A Google-Proof Collection of French Winograd Schemas

Pascal Amsili Olga Seminck

Laboratoire de Linguistique Formelle
Université Paris Diderot

CORBON Workshop, april 2017

Winograd Schemas

(Levesque et al., 2011)

- a sentence containing an anaphor & at least two possible antecedents

(1) Nicolas could not carry his son because he was too weak.
Who was too weak?

R0 : Nicolas
R1 : his son

Winograd Schemas

(Levesque et al., 2011)

- a sentence containing an anaphor & at least two possible antecedents

(1) Nicolas could not carry his son because he was too weak.
Who was too weak?

R0 : Nicolas
R1 : his son

- the "correct" answer is obvious for humans
- an alternative sentence is obtained by substituting one specific expression:

Winograd Schemas

(Levesque et al., 2011)

- a sentence containing an anaphor & at least two possible antecedents

(1) Nicolas could not carry his son because he was too weak.
Who was too weak?

R0 : Nicolas
R1 : his son

- the "correct" answer is obvious for humans
- an alternative sentence is obtained by substituting one specific expression:

(2) Nicolas could not carry his son because he was too heavy.
Who was too heavy?

Winograd Schemas

(Levesque et al., 2011)

- a sentence containing an anaphor & at least two possible antecedents

(1) Nicolas could not carry his son because he was too weak.
Who was too weak?

R0 : Nicolas
R1 : his son

- the "correct" answer is obvious for humans
- an alternative sentence is obtained by substituting one specific expression:

(2) Nicolas could not carry his son because he was too heavy.
Who was too heavy?

- the "correct" answer now changes (still obvious for humans)

General Format

(3) Frank was upset with Tom because the toaster [he] had ⟨bought from/sold to⟩ him didn't work.
Who had ⟨bought/sold⟩ the toaster?

R0 : Frank
R1 : Tom

Conventions:

- special ; alternate
- R0 is the first NP, R1 the second NP
- Item-Spe: item formed with the *special* expression
- Item-Alt: item formed with the *alternate* expression
- Correct answer Item-Spe : R0 ; correct answer Item-Alt : R1

Test for Artificial Intelligence

Winograd Schemas Challenge (WSC) :

alternative to the Turing Test (Levesque et al., 2011)

- requires reasoning capacity + encyclopedic knowledge
- solves issues with the Turing Test (TT):
  - *deception*: to pass the TT, a machine has to pretend it is human
  - *conversation*: in a conversation, a machine can use evasive strategies (as Eliza)

Actual Challenge(s)

- 2016: first Winograd Schema Challenge (Morgenstern et al., 2016)
- task: pronoun disambiguation problem (PDP) inspired by the format of Winograd Schemas
- collection of items like (4)

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

- not always grouped by pairs
- more than 2 antecedent candidates
  ⇒ baseline (chance level) around 45% (Liu et al., 2016)

Actual challenge(s): results

- winning system: Liu et al. (2016) : 58% success rate
  - unsupervised feature extraction
  - commonsense Knowledge Enhanced Embeddings
- more recent version by the same group: 66,7% success rate

Other attempts on specific subsets :

- Bailey et al. (2015): explicit inference rules and axioms to deal with schemas where discourse relations play a decisive role ;
- Schüller (2014): WS tackled by Formalizing Relevance Theory in Knowledge Graphs ;
- Sharma et al. (2015): deal with the $\approx$ 25% of the schemas that exhibit causal relations, achieve $\approx$ 75% accuracy

For the upcoming years, solving Winograd Schemas is likely to remain a challenge for NLP and AI communities.

Project

- Provide a data set for French
- Allow for cross-linguistic comparison
- Propose a systematic account for Google-proofness

Languages

- Original collection : 144 schemas in English (Davis et al., 2015)
- Translation of the whole collection into Japanese
  (with or without adaptation of the proper nouns)
- 12 schemas translated into Chinese

$\Rightarrow$

- `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`
- Not documented (literal/non literal translation)

Languages

- Original collection : 144 schemas in English (Davis et al., 2015)
- Translation of the whole collection into Japanese
  (with or without adaptation of the proper nouns)
- 12 schemas translated into Chinese

  $\Rightarrow$
  - `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`
  - Not documented (literal/non literal translation)
- 107 schemas in French translated/adapted from the original set.

  $\Rightarrow$ `http://www.llf.cnrs.fr/winograd-fr`

Adaptation examples (i)

- Gender/number features

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

Direct translation not available : the word 'hair' in French (*cheveux*) is plural, while 'drain' (*siphon*) is singular.
We replaced 'hair' with 'soap' (*savon*).

(6) Il y a du savon dans le siphon de douche. Il faut le [retirer/nettoyer].
There is soap in the shower drain. It has to be be ⟨removed/cleaned⟩

Adaptation examples (ii)

- Lexical difficulties

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

French translation for 'indiscreet': *indiscrète*.
However, in French *une personne indiscrète* can be:

- a person who reveals things that should stay secret
- a person who tries insistently to find out what should stay secret

Adaptation examples (ii)

- Lexical difficulties

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

French translation for 'indiscreet': *indiscrète*.
However, in French *une personne indiscrète* can be:

- a person who reveals things that should stay secret
- a person who tries insistently to find out what should stay secret

→ a nosy person!

Adaptation examples (ii)

- Lexical difficulties

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

French translation for 'indiscreet': *indiscrète*.
However, in French *une personne indiscrète* can be:

- a person who reveals things that should stay secret
- a person who tries insistently to find out what should stay secret

→ a nosy person!

In the French version of (7) we therefore changed the alternate to ⟨bavarde⟩ *(talkative)*

(8) Sylvie est au courant de tous les problèmes personnels de Marie car elle est ⟨curieuse/bavarde⟩.
Sylvie knows all Mary's personal problems because she is ⟨curious/talkative⟩

Adaptation examples (iii)

- Infinitival purpose phrases: language preferences

(9) Mary tucked her daughter Anne into bed, so that [she] could ⟨work/sleep⟩.
Who is going to ⟨work/sleep⟩?

R0 : Mary
R1 : Anne

in French, a purpose phrase about the subject can only be expressed via an infinitival clause (literal equivalent of *in order to work*).

⇒ the French counterpart of (9) unable to generate two questions where both NPs are possible antecedents.

Method

- translation done by two interns,
- validated by another intern while computing the Google-proof figures
- finally checked by both authors.

- most natural sounding solutions preferred over closeness to the original
- long translations avoided
- items for which no consensus could be found were simply removed

Outcome

- 107 schemas in `xml` format.
- a reference to the English counterpart will be included (when applicable)

```
<schema id="9" engn="46">
  <text>
    <txt1> Si l'escroc avait réussi à tromper Samuel, il aurait pu </txt1>
    <wordA>gagner</wordA>
    <wordB>perdre</wordB>
    <txt2> beaucoup d'argent. </txt2>
  </text>
  <question>
    <qn1>Qui aurait pu </qn1>
    <qwordA>gagner</qwordA>
    <qwordB>perdre</qwordB>
    <qn2> beaucoup d'argent ?</qn2>
  </question>
  <answer1>l'escroc</answer1>
  <answer2>Samuel</answer2>
</schema>
```

Google-proofness

- by design, schemas cannot be resolved without reasoning about world knowledge

"... there should be no obvious statistical test over text corpora that will reliably disambiguate [the anaphor of a Winograd item] correctly."
(Levesque et al., 2011)

(10) Many astronomers are engaged in the search for distant galaxies. They are spread all over the universe.

- Even though some items of the English collection have been checked for Google-proofness,
- we wanted a systematic test applicable to the whole collection,
- so we devised a simple statistic measure based on Mutual Information.

Mutual Information

Mutual Information: concept from Information Theory (Shannon and Weaver, 1949) that measures the mutual dependence of two random variables.

Mutual Information can be used to measure word association: when two words x and y are mutually dependent, the probability of their cooccurrence $P(x, y)$ will be higher than the probability of observing them together by chance : $MI(x, y)$ will be positive:

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

(Ward Church and Hanks, 1990)

Computation

(11) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Computation

(11) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Item Spe $MI(\text{sculpture}, \text{encombrer}) = 4.23$
$MI(\text{étagère}, \text{encombrer}) = 10.01$

Computation

(11) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Item Spe ~~*MI*(sculpture, encombrer)~~ = 4.23
MI(étagère, encombrer) = 10.01

Computation

(11) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23
	MI(étagère, encombrer)	= 10.01
Item Alt	*MI*(sculpture, lourd)	= 2.41
	MI(étagère, lourd)	= 4.03

Computation

(11) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23
	MI(étagère, encombrer)	= 10.01
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41
	MI(étagère, lourd)	= 4.03

Computation

(11) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/lourde⟩.
The sculpture fell off the shelf because it was too ⟨cluttered/heavy⟩

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23
	MI(étagère, encombrer)	= 10.01
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41
	MI(étagère, lourd)	= 4.03

Reliability of scores differences: introduction of a threshold.

Applicability: lexeme extraction

- Extraction of relevant expressions:

– Easy case: expected answers (R0/R1) + special/alternate

In fact we want to make a choice between possible answers:

(12) **item Spe:**
The sculpture fell off the shelf because it was too cluttered.
What was too cluttered?
R0: the sculpture... was too cluttered
R1: the shelf... was too cluttered

(13) **item Alt:**
The sculpture fell off the shelf because it was too heavy.
What was too heavy?
R0: the sculpture... was too heavy
R1: the shelf... was too heavy

Applicability: excluded items (1)

– Difficult case:

(14) **Item Spe:**
In the middle of the outdoor concert, the rain started falling, and it continued until 10. What continued until 10?
R0: the rain... continued until 10
R1: the concert... continued until 10

(15) **Item Spe:**
In the middle of the outdoor concert, the rain started falling, but it continued until 10. What continued until 10?
R0: the rain... continued until 10
R1: the concert... continued until 10

15 schemas of this form were excluded from our study.

Applicability: excluded items (2)

- Fancy schemas:

Look! There is a ⟨shark/minnow⟩ swimming right below that duck!
It had better get away to safety fast! *(Davis et al., 2015, ex(93))*
What needs to get away to safety?
Answer Pair A: The shark/The duck.
Answer Pair B: The minnow/The duck.

The pair of possible answers depends on the choice of words, since the special and alternate words are possible referents.

2 schemas of this form were excluded from our study.

Applicability: proper nouns

- Proper nouns

(16) Steve follows Fred's example in everything.
He ⟨admires/influences⟩ him hugely.
Who ⟨admires/influences⟩ whom?

⇒ Google-proof by design

- 44 schemas of this sort, still included in the scores

Probability estimation

All together, we measured Mutual Information for 90 schemas (180 items)

- unsmoothed frequency counts from FrWaC Baroni et al. (2009) (1.6 billion tokens from the `.fr` domain of the Internet)
- window for cooccurrence measures: 2× 5 tokens
- multiword expressions: lexical head
- lemmas rather than word-forms (except in a couple of exceptional cases)

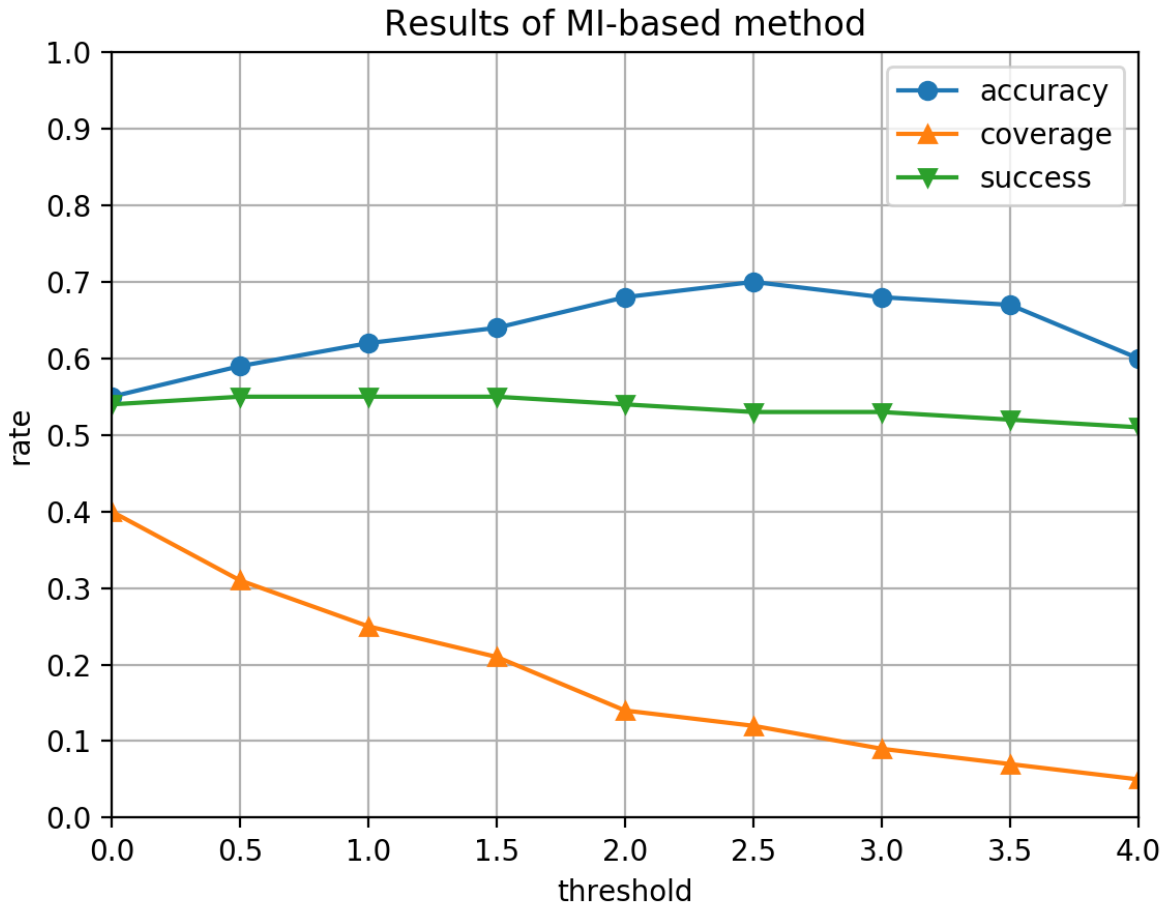
We used a fixed corpus and not the Google search engine because the counts on Google are not stable in time and also optimization algorithms could alter the counts (Lapata and Keller, 2005).

Results (table)

Threshold	# Items	Accuracy	Coverage
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' indicates the number of items that the method could answer to
- 'Accuracy' is the accuracy of the method on the items that could be answered
- 'Coverage' gives the accuracy on the 180 items we tried to solve with *MI*

Results (plot)



'Success' is the theoretical success rate that would obtain a strategy consisting in using mutual information for the questions for which the Δ is over the threshold, and replying by chance for the other questions.

Discussion

- answering at random would give an accuracy around 50%
- accuracy with no threshold not satisfactory (55%)
- accuracy reaches 70% with Δ 2.5 **but** for less than 15% of the items.
- using the best accuracy does not help the overall success rate to pass 55%

- As a whole, our collection is (in a sense) Google-proof.
- No claim about more sophisticated methods.
- Post hoc exploitation: remove schemas that are too easy

The audience refused to thank the speakers, they were too ⟨bored/boring⟩.

Acknowledgments:

Sarah Ghumundee,
Biljana Knežević,
Nicolas Bénichou
3 anonymous reviewers
École Doctorale Frontières du Vivant — Programme Bettencourt

`http://www.llf.cnrs.fr/winograd-fr`

References I

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 *MAICS*, pages 39–45.

Davis, E. (2015). A difference of a factor of 70,000 between hit counts and results returned in google. Unpublished note available on the author's web page.

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

Hemforth, B., Konieczny, L., Scheepers, C., Colonna, S., and Pynte, J. (2010). Language specific preferences in anaphor resolution: Exposure or gricean maxims? In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 2218–2223, Portland, USA.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):3.

Levesque, H. J., Davis, E., and Morgenstern, L. (2011). The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2016). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *arXiv preprint arXiv:1611.04146*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

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*.

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*.

Ward Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*.

Obvious for humans?

Bender (2015) found a 92% success rate for humans on the English collection.
See also:

`http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS2016SubjectTests.pdf`

- a subset of our schemas was used by psychology students for a self-paced reading experiment:
  replication of previous findings about language specific preferences in anaphora resolution (Hemforth et al., 2010)
- the whole set has been tested for human performance:
  - online questionnaires (Ibex Farm)
  - 22 participants recruited through RISC platform
  - removed data points where RT over $10''$ (and under 200 ms)
  - overall performance: 92.3% success rate
  - per item analysis in progress

google-proofness

Although we translated our schemas from the English collection of Levesque et al. (2011) that were at least partially checked to be Google-proof:

> *"In some cases where we were uncertain whether the schema was Google-proof, we have done some experiments with searches using Google's count of result pages. These counts, however, are notoriously unreliable (Lapata and Keller, 2005; Davis, 2015), so these "experiments" should be taken with several grains of salt."*

... we wanted to investigate further and more systematically whether obvious statistics does not help to solve our items.
We therefore defined a simple statistic test based on Mutual Information.

Non google-proof examples

(17) Many astronomers are engaged in the search for distant galaxies. They are spread all over the universe.
What are spread all over the universe?

(18) Pendant la tempête, l'arbre est tombé et s'est écrasé sur le toit de ma maison. Maintenant je dois le ⟨déplacer/réparer⟩.
Qu'est-ce que je dois ⟨déplacer/réparer⟩ ?
During the storm, the tree fell and crashed on the roof of my house. Now I have to ⟨remove/repair⟩ it.

Examples of spe/alt pairs

(19) J'ai sorti le portable de mon sac pour qu'[il] soit ⟨plus accessible/moins lourd⟩. *(101)*

(20) Le frère jumeau de Joël arrive toujours à le battre au tennis, même s'[il] a suivi deux ans de cours en ⟨moins/plus⟩. *(99)*

(21) Sandrine a appris que le fils d'Anne avait eu un accident ⟨donc/car⟩ [elle] l'a prévenue. *(98)*

(22) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'[ils] venaient de plus loin. *(93)*

(23) Fred est le seul homme encore vivant à se rappeler de mon arrière grand-père. [C'] ⟨est/était⟩ un homme remarquable. *(25)*