Natural Language Inference for Fact-checking in Wikipedia

Mykola Trokhymovych[1] and Diego Saez Trumper[2]

[1]Ukrainian Catholic University, Faculty of Applied Sciences, Lviv, Ukraine
trokhymovych@ucu.edu.ua
[2]WikiMedia Foundation, Barselona, Spain
diego@wikimedia.org

Abstract. The incoming flow of information is continuously increasing along with the disinformation part that can harm society. In the context of Wikipedia, automatically filtering unreliable content is very important to help the editors keep Wikipedia as free as possible of disinformation. This project aims to implement software as an open API that will automatically perform a facts validation process. In Natural Language Processing (NLP), this task is related to Natural Language Inference (NLI), where a claim is compared with reference to determine whether the claim is correct, incorrect, or unrelated. In this work, we analyze and compare state-of-the-art (SOTA) approaches. Although there were recently many advances in the precision of NLI models, efficiency remains an open problem. Our goal is to build a production-ready model with both high accuracy and efficiency. We observe the best works in word-based models and transfer approaches to sentence based models to achieve our research goal.

Keywords: fact-checking · fact-verification · natural language inference · natural language processing · textual entailment.

1 Introduction and motivation

1.1 Problem observation

Disinformation can influence elections, stock prices, and even how we treat ourselves from a virus. False facts are spreading faster than the truth and can negatively impact society and business [22].

Social networks allow people to post information whatever and whenever they want. Research showed that in 2016, Trump supporters' activity influenced the dynamics of the top fake news spreaders, which had an impact on US president elections [3]. In 2013, $130 billion in stock value were lost just because of one fake tweet about an "explosion" that injured Barack Obama [1].

Manual fact-checking is time-consuming and can come too late. Automation of this process reduces time to "stick" in the audience's minds [5]. It can prevent propaganda by filtering manipulation and false facts in nearly real-time. In the case of Wikipedia, we have more than 20 million user edits per month according

to official stats[1], which requires approximately eight fact-checking procedures per second, which is difficult to do manually.

1.2 Motivation

Automated fact-checking and NLI has a significant social impact, and it is currently developing very fast in academia. However, there is a gap between research achievements and applicability in real life. A project aims to observe the state-of-the-art solutions to the NLI problem, reproduce the results, propose possible improvements, and develop an open-source tool to perform the automated fact-checking. We also aim to learn the recent NLI field changes to make the project results comparable to the most valuable works in this field and show our possible contribution.

2 Related work

In this section, we observe papers in four major categories: *(i)* fact-checking problem formulation and observation; *(ii)* language modeling; *(iii)* NLI state of the art solutions. For each paper, we observed the main contribution and realization details. Such analysis allows us to have a general overview of that topic and the most recent results.

2.1 Problem formulation and datasets observation

The problem of fact-checking was initially used in journalism as an essential part of news reporting. The first published datasets were collected from the political domain. The collection consisted of 221 labeled claims checked by Politifact and Channel4 with related sources of evidence [21]. After that, Wang released similar data collection but much larger, containing 12.8K labeled claims from Politifact [24]. However, this is still a too small collection of data to train large models on them. If we speak about the NLI task, the data collection procedure is complicated, as it requires manual annotation. It is almost impossible to collect massive datasets for model training like the 160Gb dataset collected RoBERTa language model training [14].

In 2015 SNLI dataset was presented and became the primary benchmark dataset used for NLI problem [4]. Even though it is not specialized on some specific topic like politics, it is large enough (570K pairs of sentences) to train large models. SNLI consists of pairs of sentences with relation labels: entailment, contradiction, or neutral. It is created by presenting crowd workers with a sentence, asking them to generate three new sentences (hypotheses) for each entailment class [4].In 2018 MultiNLI dataset was presented, which is almost the same with SLNI [4], but has improved topics coverage and difficulty of claims [25].

As the primary benchmark dataset, SNLI has one important drawback: models that did not even look at the evidence perform well on the NLI task. This

[1] Wikimedia Statistics dashboard `https://stats.wikimedia.org`.

behavior is explicitly observed in [11], where authors reveal linguistic annotation artifacts in SNLI. They show specific words in texts which are highly correlated with certain inference classes.

There are also alternatives for SNLI and MNLI. In [18] researchers present their dataset *WIKIFACTCHECK-ENGLISH* (124K triplets of sentences), which consists of real-world claims from Wikipedia. We should also mention FEVER dataset [20], which has a more complex structure based on the Wikipedia dump.

2.2 Masked language modeling

The most crucial part of a modern NLI solution is language models. The recent state-of-the-art solutions are built on top of them. To create a valuable NLI model, we also need to observe the literature about language modeling as a base for further research. The most recent language modeling results are based on transformers architecture.

One of the most valued recent contributions to NLP is the BERT architecture, which stands for Bidirectional Encoder Representations from Transformers [9]. BERT model made revolution in NLP field. It significantly moved state-of-the-art scores for several NLP tasks by presenting new architecture for language modeling. It showed point absolute improvement 7.7% on GLUE score [23]. The authors present a model allowing it to be bidirectional and utilize the masked language model (MLM) as a pretraining objective. The training process is built on masking some of the tokens and predicting them based only on their context [9]. Then RoBERTa model was presented, which improved previous results of BERT. The authors of [14] introduce a replication study of BERT. The research is built on removing the NSP loss, using a much bigger dataset for training that consists of 160GB of text, and increasing the number of pretraining steps from 100K to 500K.

In this section, we also describe Sentence-BERT. Authors present a way how to train sentence embeddings instead of word embeddings using pretrained transformer model and Siamese network [17]. This approach allows the dump of precalculated sentence embeddings, reusing them for different tasks, improving the model's efficiency, and making transformer models like BERT and RoBERTa possible to use in high-load production tasks.

2.3 State of the art solution

We divided SOTA solutions in two groups: *(i)* sentence based and *(ii)* word-based. The difference is that in the case of a word-based solution, sentences are represented as a set of word vectors, while in the sentence based we are trying to build a single vector as a sentence representation and then use it for building a model that will solve NLI task. Sentence based solutions are usually faster, more applicable in real life as vectors can be cached. However, word-based solutions are more precise. For each paper, we defined the contribution along with the approach and SNLI score. We structured this analysis in two tables 1, 2.

Table 1: Observation of literature, word based solutions.

Name, source	**Description, contribution**	**Explanation of approach**	**SNLI Score**
Neural Natural Language Inference Models Enhanced with External Knowledge [6]	Use external knowledge from words meaning. State-of-the-art performance with a relatively small number of parameters of 4.3m.	Use information about synonymy, antonym, hypernym and hyponymy existance in attention layer.	88.6
Natural Language Inference over Interaction Space [10]	Combines both NLP and computer vision (CV) approaches. Based on a high-level understanding of the sentence pair relation.	Create a tensor representation of pairs of texts using their word embeddings, manipulate it to extract semantic features, and do the classification.	88.9
Multi-Task Deep Neural Networks for Natural Language Understanding [13]	Training BERT model on multiple natural language understanding (NLU) tasks simultaneously, benefiting from a regularization.	Use BERT model along with Lexicon encoder, adding extra information about position and word's segment. Fine-tune the model simultaneously for four different NLP tasks	91.6
Semantics-aware BERT for Language Understanding [26]	Integrates contextualized features into language model. Extend the language representation model with semantics. State-of-the-art result for SNLI.	Use semantic role labeling (SRL) model with BERT to parse the predicate-argument structure. Fine-tune the model separately for different tasks.	91.9
Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data [16]	The latest state-of-the-art result, using idea and results of [13]. The current approach to learning different sets of parameters while fine-tuning different NLP tasks. Demonstrate faster fine-tuning as most parameters are frozen and dataset balanced across different tasks that reduce data to around 60%.	Upgrade standard transformer architecture with five additions: conditional attention, conditional alignment, conditional layer normalization, conditional adapters, and multi-task uncertainty sampling to have a specific approach to each NLP task saving unified architecture.	92.1

Works [10,6] represent models that are not based on BERT, as were created earlier. However they present very different approaches along with good results. When [13,26,16] present results of models based on BERT architecture with various modifications and learning strategies.

Most of the best sentence based models are build using LSTM architecture. Also, we see that scores for those types of models are lower than word-based. However, such models can be used in production high-load tasks as they are

Table 2: Observation of literature, sentence based solutions.

Name, source	**Description, contribution**	**Explanation of approach**	**SNLI Score**
Sentence Embeddings in NLI with Iterative Refinement Encoders [19]	Hierarchical BiLSTM model with Max Pooling for building sentence embeddings and further tuning for NLI task. Present error analysis.	Use advanced architecture based on iterative refinement strategy. Build sentence embeddings and then use the MLP model to use those vectors in NLI task	86.6
Dynamic Meta-Embeddings for Improved Sentence Representations [12]	Project utilize dynamic meta-embeddings for sentence embeddings composition and building further model on top of them.	Use composition of different embeddings like Word2Vec [15] or fasttext [2]. Learns the weights for the composition of defined vectors and uses BiLSTM to compose the sentence embeddings used for the NLI task.	86.7
Enhanced LSTM for Natural Language Inference [7]	Present carefully designing sequential inference that outperforms complicated network architectures and state new state-of-the-art result for sentence based models	Use the BiLSTM block to represent a word and its context and inference composition before the final prediction. Use Local Inference Modeling for determining the overall inference between these pair of texts.	88.6

lighter, faster. Better efficiency of sentence based models is caused by their ability to cache intermediate results like sentence embeddings and lighter architecture.

3 Research plan

3.1 Open problems

The most recent research for sentence based NLI models like [7,12,19] are not using BERT models that made a great boost in word-based models case. Also, sentence based models have significantly worse accuracy comparing to word-based models on the SNLI dataset. One of the open challenges is using transformer models for sentence-based NLI models and improving their performance.

One more open problem is dataset selection and filtering. Most of the real datasets for fact-checking are too small to train machine learning models to automate a process. There are a few relatively big datasets like SNLI. However, due to their nature, they are not perfect as they have many artifacts left by crowd workers that were creating datasets. Another open problem is collecting appropriate data, preparing and filtering content used for training a model.

3.2 Research goal

In order to solve the fact-checking problem, we will concentrate on sentence based NLI models and try to use the results of [9,14] for them. Also, we will use [17] as a first approximation of the method for sentence embeddings creation. There is also an idea to transfer the approach from best word-based models [13], and [16] to sentence based model so that we can use semantic information and multitask-learning for fine-tuning. The primary dataset for evaluation will be SNLI using an accuracy metric. We understand all the drawbacks presented in [11], so another aim of the research is to do the explicit errors observation and model interpretation tool.

To sum up, we formulated such research objectives:

1. Reproduce SOTA results. Measure the efficiency of such solutions.
2. Research and implement sentence based model using masked language models like BERT
3. Experiment with dataset filtering techniques and measure how does it influence the performance of models
4. Implement API to allow the community to use the research results

Additionally, as for further research, we consider transferring the results of research on pair classification problem formulation as in [4] to the more realistic scenario of searching the proof to the claim from defined knowledge base FEVER formulation task [20].

3.3 Plan for the Research

Important part of the research work is planning. We should mention that currently we are finishing first part of research. Our estimated schedule with main checkpoints is presented in the table 3.

Table 3: Plan for research.

Checkpoint	**Tasks**	**Status**
10 November	Kick off call with mentor and beginning of project	Done
30 November	Review related literature, find open research problems and refine research goal	Done
10 December	Decide about what dataset to use for research and make exploratory data analysis. Prepare submissions for master symposium	Done
20 December	Review and reproduce SOTA results	In progress
20 January	Implement sentence based model and evaluate it on SNLI and MNLI datasets	to be done
1 February	Create an API for online fact checking and deploy it to Wikipedia servers	to be done
28 February	Prepare the technical report	to be done

4 Background and Results to Date

This section describes the datasets that we are going to use during the training and validation procedure. Also, we report our first results up to date.

4.1 Exploratory data analysis

The most recent research in the NLP field is highly tacked to data. The SOTA results are achieved not only because of innovative models but also because of significant amounts of data, acute filtering techniques, and understanding of data nature. The next step in the research was to observe and analyze data to get useful insights from them. As for our research, we decided to use those two datasets (SNLI and MNLI) datasets, as they are big enough and allow us to compare results with the most recent SOTA results. SNLI and MNLI datasets are consist of claim, related hypothesis, and label, which is either neutral, contradiction, or entailment 4.

Table 4: Data example from MNLI and SNLI datasets

Dataset	**Claim**	**Hypothesis**	**Label**
MNLI	The Old One always comforted Ca'daan, except today.	Ca'daan knew the Old One very well.	neutral
MNLI	At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	People formed a line at the end of Pennsylvania Avenue.	entailment
SNLI	A man inspects the uniform of a figure in some East Asian country.	The man is sleeping	contradiction
SNLI	An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral

Important to mention that all classes are well balanced and have almost the same amount of samples. We analyzed distributions of the length of three classes' claims and hypotheses and found out that the claims' length is equally distributed. At the same time length of the hypothesis are different within different classes. In the figure 1 we can see that hypothesis of entailment class is usually shorter than others. It can influence the model that will learn the length of a sentence instead of its meaning. The MNLI dataset situation is much better as distributions of the length of texts are more balanced, but the neutral class sentences are usually longer. Moreover, we see that the MNLI hypothesis is larger than SNLI, making it closer to the real world.

In our exploration, we found out the reason why certain words in the hypothesis are highly correlated with specific classes as it was discussed that [11]. We defined top-15 the most frequent hypothesis used by annotators and analyzed

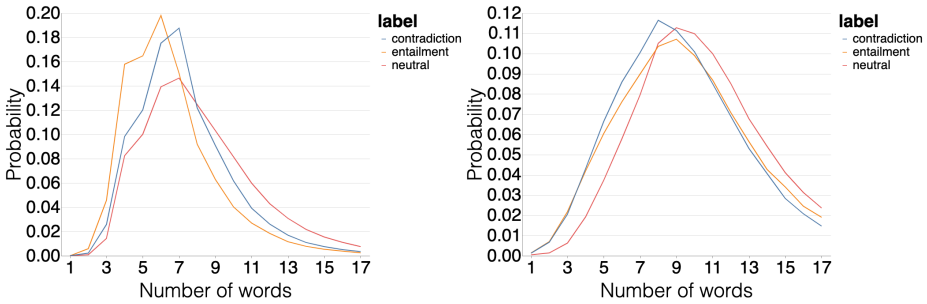

Fig. 1: Distribution of length of hypothesis in training dataset of SNLI (left picture) and MNLI (right picture)

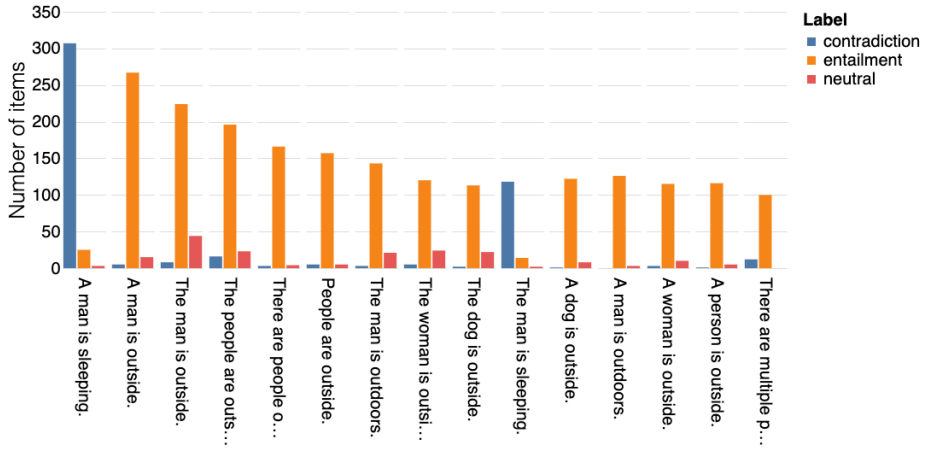

Fig. 2: SNLI dataset top-15 the most frequent hypothesis and their classes counts

the classes to which they correspond. We found out that frequent hypotheses are usually used in either entailment or contradiction class, represented in figure 2. It is also not natural behavior as the model will learn the only sentiment of hypothesis instead of the desired relation between claim and hypothesis. We will analyze how filtering out such patterns from training will influence the models' validation results in further research.

4.2 Experiments and results up to date

Reproducing SOTA results. The initial experiment was to reproduce the results of the SOTA models and measure their efficiency. In order to have a general overview, we decided to pick both word and sentence based models. For the first iteration we picked SemBERT [26] and HBMP [19] models. Those models

represent single model architecture without an ensemble, have top-performing results, and have official repositories with code[23].

The experiment of reproducing results is computationally expensive, so we used Microsoft Azure NC6 Promo virtual machine with six cores, 56 GiB RAM, and one K80 GPU. We are using the same configuration for all further experiments for unification. In our experiments, we measure accuracy on the SNLI test set and computation speed on inference.

Train set filtering experiment. Additionally, we decided to experiment with train set filtering techniques. During EDA, we mentioned that the most frequent hypothesis in the training dataset usually corresponds to only one class. So, we decided to build a new training dataset with such a filtering logic:

1. Picked top 1000 most frequent hypothesis
2. Find the distribution of classes for each hypothesis from picked
3. In case one class represented more than 80% of all items, we did an undersampling. After this procedure, it had the same amount of values as the second most frequent one.

As a result, we filtered out 16493 pairs of texts (3%) with 723 unique hypotheses and trained models with the created train dataset.

Building initial sentence based model. One of our research objectives is to build a sentence-based model that will use the BERT model for embeddings creation. Such an approach enables to cache intermediate results and reuse them for online prediction. We used the NLI model approach as a Siamese network using a bert-uncased-base model as a trainable encoder for sentences. The model architecture is presented in picture 3. The idea goes from [8], with the difference that we are not using multiplication of sentence vectors in concatenation layer, but using only original vectors and their absolute difference.

For example, we are planning to try different pre-trained BERT models, both case dependent and independent. We plan to use additional information like semantics and a more advanced softmax classifier instead of only one dense, fully connected layer in our initial architecture. Currently, we measured the accuracy and efficiency of the models trained on full and filtered SNLI trainsets.

Experiments results and discussion. We evaluated our models on 9824 samples from the SNLI testset. SemBERT was tested only on inference so we have missing accuracy on our custom filtered trainset. The results for discussed experiments are presented in the comparison table 5. Efficiency of the model is measured as average time in seconds required to classify one pair of texts from SNLI test dataset. As we can see, SemBert has the best accuracy results. However, it is 25x times slower on inference then sentence based model HBMP.

[2] Github repository for SemBERT model `https://github.com/cooelf/SemBERT`.
[3] Github repository for HBMP model `https://github.com/Helsinki-NLP/HBMP`.

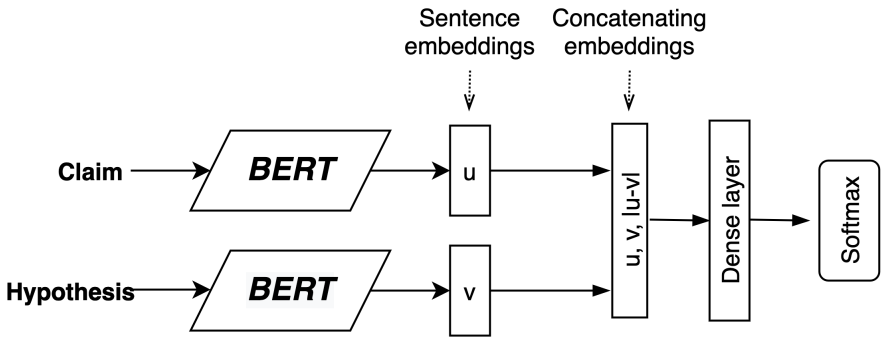

Fig. 3: Sentence based Siamese classifier with BERT encoder.

Table 5: Experiments results on SNLI datasets

Model	**Accuracy on full train**	**Accuracy on filtered train**	**Efficiency on inference**	**Efficiency on inference with caching**
SemBERT	91.9%	-	0.51 $\frac{s}{sample}$	-
HBMP	86.6%	83.5%	0.02 $\frac{s}{sample}$	0.012 $\frac{s}{sample}$
Bert sentence based NLI classifier	84.3%	84.0%	0.021 $\frac{s}{sample}$	0.0001 $\frac{s}{sample}$

It does not allow caching, as we usually can not precalculate relations for all possible pairs of text. However, in sentence based models, we can do so, and as we see, it significantly improves efficiency both for HBMP and our custom sentence based model. Important early result is that the BERT sentence-based model is less sensitive to removed annotation artifacts from trainset than HBMP, which shows the model's ability to generalize the data better.

5 Summary and discussion

Demand for reliable automatic fact-checking systems will grow with the increase in the amount of information available online. Recent research showed significant improvements in the field of NLI. However, efficiency is still an open problem for SOTA solutions. Also, there is a lack of good quality data that can be used for training, so a possible contribution to this field could be a good-quality annotated dataset.

During our research, we found out that sentence based models can have significantly better efficiency with lower accuracy at the same time. We have already built the validation pipeline and initial BERT sentence based model. An important result of initial research is that the BERT sentence based model is more

stable to dataset changes than previous sentence based architectures. We assume that such architecture is less dependent on annotation artifacts. In further research, we are planning to continue experiment with different BERT sentence-based architectures and additional features like semantics to improve models' accuracy. Also, we are planning to validate our results using other datasets.

6 Acknowledgments

We gratefully thank the Ring company for partially supporting studies in UCU with Teacher Assistance Stipend. We are also grateful to Microsoft Azure for providing student sponsorship for their services.

References

1. Forbes, Can 'Fake News' Impact The Stock Market? (2017 (accessed November 24, 2020)), https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/?sh=3f8630ae2fac
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017). https://doi.org/10.1162/tacl$_{a0}$0051, https://www.aclweb.org/anthology/Q17-1010
3. Bovet, A., Makse, H.: Influence of fake news in twitter during the 2016 us presidential election. Nature Communications **10** (01 2019). https://doi.org/10.1038/s41467-018-07761-2
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. CoRR **abs/1508.05326** (2015), http://arxiv.org/abs/1508.05326
5. Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., Tannier, X.: A content management perspective on fact-checking. In: Companion Proceedings of the The Web Conference 2018. p. 565–574. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3184558.3188727, https://doi.org/10.1145/3184558.3188727
6. Chen, Q., Zhu, X., Ling, Z., Inkpen, D., Wei, S.: Natural language inference with external knowledge. CoRR **abs/1711.04289** (2017), http://arxiv.org/abs/1711.04289
7. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H.: Enhancing and combining sequential and tree LSTM for natural language inference. CoRR **abs/1609.06038** (2016), http://arxiv.org/abs/1609.06038
8. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. pp. 670–680 (09 2017). https://doi.org/10.18653/v1/D17-1070
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
10. Gong, Y., Luo, H., Zhang, J.: Natural language inference over interaction space. CoRR **abs/1709.04348** (2017), http://arxiv.org/abs/1709.04348

11. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. CoRR **abs/1803.02324** (2018), http://arxiv.org/abs/1803.02324
12. Kiela, D., Wang, C., Cho, K.: Dynamic meta-embeddings for improved sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1466–1477. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1176, https://www.aclweb.org/anthology/D18-1176
13. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4487–4496. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1441, https://www.aclweb.org/anthology/P19-1441
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
15. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
16. Pilault, J., Elhattami, A., Pal, C.: Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters less data (2020)
17. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. CoRR **abs/1908.10084** (2019), http://arxiv.org/abs/1908.10084
18. Sathe, A., Ather, S., Le, T.M., Perry, N., Park, J.: Automated fact-checking of claims from Wikipedia. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6874–6882. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.849
19. Talman, A., Yli-Jyrä, A., Tiedemann, J.: Sentence embeddings in nli with iterative refinement encoders. Natural Language Engineering **25**(4), 467–482 (Jul 2019). https://doi.org/10.1017/s1351324919000202, http://dx.doi.org/10.1017/S1351324919000202
20. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. CoRR **abs/1803.05355** (2018), http://arxiv.org/abs/1803.05355
21. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In: LTCSS@ACL (2014)
22. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). https://doi.org/10.1126/science.aap9559, https://science.sciencemag.org/content/359/6380/1146
23. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. CoRR **abs/1804.07461** (2018), http://arxiv.org/abs/1804.07461
24. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. CoRR **abs/1705.00648** (2017), http://arxiv.org/abs/1705.00648
25. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. CoRR **abs/1704.05426** (2017), http://arxiv.org/abs/1704.05426
26. Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., Zhou, X.: Semantics-aware bert for language understanding (2020)