Attention-based Contrastive Learning for Winograd Schemas

Tassilo Klein
SAP AI Research
tassilo.klein@sap.com

Moin Nabi
SAP AI Research
m.nabi@sap.com

Abstract

Self-supervised learning has recently attracted considerable attention in the NLP community for its ability to learn discriminative features using a contrastive objective (Qu et al., 2020; Klein and Nabi, 2020). This paper investigates whether contrastive learning can be extended to Transfomer attention to tackling the Winograd Schema Challenge. To this end, we propose a novel self-supervised framework, leveraging a *contrastive loss* directly at the level of *self-attention*. Experimental analysis of our attention-based models on multiple datasets demonstrates superior commonsense reasoning capabilities. The proposed approach outperforms all comparable unsupervised approaches while occasionally surpassing supervised ones.[1]

1 Introduction

Pre-trained language models have propelled the domain of NLP to a new era. Specifically, Transformer-based models are the driving force behind recent breakthroughs. However, despite all the recent success in text understanding, the task of commonsense reasoning is still far from being solved (Marcus, 2020; Kocijan et al., 2020). In order to assess the commonsense reasoning capabilities of automatic systems, several tasks have been devised. Among them is the popular Winograd Schema Challenge (WSC) (Levesque et al., 2012). WSC frames commonsense reasoning as a pronoun co-reference resolution problem (Lee et al., 2017), which consists of twin-pair sentences. Experts curated the twin pairs manually to be "Google-proof", e.g., simple statistical biases from large data should be insufficient to resolve the pronouns. Hence, solving WSC was expected to require diverse reasoning capabilities (e.g., relational, causal). Sentences in the twin pairs differ only in the "trigger word". Furthermore, trigger words are responsible for switching the correct answer choice between the questions. Below is a popular example from WSC. In the example, the trigger word is underlined. The challenge entails resolving the pronoun "it" with a noun from the candidate set ("suitcase", "trophy"):

Sentence-1: *The trophy doesn't fit in the suitcase because* ***it*** *is too small.*
Answers: A) the trophy ***B)*** the suitcase

Sentence-2: *The trophy doesn't fit in the suitcase because* ***it*** *is too big.*
Answers: A) the trophy ***B)*** the suitcase

The research community has recently experienced an abundance of methods proposing to utilize the latest language model (LM) for commonsense reasoning (Kocijan et al., 2019b; He et al., 2019; Ye et al., 2019; Ruan et al., 2019; Trinh and Le, 2018; Klein and Nabi, 2019; Tamborrino et al., 2020). Models learned on large text corpora were hoped to internalize commonsense knowledge implicitly encountered during training. Most of such methods approach commonsense reasoning in a two-stage learning pipeline. Starting from an initial self-supervised learned model, commonsense enhanced LMs are obtained in a subsequent fine-tuning (ft) phase. Fine-tuning enforces the LM to solve the downstream WSC task only as a plain co-reference resolution task. Despite some initial success in this direction, we hypothesize that the current self-supervised tasks used in the pre-training phase are too "shallow" to enforce the model to capture a "deeper" notion of commonsense (Kejriwal and Shen, 2020; Elazar et al., 2021). Shortcomings of models obtained in such a fashion can partially be attributed to the training corpora itself. Standard training sets such

[1]The source code can be found at: https://github.com/SAP-samples/emnlp2021-attention-contrastive-learning/

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2428–2434

as Wikipedia barely contain commonsense knowledge, so supervised fine-tuning only promotes the discovery of "artificial" cues and language biases to tackle commonsense reasoning (Trichelair et al., 2018; Saba, 2018; Trichelair et al., 2019; Emami et al., 2019; Kavumba et al., 2019). This is the main reason why supervised methods pre-trained on large datasets (e.g., WinoGrande) can transfer effectively to smaller target datasets (e.g., WSC) yet do not show the same performance level on the source dataset.

In an attempt to avoid the utilization of shallow commonsense reasoning cues, very recently (Klein and Nabi, 2020) introduced a Contrastive Self-Supervised (CSS) learning method, leveraging the mutual-exclusivity of WSC pairs. Despite almost reaching state-of-the-art performance, the approach does not require external knowledge for training. However, the authors observed that leveraging the contrastive loss directly on the Transformer-backbone at the *LM-level* can destabilize the self-supervised optimization.

We propose a novel self-supervised loss to address this, introducing an abstraction layer between the backbone and the downstream task. Our approach smoothly manipulates the attentions to achieve this goal in a Transformer-like fashion while avoiding destabilization of the intrinsics. To do so, we make use of the non-identifiability property of attention, which implies that the attention values are not uniquely determined from the head's output, and vice versa. Consequently, various attention patterns across the Transformer can result in identical outcomes and permit regularization - see for details (Brunner et al., 2020). Intuitively, the proposed contrastive attention mechanism does not overwrite the low-level semantics captured in the pre-trained model. Instead, it induces modest adjustments via attention patterns. In the context of Winograd schemas, the proposed approach shifts the attention from the wrong answer candidate to the right candidate. Simultaneously, the attention contrast forces the LM to be more rigorous across attention heads while consistent over the samples.

In summary, our contributions are the following: **First**, we propose a contrastive loss enforced on the Transformer attention, which helps for the emergence of commonsense patterns. **Second**, we present empirical evidence showcasing the viability of the approach, outperforming comparable state-of-the-art.

2 Attention-based Contrastive Learning

Preliminaries: The proposed approach extends the contrastive self-supervised method (Klein and Nabi, 2020) to facilitate commonsense reasoning for Winograd schemas at the attention level. In the context of data, we assume that $\mathcal{D}$ with $N = |\mathcal{D}_c|$ is a dataset constructed from contrastive twin-pairs samples, $(s_i, s_{i+1}) \in \mathcal{D}_c$, with c_j and c_{j+1} denoting answer candidates. The difference between the sentence pairs is the so-called "trigger words" responsible for flipping the answer in pronoun disambiguation. Thus, this trigger-word structure induces a mutual-exclusive candidate answer relationship at the pair level. In the context of the model, we employ a Transformer-based LM for Masked Token Prediction (Devlin et al., 2018). Given a sentence with a `[MASK]` token, the LM provides the likelihood of sentence s_i with the token replaced by candidate tokens $c_j \in \{$`[CANDIDATE-1]`, `[CANDIDATE-2]`$\}$ denoted as $p\,(c_j|s_i) = p_{i,j}$, assuming that the dataset consists of $i \in \frac{N}{2}$ distinct twin-pairs. Besides sentence likelihoods, the Transformer architecture also provides an attention tensor $\mathcal{A}(x) \in \mathbb{R}^{H\times L\times \mathcal{C}\times \mathcal{C}}$, for a given an input x with $|x| = \mathcal{C}$, where L denotes the number of layers, and H the number of heads. Then the tensor decomposes into elements $a^{h,l}_{i,j}(x)$, gauging the influence of token i w.r.t. token j in layer l of attention head h.

2.1 Method

Inspired by (Klein and Nabi, 2020), we make use of the structural prior of Winograd schemas and their within-pair mutual-exclusivity. We formulate this as in context of Transformer-based LM as a multi-task optimization problem defined as:

$$\mathcal{L}(f_\theta) = \mathcal{L}(f_\theta)_{CM} + \mathcal{L}(f_\theta)_{CM}$$

Here f denotes the underlying LM parameterized by θ. The first term, $\mathcal{L}_{CM}$ leverages the contrast arising from twin pairs enforcing mutual-exclusivity on attentions. The second term, $\mathcal{L}_{CM}$, seeks to further reduce ambiguity at the LM level by maximization between the differences of the likelihoods for the answer candidates. It should be noted that although the proposed approach leverages the structural prior of twin pairs and it does not make use of any class label information explicitly, similar to (Klein and Nabi, 2020). See Fig. 1 for a schematic illustration of the proposed method.

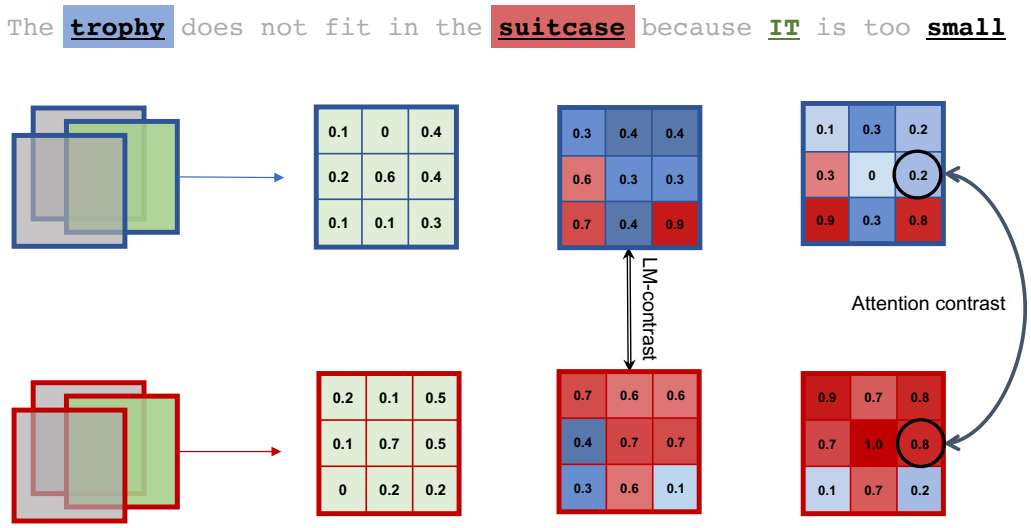

Figure 1: Schematic illustration of contrastive learning for a particular sentence, where colors show attention maps for different words of a mock setup with 3 heads and 3 layers. Squares with blue/red frames correspond to specific sliced attention 3×3 matrix for candidates, establishing the relationship 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. The **Attention-based Contrast** shows a more consistent disambiguation attention for the correct candidate compare to the **LM-based Contrast** (Klein and Nabi, 2020).

2.1.1 Contrastive Attention

The contrastive mechanism targets regularizing self-attention patterns emerging by invoking the LM on an input sequence, thus providing the model with commonsense reasoning capabilities. Specifically, the proposed approach seeks to induce *consistently* higher attention values across all attention heads and layers for the right candidate as opposed to the wrong one. This contrasts with the LM-level MEx (Klein and Nabi, 2020), where only the overall value of attention is enforced to be higher for the right candidate - see Fig. 1 for an illustration. Hence, the proposed approach promotes the emergence of diverse attention patterns between the attention heads, avoiding issues such as the collapse to a single dominant head. To this end, our proposed approach invokes twin-pair contrast on *attention level* for samples in $\mathcal{D}_c$. This pushes for the superior establishment of distant dependencies more indirectly than enforcing it directly on the LM. Given the observation of (Brunner et al., 2020) that distant relationships are formed towards the end of the transformer stack, we restrict instantiation of the contrastive attention loss on the last layers. This, in combination with the non-uniqueness of Transformer attentions w.r.t. output, operating on attention level suggests comparably smoother behavior. In order to resolve ambiguity in the attention mechanism w.r.t. candidates, we tie mutual exclusivity together with a binarization scheme. Here binarization refers to a simple form of mutual exclusivity loss applied in binary classification cases (such as WSC), defined as:

$$\mathcal{L}_{CA} = -\lambda \sum_{\substack{i=1,j=1 \\ i+=2}}^{N,2} \left(\mathbf{a}_{i,j} - \frac{\mathbf{e}}{2}\right)^2 + \left(\mathbf{a}_{i+1,j} - \frac{\mathbf{e}}{2}\right)^2 + 1 - (\mathbf{a}_{i,j} - \mathbf{a}_{i+1,j})^2 + 1 - [(1 - \mathbf{a}_{i,j}) - (1 - \mathbf{a}_{i+1,j})]^2$$

Here $\mathbf{a} \in \mathbb{R}^H$ denotes a vector containing the attentions of all heads. Assuming attentions to be *normalized* w.r.t. candidates, i.e., $\sum_j a_{i,j} = 1$, effectively turns them into pseudo-likelihoods. Furthermore, $\mathbf{e} \in \mathbb{R}^H$ is vector with all elements 1, and $\lambda \in \mathbb{R}$ a hyperparameter.

2.1.2 Contrastive Margin

To stabilize optimization, we leverage consistency between sentences of each contrastive pair. On the one hand, it leads to faster convergence. On the other hand, it enforces smoothness on the loss surface and decreases the overall gradient fluctuation. The CM term seeks to maximize the margin between the LM likelihoods for each candidate in a pair:

$$\mathcal{L}_{CM} = -\alpha \sum_{i,j}^{N,2} \max\left(0, |p_{i,j} - p_{i,j+1}| + \beta\right),$$

Method	WSC	DPR	W.G.	K.Ref	W.Gen.
Bi-LSTM (Opitz and Frank, 2018)	56.0	63.0	-	-	-
BERT (DPR-ft)	69.8	-	50.2	61.0	59.2
BERT (MaskedWiki-DPR-ft) (Kocijan et al., 2019b)	67.0	83.3	50.2	-	**79.2**
BERT (WikiCREM-DPR-ft) (Kocijan et al., 2019a)	71.8	84.8	-	-	-
RoBERTa (DPR-ft)	83.1	-	**59.4**	84.2	-
RoBERTa (WG-ft) (Sakaguchi et al., 2019)	**90.1**	**92.5**	-	**85.6**	-
(Rahman and Ng, 2012)	58.0	73.0	-	-	-
(Peng et al., 2015)	-	76.4	-	-	-
Knowledge Hunter (Emami et al., 2018)	57.1	-	-	-	-
E2E (Emami et al., 2019)	-	-	-	58.0	-
MAS (Klein and Nabi, 2019)	60.3	-	-	-	-
Ensemble LM (Trinh and Le, 2018)	63.8	-	-	-	-
BERT (zero-shot) (Vaswani et al., 2017)	62.6	58.5	51.7	62.3	62.5
RoBERTa (zero-shot) (Liu et al., 2019)	67.7	70.3	53.7	60.4	61.6
Self-supervised Ref. (BERT) (Klein and Nabi, 2021)	61.5	61.3	52.3	62.4	62.0
Self-supervised Ref. (RoBERTa) (Klein and Nabi, 2021)	71.7	76.9	55.0	63.9	69.1
CSS (BERT) (Klein and Nabi, 2020)	69.6	80.1	50.9	65.5	69.5
CSS (RoBERTa) (Klein and Nabi, 2020)	79.8	**90.6**	57.7	68.0	76.2
Our Proposed Method	**84.1**	90.0	**60.8**	**69.9**	**93.3**

Table 1: Results on different tasks: WSC, DPR, WinoGrande(W.G.), KnowRef (K.Ref) and WinoGender (W.Gen). Task performances in accuracy (%) are subdivided into two parts. Top: supervised (ft), bottom: unsupervised.

with $\alpha, \beta \in \mathbb{R}$ being hyperparameters.
When training the language model, the algorithm will look for a pattern of consistency in the attention heads and layers rather than force-fit supervisory signals from labels. Assuming the answer of the first sentence is `[CANDIDATE-1]`, it follows the answer for the second one is `[CANDIDATE-2]`. This restricts the answer space. As the model is forced to leverage the pairwise relationship to resolve the ambiguity, it needs to generalize w.r.t. commonsense relationships. Intuitively speaking, as no labels are provided to the model during training, the model seeks to make the answer probabilities less ambiguous. It should be noted that the proposed approach leverages the structural prior of twin pairs, not making use of any label.

3 Experiments and Results

3.1 Setup

We leverage RoBERTa (Liu et al., 2019) as Language Model for Masked Token Prediction, and DPR (Rahman and Ng, 2012) as dataset for training. Specifically, we use the Hugging Face (Wolf et al., 2019) implementation of RoBERTa. The model is trained for 22 epochs using a batch size of 18 (pairs). Hyperparameters are $\alpha = 0.05$, $\beta = 0.02$, $\lambda = 1.0$. For optimization Adam was selected with a learning rate of 10^{-5}. Commonsense reasoning is approached by first fine-tuning the pre-trained RoBERTa (*large*) masked-LM model on the DPR (Rahman and Ng, 2012).

3.2 Results

While observing loss fluctuations by learning mutual-exclusivity at LM model directly via log-likelihood (MEx) (Klein and Nabi, 2020), such fluctuations are less pronounced when operating at attention level (proposed approach).

We evaluate the performance on different tasks - see Tab. 1. As can be seen, the proposed approach outperforms other unsupervised methods by a significant margin, outperforming some supervised methods or at least significantly reducing the gap between supervised and unsupervised approaches. The results are discussed separately for each benchmark below:

WSC (Levesque et al., 2012): the most well-known pronoun disambiguation benchmark. Our method outperforms the strongest unsupervised baseline CSS(BERT) margin of $(+14.5\%)$ and CSS(RoBERTa) by $(+4.3\%)$.

DPR (Rahman and Ng, 2012): this pronoun disambiguation benchmark resembles WSC, yet sig-

nificantly larger in size. According to (Trichelair et al., 2018), less challenging due to inherent biases. Here the proposed approach outperforms the unsupervised baseline CSS(BERT) by a margin of $(+9.9\%)$, while observing a slight drop of (-0.6%) compared to CSS(RoBERTa).

WinoGrande (W.G.) (Sakaguchi et al., 2019): the largest dataset for Winograd co-reference resolution. Our method outperforms the unsupervised baseline CSS(BERT) by $(+9.9\%)$ and CSS(RoBERTa) by $(+3.1\%)$, even surpassing supervised RoBERTa(DPR-ft) by $(+1.4\%)$.

KnowRef (Emami et al., 2019): a co-reference corpus addressing gender and number bias. The proposed approach outperforms the unsupervised baseline CSS(BERT) by a margin of $(+4.4\%)$ and CSS (RoBERTa) by $(+1.9\%)$.

WinoGender (Rudinger et al., 2018): a gender-balanced co-reference corpus. The proposed approach outperforms the unsupervised baseline CSS(BERT) by a margin of $(+23.8\%)$ and CSS(RoBERTa) by $(+17.1\%)$.

3.2.1 Attention-level Analysis

Inspired by (Vig and Belinkov, 2019), we assess the impact of the attention mechanism by analyzing the attention tensor which is obtained by querying the attention of the *MASK* token w.r.t. right/wrong candidate over all layers and heads. The tensor decomposes into elements $a_{i,j}^{h,l}(x)$, gauging the influence of token i w.r.t. token j in layer l of attention head h. Aggregating the attention of *MASK* token i for the tokens c_j for the right and wrong candidates by summation, slices the tensor into matrices $A_r, A_w \in \mathbb{R}^{H \times L}$ generating attention maps.Here A_r, A_w corresponds to the attention maps w.r.t. the *right* answer and the *wrong* answer, respectively. Following (Brunner et al., 2020), we also investigated the maps of the *last* k-layers, denoted as $A_r^{[k]}$ and $A_w^{[k]}$. We then computed the attention difference and entropy $H(.)$ difference on the attention maps of all DPR (Rahman and Ng, 2012) samples, and presented the statistics in Tab. 2.

We observed a significant concentration of attention for the right candidates for the proposed approach compared to the wrong ones. This pattern is even more pronounced for the last 3 layers. Specifically, we observed the manifestation of an average entropy of 3.41 *(right)* nats vs. 2.1 nats *(wrong)* on the last 3 layers, giving rise to the emergence of the desired pattern of more concerted attention on the right candidate. See supplementary material for more detailed results.

	RoBa	CSS	**Ours**
$\lvert H(A_r) - H(A_w)\rvert$	0.024	**0.097**	0.078
$\lvert H(A_r^{[3]}) - H(A_w^{[3]})\rvert$	0.005	0.772	**1.328**
$\lvert \bar{A}_r - \bar{A}_w\rvert$	0.009	0.010	**0.061**
$\lvert \bar{A}_r^{[3]} - \bar{A}_w^{[3]}\rvert$	0.020	0.034	**0.306**

Table 2: Attention analysis of different models on DPR, and $k = 3$. Top: entropies, Bottom: mean statistics.

Method	**WSC**	**W.G.**
RoBERTa (Liu et al., 2019)	67.76	53.75
CSS (RoBERTa)	79.85	57.77
Our Method (CM)	60.81	52.88
Our Method (CA)	80.95	57.14
Our Method (CA+CM)	**84.10**	**60.80**

Table 3: Ablation study, performance in accuracy (%)

3.2.2 Ablation Study

To assess the contribution of each component, we evaluated the performance of each module separately, gradually adding components to the loss. See Tab. 3 for the ablation study on WSC and WinoGrande. Pre-trained RoBERTa (*large*) constitutes the baseline. MEx denotes the mutual-exclusive loss on the sentence log-likelihoods (Klein and Nabi, 2020), CA denotes the contrastive attention defined in Sec. 2.1.1, CM denotes the contrastive-margin defined in Sec. 2.1.2. While the CA term alone already suggests strong performance, this does not apply to the CM term. Given the regulatory nature of the CM term, optimizing it in isolation yields a model with inferior accuracy.

4 Conclusion

In this paper, we introduce an attention-level self-supervised learning method for commonsense reasoning. Specifically, we propose a method that enforces a contrastive loss on the attentions produced by transformer LM while pushing the likelihood of the candidates towards the extremities. The experimental analysis demonstrates that our proposed system outperforms the previous unsupervised state-of-the-art in multiple datasets.

References

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

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*.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Bias detection, training and commonsense disentanglement in the winograd schema.

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.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961.

Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. 2019. A hybrid neural network model for commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 13–21, Hong Kong, China. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics. [link].

Mayank Kejriwal and Ke Shen. 2020. Do fine-tuned commonsense language models really generalize?

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7517–7523, Online. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2021. Towards zero-shot commonsense reasoningwith self-supervised refinement of language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019a. Wikicrem: A large unsupervised corpus for coreference resolution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong.

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

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

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*.

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

Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Juri Opitz and Anette Frank. 2018. Addressing the winograd schema challenge as a sequence ranking task. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 41–52.

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, Denver, Colorado. Association for Computational Linguistics.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. *arXiv preprint arXiv:2010.08670*.

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, Jeju Island, Korea. Association for Computational Linguistics.

Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. 2019. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *arXiv preprint arXiv:1904.09705*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641.

Alexandre Tamborrino, Nicola Pellicano, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pretraining is (almost) all you need: An application to commonsense reasoning. *arXiv preprint arXiv:2004.14074*.

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. *CoRR*, abs/1811.01778.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. 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 3380–3385, Hong Kong, China. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/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.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

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.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint arXiv:1908.06725*.