EPiC Series in Computing

Volume 55, 2018, Pages 107–120

GCAI-2018. 4th Global Conference on Artificial Intelligence

EPiC Computing

A Data-Driven Metric of Hardness for WSC Sentences

Nicos Isaak[1] and Loizos Michael[2]

[1] Open University of Cyprus, Nicosia, Cyprus
nicos.isaak@st.ouc.ac.cy

[2] Open University of Cyprus, Nicosia, Cyprus
& Research Center on Interactive Media,
Smart Systems, and Emerging Technologies
loizos@ouc.ac.cy

Abstract

The Winograd Schema Challenge (WSC) — the task of resolving pronouns in certain sentences where shallow parsing techniques seem not to be directly applicable — has been proposed as an alternative to the Turing Test. According to Levesque, having access to a large corpus of text would likely not help much in the WSC. Among a number of attempts to tackle this challenge, one particular approach has demonstrated the plausibility of using commonsense knowledge automatically acquired from *raw text* in English Wikipedia.

Here, we present the results of a large-scale experiment that shows how the performance of that particular automated approach varies with the availability of training material. We compare the results of this experiment with two studies: one from the literature that investigates how adult native speakers tackle the WSC, and one that we design and undertake to investigate how teenager non-native speakers tackle the WSC. We find that the performance of the automated approach correlates positively with the performance of humans, suggesting that the performance of the particular automated approach could be used as a metric of hardness for WSC instances.

1 Introduction

One of the essential challenges in Computer Science is to understand how we can create systems that procure and manage commonsense knowledge [18]. A number of challenges have been proposed that aim for systems that will replace or substitute basic human abilities, so that we can relate and interact with them. One of these challenges is the Winograd Schema Challenge (WSC) [8], an alternative to the well-known Turing test believed to be able to provide a more meaningful measure of machine intelligence [2].

The WSC is effectively a carefully-crafted pronoun resolution task. One potentially promising approach to handle this challenge builds natural language representations and supports the necessary reasoning with the available information by acquiring knowledge in the form of general inference rules [6, 10, 16]. This paper presents the results of a large-scale experiment to see how this kind of approach can be used as a data-driven metric of hardness for WSC sentences. We compare the results of this experiment with two studies: one from the literature [3] that investigates how adult native speakers tackle the WSC, and one that we design and undertake

D. Lee, A. Steen and T. Walsh (eds.), GCAI-2018 (EPiC Series in Computing, vol. 55), pp. 107–120

to investigate how teenager non-native speakers tackle the WSC. To date, no study has looked specifically at how the amount of training material for a learning-based approach to the WSC can be used as a data-driven metric of hardness for WSC sentences, and any evidence for this has been mainly anecdotal.

According to Bender [3], certain people are unfamiliar with certain concepts in WSC sentences, and their performance ends up being correlated with this familiarity. Instead of being oblivious to such issues, the use of our proposed metric, trained on appropriately selected training data, can be used to provide an a priori level of objective hardness of WSC sentences so that the challenge can be personalized to the strengths and weaknesses of a particular group of *human* participants. We are not claiming, however, that this metric can be used to anticipate how hard it is for *machines* to resolve certain WSC sentences, nor, by extension, that it can be used to select material for WSC competitions that test the progress of machines on the WSC.

The purpose of this work is to investigate whether one can design an automated system — and, in fact, we reuse an existing WSC system to that end — whose performance *varies* across WSC sentences in the same way that *human* performance varies across WSC sentences. Showing a positive correlation of the performance of the system with the performance of humans would suffice to offer evidence that the system can be used to automatically differentiate between WSC sentences based on their perceived hardness for humans. The system does not purport to replicate the cognitive mechanisms used by humans when solving the WSC, but only to offer a phenomenological account of this perceived hardness.

The system considered in this work is one that effectively improves its behavior as it gets more training data. Since the WSC is claimed to require commonsense knowledge to be solved by humans, this might suggest that WSC instances that are harder for humans are the ones that require more training, and hence more effort to identify the right knowledge; or, put differently, harder instances are the ones that require the use of commonsense knowledge that is less common. Our experiment supports this hypothesis by showing that the system's performance is correlated with human performance. In particular, adults asked to solve the WSC are shown to perform better than teenagers. Since age is generally correlated with more experiences, and thus the acquisition of knowledge that might be less common, this is in line with the above hypothesis.

In the sections that follow, we proceed to present the WSC with some highlights on previous work, and focus, in particular, on the system that we use for our experiment. We continue to outline the methodology we follow to demonstrate how the size of the training corpus affects the performance of the system, and then present a study that we have undertaken to measure the performance of teenagers on the WSC. We finally discuss how our findings support the use of the system for determining the hardness of WSC sentences, along with potential implications and directions for future research.

2 The Winograd Schema Challenge

Each schema in the WSC [8] comprises two nearly identical sentences with clear but very different meanings (twin sentences), both sharing a definite pronoun and two potential co-referents. Due to the difference of a certain key phrase in the two sentences, the pronoun is naturally resolved to a different co-referent in each sentence. Given one of the two sentences, then, the task is to resolve the definite pronoun to the correct co-referent. To avoid trivializing the task, the co-referents are of the same gender and number, and one has to rely critically on the key phrase to determine the right answer.

The following WSC schema (*catch example*) illustrates how difficult the problem can be: *1.)*

The cat caught the mouse because it was clever. Question: Who is clever? Answers: cat, mouse 2.) The cat caught the mouse because it was careless. Question: Who is careless? Answers: cat, mouse. It has been argued that to reliably answer such questions, machines might need to engage in human-like reasoning and capitalize on the use of commonsense knowledge.

A number of approaches have been proposed in the literature to tackle the WSC, while the AI community has sought to promote the WSC through specialized competitions, the first of which was organized by Nuance Communications during the 2016 edition of IJCAI [1]. Below we review some of the existing WSC tools and techniques, focusing on how they acquire knowledge.

Rahman and Ng's system [15] attempts to identify the most probable co-referent through a number of lexicalized statistical techniques, using an *SVM ranking-based approach* that combines the features derived from different knowledge resources like Web Queries, Framenet, OpinionFinder, English Giga world, BLLIP and Reuters. Related is the approach taken by the Budukh system [5], which uses an aggregation mechanism over four answering modules that, correspondingly, use world knowledge from ConceptNet, Web Queries, Narrative chains and sentiment analysis.

Another work [14] approaches the problem as an instance of Integer Linear Programming, and acquires statistics in an unsupervised manner from multiple knowledge resources, like Gigaword corpus, Wikipedia Wikifier, Web Queries and polarity information. Sharma's system [17] is based on Answer Set Programming, and attempts to retrieve the needed background knowledge directly from the *Google* search engine, through the use of certain queries.

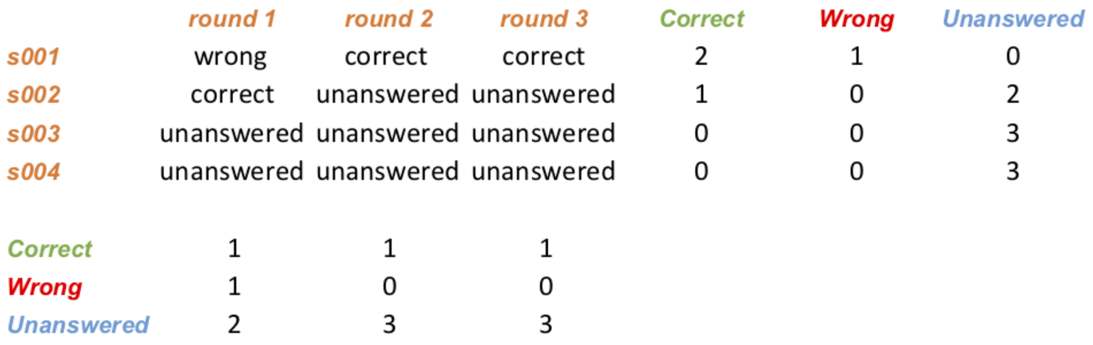
2.1 The Wikisense Approach

In this work we focus on the *Wikisense* system [6] for the WSC. Unlike certain other WSC systems [14, 15, 17], the *Wikisense* system has a particular online flavor, in that it first considers the WSC sentence at hand, and then retrieves the relevant *training material* on which it is trained. It is, therefore, straightforward to adapt the amount of training material that will be made available to the system, and consider the effects of data availability on its performance.

The *Wikisense* system is based on the *Websense* engine [12], which uses raw text as a source of training material [10, 13], and a form of supervised learning, called *autodidactic* [11], to acquire background knowledge in the form of logical inference rules that can be reasoned with. The *Wikisense* system focuses on the use of the English Wikipedia as a training corpus, and selects the relevant pieces of text that will be retrieved for training, based on the WSC sentence at hand. It, then, uses the acquired background knowledge to respond with the answer that is implied by the WSC sentence.

To select relevant text from Wikipedia, the *Wikisense* system creates multiple keyword-queries based on the given WSC sentence. For instance, for the first sentence in the *catch example*, it creates the following set of four queries *catch/clever, cat/mouse/catch, cat/clever, mouse/clever.* For every query in turn, the system retrieves a number of sentences from Wikipedia that match the query, as specified by a parameter of the system. Using those sentences as training material, the system determines if it can conclude that one of the two answers of the WSC sentence can be inferred. If not, it attempts to use the subsequent query and repeats the process.

Starting from the raw text training material, the *Wikisense* system utilizes the dependency parsers *Spacy* and *Stanford Parser* to turn raw text into semantic relations. These relations act, in turn, as the features of learning examples from which inference rules are induced, following the approach in [13]. In case sufficiently confident rules are identified (based on the weights given to the rules by the learning algorithm), those rules are used to draw inferences about the



	round 1	round 2	round 3	Correct	Wrong	Unanswered
s001	wrong	correct	correct	2	1	0
s002	correct	unanswered	unanswered	1	0	2
s003	unanswered	unanswered	unanswered	0	0	3
s004	unanswered	unanswered	unanswered	0	0	3

Correct	1	1	1
Wrong	1	0	0
Unanswered	2	3	3

Figure 1: A snapshot of the results of the *Wikisense* system with $S = 1 \cdot 10^1$.

WSC sentence. More details about the process can be found in the paper that introduced the *Wikisense* system [6].

3 Corpus-Level Analysis

To evaluate how the size of the training corpus affects the performance of the *Wikisense* system, we ran the system with varying values of the parameter S that specifies how many Wikipedia sentences are retrieved for training purposes. In particular, we let S range over the following 12 values: $1 \cdot 10^1$, $2 \cdot 10^1$, $5 \cdot 10^1$, $1 \cdot 10^2$, $2 \cdot 10^2$, $5 \cdot 10^2$, $1 \cdot 10^3$, $2 \cdot 10^3$, $5 \cdot 10^3$, $1 \cdot 10^4$, $2 \cdot 10^4$, $5 \cdot 10^4$. We tested the system on the first 100 WSC sentences (labeled $s001$ to $s100$) from a standard WSC Library.[1] For each WSC sentence, and for each value of S, the system was run for 100 rounds, and the number of times that the system responded correctly, responded incorrectly, or did not respond was recorded (cf. Figure 1).

3.1 Corpus Analysis Results

Figure 2 reveals that as the size of the training set increases, the number of unanswered WSC sentences decreases, while the numbers of both the correctly answered and incorrectly answered WSC sentences increase, with the latter seemingly increasing at a lower rate. The null hypothesis that the size of the training set does not affect the performance of the *Wikisense* system in terms of the correct answers it produces can, therefore, be rejected using an ANOVA analysis that gives F = 20.860 > Fcrit = 3.2849, showing that the means of the three populations (correct, wrong, unanswered) in Figure 2 are not equal.

As shown in Figures 2 and 3, the number of unanswered WSC sentences monotonically reduces as the training set size increases. The only exception to this monotonicity is when $S = 1 \cdot 10^4$, where we observe an increase of 0.36%, benefiting the number of correctly answered WSC sentences. This is the point in the graph where the distance between the correctly answered and the incorrectly answered WSC sentences is the largest (8%).

Comparing the performance of the system when $S = 1 \cdot 10^3$ — this being the default value used in earlier work [6] — to the system's performance when $S = 5 \cdot 10^4$, we can see a measurable increase of 5%, which suggests that the performance of Wikisense as reported in earlier work can be further improved with the simple adjustment of the training set.

[1] http://www.cs.nyu.edu/faculty/davise/papers/OldSchemas.xml

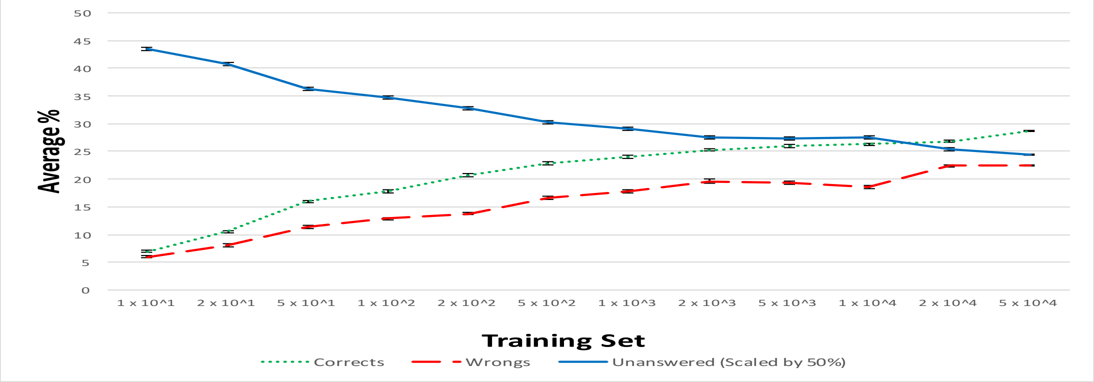


Figure 2: Performance evaluation (along with standard errors) on the entire corpus across different values of S.

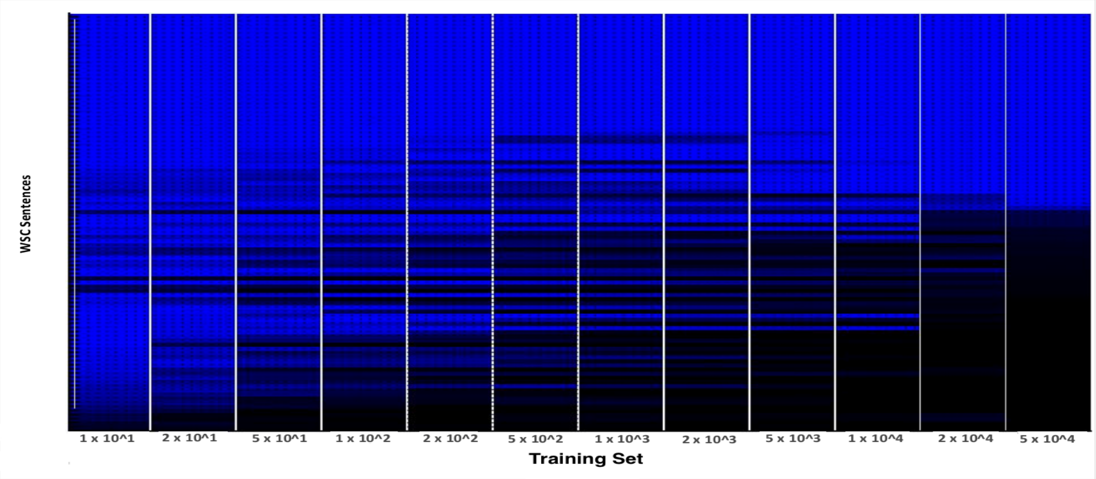


Figure 3: Color intensity shows how often (among 100 rounds) each WSC sentence on the Y axis has been answered (correctly or incorrectly), for each value of S on the X axis. The WSC sentences on the Y axis have been reordered based on the percentage with which they have been answered when $S = 5 \cdot 10^4$.

Finally, Figure 4 shows the system's performance for each of the 100 rounds that were run for the two extreme values of S, demonstrating a consistent (not simply on average, but on each individual round) ability of the system to answer correctly more often than incorrectly when S is larger. Thus, not only larger training sets lead to less unanswered WSC sentences, but among those that are answered, the percentage of the correctly answered ones tends to become larger than the percentage of the incorrectly answered ones.

Overall, larger training sets seem to lead the *Wikisense* system to answer more WSC sentences, and among those answered, to answer correctly more often (cf. Figure 5). Given the knowledge-based workings of the *Wikisense* system, this could be taken as an indication that richer and more useful knowledge is acquired from larger training sets. This, of course, should not be taken to conflict with the *Google-proofness* of the WSC: the claim that statistical in-

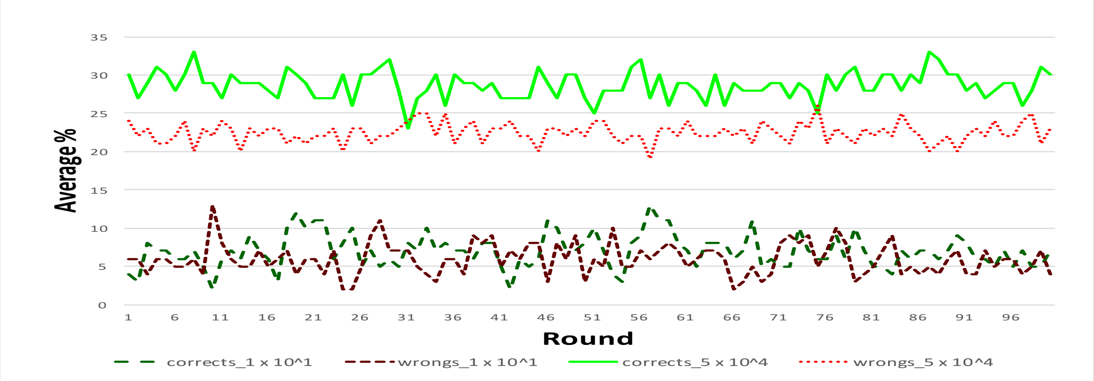


Figure 4: Percentages of the correctly answered and incorrectly answered WSC sentences in each round. The plot shows these percentages for the two extreme values of S.

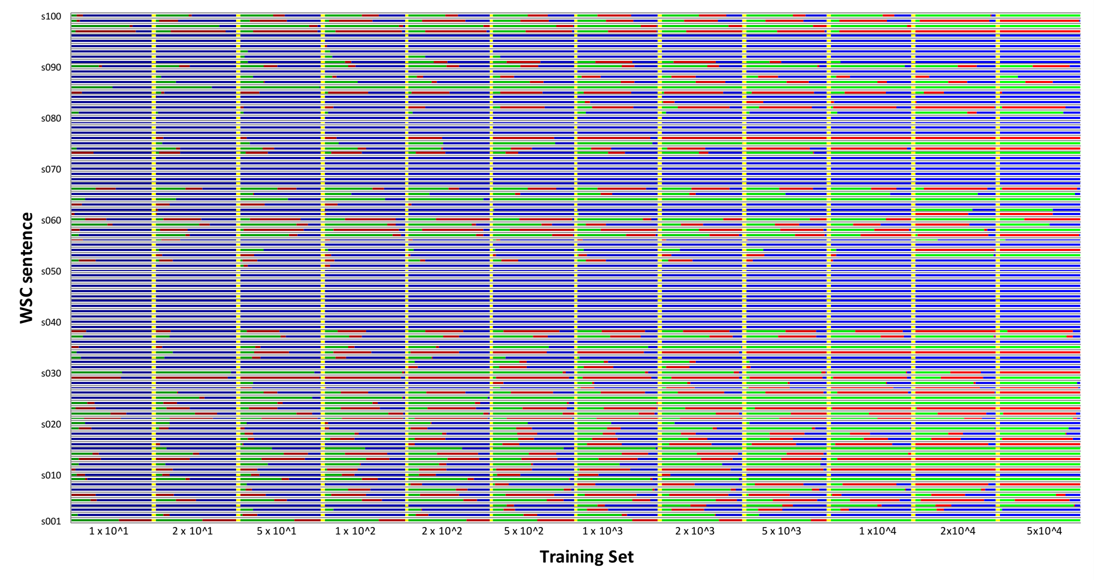


Figure 5: The coloring of each horizontal bar indicates the percentage of rounds in which each WSC sentence on the Y axis was correctly answered (green color), incorrectly answered (red color), or unanswered (blue color), for each value of S on the X axis.

formation from Google search results is insufficient to reliably address the WSC. Our results indicate only that more information helps *improve* performance, not that it suffices to achieve *human-level* performance. On the other hand, we will offer evidence in the sequel that even without achieving human-level performance, one can still use the *Wikisense* system to determine how hard a WSC sentence might be for humans.

4 Human Performance on the WSC

In this section we present evidence from two studies in support of the claim that the performance of the *Wikisense* system *varies* across WSC sentences in a manner analogous to the performance of humans. The first study comes from the literature, and concerns adult native speakers, while the second study was designed as part of this work, and concerns teenager non-native speakers.

4.1 Adult Native Speakers

In terms of the performance of humans on the WSC, the literature [3] establishes a baseline with adult speakers — residents of the United States — who speak English fluently. Using Amazon's Mechanical Turk to run the experiment on all sentences from the standard WSC Library that we have also used in our analysis of the *Wikisense* system, Bender shows that native English speakers are, on average, able to correctly resolve 92.1% of the WSC sentences (91%, if we consider only the first 100 WSC sentences). A detailed analysis of human performance on each individual WSC sentence is available from: `https://github.com/benderdave/wsc-exp.git`.

4.2 Teenager Non-Native Speakers

We undertook an analogous study to that of Bender in December 2017, albeit the study was carried out in a lab setting, with the voluntary participation of 126 teenager (aged 11-15) English-speaking students of secondary education in Cyprus. In terms of their knowledge of the English language, 37 reported that it was "good", 66 that it was "very good", and 23 that they speak English fluently (out of which, 9 mentioned that English is their mother tongue). All participants had experience with the WSC, as they had previously participated in another study that involved the WSC (although that study was in Greek).

We split the 100 WSC sentences that were used in the evaluation of the *Wikisense* system into four equal sets, ensuring that no set included both twin sentences from the same schema. Participants were asked to answer the questions of the WSC sentences in one of the sets. The participation was anonymous, and it lasted about 10 minutes during school break-time between lessons. The study was undertaken in the school's Computer Science labs under supervision by a teacher. Each WSC sentence was displayed on a screen, followed by the question. Two choices were displayed side-by-side with a comment box below each question. Access to translation services was not allowed, and each participant was instructed to write any remarks (on whether a question was confusing or non-intuitive) in the comment box. Participants were offered a €0.50 chocolate bar as a compensation for their time.

Based on the results of the study, teenagers scored a mean accuracy of 60.77% ($\sigma = 0.16$). The 9 teenagers with English as their mother tongue scored a mean accuracy of 54.83%, offering an indication that the lower performance of the teenager group compared to the adult group might not be a result of the teenagers being non-native speakers, but a result of their age. Beyond the performance difference, it is worth noting that the two groups had a positive correlation of 0.43, suggesting that those WSC sentences that were harder to answer by one group were also harder to answer by the other group.

5 Measuring WSC Sentence Hardness

Using the data from the two aforementioned studies, we examine in this section whether the performance of the *Wikisense* system can be predictive of the hardness of the WSC sentences for humans. As a baseline, we compare the predictive ability of the system against that of other co-reference resolution systems.

5.1 A Boolean Hardness Metric

We start with the simple approach of characterizing a WSC sentence as either “easy” or “hard” depending on whether it can be resolved correctly or incorrectly by an automated system.

For our *Wikisense*-based approach, we proceed as follows: For a given WSC sentence, and for a given value of S, we run the *Wikisense* system for 100 rounds and record the most frequent result returned by the system. Thus, we are able to determine if most of the time the system responded with the first answer, with the second answer, or abstained from responding. We repeat the process for all 12 possible values of S, as described in preceding sections. If the majority of these 12 repetitions yield the same answer, then we take that to be the answer of the approach. We, then, check to see if the answer is correct or not, characterizing, respectively, the WSC sentence as “easy” or “hard”; some WSC sentences remain uncharacterized.

To compare this boolean hardness metric against what can be derived from other systems, we consider three co-reference resolution systems. For each system, a WSC sentence is characterized as “easy” or “hard” (or remains uncharacterized) depending on whether the system is able to correctly or incorrectly resolve the WSC sentence (or does not produce an answer).

Based on the characterizations of WSC sentences by each of the four considered approaches, we group the WSC sentences into an “easy” and a “hard” group, and compare the performance of humans on these two groups to see whether their performance varies. The results are summarized in Table 1, which shows that the boolean hardness metric derived from the *Wikisense*-based approach can discriminate better between what humans find easy and hard in the WSC. In particular:

***Wikisense*-Based Approach:** The WSC sentences characterized as “easy” and “hard” can be resolved by adults with a mean accuracy of 93% ($\sigma = 0.08$) and 87% ($\sigma = 0.12$), respectively, compared to their overall mean accuracy of 91%. Analogously, the WSC sentences characterized as “easy” and “hard” can be resolved by teenagers with a mean accuracy of 66% ($\sigma = 0.16$) and 57% ($\sigma = 0.17$), respectively, compared to their overall mean accuracy of 60.77%.

Stanford Core NLP [9]: The WSC sentences characterized as “easy” and “hard” can be resolved by adults with a mean accuracy of 90% ($\sigma = 0.12$) and 93% ($\sigma = 0.08$), respectively, showing a negative correlation with the human performance. The same phenomenon appears with teenagers, where the WSC sentences characterized as “easy” and “hard” can be resolved with a mean accuracy of 60% ($\sigma = 0.14$) and 62% ($\sigma = 0.17$), respectively.

Illinois Co-reference Resolver [4, 14]: The WSC sentences characterized as “easy” and “hard” can be resolved by adults with a mean accuracy of 93% ($\sigma = 0.07$) and 91% ($\sigma = 0.10$), respectively, showing a smaller discriminatory power than our proposed approach. This is even more evident with teenagers, where the WSC sentences characterized as “easy” and “hard” can be resolved with a mean accuracy of 62% ($\sigma = 0.16$) and 61% ($\sigma = 0.14$), respectively.

Knowledge Parser [17]: This system was built for the WSC, yet its performance seems to be non-predictive of human performance. The WSC sentences characterized as “easy” and “hard” can be resolved by adults with a mean accuracy of 89% ($\sigma = 0.13$) and 93% ($\sigma = 0.08$), respectively. Analogously, the WSC sentences characterized as “easy” and “hard” can be resolved by teenagers with a mean accuracy of 57% ($\sigma = 0.14$) to 62% ($\sigma = 0.16$), respectively, showing an important gap in the wrong direction.

The results ultimately show that the performance of the *Wikisense*-based approach *varies*

	adults		teenagers	
	“easy”	“hard”	“easy”	“hard”
Stanford NLP	0.90	0.93	0.60	0.62
Illinois Coref.	0.93	0.91	0.62	0.61
Kparser	0.89	0.93	0.57	0.62
Wikisense-based	0.93	0.87	0.66	0.57

Table 1: Predictive behavior of human performance from simple boolean hardness metrics derived from automated systems.

across WSC sentences in a manner that resembles the variability of the human performance more closely than what other systems can achieve.

5.2 A Real-Valued Hardness Metric

As afforded by the *Wikisense* system’s online access to training material, and aiming to derive a more fine-grained metric of hardness, we consider next a certain way of deriving a real-valued (as opposed to a boolean) hardness index for each WSC sentence.

As in the previous subsection, given a WSC sentence and a value of S, we run the *Wikisense* system for 100 rounds and record the most frequent result returned by the system. Thus, we are able to determine if most of the time the system responded with the first answer, with the second answer, or abstained from responding. For each case where the response is one of the two answers, we check and mark the answer as correct or incorrect. We repeat the process for all 12 possible values of S, and end up with a set of 12 labels. Intuitively, if all of these labels are “unanswered”, we do not have enough information to give a hardness index to the sentence. This particular approach ends up giving a hardness index to 57 out of the 100 WSC sentences under consideration, and our subsequent discussion refers to only these 57 instances.

Now, consider the case where at least one label is “correct”, and therefore, the system has identified, at least once, knowledge that is relevant to, and *appropriate* for, the particular WSC sentence. The more “correct” labels one has, then, the easier it would seem that this WSC sentence is. Taking into account that out of the cases with an “unanswered” label one could randomly guess the correct answer half of the time, we can adjust the number of “correct” labels to also include half of the “unanswered” labels. Normalizing this value by dividing by 12, we end up with a number in the interval $[0, 1]$ that is higher for easier WSC sentences. Taking 1 minus this value gives us the hardness index of the sentence.

If none of the labels is “correct”, and since we compute a hardness index only if there is at least one label that is not ‘unanswered”, it must be the case that there exists at least one “incorrect” label. Therefore, the system has identified, at least once, knowledge that is relevant to, but *inappropriate* for, the particular WSC sentence. One could argue that the more “incorrect” labels one has, then, the harder this WSC sentence should be. But given the simple approach that the *Wikisense* system follows in retrieving relevant training data, one could also make another argument. Since there are *no* “correct” labels, the more “incorrect” labels one has should simply be taken as an indication of the availability of more relevant knowledge, ignoring the fact that it led to the wrong answer. The availability of knowledge suggests, then, an easier WSC sentence. Taking into account, as before, that out of the cases with an “unanswered” label one could randomly guess the incorrect answer half of the time, we can adjust the number of “incorrect” labels to also include half of the “unanswered” labels. Normalizing this value by dividing by 12, we end up with a number in the interval $[0, 1]$ that is

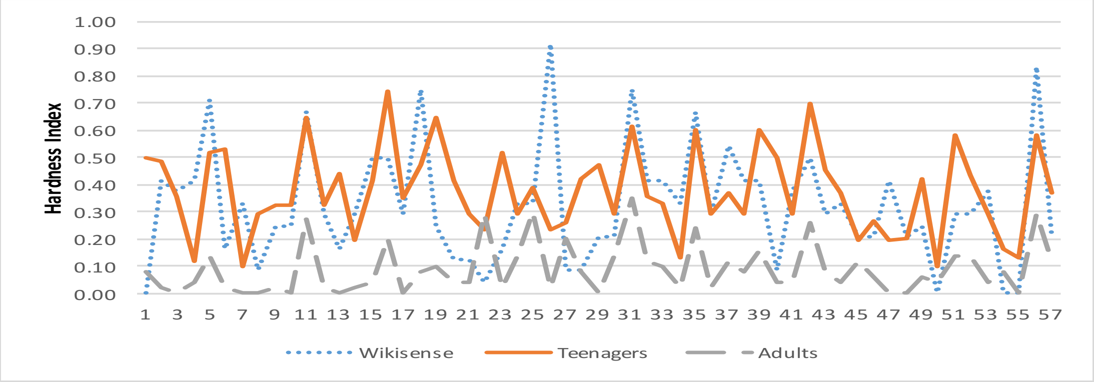


Figure 6: Variability of our developed *Wikisense*-based hardness index across the 57 WSC sentences on which it was computed, in relation to the variability of the human hardness index for adults and teenagers.

higher for easier WSC sentences. Taking 1 minus this value gives us the hardness index of the sentence.

In terms of the human performance data, we treat the human hardness index of a WSC sentence to be the percentage of people from a certain group that resolved the sentence incorrectly. Our computed hardness index and the human hardness index for the adult and the teenager groups in our discussed studies have correlation coefficients of 0.38 and 0.37, respectively. Both results offer evidence that our proposed computed hardness index might be indicative of how humans perceive the hardness of the WSC, and that this indication might not be significantly affected across different human groups.

Figure 6 shows in more detail how the computed hardness index and the human hardness index vary across WSC sentences, suggesting that indeed, certain WSC sentences that are more easy or hard for humans are accordingly labeled as such by the computed hardness index. The figure also shows that despite the teenager group performing almost consistently worse that the adult group, their performance across WSC sentences seems to vary analogously.

Our developed tool, which takes as input a WSC sentence and outputs its hardness index, is available online at `http://cognition.ouc.ac.cy/ws_hardness`. The tool can adjust the conditions under which it chooses to produce a hardness index or abstain from producing one. For example, if the parameters are appropriately adjusted to compute a hardness index for only 10% of the tested WSC sentences, the correlation coefficient against the teenager group becomes 0.70.

6 Qualitative Analysis

Based on the remarks submitted by the teenager participants in our study, we present below a qualitative analysis that relates those remarks to the performance of the *Wikisense*-based approach that we have developed.

Unanswered Sentences. 27 WSC sentences remained *unanswered* in all rounds across all training sets. The sentence *I couldn't put the pot on the shelf because it was too tall. Question: What was too tall?* was accompanied by a remark that it was very confusing; the mean adult accuracy was 45% and the mean teenager accuracy was 37%. The sentence *Frank was upset with*

Tom because the toaster he had bought from him didn't work. Question: Who had bought the toaster? was accompanied by a remark that it was very difficult; the mean adult accuracy was 75% and the mean teenager accuracy was 50%. For the sentence *Pete envies Martin although he is very successful. Question: Who is very successful?* the mean adult accuracy was 84% and the mean teenager accuracy was 35%. The sentence *The lawyer asked the witness a question, but he was reluctant to repeat it. Question: Who was reluctant to repeat the question?* was accompanied by a remark on not understanding the meaning of "reluctant"; the mean adult accuracy was 63% and the mean teenager accuracy was 32%.

The *Wikisense* system was not able to formulate a query to retrieve training data in 4 of these 27 sentences. In some other cases, despite formulating a query (e.g., lie/cautious), the system was unable to retrieve enough training data to create the necessary knowledge. The sentence *The cat was lying by the mouse hole waiting for the mouse, but it was too cautious. What was too cautious?* was accompanied by a remark on not understanding its meaning; the mean adult accuracy was 90% and the mean teenager accuracy was 42%. The sentence *In the middle of the outdoor concert, the rain started falling, but it continued until 10. Question: What continued until 10?* was accompanied by the remak that it was an interesting sentence; the mean adult accuracy was 60% and the mean teenager accuracy was 53%.

Correctly-Resolved Sentences. There were 3 WSC sentences that were correctly resolved across all training sets: i) *The city councilmen refused the demonstrators a permit because they feared violence. Question Who feared violence?*, ii) *Bob paid for Charlie's college education, but now Charlie acts as though it never happened. He is very ungrateful. Question: Who is ungrateful?*, iii) *Anne gave birth to a daughter last month. She is a very charming baby. Question: Who is a charming baby?*.

The *Wikisense* system resolved the first sentence through the query refuse/fear. It might be considered as an easy sentence, because the subject of the verb "refuse" is the one who fears that something is going to happen, and that the query directly leads to the correct pronoun target. 50% of the teenagers, though, did not manage to resolve the pronoun correctly, compared to only 8% of adults. Two teenagers commented that they found it very difficult, with one specifying that they did not know the meaning of the word "councilmen". No remarks were received on the second and third sentences. On the second sentence the mean teenager accuracy was 90% and the mean adult accuracy was 96%, while on the third sentence the mean teenager accuracy was 87% and the mean adult accuracy was 100%.

There were sentences that the system was able to resolve correctly only when the size of the training set was sufficiently large. For example, the sentence *Jim yelled at Kevin because he was so upset. Question: Who was upset?* was correctly resolved only with the two largest training set sizes. Three teenagers remarked that the sentence was difficult; the mean teenager accuracy was 53% and the mean adult accuracy was 100%. As another example, the sentence *Paul tried to call George on the phone, but he wasn't successful. Question: Who was not successful?* was correctly resolved from the fourth training set size onwards; the mean teenager accuracy was 43% and the mean adult accuracy was 98%. Finally, for the sentence *There is a gap in the wall. You can see the garden behind it. Question: You can see the garden behind what?* only 40% of the teenager managed to resolve it, compared to 85% of the adults.

Incorrectly-Resolved Sentences. On the other hand, there were queries that led the system to wrong conclusions. For example, the sentence *Anne gave birth to a daughter last month. She is a very charming woman. Question: Who is a charming woman* was wrongfully resolved across all training set sizes. The query give/charming ended up producing more training data in support of the inference *daughter*, as there seem to be more training sentences for charming children than for charming adults. On the other hand, humans do not typically refer to a female newborn as a woman; the mean teenager accuracy was 84% and the mean adult accuracy was 92%.

The sentence *Alice tried frantically to stop her daughter from chatting at the party, leaving us to wonder why she was behaving so strangely. Question: Who was behaving strangely?* was accompanied by the remark that it was odd; the mean teenager accuracy was 40% and the mean adult accuracy was 71%.

There were also WSC sentences that were correctly resolved with smaller training sets but incorrectly resolved with larger training sets. For instance, the sentence *Tom threw his schoolbag down to Ray after he reached the bottom of the stairs. Question: Who reached the bottom of the stairs?* was correctly resolved only until the ninth training set. Teenagers correctly resolved the sentence 35% of the time, while adults 90% of the time.

Confusing Sentences. For certain WSC sentences there was no obvious relation between the system's performance and the training set size. For instance, in the WSC sentence *Frank felt vindicated when his longtime rival Bill revealed that he was the winner of the competition. Question: Who was the winner of the competition?* the performance of the system across the 12 training set sizes started with not producing an answer and flipped back and forth between producing the right and the wrong answers as the training set sizes increased. Such sentences might be confusing even for humans; the mean teenager accuracy was 35% and the mean adult accuracy was 73%. A teenager remarked that this sentence was very difficult.

As another example, the mean accuracy on the sentence *The sack of potatoes had been placed below the bag of flour, so it had to be moved first, Question: What had to be moved first?* was 35% for teenagers and 69% for adults, whereas the mean accuracy on the sentence *My meeting started at 4:00 and I needed to catch the train at 4:30, so there wasn't much time. Luckily, it was delayed so it worked out. Question: What was delayed?* was 30% for teenagers and 74% for adults, with two teenagers remarking that it was very confusing.

Twin Sentence Issues. In analyzing the behavior of the *Wikisense* system on twin sentences within a schema, we have observed that it was never the case that the two sentences were both resolved correctly. We speculate that this happens because the simple form of queries that we have used in the context of this work effectively missed the small differences between twin sentences, giving rise to the same query for both sentences. This directly points to an opportunity to further improve the performance of the system through the creation of more nuanced queries. If this improvement ends up yielding a worse metric of hardness, this might be an indication that humans might also, to some extent, ignore parts of a WSC sentence that might be critical in its correct resolution.

Another observation worth reporting is that the mean accuracy of teenagers when tested on the first sentence in the schemas versus their mean accuracy when tested on the second sentence in the same schemas has a gap of 20% ($\sigma = 0.15$), suggesting that most of the twin sentence pairs do not include sentences of the same hardness.

7 Conclusion and Future Work

We have shown in this work that a particular existing system that was developed for the WSC can form the basis for deriving a data-driven metric of hardness for WSC sentences. Evidence that the system's computed hardness index is correlated with the perceived human hardness was offered through two studies, one from the literature and one designed as part of this work.

We envision that our developed approach can be used by researchers or challenge organizers, who wish to group sentences in terms of their human hardness. As an example application, the designers of CAPTCHAs could utilize the WSC as a test to distinguish humans from machines (as pursued, for example, in [7]), and could use our system to ensure that the generated tests are not overly demanding for human users.

References

[1] Evan Ackerman. Winograd Schema Challenge Results: AI Common Sense Still a Problem, for Now. *Spectrum*, 2016.

[2] Dan Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. The Winograd Schema Challenge and Reasoning about Correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*, 2015.

[3] David Bender. Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*, pages 39–45, 2015.

[4] Eric Bengtson and Dan Roth. Understanding the Value of Features for Coreference Resolution. In *EMNLP*, 10 2008.

[5] Tejas Ulhas Budukh. An intelligent co-reference resolver for Winograd schema sentences containing resolved semantic entities, 2013.

[6] Nicos Isaak and Loizos Michael. Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In David Pearce and Helena Sofia Pinto, editors, *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press, 2016.

[7] Nicos Isaak and Loizos Michael. Using the Winograd Schema Challenge as a CAPTCHA. In *Proceedings of the 4th Global Conference on Artificial Intelligence (GCAI 2018)*. EasyChair, 2018.

[8] Hector J. Levesque. The Winograd Schema Challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, number SS-11-06. American Association for Artificial Intelligence, 2011.

[9] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[10] Loizos Michael. Reading Between the Lines. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1525–1530, July 2009.

[11] Loizos Michael. Partial observability and learnability. *Artif. Intell.*, 174(11):639–669, 2010.

[12] Loizos Michael. Machines with Websense. In *Proc. of 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 13)*, 2013.

[13] Loizos Michael and Leslie G. Valiant. A First Experimental Demonstration of Massive Knowledge Infusion. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 378–388. AAAI Press, September 2008.

[14] Haoruo Peng, Daniel Khashabi, and Dan Roth. Solving Hard Coreference Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, 2015.

[15] Altaf Rahman and Vincent Ng. 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*, EMNLP-CoNLL '12, pages 777–789, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[16] Adam Richard-Bollans, L Gomez Alvarez, and Anthony G Cohn. The Role of Pragmatics in Solving the Winograd Schema Challenge. In *Proceedings of 13th International Symposium on Commonsense Reasoning (Commonsense-2017)*. CEUR Workshop Proceedings, 2017.

[17] Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. Towards Addressing the Winograd Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module. In

Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, pages 25–31, 2015.

[18] Leslie G. Valiant. Knowledge Infusion. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1546–1551. AAAI Press, 2006.