Attention Is (not) All You Need for Commonsense Reasoning

Tassilo Klein[1], **Moin Nabi**[1]
[1]SAP Machine Learning Research, Berlin, Germany
{tassilo.klein, m.nabi}@sap.com

Abstract

The recently introduced BERT model exhibits strong performance on several language understanding benchmarks. In this paper, we describe a simple re-implementation of BERT for commonsense reasoning. We show that the attentions produced by BERT can be directly utilized for tasks such as the *Pronoun Disambiguation Problem* and *Winograd Schema Challenge*. Our proposed attention-guided commonsense reasoning method is conceptually simple yet empirically powerful. Experimental analysis on multiple datasets demonstrates that our proposed system performs remarkably well on all cases while outperforming the previously reported state of the art by a margin. While results suggest that BERT seems to implicitly learn to establish complex relationships between entities, solving commonsense reasoning tasks might require more than unsupervised models learned from huge text corpora.

1 Introduction

Recently, neural models pre-trained on a language modeling task, such as ELMo (Peters et al., 2018b), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018), have achieved impressive results on various natural language processing tasks such as question-answering and natural language inference. The success of BERT can largely be associated to the notion of context-aware word embeddings, which differentiate it from common approaches such as word2vec (Mikolov et al., 2013) that establish a static semantic embedding. Since the introduction of BERT, the NLP community continues to be impressed by the amount of ideas produced on top of this powerful language representation model. However, despite its success, it remains unclear whether the representations produced by BERT can be utilized for tasks such as commonsense reasoning. Particularly, it is not clear whether BERT shed light on solving tasks such as the *Pronoun Disambiguation Problem* (PDP) and *Winograd Schema Challenge* (WSC). These tasks have been proposed as potential alternatives to the Turing Test, because they are formulated to be robust to statistics of word co-occurrence (Levesque et al., 2012).

Below is a popular example from the binary-choice pronoun coreference problem (Lee et al., 2017) of WSC:

Sentence: *The trophy doesn't fit in the suitcase because **it** is too small.*
Answers: A) the trophy ***B)*** the suitcase

Humans resolve the pronoun “it” to “the suitcase” with no difficulty, whereas a system without commonsense reasoning would be unable to distinguish “the suitcase” from the otherwise viable candidate, “the trophy”.

Previous attempts at solving WSC usually involve heavy utilization of annotated knowledge bases (KB), rule-based reasoning, or hand-crafted features (Peng et al., 2015; Bailey et al., 2015; Schüller, 2014; Sharma et al., 2015; Morgenstern et al., 2016). There are also some empirical works towards solving WSC making use of learning (Rahman and Ng, 2012; Tang et al., 2018; Radford et al., 2018). Recently, (Trinh and Le, 2018) proposed to use a language model (LM) to score the two sentences obtained when replacing the pronoun by the two candidates. The sentence that is assigned higher probability under the model designates the chosen candidate. Probability is calculated via the chain rule, as the product of the probabilities assigned to each word in the sentence. Very recently, (Emami et al., 2018) proposed the knowledge hunting method, which is a rule-based system that uses search engines

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4831–4836
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

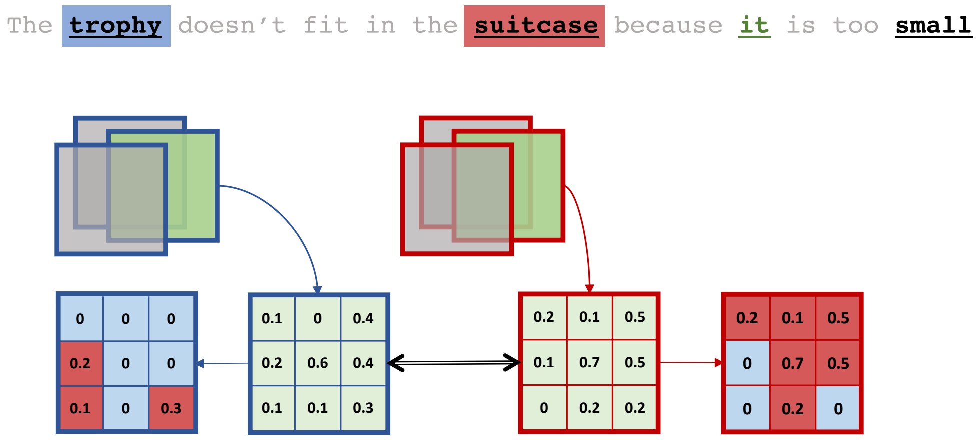


Figure 1: Maximum Attention Score (MAS) for a particular sentence, where colors show attention maps for different words (best shown in color). Squares with blue/red frames correspond to specific sliced attentions A_c for candidates c, establishing the relationship to the reference pronoun indicated with green. Attention is color-coded in blue/ red for candidates "trophy"/ "suitcase"; the associated pronoun "it" is indicated in green. Attention values are compared elementwise (black double arrow), and retain only the maximum achieved by a masking operation. Matrices on the outside with red background elements correspond to the masked attentions $A_c \circ M_c$.

to gather evidence for the candidate resolutions without relying on the entities themselves. Although these methods are interesting, they need fine-tuning, or explicit substitution or heuristic-based rules. See also (Trichelair et al., 2018) for a discussion.

The BERT model is based on the "Transformer" architecture (Vaswani et al., 2017), which relies purely on attention mechanisms, and does not have an explicit notion of word order beyond marking each word with its absolute-position embedding. This reliance on attention may lead one to expect decreased performance on commonsense reasoning tasks (Roemmele et al., 2011; Zellers et al., 2018) compared to RNN (LSTM) models (Hochreiter and Schmidhuber, 1997) that do model word order directly, and explicitly track states across the sentence. However, the work of (Peters et al., 2018a) suggests that bidirectional language models such as BERT implicitly capture some notion of coreference resolution.

In this paper, we show that the attention maps created by an out-of-the-box BERT can be directly exploited to resolve coreferences in long sentences. As such, they can be simply repurposed for the sake of commonsense reasoning tasks while achieving state-of-the-art results on the multiple task. On both PDP and WSC, our method outperforms previous state-of-the-art methods, without using expensive annotated knowledge bases or hand-engineered features. On a Pronoun Disambiguation dataset, PDP-60, our method achieves 68.3% accuracy, which is better than the state-of-art accuracy of 66.7%. On a WSC dataset, WSC-273, our method achieves 60.3%. As of today, state-of-the-art accuracy on the WSC-273 for single model performance is around 57%, (Emami et al., 2018) and (Trinh and Le, 2018). These results suggest that BERT implicitly learns to establish complex relationships between entities such as coreference resolution. Although this helps in commonsense reasoning, solving this task requires more than employing a language model learned from large text corpora.

2 Attention Guided Reasoning

In this section we first review the main aspects of the BERT approach, which are important to understand our proposal and we introduce notations used in the rest of the paper. Then, we introduce Maximum Attention Score (MAS), and explain how it can be utilized for commonsense reasoning.

2.1 BERT and Notation

The concept of BERT is built upon two key ingredients: (a) the transformer architecture and (b) unsupervised pre-training.

The transformer architecture consists of two main building blocks, stacked encoders and de-

Method	Acc.
Unsupervised Semantic Similarity Method (USSM)	48.3 %
USSM + Cause-Effect Knowledge Base (Liu et al., 2016)	55.0 %
USSM + Cause-Effect + WordNet (Miller, 1995) + ConceptNet (Liu and Singh, 2004) KB	56.7 %
Subword-level Transformer LM (Vaswani et al., 2017)	58.3 %
Single LM (partial) (Trinh and Le, 2018)	53.3 %
Single LM (full) (Trinh and Le, 2018)	60.0 %
Patric Dhondt (WS Challenge 2016)	45.0 %
Nicos Issak (WS Challenge 2016)	48.3 %
Quan Liu (WS Challenge 2016 - **winner**)	58.3 %
USSM + Supervised DeepNet	53.3 %
USSM + Supervised DeepNet + 3 KBs	66.7 %
Our Proposed Method	**68.3 %**

Table 1: Pronoun Disambiguation Problem: Results on (top) Unsupervised method performance on PDP-60 and (bottom) Supervised method performance on PDP-60. Results other than ours are taken from (Trinh and Le, 2018).

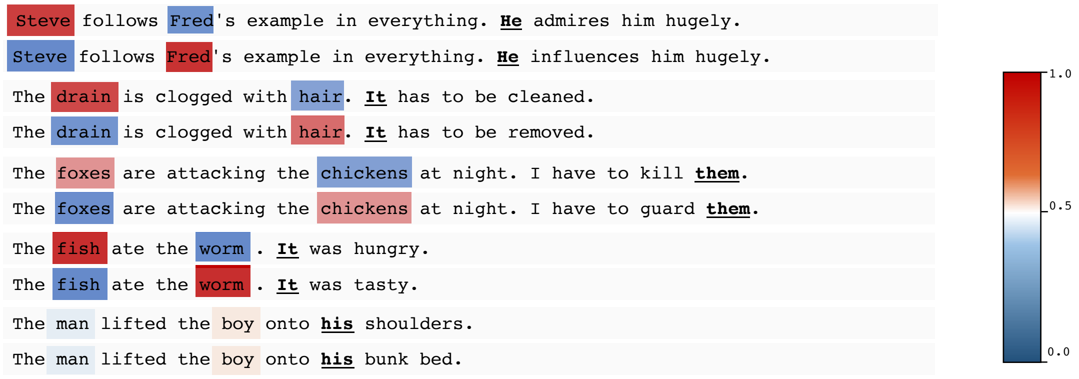
Method	Acc.
Random guess	50.0 %
USSM + KB	52.0%
USSM + Supervised DeepNet + KB	52.8 %
Single LM (Trinh and Le, 2018)	54.5 %
Transformer (Vaswani et al., 2017)	54.1 %
Know. Hunter (Emami et al., 2018)	57.1 %
Our Proposed Method	**60.3 %**

Table 2: Winograd Schema Challenge. The other results are taken from (Trichelair et al., 2018) and (Trinh and Le, 2018).

coders, which are connected in a cascaded fashion. The encoder is further divided into two components, namely a self-attention layer and a feed-forward neural network. The self-attention allows for attending to specific words during encoding and therefore establishing a focus context w.r.t. to each word. In contrast to that, the decoder has an additional encoder-decoder layer that switches between self-attention and a feed-forward network. It allows the decoder to attend to specific parts of the input sequence. As attention allows for establishing a relationship between words, it is very important for tasks such as coreference resolution and finding associations. In the specific context of pronouns, attention gives rise to links to m candidate nouns, which we denote in the following as $\mathcal{C} = \{c_1, .., c_m\}$. The concept of self-attention is further expanded within BERT by the idea of so called multi-head outputs that are incorporated in each layer. In the following, we will denote heads and layers with $h \in H$ and $l \in L$, respectively. Multi-heads serve several purposes. On the one hand, they allow for dispersing the focus on multiple positions. On the other hand, they constitute an enriched representation by expanding the embedding space. Leveraging the nearly unlimited amount of data available, BERT learns two novel unsupervised prediction tasks during training. One of the tasks is to predict tokens that were randomly masked given the context, notably with the context being established in a bi-directional manner. The second task constitutes next sentence prediction, whereby BERT learns the relationship between two sentences, and classifies whether they are consecutive.

2.2 Maximum Attention Score (MAS)

In order to exploit the associative leverage of self-attention, the computation of MAS follows the notion of max-pooling on attention level between a reference word s (e.g. pronoun) and candidate words c (e.g. multiple choice pronouns). The proposed approach takes as input the BERT attention tensor and produces for each candidate word a score, which indicates the strength of association. To this end, the BERT attention tensor $A \in \mathbb{R}^{H \times L \times |\mathcal{C}|}$ is sliced into several matrices $A_c \in \mathbb{R}^{H \times L}$, each of them corresponding to the attention between the reference word and a candidate c. Each A_c is associated with a binary mask matrix M_c. The mask values of M_c are obtained



Steve follows Fred's example in everything. He admires him hugely.

Steve follows Fred's example in everything. He influences him hugely.

The drain is clogged with hair. It has to be cleaned.

The drain is clogged with hair. It has to be removed.

The foxes are attacking the chickens at night. I have to kill them.

The foxes are attacking the chickens at night. I have to guard them.

The fish ate the worm . It was hungry.

The fish ate the worm . It was tasty.

The man lifted the boy onto his shoulders.

The man lifted the boy onto his bunk bed.

Figure 2: Maximum Attention Score (MAS) for some sample questions from WSC-273: The last example is an example of failure of the method, where the coreference is predicted incorrectly.

at each location tuple (l, h), according to:

$$M_c(l,h) = \begin{cases} 1 & \operatorname{argmax} A(l,h) = c \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Mask entries are non-zero only at locations where the candidate word c is associated with maximum attention. Limiting the impact of attention by masking allows to accommodate for the most salient parts. Given the A_c and M_c matrix pair for each candidate c, the MAS can be computed. For this purpose, the sum of the Hadamard product for each pair is calculated first. Next, the actual score is obtained by computing the ratio of each Hadamard sum w.r.t. all others according to,

$$MAS(c) = \frac{\sum_{l,h} A_c \circ M_c}{\sum_{c \in \mathcal{C}} \sum_{l,h} A_c \circ M_c} \in [0,1]\,. \tag{2}$$

Thus MAS retains the attention of each candidate only where it is most dominant, coupling it with the notion of frequency of occurrence to weight the importance. See Fig. 1 for a schematic illustration of the computation of MAS, and the matrices involved.

3 Experimental Results

We evaluate our method on two commonsense reasoning tasks, PDP and WSC.

On the former task, we use the original set of 60 questions (PDP-60) as the main benchmark. The second task (WSC-273) is qualitatively much more difficult. The recent best reported result are not much above random guess. This task consists of 273 questions and is designed to work against traditional linguistic techniques, common heuristics or simple statistical tests over text corpora (Levesque et al., 2012).

3.1 BERT Model Details

In all our experiments, we used the out-of-the-box BERT models without any task-specific fine-tuning. Specifically, we use the PyTorch implementation of pre-trained $bert-base-uncased$ models supplied by Google[1]. This model has 12 layers (i.e., Transformer blocks), a hidden size of 768, and 12 self-attention heads. In all cases we set the feed-forward/filter size to be 3072 for the hidden size of 768. The total number of parameters of the model is 110M.

3.2 Pronoun Disambiguation Problem

We first examine our method on PDP-60 for the Pronoun Disambiguation task. In Tab. 1 (top), our method outperforms all previous unsupervised results sharply. Next, we allow other systems to take in necessary components to maximize their test performance. This includes making use of supervised training data that maps commonsense reasoning questions to their correct answer. As reported in Tab. 1 (bottom), our method outperforms the best system in the 2016 competition (58.3%) by a large margin. Specifically, we achieve 68.3% accuracy, better than the more recently reported results from (Liu et al., 2017) (66.7%), who makes use of three KBs and a supervised deep network.

3.3 Winograd Schema Challenge

On the harder task WSC-273, our method also outperforms the current state-of-the-art, as shown in Tab. 2. Namely, our method achieves an accuracy of 60.3%, nearly 3% of accuracy above the

[1] https://github.com/huggingface/pytorch-pretrained-BERT

previous best result. This is a drastic improvement considering the best system based on language models outperforms random guess by only 4% in accuracy. This task is more difficult than PDP-60. First, the overall performance of all competing systems are much lower than that of PDP-60. Second, incorporating supervised learning and expensive annotated KBs to USSM provides insignificant gain this time (+3%), comparing to the large gain on PDP-60 (+19%). Finally, for the sake of completeness, (Trinh and Le, 2018) report that their single language model trained on a customized dataset built from CommonCrawl based on questions used in comonsense reasoning achieves an higher accuracy than the proposed approach with 62.6%.

We visualize the MAS to have more insights into the decisions of our resolvers. Fig. 2 displays some samples of correct and incorrect decisions made by our proposed method. MAS score of different words are indicated with colors, where the gradient from *blue* to *red* represents the score transition from low to high.

4 Discussion

Pursuing commonsense reasoning in a purely unsupervised way seems very attractive for several reasons. On the one hand, this implies tapping the nearly unlimited resources of unannotated text and leveraging the wealth of information therein. On the other hand, tackling the commonsense reasoning objective in a (more) supervised fashion typically seems to boost performance for very a specific task as concurrent work shows (Kocijan et al., 2019). However, the latter approach is unlikely to generalize well beyond this task. That is because covering the complete set of commonsense entities is at best extremely hard to achieve, if possible at all. The data-driven paradigm entails that the derived model can only make generalizations based on the data it has observed. Consequently, a supervised machine learning approach will have to be exposed to all combinations, i.e. replacing lexical items with semantically similar items in order to derive various concept notions. Generally, this is prohibitively expensive and therefore not viable. In contrast, in the proposed (unsupervised self-attention guided) approach this problem is alleviated. This can be largely attributed to the nearly unlimited text corpora on which the model originally learns, which makes it likely to cover a multitude of concept relations, and the fact that attention implicitly reduces the search space. However, all these approaches require the answer to explicitly exist in the text. That is, they are unable to resolve pronouns in light of abstract/implicit referrals that require background knowledge - see (Saba, 2018) for more detail. However, this is beyond the task of WSC. Last, the presented results suggest that BERT models the notion of complex relationship between entities, facilitating commonsense reasoning to a certain degree.

5 Conclusion

Attracted by the success of recently proposed language representation model BERT, in this paper, we introduce a simple yet effective re-implementation of BERT for commonsense reasoning. Specifically, we propose a method which exploits the attentions produced by BERT for the challenging tasks of *PDP* and *WSC*. The experimental analysis demonstrates that our proposed system outperforms the previous state of the art on multiple datasets. However, although BERT seems to implicitly establish complex relationships between entities facilitating tasks such as coreference resolution, the results also suggest that solving commonsense reasoning tasks might require more than leveraging a language model trained on huge text corpora. Future work will entail adaption of the attentions, to further improve the performance.

References

Daniel Bailey, Amelia J Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. The winograd schema challenge and reasoning about correlation. In *2015 AAAI Spring Symposium Series*.

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

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1958, Brussels, Belgium. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997.

Long short-term memory. *Neural computation*, 9(8):1735–1780.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019*. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. Combing context and commonsense knowledge through neural networks for solving winograd schema problems. In *2017 AAAI Spring Symposium Series*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Leora Morgenstern, Ernest Davis, and Charles L Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1):50–54.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018a. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Walid S. Saba. 2018. A simple machine learning method for commonsense reasoning? A short commentary on trinh & le (2018). *CoRR*, abs/1810.00521.

Peter Schüller. 2014. Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. 2015. Towards addressing the winograd schema challengebuilding and using a semantic parser and a knowledge hunting module. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*.

Paul Trichelair, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz. 2018. On the evaluation of common-sense reasoning in natural language understanding. *arXiv preprint arXiv:1811.01778*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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