Winograd Schemas in Portuguese

Gabriela S. de Melo, Vinicius A. Imaizumi, Fabio G. Cozman

[1]Universidade de Sao Paulo, Escola Politecnica
Av. Prof. Mello Moraes 2231, 05508-900 Sao Paulo, Brazil

{gabriela.melo,vinicius.imaizumi, fgcozman}@usp.br

Abstract. *The Winograd Schema Challenge has become a common benchmark for question answering and natural language processing. The original set of Winograd Schemas was created in English; in order to stimulate the development of Natural Language Processing in Portuguese, we have developed a set of Winograd Schemas in Portuguese. We have also adapted solutions proposed for the English-based version of the challenge so as to have an initial baseline for its Portuguese-based version; to do so, we created a language model for Portuguese based on a set of Wikipedia documents.*

1. Introduction

The Winograd Schema Challenge (WSC) consists of questions whose answer depends on relating a particular word to one of two possible antecedents [Levesque et al. 2012]. One example of a Winograd Schema is:

> The trophy doesn't fit into the brown suitcase because it is too large. What is too large?

To answer this, one would have to understand that, in the snippet "it is too large", "it" refers to the trophy, and not to the suitcase.

The WSC has been advocated as an alternative to the Turing Test. This is due to the fact that it contains particularly difficult coreference resolution problems that can only be solved using commonsense knowledge. These questions are simple for humans: in a test run in 2015, humans displayed 92% accuracy [Bender 2015]. But the questions are challenging for computers; standard coreference resolution solvers do not work well on the challenge [Peng et al. 2015]. Another difficulty is the fact that there are few examples of sentences that actually qualify as Winograd Schemas (based on the rules proposed by [Levesque et al. 2012]); solutions that depend on training with very large datasets of Schemas do not work.

Interest in the WSC has also emerged in the field of Natural Language Processing (NLP), such as [Radford et al. 2019], where Winograd Schemas are used as benchmarks. Pronoun resolution as needed in the WSC is very relevant to the development of NLP. Other tasks, such as machine translation, also depend on resolving ambiguous sentences as displayed in the WSC [Davis 2016].

In order to stimulate the development of NLP research in Portuguese, we have created a set of Winograd Schemas in Portuguese. This task is not as simple as translating each Schema in English word by word; there are several rules to consider when developing a Schema, and their effect changes from language to language. We have also

developed a system for solving Winograd Schemas in Portuguese; this system should serve as an initial baseline for the task.

The paper is organized as follows. Section 2 reviews related work, and Section 3 reports on our collection of Portuguese-based Winograd Schemas. Then, in Section 4 we present the solver that we have developed for establishing baseline results for the Portuguese-based collection of Schemas, and Section 5 explains how we obtained the results; Section 6 presents those results. We end by discussing conclusions and future work in Section 7.

2. Related Work

A set of rules has been established for a sentence to be considered a Winograd Schema [Levesque et al. 2012]. In short, the rules determine: that the possible antecedents are noun phrases of the same gender; that a pronoun or possessive adjective refers to one of these antecedents, but is also of the correct type for the other possible antecedent; that answer 0 is always the first party mentioned in the sentence, answer 1 is the second; that there is a *special word*, that, when changed to the *alternate word*, the sentence is still perfectly valid, but the answer switches. Also, a Schema cannot be too obvious, in the sense that a simple statistical check of whether the special word happens more frequently with one of the possible answers than the other must not be able to solve the Schema. Finally, the sentences must not be too ambiguous; that is, fluent speakers of the language must be able to correctly answer the sentence without doubts.

Several systems have been developed to solve the WSC. A number of them are based on understanding how the sentences are structured, and from this, to use examples or rules to resolve the ambiguity. Some of these systems derive sets of features from the sentences and use them for training the model [Rahman and Ng 2012]. Others are based on linguistic tools for parsing sentences [Sharma et al. 2015, Emami et al. 2018] or fitting them into predicate schemas [Peng et al. 2015] and using these strategies on the Winograd sentence and on results of queries on search engines. Yet others are based on relevance theory and knowledge graphs [Schüller 2014] and some on logic rules based on correlation formulas [Bailey et al. 2015].

Another type of solver leverages the fact that models trained on vast amounts of language corpora indirectly incorporate commonsense knowledge when learning word relations. These are relatively recent solutions, mostly based on Deep Learning, using neural networks based on embeddings [Liu et al. 2016], siamese networks [Opitz and Frank 2018], language model networks [Trinh and Le 2018], and transformer architectures such as GPT-2 [Radford et al. 2019] and BERT [Kocijan et al. 2019, Ruan et al. 2019]. Traditional linguistic tools for extracting dependency graphs in the sentences are also employed by [Ruan et al. 2019], who use this extracted information to complete the traditional transformer model [Vaswani et al. 2017].

Some of the existing solutions apply to a subset of Schemas, restricting its usability on a more general pronoun resolution scenario [Schüller 2014, Bailey et al. 2015, Peng et al. 2015, Sharma et al. 2015].

To deal with the fact that we now have a relatively small number of instances of Winograd schemas, some of the proposals in the literature have developed their own

custom datasets to help with training. [Rahman and Ng 2012] have developed a set of relaxed Winograd schemas, containing 941 sentence pairs. This dataset has also been used by [Peng et al. 2015, Opitz and Frank 2018, Kocijan et al. 2019, Ruan et al. 2019]. [Trinh and Le 2018] and [Kocijan et al. 2019] have also developed custom datasets, based on text corpora.

Recently, new evaluation criteria for the challenge have been proposed by [Trichelair et al. 2018]. These criteria are based on dividing the data into two new sets, based on associativity and switchability characteristics of the sentences. This allows for further insights into model performance and facilitates understanding robustness to slight variations in sentences.

There has been work translating Winograd Schemas to other languages, specifically French, Japanese and Chinese; these sets are available at the Challenge's official website.[1] The method used to generate French translations has been reported by [Amsili and Seminck 2017].

3. A Collection of Portuguese-based Schemas

Our Portuguese-based collection of Schemas was developed following the rules proposed by [Levesque et al. 2012], mentioned in Section 2. To develop our set, we used as a base the set of 285 original English-based Schemas that are available online [2] and manually translated each of them. Our translated set is also available both in a JSON format and in a more visually pleasing, HTML format.[3] Note that a few additional tags are present in the JSON format, and these are described in Section 5.3.

Three native Portuguese speakers worked on translating the sentences, all of whom were familiar with the Winograd Schemas Challenge and the rules regarding the consideration of a sentence as a Winograd Schema. Each sentence was translated by one of the speakers and validated by the other two. For eight sentences we could not find a suitable translation, and hence these were discarded in the Portuguese set.

Some of the Schemas had to be adapted. For instance, consider the sentence:

The trophy doesn't fit into the brown suitcase because it is too large.

Its literal translation would be:

O troféu não cabe na maleta porque ele é muito grande.

In Portuguese, however, objects are not gender-neutral, and, in this case, *troféu* is of masculine gender, while *maleta* is of feminine gender. This would make the pronoun *ele* to be very easily resolved, given that it refers to a masculine object, and the only masculine object in the sentence is *troféu*. We adapted such sentences so that they would follow all the rules for being a Winograd Schema, as long as there was a plausible adaptation that would not change the meaning of the sentence. For instance, we produced:

A medalha não cabe na maleta porque ela é muito grande.

[1] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

[2] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.xml

[3] https://github.com/gabimelo/portuguese_wsc/blob/master/data/processed/portuguese_wsc.json and https://github.com/gabimelo/portuguese_wsc/blob/master/data/raw/portugues_wsc.html

Given that *medalha* is of feminine gender, it now becomes possible for the pronoun to refer to either *medalha* or *maleta*.

It is also worth noting that there are some significant differences between Brazilian Portuguese and Portuguese in Portugal. We have generated a collection of Schemas in Brazilian Portuguese, and we have not evaluated how natural they sound to Portuguese speakers from Portugal.

At first, we keep names as they were in the original sentences. Nevertheless, the fact that many of these names are not commonly found in Portuguese speaking countries might interfere with the task. Hence, we have developed an additional collection where names have been replaced by popular names in Brazil. The only restriction for these substitutions was that gender would be kept. Also, names of famous personalities that appear in the Schemas, such as Madonna or Shakespeare, were kept as in the original collection. This set is available in HTML and JSON formats.[4]

4. A Baseline Solver for the Portuguese-based WSC

To build a baseline solver for the Portuguese-based WSC we employed a language model as proposed by [Trinh and Le 2018]. Their solution is an ensemble of models that are so large that actually running a test is a nontrivial matter. We thus pursued a much simpler language model than they did, and also a single model instead of an ensemble of models, as we were mostly interested in establishing an initial baseline that other researchers can easily run.

For our language model, we used a neural network with input and output layers with hidden unit size equal to that of the vocabulary, acting as encoding and decoding layers, using an embedding size of 200, and two LSTM layers, with 200 hidden units each. We used dropout between layers, with a probability of dropout equal to 0.2. The code for our solution has been made publicly available, in its entirety.[5]

The use of a language model for resolving Winograd Schemas goes as follows:

1. Each one of the candidate antecedents is substituted in place of the pronoun to be resolved. For instance, in this sentence (examples in English):

   > The trophy doesn't fit into the brown suitcase because it is too large.

   We would then generate two sentences:

   > The trophy doesn't fit into the brown suitcase because the trophy is too large.
   >
   > The trophy doesn't fit into the brown suitcase because the suitcase is too large.

2. Each of these generated sentences is passed on to the model. In this step, all words in the sentence are sequentially sent to the language model; at each step, the generated probability for the next word in the sentence is stored.

[4]These datasets can be found at https://github.com/gabimelo/portuguese_wsc/blob/master/data/raw/portuguese_wsc_portuguese_names.html and https://github.com/gabimelo/portuguese_wsc/blob/master/data/processed/portuguese_wsc.json

[5]https://github.com/gabimelo/portuguese_wsc

3. The joint probability of each sentence is calculated. The sentence with the highest probability is assigned as the correct one. We employed two ways to calculate this probability, as [Trinh and Le 2018]. We describe these two ways of scoring the sentences later in Section 5.1.

Our language model was trained using a corpus that we created for this task. There seems to be no corpus currently established as a benchmark for language models in Portuguese. We derived our own from a Wikipedia dump which was the latest as of April 23rd, 2019 [6]. We used a subsection of the dump for training the model, equivalent to 15 MB (out of the original 2.3 GB), which contains 2,018,034 training tokens, 389,541 validation tokens, and 373,508 test tokens. The vocabulary consists of every word that appeared more than 5 times in the dataset, resulting in 32,032 unique tokens. Words outside of the vocabulary were replaced by an <unk> token and end of sentences were represented by <eos>. We have also made this corpus available.[7]

We also trained a similarly simple language model for English, and applied that to the original set of English Schemas, so as to compare how the approach works for these two different languages with similarly sized models. For this, the model was trained with the Wikitext-2 dataset [Merity et al. 2016]. This dataset contains 33,278 unique tokens, having 2,075,677 training tokens, 216,347 validation tokens, and 244,102 test tokens. This model follows the same architecture as the Portuguese model.

To train the models, we used learning rate annealing and started with an initial learning rate of 20. Input and output embeddings were tied, as proposed by [Press and Wolf 2017, Inan et al. 2016]. Gradients were clipped at 0.25 and training ran for 40 epochs. Sentences were organized into sequences of length 35.

5. An Empirical Analysis of Performance

This section presents details on how the performance of our solver was evaluated. It starts by explaining the two different scoring methods we utilized. We then indicate the metrics used when presenting the results for our models. As we mentioned in Section 2, new subsets for the evaluation of WSC solvers have been proposed by [Trichelair et al. 2018], and have, since then, been used for reporting the performance of such solvers. Section 5.3 discusses the incorporation of this approach in our results. Lastly, we argue that grammatical mistakes introduced by the automatic substitution of candidate antecedents in place of the pronouns to be resolved might impact on the performance of solvers, and explain how we measured this impact.

5.1. Scoring of Sentences

There are two approaches proposed in [Trinh and Le 2018] for the scoring of the sentences. The first type is called full scoring. It is the ordinary joint probability of the sentence. That is:

$$Score_{full}(w_k \leftarrow c) = P_\theta(w_1, w_2, ..., w_{k-1}, c, w_{k+1}, ..., w_n)$$

[6] https://dumps.wikimedia.org/ptwiki/latest/ptwiki-latest-pages-articles.xml.bz2

[7] https://github.com/gabimelo/portuguese_wsc/tree/master/data/processed

Where $w_k \leftarrow c$ indicates the word at position k being substituted by candidate c. The second way of scoring the model is with partial scoring, which is described as:

$$Score_{partial}(w_k \leftarrow c) = P_\theta(w_{k+1}, ..., w_n | w_1, ..., w_{k-1}, c)$$

We used the two approaches and present the results for each of them.

5.2. Metrics

Some of the works related to solving the Winograd Schema Challenge present their results in terms of the accuracy or precision metrics - accuracy being used when all of the WSC sentences are answered by the model, and precision when otherwise. [Emami et al. 2018] argued that the F1 Score would be a more suitable metric when the solution being used presents answers for only some, but not all, of the sentences in the Winograd Schema set. In these cases, they constructed the F1 Score by having the values for recall and precision being defined as:

$$recall = \frac{\#Correct}{Size\ of\ Winograd\ Set}\ ; \quad precision = \frac{\#Correct}{\#Answered}$$

It can be noted that the definition of recall is the same as for the accuracy on the full set. When the amount of answered Schemas is the same as the size of the full set, $recall$ and $precision$ both become the same, and in this case, based on the definition of the F1 Score, we can see that accuracy and F1 Score are equivalent.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Given that this is the case for our system (all sentences can get answered by our model), for evaluating the results, we used as our metric the accuracy of the answers.

We also use the consistency metric, which calculates for how many of the switchable sentences the system provided the same answer both for the original sentence and the switched version of it.

5.3. Subsets of Schemas

Based on the criteria established by [Trichelair et al. 2018], we have used the subsets of the datasets made available by their work. These subsets propose a way for better understanding the robustness of solvers and were called the switchable and the associative sets.

The switchable set consists of sentences that can have the antecedents switched; the sentence is still valid and the answer to it switches accordingly. Therefore, for this switchable subset, we can have the unswitched (original) sentences, and the switched ones. The associative set consists of sentences where one of the antecedents relates more strongly to the special word than the other (although this is one of the rules for the Winograd Schema Challenge collection, there has not been a strong check on whether all sentences follow this rule, and therefore some can be considered as being associative). Thus,

all sentences in the collection are divided into the associative and the non-associative subsets.

For the Portuguese set, we considered as associative and as switchable the same sentences that were marked as that on their work. This resulted in 35 associative sentences and 135 switchable sentences. We have two associative sentences less than the work in English (two of the sentences that were not able to be translated into Portuguese were considered associative).

Our collection of Schemas with names translated to Portuguese was also subdivided into these subsets.

5.4. Manual Corrections

When doing automatic substitutions of the pronouns, some of the cases become grammatically incorrect. For instance, in the sentence:

> Jim signaled the barman and gestured toward his empty glass.

Which has as pronoun to be resolved *his* and as possible answers *A. Jim* or *B. The barman*, the automatic substitutions becomes:

> Jim signaled the barman and gestured toward Jim empty glass.
> Jim signaled the barman and gestured toward the barman empty glass.

These sentences are missing the "'s" after the substituted candidates. We noticed that these sentences were left without any further corrections on the dataset of sentences after substitutions released by [Trinh and Le 2018] and that there has not been any work analyzing if such errors might impact on the performance of WSC solvers.

In Portuguese, there are even more cases in which the automatic substitution does not work than in the English ones (where these cases are restricted to the usage of his/her pronouns). In Portuguese, the first type of sentence where this sort of error occurs are sentences with the usage of possessive pronouns, such as:

> Há uma fenda na parede. É possível enxergar o jardim através **dela**.

Where the substitution for the first possible antecedent becomes:

> Há uma fenda na parede. É possível enxergar o jardim através **a fenda**.

When it should instead be:

> Há uma fenda na parede. É possível enxergar o jardim através **da fenda**.

Another case is that where the pronoun appears joined with the verb (a common Portuguese sentence construction):

> Eu estava tentando abrir o cadeado com a chave, mas alguém havia preenchido a fechadura com goma de mascar, e eu não conseguia **removê-la**.

Resulting in:

> Eu estava tentando abrir o cadeado com a chave, mas alguém havia preenchido a fechadura com goma de mascar, e eu não conseguia **removê-a goma de mascar**.
>
> Instead of:
>
> Eu estava tentando abrir o cadeado com a chave, mas alguém havia preenchido a fechadura com goma de mascar, e eu não conseguia **remover a goma de mascar**.

The last case where there were issues with the automatic substitution were those where the pronoun appears before the verb:

> Eu usei um pano velho para limpar o alicate, e então **o coloquei** no lixo.
>
> Resulting in:
>
> Eu usei um pano velho para limpar o alicate, e então **o pano coloquei** no lixo.
>
> But when the pronoun gets substituted it should instead be:
>
> Eu usei um pano velho para limpar o alicate, e então **coloquei o pano** no lixo.

For the English-based collection of Schemas, we kept these substituted sentences unchanged — that is, in the same manner as they were used by previous work that utilized this set. This was done so in order to have accurate comparisons of results. For the Portuguese set, we developed, in addition to the collection of sentences after the automatic substitutions, a collection containing the manually fixed sentences, in order to assess whether this sort of treatment for the automatic substitution of the pronouns would make any difference. This helps understand how well the solution proposed here could be expanded to less structured pronoun resolution problems - where manual fixes might not be feasible.

6. Results

Table 1 presents the main results from our experiments. Each of the rows, except for the last, represent the accuracy for a subset of the dataset, for each of the scoring techniques (full and partial). The last row refers to the consistency, which was defined in Section 5.2.

Using the same model with very similar vocabulary sizes for the English and Portuguese languages, the English language shows better results on the Winograd Schema Challenge. The fluctuation of results between each of the subsets was similar for the two languages. Because the actual corpus used for training the model is of great influence on its performance, the fact that for the Portuguese model we used a corpus naively derived from the Wikipedia dump while for the English model we used a dataset that is a benchmark for language model work might be important.

Table 1. Main Results

		English	Portuguese
Original (Full Dataset)	Full	50.55%	45.13%
	Partial	49.08%	44.77%
Associative	Full	45.95%	34.29%
	Partial	54.05%	51.43%
Non-Associative	Full	51.27%	46.69%
	Partial	48.3%	43.80%
Switched	Full	48.09%	40.74%
	Partial	51.14%	39.26%
Unswitched	Full	48.85%	42.22%
	Partial	46.56%	42.96%
Consistency	Full	2.20%	9.03%
	Partial	4.03%	13.72%

6.1. Comparison of English Models

It is important to compare the performance of our model for the English set of Winograd Schemas to that of the work from which we based our solution from [Trinh and Le 2018]; this comparison is present in table 2. Given that their publication was previous to that of the work introducing the associative and switchable subsets, we extracted the results from the latter [Trichelair et al. 2018]. We only compare the results from their single language model solution (their better solution involves an ensemble of multiple models). We also only compare to the partial scores, as the results from full scoring were not disclosed.

Table 2. English Results - Partial Scoring

	English - Single LM [Trichelair et al. 2018]	English - Ours
Original (Full Dataset)	54.58%	49.08%
Associative	73.0%	54.05%
Non-Associative	51.7%	48.30%
Switched	54.20%	51.14%
Unswitched	54.96%	46.56%
Consistency	56.49%	4.03%

In terms of model size, their model consists of almost 1.8 billion parameters, while ours has 7.3 million; the difference in the capacity of the models is very significant. These results show that an improved language model can perform substantially better in the English dataset [Trinh and Le 2018], which leads us to believe that reaching better performance for the Portuguese collection might also be achieved through the improvement of the language model being used (our current Portuguese model is very similar in size to our English model, both consisting of 7 million parameters).

6.2. Impact of Translation of Names and Manual Fixes

As reported in Section 5.4, we made some manual fixes to the automatic substitution of candidate antecedents in place of pronouns, to analyze whether this would be a necessary

measure when solving the WSC utilizing language models. Table 3 shows the impact of manual fixes and also that of translating the names for some more commonly found in the Portuguese language.

Table 3. Portuguese Names and Manual Fixes

	Full	Partial
Portuguese	45.13%	44.77%
Portuguese - Manually Fixed	44.77%	45.13%
Portuguese - Portuguese Names	45.49%	44.04%
Portuguese - Portuguese Names - Manually Fixed	45.49%	44.77%

For our current approach, there was little difference in the result between the original set and the ones with these changes. The fact that manual corrections did not imply in a great difference in results suggests it might not be necessary to spend much effort into trying to improve the automatic substitution of candidates method. Nevertheless, given that these aspects might have more of an influence if other approaches for solving the challenge were being used, we still find it relevant to release the collection of Portuguese Schemas with these translated names and manual fixes, in addition to the base collection.

7. Conclusion and Future Work

In this paper we have described a collection of Portuguese-based Winograd schemas; to the best of our knowledge, no similar collection exists today. We have also created a baseline for solving the Portuguese-based WSC, based on an approach that has produced good results for the English-based version of the WSC [Trinh and Le 2018].

The results obtained by this baseline show just how difficult it is to solve the Winograd Schema Challenge. It is worth noting that the solvers for the English challenge that had substantially larger performance than ours are all based on very large language models or linguistic models such as BERT [Devlin et al. 2018]. This demonstrates that generic Natural Language Processing in Portuguese can benefit from such language models.

In future work, the language model could be improved, by increasing model capacity or by developing models such as BERT [Devlin et al. 2018] for the Portuguese language.

Additionally, other methods used for the English-based challenge could be employed. An extended collection of relaxed Schemas could be developed in Portuguese, similarly to the one released by [Rahman and Ng 2012]. A customized corpus for fine-tuning models such as the one developed by [Kocijan et al. 2019] could also be developed for Portuguese. Another technique that could be tested is that of anonymizing the Schemas that mention proper names, which was tested by works for the English collection such as [Rahman and Ng 2012, Opitz and Frank 2018].

Acknowledgements

This work was carried out with the support of Itaú Unibanco S.A.; the second author has been supported by the Itaú Scholarship Program (PBI), linked to the Data Science Center (C2D) of the Escola Politécnica da Universidade de São Paulo.

The third author has been partially supported by the Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq), grant 312180/2018-7. The work was also supported by the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), grant 2016/18841-0, and also by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) - finance code 001. We are also grateful to the Center for Inovation at Universidade de São Paulo (InovaUSP) for hosting our lab.

References

Amsili, P. and Seminck, O. (2017). A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Bailey, D., Harrison, A., Lierler, Y., Lifschitz, V., and Michael, J. (2015). The winograd schema challenge and reasoning about correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*. AAAI Press.

Bender, D. (2015). Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference 2015, Greensboro, NC, USA, April 25-26, 2015.*, pages 39–45.

Davis, E. (2016). Winograd schemas and machine translation. *CoRR*, abs/1608.01884.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emami, A., De La Cruz, N., Trischler, A., Suleman, K., and Cheung, J. C. K. (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.

Inan, H., Khosravi, K., and Socher, R. (2016). Tying word vectors and word classifiers: A loss framework for language modeling. *CoRR*, abs/1611.01462.

Kocijan, V., Cretu, A., Camburu, O., Yordanov, Y., and Lukasiewicz, T. (2019). A surprisingly robust trick for winograd schema challenge. *CoRR*, abs/1905.06290.

Levesque, H. J., Davis, E., and Morgenstern, L. (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.

Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2016). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *CoRR*, abs/1611.04146.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Opitz, J. and Frank, A. (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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Peng, H., Khashabi, D., and Roth, D. (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.

Press, O. and Wolf, L. (2017). Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rahman, A. and Ng, V. (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.

Ruan, Y., Zhu, X., Ling, Z., Shi, Z., Liu, Q., and Wei, S. (2019). Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *CoRR*, abs/1904.09705.

Schüller, P. (2014). Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'14, pages 358–367. AAAI Press.

Sharma, A., Vo, N. H., Aditya, S., and Baral, C. (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*, IJCAI'15, pages 1319–1325. AAAI Press.

Trichelair, P., Emami, A., Cheung, J. C. K., Trischler, A., Suleman, K., and Diaz, F. (2018). On the evaluation of common-sense reasoning in natural language understanding. *CoRR*, abs/1811.01778.

Trinh, T. H. and Le, Q. V. (2018). A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.