Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Investigating associative, switchable and negatable Winograd items on renewed French data sets

Xiaoou Wang[1] Olga Seminck[2] Pascal Amsili[2]
(1) CENTAL, Place Montesquieu 3, 1348 Louvain-la-Neuve, Belgium
(2) Lattice, 1 rue Maurice Arnoux, 92120 Montrouge, France
xiaoouwangfrance@gmail.com, olga.seminck@cnrs.fr, pascal.amsili@ens.fr

RÉSUMÉ

Le Winograd Schema Challenge (WSC) regroupe des problèmes de résolution d'anaphore nécessitant un raisonnement sur les connaissances du monde. Cet article décrit la mise à jour des items français existants et la création de trois sous-ensembles permettant une évaluation plus robuste et plus fine du WSC en français (FWSC) : un sous-ensemble *associatif* (items pouvant être résolus avec de la simple co-occurrence lexicale), un sous-ensemble *commutable* (items où l'inversion de mots-clés inverse la réponse) et un sous-ensemble *niable* (items où l'application d'une négation inverse la réponse). Sur ce jeu de données, nous obtenons des performances SOTA grâce à l'utilisation de CamemBERT. Notre protocole d'évaluation montre par ailleurs que cette performance peut être expliquée par l'existence d'items associatifs et que si augmenter la taille du corpus d'entraînement améliore la capacité du modèle à traiter les items commutés, cela affecte peu la performance sur les items niés.

ABSTRACT

The Winograd Schema Challenge (WSC) consists of a set of anaphora resolution problems resolvable only by reasoning about world knowledge. This article describes the update of the existing French data set and the creation of three subsets allowing for a more robust, fine-grained evaluation protocol of WSC in French (FWSC) : an *associative* subset (items easily resolvable with lexical co-occurrence), a *switchable* subset (items where the inversion of two keywords reverses the answer) and a *negatable* subset (items where applying negation on its verb reverses the answer). Experiences on these data sets with CamemBERT reach SOTA performances. Our evaluation protocol showed in addition that the higher performance could be explained by the existence of associative items in FWSC. Besides, increasing the size of training corpus improves the model's performance on switchable items while the impact of larger training corpus remains small on negatable items.

MOTS-CLÉS : Schémas Winograd, connaissances du monde, inférence automatique, négation, français, CamemBERT.

KEYWORDS: Winograd Schema Challenge, world knowledge, commonsense reasoning, negation, French, CamemBERT.

1 Introduction

A Winograd schema (Levesque, 2011) consists of two anaphora resolution problems (items) differing by two keywords (*successful/available* in (1)) which change the answer to a question targeting the referent (*Paul* and *George*) of an ambiguous anaphor (*he*). A Winograd item is supposed to be Google-proof or non associative, meaning that it should be insensitive to simple statistics such as lexical co-occurrence [1]. The idea behind this challenge is that if a system is capable of solving these schemas, it should be capable of commonsense reasoning and hence, could be called "intelligent".

(1) Paul tried to call George on the phone, but *he* wasn't *successful/available*.
Question : Who wasn't *successful/available* ? Response : Paul/George

Since the launch of the first Winograd Schema Challenge (Morgenstern *et al.*, 2016), the reference data set in English has evolved from 273 (WSC273) to 285 items (WSC285). [2] A large variety of methods have been explored to resolve the WSC, including logical formalisms (e.g. Bailey *et al.*, 2015), information retrieval approaches (e.g. Emami *et al.*, 2018), neural networks (e.g. Liu *et al.*, 2017) and neural language models (e.g. Trinh & Le, 2018). The recent literature is dominated by the use of large pretrained language models such as GPT (Radford *et al.*, 2019). SOTA performance has been reached by Sakaguchi *et al.* (2020), who fine-tuned RoBERTa (Liu *et al.*, 2019) on a large data set they crowd-sourced (WinoGrande) and achieved 90.1% accuracy on WSC273, *vs.* 92.1% for humans (Bender, 2015) and 50% at random.

Despite the accuracy of models which is now close to human performance, it can be questioned whether systems have become truly capable of commonsense reasoning. Trichelair *et al.* (2019) found that not all items are equally robust and categorized items into two subsets : associative and switchable. An associative item is an item where the correct answer can be deduced by solely looking at the clause containing the pronoun/possessive adjective (*but he wasn't successful/available* in (1)). In a switchable item, the referents can be switched (*Paul* and *George* in (1)), causing the correct answer to shift accordingly. It was demonstrated that the then-state-of-the-art performance by Trinh & Le (2018) was mainly due to the simpler associative subset and on the other hand, insensitive to the switching operation. The evaluation on separate subsets makes it possible to investigate the performance of a system on difficult items and, moreover, tests the robustness of a model's decisions when items are slightly modified.

When the French version of WSC (214 items, henceforth FWSC214) was developed (Amsili & Seminck, 2017a), it was found that some items were associative. However, it was considered that this would not be of much influence, as a system trying to exploit this feature could obtain at most 55% accuracy, *vs.* a human baseline of 93.6% (Amsili & Seminck, 2017b). A more recent approach (Seminck *et al.*, 2019) using small pretrained language models (without fine-tuning on WSC-problems) also pointed at some associativity in the data set, but failed on a large number of other items, performing only at 52% accuracy.

The aim of the present work is to get a better understanding of Winograd items by dividing FWSC into three subsets. After transforming FWSC214 to FWSC285 based on the most recent version of WSC [3], we identified an associative and a switchable subset inspired by Trichelair *et al.* (2019). Moreover,

[1] Even though nowadays Google-proofness and associativity are often taken to refer to the same property, there is still a difference in the methods used to ensure Google-proofness *vs.* non associativity. In the first case, the idea is that counting co-occurrences in a corpus shouldn't suffice to choose the appropriate answer —for instance in the schema *A tree fell on the roof, we'll have to remove/fix it*, any search in a corpus will give a higher co-occurrence count to the pair *roof/fix vs. roof/remove* ; in the second case the idea is that a speaker hearing only the question and the possible answers will be biased towards one of the answers. The bias may come from lexical co-occurrence or from world knowledge.

[2] `https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html`, consulted February 23, 2022.

[3] We extended FWSC214 based on English items in the spirit of ensuring a certain comparability with the English data set, French-only items would be added in the future.

we proposed a new *negatable* subset which can help test a model's sensitivity to negation. Then, we fine-tuned the CamemBERT model (Martin *et al.*, 2020) on a machine-translation of WinoGrande and compared its performance on our three different subsets. [4]

2 Update of the current French data set

In this section, we will explain how we first adapted the existing FWSC214 to FWSC285 and how we created the associative, switchable and negatable subsets.

2.1 From FWSC214 to FWSC285

12 items were first removed from FWSC214 because they were neither in WSC273 nor in WSC285. Minor modifications were then made so that the items were closer to the English version. For example, "Nicolas" was replaced by "L'homme" in (2) to better match the English version. We also modified certain items to improve their naturalness. For instance, we changed "était" (*was*) to "avait l'air" (*looked*) in (3) since it is odd to assume a fish's feelings.

(2) FWSC214 : Nicolas n'a pas pu soulever son fils car il était trop faible/lourd.
WSC285 : The man couldn't lift his son because he was so weak/heavy.
FWSC285 : L'homme n'a pas pu soulever son fils car il était trop faible/lourd.

(3) FWSC214 : Le poisson a mangé le ver. Il était affamé/délicieux.
FWSC285 : Le poisson a mangé le ver. Il avait l'air affamé/délicieux.

Then, a total of 83 new items were translated from WSC285, with significant adaptations of 13 items. Since an item is only valid if the candidate answers have the same number and gender, we had to replace some answers with nouns of opposite gender as in (4) where "le plateau de théâtre" (*the stage* masculine) was chosen instead of "la scène" (feminine). Besides, some oppositions in English do not have an equivalent in French. In these cases, we established opposition on other elements as in (5).

(4) WSC285 : There is a pillar between me and the stage, and I can't see/see around it.
FWSC285 : Il y a un pilier entre moi et le plateau de théâtre, et je n'arrive pas à le voir/contourner.

(5) WSC285 : They broadcast an announcement, but a subway came into the station and I couldn't hear/hear over it.
FWSC285 : Ils ont diffusé une annonce quand une voiture est arrivée dans le parking souterrain. La voiture/l'annonce était trop bruyante et je n'ai pas pu l'entendre.

Nuances related to some English verbs are also hard to translate. (6) gives an example where Shakespeare can either refer to the author or his writings depending on the birthdate of the other

[4] All our data sets and subsets, along with the machine-translated WinoGrande, are available at https://github.com/xiaoouwang/FWSC285.

author. We adapted the item by establishing a new opposition based on word order. The last type of problems are related to naturalness.

(6) This book introduced Shakespeare to Ovid/Goethe ; it was a major influence on his writing.
Adaptation : Ce livre a fait découvrir (Ovide à Shakespeare)/(Shakespeare à Ovide) ; il a eu une influence majeure sur son écriture.

All the new items were first translated with DeepL[5]. The main author of this work made a first adapted version which was in turn improved and validated by a native French speaker who speaks also English. Finally, a monolingual native French speaker was consulted to improve the naturalness of the translated items without changing the meaning.

2.2 Associative, switchable and negatable subsets

We designed a psycholinguistic questionnaire to determine which items are associative. Human participants were presented only the question and possible answers and were asked whether, with no context, one answer seemed more likely than the other. Participants were explicitly instructed to look for biases with the possibility of answering "no bias" (7). While Trichelair *et al.* (2019) considered only the association between a keyword and the right answer, our design differentiates positive and negative associativity, demonstrated respectively by (8-a) and (8-b).

(7) Qu'est-ce qui est trop grand ? (*What is too large ?*)
1. la coupe (*the trophy*) 2. la valise (*the suitcase*) 3. pas de biais (*no bias*)

(8) a. Positively associative : Qu'est-ce que je dois réparer ? (*What should I repair ?*)
Correct answer : le toit (*the roof*) Wrong answer : l'arbre (*the tree*)
b. Negatively associative : Qu'est-ce qui avait l'air délicieux ? (*What looks delicious ?*)
Correct answer : le ver (*the worm*) Wrong answer : le poisson (*the fish*)

Among our 42 participants, two were excluded because the average response time was too short ($< 1''$). The experiments lasted 23 to 33 minutes. Figure 1 shows the distribution of items according to the percentage of subjects answering correctly (which we consider as correlated to associativity) and the isolated right cluster indicates clearly the existence of positively associative items for which more than 76% of the subjects chose the right answer without looking at the context. Using the same method, we identified 3 negatively associative items where more than 81% of the participants chose the wrong answer. We also established a switchable subset of 141 items and a negatable subset of 38 items. An item is negatable if a verb in the item can be negated without inducing semantic awkwardness. Naturally, both switching and negation should also induce an answer switch. We consulted two native French speakers and an item is only classified as switchable or negatable if both speakers come to an agreement. (9) shows these two operations on one item.

(9) a. Le bus scolaire a dépassé le scooter, car il roulait trop vite.
(*The school bus surpassed the scooter because it was going too fast*)
Switched version : Le scooter a dépassé le bus scolaire, car il roulait trop vite.

[5] https://www.deepl.com/translator

(*The scooter surpassed the school bus because it was going too fast.*)

b. Negated version : Le scooter n'a pas dépassé le bus scolaire, car il roulait trop vite. (*The school bus didn't surpass the scooter because it was going too fast.*)

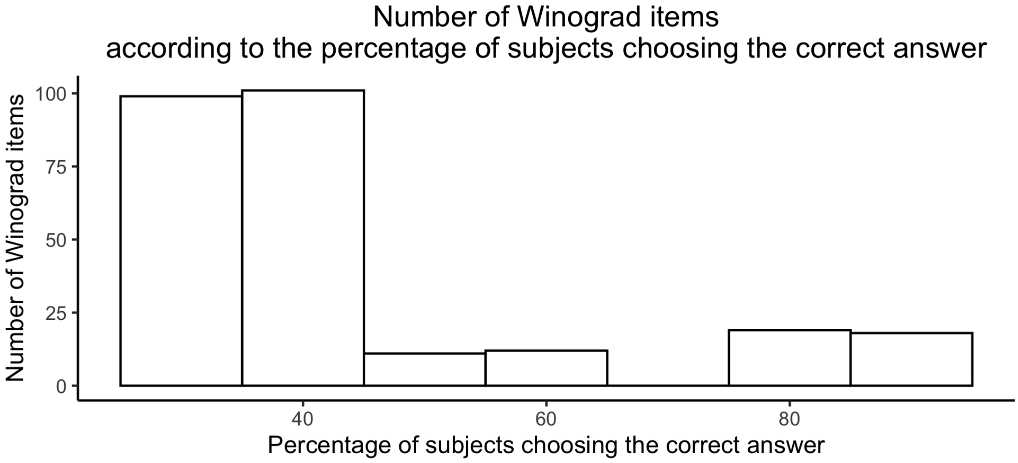

FIGURE 1 – Number of Winograd items *vs.* percentage of subjects choosing the correct answer.

3 Experiments with CamemBERT on FWSC285 and its subsets

We fined-tuned *CamemBERT large* on the machine-translated version of WinoGrande. Concretely, we reframed the WSC task as a sentence classification problem by replacing the pronoun with the two candidate answers. The correct sentence is marked 1 and the wrong one 0 (10) :

(10) Original item : Claire a frappé à la porte de Sylvie, mais elle n'a pas eu de réponse.
1 → Claire a frappé à la porte de Sylvie, mais Claire n'a pas eu de réponse.
0 → Claire a frappé à la porte de Sylvie, mais Sylvie n'a pas eu de réponse.

We used DeepL to translate WinoGrande into French. The first issue of this approach is that the second occurrence of the answer candidate (*Claire* or *Sylvie* in (10)) is sometimes translated back to a pronoun. It is also important to note that the translated corpus does not comply strictly with the definition of WSC items, because DeepL may not translate the same word consistently. For instance, *because* can be translated *parce que* in one sentence and *car* in another sentence. Since WinoGrande training corpora come in five sizes, from xs (160 items) to xl (40 938 items), included into one another, we used all these partitions in various experiments on our data sets (Table 1).

As shown by Table 1, we achieved 68% accuracy on FWSC285 and 66% on FWSC214 with the xl training set. Our model, using a simple classification protocol, achieved thus a new SOTA performance. After testing on the entire FWSC285, we ran the best model (trained on xl) on the positively associative and non-associative subsets of FWSC285. The best model scored 90% accuracy on the positively associative subset, while a test on 10 random samples of the same size of the positively associative subset, all drawn from the non-associative items, yielded only 59% accuracy on average. These experiments highlight the significant contribution of the associative subset to the overall performance, observed also in studies regarding WSC in English (Trichelair *et al.*, 2019). It is also worth noting

training size	FWSC285	positively associative	non-associative	unswitched	switched	unnegated	negated
xs (160)	51%	-	-	51%	49%	50%	50%
s (640)	60%	-	-	61%	57%	58%	52%
m (2 558)	66%	-	-	66%	61%	64%	56%
l (10 234)	68%	-	-	66%	63%	63%	56%
xl (40 938)	68%	90%	59%	67%	67%	64%	55%

TABLE 1 – Accuracy on FWSC285 depending on the data set and the size of training set

that the best model fails on all the 3 negatively-associative items. Although the sample is too small to suggest that the model has simply used co-occurrence as main cues to tackle FWSC, it would be interesting to build more negatively associative items in the future to test the validity of this hypothesis.

We evaluated further the model's robustness against perturbations on the switchable and negatable subsets. It can be seen from Table 1 that the accuracy for the switched subset improves consistently when the train set size is enlarged, while the score for the negated subset remains unchanged beyond the medium size training set. Besides, the accuracy for the negated subset is significantly lower than the unnegated subset, even when more and more training data are used. This highlights the interest of our negatable subset. Although enlarging the size of our training corpus does improve the robustness of our model to the switching operation, probably because a large amount of data facilitates a more abstract and general representation of the candidate answers. The sensitivity to negation, however, doesn't increase even when the largest corpus is used. This insensitivity is reminiscent of the study of Ettinger (2020) where BERT, in a zero-shot setting, fails to understand negation in a cloze task (fill-in-the-blank sentences). Since a robust commonsense reasoning system must include the understanding of negation, it would be interesting to build more negatable items in the future to test models on their ability to understand negation.

4 Discussion

Although we achieved SOTA performances on FWSC214 (66%) and FWSC285 (68%), these performances are expected since we used a SOTA pretrained language model fine-tuned on a large corpus while previous studies on FWSC used models trained on much smaller corpora and could not leverage the power of transfer learning. Besides, our performances were partly due to the existence of associative items. The main contribution of the present work is thus the update of current Winograd items in French and the creation of three subsets allowing for a more robust evaluation protocol of the Winograd Schema Challenge. More advanced metrics could be derived using our subsets, such as the *group-scoring* method used by Elazar *et al.* (2021) which assigns a point only if both items of one schema get solved. In our case, an item could be considered as solved only if its switched and negated versions (if they exist) get solved as well. The negatable subset of Winograd items seems particularly challenging, it would be interesting to see if models enhanced with information about the syntactic structure of items (Xu *et al.*, 2021, e.g.) can perform better. It is also worth noting that both operations (switching and negation) proposed in this study lead to an answer switch. Abdou *et al.* (2020) proposed seven perturbations (tense switch, number switch, etc.) which didn't alter the answer and showed that language models were more likely to switch the answer in case of some perturbations (e.g., number or gender alternations) than humans. Similar data sets could be created in French to allow further investigations into humans' and language models' sensitivity to linguistic perturbations.

One major limitation related to our training process is the quality of the fine-tuning corpus. It is difficult to know if the insensitivity to negation observed in Table 1 is due to our training strategy or if the general quality of our corpus hinders the model from learning. It is thus necessary to build a pretraining corpus of higher quality either by improving the current machine-translated items, or by designing a crowdsourcing procedure as Sakaguchi *et al.* (2020).

A final note concerns the WSC task itself. With the quasi-human performance achieved on WSC, a debate has been raised on whether the challenge has been defeated or not (Kocijan *et al.*, 2022). However, the same performance is far from being reached on FWSC (the best performance is 68% in our work, *vs.* 93.6% achieved by humans (Amsili & Seminck, 2017b)). Also, Elazar *et al.* (2021) raised the question of whether the commonsense reasoning ability is inherent to the language model or learned during the fine-tuning process and called for more studies using a zero-shot setting. We'd like to point out that using fine-tuning or not, the same evaluation protocol is always necessary to test the robustness of a model's decisions.

Références

ABDOU M., RAVISHANKAR V., BARRETT M., BELINKOV Y., ELLIOTT D. & SØGAARD A. (2020). The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7590–7604, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.679.

AMSILI P. & SEMINCK O. (2017a). A Google-Proof Collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, p. 24–29, Valencia, Spain : Association for Computational Linguistics.

AMSILI P. & SEMINCK O. (2017b). Schémas Winograd en français : une étude statistique et comportementale (Winograd schemas in French : a statistical and behavioral study). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts*, p. 28–35, Orléans, France : ATALA.

BAILEY D., HARRISON A., LIERLER Y., LIFSCHITZ V. & MICHAEL J. (2015). The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

BENDER D. (2015). Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*, p. 39–45.

ELAZAR Y., ZHANG H., GOLDBERG Y. & ROTH D. (2021). Back to square one : Artifact detection, training and commonsense disentanglement in the winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10486–10500.

EMAMI A., TRISCHLER A., SULEMAN K. & CHEUNG J. C. K. (2018). A generalized knowledge hunting framework for the winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 25–31.

ETTINGER A. (2020). What BERT Is Not : Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, **8**, 34–48. DOI : 10.1162/tacl_a_00298.

KOCIJAN V., DAVIS E., LUKASIEWICZ T., MARCUS G. & MORGENSTERN L. (2022). The defeat of the Winograd Schema Challenge. arXiv, 2201.02387.

LEVESQUE H. J. (2011). The Winograd Schema Challenge. In *Papers from the 2011 AAAI Spring Symposium*, volume Technical Report SS-11-06, p. 63–68.

LIU Q., JIANG H., LING Z.-H., ZHU X., WEI S. & HU Y. (2017). Combing Context and Commonsense Knowledge Through Neural Networks for Solving Winograd Schema Problems. In *AAAI Spring Symposia*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

MORGENSTERN L., DAVIS E. & ORTIZ JR. C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, **37**(1), 50–54.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

SAKAGUCHI K., LE BRAS R., BHAGAVATULA C. & CHOI Y. (2020). WinoGrande : An Adversarial Winograd Schema Challenge at Scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(05), 8732–8740.

SEMINCK O., SEGONNE V. & AMSILI P. (2019). Modèles de langue appliqués aux schémas Winograd français (Language Models applied to French Winograd Schemas). In *Actes de La Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles Courts*, p. 343–350.

TRICHELAIR P., EMAMI A., TRISCHLER A., SULEMAN K. & CHEUNG J. C. K. (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)*, p. 3382–3387, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1335.

TRINH T. H. & LE Q. V. (2018). A simple method for commonsense reasoning. *arXiv preprint arXiv :1806.02847*.

XU Z., GUO D., TANG D., SU Q., SHOU L., GONG M., ZHONG W., QUAN X., JIANG D. & DUAN N. (2021). Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 5412–5422.