Schémas Winograd: données et modèles

Pascal Amsili

Université Sorbonne Nouvelle
& Lattice (CNRS/PSL-ENS/SN)

séminaire Lattice, février 2022

Sorbonne Nouvelle
Lattice

Tout commence...

(1) The city councilmen refused the demonstrators a permit because they ⟨feared/advocated⟩ violence. *(Winograd, 1972)*

(2) Les gardiens ont donné les fruits aux singes parce qu'ils étaient [pourris/affamés/rassasiés].

Pas d'influence de la structure syntaxique

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(3) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?
R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

(4) Nicolas n'a pas pu soulever son fils parce qu' il était trop lourd.
Qui était trop lourd ?

- ... la bonne réponse change et elle est aussi évidente

$\Rightarrow$ indices linguistiques insuffisants

- Terminologie : 1 schéma = 2 items spe(cial) et alt(ernate)

Sorbonne Nouvelle
Lattice

Test d'intelligence artificielle

Alternative au test de Turing (simulation d'une conversation par une IA)

- besoin de raisonnement et de connaissances encyclopédiques
- résoud des problèmes liés au test de Turing :
  - *machine non humaine*
  - *détournement de la conversation* (Levesque et al., 2012)

- 2016 : premier Winograd Schema Challenge (Morgenstern et al., 2016)
  - 1er round : désambiguisation des pronoms (5)
  - 2e round : schémas Winograd

(5) Mrs. March gave the mother tea and gruel, while she dressed the little baby as tenderly as if it had been her own.

- 1er round : baseline (chance) : **42%**

→ meilleur système : Liu et al. (2017) : **58%**
(**66,7%** dans une version plus récente)

→ pas de 2e round

Sorbonne Nouvelle
Lattice

Contraintes sur les anaphores

(6) a. The trophy doesn't fit inside the suitcase because it is too ⟨large/small⟩.

b. *The trophy doesn't fit in the suitcase because the trophy is an awkward shape and it is too small.

(7) a. Ann has no children, but Barbara has two sons, Carl and David. Her children are in elementary school.

b. *Barbara has two sons, Carl and David, but Ann has no children. Her children are in elementary school.

(Kocijan et al., 2022)

Contraintes discursives

(8)
a. ?The demonstrators were denied a permit by the city council because they feared violence.
b. ?The city council denied the demonstrators a permit because they felt strongly that the best way to draw attention to current political issues is to advocate violence.
c. ?Norm lent his car to his brother's girlfriend. He doesn't own one.
d. ?Margaret Thatcher admires Hillary Clinton, and George W. Bush absolutely worships her.

(Kehler et al., 2008)

Jeu de données initial

- (Kehler et al., 2008) : 32 item expérimentaux inspirés de Winograd.
- (Davis et al., 2015) : publication de 143 schémas
- 2017 : augmentation de la collection : WSC273
- 2018 : collection de référence : WSC285

Traduction/adaptation d'un jeu de données français

- traduction par deux stagiaires
- validation par un autre stagiaire
- vérification par les auteurs (Seminck and Amsili, 2017)
- accent mis sur la naturalité des propositions
- élimination des items non consensuels

Résultat

107 schémas en français avec référence aux schémas anglais.

Wang (2021) a repris ce travail et a augmenté la collection en relation plus directe avec les items anglais.

Examples d'Adaptation (i)

- Traits de genre et de nombre

(9) The drain is clogged with hair. It has to be ⟨cleaned/removed⟩.

- Traduction directe n'est pas possible, car *cheveux* est toujours au pluriel, alors que *siphon* est au singulier.
- *cheveux* $\longrightarrow$ *savon*

(10) Il y a du savon dans le siphon de douche. Il faut le ⟨retirer/nettoyer⟩.

Examples d'Adaptation (ii)

- Non équivalence des constructions

(11) Mary tucked her daughter Anne into bed, so that she could sleep.

En français, la disponibilité de la construction infinitive (12-c) rend l'interprétation (12-b) peu plausible.

(12) a. Marie a couché sa fille$_i$ pour qu'elle$_i$ dorme.
b. ?Marie$_i$ a couché sa fille pour qu'elle$_i$ dorme.
c. Marie a couché sa fille pour dormir.

Autre exemple (Wang, 2021)

(13) a. They broadcast an announcement, but a subway came into the station and I couldn't ⟨hear/hear over⟩ it.
b. 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.

Exemples d'adaptation (iii)

- Problèmes lexicaux

(14) Susan knows all about Ann's personal problems because [she] is ⟨nosy/indiscreet⟩.

Traduction française pour 'indiscreet' : *indiscrète*.
Malheureusement, en français, *une personne indiscrète* peut être :

- quelqu'un qui révèle des secrets
- quelqu'un qui cherche avec insistance à découvrir des secrets

→ a nosy person !

(15) Sylvie est au courant de tous les problèmes personnels de Marie car [elle] est ⟨curieuse/bavarde⟩.

Sorbonne Nouvelle
Lattice

Autres traductions

Des collections de schémas, le plus souvent traduits ou adaptés de WSC273, existent dans plusieurs langues :

- Français : (Amsili and Seminck, 2017)
- Portugais : (Melo et al, 2020)
- Japonais et Chinois (incomplets)
- Chinois : (Bernard and Han, 2020)
- Russe, Arabe (non publiés)
- Hebreu (non publié)

Autres jeux de données

- Résolutions d'anaphores moins contraintes : DPR (2000 exemples, (Rahman and Ng, 2012)) ; PDP (122 problèmes en lien avec le WSC)
- Utilisation de schémas Winograd pour d'autres tâches : Wnli (16) (1000 exemples), WinoGender et WinoBias (17) (200-2000 exemples), etc.

(16)

prémisse	The city councilmen refused the demonstrators a permit because they feared violence.
conclusion	The demonstrators feared violence.
réponse	**non valide**

(17) The surgeon operated on the child with great care ; ⟨his/her⟩ ⟨tumor/affection⟩ had grown over time.

Changement d'échelle : WinoGrande

- motivation : performances au plafond sur les jeux de données existants... mais toujours pas de raisonnement naturel... *(Sakaguchi et al., 2019)*
- → banc d'essai à la bonne échelle pour les systèmes d'IA actuels
- collecte de 44 000 exemples avec Amazon MTurk
- filtrage adversatif : 12 300 exemples restant

Critiques (Kocijan et al., 2022)

- Malgré le filtrage, certains items sont associatifs (18)
- Certains items sont particulièrement compliqués (19)
- Le nouveau banc d'essai a été « résolu » presque aussitôt

(18) The doctor diagnosed Justin with bipolar and Robert with anxiety. He had terrible nerves recently.

(19) George opted for both of them to use a knife instead of a gun in the duel because it could partially injure them.

Sorbonne Nouvelle Lattice

Validation des schémas

Idéalement, les schémas ne doivent être

- ni trop faciles (pas associatifs)
- ni trop difficiles (trop difficiles pour les humains)

Google-proofness

> *"... there should be no obvious statistical test over text corpora that will reliably disambiguate [the anaphor of a Winograd item] correctly."*
>
> *(Levesque et al., 2012)*

- Les schémas doivent donc être à l'épreuve de Google (google-proof), ou comme on dit maintenant, non associatifs — ce n'est pas le cas de (20) (restrictions de sélection), ni de (21) (associations lexicales).

(20) Un arbre est tombé sur le toit, il va falloir l(e) ⟨enlever/réparer⟩

(21) Le bolide a dépassé le bus scolaire, parce qu'il roulait très vite.

- Certains items en anglais ont été vérifiés manuellement pour leur Google Proofness
- Pour réaliser un test plus sytématique :
  ⇒ On peut utiliser des mesures basées sur l'information mutuelle

Sorbonne Nouvelle Lattice

Information Mutuelle

- théorie de l'information (Shannon and Weaver, 1949)
- mesure la dépendance entre deux variables aléatoires
- peut mesurer la dépendance entre deux mots x et y (Church and Hanks, 1990)
- $MI(x, y)$ est positive si $P(x, y) > P(x) \times P(y)$

$$MI(x, y) = log_2 \left(\frac{P(x, y)}{P(x) \times P(y)} \right) \quad (1)$$

- comptages de fréquences (non-lissées) dans FrWaC (1.6 milliard de tokens du domaine `.fr`) (Baroni et al., 2009)
- fenêtre de voisinage : $-2 + 2$
- expressions multi-mots : tête lexicale
- lemmes plutôt que mots-formes (sauf cas particulier)

Sorbonne Nouvelle Lattice

Illustration du calcul

(22) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/lourde⟩.

Item Spe

MI(sculpture, encombrer)	$= 4.23$	
~~MI(sculpture, encombrer)~~	$= 4.23$	
MI(étagère, encombrer)	$= 10.01$	✓

Item Alt

MI(sculpture, lourd)	$= 2.41$	
~~MI(sculpture, lourd)~~	$= 2.41$	
MI(étagère, lourd)	$= 4.03$	✗

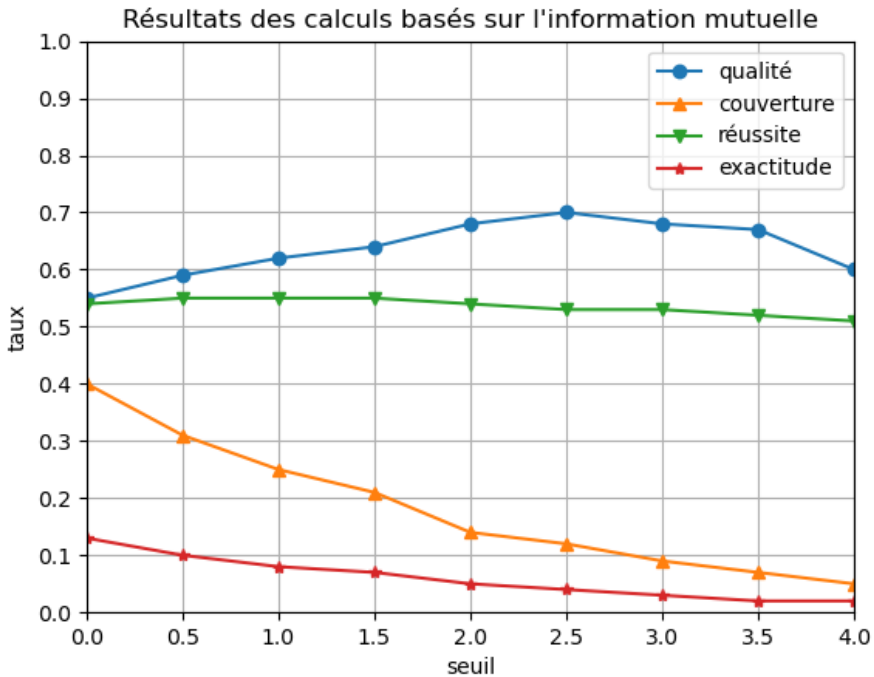
- Introduction d'un seuil de confiance (écart entre les deux valeurs)
- Au total, la méthode a pu être appliquée à 180 items (sur ≈ 220)
- parmi lesquels 131 ont pu recevoir un score d'information mutuelle (pas assez d'occurrences pour les items restants)

Sorbonne Nouvelle Lattice

Résultats

Seuil	# Items	Qualité	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Couverture	pourcentage d'items ayant reçu une mesure
Qualité	taux de réussite sur les items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items
Exactitude	mesure globale sur la collection complète



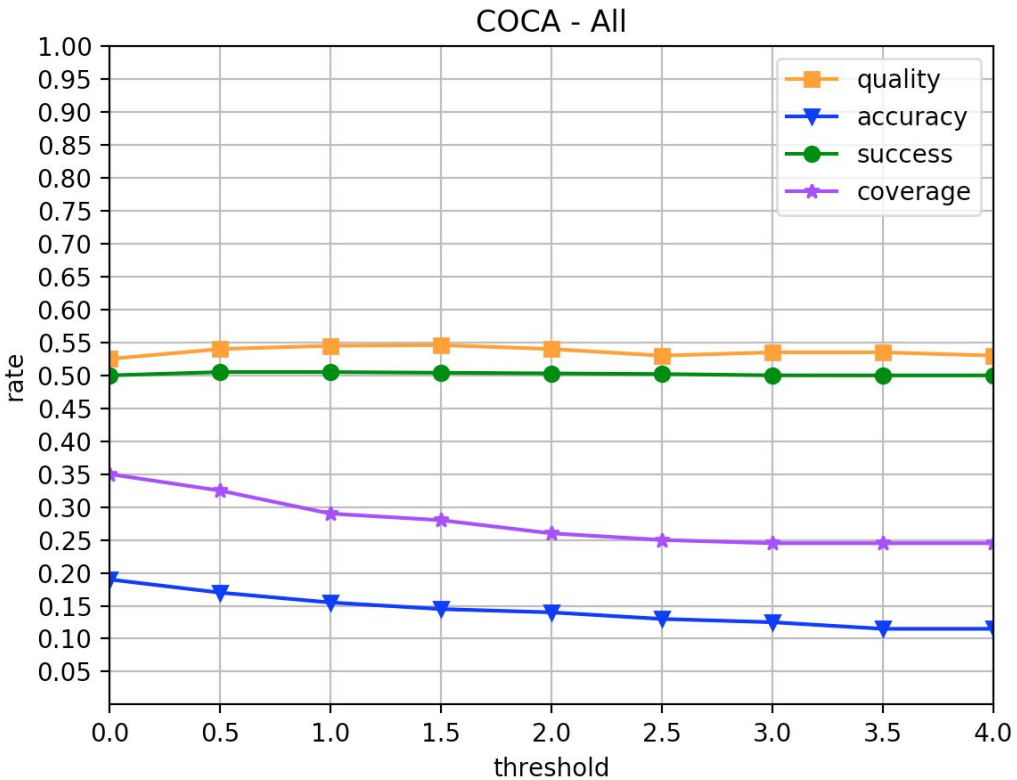
- répondre au hasard donne $\approx$ 50% de réussite
- le taux de réussite sans application de seuil reste bas (55%)
- réussite dans 70% des cas avec Δ 2.5 **mais** $<$ 15% des items.
- Dans son ensemble la collection est "Google-proof".
- Question ouverte : robustesse face à des méthodes plus sophistiquées

Sorbonne Nouvelle Lattice

Collection anglaise

- Calcul automatisé pour la collection WSC286
- Corpus COCA (10^9 mots, Americain contemporain)
- Résultats comparables

stage Patricia Rozé, 2019



$$accuracy = \frac{h}{n}$$

$$coverage = \frac{n-\theta}{n}$$

$$success = \frac{h+\frac{\theta}{2}}{n}$$

$$quality = \frac{h}{n-\theta}$$

n = 286
h : bonnes réponses
θ : non-réponses

Collection chinoise

Jeu de données Mandarinograd (Bernard and Han, 2020) :
Corpus tiré du Wikipedia chinois, $\approx$ 250 millions de mots
Construction manuelle des paire de termes, avec des cas impossibles comme (23).
Seuil $\Delta = 0$

(23) a. Ann asked Mary what time the library closes, ⟨but/because⟩ she had forgotten.
b. Jane gave Joan candy because she was ⟨∅/not⟩ hungry.

nombre items	$n =$	308
nombre hits	$h =$	93
nombre non-rep	$\theta =$	155
exactitude :	h/n	30,2 %
qualité :	$h/(n-\theta)$	60,7 %
réussite :	$(h+\theta/2)/n$	55,3 %
couverture	$(n-\theta)/n$	49,7 %

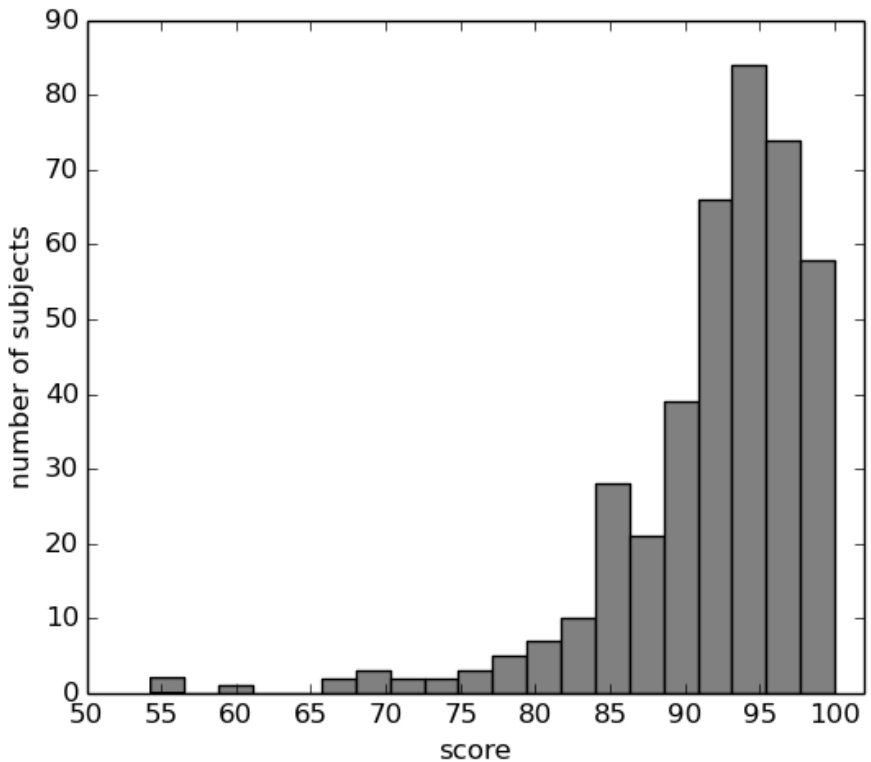
Facilité

Performance humaine sur les schémas anglais (Bender, 2015) :

- 430 participants sur MTurk (407 après élimination)
- 160 schémas (dont 17 “faciles”)
- 40 questions par participants, total 16 000 points de données, au moins 50 réponses par item.

Score moyen : 92.1% de réussite

Temps moyen : 10,2′ pour 40 questions
1 p. sur 7 ont répondu à 100%
Items faciles : 98,6%



Analyse par schéma

- 75% des questions ont reçu une réponse correcte de plus de 90% des participants
- 17% des items n'ont jamais reçu la bonne réponse

Pistes pour une explication (après entretiens post-passassion) :

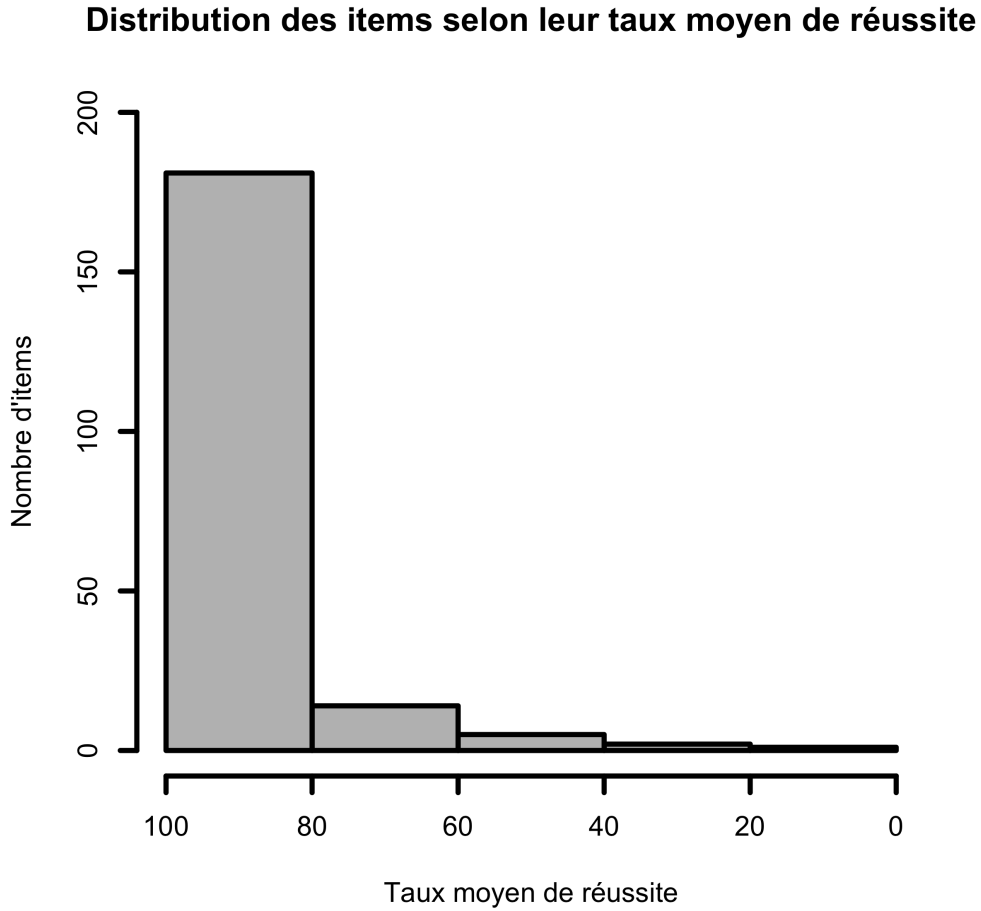
- utilisation d'indices externes aux items (une réunion a plus de chance qu'un train d'être retardé)
- réponses trop rapides
- problèmes de vocabulaire
- "ambiguïtés unidirectionnelles"

Schémas français

Expérience en ligne :

- 22 participants sur IbexFarm
- items randomisés
- réponses avec RT $> 1''$ et $< 60''$

93,6 %



Type d'items difficiles

Ambigus :

(24) Pierre et Marc sont poursuivis pour diffamation. Pierre a écrit dans leur livre plusieurs faux témoignages que Marc a colportés. Il aurait dû être plus ⟨prudent/honnête⟩.

Réussite humaine : ⟨50%/18%⟩

Connecteur :

(25) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'ils venaient de plus loin.

Réussite humaine : ⟨70%/75%⟩

Complexes :

(26) Pierre jouait aux cartes avec Adam qui menait au score. Si la chance d'Adam n'avait pas tourné il aurait ⟨perdu/gagné⟩.

Réussite humaine : ⟨50%/67%⟩

Sorbonne Nouvelle Lattice

Sous-classes de schémas

Trichelair et al. (2018) ont initié une série d'études pour établir des sous-classes de schémas.

- Associativité (27) établie par annotation humaine (13% de WSC273)
- sous-classe repliquée par Elazar et al. (2021) de façon automatique

(27) Joe has sold his house and bought a new one a few miles away. He will be moving out of it on Thursday.

- Switchability : 48% de WSC273 forment des énoncés pertinents quand on intervertit les antécédents (28)
- Un système qui résout bien (ie pas par chance) le schéma initial devrait résoudre le schéma inversé

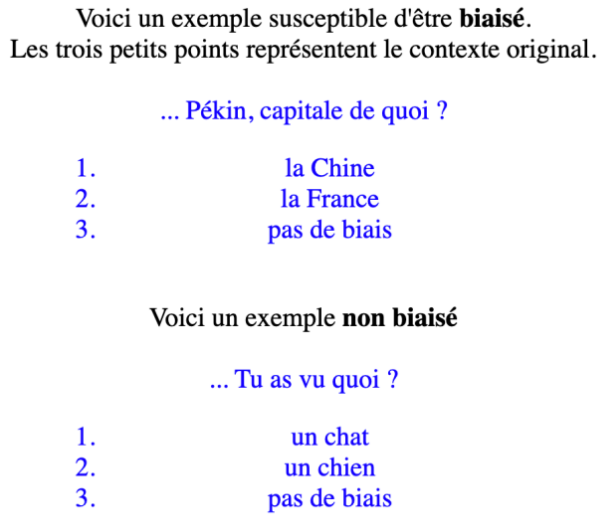
(28) Bob collapsed on the sidewalk. Soon he saw Carl coming to help. He was very ⟨ill/concerned⟩.

Sorbonne Nouvelle Lattice

Pour le français

Wang (2021) : recherche de l'association positive ou négative avec un questionnaire

- seulement 2 annotateurs pour Trichelair et al. (2018)
- difficulté méthodologique
- 285 items (en français)
- 45 participants



Voici un exemple susceptible d'être **biaisé**.
Les trois petits points représentent le contexte original.

... Pékin, capitale de quoi ?

1. la Chine
2. la France
3. pas de biais

Voici un exemple **non biaisé**

... Tu as vu quoi ?

1. un chat
2. un chien
3. pas de biais

- 37 items associatifs ($\geq$ 76% de bonnes réponses)
- 3 items négativement associatifs :

(29) a. Qu'est-ce qui avait l'air délicieux ? ⟨le ver/le poisson⟩
b. Qu'est-ce qui s'est retrouvé plein d'encre ? ⟨le café/le stylo⟩

Sous-classe *negatable*

Définition (Wang, 2021) : un item est *niable* si le verbe principal peut être nié sans altérer l'intelligibilité de l'item, en causant une inversion des réponses attendues.

(30) a. Le scooter a dépassé le bus scolaire, car [il] roulait trop vite.
b. Le scooter n'a pas dépassé le bus scolaire, car [il] roulait trop vite.

FrWSC285 :

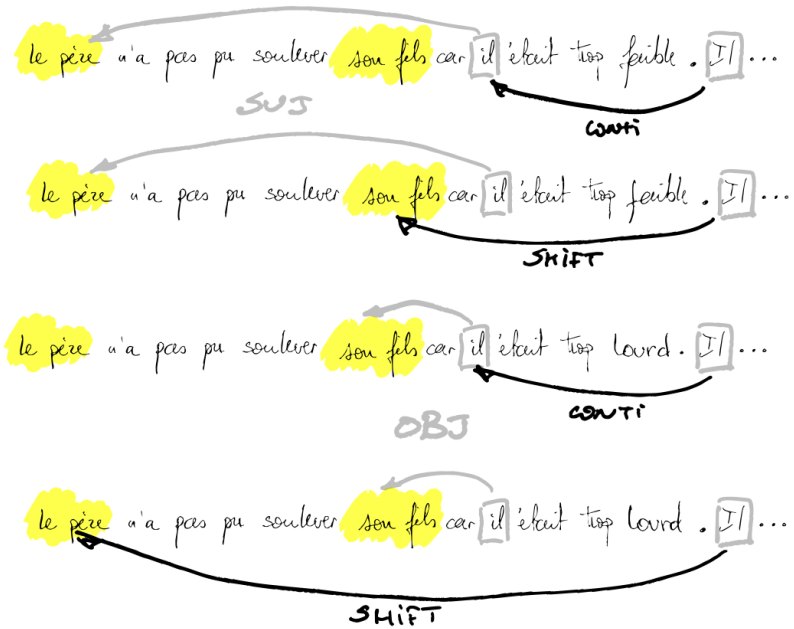
- 37+3 items associatifs
- 141 items switchable
- 38 items negatable

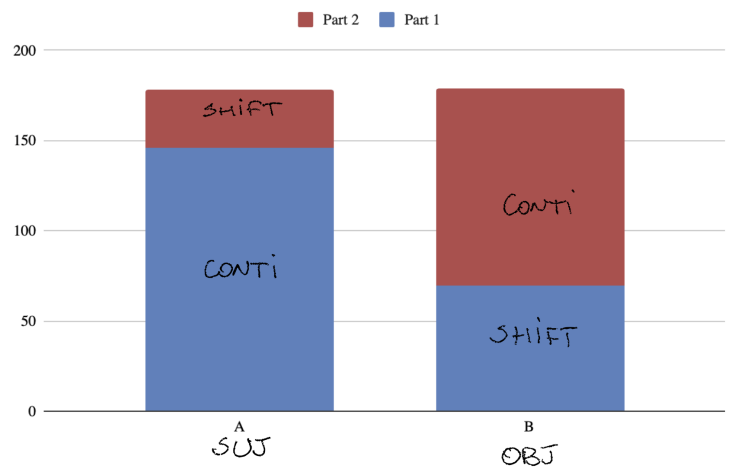
Objectif : raffiner les études de performance

Sorbonne Nouvelle
Lattice

Schémas Winograd comme items expérimentaux

Exemple : influence de la position sujet sur la topicalité. Cloze task.





merci à Olga Seminck et les stagiaires du Lattice – printemps 2020

Approches feature-based

(31) Le poisson mangea le ver. Il était ⟨affamé/délicieux⟩.

1. Analyse linguistique de l'item
   - extraction de mots-clés
   - pattern-matching
   - parsing en dépendance
2. Injection de connaissances du monde
   - à partir de bases de connaissances
   - à partir de requêtes sur Internet
   - à partir de corpus (collocations, fréquences...)
3. Raisonnement "naturel"
   - logique (th. provers)
   - transformations de graphes
   - SVM-rankers
   - ...

Performances

- (Sharma et al., 2015) : 84% des exemples sont résolus par le système avec un pré-traitement manuel, le score tombe à 50% avec pré-traitement automatique
- (Emami et al., 2018) : premier système à être significativement meilleur que la chance sur l'ensemble du jeu de données (57%).

Knowledge hunting : décomposition de l'item ; génération de requêtes et recherches sur Internet ; résolution des pronoms.

Pair	$Pred_C$	E_1	E_2	$Pred_Q$	P	Alternating Word (POS)
1	couldn't lift	the man	his son	was so heavy	he	weak/heavy (adjective)
2	were bullying	the older students	the younger ones	punished	them	punished/rescued (verb)
3	tried to paint	shepherds	sheep	ended up .. like	they	golfers/dogs (noun)

Table 2: Winograd sentence pairs from Table 1.

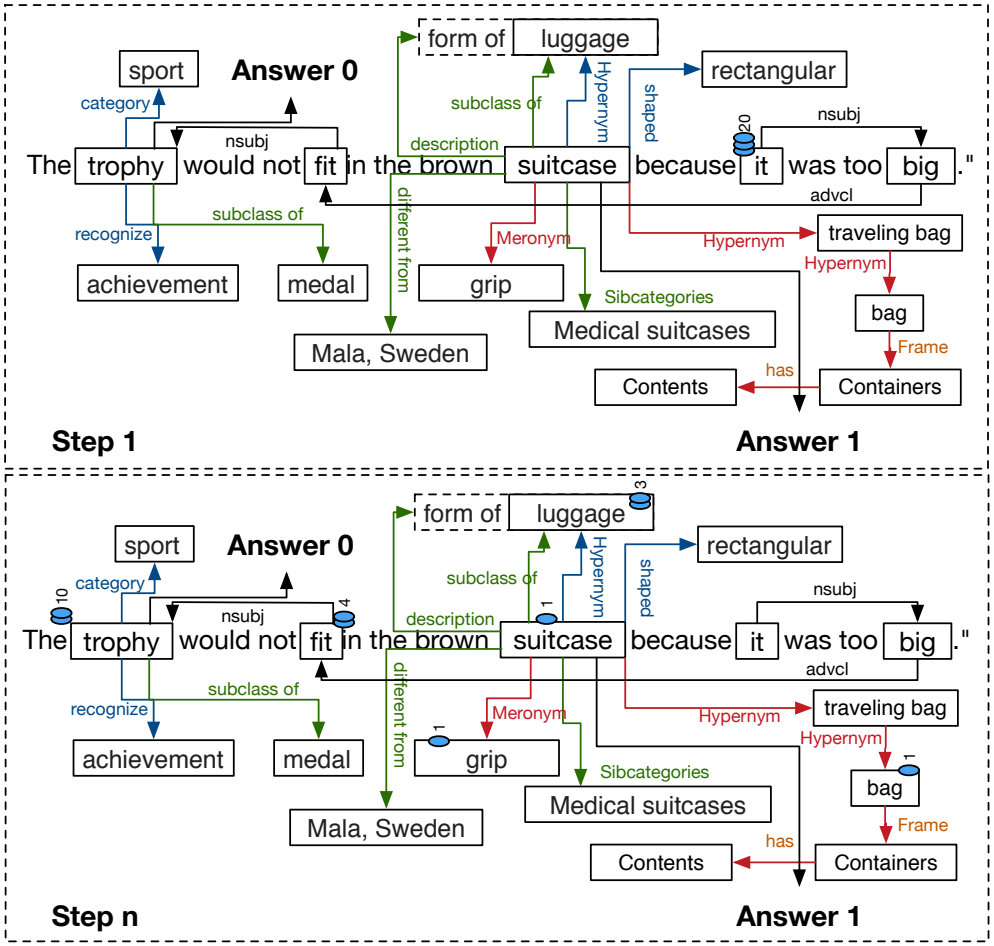
Sentence: The trophy doesn't fit into the brown suitcase because it is too large.		
Query Generation Method	C	Q
Automatic	{"doesn't fit into", "brown", "fit" }	{"large", "is too large"}
Automatic, with synonyms	{"doesn't fit into", "brown", "accommodate", "fit", "suit" }	{"large", "big", "is too large" }
Manual	{"doesn't fit into", "fit into","doesn't fit" }	{"is too large", "too large"}

Table 3: Query generation techniques on an example Winograd sentences, where C and Q represent the sets of queries that capture the context and query clauses of the sentence, respectively.

Sorbonne Nouvelle Lattice

Knowledge intensive methods

Performance correcte ($\approx$ 75%)

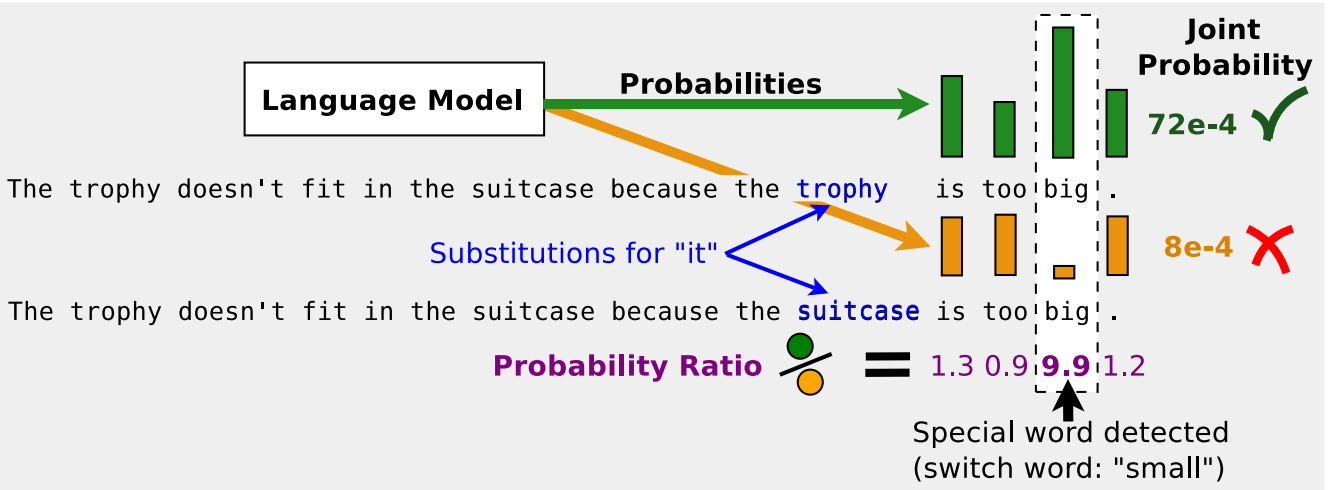


WordNet, Wiktionnary, Wikidata

(Fähndrich et al., 2018)

Modèles de langue

Modèles de langue : systèmes d'apprentissage (milliers de paramètres) entraînés à prédire le prochain mot dans un texte (apprentissage auto-supervisé). Permet (une fois entraîné) d'attribuer un score à n'importe quelle suite de mots.



Premier système en 2018 (non neuronal) : 63% *(Trinh and Le, 2018)*

Modèle de langue récurrent

- Méthode inspirée de Trinh and Le (2018)
- Génération d'une paire de phrases en remplaçant le pronom
- Ré-agencement de phrase si nécessaire
- Calcul de probabilité jointe donnée par modèle de langue
  - *Long Short Term Memory* (LSTM)
  - Entraîné sur la version française de Wikipédia (Coavoux, 2017)
  - Vocabulaire = 100K mots les plus fréquents
  - Taille des représentations vectorielles de mots = 1024
  - Taille des couches cachées = 2048
  - Minimisation de perte algorithme *Adagrad* (Duchi et al., 2011)

Simon a expliqué sa théorie à Marc, mais [il] ne l'a pas ⟨convaincu/compris⟩.

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.

Sorbonne Nouvelle Lattice

Modèle contextuel

- Context2vec (C2V) (Melamud et al., 2016)
  - Représenter le contexte d'un token par réseaux de neurones récurrents bi-directionnels
  - Estimer à quel point un token est attendu dans un contexte
- A nouveau 2 versions des items Winograd
- Laquelle des deux réponses s'insère le mieux dans le contexte ?
- Similarité entre les réponses et les contextes
  - Entraîné sur la version française de Wikipédia (Coavoux, 2017)
  - Taille vecteurs = 300
  - Optimisation avec *Adagrad* (Duchi et al., 2011)
    pas d'apprentissage = 0,001
  - 2 méthodes de calcul de similarité :
    Similarité cosinus
    Similarité log sigmoïde

Simon a expliqué sa théorie à Marc, mais **Simon** ne l'a pas comprise.

Simon a expliqué sa théorie à Marc, mais **Marc** ne l'a pas comprise.

Limites

- Absence d'expression anaphorique

Joël a vendu sa maison et en a acheté une nouvelle à quelques kilomètres.
Il va ⟨déménager/emménager⟩ ce jeudi.
Joël va ⟨déménager de/emménager dans⟩ quelle maison ?

R0 : son ancienne maison
R1 : sa nouvelle maison

- Item Winograd sur plusieurs phrases

Fred est le seul homme encore vivant à se rappeler de mon arrière grand-père. C'
⟨est/était⟩ un homme remarquable.
Qui ⟨est/était⟩ remarquable ?

R0 : Fred
R1 : mon arrière grand-père

- Mots de réponse qui sont inconnus → créent des non-réponses

Résultats

	n	h	θ	réus
Collection en anglais				
Emami et al. (2018) AGQS	273	106	83	0,54
Emami et al. (2018) MGQ	273	118	76	0,57
Sharma et al. (2015)	283	49	230	0,58
Trinh and Le (2018) Word-full	273	147	0	0,54
Trinh and Le (2018) 10 modèles	273	168	0	0,62
Trinh and Le (2018) 14 modèles	273	174	0	0,64
Collection en français				
LSTM Seminck et al. (2019)	214	65	88	0,51
C2V Seminck et al. (2019)	214	33	158	0,52
Amsili and Seminck (2017)	180	72	49	0,54

C'était il y a 3 ans...

Sorbonne Nouvelle Lattice

	WSC273	WNLI	PDP	WINOGRANDE
Trinh and Le (2018)	63.7%	–	70%	–
Radford et al. (2019)	70.7%	–	–	–
Klein and Nabi (2019)	60.3%	–	–	–
Prakash et al. (2019)	70.17%	–	–	–
Kocijan et al. (2019b)	72.5%	74.7%	–	–
Kocijan et al. (2019a)	71.8%	74.7%	86.7%	–
Ruan et al. (2019)	71.1%	–	–	–
He et al. (2019)	75.1%	89%	90.0%	–
Ye et al. (2019)	75.5%	83.6%	–	–
Sakaguchi et al. (2020)	90.1%	85.6%	87.5%	79.1%
Brown et al. (2020)	88.3%	–	–	77.7%
Yang et al. (2020)	–	–	–	80.0%
Lin et al. (2020)	–	–	–	84.6%
Khashabi et al. (2020)	–	–	–	89.4%
Lourie et al. (2021b)	–	–	–	91.3%

(Kocijan et al., 2022)

Sorbonne Nouvelle Lattice

The defeat of the WSC

(Kocijan et al., 2022) : le défi est terminé.

Mais les réalisations impressionantes de la communauté reflètent-t-elles une percée de l'IA ?

Selon (Elazar et al., 2021), la réussite des modèles actuels est un artéfact, dû :

- à des mesures de performance trop accommodantes ;
- à des artefacts dans les jeux de données, qui persistent malgré les efforts ;
- à des "fuites" de connaissance et de raisonnement qui découlent des immenses quantités de données d'entraînement.

Dans leurs manipulations, ils montrent que lorsque la définition de la tâche, le régime d'entraînement, les données d'entraînement et la métrique d'évaluation sont modifiés pour corriger ces points, la performance chute significativement.

(32) a. The large ball crashed right through the table because it was made of steel.
b. I bought a steel property at the same time as my wooden property. The — property was harder.

Objectifs

Le but du défi des schémas Winograd était de fournir une épreuve

- facile pour les humains et "*commonsensical*" ✓
- facile à évaluer ✓
- pertinente pour évaluer le "raisonnement naturel" ✗

- malédiction du benchmark (Kocijan et al., 2022)
- continuons à faire de la linguistique...
- quant à l'intelligence artificielle...

Doug Hofstadter, 2011

The problem is that I believe that what will happen is that you will simply wind up spawning a whole host of new and ultra-clever brute-force techniques to solve the "Winograd Challenge" without solving the problem of understanding whatsoever. I always liked Terry Winograd's sample sentence, but I hardly think that it represents the epitome or the essence of what is wrong with today's computer approaches to language.

Remerciements

Olga Seminck

Nous remercions nos stagiaires pour les items parce qu'on ne peut pas les ⟨oublier/résoudre⟩.

Sarah Ghumundee
Biljana Knežević
Nicolas Bénichou
Hugo Taquet
Dara Nguyen
Ryan Hunt
Yann Castellvi
Lara Perinetti
Léonard Fromond
Pierre-Louis Lugiery
Diana Khabarova
Quentin Gliosca
Pauline Bossard
Patricia Roze
Victoriane Djaroudi
Jonas Noblet
Ward El Mouna Belarbi
Léopold Irion-Dewavrin
Emma Dubois
Leia Gilanton
Brittany Guannel
Ilona Le Cors
Manon Ignaczak
Xiaoou Wang

Labex « Empirical Foundations of Linguistics » (ANR-10-LABX-0083)
École Doctorale Frontières du Vivant — Programme Bettencourt

Références I

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. EACL, 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 *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3) :209–226.

Bender, D. (2015). Establishing a human baseline for the winograd schema challenge. In Glass, M. and Hee, K. J., editors, *Proceedings of the 26th Modern AI and Cognitive Science Conference (MAICS 2015)*, pages 39–45, Greensboro, NC, USA.

Bernard, T. and Han, T. (2020). Mandarinograd : A Chinese collection of Winograd schemas. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 21–26, Marseille, France. European Language Resources Association.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29.

Coavoux, M. (2017). *Discontinuous Constituency Parsing of Morphologically Rich Languages*. PhD thesis, Univerité Paris Diderot.

Davis, E., Morgenstern, L., and Ortiz, C. (2015). A collection of winograd schemas. Web page collecting 144 Winograd pairs, with comments and references.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul) :2121–2159.

Sorbonne Nouvelle
Lattice

Références II

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emami, A., Trischler, A., Suleman, K., and 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*, pages 25–31. Association for Computational Linguistics.

Fähndrich, J., Weber, S., and Kanthak, H. (2018). A marker passing approach to winograd schemas. In *Proceedings of the 8th Joint International Conference, JIST 2018*, pages 165–181, Awaji, Japan. Fähndrich J., Weber S., Kanthak H. (2018) A Marker Passing Approach to Winograd Schemas. In : Ichise R., Lecue F., Kawamura T., Zhao D., Muggleton S., Kozaki K. (eds) Semantic Technology. JIST 2018. Lecture Notes in Computer Science, vol 11341. Springer, Cham.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *Journal of semantics*, 25(1) :1–44.

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

Levesque, H. J., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge. In *Proceedingqs of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, North America.

Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2017). Combing context and commonsense knowledge through neural networks for solving Winograd schema problems. In *2017 AAAI Spring Symposium Series*.

Sorbonne Nouvelle

Lattice

Références III

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec : Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Morgenstern, L., Davis, E., and Ortiz Jr., C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1) :50–54.

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.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). Winogrande : An adversarial winograd schema challenge at scale. *arXiv preprint arXiv :1907.10641*.

Schüller, P. (2014). Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Seminck, O. (2018). *Cognitive Computational Models of Pronoun Resolution*. Phd dissertation, Université Paris Diderot.

Seminck, O. and Amsili, P. (2017). A computational model of human preferences for pronoun resolution. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 53–63. Association for Computational Linguistics.

Seminck, O., Segonne, V., and Amsili, P. (2019). Modèles de langue appliqués aux schémas winograd français. In *Actes de la conférence TALN 2019*, Toulouse.

Shannon, C. E. and Weaver, W. (1949). The mathematical theory of information.

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 Twenty-Fourth International Joint Conference on Artificial Intelligence. AAAI*.

Sorbonne Nouvelle
Lattice

Références IV

Trichelair, P., Emami, A., Kit Cheung, J. C., Trischler, A., Suleman, K., and Diaz, F. (2018). On the Evaluation of Common-Sense Reasoning in Natural Language Understanding. *ArXiv e-prints*.

Trinh, T. H. and Le, Q. V. (2018). A Simple Method for Commonsense Reasoning. *ArXiv e-prints*.

Wang, X. (2021). *SOTA performance on French Winograd Schemas with analysis of robustness of fine-tuned language models (CamemBERT) to entity switching and negation*. Master dissertation, Paris Nanterre, Master PluriTAL.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP : A Balanced Corpus of Gendered Ambiguous Pronouns. *ArXiv e-prints*.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press.

Sorbonne Nouvelle

Lattice

Doug Hofstadter, 2011, citation complète

The problem is that I believe that what will happen is that you will simply wind up spawning a whole host of new and ultra-clever brute-force techniques to solve the "Winograd Challenge" without solving the problem of understanding whatsoever. I always liked Terry Winograd's sample sentence, but I hardly think that it represents the epitome or the essence of what is wrong with today's computer approaches to language. It is just one type of example among thousands of types. It's a great example but it's misleading to focus on it as if it were really the crux of the matter. Getting people to spend huge amounts of time on just one kind of challenge is not going to be helpful. In fact, I fear it will be counterproductive, because I don't think that anyone who will be moved to tackle this particular challenge is likely to take up the deeper and more general challenge of what language understanding really is. People are daunted by that, as well they should be, and no one is going to be motivated by a prize to suddenly tackle that gigantic challenge. Instead, very smart engineering types are going to be motivated to seek clever tricks that will allow computers to solve this very narrow type of linguistic disambiguation problem with a high degree of accuracy.