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

Xiaoou Wang[1], Olga Seminck[2], Pascal Amsili[2]

1) CENTAL, Université catholique de Louvain
2) Lattice, CNRS, ENS & Université Sorbonne Nouvelle

TALN 2022, Avignon

Exemple d'un schéma Winograd

Simon a expliqué sa théorie à Marc, mais il ne l'a pas convaincu.
Qui n'a pas convaincu l'autre ?
A : Simon
B : Marc

Simon a expliqué sa théorie à Marc, mais il ne l'a pas compris.
Qui n'a pas compris l'autre ?
A : Simon
B : Marc

Simon a expliqué sa théorie à Marc, mais il ne l'a pas ⟨convaincu/compris⟩.
Qui n'a pas ⟨convaincu/compris⟩ l'autre ?
R0 : Simon
R1 : Marc

Histoire des Winograd Schema pour l'anglais

Une alternative au Test de Turing (Levesque et al., 2012)

- Première version de la collection en anglais (273 items)
- Besoin de raisonnement et de connaissances encyclopédiques
- 2016 : premier Winograd Schema Challenge (Morgenstern et al., 2016)
  - *Résultats au niveau de la chance*
- 2018 : amélioration grâce aux systèmes basés sur des LMs
  - *Trinh and Le (2018): 14 LMs, 64% d'exactitude*
  - *Radford et al. (2019): using GPT-2, 70% d'exactitude*
- 2020 : obtention de 90% par Sakaguchi et al. (2020)
  - *Corpus d'entraînement de 44k items 'à la Winograd'*
  - *Finetuning de RoBERTa*
- 2022 : Fin de la tâche ? (Kocijan et al., 2022)

Histoire des Winograd Schema pour le français

- 2017 : Adaptation des schémas au français
  Test de 'Google Proofness' par information mutuelle
  Mesure de performance humaine (93 %)
  (Amsili and Seminck, 2017a,b)
- 2019 : Tentative de résolution par LMs (Seminck et al., 2019)
  - *Scores qui ne dépassent pas la baseline de la chance (petits modèles)*

Simon a expliqué sa théorie à Marc, mais [Simon] ne l'a pas convaincu.
Simon a expliqué sa théorie à Marc, mais [Marc] ne l'a pas convaincu.

Entre temps... questionnement sur ce que les LMs apprennent

Est-ce que les modèles sont réellement intelligents ? Ont-ils appris à raisonner ?

- 2019 : Trichelair et al. (2019) trouvent
  - *qu'il y a 13% d'items associatifs pour les données en anglais*

    Un arbre est tombé sur le toit, il va falloir le réparer.
  - *que la réponse des LMs n'est pas forcément cohérente quand on inverse les antécédents potentiels*

    Original : Emma did not pass the ball to Janie although she saw that she was open.
    Switched : Janie did not pass the ball to Emma although she saw that she was open.
- 2020 : *Emami et al. (2020) trouvent qu'il y a un large chevauchement entre le test de Winograd en anglais et les items dans le corpus d'entraînement utilisé pour le fine-tuning.*
- 2021 : *Elazar et al. (2021) proposent de nouvelles métriques d'évaluation et trouvent que sans beaucoup de données d'entraînement, la performance des LMs est au niveau de la chance.*

Objectifs de notre étude

- Rendre les données du français plus semblables à celles de l'anglais pour avoir une correspondance item à item et étoffer la collection
- Catégoriser les items selon des caractéristiques (associatif, commutable et **négatable**)
- Tester la méthode état de l'art pour l'anglais (fine-tuning d'un modèle transformer) sur les données françaises
- Évaluer la performance du modèle selon nos trois catégories

Mise à jour des données

- 214 items (FWSC214) $\Rightarrow$ 285 (FWSC285)
- Plus proche de la collection anglaise (WSC285)

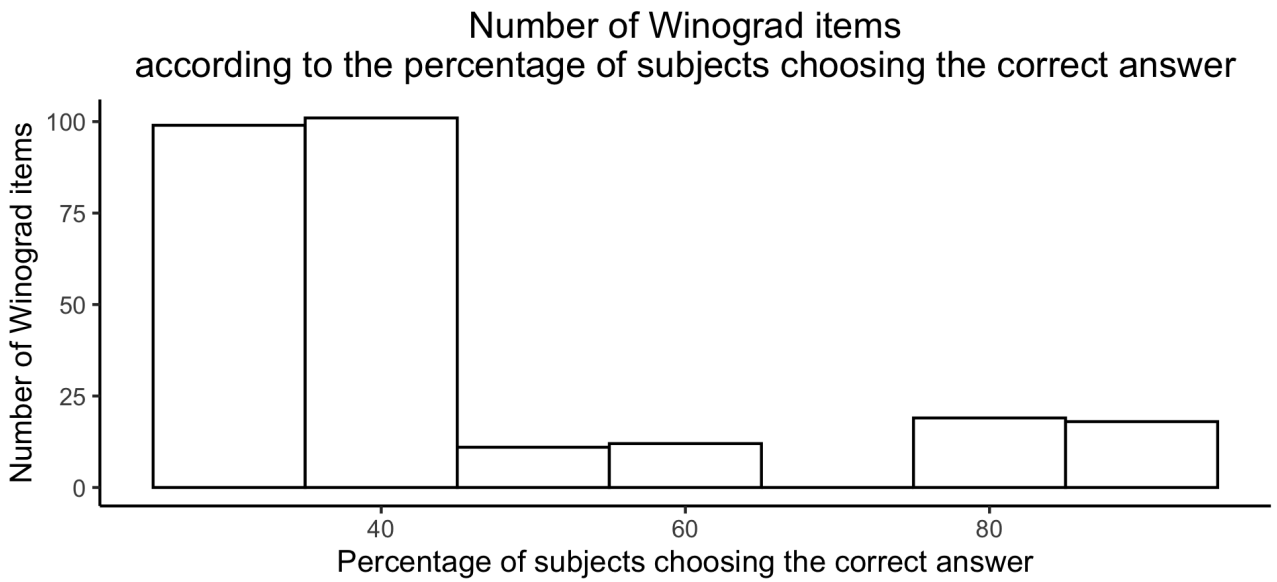
FWSC214 :
Nicolas n'a pas pu soulever son fils car [il] était trop ⟨faible/lourd⟩.
WSC285 :
The man couldn't lift his son because [he] was so ⟨weak/heavy⟩.
FWSC285 :
L'homme n'a pas pu soulever son fils car [il] était trop ⟨faible/lourd⟩.

- Tous les items ont d'abord été traduits par DeepL
- Ensuite adaptés par Xiaoou
- Ensuite validés par une linguiste locutrice native du français parlant anglais et français
- Validés enfin par une locutrice native monolingue pour garantir la naturalité des items

Catégorisation des items : associatifs

Qu'est-ce que je dois réparer ?
Bonne réponse : le toit Mauvaise réponse : l'arbre

- La réponse est associée sémantiquement à la question
- Expérience psycholinguistique (n = 40)
- 37 items associatifs



Catégorisation des items : commutatifs

Simon a expliqué sa théorie à Marc, mais il ne l'a pas ⟨convaincu/compris⟩.
Qui n'a pas ⟨convaincu/compris⟩ l'autre ?

R0 : Simon
R1 : Marc

Marc a expliqué sa théorie à Simon, mais il ne l'a pas ⟨convaincu/compris⟩.
Qui n'a pas ⟨convaincu/compris⟩ l'autre ?

R0 : Marc
R1 : Simon

- Classification d'items comme commutables si deux annotateurs natifs du français considéraient que cela ne rendait pas l'item sémantiquement discutable.
- 141 items commutables

Catégorisation des items : négatables

Si l'escroc avait réussi à tromper Samuel, il aurait pu gagner beaucoup d'argent

Si l'escroc n'avait pas réussi à tromper Samuel, il aurait pu gagner beaucoup d'argent.

- Classification d'items comme négatable si deux annotateurs natifs du français étaient d'accord que cela ne rendait l'item pas sémantiquement discutable.
- 38 items négatables
- Nous sommes la première équipe à enquêter sur cette propriété

Modèle Transformer pour le français

- Traduction des 44K items Winogrande par DeepL `https://github.com/xiaoouwang/FWSC285`
- *Fine-tuning* de CamemBERT large (Martin et al., 2020)

Évaluation selon la catégorie

taille entraÎn.	FWSC285	assoc.	non assoc.	avant comm.	après comm.	non nég.	nég.
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: Accuracy on FWSC285 depending on the data set and the size of training set

1 Introduction

2 Mise à jour des données

3 Catégorisation des items

4 Modèle SOTA

5 Évaluation selon catégorie

6 Conclusions

Conclusions

- Nouvelle performance SOTA grâce à la méthode 'fine-tuning d'un modèle transformer'
- Notre catégorisation a permis de voir que la performance dépend principalement des items associatifs
- La performance sur les items négatables dépasse à peine la performance aléatoire
- L'évaluation sur différentes catégories permet de voir plus clairement ce dont le modèle est capable et s'il a appris à raisonner.

References I

Amsili, P. and Seminck, O. (2017a). A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017*, pages 24–29.

Amsili, P. and Seminck, O. (2017b). Schémas Winograd en français: une étude statistique et comportementale. In *Conférence sur le Traitement Automatique du Langage Naturel*, volume 2, pages 28–35, Orléans. Association pour le Traitement Automatique des Langues.

Elazar, Y., Zhang, H., Goldberg, Y., and 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*, pages 10486–10500.

Emami, A., Suleman, K., Trischler, A., and Cheung, J. C. K. (2020). An analysis of dataset overlap on Winograd-style tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., and Morgenstern, L. (2022). The defeat of the Winograd Schema Challenge. arXiv, 2201.02387.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and 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., and 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., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and 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., and 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*, pages 343–350.

Trichelair, P., Emami, A., Trischler, A., Suleman, K., and 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)*, pages 3382–3387, Hong Kong, China. Association for Computational Linguistics.

Trinh, T. H. and Le, Q. V. (2018). Do language models have common sense?