the combination of QQP and MRPC gives us the best results across all MTransSAN models. This suggests that the paraphrase detection might be the most useful task type among the NLP tasks in terms of boosting stance prediction performance. Question answering and sentiment analysis (rows 15–18), on the other hand, give lower performance improvements compared to paraphrase detection. Models trained on joint tasks (rows 19–21) do not outperform our best model either.

Overall, we find that utilizing the BERT model results in large improvements compared to the baselines, which is not unexpected given the success of BERT. We also show that our multi-task learning approach gives even further improvements upon BERT by a wide margin.

4 Related Work

Stance Prediction. This task is an important component for fact checking and veracity inference. To address stance prediction, (Riedel et al., 2017) used a Multi-Layer Perceptron (MLP) with bag-of-words and similarity features extracted from input documents and claims, and (Hanselowski et al., 2018) presented a deep MLP trained using a rich feature representation, based on unigrams, non-negative matrix factorization, latent semantic indexing. (Baird et al., 2017) presented an ensemble of gradient-boosted decision trees and a deep convolutional neural network, while (Chopra et al., 2017) proposed a model based on bi-directional LSTM and attention mechanism. While, these works utilized a rich handcrafted features, (Mohtarami et al., 2018, 2019) proposed strong end-to-end feature-light memory networks for stance prediction in mono- and cross-lingual settings. Recently, (Xu et al., 2018) presented a state-of-the-art model based on adversarial domain adaptation with more labeled data, but they limited their model to only using data from the same stance prediction task. In this work, we remove this limitation and used labeled data from other tasks that are similar to stance prediction through multi-task learning.

Multi-task and Transfer Learning. Multi-task and transfer learning have been long-studied problems in machine learning and NLP (Caruana, 1997; Collobert and Weston, 2008; Pan and Yang, 2010). More recently, numerous methods on unsupervised pre-training of deep contextualized models for transfer learning have been proposed (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Radford et al., 2019; Dai et al., 2019; Liu et al., 2019), and (Conneau et al., 2017; McCann et al., 2017) presented supervised pre-training methods for NLI and translation. Recent work on multi-task learning has focused on designing effective neural architectures (Hashimoto et al., 2017; Søgaard and Goldberg, 2016; Sanh et al., 2018; Ruder et al., 2017). Combining these two lines of work, (Liu et al., 2019; Clark et al., 2019) explored fine-tuning the contextualized models with multiple natural language understanding tasks. In this work, we depart from previous works by specifically studying the effects of multi-task fine-tuning for the stance prediction task with pre-trained models.

5 Conclusion and Future Work

We present an effective multi-task learning model that transfers knowledge from existing NLP tasks to improve stance prediction. Our model outperforms state-of-the-art systems by 6.0 and 14.4 points in weighted accuracy and macro-F1 respectively on the FNC-1 benchmark dataset. In future, we plan to further investigate our model to more specifically identify and illustrate its source of improvement, improve our transfer learning approach for better fine-tuning, and investigate the utility of our model in other fact-checking subproblems such as evidence extraction.

Acknowledgments

We thank the anonymous reviewers for their insightful comments, suggestions, and feedback. This research was supported in part by HBKU Qatar Computing Research Institute (QCRI) and DSTA of Singapore.

References

Sean Baird, Doug Sibley, and Yuxi Pan. 2017. Talos targets disinformation with fake news challenge victory. https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html.

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.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

Sahil Chopra, Saachi Jain, and John Merriman Sholar. 2017. Towards automatic identification of fake news: Headline-article stance detection with lstm attention models.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2019. Bam! born-again multi-task networks for natural language understanding. In *ACL*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704. Association for Computational Linguistics.

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 e-prints*, page arXiv:1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. 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 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, New Orleans, Louisiana.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *ArXiv:1707.03264*.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Latent multi-task architecture learning.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A hierarchical multi-task approach for learning embeddings from semantic tasks. *CoRR*, abs/1811.06031.

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.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv e-prints*, page arXiv:1609.08144.

Brian Xu, Mitra Mohtarami, and James Glass. 2018. Adversarial doman adaptation for stance detection. In *Proceedings of the Thirty-second Annual Conference on Neural Information Processing Systems (NIPS)–Continual Learning*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv e-prints*, page arXiv:1906.08237.

Yukun Zhu, Ryan Kiros, Rich 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. In *The IEEE International Conference on Computer Vision (ICCV)*.