Hard for humans = hard for machines? An analysis of the WSC data

Wiebke Petersen & Katharina Spalek
University of Düsseldorf

1 Introduction

The Turing Test has been introduced as a measure of a machine's ability to exhibit human-like intelligent behavior (Turing, 1950). However, the standard Turing Test, which involves a human evaluator engaging in a natural language conversation with a machine and attempting to distinguish its responses from those of a human, has been criticized for being too easy to pass as it may rely on the machine's ability to mimic human conversation. Thus, a machine could potentially pass the test without demonstrating any true cognitive abilities or language understanding (Weizenbaum, 1966).

In response to this, Levesque et al. (2012) proposed the use of sentences in the Winograd Schema as an alternative Turing Test which require machines to understand and reason about natural language statements that involve common sense knowledge and the ability to resolve ambiguous pronouns. The idea is that the Winograd Schema Challenge (WSC) can be used as a more challenging and reliable measure of a machine's ability to exhibit human-like intelligence than a conversational Turing test. Winograd Schema sentences come in pairs which differ only by a discriminatory segment that flips the correct referent of an ambiguous pronoun. Examples are given in (1) and (2):

(1) Winograd (1971, p. 441)
   a. I put the heavy book on the table and it broke.
   b. I put the butterfly wing on the table and it broke.

(2) Winograd (1972, p. 33)
   a. The city councilmen refused the demonstrators a permit because they feared violence.
   b. The city councilmen refused the demonstrators a permit because they advocated violence.

The original challenge consists in the task of answering a question asking the antecedent of the ambiguous pronoun given two answer choices corresponding to the noun phrases; e.g., for (1) 'What broke? The heavy book or the table?'

The sentences are chosen in such a way that the anaphora resolution is (a) non-ambiguous for humans (having a "natural reading") and (b) not solvable on the basis of selectional restrictions or (c) by matching based on frequency of co-occurrence. Condition (b) and (c) excludes examples as (3-a) and (3-b), respectively ('zooming by' and 'racecar' co-occur more frequently than 'zooming by' and 'school bus' and 'pills' cannot be 'pregnant').

(3)
   a. The women stopped taking the pills because they were pregnant/carcinogenic.
   b. The racecar zoomed by the school bus because it was going so fast/slow.
   c. The woman looked for a different vase for the bouquet because it was too small/big.

The original published Winograd Schema Challenge (WSC) dataset WSC273 consists of 273 manually constructed sentence pairs that are claimed to fulfill the above conditions. Since the WSC provides challenging problems for machine translation and anaphora resolution, it has become one of the tasks included in the state-of-the-art General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). Larger datasets in WSC-style have been compiled for training and testing language models. The WinoGrande dataset consists of 44k problems which are formulated as a fill-mask task by replacing the ambiguous pronoun with a blank that has to be filled with the appropriate referent (Sakaguchi et al., 2021). The data has been crafted and evaluated by crowdsourcing. One of the aims was to provide more challenging problems as the original WSC problems. However, examples like (3-c) are included that violate the non-ambiguity condition (a) as a vase can be too small or too big for a bouquet.

Our research questions are: (1) Does the WSC data family fulfill the central claim of being easy for humans but difficult for machines? (2) Is there a connection between WSC problems that are difficult for people and those that are difficult for machines? Such a connection would support the claim that machines show human-like behavior when defeating the Winograd Schema Challenge.

2 Experiments

2.1 Behavioral experiment

Materials. We selected fifty WSC-style sentence pairs. Twenty-five were German translations of original WSC273 problems using DeepL as a first approximation. In cases where the translation struck us as awkward or non-idiomatic, we slightly changed the wording. The other twenty-five pairs came from the German examples of Wino-X, a multilingual dataset that contains translations of WinoGrande problems (Emelin and Sennrich, 2021) and is designed for training translation models. All sentences were edited in such a way that the gap filler is required to be a nominative singular NP and that the two potential referents have different grammatical gender in German. Instead of asking participants which referent the pronoun refers to, we asked them to choose the correct pronoun for a gap in the text. (4) gives the German translation of Example (1-a) with a gap in the place of the pronoun.

(4) Ich legte das schwere Buch auf den Tisch und _____ zerbrach.

If participants put the masculine gender pronoun *er* in the gap, we can infer that they reasoned that the table broke. If participants put the neuter gender pronoun *es* in the gap, we can infer that they assume that the book broke.

Design and Procedure. We created ten experimental lists. Each list contains ten WSC problems presented as in (4). For a given problem, only one version was included in a list. Fifteen filler items were included to prevent participants from detecting the logic behind the WSC-problems. As filler items we used sentences with only one possible antecedent and varied whether the antecedent has a fixed gender in German or not.

For each problem (including the filler items), participants have to choose which of the three German pronouns *er*, *sie*, *es* belongs in the gap. After their decision, participants have to indicate on a scale from 1 to 5 how confident they are that their decision was correct.

We reason that hard choices can now be reflected by higher variance in inter-individual accuracy for a given problem and lower intra-individual confidence ratings. In addition, we measured how long it took participants to choose a pronoun. Reaction times tend to increase if a cognitive task is more difficult.

The experiment is carried out online. It was programmed in PsychoPy (Peirce, 2007) and is distributed with the platform Clickworker.

Participants We will test 50 participants such that we will have 5 responses per (version of a) WSC problem. Participants are native speaker of German between the age of 18 and 55. We will test equal numbers of men and women. Testing is on-going, at the time of writing the abstract, 34 participants have already completed the online study.

2.2 Language model experiments

The same task as in the behavioral experiment was given to a BERT (Devlin et al., 2018) and a GPT (Radford et al., 2018) language model. Standard general language models without specialization on coreference resolution have been chosen, as we are interested in how machines handle challenging anaphora problems without being fine-tuned on such tasks.

BERT model. For BERT, we use the checkpoint 'bert-base-german-uncased' from HuggingFace and the standard fill-mask task, i.e. the task to predict a masked token in a sentence. This type of task was part of BERT's original training paradigm. We apply softmax to the returned logits over BERT's entire vocabulary (31.102 tokens) and interpret the softmax value of a token as its probability to be chosen and as the confidence of the machine that this token is the correct filler for the masked token. For each item we collect the top scoring 10 predictions. Using handcrafted lists we match these top predictions by their gender to the classes *fem*, *masc*, *neut* and to the class *other* and sum the softmax values for each class

up. Thereby we allow not only for the gap to be filled by the personal pronouns 'er', 'sie', 'es', but also by corresponding demonstrative pronouns, proper nouns and the like.

GPT model. For GPT we use the GPT-3.5 'text-davinci-003' from the OpenAI playground (temperature = 0.3). As instruction prompt we use the instructions from the behavioral experiment shortened by information concerning only the use of the online experiment.

3 Results and Discussion

Experiments. Since only two third of the intended participants of the behavioral study have yet done the experiment, data analyses are in a preliminary state at the moment. Currently the human accuracy (percentage of correct answers) is 74% and 76% if majority voting[1] is applied. If full agreement of all participants on the correct response is required, the accuracy rate drops to 47%. Comparing the accuracies for the WSC273 and the Wino-X items separately, the accuracy for WSC273 is 85% and for Wino-X 62%. Thus the Wino-X data is harder than the WSC273 data. These are a strikingly low values compared to the reported 96.5% accuracy for the original WSC problems and 94% for WinoGrande (Sakaguchi et al., 2021). However, the latter was based on majority voting of three online participants, no filler items were used and participants had to complete a training phase first. Our results show that humans struggle more if they are confronted with WSC problems intermixed with filler problems.

In a first descriptive analysis, we compared confidence ratings for problems grouped by accuracy. For those cases where all participants gave the wrong response, the average rating was 4.10 (sd = 1.07) compared to those cases were all participants gave the right response (mean = 4.44, sd = 0.91). Thus, problems that are hard (as evidenced by low accuracy), were also replied to less confidently. We also observed a significant correlation between the confidence ratings and the reaction times (r = -0.21, p < 0.01). That is, when humans felt less confident (i.e., ratings went down), it took them longer to make a decision (i.e., reaction times went up).

Looking at the machine models we found an accuracy score of 56% for BERT and of 48% for GPT. The former is remarkable high as Wang et al. (2018) report that non-finetuned BERT models do not outperform most-frequent-class guessing (here 42%). Note that the GPT results have been obtained without any prompt engineering. Again higher accuracies are found for the WSC273 data (BERT: 60%, GPT: 66%) than for the Wino-X data (BERT: 52%, GPT: 30%). Looking at BERT's confidence in its decision (i.e. softmax values) no difference can be found between correct and incorrect predictions (correct: 0.755, incorrect: 0.753).

In order to investigate whether problems that are hard for humans are also hard for machines, we looked at the correlation of human accuracy, ratings and reaction times with BERT's confidence in the predicted and the correct fillers (i.e. softmax values). The following significant correlations have been found: (a) Human accuracy correlates with Bert's confidence in the correct filler (r = 0.33, p < 0.01). Thus if humans tend to give the wrong answer, BERT has low confidence in the correct one and vice versa. Thus the more difficult a problem is for humans the more difficult it is for BERT as well. (b) Concerning BERT's confidence in its prediction we found a correlation with the reaction times of humans: The higher the machine confidence score, the lower the human reaction time or the longer humans need to come to a decision the less confident BERT is in its decision. (r = -0.26, p < 0.01). Put differently, for those problems where the machine made the prediction with high confidence, a human could make the decision quickly, that is, with little cognitive effort. However a significant correlation between human confidence ratings and BERT's confidence in its predictions could not be found.

If we group the problems by human accuracy and focus on those problems that were consistently solved incorrectly by humans we find very low machine accuracy values (GPT: 0.0, BERT: 0.2). That indicates that machines and humans struggle with the same problems. For the problems solved by all participants correctly we find higher machine accuracy values (GPT: 0.68, BERT: 0.64), i.e., machines struggle with some problems that are easy for humans.

Data inspection: error analysis. A look into the data that turned out to be problematic for humans as well as for machines exhibits that some of the Wino-X data are not carefully crafted. Two examples will be discussed:

[1] An item is said to be correctly answered if the majority of participants answered it right.

(5) 3 Autos konnten in der Garage parken, aber nur 2 im Carport, da _____ kleiner / größer war.
3 cars could park in the garage but only 2 in the carport because it was smaller / larger.

For example (5) all humans have given an answer that is considered 'wrong' in the Wino-X dataset. In both versions all humans (and 'GPT and BERT) answered 'es', while according to Wino-X the correct answer would be 'er' (referring to the carport) for 'kleiner' and 'sie' (referring to garage) for 'größer'. However, the sentence is problematic for two reasons: First, 'Carport' is a loanword in German without fixed gender; some dictionaries list neuter as alternative gender. Second, car is neuter in German and the size of cars could be the reason for the parking situation as well.

For the second example, the translation from Wino-X was incorrect and we used a corrected version in our experiments:

(6) Clara beschloss, Gemüse im Ofen anstatt ~~auf~~[in] der Mikrowelle zu kochen, weil _____ das Gemüse ~~feuchter~~ [saftiger] / knuspriger schmecken ließ.
Clara decided to boil vegetables on the stove instead of the microwave because it made the vegetables taste soggier / crunchier.

Again, none of the humans got these examples correct according to Wino-X ('sie' referring to microwave for 'saftiger'; 'er' referring to stove for 'knuspriger'). The common human answer was 'es' (also chosen by GPT). Here the problem is, that neuter 'es' can refer to the whole cooking situation.

Discussion. We believe that two conclusions merit discussion:
(a) The premise of the Winograd Schema Challenge is problematic: The original assumption was that the problems are non-ambiguous for humans but challenging for machines. However, as we have started to show with our data, some of the problems are difficult or even unsolvable for human language users.
(b) At the same time, the 'problem' described above might make the WSC an even better alternative Turing test. If it turns out that not all WSC problems are equally easy for humans AND that those problems humans struggle with are the same ones machines struggle with, this meets the requirement that a machine is able to act like a human even better than a perfect score for the WSC. The finding that measures for human effort and machine effort correlate suggests that this is a useful avenue for future investigations.

References

J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

D. Emelin and R. Sennrich. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of EMNLP 2021*, pages 8517–8532. ACL, 2021.

H. J. Levesque, E. Davis, and L. Morgenstern. The Winograd Schema Challenge. In *Proceedings of KR'12*, page 552–561. AAAI Press, 2012.

J. W. Peirce. PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, 162(1-2): 8–13, 2007.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training, 2018. URL https://openai.com/research/language-unsupervised.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

J. Weizenbaum. ELIZA -— a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

T. Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, Massachusetts Inst Of Tech Cambridge Project Mac, 1971.

T. Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972.