Contrastive Self-Supervised Learning for Commonsense Reasoning

Tassilo Klein   **Moin Nabi**
SAP AI Research, Berlin, Germany
{tassilo.klein, m.nabi}@sap.com

Abstract

We propose a self-supervised method to solve *Pronoun Disambiguation* and *Winograd Schema Challenge* problems. Our approach exploits the characteristic structure of training corpora related to so-called "trigger" words, which are responsible for flipping the answer in pronoun disambiguation. We achieve such commonsense reasoning by constructing pairwise contrastive auxiliary predictions. To this end, we leverage a *mutual exclusive loss* regularized by a *contrastive* margin. Our architecture is based on the recently introduced transformer networks, BERT, that exhibits strong performance on many NLP benchmarks. Empirical results show that our method alleviates the limitation of current supervised approaches for commonsense reasoning. This study opens up avenues for exploiting inexpensive self-supervision to achieve performance gain in commonsense reasoning tasks. [1]

1 Introduction

Natural language representation learning (e.g., BERT (Devlin et al., 2018), etc.) can capture rich semantics from text and consistently improve the performance of downstream natural language processing (NLP) tasks. However, despite the recent progress, the task of *commonsense reasoning* is still far from being solved. Among many factors, this can be attributed to the strong correlation between attainable accuracy and training corpora size and quality. A particular case in point is the Winograd Schema Challenge (WSC) (Levesque et al., 2012). Despite its seeming simplicity for humans, it is still not solved by current algorithms.

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

---

[1]Code available at https://github.com/SAP-samples/acl2020-commonsense/

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

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

For humans resolving the pronoun "it" to "the suitcase" is straightforward. However, a system without the capacity of commonsense reasoning is unable to conceptualize the inherent relationship and, therefore, unable to distinguish "the suitcase" from the alternative "the trophy".

Recently, the research community has experienced an abundance in methods proposing to utilize latest word embedding and language model (LM) technologies for commonsense reasoning (Kocijan et al., 2019; He et al., 2019; Ye et al., 2019; Ruan et al., 2019; Trinh and Le, 2018; Klein and Nabi, 2019). The underlying assumption of these methods is that, since such models are learned on large text corpora (such as Wikipedia), they implicitly capture to a certain degree commonsense knowledge. As a result, models permit reasoning about complex relationships between entities at inference time. Most of the methods proposed a two-stage learning pipeline. They are starting from an initial self-supervised model, commonsense-aware word embeddings are then obtained in a subsequent fine-tuning (ft) phase. Fine-tuning enforces the learned embedding to solve the downstream WSC task only as a plain co-reference resolution task.

However, solving this task requires more than just employing a language model learned from large text corpora. We hypothesize that the current self-supervised pre-training tasks (such as *next sentence prediction*, *masked language model*, etc.) used in the word embedding phase are too "easy" to

enforce the model to capture commonsense. Consequently, the supervised fine-tuning stage is not sufficient nor adequate for learning to reason commonsense. This is particularly more severe when pre-training on commonsense-underrepresented corpora such as Wikipedia, where the authors often skip incorporating such information in the text, due to the assumed triviality. In this case, the supervised fine-tuning does not seem to be enough to solve the task, and can only learn to "artificially" resolve the pronoun based on superficial cues such as dataset and language biases (Trichelair et al., 2018; Saba, 2018; Trichelair et al., 2019; Emami et al., 2019; Kavumba et al., 2019).

In this work, we propose to use minimal existing supervision for learning a commonsense-aware representation. Specifically, we provide the model with a supervision level identical to the test time of the Winograd challenge. For that, we introduce a self-supervised pre-training task, which only requires pair of sentences that differ in as few as one word (namely, "trigger" words). It should be noted that the notion of trigger words is inherent to the concept of Winograd Schema questions. Trigger words are responsible for switching the correct answer choice between the questions. In the above example, the adjectives big and small act as such trigger words. Given the context established by the trigger word, candidate answer A is either right in the first sentence and wrong in the second, or vice-versa. As is evident from the example, trigger words give rise to the mutual-exclusive relationship of the training pairs. The proposed approach targets to incorporate this pairwise relationship as the only supervisory signal during the training phase. Training in such a contrastive self-supervised manner is inducing a commonsense-aware inductive bias. This can be attributed to several factors. Optimization enforces the classifier to be more rigorous in its decision as well as consistent across pairs while being discriminative. Specifically, in the absence of strong individual sentence signals, the model seeks to combine weak signals across pairs. This unsupervised task is much harder to learn compared to the supervised task, and resolving the respective associations requires a notion of commonsense knowledge. Consequently, we postulate that training with contrastive self-supervised fashion allows for learning more in-depth word relationships that provide better generalization properties for commonsense reasoning.

For that, we propose to incorporate a Mutual Exclusive (MEx) loss (Sajjadi et al., 2016) during the representation learning phase by maximizing the mutual exclusive probability of the two plausible candidates. Specifically, given a pair of training sentence, the pronoun to be resolved is masked out from the sentence, and the language model is used to predict such only one of the candidates can fill in the place of masked pronoun while fulfilling the mutual-exclusivity condition. In this self-supervised task, the labels (i.e., correct candidates) do not have to be known a priori. Thus it allows learning in an unsupervised manner by exploiting the fact that the data is provided in a pairwise fashion.

Our contributions are two-fold: **(i)** we propose a novel self-supervised learning task for training commonsense-aware representation in a minimally supervised fashion. **(ii)** we introduce a pair level mutual-exclusive loss to enforce commonsense knowledge during representation learning.

2 Previous Works

There is a wealth of literature on commonsense reasoning, but we only discuss here the ones most related to our work and refer the reader to the recent analysis paper by (Trichelair et al., 2019).

Traditional attempts on commonsense reasoning usually involve heavy utilization of annotated knowledge bases (KB), rule-based reasoning, or hand-crafted features (Bailey et al., 2015; Schüller, 2014; Sharma et al., 2015). Only very recently and after the success of natural language representation learning, several works proposed to use *supervised learning* to discover commonsense relationships, achieving state-of-the-art in multiple benchmarks (see, e.g., (Kocijan et al., 2019; He et al., 2019; Ye et al., 2019; Ruan et al., 2019)). As an example, (Kocijan et al., 2019) has proposed to exploit the labels for commonsense reasoning directly and showed that the performance of multiple language models on Winograd consistently and robustly improves when fine-tuned on a similar pronoun disambiguation problem dataset. Despite the success of these methods, we posit that *unsupervised learning* is still more attractive for commonsense reasoning tasks, because curating a labeled dataset entailing all existing commonsense is likely to be an unattainable objective. Very recently, unsupervised learning has also been applied successfully to improve commonsense reasoning in a few works (Trinh and
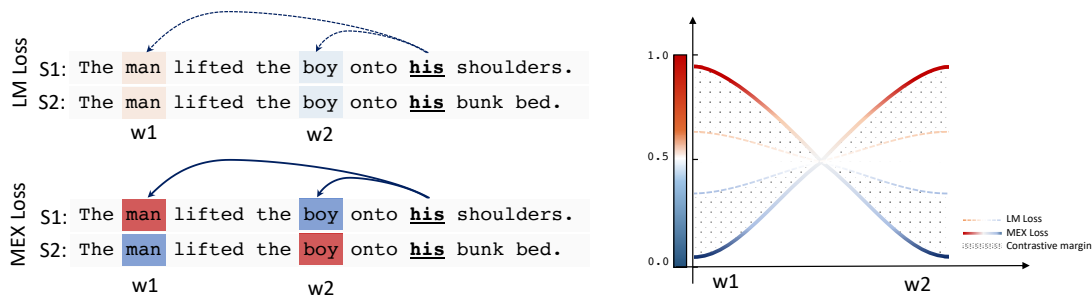
Figure 1: Contrastive Self-supervised Learning for a particular sentence. Colors show the likelihood of different words. Weak commonsense signal manifests in the likelihood of both candidates to be around 0.5 for the LM-only loss (shown in dash lines); incorporating the MEx loss (shown in solid lines) leverages mutual exclusivity of the candidates, enforcing the classifier to be more rigorous and consistent across pairs (best shown in color).

Le, 2018; Klein and Nabi, 2019). The most pioneering work in this space is probably by (Trinh and Le, 2018), where the authors proposed to use BERT as a (pseudo) language model to compute the likelihood of candidates replacing the pronoun, and the corresponding ratio giving rise to answer. In another recent work, (Klein and Nabi, 2019) proposed a metric based on the maximum attention score for commonsense reasoning. While these papers show that BERT can implicitly learn to establish complex relationships between entities, our results suggest that solving commonsense reasoning tasks require more than unsupervised models learned from massive text corpora. We note that our model is different from all of the methods above. A key difference is that they require fine-tuning, or explicit substitution or heuristic-based rules, whereas our method learns a commonsense-aware representation in self-supervised fashion.

3 Contrastive Self-supervised Reasoning

The goal of the proposed approach is to exploit the mutual-exclusive nature of the training samples of commonsense reasoning corpora. Given two sentences where the only difference between them is the trigger word(s), we postulate that the pairwise pronoun disambiguation is mutually exclusive. We formulate this idea using a contrastive loss and use this to update the language model. The proposed contrastive loss decomposes into two components:

$$\mathcal{L}(f_\theta) = \mathcal{L}(f_\theta)_{MEx} + \mathcal{L}(f_\theta)_{CM} \quad (1)$$

Here f is the language model parameterized by θ. The first term, $\mathcal{L}_{MEx}$ enforces the Mutual Exclusivity of the answers across pairs. As such, it is a relaxation of the Exclusive-OR (XOR) operator w.r.t. candidates. The second term, $\mathcal{L}_{CM}$

constitutes the Contrastive Margin. It enforces a margin between the candidate likelihoods from the language model. Whereas $\mathcal{L}_{MEx}$ operates across pairs, $\mathcal{L}_{CM}$ considers the candidates of each pair. Although both terms encourage the same property (mutual exclusivity of the answers), we empirically observed that adding CM increases stability. It should be noted that the proposed approach does not make use of any class label information explicitly. Rather, it solely exploits the structural information of the data. In terms of the language model, we leverage BERT for Masked Token Prediction (Devlin et al., 2018). This entails replacing the pronoun by a mask, i.e., *[MASK]*. As a result, we yield probabilities for the candidates of each sentence.

Preliminaries: Given an associated pair of training sentences, i.e., (s_j, s_{j+1}), where the difference between the sentence pairs are the trigger words. Let c_i and c_{i+1} be the two answer candidates for the masked pronoun resolution task. Then employing BERT for Masked Token Prediction (Devlin et al., 2018) provides $p(c_i|s_j)$ and $p(c_{i+1}|s_j)$, i.e., the likelihood of the first and the second candidate being true in sentence s_j, respectively. It should be noted, if a candidate consists of several tokens, the corresponding number of *[MASK]* tokens is used in the masked sentence. The candidate probability then corresponds to the average of log-probabilities of each composing token.

Since a candidate cannot be the right answer for the first and second sentence in the pair, we yield a logical term that holds *true* for viable answers. It is worth noting that the logical expression is not unique as many logical equivalents exist:

$$(c_{i,1} \oplus c_{i+1,1}) \wedge (c_{i,2} \oplus c_{i+1,2}) \wedge (c_{i,1} \oplus c_{i,2}) \quad (2)$$

Here $\oplus$ denotes the XOR operator and $\mathbf{c}_{i,j} \in$

7519

$\{0, 1\}$ denotes the binary state variable corresponding to candidate $\mathbf{c}_i$ in sentence $\mathbf{s}_j$.

Mutual-Exclusive Loss: In order to be differentiable, the discrete logical term of Eq. 2 has to be converted into a "soft" version. To this end, we replace the binary variables with their corresponding probabilities. Similarly, the logical operators are replaced accordingly to accommodate for the probabilistic equivalent.

With $a \oplus b = (a \wedge \neg b) \vee (\neg a \wedge b)$ a logical decomposition of the XOR operator, we adopt the following replacement scheme: **(i)** $\bigwedge_i^k x_i$ is replaced by $\prod_i^k x_i$, **(ii)** $\bigvee_i^k x_i$ is replaced by $\sum_i^k x_i$, **(iii)** the not operation of a binary variable $\neg x_i$ is replaced by $1 - x_i$. Thus, transforming all the logical terms of Eq. 2, we yield the following soft-loss equivalent:

$$\mathcal{L}_{MEx} = -\gamma \sum_{i=i+2,}^{N} \mathbf{p}_{i,1}\mathbf{p}_{i+1,2}\left(1 - \mathbf{p}_{i,2}\mathbf{p}_{i+1,1}\right)$$
$$+ \mathbf{p}_{i,2}\mathbf{p}_{i+1,1}\left(1 - \mathbf{p}_{i,1}\mathbf{p}_{i+1,2}\right) \quad (3)$$

Here $\mathbf{p}_{i,j} = p\left(c_i | s_j\right) \in [0, 1]$ denotes the probability of candidate $\mathbf{c}_i$ being the right answer in sentence $\mathbf{s}_j$, γ is a hyperparameter, and N corresponds to the number of training samples. Intuitively speaking, as no labels are provided to the model during training, the model seeks to make the answer probabilities less ambiguous, i.e., approximate binary constitution. As the model is forced to leverage the pairwise relationship in order to resolve the ambiguity, it needs to generalize w.r.t. commonsense relationships. As such, the task is inherently more challenging compared to, e.g., supervised cross-entropy minimization.

Contrastive Margin: In order to stabilize optimization and speed-up convergence, it is beneficial to augment the MEx loss with some form of regularization. To this end, we add a contrastive margin. It seeks to maximize difference between the *individual* candidate probabilities of the language model and is defined as,

$$\mathcal{L}_{CM} = -\alpha \cdot \max\left(0, |p_{i,j} - p_{i,j+1}| + \beta\right), \quad (4)$$

with α, β being hyperparameters. See Fig. 1 for a schematic illustration of the proposed method.

4 Experiment & Results

In this work, we use the PyTorch (Wolf et al., 2019) implementation of BERT. Specifically, we employ a pre-trained BERT *large-uncased* architecture. The model is trained for 25 epochs using a batch size of 4 (pairs), hyperparameters $\alpha = 0.05$, $\beta = 0.02$ and $\gamma = 60.0$, and Adam optimizer at a learning rate of 10^{-5}. We approach commonsense reasoning by first fine-tuning the pre-trained BERT LM model on the DPR training set (Rahman and Ng, 2012). Subsequently, we evaluate the performance on four different tasks.

Pronoun Disambiguation Problem: The first evaluation task is on PDP-60 (Davis et al., 2016), which aims the pronoun disambiguation. As can be seen in Tab. 1 (top), our method outperforms all previous unsupervised results by a significant margin of at least (+15.0%). Next, we have the alternative approaches making use of a supervisory signal during training. Here, our method outperforms even the best system (78.3%) by (+11.7%).

Winograd Schema Challenge: The second task is WSC-273 (Levesque et al., 2012), which is known to be more challenging than PDP-60. Here, our method outperforms the current *unsupervised* state-of-the-art (Trinh and Le, 2018) (62.6%), as shown in Tab. 1 (middle). Specifically, our method achieves an accuracy of (69.6%), which is (+7%) above the previous best result. Simultaneously, the proposed approach is just slightly lower than the best *supervised* approach (Kocijan et al., 2019).

Definite Pronoun Resolution: The third task is DPR (Rahman and Ng, 2012), which resembles WSC. Compared to the latter, it is significantly larger in size. However, according to (Trichelair et al., 2018), it is less challenging due to several inherent biases. Here the proposed approach outperforms the best alternative by a margin of (+3.7%), as can be seen in Tab. 1 (lower part).

KnowRef: The fourth task is KnowRef (Emami et al., 2019), which is a coreference corpus tailored to remove gender and number cues. The proposed approach outperforms the best alternative by a margin of (+4.5%), as can be seen in Tab. 1 (bottom).

Ablation study on contrastive margin: The contrastive margin term was incorporated in our method as a regularizer, mainly for the sake of having faster convergence. As such, discarding it during optimization has a minor impact on the accuracy of most benchmarks (less than 1% on WSC, DPR, KnowRef). However, on PDP, we noticed a wider margin of more than 10%.

5 Discussion

In contrast to supervised learning, where semantics is directly injected through "labels", the self-

PDP-60 (`sup.`) (Davis et al., 2016)	
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 KB	66.7 %
BERT-ft (Kocijan et al., 2019)	78.3 %
PDP-60 (`unsupervised`)	
Unsupervised Sem. Similarity (USSM)	55.0 %
Transformer LM (Vaswani et al., 2017)	58.3 %
BERT LM (Trinh and Le, 2018)	60.0 %
MAS (Klein and Nabi, 2019)	68.3 %
DSSM (Wang et al., 2019)	75.0 %
CSS (Proposed Method)	**90.0 %**
WSC-273 (`sup.`) (Levesque et al., 2012)	
USSM + KB	52.0%
USSM + Supervised DeepNet + KB	52.8 %
Transformer (Vaswani et al., 2017)	54.1 %
Know. Hunter (Emami et al., 2018)	57.1 %
GPT-ft (Kocijan et al., 2019)	67.4 %
BERT-ft (Kocijan et al., 2019)	71.4 %
WSC-273 (`unsupervised`)	
Single LMs (Trinh and Le, 2018)	54.5 %
MAS (Klein and Nabi, 2019)	60.3 %
DSSM (Wang et al., 2019)	63.0 %
Ensemble LMs (Trinh and Le, 2018)	63.8 %
CSS (Proposed Method)	**69.6 %**
DPR (Rahman and Ng, 2012)	
(Rahman and Ng, 2012)	73.0%
(Peng et al., 2015)	76.4 %
CSS (Proposed Method)	**80.1 %**
KnowRef (Emami et al., 2019)	
E2E (Emami et al., 2019)	58.0 %
BERT-ft (Emami et al., 2019)	61.0 %
CSS (Proposed Method)	**65.5 %**

Table 1: Results on different tasks. From Top to bottom: PDP, WSC, DPR, KnowRef. The first two task performances are subdivided into two parts. Upper part: supervised, lower part: unsupervised.

supervised-learning paradigm avoids labels by employing a pre-text task and exploits the structural "prior" of data as a supervisory signal. In this paper, this prior corresponds to the Winograd-structured twin-question pairs, and the pre-text task is to switch the correct answer choice between the pairs using "trigger" words. We postulate that training in such a contrastive self-supervised manner allows for learning more commonsense-aware word rela-

tionships that provide better generalization properties for commonsense reasoning. We acknowledge that this prior is strong in terms of data curation, i.e., expert-crafted twin pairs. However, during training, we provide the model to have access to a supervision level equal to the test time, i.e., not making use of the labels. Therefore, maximizing the mutual exclusive probability of the two plausible candidates is inducing a commonsense-aware inductive bias without using any label information and by merely exploiting the contrastive structure of the task itself. This is confirmed by our approach, reaching the performance of the most recent supervised approaches on multiple benchmarks. At last, we note that our model is different from the self-supervised contrastive learning methodology in (Chen et al., 2020), which focuses on learning powerful representations in the self-supervised setting through batch contrastive loss. A key difference compared to this method is that they generate the contrastive pairs as data augmentations of given samples, whereas in our setting the auxiliary task of "mutual exclusivity" is enforced on given contrastive pairs.

6 Conclusion

The proposed approach outperforms all approaches on PDP and DPR tasks. At the more challenging WSC task, it outperforms all unsupervised approaches while being comparable in performance to the most recent supervised approaches. Additionally, it is less susceptible to gender and number biases as the performance on KnowRef suggests. All this taken together confirms that self-supervision is possible for commonsense reasoning tasks. We believe in order to solve commonsense reasoning truly, algorithms should refrain from using labeled data, instead exploit the structure of the task itself. Therefore, future work will aim at relaxing the prior of Winograd-structured twin-question pairs. Possibilities are automatically generating an extensive collection of similar sentences or pre-training in a self-supervised fashion on large-scale Winograd-structured datasets, such as the recently published WinoGrande (Sakaguchi et al., 2019). Furthermore, we seek to investigate the transferability of the obtained inductive bias to other commonsense-demanding downstream tasks, which are distinct from the Winograd-structure.

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*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Ernest Davis, Leora Morgenstern, and Charles Ortiz. 2016. Human tests of materials for the winograd schema challenge 2016.

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.

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].

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.

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

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*.

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.

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*.

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

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1163–1171. Curran Associates, Inc.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

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 challenge: Building and using a semantic parser and a knowledge hunting module. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI15, pages 1319–1325. AAAI Press.

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.

Shuohang Wang, Sheng Zhang, Yelong Shen, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Jing Jiang. 2019. Unsupervised deep structured semantic models for commonsense reasoning. 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 882–891, Minneapolis, Minnesota. 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*.