Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 21–26
Marseille, 11–16 May 2020

Mandarinograd: A Chinese Collection of Winograd Schemas

Timothée Bernard, Ting Han
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{timothee.bernard, ting.han}@aist.go.jp

Abstract

This article introduces Mandarinograd, a corpus of Winograd Schemas in Mandarin Chinese. Winograd Schemas are particularly challenging anaphora resolution problems, designed to involve common sense reasoning and to limit the biases and artefacts commonly found in natural language understanding datasets. Mandarinograd contains the schemas in their traditional form, but also as natural language inference instances (ENTAILMENT or NO ENTAILMENT pairs) as well as in their fully disambiguated candidate forms. These two alternative representations are often used by modern solvers but existing datasets present automatically converted items that sometimes contain syntactic or semantic anomalies. We detail the difficulties faced when building this corpus and explain how we avoided the anomalies just mentioned. We also show that Mandarinograd is resistant to a statistical method based on a measure of word association.

Keywords: Winograd Schemas, common sense reasoning, anaphora, natural language inference

1. Introduction

Winograd Schemas (henceforth "WS"), introduced by Levesque et al. (2012), are pairs of short reading comprehension problems that usually amount to finding the antecedent of an anaphoric expression. They constitute a theoretically motivated benchmark for Natural Language Understanding (NLU) that is also one of the most challenging today (Nangia and Bowman, 2019). WS are, however, hard to collect and few datasets are publicly available. As good quality datasets are essential to evaluate and compare computational systems, and as NLU should not be (and indeed is not) confined to the English language, we present here Mandarinograd, the first collection of WS in Mandarin Chinese. Mandarinograd is publicly available at `https://gitlab.com/vanTot/mandarinograd/`.

In Section 2, we briefly explain what WS are and why they are considered useful. In Section 3, we detail how we built Mandarinograd. In Section 4, we present a simple statistical baseline on this dataset.

2. Winograd Schema

2.1. Definition

Each element of a WS is a text mentioning two entities (e.g., a trophy and a suitcase) and containing a referential expression — usually a pronoun (e.g., *it*) — which, based only on syntax and basic semantic selection restrictions, is ambiguous as to which of the two entities it refers to. Each text of the pair should have a commonly accepted most natural interpretation and these interpretations should be such that the ambiguous expressions do not refer to the same entity. For each text, the problem is to determine to which entity the ambiguous expression refers to — which is usually expressed as a question. (1) presents such a WS.[1]

(1) a. The trophy doesn't fit into the brown suitcase because it's too [small/large].
b. What is too [small/large]?
c. the suitcase/the trophy

In order to prevent the introduction of biases that would make the disambiguation doable in practice without "proper reasoning", the two texts of the pair should be as similar to each other as possible (think about, e.g., salience-based algorithms such as the classic one presented by Lappin and Leass (1994), or about how modern NLU systems exploit artefacts present in the datasets (Gururangan et al., 2018)). Ideally, the two texts only differ by one word, called the *special* word in one version and the *alternate* word in the other, as is the case in (1). Many WS do not follow this strict pattern, however, and we will call one version of the WS the *special* version, and the other the *alternate* version.

2.2. AI and Natural Language Understanding

While non-problematic for any competent speaker of the language they are expressed in, WS arguably require non-trivial forms of reasoning, in particular *common sense reasoning*, involving a wide range of linguistic knowledge (e.g., synonymy, hypernymy) as well as world-knowledge (about, e.g., geometry, time, causality or human interactions). That is one of the main reasons why the task of solving WS has been proposed by Levesque et al. (2012) as a practical alternative to the Turing Test (also known as the "Imitation Game"; (Turing, 1950)).[2] This is also why WS have been included in the General Language Understanding Evaluation (GLUE) dataset (Wang et al., 2019), a benchmark commonly used today to evaluate the performance of Natural Language Understanding (NLU) systems in English. In GLUE, WS are present under the form of (typically four) Natural Language Inference (NLI) problems consisting of determining whether the first of two texts

[1] Following the usual notation for WS, the two elements are here factorised using square brackets in the obvious way.

[2] An argument given against the Turing Test is that it in fact tests the deception skill of its subject, which has to pretend to be human. While the ability to lie convincingly can indeed be related to intelligence, it is arguably not the quality we are primarily interested in when doing AI. In contrast, solving WS does not suffer from this shortcoming.

naturally entails the second or not. For example, the special version of the WS in (1) corresponds to the two NLI problems in (2).

(2) a. (i) The trophy doesn't fit into the brown suitcase because it's too small.
(ii) The suitcase is too small.
(iii) ENTAILMENT
b. (i) The trophy doesn't fit into the brown suitcase because it's too small.
(ii) The trophy is too small.
(iii) NO ENTAILMENT

3. A Chinese collection of WS

In this section, we detail how we built Mandarinograd, our collection of 154 WS in simplified Mandarin Chinese.

3.1. Collecting WS

To the extent of our knowledge, there is currently no known technique to automatically detect or generate WS. As a consequence, WS are either hand-crafted or manually translated from WS in another language. Both tasks are particularly challenging as many conditions must be met in a text to ensure it forms a proper WS. In particular, translation is made difficult by the fact that some WS rely on linguistic phenomena that play differently in different languages. For example, (3) relies on the fact that *A introduces B to C* is underspecified about whether B is introduced to C or the contrary (or both). An appropriate translation for *introduce* might not exist in every language, and indeed this WS has not been translated in the French collection of Amsili and Seminck (2017). Similarly, (4) is a WS thanks to the fact that, in English, the third person plural subject pronoun (*they*) is unaffected by the animacy of its reference. In a language such as Mandarin Chinese, however, two different forms would have to be used (他们 for the judges and 它们 for the chatbots).

(3) a. This book introduced Shakespeare to [Ovid/Goethe]; it was a major influence on his writing.
b. Whose writing was influenced?
c. Shakespeare/Goethe

(4) a. At the Loebner competition the judges couldn't figure out which respondents were the chatbots because they were so [advanced/stupid].
b. Who were so [advanced/stupid]?
c. the chatbots/the judges

The Winograd Schema Challenge's website[3], which compiles information about WS, lists, in addition to 151 English WS[4], translations of 145 of them in Japanese, 107 in French (Amsili and Seminck, 2017) and 12 in Chinese. During the redaction of the present paper, a collection of Brazilian Portuguese translations was released by Melo et al. (2020), containing 285 items (which amounts to 142.5 schemas).

In addition to these carefully hand-crafted but small datasets, there exist other sets of WS-style problems in English which, however, still required a large amount of human supervision. For example, Rahman and Ng (2012) lead a group of 30 students to produce a set of 941 pairs satisfying slightly relaxed constraints. Even though still harder than most cases of pronominal resolutions, they have appeared easier to solve automatically than true WS. More recently, Sakaguchi et al. (2019) released WinoGrande, which contains 12,282 WS-style sentences generated by crowdworkers and then automatically filtered to remove statistical biases. They show that this dataset is particularly challenging for state-of-the-art systems, but is also harder for human beings: their methodology estimates human performance at an accuracy of 96.5% on the WS of Levesque et al. (2012) and 94.0% on WinoGrande.

Let us now turn to Mandarinograd. We first checked the 12 existing Chinese WS and corrected a few errors.[5] We then translated the remaining 138 English WS into Chinese. In the remainder of this section, we detail the translation procedure which consists of an initial translation by the authors of this paper, followed by a checking phase performed by four other native Chinese speakers, before a final validation by the authors. For each WS, our dataset indicates not only its traditional form (as in (1)), but also the four corresponding NLI pairs as found in GLUE and the four disambiguated versions as used by some language-model based WS solvers such as the one of Trinh and Le (2018).[6]

3.2. Initial translation

We tried to translate each English WS into Chinese as closely as possible in terms of vocabulary and sentence structure while ensuring that the resulting texts (i) were natural, (ii) formed a proper WS — i.e., it contained a referential ambiguity easily solvable via common sense reasoning — and (iii) gave rise to clear ENTAILMENT/NO ENTAIL-

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

[4] The website only lists 150 WS in English, but, following Bender (2015), we have considered number 135 as two different ones.

[5] In one WS, for instance, the answers were initially specified as 铁 (*iron*) and 泡沫塑料 (*plastic foam*) instead of the correct 球 (*ball*) and 桌子 (*table*).

[6] The disambiguated texts are obtained by replacing the ambiguous expression in a WS by an expression referring unambiguously to either of the two entities, as in (i) for (1).

(i) a. The trophy doesn't fit into the brown suitcase because the suitcase is too small.
b. The trophy doesn't fit into the brown suitcase because the trophy is too small.
c. The trophy doesn't fit into the brown suitcase because the trophy too large.
d. The trophy doesn't fit into the brown suitcase because the suitcase is too large.

These texts are used only for computational purpose and do not always sound entirely natural due to possible repetitions. However, while Trinh and Le (2018) performed mere substitutions without correction — generating anomalous sentences, for instance, when possessive pronouns are replaced by noun phrases as in *Jim signaled the barman and gestured toward Jim empty glass*, with *Jim* instead of *Jim's* —, the texts in Mandarinograd are all grammatically correct.

MENT pairs in their NLI form (the form used in the GLUE dataset; see (2)). During this process, we paid particular attention to the items that Bender (2015), who performed a crowd-sourced experiment to estimate the human ability to solve WS (in English), observed to be confusing to native speakers.

In many cases, as for (1) translated as indicated in (5), a quasi-literal translation was possible.

(5) a. 奖杯 无法 放进到 棕色的 箱子 里 ,
trophy can't be-placed brown suitcase in ,
因为 它 太 [小 / 大] 了 。
because it too [small / big] PERF .
b. 什么 东西 太 [小 / 大] 了 ？
what thing too [small / big] PERF ?
c. 箱子 / 奖杯
suitcase / trophy

As explained above, however, a direct translation was not always possible. In these cases, we tried to produce a WS thematically related to the original one and based on the same form of reasoning. For example, (4) was adapted to (6). (检测出 is a sort of phrasal verb which can be translated as *to detect*.)

(6) a. 安全 机器人 [/ 无法] 成功 检测 出
security bots [/ can't] succeed detect *out*
病毒 , 因为 他们 不够 智能 。
virus , because they not-enough smart .
b. 谁 不够 智能 ？
what not-enough smart ?
c. 病毒 / 安全 机器人
virus / security bots

Similarly, (7) does not admit a straightforward translation because both *high* (for a shelf) and *tall* (for a pot) would be translated as 高. As a consequence, we adapted it to (8).

(7) a. I couldn't put the pot on the shelf because it was too [high/tall].
b. What was too [high/tall]?
c. the shelf/the pot

(8) a. 我 没有 办法 把 壶 放在 架子 上 ,
I no way take pot put shelf on ,
因为 太 [高 / 矮] 了 。
because too [high / short] PERF .
b. 什么 太 [高 / 矮] 了 ？
what too [high / short] PERF ?
c. 架子 / 我
shelf / me

In some other cases, we decided to adapt the WS not because of the WS itself, but because of its NLI form. Consider for example the NO ENTAILMENT pair of the alternate version of WS (9), present as item 116 of the training portion of the WNLI section of GLUE and given in (10). While we agree that in the most natural interpretation of (10a), the pronoun *it* refers to the path and not the lake, it also seems to us (an intuition confirmed by some native speakers of English) that an entailment relation can naturally be seen in (10), the reasoning being that if the path to the lake was blocked so that we could not use it (the path), then we could not use the lake either. (Keep in mind that in the NLI task, the four possible pairs have to be considered *independently*.)

(9) a. The path to the lake was blocked, so we couldn't [reach/use] it.
b. What couldn't we [reach/use]?
c. the lake/the path

(10) a. The path to the lake was blocked, so we couldn't use it.
b. We couldn't use the lake.
c. NO ENTAILMENT

Nangia and Bowman (2019) report a human performance of 95.9% on the WNLI section of GLUE, which is composed of WS in their NLI form as (10). This number has to be interpreted with caution as it corresponds to the accuracy of a majority vote across five crowd-sourced annotations, which means that a given item was still considered correctly answered if two out of five (i.e., 40%) of the human annotators disagreed with the expected answer. This estimation does show, however, that 4.1% of the NLI items are more often than not interpreted in conflict with the "gold" annotation. Concerning (9), different adaptations seemed possible to fix this issue, that is why Mandarinograd contains two different WS inspired from (9) but based on two different forms of reasoning. The first one, translated in (11), simply describes in the alternate text the path as the first one. Because of this, we expect the reader to infer the existence of other paths, invalidating the troubling entailment relation discussed above. The other one, translated in (12), uses a different set of motion verbs instead.

(11) a. The [only/first] path to the lake was blocked, so we couldn't [reach/use] it.
b. What couldn't we [reach/use]?
c. the lake/the path

(12) a. The path to the lake was blocked, so we [stopped/couldn't get] there.
b. Where [did we stop at/couldn't we get to]?
c. the path/the lake

As a result, we translated or adapted all WS but one (number 84) into Chinese, four of them (numbers 72, 85, 99 and 135) in two different ways. For all non-trivial modifications, we provide an explanation alongside the WS. In addition, we added the new schema in (13). In total, Mandarinograd contains 154 WS.

(13) a. Dan 为 Bill 做 了 晚饭 , 因为
Dan for Bill make PERF dinner , because
他 打赌 [赢 / 输] 了 。
he bet [win / lose] PERF .
b. 谁 打赌 [赢 / 输] 了 ？
who bet [win / lose] PERF ?
c. Bill/Dan

The texts in Mandarinograd have been tokenised using Jieba[7] and then manually corrected (more details on the

[7] https://github.com/fxsjy/jieba

project's website). We kept the original Western proper names (e.g., *Dan*, *Bill*) during the translation process, but we also implemented a script to replace them with Chinese ones (e.g., 老王, 俊杰).

3.3. Validation

To ensure the quality of the WS, we then asked four native Chinese speakers to check our translations. These participants were fully literate in Chinese (holders of a master's degree or higher).

This checking process was built around a *testing* phase, using an electronic questionnaire similar to the one used by (Bender, 2015). This questionnaire consisted of a list of problems, each of them being composed of one version (special or alternate) of a WS, the corresponding question and the two answers (presented in randomised order). The goal of each problem was to select the correct answer to the question. Each participant had to answer either the special or the alternate version of 144 WS, but never both; the 10 remaining WS were used for the training phase described below. Two of the participants answered the special versions of the WS, the other two answered the alternate ones. The questionnaire was presented as different screens of 10 such problems. The participants had the possibility to interrupt the session and resume it at any time.

As none of the participants had any prior experience with WS, this testing phase was preceded by a *training* phase, providing a description of the task and interface as well as 10 problems. Contrary to what happened during the testing phase, when the participants answered these problems, the correct answer was shown right away and an explanation was provided. These 10 problems corresponded to the 10 first WS of Mandarinograd, that we considered uncontroversial. After having completed this testing phase, the participant had to complete the training phase, which was then followed by a *verification* phase.

The verification phase showed to each participant all problems for which their answer disagreed with the expected one. They were asked to reconsider each of them and to either check a box indicating they had made a mistake, or explain why they disagreed.

We thus collected 576 answers, 527 of which (91%) were in agreement with the expected answers before reconsideration. After reconsideration, the participants agreed on 556 (97%) problems, decreasing the number of disagreements to 20. We studied attentively all these conflicting items and the explanations provided in order to correct the WS when necessary. All modifications are documented in the dataset.[8]

4. Statistical association baseline

In this section, we show that the WS in Mandarinograd are resistant to simple corpus statistics. The baseline proposed by Amsili and Seminck (2017) achieves an accuracy of 55%.

4.1. Pointwise mutual information

As mentioned by Levesque et al. (2012), WS should be "Google-proof", which means that "there should be no obvious statistical test over text corpora that will reliably disambiguate these correctly". In this section, following Amsili and Seminck (2017), we propose to check the resistance of our dataset to a method based on *pointwise mutual information* (PMI).

The PMI of two events e_1 and e_2 is defined as follows:[9]

$$PMI(e_1, e_2) = \log_2\left(\frac{P(e_1 \cap e_2)}{P(e_1)P(e_2)}\right) \quad (14)$$

The events that we consider here are the presence of a given word in a window of fixed size s. For example, $P(e_{\text{win}})$ is the probability of the word *win* appearing in a natural sequence of s words in English. Similarly, $P(e_{\text{win}}, e_{\text{race}})$ is the probability of both words *win* and *race* appearing in the same window of length s. We can estimate these probabilities by counting in a corpus.

If the presence of two words w and w' are statistically independent, then $P(e_w \cap e_{w'}) = P(e_w)P(e_{w'})$ and $PMI(e_w, e_{w'}) = 0$. If, on the contrary, the two words tend to be found together, $P(e_w \cap e_{w'}) > P(e_w)P(e_{w'})$ and $PMI(e_w, e_{w'}) > 0$, while if they tend to *not* be found together, $P(e_w \cap e_{w'}) < P(e_w)P(e_{w'})$ and $PMI(e_w, e_{w'}) < 0$, with the extreme case $PMI(e_w, e_{w'}) = -\infty$ when the two words are never found together. $PMI(e_w, e_{w'})$ is a measure of the association between w and w' (Church and Hanks, 1990).

4.2. Method

To solve a given version of a WS, the idea exposed by Amsili and Seminck (2017) relies, when possible, on representing the two possible answers of the question by two single lemmas w_1 and w_2 (e.g., *suitcase* and *trophy*) and on selecting another lemma w (e.g., *small*) in order to estimate the association of this word, that we call here the *reference*, with the two answers as defined by PMI.[10] Then, whichever answer is the most associated with the reference is selected; when this process is not applicable, a random answer is given.[11] This algorithm might be simple but it is

[8] In addition to this validation process, we are interested in performing a large-scale experiment aimed at evaluating human performance on Mandarinograd, as done by Bender (2015) in English. However, because reaching native Chinese speakers (e.g., on crowd-sourcing platforms) appears significantly harder than English ones, we decided to leave this for future work.

[9] It seems that PMI is sometimes referred to as simply "mutual information" which might be confusing because the mutual information of two *random variables* is a related but not identical concept.

[10] We are working with Chinese, and as in English, there is no agreement between an adjective and the noun it modifies. Additionally, there is no agreement between a verb and its subject. Tense or mode are marked with particles that are not part of the verbs (at least not with the word segmentation scheme that we use).

[11] Because small differences in PMI might be unreliable, Amsili and Seminck (2017) experiment with a threshold t: the first answer is selected if $PMI(e_w, e_{w_1}) - PMI(e_w, e_{w_2}) > t$, the second answer if $PMI(e_w, e_{w_1}) - PMI(e_w, e_{w_2}) < -t$, and a random answer is given otherwise. They observe, however, that the best performance (55%) is obtained with $t = 0$, which corresponds to the method used here.

deliberately so and yet non trivial. WS have to be solvable by intelligent agents; they are just required to necessitate common sense reasoning in a relatively challenging way. Determining what aspects of language and what kinds of world knowledge are captured by recent language models such as the ones employed by Yang et al. (2019), Kocijan et al. (2019) or Trinh and Le (2018) (among others) is notoriously hard. These are highly engineered systems that cannot be considered as baselines in the sense relevant here. As mentioned by Amsili and Seminck (2017), association with arbitrary proper names (e.g., *Jane*) is irrelevant. As a consequence, to make our baseline stronger, we sometimes represented a given answer by a word that is not part of this answer but that is syntactically or semantically associated with it in the text. For example, in (15), we represented *Bob* by *pay* and so based the algorithm's decision on $PMI(e_{\text{generous}}, e_{\text{pay}}) - PMI(e_{\text{generous}}, e_{\text{Charlie}})$ for the special version of the WS and $PMI(e_{\text{grateful}}, e_{\text{pay}}) - PMI(e_{\text{grateful}}, e_{\text{Charlie}})$ for the alternate version.

(15) a. Bob paid for Charlie's college education. He is very [generous/grateful].
b. Who is [generous/grateful]?
c. Bob/Charlie

As a general rule, we hand-selected meaningful terms as often as possible, even when the question was invariant between the two versions of the WS. For example, in (16), we selected *stand* (resp. *sing*) as the reference for the special (resp. alternate) version of the WS. If *broken* instead had been selected, all selected terms being constant, the algorithm would have given the same answer for both versions of the WS and would then necessarily be right for one version and wrong for the other.

(16) a. Sam pulled up a chair to the piano, but it was broken, so he had to [stand/sing] instead.
b. What was broken?
c. the chair/the piano

Still, it was not always possible to find relevant terms. We rejected in particular cases for which the two versions only differ by the substitution of a discourse connective (as in (17)), of a preposition, or the addition of a word (as in (18)). In total, 29 WS were not assigned terms and were answered randomly by this baseline.[12]

(17) a. Ann asked Mary what time the library closes, [but/because] she had forgotten.
b. Who had forgotten?
c. Ann/Mary

(18) a. Jane gave Joan candy because she was [/not] hungry.
b. Who was [/not] hungry?
c. Joan/Jane

[12]In order to take into account modifiers (such as negation) which could intuitively lead to a better baseline, we also adapted this method to strings of characters rather than words directly. This is relatively easy to implement because in Chinese words are not explicitly segmented. We did not, however, obtain a stronger baseline this way.

To compute the pointwise mutual information values between pairs of terms, we estimated the corresponding probabilities by counting on the Chinese version of Wikipedia. The Chinese Wikipedia being stored using a mix of different scripts, we converted all articles to simplified script (the one used for Mandarinograd) using OpenCC.[13] We then word-segmented the result using Jieba. Roughly, this corpus contains 450 million characters, representing 250 million words.

4.3. Results

Out of the 2×154 WS versions of our corpus, the method described above allowed us to define PMI differences for 153 items.[14] On these 153 items, the algorithm had a precision of 61% (93 correct answers). Overall, answering randomly for other items, this baseline obtains an accuracy of 55%, which happens to be the same as the one obtained by Amsili and Seminck (2017) on their French collection. These numbers are clearly lower than the 91% and 97% agreement rates observed during the validation process, or the 92% human baseline determined by Bender (2015) on English WS. We consider the WS in Mandarinograd to be resistant in the sense discussed by Levesque et al. (2012).

5. Conclusion

WS represent hard cases of anaphora resolution problems, designed to require some forms of common sense reasoning. Because the availability of evaluation data is essential to the advance of AI, we have presented the first collection of WS in (Mandarin) Chinese. Such a dataset will allow cross-linguistic comparison as well as the evaluation of Chinese-specific systems. We have explained the various challenges we faced when translating and adapting the WS from English, in relation to the different forms (traditional or NLI) used by WS solvers in the literature. We have shown that the resulting collection was resistant to simple statistical methods, satisfying the requirements specified by Levesque et al. (2012).

6. Acknowledgements

We thank Hong Chen, Yifan Deng, Houqing Du and Lili Wang who accepted to participate in the validation process, as well as Goran Topić who implemented the questionnaire's web interface. This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

7. Bibliographical 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, April. Association for Computational Linguistics.

[13]`https://github.com/BYVoid/OpenCC`

[14]The PMI difference of an item $(w, (w_1, w_2))$ is undefined if neither w_1 nor w_2 co-occurs with w in the corpus, as both PMI are then negative infinities.

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.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. 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 107–112. Association for Computational Linguistics. event-place: New Orleans, Louisiana.

Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., and Lukasiewicz, T. (2019). A Surprisingly Robust Trick for Winograd Schema Challenge. *arXiv:1905.06290 [cs]*, May.

Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 552–561.

Melo, G., Imaizumi, V., and Cozman, F. (2020). Winograd Schemas in Portuguese. pages 787–798. SBC, January.

Nangia, N. and Bowman, S. R. (2019). Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy, July. Association for Computational Linguistics.

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*, EMNLP-CoNLL '12, pages 777–789, Stroudsburg, PA, USA. Association for Computational Linguistics. event-place: Jeju Island, Korea.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv:1907.10641 [cs]*, November.

Trinh, T. H. and Le, Q. V. (2018). A Simple Method for Commonsense Reasoning. *arXiv:1806.02847 [cs]*, June.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):443–460, October.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR 2019*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*, June.