A Google-Proof Collection of French Winograd Schemas

Pascal Amsili and **Olga Seminck**
Laboratoire de Linguistique Formelle
Université Paris Diderot & CNRS
amsili@linguist.univ-paris-diderot.fr
olga.seminck@cri-paris.org

Abstract

This article presents the first collection of French Winograd Schemas. Winograd Schemas form anaphora resolution problems that can only be resolved with extensive world knowledge. For this reason the Winograd Schema Challenge has been proposed as an alternative to the Turing Test. A very important feature of Winograd Schemas is that it should be impossible to resolve them with statistical information about word co-occurrences: they should be *Google-proof*. We propose a measure of Google-proofness based on Mutual Information, and demonstrate the method on our collection of French Winograd Schemas.

1 Introduction

1.1 Winograd Schemas

Anaphora resolution depends on many factors from different linguistic levels. For example, grammatical role, number, gender, syntactic structure, phonological stress, distance between the referent and the anaphor and world knowledge all play a role. However, in automatic systems for anaphora resolution, rich semantics (world knowledge) is not often used. State of the art systems on the coreference task (Clark and Manning, 2016; Wiseman et al., 2015, *i.a.*) rely mostly on grammatical features, string matching features and some lexical semantic information (e.g., WordNet (Miller, 1995), named entities, or distributional semantics).

Winograd Schemas[1], as proposed by Levesque et al. (2011), form a special anaphora resolution challenge, because they cannot be resolved without a reasoning about world knowledge. A Winograd Schema is formed with a sentence containing an anaphor, along with a question about its antecedent and two possible answers (1). The correct answer should be obvious for a human.

(1) Nicolas could not carry his son because he was too ⟨weak⟩. Who was too ⟨weak⟩?
R0 : Nicolas
R1 : his son

The first sentence contains a word or an expression (labeled *special*) which can be replaced by another word or expression (*alternate*) in such a way that the sentence still makes sense, but the right answer to the question changes. For example in (1) if the special word 'weak' is replaced by 'heavy' (both in the sentence and in the question), the correct answer to the question is no longer R0, but R1. This property ensures that nothing in the overall structure of the schema prevents any NP to function as a possible antecedent.

According to this definition, for each Winograd Schema we get in fact two (related) questions (that we also call *items* in the rest of this paper).

1.2 Google-Proofness

Levesque et al. (2011) underline that the type of knowledge needed to resolve Winograd Schemas could be characterized as *thinking*, or *reasoning* — see also Levesque (2014). The idea is that Winograd Schemas cannot be resolved with only grammatical, or statistical information, nor any other non-semantic feature often used in standard coreference resolution systems. So in particular, the schemas should not be resolvable by typing the question and the answers into a search engine, such as Google, or by doing any obvious statistic test on a corpus. This feature is called *Google-proofness*. For instance, the item (2) is probably not Google-proof, because it is imaginable that

[1] Winograd Schemas are named after the examples Winograd (1972) used to illustrate the difficulty of natural language understanding.

Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017, pages 24–29,
Valencia, Spain, April 4, 2017. ©2017 Association for Computational Linguistics

"Galaxies are spread all over the universe" gets more Google hits than *"Astronomers are spread all over the universe."*.

(2) [2]Many astronomers are engaged in the search for distant galaxies. They are spread all over the ⟨universe⟩. What are spread all over the ⟨universe⟩?
R0 : the astronomers
R1 : the galaxies

On the other hand, for humans, Winograd Schemas should be obvious to resolve. Consequently, human performance should be near 100%. Indeed, Bender (2015) found a 92% success rate for humans on the English collection.

1.3 Test for Artificial Intelligence

Winograd Schemas can be seen as a difficult test of artificial intelligence. Indeed, Levesque et al. (2011) proposed the Winograd Schema Challenge (WSC) as an alternative to the Turing Test (Turing, 1950) —according to which a successful artificial intelligence system should be able to convince a human judge that it is human by conversing with him or her. In addition to the fact that resolving Winograd Schemas requires sophisticated reasoning, Levesque et al. (2011) argue that they overcome two majors issues of the Turing test. The first issue is that in order to pass the Turing test, a computer has to pretend to be human, in order to give human-like answers to questions like "How old are you?" or "Do you like chocolate?". The capacity to imitate a human behavior is in this respect orthogonal to the question of intelligence. The second issue of the Turing Test is the format of free conversation, which allows a system to use strategies to avoid answering difficult questions, for example by changing the subject, or making a joke. Winograd Schemas on the other hand, force the system to answer and do not allow evasive behavior.

1.4 State of the Art

In 2016 the first Winograd Schema Challenge was organized (Morgenstern et al., 2016). The task consisted of a pronoun disambiguation problem inspired by the format of Winograd Schemas.

Liu et al. (2016) submitted the winning system. It was based on unsupervised learning upon common sense knowledge bases and performed at a 58% success rate. After the WSC took place, the same group elaborated their system further, so that it was able to even attain a 66.7% success rate. It should be noted that the items that were used for the challenge were slightly different from the schemas that we have presented above: the items were not built in pairs through a common schema, and there were sometimes more than two antecedent candidates (3). As a consequence the baseline (chance level) for pronoun disambiguation problems was lower than 50% : 45% according to Liu et al. (2016).

(3) Mrs. March gave the mother tea and gruel, while she dressed the little baby as tenderly as if it had been her own.
As if <u>it</u> had been: tea / gruel / baby

Whereas the state of the art established by Liu et al. (2016) is much higher than the baseline, the result is still very far from the near 100% expected human score. Other systems that were not submitted to the competition can be found in literature; they often concentrate on a subset of schemas for which they developed a strategy for which we don't know how well it would generalize to the complete collection (Bailey et al., 2015; Schüller, 2014; Sharma et al., 2015, *i.a.*).

1.5 Our Contribution

Since the anaphora in the WSC can only be resolved with world knowledge, working on Winograd Schemas is an excellent way to develop models with rich semantic representations. We decided to provide a first collection of French schemas to encourage the development of these types of model for the French language. Having a French collection of schemas also enables more cross-linguistic comparison. Today there is a collection of 144 schemas for English that has been entirely translated into Japanese and 12 of the English schemas have also been translated into Chinese[3], but no documentation about the translation/adaptation method is provided. Our collection is also translated (or, rather, adapted) from the English set. We will say a few words about the adaptation process below.

While working on the adaptation of the English set, we also wanted to take seriously the constraint that Winograd Schemas should be Google-proof and therefore checked that our schemas were not

[2] taken from `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSfailed.html`

[3] These collections can be found on `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`

(too) sensitive to simple statistical information. We developed a simple method that uses corpus statistics to see if one of the two answers is more likely to be the correct one. We show that our method is not able to get a good score on the collection of schemas as a whole, even though certain items are more sensitive to statistics than others and if carefully selected, could rise the theoretical baseline score of 50% by 4 to 5 %.

2 Collection of French Schemas

Our collection contains 107 Winograd Schemas, which yields 214 questions. We based our set of French Winograd Schemas on the English collection of Levesque et al. (2011). We started by trying to translate the English version. This was challenging, because the schemas are only valid if they contain an anaphor with the same number and gender as the two possible answers. For example, for an item like (4), we cannot use a direct translation, as the word 'hair' in French (*cheveux*) is plural, while 'drain' (*siphon*) is singular.

(4) The drain is clogged with hair. It has to be [cleaned/ removed].

If the straightforward translation was not available, we tried to find another word that met the gender and number criteria, for example in (4) we replaced 'hair' with 'soap' (*savon*).

A second problem was that a literal translation could make one of the two versions of the schema ambiguous. Consider (5) with the alternate word ⟨indiscreet⟩. The French translation for 'indiscreet' is *indiscrète*. It turns out that in French *une personne indiscrète* — besides a person that reveals things that should stay secret — can also be somebody who *tries insistently to find out what should stay secret*, that is, a nosy person. In the French version of (5) we therefore changed the alternate to ⟨bavarde⟩ *(talkative)*.

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

We always privileged the most natural sounding solution and avoided long translations. Every item had to be validated by three native speakers of French. First, two interns translated the English schemas into French. Second, a third intern validated and improved the collection. And in the end, the entire collection was validated by the authors. Items that we could not find a solution for were excluded from our final set.

All our 107 Winograd Schemas can be freely downloaded from the following webpage: `http://www.llf.cnrs.fr/winograd-fr`. Every schema has a reference to the English schema it was translated from or inspired by.

3 Test of Google-Proofness

By Google-proofness we understand that there should be no obvious statistical test over text corpora that will reliably disambiguate the anaphor of an item correctly (Levesque et al., 2011).

Although we translated our schemas from the English collection of Levesque et al. (2011) that were at least partially checked to be Google-proof, we wanted to investigate further if obvious statistics does not help to solve our items. We therefore defined a simple statistic test based on Mutual Information.

3.1 Mutual Information

Mutual Information is a 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 (equation 1). (Ward Church and Hanks, 1990)

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

To test the Google-proofness of our schemas, for each question we measured the Mutual Information between the lexemes of the answers and the special, or the alternate. For example in the first item of (6), we measured *MI* between *sculpture* and *encombrer* and *étagère* and *encombrer* (7).

(6) La sculpture est tombée de l'étagère car elle était trop [encombrée/lourde].
Qu'est-ce qui était trop [encombré/lourd]?
R0 : la sculpture
R1 : l'étagère
The sculpture fell off the shelf because it was too [cluttered/heavy].
What was too [cluttered/heavy]?
R0 : the sculpture
R1 : the shelf

(7) *MI(sculpture, encombrer)* = 4.23
MI(étagère, encombrer) = 10.01

The simplest way to exploit these scores is to choose the answer which maximizes *MI* scores, so here for instance R1, which turns out to be the correct answer. However, the difference between the two scores, which ranges from .01 to around 10 in our data set, is likely to be, in some cases, too small to be reliable.

Therefore we introduce various thresholds of minimal difference between *MI* scores. We vary the threshold from 0 to 4 and observe the impact on accuracy.

3.2 Applicability of the measure

It should be noted that many items, in the original set as well as in ours, have proper nouns as possible answers (8). This in itself should ensure Google-proofness since cooccurrence frequencies of proper nouns with lexical nouns is likely to be random. In our set 44 schemas are of this sort, but we have decided to include them in the scores.

(8) [4]Steve follows Fred's example in everything. He [admires/influences] him hugely. Who [admires/ influences] whom?

An important aspect of our method is that it requires that there be a way to extract the words between which *MI* is to be computed. This method is in fact based on the comparison between the two possible answers. For instance, with (6), the two possible full answers for the question formed with the special word are:

- the sculpture was too cluttered (R0, special)
- the shelf was too cluttered (R1, special)

while the two possible answers for the question formed with the alternate word are:

- the sculpture was too heavy (R0, alternate)
- the shelf was too heavy (R1, alternate)

In such a case, it is obvious to find the pairs of words for which we want to compute *MI*.

However, some schemas do not offer the same possibility. Consider (9). In this case, since the answers do not include the special/alternate word, the pair of possible answers is exactly the same for both questions derived from this schema. So any *MI* score that could be computed are going to be the same for both questions, to which the correct answers are by construction different. We haven't included the 30 items of this sort in our scores.

(9) In the middle of the outdoor concert, the rain started falling, [and/but] it continued until 10. What continued until 10?
R0 : The rain
R1 : the concert

We have 107 schemas in our collection, which yields 214 questions. 30 items were removed for the reasons we have just exposed, and 2 more schemas were removed because the possible answers R0 and R1 comprise the special/alternate words. All together, we measured Mutual Information for 90 schemas (180 items).

3.3 Probability Estimation

To estimate Mutual Information we used unsmoothed frequency counts from FrWaC, the French version of Web as a Corpus (Baroni et al., 2009), which is a corpus of 1.6 billion tokens from the `.fr` domain of the Internet. If the answers, the special, or the alternate were formed by multiple words, we took the frequency counts of the lexical head. Except in a few special cases, we measured the frequencies of lemmas rather than word-forms. 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).

3.4 Results

In Table 1 we can see the accuracy of the statistical method based on *MI* for different thresholds of difference in the scores of the two answers. Out of the 180 items we considered, 49 items could not get a score, because either one of the words did not appear at all or the coocurrence was not found in the corpus.

One should keep in mind that answering at random would give an accuracy around 50%. So the accuracy we get when no threshold is applied (55%) is clearly not satisfactory, and suggests, as we expected, that using any difference in *MI* scores is very similar to answering at random. The accuracy score reaches 70% however for a threshold of Δ 2.5, which is much better, but then the number of items to which the method applies is drastically small, namely less than 15% of the items.

The curves on Figure 1 plot accuracy and coverage as given in Table 1, along with another measure, that we call success rate. It is the theoretical accuracy that we would get by answering at random for items for which the *MI* difference is be-

[4] taken from `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`

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

Table 1: Results of the statistical method based on Mutual Information. Different thresholds give the minimal difference between the scores $I(R0, \mathit{special})$ and $I(R1, \mathit{special})$ that should be attained before the system can answer. '# 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 Mutual Information; if the method did not respond due to lack of counts or too high a threshold, this was counted as an error.

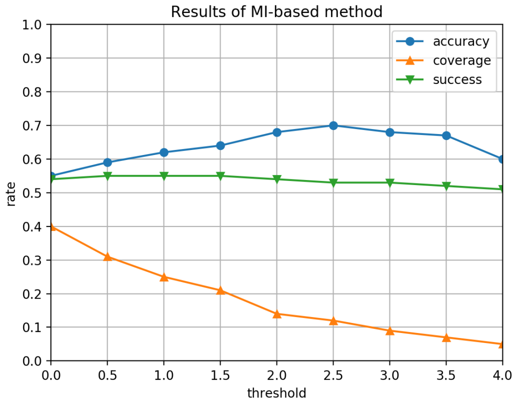

Figure 1: Results of the statistical methods based on Mutual Information. 'Accuracy' is the number of correct answers for the questions for which the method applies, while 'coverage' corresponds to the number of correct answers divided by the total number of questions. 'Success' is the theoretical success 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.

low the threshold, and using the MI difference for items for which it is above the threshold. We can see that this success rate never goes over 55%.

3.5 Discussion

Our collection as a whole seems to be Google-proof. Using Mutual Information as a strategy to resolve the schemas, we could not exceed a score of 55% success rate on the entire corpus, whichever threshold we used. However, there are a few cases where Mutual Information can be helpful (when the difference is high enough), which might still bring an improvement to a WSC system and one can easily imagine more sophisticated methods that would do better.

However, we would like to underline that we chose specifically not to use a sophisticated method. According to the concept of Google-proofness, Winograd Schemas should not be resolvable by obvious statistics. This raises the question where the boundary between obvious and smart statistics lies. For example, can we consider that a method such as word2vec (Mikolov et al., 2013) falls into the category of obvious statistics? Because we are not sure, we do not make the claim that the collection would resist any statistical test. But we are confident that it resists statistical test of the same level of simplicity as our mutual information measure.

4 Conclusion

Winograd Schemas, often referred to as a *new* Turing test, form an interesting AI problem. The schemas represent anaphora resolution problems that can only be resolved by rich semantic representations. To encourage research on the problems Winograd Schemas pose, we developed the first French Winograd Schema Collection. We investigated if our schemas could resist an obvious statistical method of resolution based on Mutual Information. It appeared that our collection is robust: only a small gain of 4 to 5% could be obtained by using the method.

Acknowledgments

We thank Sarah Ghumundee, Biljana Knežević, and Nicolas Bénichou who helped us prepare the items and compute mutual information scores. We also thank the three anonymous reviewers of the CORBON workshop for their feedback on our article. This work was supported in part by the École Doctorale Frontières du Vivant — Programme Bettencourt.

References

Dan Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

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

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *MAICS*, pages 39–45.

Kevin Clark and D. Christopher Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653. Association for Computational Linguistics.

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

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Hector J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212:27–35.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *arXiv preprint arXiv:1611.04146*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

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

Claude E. Shannon and Warren Weaver. 1949. The mathematical theory of information.

Arpit Sharma, Nguyen H. Vo, Somak Aditya, and Chitta Baral. 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*.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*.

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

Sam Wiseman, M. Alexander Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426. Association for Computational Linguistics.