UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

SEMINAR

The Transformer Architecture: Facilitating Effective Pretraining and Task Adaptation

Antonio Lukić
Mentor: *prof. dr. sc. Jan Šnajder, mag. ing. David Dukić*

Zagreb, February 2024.

CONTENTS

1. Introduction

A good portion of the population has heard, encountered, or used ChatGPT or other applications powered by large language models. The field of Natural Language Processing (NLP) went through an intense transformation in recent years, largely driven by the models based on transformer architecture. These models, such as BERT (Devlin et al., 2018) and GPT (Generative Pre-Training) (Radford et al., 2018) have left a mark on how we interact with language technologies. From the appearance of these models, they have transformed society by automating tasks and enhancing communication and content creation. It has also raised concerns about job displacement, misinformation, and ethics. Understanding how to effectively employ these language technologies is equally important. To address these challenges, it is important to analyze where these models came from and what their direction of development is. Foundational papers, such as "Attention Is All You Need" (Vaswani et al., 2017), laid the groundwork for the Transformer architecture, enabling further advancements.

Throughout this work, I will analyze five papers associated with Transformer model architecture. The analysis will start with foundational papers and then some of the contemporary developments in the NLP field, focusing on the effective pretraining and adaptation capabilities of the Transformer architecture. I will study different approaches presented in the papers and analyze their results. Finally, I will formulate conclusions about the current state of the field and what could hypothetically be my contribution to further development.

2. Attention Is All You Need (Vaswani et al., 2017)

This paper proposes a novel neural network architecture, the Transformer, based for the most part on the attention mechanisms, dispensing with recurrence and convolutions entirely (Vaswani et al., 2017). It describes well-thought-out Transformer architecture. Additionally, the authors have presented the results of the English-to-German, English-to-French translation task, and English constituency parsing task. The motivation for using only self-attention has three reasons: computational complexity per layer, amount of computation that can be parallelized, and path length between long-range dependencies in the network.

The previous works used Recurrent neural networks (Chung et al., 2014; Bahdanau et al., 2014) as model architecture and were state-of-the-art. However, these models divide the computation into sequential components. This sequential computation precludes parallelization and is sensitive to order and distance between words. On the other hand, this paper shows that the Transformer can model the dependencies regardless of the order and can be made parallelizable.

The Transformer architecture can be seen in Figure 2.1. It consists of encoder and decoder stacks. The encoder stack consists of layers. Each layer has two sublayers. The first sublayer is a multi-head self-attention mechanism, and the other is a feed-forward network. Each sublayer has a layer of normalization, so the output of a sublayer is LayerNorm(x + Sublayer(x)). The decoder stack is similar to the encoder, but it has three sublayers. The additional middle sublayer is the self-attention mechanism over the output of the encoder stack. The Transformer has auto-regression property. It means that during training, each position in the input sequence can only attend to positions before it, preventing information flow from future positions. To ensure that, these sublayer inputs of the self-attention are modified to make the self-attention equal to 0. Words that enter the model are turned into vectors with embeddings, and then the positional encodings are added to those vectors. The self-attention is defined as

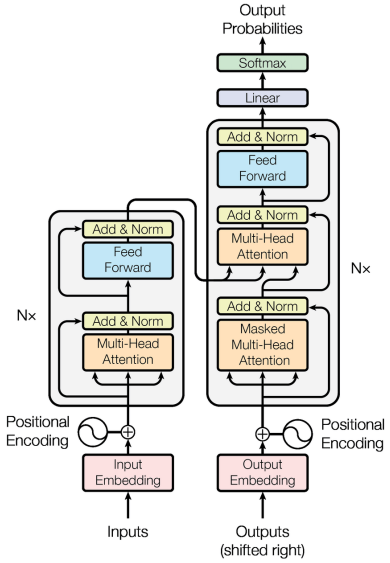

Figure 2.1: The Transformer - model architecture

$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where Q is the query matrix, K key matrix, V value matrix and d_k is the number of columns in matrix K. Multi-head attention differs in a way that these matrices are multiplied by projection matrices and the attention of each head is concatenated into one multi-head. Finally, to get the output, apply a linear and softmax layer.

The datasets used for translation tasks were the WMT 2014 English-German dataset and the WMT 2014 English-French dataset. English-German dataset consisted of about 4.5M sentence pairs while the English-French consisted of 36M sentence pairs. For English Constituency Parsing, the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1993) consisting of about 40K training sentences was used.

The results in the paper show that Transformer is state-of-the-art by achieving the highest BLEU score (Papineni et al., 2002) on both translation tasks and at minimal training cost. On the English constituency task, the Transformer was only beaten by Recurrent Neural Network Grammar (Kuncoro et al., 2016).

From my perspective, I think that the paper is well-structured and well-researched. Moreover, the authors released the code along with the paper, which contributes to the trustworthiness of their work. On the other hand, for a person who encounters an attention mechanism for the first time, hyperparameter choices such as positional encoding and the meaning of Q, K, and V matrices could be explained in more detail. In conclusion, the paper showed the crucial role of attention mechanisms but could have placed a bigger focus on simplifying the architectural choices and giving more intuitive explanations.

3. Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)

The paper uses the Transformer architecture and represents a continuation of the previous paper. This paper emphasizes natural language understanding, highlighting tasks like textual entailment, question answering, semantic similarity, and document classification. The idea is to first do *generative auto-regressive pre-training* on a diverse corpus of unlabeled data and then perform *discriminative fine-tuning* on each specific task with labeled data (Radford et al., 2018). The motivation for this approach is that plentiful unlabeled data is available but rarely labeled data that is expensive and time-consuming to obtain.

In the related work for semi-supervised learning for NLP, some approaches used unlabeled data to compute word-level statistics, which are then used as features in a supervised model (Liang, 2005). This paper focuses on obtaining higher-than-word-level semantics. Moreover, some similar works involve pre-training a neural network using language modeling objectives and then fine-tuning it on a target task with supervision. However, those works cannot capture long-range relationships or use hidden representations from a pre-trained model (Peters et al., 2017), which requires a significant amount of change to the model architecture depending on a task. Furthermore, this paper uses an auxiliary training objective which has been shown to enhance performance (Rei, 2017).

The authors of the paper use the Transformer decoder as the model to be able to generate text. For unsupervised pre-training with an unlabeled corpus of tokens $\mathcal{U} = \{u_1, ..., u_n\}$, the authors aim to maximise the likelihood:

$$L_1(\mathcal{U}) = \sum_i logP\left(u_i | u_{i-k}, ..., u_{i-1}; \Theta\right) \tag{3.1}$$

where k is the window size and Θ are the parameters of the neural network which

models the probability P. After unsupervised pre-training, the next step involves supervised fine-tuning. It assumes the availability of a labeled dataset $\mathcal{C} = x_1, ..., x_n$ with corresponding labels y, aiming to maximize:

$$L_2(\mathcal{C}) = \sum_{(x,y)} logP\left(x_i | x_{i-k}, ..., x_{i-1}\right). \tag{3.2}$$

Additionally, the axillary objective has been examined: $L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$.

The paper outlines a two-step approach, emphasizing that the model can be fine-tuned for textual entailment, similarity, and question answering with minimal input modification. Specifically for classification, the authors propose direct fine-tuning by adding only randomly initialized start and end tokens to the original input.

Unsupervised pre-training is done on the BooksCorpus dataset (Zhu et al., 2015) which contains over 7000 unique unpublished books. For each task, various datasets are used for which I am going to mention only a few:

Natural language inference: SNLI(Bowman et al., 2015), RTE(Bentivogli et al., 2009)

Question answering: RACE(Lai et al., 2017), Story Cloze(Mostafazadeh et al., 2017)

Sentence similarity: MRPC(Dolan and Brockett, 2005), QQP(Chen et al., 2018)

Classification: Stanford Sentiment Treebank-2(Socher et al., 2013), CoLA (Warstadt et al., 2018)

The results provided show that this approach outperforms all compared models on all datasets on question answering and commonsense reasoning. For classification and semantic similarity, a model exists that performs better only on CoLa and a model exists that performs better only on MRPC. It achieves new state-of-the-art results. The authors have also done ablation studies that are worth mentioning. It shows that the L_3 objective where L_1 is the auxiliary objective, is crucial for big datasets. Additionally, the authors observe a 5.6 average score drop when using the LSTM instead of the Transformer. Finally, when the model is trained directly without pre-training, the results are significantly worse, showing a decline of 14.8%.

From my perspective, the paper provided an effective way of transfer learning, allowing the model to learn general language representations before adapting to specific tasks. However, while the methods and objective functions are well described, there are some inconsistencies between the figure of input transformations for specific tasks and what is described in the text. Without the code and enough resources, it's hard to replicate the results reliably. The methods presented in this paper are well-supported by

experimental results, thereby establishing a new advancement. The paper introduces the first significant pre-trained decoder transformer.

4. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)

This paper proposes a new language model called BERT, which stands for Bidirectional Encoder Representations from Transformers. The novel model is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right context in all layers (Devlin et al., 2018). The motivation is that unidirectional language models can only attend previous tokens and thereby are limited in making better pre-trained representations and fine-tuning for contextual natural language understanding.

In the related work, feature-based approaches such as ELMo (Peters et al., 2018) need task-specific architecture and the representation is a linear combination of the hidden states from both the forward and backward LSTMs as context whereas BERT considers all positions in the sequence simultaneously capturing both left and right contexts during training. Moreover, fine-tuning approaches such as (Radford et al., 2018), benefit from not learning the representations from scratch.

As in the previous paper, BERT also uses a two-step approach of pretraining over unlabeled data and then fine-tuning for each specific task. This process is called transfer learning. BERT's architecture is a multi-layer bidirectional Transformer encoder based on the encoder part of the transformer model presented in the paper (Vaswani et al., 2017). The input could be one or more sentences but the BERT packs them all into one sequence unlike (Radford et al., 2018). The authors used WordPiece embeddings (Wu et al., 2016). The first token of every sequence is a special token [CLS]. It is used to aggregate sequence representation for classification tasks. To separate the sentences the authors used the [SEP] token. The input is the sum of 3 embeddings.

Learned positional embedding in a sequence, segment embeddings that annotate from which sentence they are and token embeddings.

Unlike the previous models, BERT is pre-trained using two unsupervised tasks. **Masked language modeling** (MLM) is masking the input tokens for the model to guess the masked tokens. The authors mask 15% of the token input. Of this, 80% are replaced with [MASK] token, 10% with a random token and 10% is unchanged to avoid a mismatch between pre-training and fine-tuning. **Next Sentence Prediction** (NSP) is used to enhance understanding of sentence relationships. BERT uses a binarized next sentence prediction task, where pairs of sentences are selected, and 50% of the time, the second sentence is the actual next sentence (IsNext), while 50% it's a random sentence (NotNext). Pre-training is done on the already mentioned bookcorpus (Zhu et al., 2015) and English Wikipedia. Fine-tuning BERT for tasks is simplified through a straightforward adaptation, achieved by swapping out task-specific inputs and outputs.

The datasets used for evaluation are various. Firstly, the GLUE benchmark covers all datasets used in (Radford et al., 2018). The results show an improvement on all tasks over OpenAI GPT, setting new state-of-the-art results. Moreover, a collection of question/answer pairs SQuAD 1.1(Rajpurkar et al., 2016) and SQuAD 2.0 were analyzed. In both datasets, BERT was again a new state-of-the-art system but on SQuAD 2.0, it did not outperform humans. Furthermore, the SWAG dataset (Zellers et al., 2018), a collection of selecting the most plausible continuation among four choices, was analyzed. BERT here also sets state-of-the-art results.

Additionally, the authors conducted ablation studies. They showed that removing NSP significantly hurts QNLI, MNLI, and SQUAD 1.1. Moreover, MLM without NSP performs better than the left-to-right context model without NSP.

From my perspective, the paper presented new and original research. The paper introduced new architecture, thus new pre-training unsupervised tasks and new datasets for analysis. However, while intriguing these novelties are, I struggled to understand how BERT processes tokens for evaluating some datasets. Furthermore, the intuition behind pre-training unsupervised tasks could be more clarified. Additionally, it would have been helpful if the paper compared BERT more directly to the OpenAI GPT model (Radford et al., 2018). This would give a clearer picture of how BERT improves upon the earlier model and what makes it stand out. Overall, BERT provides a step further in NLP with detailed results on various datasets. The paper was composed as a response to the (Radford et al., 2018) study, surpassing the OpenAI GPT model and thereby establishing new state-of-the-art results.

5. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks (Gururangan et al., 2020)

This paper analyses whether it is still helpful to adapt the model, after pre-training to a domain or target task. The motivation is to investigate whether the latest large pre-trained models have consistent performance across diverse domains, tasks and datasets. Specifically, the model that is used is RoBERTa (Liu et al., 2019).

The prior works have shown the benefits of continued pre-training in the domain (Lee et al., 2020), but in contrast to this paper, they used one domain at a time and a smaller and less diverse corpus. Additionally, previous works have also shown the advantage of continued pretraining through *task-adaptive pretraining* (TAPT) for improved end-task performance (Howard and Ruder, 2018; Phang et al., 2018). This paper additionally explores its comparison with *domain-adaptive pretraining* (DAPT), examining their interaction in transferability to other tasks and multi-phase pretraining strategies.

The authors focused on four domains which are centered around genres and forums. Domains are biomedical (BIOMED) papers, computer science (CS) papers, news text from REALNEWS, and AMAZON reviews. The tasks they focused on were two classification tasks from each domain. Before conducting the additional pre-training, the authors compared the similarity between the four mentioned domains and the original domain of RoBERTa. Consequently, showing that original RoBERTa's domain is most similar to News and Reviews, therefore, the most benefits could get BIOMED and CS.

After training on all domains together and additionally on each domain separately, the authors show that loss only increases in News. Furthermore, the F_1 score across all tasks rises by 2% on average except for one dataset from NEWS where it stays the same and for one dataset from CS where it has risen by 12.5%. The baseline model RoBERTa is not far behind. Additionally, when doing DAPT on the most dissimilar

domain the F_1 score drops from the baseline for all except the CS domain. This goes to show that better performance is not obtained by simply exposure to more data regardless of the domain. The authors pointed out another interesting point of view that since domain overlapping in the datasets leads to unexpected positive transfer, training on unconventional domains beyond conventional boundaries could yield better results for DAPT.

Furthermore, the paper shows that training on the task data, a narrowly defined subset of the domain that is more relevant to the task, is beneficial. TAPT raises F_1 score on all domain tasks. DAPT is more resource-intensive but TAPT manages to keep up. Finally, when applying DAPT on RoBERTa and then TAPT yields the best results on tasks. Another analysis for TAPT is to pre-train on one task and then fine-tune on the other task in the same domain. The paper shows a clear drop in performance across all tasks. It shows that data distributions within the domain of tasks differ and that adapting to a broad domain is not sufficient.

The last set of analyses deals with augmenting training data for TAPT for BIOMED and CS domains. It is the same principle as TAPT but on a larger dataset curated by humans or it is automated. DAPT + human curated-TAPT outperforms DAPT + TAPT and only human curated-TAPT on these tasks suggests that curating large amounts of data is beneficial. For automated curated-TAPT, the authors use VAMPIRE (Gururangan et al., 2019) bag-of-words language model. The goal is to find task-relevant data from the domain by embedding text from task and domain in a shared space and then finding the candidates from the domain by using queries and VAMPIRE from task data. The candidates were chosen by the *k*-NN algorithm or randomly. The paper shows that both types of candidate selection are better than TAPT but worse than DAPT.

This paper has had some creative and interesting analysis. The paper is well-written and I was fascinated by the combination of the usage of DAPT and TAPT as well as both human-curated and automated methods for expanding the data. On the other hand, the paper could use a more detailed discussion of the dataset's biases and representativeness. The choice of hyperparameters or the choice of specific architectures such as for VAMPIRE is not explored enough in my opinion. The analysis of the combination of DAPT and TAPT shows improved results but could be further investigated on the trade-offs between domain diversity and task relevance. In conclusion, the work provides valuable contributions to the understanding of continued pretraining strategies, domain adaptation, and task-specific adaptability.

6. TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning (Su et al., 2021)

This paper proposes a continual novel and effective pre-training approach called TaCL (Token-aware Contrastive Learning). TaCL pre-training is unsupervised and requires no additional data. The motivation for this is that pre-trained MLMs like BERT (Devlin et al., 2018) output an anisotropic distribution of token representations that live in a narrow subset of the entire representation space. This kind of token representation is not ideal for discriminative tasks.

Prior work showed the benefits of continual pre-training (Gururangan et al., 2020). Furthermore, related works used sentence-level contrastive learning with sentence-level embeddings (Giorgi et al., 2020) but have not been tested on improving general-purpose token representations and have no evidence that they will work on more complex datasets such as SQuAD (Rajpurkar et al., 2016).

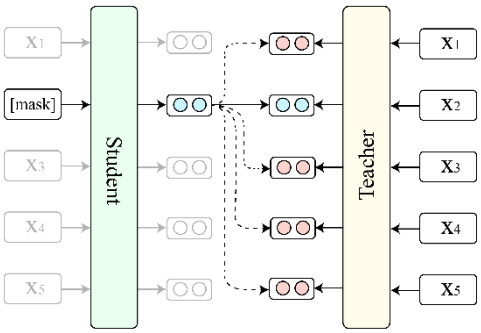

Figure 6.1: TaCL objective

The authors use two BERT models in their research. One is called the student and the other one is called the teacher and they are the same pre-trained BERT model. The concept of learning is to freeze the teacher and optimize the student model. The student model is optimized with the original BERT MLM and NSP objective (Devlin

et al., 2018) and a new TaCL objective. TaCL's objective, shown in Figure 6.1, is to contrast the masked token from the student to the unmasked same token from the teacher by getting closer to the teacher's unmasked token and getting further from the student's reference tokens in the same sequence. The ith token in the input sequence is represented by x_i, $\tilde{x_i}$ is the masked input token, h_i is the ith output token of the teacher and $\tilde{h}_i$ is the ith token output of the student. The TaCL with the described objective is then:

$$\mathcal{L}_{TaCL} = -\sum_{i=1}^{n} \mathbb{1}(\tilde{x}_i) log \frac{exp(sim(\tilde{h}_i, h_h))/\tau}{\sum_{j=1}^{n} exp(sim(\tilde{h}_i, h_j)/\tau)} \tag{6.1}$$

where $\mathbb{1}$ is the indicator function being 1 when $\tilde{x}_i$ is a masked token, the *sim* is cosine similarity and τ is a hyperparameter. Another measurement used in this paper is *self-similarity* which is the sum of all cosine similarities of each token to all other tokens in the model output. The lower the self-similarity the more discriminative the tokens are. Since this model operates on the improvement of token-level representations, it can be quite beneficial for token classification tasks.

The datasets used for the English benchmark are the same datasets used in (Devlin et al., 2018) to be comparable to the BERT model. For the Chinese benchmark, the authors evaluate the model in tasks on various datasets. Naming only a few, for named entity recognition (NER) (Weischedel et al., 2011) (Levow, 2006) and for Chinese word segmentation (CWS) (Emerson, 2005).

TaCL outperforms BERT on the majority of English datasets, showing improvements ranging from 0.5 to 2 in the F_1 score. However, its performance is slightly worse on QNLI and CoLA. For Chinese datasets, TaCL performs better than baseline BERT on all datasets having an improvement from 0.5 to 2. Ablation studies have also been made on the SQuAD 1.1 and SQuAD 2.0, showing that training on sentence-level contrastive objective as proposed by (Liu et al., 2021), is worse by 0.6 and 1. Moreover, training with only the TaCL objective shows that the model suffers a loss of 0.3 and 0.4. Furthermore, results of self-similarity on SQuAD show that by all layers, the self-similarity in baseline BERT is higher than training with TaCl objective except in the last output layer. In the last layer of TaCL, the self-similarity significantly drops which shows that the learned representations are more discriminative.

From my perspective, this paper provides a good view of the limitations of token representations from BERT. It was very intriguing to see that without any need for additional data, the model could be even more improved. The authors gave a clear explanation of methods, objectives, and what their goal was. However, the paper did not conduct hyperparameter tuning. I would have liked to see a deeper analysis of why

the self-similarity is lower only in the final layer. Additionally, the detailed analysis of self-similarity and ablation studies is done only on SQuAD, which could show only this approach's benefits. To sum up, TaCL has the potential to make token representations better, but it needs further exploration.

7. Conclusion

All the papers throughout this work were intriguing. The papers provided how rapidly the architecture and pre-training approaches have been evolving over the years. I was especially impressed by the variety and cleverness of the approaches, such as DAPT, TAPT, and token-aware contrastive learning. Moreover, I liked how token-aware contrastive learning was made without the need for additional data. From the appearance, the Transformers are the state-of-the-art model in the field of NLP. Today, the two-step approach of pre-training and fine-tuning is a typical learning paradigm. Further development could explore more advanced pre-training techniques regarding biases of the domain. Hypothetically, my contribution could involve developing some domain-specific pre-training methods, ensuring that models are performing various tasks within a specific domain. A step in this direction might involve experimenting with transfer learning approaches across related domains. To be able to contribute I would need a much deeper understanding of domain adaptation and transfer learning. Another contribution could be linked to model interpretability which could explain model decisions with higher granularity.

8. Bibliography

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1, 2009.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Z. Chen, H. Zhang, X. Zhang, and L. Zhao. Quora question pairs, 2018. URL `https://data.quora.com/First-QuoraDataset-Release-Question-Pairs`.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

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

Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. U *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

Thomas Emerson. The second international chinese word segmentation bakeoff. U *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, 2005.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*, 2019.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith. What do recurrent neural network grammars learn about syntax? *arXiv preprint arXiv:1611.05774*, 2016.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. U *Proceedings of the Fifth SIGHAN workshop on Chinese language processing*, stranice 108–117, 2006.

Percy Liang. *Semi-supervised learning for natural language*. Doktorska disertacija, Massachusetts Institute of Technology, 2005.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*, 2021.

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

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. U *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, stranice 46–51, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. U *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, stranice 311–318, 2002.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. U Marilyn Walker, Heng Ji, and Amanda Stent, urednici, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, stranice 2227–2237, New Orleans, Louisiana, Lipanj 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Marek Rei. Semi-supervised multitask learning for sequence labeling. *arXiv preprint arXiv:1704.07156*, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. U *Proceedings of the 2013 conference on empirical methods in natural language processing*, stranice 1631–1642, 2013.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv preprint arXiv:2111.04198*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Corpus of linguistic acceptability, 2018.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 2011.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. U *Proceedings of the IEEE international conference on computer vision*, stranice 19–27, 2015.