Pre-training Is (Almost) All You Need: 
An Application to Commonsense Reasoning

Alexandre Tamborrino∗, Nicola Pellicanò∗, Baptiste Pannier∗, Pascal Voitot, Louise Naudin
Samsung Strategy and Innovation Center
{a.tamborrino,n.pellicano,b.pannier,p.voitot,l.naudin}@samsung.com

Abstract
Fine-tuning of pre-trained transformer models has become the standard approach for solving common NLP tasks (Devlin et al., 2019). Most of the existing approaches rely on a randomly initialized classifier on top of such networks. We argue that this fine-tuning procedure is sub-optimal as the pre-trained model has no prior on the specific classifier labels, while it might have already learned an intrinsic textual representation of the task. In this paper, we introduce a new scoring method that casts a plausibility ranking task in a full-text format and leverages the masked language modeling head tuned during the pre-training phase. We study commonsense reasoning tasks where the model must rank a set of hypotheses given a premise, focusing on the COPA (Gordon et al., 2012), Swag (Zellers et al., 2018), HellaSwag (Zellers et al., 2019) and CommonsenseQA (Talmor et al., 2019) datasets. By exploiting our scoring method without fine-tuning, we are able to produce strong baselines (e.g. 80% test accuracy on COPA) that are comparable to supervised approaches. Moreover, when fine-tuning directly on the proposed scoring function, we show that our method provides a much more stable training phase across random restarts (e.g. ×10 standard deviation reduction on COPA test accuracy) and requires less annotated data than the standard classifier approach to reach equivalent performances.

1 Introduction
Recent advances in natural language processing have been made using sequential transfer learning over large pre-trained transformer models. From these models, most NLP tasks can be addressed by adding a classifier on top of the transformer embedding outputs (Devlin et al., 2019; Liu et al., 2019).

∗Equal contribution.
tuned during the pre-training phase (see Figure 1 for an overview of the proposed approach). This method produces strong zero-shot baselines on the COPA (Gordon et al., 2012), Swag (Zellers et al., 2018), HellaSwag (Zellers et al., 2019) and CommonsenseQA (Talmor et al., 2019) datasets. Then, we fine-tune this new scoring function with a margin-based loss as proposed in (Li et al., 2019). Using RoBERTa LARGE, our results reveal that this new training procedure leads to better accuracy and much more stable training trajectories which is an important feature since large MLM models are known to be unstable on several tasks (Devlin et al., 2019; Phang et al., 2018). Finally, we find that a progressive decrease of the training dataset size results in a progressive increase of the accuracy gap between our proposed method and the standard classifier ones. This makes our method advantageous in small dataset context.

2 Related Work

In (Trinh and Le, 2018), researchers have shown that a RNN Language Model pretrained on a large amount of data can be used to efficiently score sentences in a zero-shot setting. They used the Winograd Schema Challenge (WSC-273) dataset (Levesque et al., 2012) which mostly consists of a pronoun disambiguation task that requires commonsense reasoning. In their approach, the pronoun to disambiguate is replaced by the different candidates. Then, each version of the sentence is scored using the likelihood of the sequence under the forward autoregressive factorization. They showed that targeting the likelihood of the tokens placed after the candidate words performs better than a full-sentence likelihood estimation. This result highlights the fact that the choice of the targeted sub-sequence for the likelihood estimation has an important impact on the overall performance of the model. More recently, analysis of relational knowledge contained in pre-trained BERT models has been the subject of different studies (Petroni et al., 2019; Poerner et al., 2019). Results have shown evidences that BERT models memorize reasoning about entity names and commonsense knowledge, making MLM models appropriate candidates to commonsense oriented tasks.

From a supervised learning perspective, (Li et al., 2019) proposed to replace the traditional cross-entropy loss with a margin-based one on the COPA dataset. The authors argued that cross-entropy based methods are not adapted for plausibility ranking tasks since they force the scores to adopt extreme values (near 0 or 1). In contrast, a margin-based objective function appeared to be a natural way to rank a set of hypotheses. Both approaches were compared using the [CLS] token of the BERT-base model and a separated-sentence input format. The margin-based objective function surpassed the cross-entropy one by increasing the Test set accuracy from 73.4% to 75.4%.

Adopting a token level scoring approach (Kocijan et al., 2019) used a BERT model with a mixture between a margin-based and a MLM loss on WSC-273 to score the different pronouns to disambiguate. This approach allows the authors to improve the previous state of the art by 8.8%. Despite being the closest method to the one proposed in this paper, our approach differs from three points:

- We generalize the scoring method by targeting different contiguous sub-sequences for the likelihood estimation. To do so, different datasets are recasted in a full-text format.
- We also focus on targeting the premise avoiding inner statistical biases of different hypotheses (e.g. word frequencies, punctuation, variable sequence lengths etc...).
- The objective of the present paper is to propose a direct comparison in terms of accuracy and training stability across random restarts between the proposed method and standard classifiers.

3 Method

3.1 Problem Formulation

Given an input premise \( p = (p^{(1)}, p^{(2)}, \ldots, p^{(L_p)}) \), and a set of candidate hypotheses:

\[
H = \{ h_i = (h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(L_i)}) \}_{i=1 \ldots n}
\]

we aim to identify the fitting hypothesis \( h^* \in H \) which correctly matches \( p \). The values \( L_p \) and \( \{L_i\}_{i=1 \ldots n} \) are the sequence lengths of premise and hypotheses respectively. In a commonsense settings, such problem corresponds to find premise-hypothesis implications by exploiting some prior commonsense knowledge. Since

---

1For the following of our paper, we will note as zero-shot setting the use of the pre-trained model without fine-tuning.
our scoring method consumes input sequences in a full-text format (see Section 3.2), our method is formulated on a commonsense task but not limited to it.

3.2 Sequence Scoring Method

The proposed Sequence Scoring Method (SSM), takes as input a pair \( \langle p, h_i \rangle \) returns a score representing the likelihood of \( h_i \) of being implied by \( p \).

First, a transform operator \( T \) converts \( \langle p, h_i \rangle \) pair into a full-text input. Such operator, in its simplest form, just concatenates the two sequences. However, in general \( T \) can be constrained on the task \( t \).

\[
s_i = T^t(p, h_i) = (c^l_i, p, c^m_i, h_i, c^r_i),
\]

where \( s_i \) is the resulting full-text input, while \( c^l_i, c^m_i, \) and \( c^r_i \) are left, middle and right conjunction sequences of the task. For example, Swag will have no conjunction, since the correct hypothesis is the natural continuation of the premise, while COPA will have because/so middle conjunctions due to its cause/effect nature (see Section 4).

Given the full-text input, the scorer aims to exploit the pre-training task of word masking in order to compute its result. Let us consider the masking of a word \( w \) which contributes to make sense of the matching between \( p \) and \( h_i \). The intuition is that the confidence of the network in recovering such word is directly related to the score of \( \langle p, h_i \rangle \). Let us define, inspired by the notation of (Song et al., 2019), \( s_i \wedge w \) as the sentence \( s_i \) with the tokens of \( w \) replaced by the \([\text{MASK}]\) token.

The target premise score is calculated as follows:

\[
S^p_i = \frac{1}{L_p} \sum_{k=1}^{L_p} \log P \left( p^{(k)} \mid s_i \wedge p^{(k)} \right),
\]

where premise words are masked one by one in order to compute their relevance with respect to the given hypothesis. Masked word probability is estimated from direct inference on a model pre-trained on MLM task. The computational complexity of such method grows linearly with \( L_p \) (requiring \( L_p \) examples per forward pass). Alternatively, the target hypothesis score is computed as:

\[
S^h_i = \frac{1}{L_i} \sum_{k=1}^{L_i} \log P \left( h_i^{(k)} \mid s_i \wedge h_i^{(k)} \right).
\]
The target hypothesis score needs normalization by $L_i$ in order to allow comparison between variable candidate hypothesis length. The best hypothesis will be taken as the one maximizing the target premise (or hypothesis) score:

$$h^* = h_j \in H \text{ s.t. } \max_{i=1,...,n} S^p_i = S^p_j. \quad (4)$$

As demonstrated in Section 5.2, the target premise score allows for a fairer comparison between different hypotheses. In fact, they present inherent differences in terms of statistical frequency of words, sequence length or may exhibit more or less strong inter-dependency between words (e.g. composite words reinforce each other confidence). Such variance could introduce a bias in the relative significance of each hypothesis alone (independently from the premise). On the opposite, different probabilities on the same target premise word can only be affected by the change of hypothesis context.

**N-grams sequence scoring**

We can extend the proposed SSM by scoring the reconstruction not only of single words, but of entire n-grams. Adding n-grams probabilities to the joint probability of (dependent) close words, especially in a zero-shot setting. Let us note as $p^{(uv)}$ the sub-sequence of $p$, spanning between indexes $u$ and $v$ (included). The partial target premise score for g-grams (i.e. mask windows of size $g$) can be expressed as:

$$S^{p,g}_i = \sum_{k=1}^{L_p-g+1} \log \left( P \left( p^{(k:k+g-1)} \mid s_i^{p^{(k:k+g-1)}} \right) \right).$$

By definition the target premise score in Equation 2 is equivalent to 1-gram partial target premise score (i.e. $S^p_i \equiv S^{p,1}_i$). The n-gram sequence scoring accumulates masked language model probabilities from every gram size till $n$.

$$S^{p,[1,n]}_i = \sum_{g=1}^{n} S^{p,g}_i. \quad (5)$$

### 3.3 SSM-based fine-tuning

The proposed score function, since it does not imply any addition of a head module, can be directly applied without any retraining (see Section 5.2). It can also be directly used when fine-tuning on the task. The different masked inputs needed to compute the target premise score, $\left\{ s_i^{p^{(j)}} \right\}_{j=1,...,L_p}$, are batched together in order to compute score $S^p_j$ in one forward pass. The model acts as a siamese network that performs independent computation of target premise score for each hypothesis $h_i$.

**Loss function**

As already noted in (Li et al., 2019), multiple choice tasks (e.g. COPA) are more naturally expressed as learning to rank problems. For this reason we adopt as objective function a margin-based loss in contrast to cross-entropy loss. Given ground truth sentence index $i^*$, the loss is specified as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \max(0, \eta - S^{p,i^*}_i + S^{p,i}_i), \quad (6)$$

where $\eta$ is a margin threshold hyperparameter.

According to our preliminary experiments, we do not add a second MLM component in the general loss (as in (Kocijan et al., 2019)), since it always leads to a decrease of the model performance for various weighted contributions of the MLM term.

### 4 Datasets

The commonsense reasoning datasets that we focus on are COPA (Gordon et al., 2012), Swag (Zellers et al., 2018), HellaSwag (Zellers et al., 2019) and CommonsenseQA (Talmor et al., 2019). All these datasets share the premise-hypothesis task format. Table 1 shows examples of full-text format and separated-sentence format for all datasets.

**COPA**

COPA (Choice of Plausible Alternatives) (Gordon et al., 2012) is a commonsense causal reasoning task where two candidate hypotheses are given. COPA itself is composed of two sub-tasks: effect samples and cause samples. The effect and cause samples have respectively implies and implied by relation with the correct hypothesis. The full-text format of COPA is built by using the conjunction words because (resp. so) as middle conjunctions for cause (resp. effect) samples. Concerning the separated-sentence format, we reverse the premise and hypothesis order for cause samples in
| Dataset        | Full-text format                                                                 | Separated-sentence format                                                                 |
|---------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| COPA (effect) | [CLS] I knocked on my neighbor’s door so my neighbor invited me in. [SEP]      | [CLS] I knocked on my neighbor’s door. My neighbor invited me in. [SEP]                  |
| COPA (cause)  | [CLS] The man broke his toe because he dropped a hammer on his foot. [SEP]      | [CLS] He dropped a hammer on his foot. The man broke his toe. [SEP]                     |
| CommonsenseQA | [CLS] Q: Where on a river can you hold a cup upright to catch water on a sunny day? [SEP] | [CLS] Q: Where on a river can you hold a cup upright to catch water on a sunny day? [SEP] |
| Swag          | [CLS] We notice a man in a kayak and a yellow helmet coming in from the left. He approaches, his kayak flips upside-down. [SEP] | [CLS] We notice a man in a kayak and a yellow helmet coming in from the left. [SEP] As he approaches, his kayak flips upside-down. [SEP] |
| HellaSwag     | [CLS] A man is standing in front of a camera. He starts playing a harmonica for the camera. He rocks back and forth to the music as he goes. [SEP] | [CLS] A man is standing in front of a camera. He starts playing a harmonica for the camera. He rocks back and forth to the music as he goes. [SEP] |

Table 1: Examples of full-text format and separated-sentence format for gold premise-hypothesis pairs. Left conjunction $c_1$ is highlighted in italic blue, middle conjunction $c_m$ in bold red.

order to convert all cause samples into effect samples. This has the benefit to present a unique task to the model, and our experiments show that this give better results than keeping cause samples and effect samples unmodified. We choose the SuperGLUE split (Wang et al., 2019).

CommonsenseQA

CommonsenseQA (Talmor et al., 2019) is a multiple-choice commonsense question answering dataset where each question has one correct answer and four distractor answers. To create the full-text format, we prepend $Q:$ to the question, $A:$ to the answer, and then concatenate the question and the answer ($\_$ stands for space character). For the separated-sentence format, we also use the $Q:$ and $A:$ prefixes to follow the best recommendation from the FairSeq repo on how to fine-tune RoBERTa on CommonsenseQA. Since the benchmark Test set is private, we evaluate our zero-shot setting on the Val set (we do not perform a fine-tuning study on Swag and HellaSwag as explained in Section 5.3).

5 Experiments

In this section we first apply our scoring method in a zero-shot setting on the four aforementioned datasets. Then we fine-tune our scoring method while varying the percentage of the training data used and compare it to approaches that use a randomly initialized classifier head. We use RoBERTa$\text{LARGE}$ (Liu et al., 2019) for our pretrained model as RoBERTa$\text{LARGE}$ fine-tuned with a classification layer on top has very competitive results on those datasets. Our implementation use PyTorch and the HuggingFace Transformers library (Wolf et al., 2019).

5.1 Task probing

Before assessing our zero-shot and fine-tuning results, we perform a task probing by evaluating the zero-shot score we obtain by removing the premise from the input and only scoring the hypotheses. If the score is significantly better than a random baseline, it means that the task is not actually solved by commonsense reasoning, but by using statistical biases in the hypotheses. This probing method has been already used on several dataset.

Swag and HellaSwag

Swag (Situations With Adversarial Generations) (Zellers et al., 2018) is a multiple choice commonsense dataset about grounded situations. Each premise is a video caption with four answer choices about what might happen next in the scene. The correct answer is the video caption for the next event in the video. The other negative answers are created via Adversarial Filtering: generated by language modeling models and filtered by discriminator models. HellaSwag (Zellers et al., 2019) is an evolved version of Swag using better generators and discriminators models for Adversarial Filtering. Since the benchmark test set is private, we evaluate our zero-shot setting on the Val set (we do not perform a fine-tuning study on Swag and HellaSwag as explained in Section 5.3).
The results of the task probing evaluation are reported in Table 2. While COPA and CommonsenseQA have a hypothesis only score close to the random baseline, the score of both Swag and HellaSwag are significantly higher than their random baseline (more than twice). This confirms the study from (Zellers et al., 2019) that shows that Swag’s false hypotheses were generated using a weak generator, therefore the authors argue that the fine-tuning process on a BERT model on Swag learns to pick up the statistical cues left by the weak generator. Our results show that RoBERTa\textsubscript{LARGE} can leverage these distributional biases without the fine-tuning phase. We argue that the human-written pre-training corpora of RoBERTa biases it to give better score to human-written language rather than model-generated sentences. As shown in (Holtzman et al., 2019), there is indeed still a strong distributional differences between human text and machine text. Furthermore, our result also highlights that HellaSwag still exhibits a strong bias due to its generation scheme when evaluated with RoBERTa\textsubscript{LARGE}.  

### 5.2 Zero-shot Results

For both COPA and CommonsenseQA, the best performing scoring method uses the target premise and 4-grams settings as shown in Tables 3 and 4. Targeting the premise gives better results than targeting the hypothesis, which reinforces our argument that targeting the hypothesis may be harder as the differences between the hypotheses make the score comparison noisier. Also, more grams give increasingly better results but the trend inverts after 4-grams, which may be due to the fact that masked models are not trained to mask large chunks of text. It is interesting to note that our zero-shot result is significantly better than a BERT\textsubscript{LARGE} cross-entropy model fined-tuned on the COPA training set (80.0% vs. 70.6% accuracy) (Wang et al., 2019), while being comparable for CommonsenseQA. Moreover, when we intentionally switch the so and because conjunction words on COPA to make the samples erroneous, the accuracy drops significantly (64.4%). We reckon this is an indicator that our scoring method effectively reuse the pre-learned representation the full-text format of the task.

Concerning Swag and HellaSwag, the target hypothesis mode is significantly better than the target premise mode (see Table 5), as expected from our task probing work in Section 5.1. For example, on HellaSwag, the target hypothesis mode is only 8% better than the hypothesis only mode (58.8% versus 50.8%), which confirms that on this setting our zero-shot method is mainly taking advantage of the bias in the hypotheses. Therefore we refrain from doing more zero-shot experiments on both datasets.

### 5.3 Fine-tuning Results

Following the strong bias of Swag and HellaSwag that was shown in Section 5.1 using our scoring method with RoBERTa\textsubscript{LARGE}, we decide to not...
Figure 2: COPA fine-tuning results on Test set. The whole training set corresponds to 400 examples.

Figure 3: CommonsenseQA fine-tuning results on Test* set. The whole training set corresponds to 9741 examples.

Dataset | Target | Val Acc (%)  
--- | --- | ---  
Swag  | premise | 48.3  
Swag  | hypothesis | **72.5**  
HellaSwag | premise | 37.1  
HellaSwag | hypothesis | **58.8**  

Table 5: Swag/HellaSwag zero-shot results (1-Gram).

include them into our fine-tuning study to be sure to compare results for which models learn the actual premise-hypothesis commonsense reasoning task.

Comparison settings

In order to make fair comparisons, we train and compare three different model settings:

- Our scoring method with target premise mode, 1-gram, margin-based loss, full-text format (ours).
- A randomly initialized classifier with cross-entropy loss and separated-sentence format (head CE). The cross-entropy loss is computed on the probability of the correct candidate, normalized over all candidates in the set (see Equation 1 in (Li et al., 2019)).
- A randomly initialized classifier with margin-based loss and full-text format (head margin).

The head margin setting is an ablated version of our scoring method to verify that our reuse of the MLM head actually provides a significant advantage over a randomly initialized head. For our method, we report results only for the best performing scoring method which is the target premise mode. Experiments showed us that varying the number of grams produce comparable re-
results, so we use the 1-gram setting for computational efficiency. We reckon that the enriched bidirectional context granted by N-gram score can be directly learned when fine-tuning on the task.

For each dataset, we train the three model settings for 20 random seeds each. For each seed, we pick the best performing model on the validation set and report its accuracy on the Test set. We then compute the max accuracy, mean accuracy and standard deviation of each model setting on the Test set. For all model settings, following the recommended hyper-parameters to fine-tune RoBERTa\textsubscript{LARGE} (Liu et al., 2019), we set a learning rate of 1e-5, a warm-up ratio of 6% of the total number of training steps, a linear learning rate decay and a weight decay of 0.01. We use a batch size of 8 for COPA (4 for the 10% training percentage setting) and 16 for CommonsenseQA. For the margin-based loss (ours and head margin), we set $\eta = 0.5$ after a few trials.

**COPA and CommonsenseQA results**

On both COPA and CommonsenseQA, our method outperforms both the head CE and head margin methods in terms of mean accuracy and max/best accuracy (see Figure 2 and Figure 3). Moreover, we find that a progressive decrease of the training dataset size results in a progressive increase of the best accuracy gap between our method and the other ones. This confirms our intuition that our methods is the most advantageous when few training data is available.

For example, when using 1% of training data of CommonsenseQA, our method achieves an accuracy of 56.7% on the Test set (vs. 40.2% for the head CE approach). Using the whole training data, our approach still outperforms other methods but by a lower margin (76.4% accuracy versus 75.4% for head CE). In addition, when evaluated on the CommonsenseQA private Test set, our approach gets 71.6% accuracy which is close to RoBERTa\textsubscript{LARGE} cross-entropy (Liu et al., 2019) under an important hyper-parameter grid search\footnote{https://github.com/pytorch/fairseq/tree/master/examples/roberta/commonsense_qa} (72.1% accuracy).

When using 100% of the COPA training set (400 train samples), our method outperforms the head CE setting per 5 points and the head margin setting per 3 points, achieving an accuracy of 92.4% on the Test set. This result allows our approach to reach the second place in the SuperGLUE leaderboard\footnote{https://super.gluebenchmark.com/leaderboard} (Wang et al., 2019) between RoBERTa\textsubscript{LARGE} (Liu et al., 2019) and the T5 model composed of 11 billions of parameters (Raffel et al., 2019) (respectively 90.6 and 94.8 % accuracy on the Test set).

We also notice that our method provides a much more stable training relative to the random seed as shown by the box plots in Figure 2 a) and 3 a). When training on the full COPA dataset, our method exhibits a $\times10$ standard deviation reduction on the test accuracy compared to the head CE setting (1.35% versus 12.8%). Our intuition is that the improved stability is due to the better reuse of the pre-trained model priors and the absence of new randomly initialized weights. This is important result towards easier experiment comparisons as fine-tuning BERT-like architectures is known to be unstable across random restarts as shown in (Phang et al., 2018).

**6 Conclusions**

In this work, we presented a new method for plausibility ranking tasks, specifically targeting commonsense ranking problem. We define a scoring function that leverages the MLM head of large pre-trained bidirectional transformer models. We establish strong results in a zero-shot setting on four commonsense reasoning datasets, comparable to supervised approaches. We then fine-tune such model using a margin-based loss on the proposed scoring function, and provide a comparative study with state of the art randomly initialized head methods. Our study demonstrates that the direct use of MLM over custom head yields increasingly superior performance gain when decreasing training data size. The proposed approach outperforms state-of-the-art training methods in terms of both test accuracy and training stability.

Future works include applying such scoring method on broader classification tasks like Natural Language Inference and Sentiment Analysis. We also think that our token-level scoring method could be used during the self-supervised pre-training phase to extend traditional next sentence prediction and sequence ordering tasks, bringing more commonsense knowledge in the model.
References

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.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiwicz. 2019. A surprisingly robust trick for the Winogrand schema challenge. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR’12, pages 552–561. AAAI Press.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4818–4823, Florence, Italy. Association for Computational Linguistics.

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

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases?. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *ArXiv* preprint arXiv:1911.03681.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv* preprint arXiv:1910.10683.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 5926–5936, Long Beach, California, USA. PMLR.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Triew H., Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1066–1076. Association for Computational Linguistics.
Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.