School of Computing

FACULTY OF ENGINEERING

UNIVERSITY OF LEEDS

Analysis of Existing Approaches to the Winograd Schema Challenge

Ruiwei Hu

Submitted in accordance with the requirements for the degree of
MSc Advanced Computer Science

2018/2019

The candidate confirms that the following have been submitted*:*

Items	Format	Recipient(s) and Date
Deliverable 1	Report	SSO (04/09/2019)

Type of Project: Theoretical Study

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student) ___________________________

Summary

With the increasing urge for development of the machine which has human-like thinking, the ability of the machine to understand human language become significant. The Winograd Schema Challenge, as an alternative to the Turing test, is a way to test this ability. This project investigates the different methods and tools that contribute to the resolution of anaphora, and the possibility of the machine automatically extracting commonsense knowledge to reason with natural language. Meanwhile, classify the Winograd Schema instances from these three aspects, namely connection word, the complexity of grammar and semantic structure, and discuss the performance of different types of the Winograd Schema instance under different tools and methods.

Acknowledgements

I would sincerely like to thank my project supervisor, Dr. Brandon Bennett, for his professional and patient weekly guidance. As well as I would like to thank my assessor, Prof. Netta Cohen, for her suggestions about clarifying the project aims.

Table of Contents

1 Introduction

1.1 Winograd schema challenge

The Winograd Schema Challenge (WSC) is a test for artificial intelligence by setting 150 sentences that may lead to ambiguity which are well understood by humans, but tricky to the computer software due to the answer of the question in schema needs basic knowledge about the world and ability of precise semantic parsing. According to Levesque [1], each instance in Winograd Schema (WS) has a pair words which lead to opposite understanding to the sentence, different word in the sentence can lead to the different answer, the alternative word in the sentence make the question still make sense. Splitting the alternative word in the sentence so that the collection of sentences is easier to be used, the total number of examples used of schemas is 285. These can be found on the webpage (https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.xml). All examples of WS schema that are mentioned in this report are taken from that page.

There is an example of a pair sentence from schema shows below,

Paul tried to call George on the phone, but he wasn't successful

Question: Who was not successful?

Answer: Paul

Paul tried to call George on the phone, but he wasn't available

Question: Who was not available?

Answers: George

From the example, the special word ("successful") and the alternative word ("available") in the same position in the sentence will lead to a different answer to the question. It can be observed that the pronoun contained clause "he was not successful or available" in the sentence which means the different adjectives followed by 'he' determines who 'he' is. If the condition is "he wasn't successful" with the precondition "Paul tried to call George on the phone". In this circumstance, the only person who can be described as successful is Paul because Paul was trying to do something, only a person wants to do something, who could be described by the word "successful". Meanwhile, it should be known that the result of calling someone. There are just two situations, 'the person called is available' or 'the person is not available'. The only person can be described as "available" or not in this circumstance is the one who is called.

Therefore, it is necessary to have basic knowledge about the result of calling someone to solve the example question.

Therefore, the Winograd schema should be designed to satisfy the following restrictions [2]:

- Humans should be able to easily eliminate the ambiguity of these issues. The aim of the system is as smart as a human, not more than human.
- They should not follow from simple grammatical and selectional restrictions concerning the objects referred to.
- They should be search-engine proof as much as possible. The Winograd pattern should be constructed so that it is not possible to use the statistical properties of a corpus (e.g. text available via the internet) to solve these problems.

The Winograd schema studying is very meaningful to the progression of artificial intelligence. But so far there is no very appropriate way to solve this schema. It is difficult to do well in the WSC. The highest reported accuracy of any approach, up to now, is 72.7% reported by Vid Kocijan et al in the paper ‘A Surprisingly Robust Trick for Winograd Schema Challenge’ in 2019. Since 50% accuracy can be obtained from a completely random choice, 72.7% is well below what would be expected from a full solution to the problem [3].

Therefore, it is clear that the reason why the WSC is difficult to be solved by computer software and can be considered as an alternative test of Turing test is that the WSC needs the understanding of natural language and ability to use commonsense knowledge.

1.2 Aim and objectives

The aim of this project is to study, analyse and compare methods and tools that can be applied to solving the Winograd Schema Challenge problem.

There are the objectives carried out in the project:

- To carry out an extensive literature review on methods and tools for solving Winograd schema from a variety of different sources, and to understanding how they work and their limitations.
- To identify the main approaches that have been taken to solving Winograd schema.
- To classify the type of Winograd schemas from the aspect of grammar and semantic and identify the certain method for the different type of schemas.
- To evaluate the tools and methods identified in the research described above, by comparing and contrasting how they perform on the schema set and relating this to the classification of schema types.

In this dissertation, carry out an extensive literature review on methods for solving Winograd schema from a variety of different sources, and to understand how they work and their limitations and identify the main methods and tools that have been taken to solving Winograd schema. Then classify a certain type of Winograd schema and compare the solutions introduced by researchers studying before. And evaluate the tools and methods identified in the research described above, by comparing and contrasting how they perform on the schema set and relating this to the classification of schema types.

2 Background

This section contains the explanation of the reason that the WSC can be considered as an updated AI test approach, the weakness of Turing test and a detailed description of basic knowledge related to resolving the WSC, namely, commonsense knowledge, pronoun disambiguation problem and natural language processing.

2.1 Turing test

The imitation game well-known as the Turing test proposed by Alan Turing has used to be considered as the best way to test artificial intelligence. This test is designed to let a judge through the screen talk to a group of people and a chat robot which be created to trick judge into thinking it is a person. If the judge cannot identify the chat robot this means that chat robot has artificial intelligence. In 2014, there is a Chabot named 'Eugene Goostman' convinced more than 30% of human judges that it is a 13-year old boy who born in Ukrainian. However, this Chabot also exposed the problem that the Turing test cannot avoid which is that the Turing test is too easy to cheat and can be easily tested by deception or pretending to be ignorant rather than the true artificial intelligence. Artificial intelligence needs more than just 'cheating'.

Therefore, the new test that we are after needs satisfy these features:

- It involves a wide range of English sentences in response to the subject;
- Adults whose native language is English can easily pass;
- It can be administered and graded without expert judges [1];
- Performance is better than Tuning test.

To imitate the way of human thinking, we should understand the difference between human thinking and machine "thinking",

Comparing artificial intelligence, we define the following three functions of human beings.

First, Human instinct and unique features. Such as walk upright, natural language understanding, especially non-standard spoken language. These are the characteristics that only humans have and seem to be essential genetic features of humans.

Second, Human instinct but not unique. For instance, identify the sound, identify image but these are function that animals also have, and even some animals are better than people in these two aspects. Dolphins are better than humans in hearing and eagles are better than humans in terms of vision.

Third, Human unique but not instinct. This cover learned abilities, such as: playing the board game Go, driving a car.

Artificial intelligence is relatively mature in the second and third categories. Robotic speech can achieve a high degree of anthropomorphism, machine image recognition is also highly accurate. Even AlphaGo achieved defeated professional Go player [4]. The self-driving car is no longer just exist with a science fiction novel.

This shows that those human instinct and unique features especially common sense and awareness of the world are hard to be imitated by machines. Therefore, the Winograd schema is introduced as a better alternative of the Turing test for testing artificial intelligence.

2.2 Pronoun disambiguation problem

According to Leora Morgenstern et al [2], the WS are mostly focused on the pronoun disambiguation problem (PDP). There are two rounds of WSC, the first round is totally about PDP well-known as PDP-60 that extracted or modified from news, essays, autobiographies, biographies or made by the organizer of the WSC competition [2]. In the second round, organizer adds some other types examples into WS.

For instance, the 24th example of WSC-150 from the Winograd schema challenge:

I poured water from the bottle into the cup until it was [full/empty]. What was [full/empty]?

Whether can figure what is the ‘it’ in the sentence is the key to solve this question. Therefore, the first step is understanding the definition of full and empty and the result of pour water from one container to another container. Then understanding the relationship between these two concepts. Such question cannot be easily solved by only mining a corpus, it is necessary that combines the understanding of pronoun and basic knowledge.

Moreover, the example of Winograd schema above about calling person is also a pronoun disambiguation problem, the key is understanding which person is ‘he’ in different circumstances. Through the above simple explanation, it can be seen that the elimination of pronoun ambiguity and commonsense knowledge are closely related. To identify what the pronoun refers to, it is necessary that having the basic understanding of the world.

The examples shown above all only have one pronoun in the sentence, but there may have more than one pronoun in the sentence, as following example 64th from WSC-150:

Mary took out her flute and played one of her favourite pieces. She has [loved/had] it since she was a child. What has Mary [loved/had] since she was a child?

Answers: The piece/the flute.

There are three pronouns in the sentence which are “she”, “her” and “it”. The “she” and “her” refer to the “Mary”, at the same time, the “it” could represent both “the piece” and “the flute” under different situations. If the verb is “love” in the second sentence of example, the “it” will refer to “the piece” because the commonsense knowledge can be extracted are “the piece cannot be had by Mary.”, “using flute to play piece which is one of Mary’s favourites.” and “ the flute belongs to Mary due to that it is her flute”. In other words, the “it” will represent “the flute” if the verb is “had”. Consequently, there is only a limited number of examples which involve multiple pronouns in the WSC because of the complex ambiguities of multiple pronoun sentence.

Thus, the reason why uses a great number of PDPs in the WSC is to take advantage of characteristics of the PDP that are:

First, the pronoun disambiguation example can be easily found in real-world human language text, we can directly take those pronoun disambiguation examples from existed resource which absolutely have the correct answer and can be understood by people. Creating the original Winograd schema example is difficult which require time, money and manpower, also we cannot make sure that those original examples can work in AI testing.

Second, the pronoun disambiguation example from a large number of text materials could cover a lot of different areas of commonsense knowledge.

Last but not least, the characteristic of pronoun disambiguation example can meet the original objective of the WSC which is the computer software uses commonsense knowledge instead of analysis grammar or semantic structure of sentence to solve the problem.

2.3 Commonsense knowledge

Commonsense knowledge refers to the knowledge that recognized by everyone and does not require explanation. Just like the example above, the basic knowledge about the result of calling people and trying to do something can be seen as commonsense knowledge. The commonsense knowledge in artificial intelligence area is growing slower than other areas. Because it is difficult that let the machine can understand some stupid questions in the human eye without thinking, such as, ‘can you see what is happening in front if the person in front of you is taller than you?’ or ‘who is taller, the father or his one-year-old son?’. Those questions are easy to be answered by human. Because we know that if the person who stand in front of you is taller than you, the sight in front of you will be blocked by that person. Also, a one-year-old baby cannot taller than an adult man. It is necessary that software can use commonsense knowledge to answer those question in the WSC.

According to Peter [6], there are main 20 categories of commonsense knowledge. Those categories can be divided into three types, which are namely, form-based categories, content-based categories and miscellaneous categories. In the form-based categories, the knowledge can be explained with a representation language, such as logical formula, semantic net [6].

There are some following form-based categories which are useful to solve the Winograd schema.

1) Cause and effect: the action or event happen after a certain action or event happen. This type of commonsense knowledge is a large proportion of all example of WS. The sentence include word “because” can be easily considered as a causal event. For example,

The city councilmen refused the demonstrators a permit because they feared violence.

Question: Who feared violence?

Answers: The city councilmen

This example shows commonsense knowledge about cause (“fear violence”) and effect (“refuse a permit”) involved in the sentence. The action in the example is “the city councilmen refuse the demonstrators a permit”, the predicate required is “the city councilmen fear violence”.

2) Precondition: A action or event will happen at a certain time, requiring a precondition action or event happened before that time. For example, in order to be a naturalised citizen of a place, the precondition is that person cannot be born that place. Involved instance of WS is,

This book introduced Shakespeare to Ovid; it was a major influence on his writing.

Question: Whose writing was influenced?

Answer: Shakespeare

To answer this instance of WS, for the fact “a major influence on his writing” the precondition we should know is that Ovid was born earlier than Shakespeare. Therefore, when a person is influenced by another person, it means that the person who influences others is older than the person influenced.

3) Simultaneous Conditions: two actions or events should occur at the same time, like action force and reaction force, those two actions or events cannot exist independently, one action or event occurred must lead to the existence of the second action or event. A real-world example of this type of commonsense knowledge is that when a couple gets married, they become each other’s spouses. The fact of getting married simultaneously occur with the fact of becoming spouse. The example from WS shows below,

I poured water from the bottle into the cup until it was empty.

Question: What was empty?

Answers: the bottle.

The action of “pour water from the bottle into the cup” is consistent with the fact that the water in the bottle is reduced. According to this commonsense knowledge, it is obvious that the bottle will be empty after pour water from bottle into the cup.

4) Prominent Relationship: this type of commonsense refer to that there is special relationship between two entities, such as, career (like the responsibility of a career), skills (for example, the painter A relates with a painting B, it is most likely that the author of the painting B is that painter A.), intrinsic ability (for example, the women can be pregnant, if a person C is pregnant, the most chance that C is female.). In the WS, the example is,

The city councilmen refused the demonstrators a permit because they advocated violence.

Question: Who advocated violence?

Answers: the demonstrators.

From the instance, the city councilman as a member of the council for a city has the responsibility to protect city. In the sentence, it is clear that violence is bad factor to a city

5) Transitivity: there is a relationship R between entity A and entity B, also the same relationship R between entity B and entity C and this relationship is transitive. Due to this type of commonsense knowledge, we can infer that there exists relationship R between A and B. There is a sentence contained a pair word in the WS which can use transitivity of commonsense knowledge.

This book introduced Shakespeare to [Ovid/Goethe]; it was a major influence on his writing.

Question: Whose writing was influenced?

Answer: Shakespeare/Goethe

In this example, relationship between Shakespeare and Ovid (R (Ovid, Shakespeare)) is that Ovid is older than Shakespeare. Similarly, R (Shakespeare, Goethe) is true, then R (Ovid, Goethe) is true.

6) Definition: it is an explanation for a word or phrase. It can be used to explain the function of an object or a thing. Such as, a “box” is a container which holds some things. There is an instance in the WS,

The father carried the sleeping boy in his bassinet.

Question: Whose bassinet?

Answer: the boy

The definition of bassinet is a specific bed for babies who only can lie or child who is lying. In other words, the bassinet is not for the adult. The answer is clear after understand the definition of function of bassinet.

Thus, making machine recognize those various categories of commonsense knowledge involved in the WSC is a big challenge to researcher. In addition, from the aspect of the commonsense reasoning, there are many kinds of relations in the commonsense knowledge, that we do not know how to represent them in a computer usable form, and let machine correctly reason them. Moreover, those commonsense reasoning is based on human understanding, sometimes the conclusion is reasonable but may not correct.

2.4 Natural language processing

Natural language processing refers to the computer software can process and understand natural language using human-like thinking. To make computer understand human language, especially, have a deeper understanding of oral expression and certain understanding involved with special culture, human habits. Usually, using machine learning algorithms train millions of data sample, such as, words, sentence, and passages, to gain a prediction of understanding of human language.

Natural language processing is a very wide concept, it is not only involved in the Winograd schema problem but also other computational linguistics problem. It is a subclass of all human language problem.

The main techniques used to complete task of natural language processing are syntax analysis and semantic analysis.

Syntax refers to understanding the grammar of the text, in the field of solving Winograd schema, understanding the grammar of the text, usually using gramma parsing generates the more structured form of a sentence and analyse the affiliation of words in a sentence.

There are some following techniques in syntax analysis,

- **Lemmatization**. This helps machine group different forms of one word and can analyse those difference form as a single item.
- **Part-of-speech tagging**. This identifies the part of speech of the word.
- **Parsing**. This involves the grammatical analysis of a sentence.

Semantic refers to understanding the meaning of the text. It normally combines with natural language understanding. Applying the computer algorithms to analyse the sentence, understand the meaning of words.

There are some sematic techniques which can be used in the WSC,

- **Named entity recognition**. This can be used in grouping the words due to the different categories. Such as, person, location, organisation.
- **Word sense disambiguation**. This involves giving the meaning for a word based on the context.

However, in the current situation, the fully understanding of human language meaning is quite an arduous process, comparing with the grammatical understanding, there are some difficulties in understanding the meaning of natural language.

Initially, as the human, our language processing center is too complex to simulate and figure out how it works. We have trained our language computer center from reading and communication in real life since an early age.

Secondly, human language is constant evolving over the time, especially in the current Internet age. As the increasing use of internet language and emoji, it is more difficult to understand the text written by human. Even human who not familiar with those usage of internet language cannot know what it means. For example, 2 is not only refer to the number, but also can equal with the word “too” or “to” due to the same pronounce. Similarly, the number 4 can be used as the word “for”, the machine definitely cannot understand that the meaning of “4ever” is “forever”. Besides, the emoji takes a big part of the communication text in real life, the semantic analysis of emoji is also a big challenge in language processing by machine. Sometimes, the emoji can completely change the mood of the text expression to lead a contrary understanding of the text.

Then, the same word in daily usage normally not only expresses one meaning. There is an infinite way of words combination, and different combination can express different meaning even through the words are same. In many cases, it is depending on the context in which is used.

Last but not the least, the meaning of the same word in different English-speaking countries can be different. For instance, the word “can” as a noun in the United State can express a closed metal container, but in the United Kingdom, that container usually is called “tin”. In addition, “the first floor” in the American English means the level of a building that is at the same level as the street, but in the United Kingdom it means the level of the building that above the street. “the ground floor” in the United Kingdom means the level as same as the street. Those cultural conditions add another layer of meaning that needs to be deciphered by the machine.

Thus, due to those difficulties, the machine natural language processing has to deal with multiple information layers. To realize an independent artificial intelligence of human language, the computer needs to independently acquire and learn from the enormous data of the internet to make sense of those information.

3 Previous approaches

Since the WS released as an alternative to the Turing Test by Hector Levesque, there is a large number of researchers show their interesting to resolve this problem. Among those previous efforts devoted to solution of the WSC, normally, it is wildly recognized that using unsupervised learning which can be used to generate simple commonsense relationships in the text [6] and language models (LMs) which is a state of the art technology to estimate the probability distribution of various linguistic units, such as, a word, a phrase or even a sentence [7], in the WSC.

To put this simple, similar with word vector which can be used to answer question that can be predicted via adjacent words in a sentence [6], language model can reach a better result due to the complexity of the Winograd schema questions. The LM can be used to capture real-world knowledge from the text by trained on enormous unlabeled data. for example, BERT (Bidirectional Encoder Representations from Transformers) introduced by Devlin in 2018 is a pre-training language presentation model [8]. The BERT is used by Yu-Ping Ruan et al to build an unsupervised pertaining method of addressing WSC [9] and Vid Kocijan et al to build a fine-tuned robust model [3].

Moreover, in 2012, Altaf Rahman and Vincent Ng [10] create an annotated training dataset of 941 sentence pairs via employ 30 underground students to compose constraints for each sentence pair. Using around 70 thousand linguistic features derived from various components, namely, Narrative Chains, Google, FrameNet, Heuristic Polarity, Machine-Learned Polarity, Connective-Based Relation, Semantic Compatibility and Lexical Features combine with machine learning to build a ranking-based model so that the correct candidate in the sentence can gain a higher rank. The goal of their research is examination of resolution of pointing out the meaning of pronouns in the question and there is no clear evidence can be seen between the answers from aspect of syntax. They evaluate the result of their system through compare with a baseline combination of the Stanford resolver and the Baseline Ranker.

However, this model uses the dataset they build instead of the Winograd schema answer-question pairs. Through analyses the advantage and disadvantage of the two most uncontributed components in the model except Lexical Features, which are Narrative Chains and Google search.

To compare the probability of the result of Google search about two candidates in the sentence, the higher probability of candidate normally is the correct candidate for the sentence. The google is strong in extracting the single simple fact, so it is more useful to resolve the sentence such as "*the city councilman refused the demonstrators a permit* *because they advocated violence",* because the fact "*the demonstrators advocated violence",* obtain higher rate comparing to the fact *"the city councilman advocated violence".* However, if I change the sentence into "*the city councilman refused the demonstrators a permit because they are unreasonable",* the google search results will fail due to that the analysis of probability of the facts "demonstrators are unreasonable" and "the city councilman are unreasonable" cannot contribute to get correct answer. Thus, the weakness of Google is generating the relationship between facts, Google cannot understand the basic knowledge of the world in the sentence via the link between the two parts of the sentence.

Another component involved in the Rahman et al model is Narrative Chain which is a representation of structured knowledge presented by Chambers and Jurafsky [11]. Narrative Chain consists of an ordered set of events (simply, verbs) about a same actor [e_1, e_2, ..., e_n] where n is the length of the Chain [11]. For example, eat-s, wash-s, cut -s, cook -s where "s" refers to "subject" role, this chain shows that the person who eats something, may wash, cut, cook it through this order. The Narrative Chain is good at extracting knowledge for one entity or event and actions related to that entity or event. The issue of the Narrative Chain is to represent the relationship between events.

Furthermore, by 2014, Peter Schuller introduce a method to using the knowledge graph and formalizing Relevance Theory [12]. Relevance Theory is a theory that attempt to explain the recognised fact because the literal word expression may lose some information compare with the utterance. This theory was first introduced by Dan Sperber and Deirdre Wilson [13], the Cognitive Principle of Relevance (human cognitive tends to adapt to the maximization of relevance) as one of basic principle of relevance theory states that the predicable information only can be obtained by combined with the certain cognitive mechanism, such as, memory, reasoning, perception, classification and so on [14]. Because understanding natural human language needs to take correct predication under a certain context, relevance theory in this approach suggests to playing a translation role that audience can make a correct assumption according to the information given by the speaker [12]. Besides, they convert the text into knowledge graph which is a more structured form. Then reason the answer via manually extract background knowledge.

Similarly, Arpit Sharma et al. [15] [16] [17] propose a method can automatically extract commonsense knowledge rather than manually encode relevance information introduced by Schuller. The main notion in Sharma's approach is to translating the text question by semantic parsing into a more formal structured representation, generating a string query and extract background knowledge using the query through the automated Google search. According to the result from Google web search, extract sentence and split the sentence. They focus on two specific type of commonsense knowledge which are direct causal events and causal attributive, there are total 71 sentence related to these two categories and their system can be able to resolve 53 out of 71 Winograd schemas sentences. In their system, they attempt to imitate the human behavior that the human can able to answer the Winograd schema question using the basic real-world knowledge learnt by years reading. Using the same reasoning rules as that Peter Schuller uses in his system, which is Answer Set Programming (ASP) [18] , for the purpose of the following characteristics of ASP: (i) simple syntax; (ii) strong theoretical foundation; the rule is defined as "if a set of atoms $a_1…a_n$ are true and $b_{1,}…,b_n$ can be assumed to be false then a is true" [19].

Daniel Bailey et al. [20] propose a framework about correlation that reason the Winograd schema using annotated knowledge bases as axioms. Because a massive number of Winograd schemas is a pronoun disambiguation problem, they substitute pronoun in the question with a special word and an alternative word in the sentence. For example, "The city councilmen refused the demonstrators a permit because they [feared/advocated] violence. Who [feared/advocated] violence?", after replace the pronoun in the sentence, the sentence pair can be divided into four sentence without pronoun which are "The city councilmen refused the demonstrators a permit because the city councilmen feared violence.", "The city councilmen refused the demonstrators a permit because the demonstrators feared violence.", "The city councilmen refused the demonstrators a permit because the city councilman advocated violence." and "The city councilmen refused the demonstrators a permit because the demonstrators advocated violence.". The correlation in the framework can be understood as the coherent of the sentence, as they mentioned in the paper, if the candidate of the pronoun can make the discourse coherent then that candidate is correct answer for the pronoun problem. As the above example, the clause of schema "The city councilmen refused the demonstrators a permit" has a positive correlation with the clauses "the city councilmen feared violence" and "the demonstrators advocated violence" which means these two facts are more reasonable than other two facts. According to Bailey, the positive correlation of the sentence is normally existing in sentence which contain the connective word "because", on other words, the negative correlation is related to the connective word "although".

They manually apply the framework on the first 100 WS examples to build an annotated corpus, 72 out of 100 WS examples show the positive or negative correlation, where 64 examples have positive correlation and 8 examples have negative correlation.

Currently, Ali Emami et al. [21] [22] introduce a method that using an automatic knowledge hunting model to extract knowledge from the internet by generating query search online, like Sharma's method, the query in the Emami's system is generated based on a semantic parsing and query filter. Moreover, to compare the performance of automatic query generation, they add a manual query construction for all Winograd schema pairs. They use the example "The trophy doesn’t fit into the brown suitcase because it is too large” to indicate the difference between various query generation method. The result of this system can achieve 58 percent accuracy on WS corpus.

Compare to the previous works on the WS solution using manual annotated features. Juri Opitz and Anette Frank [23] resolve the WS problems from a new view as a sequence ranking task and because their approach is an end to end neural approach, they address a problem that the WS solutions rely on Google search heavily, this problem is unavoidable for most previous method mentioned above. This approach can gain 63 percent accuracy on WSC.

The best result of addressing Winograd schema so far is the robust fine-tuned method introduced by Vid Kocijan et al. and achieve over all 72.2 percent accuracy on WSC-273 dataset, improving 8.5 percent on previous method. Besides, they claim that the difficult of solving Winograd schemas is not only it is hard to achieve human-like commonsense knowledge and reasoning, but also that the dataset for training neural networks is too small to acquire a more previous result.

In this section, I analyze some approaches proposed in these areas which introduced above. I pay attention on not only the performance on the WSC but also the contribution and achievement in the area of commonsense reasoning.

3.1 Methods analysis

3.1.1 Statistical methods

The statistical methods, generally speaking, is usually gathering with other approaches. It not only can use in the hand-crafted rule system but also in training of the machine learning algorithms. The statistical models using in the machine learning help to acquire a probabilistic decision based on attaching real-valued weights to each input feature. Whether the hand-crafted rule or machine learning, both using the selectional constrains which are usually specific for the certain example to filter and get the appropriate candidate.

The simplest statistical method in the analysis of Winograd schema is to manually input the text in the search engine and the text using candidates substitute the pronoun, then compare the probability of the returned result.

From the aspect of pure knowledge-free statistical method, the process generally can be divided as two parts, the first is identifying the syntactic structure of the sentence, finding the possible candidates which are coreference with the pronoun and substituting the pronoun with the candidates. Then statistic the probability of each candidates as the statistical data which can be reused.

However, the statistical method usually focuses on the aspect of syntactic structure. To realize the understanding of sentence, it is not sufficient that just match the coreference of words or syntactic analysis. Besides, statistical methods depend on the corpus, which means the different corpus will influent the result of statistic.

3.1.2 Logic-based methods

The logic-based method is using logical formula to explain the text which contains real-world commonsense knowledge. Those logic rules can integrate world knowledge and formulate knowledge to the certain extent in the training of machine reasoning.

Establish the translation from human natural language into formal language, such as, first-order logic or higher order logic to build a knowledge base for the machine.

Compare with the statistical method, the logic formula is more explicitly and can give more understanding. The problem of logic-based methods, the source of the world knowledge and how to build the knowledge base in an effective way.

3.1.2.1 Analysis on the WSC set

There are two way of logical formula to explain the sentence and complete the definition of pronoun. I analyses and compares them based on the “trophy” case introduced above in the WS.

First, the correlation form that introduced by Daniel Bailey [20] deal the “trophy” case that “the trophy does not fit into the suitcase because it is too small” by using the correlation calculus to justify the fact one object cannot fit into another object is related to which object is small. The general theory of correlation form is understanding coherence. The clause with the replacement for the pronoun is more coherent than another alternative which means the coherent one is the correct answer for the problem. Back to the “trophy” example, using the correlation formula “ $\neg fit\_into(T,S) \oplus small(x)$” where T is trophy and S is suitcase to represent the relation of commonsense fact. Combination with the axioms, where “any object can be small”, “small object fit into a larger object” and “suitcase is a physical object” which represented in the formula are

$\forall x(object(x) \rightarrow small(x))$

$\forall x(small(x) \leftrightarrow \neg large(x))$

$\forall x(suitcase(x) \rightarrow object(x))$

$fit\_into(x,y) \oplus large(y)$

to derive the answer for the "trophy" case. The axiom "$fit\_into(x, y) \oplus large(y)$" is kind of relation of commonsense facts what the machine needs to know at the first place.

However, the correlation form only considers the simplest situation under the precondition "the trophy doesn't fit into the suitcase". According to Ernest Davis [24] , there are other two more complex explanations which is (a) to someone needs to pack other things in the suitcase, "doesn't fit into" may refer to that comparing with other items that needs to be packaged, the trophy has a lower priority; (b) to a person who has already packed some items in the suitcase, the trophy doesn't fit into suitcase may means to pack the trophy, it is necessary that rearranging the items already packed make sure there has a space for the trophy.

Even just from the simplest explanation, Ernest Davis covers more possible situations and symbolize into the formula. Instead of using correlation relationship, Davis considers the problem with more detail facts, involving the comparison of object size, representing with formula smaller (a, b) and larger (a, b).

The reasonable explanation of special relation "cannot fit in" can be discussed as (a) the trophy cannot fit into the suitcase, and all items which larger than the trophy cannot fit into the suitcase, and a existed item which smaller than the trophy can fit into the suitcase; (b) from aspect of the suitcase, anything smaller than the suitcase that cannot hold the trophy, and a existed item that larger than the suitcase which can hold the trophy.

Represented into

$$\neg fit\_into(T, S) \rightarrow ((\forall x\ larger(x, T) \rightarrow \neg fit\_into(x, S)) \wedge (\exists y\ larger(T, y) \rightarrow fit\_into(y, S)))$$

$$\neg fit\_into(T, S) \rightarrow ((\forall x\ smaller(x, S) \rightarrow \neg fit\_into(T, x)) \wedge (\exists y\ smaller(S, y) \\ \rightarrow fit\_into(T, y)))$$

Therefore, the general relation in the "trophy" case can interpret as "$\alpha\ cannot\ \beta\ becasue\ it\ is\ too\ \gamma$"

$$\neg \beta(\alpha) \rightarrow ((\forall x\ \gamma(x, \alpha) \rightarrow \neg \beta(x)) \wedge (\exists y\ \gamma(\alpha, y) \rightarrow \beta(y)))$$

Where α is the entity, β is the event or precondition, and γ is the comparison relationship, in the example, it can be smaller or larger.

Although the logic formula can present relation in the text and trained by machine, specific rule is lack of generality. Every relation needs to be manually generated which is time-waste and not smart. Moreover, the correlation formula requires that substitute the pronoun with the candidates and discuss the coherent of clauses after substitution. In other words, the two clauses of the example should have relation whether a negative correlation or positive correlation. The fact is that normally the example includes a subordinating conjunction words, that example can be considered as involving correlation. As in the example with "trophy", there is conjunction word "because". And the positive correlation is often relevant with the connective word “because”, on the other hand, the negative correlation usually related to the “although”. But there is not every example in the WS which has a correlation. For example, “I put the cake away in the refrigerator. It has a lot of butter in it”, after replace the pronoun with the answer “cake”, the sentence is “I put the cake away in the refrigerator. The cake has a lot of butter in it”. But there is no clear correlation between the phrases

I put the cake away in the refrigerator

and

the cake has a lot of butter in it.

The action “put cake in the fridge” cannot indicate any relation to the fact “the cake has butter”. Substituting the pronoun with another candidate “refrigerator” instead of the answer “cake”, the phrase is “the refrigerator has a lot of butter in it” which is also plausible to the reader. The probability of correlation of the phrase cannot directly reveal the answer.

3.1.3 Machine learning methods

The machine learning method is used in resolving anaphora problem for the past two decades [25]. Since the 1990s, the methods of correlation resolution have started to use machine learning approaches instead of heuristic approaches. Moreover, the neural network shows its strength on addressing the natural language task, such as image recognition, speech processing and machine translation [26] [27] [28].

As the report that achieve the best performance in competition of 2016 [29], Liu, Jiang and Ling introduce a neural network method combined with context and commonsense knowledge. This framework can supervised learn the commonsense knowledge from a large corpus. There are two main method involved in the framework, first is using unsupervised way to extract the semantic similarity between the pronoun and all candidates mentioned in the same sentence. Second method is neural knowledge activated method which is a supervised method training a deep neural network to judge the mention pair (candidate/pronoun) whether coreference or not.

In 2019, Ruan, Zhu and Ling [9] propose an updated method which can achieve 71.1% accuracy on the WSC set. Their framework mould the dependency structure of sentences and extract the knowledge from unsupervised pretraining models.

Wolff [30] introduced the SP system which is designed to simulate the human brain function. The senses of the system can receive new information (like human’s eyes and ears) and store the information as the old information (like human’s brain). This SP system as the integrated product of the concepts of artificial intelligence, mainstream computing, mathematics, and human learning, perception, and cognition [31] can contribute in the identification the pattern of association between linguistic features. As an unsupervised learning, the SP system outline the potential ability of machine that can automatically learn the knowledge and human language without the guidance of a “teacher”.

However, as far as the current situation of machine learning about the WSC is concerned, almost all machine learning methods are only train the machine to determine the sentence structure and find the probability rather truly understand the meaning like a human.

3.2 Tools analysis

3.2.1 Google

According to Satoh and Yamana, hit counts which is the number of the search results can be used in the crude statistics analysis in the natural language processing problem, ontology construction, and analysis of social networks [32].

However, the result of google hit is less reliable to prove the relationship between entities. According to the explanation of Satoh and Yamana, index updated over time can result in inconsistencies between multiple indexes and inconsistencies between different search units. Furthermore, Davis indicates that there is a huge difference between the number of google hits and the actual number of the returned result from google search page [24]. The inconsistent in result pages returned by google itself may mislead to correctly acquire the relation between words or phrases.

Moreover, the study by Ahmet Uyar [33] shows the number of words in the query can significantly affect the accuracy of the estimate. When testing from a single word to two words in the search engine, the percentage of accurate hit count estimates is almost halved. As the number of query words increases, the estimation error increases and the number of accurate estimates decreases.

Overall, the massive number of web documents collected by the search engine is a useful dataset for the researchers for study purpose, especially, language research. However, due to the limitations of search engine, the lower accuracy of google hits, without combining other technologies, only using Google hits cannot achieve good performance on finding the relation between words or phrases.

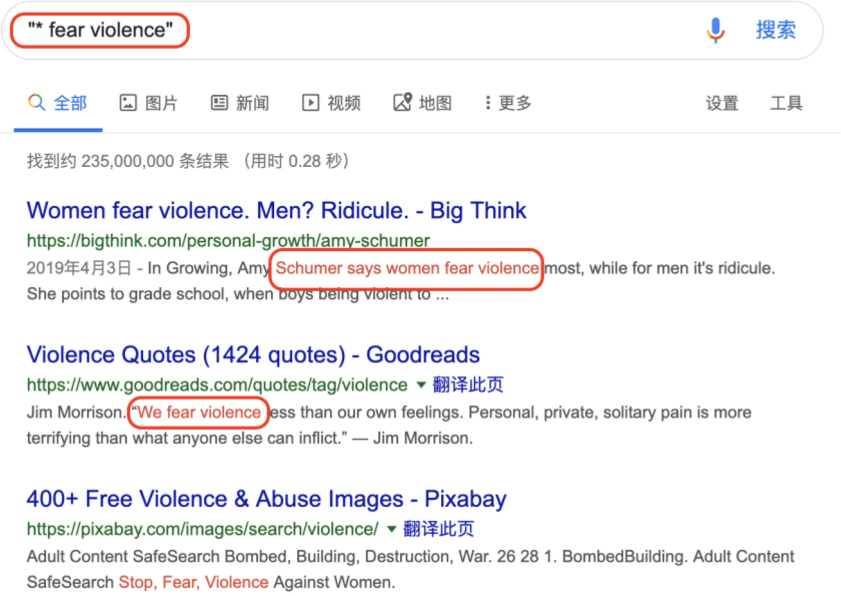
3.2.1.1 Analysis on the WSC set

I will examine several examples of the WS sentences. Because the result of query string without double quotation marks just shows the number of documents on the web which contain the words in the query string rather than that return the document which can show the relation between words. Thus, to research the WS problem, putting the phrase in the double quotes, google search will consider the documents which contain the set of searched words in a specific order. In addition, the * is powerful function can be used in the research, using the query within a *, it tells Google to try to treat the star as a placeholder for any unknown terms and then find the best matches. For instance,

EX 1 “The city councilmen refused the demonstrators a permit because **they** feared violence.”

In this example, I substitute the “they” with candidates “the city councilmen” and “the demonstrators”.

Using query [“* fear violence”] can get the amount of document match the exact this ordered string, such as, “woman fear violence”, “we fear violence” and so on. It can be as a base total number for “fear violence”. As shown in figure,



找到约 235,000,000 条结果 （用时 0.28 秒）

Women fear violence. Men? Ridicule. - Big Think
https://bigthink.com/personal-growth/amy-schumer
2019年4月3日 - In Growing, Amy Schumer says women fear violence most, while for men it's ridicule. She points to grade school, when boys being violent to ...

Violence Quotes (1424 quotes) - Goodreads
https://www.goodreads.com/quotes/tag/violence ▾ 翻译此页
Jim Morrison. “We fear violence less than our own feelings. Personal, private, solitary pain is more terrifying than what anyone else can inflict.” — Jim Morrison.

400+ Free Violence & Abuse Images - Pixabay
https://pixabay.com/images/search/violence/ ▾ 翻译此页
Adult Content SafeSearch Bombed, Building, Destruction, War. 26 28 1. BombedBuilding. Adult Content SafeSearch Stop, Fear, Violence Against Women.

Figure 3.1 Google search result for the query “* fear violence”

Subsequently, searching the query [“the city councilmen fear violence”] and [“the demonstrators fear violence”] and compare the probability of the results of those two queries among the total number of [“* fear violence”]. Sometime, as this example, the quoted query string cannot gain the result, which means the exact string order cannot be found in the web document. Due to this situation, I replace the word in the query with its synonym without changing the meaning of query. Using councilor instead of councilman acquire 6 results. On the other hand, whether the demonstrator or the its substitution words cannot get the result in exact order in the google search. To a certain extent, this can prove the “the councilman fear violence” achieve higher probability than “the demonstrator fear violence”.

Ex 2 “I was trying to open the lock with the key, but someone had filled the keyhole with chewing gum, and I couldn't get **it** in.” and “I was trying to open the lock with the key, but someone had filled the keyhole with chewing gum, and I couldn't get **it** out.”

This is a pair example, the examined result on the google search in 27th of July 2019, as shown in the table1 and table 2,

Query string	Number of results	Probability among total
“couldn’t get * in”	845,000,000	
“couldn’t get key in”	673,000	7.96×10^{-4}
“couldn’t get gum in”	6	7.10×10^{-9}

Table 1 Probability of “I was trying to open the lock with the key, but someone had filled the keyhole with chewing gum, and I couldn't get **it** in.”

Query string	Number of results	Probability among total
“couldn’t get * out”	889,000,000	
“couldn’t get key out”	1,510	7.20×10^{-6}
“couldn’t get gum out”	6,400	1.70×10^{-6}

Table 2 Probability of “I was trying to open the lock with the key, but someone had filled the keyhole with chewing gum, and I couldn't get **it** out.”

Quoted query string fail to gain the results in many examples, because the web documents that stick with the exact words and order in the query is small probability event, the google hits is too small to reflect the answer of the Winograd schema question.

3.2.2 Pronoun resolver

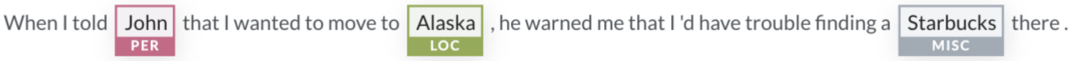
There is various open source software on the web which can automatically complete task of coreference resolution. For example, Stanford CoreNLP [34], AllenNLP, Berkeley coreference resolution, spaCy and so on.

The Stanford is a rule-based system that uses a precisely ordered sieve (filtering rules) to decide whether two references should be linked [35]. It can give the basic form of words, their part of speech, they are company names, characters, and so on, standardized dates, time and numbers, and sentence structures are constructed with phrases and syntactic dependencies, indicating that noun phrases refer to the same entity, expressing emotions, extracting specific or open relationships between entity mentions, getting quotes, and so on.

The Berkeley system is a learning-based, mention-synchronous coreference resolution system [36] that learns to link two references using surface features that capture linguistic properties of mentions and mention pairs [35].

3.2.2.1 AllenNLP

There is online demo called AllenNLP for answering the nature language questions. It has a lot of functions to deal with the nature language question such as named entity recognition, constituency parsing, dependency paring, semantic analysis and more useful coreference resolution. Entity recognition, constituency parsing, dependency parsing and semantic analysis are only for annotating a sentence, coreference resolution can work on a passage. The entity recognition can identify the name of a person, location, organization or miscellaneous in a sentence. An example from the AllenNLP demo shows in the figure 3.1:



When I told John that I wanted to move to Alaska , he warned me that I 'd have trouble finding a Starbucks there .

Figure 3.1 Named entity recognition in the AllenNLP.

The system identifies the person John, location Alaska and miscellaneous Starbucks in the sentence, which is can used in the WS sentence because there may be many words in the sentence that begin with a capital letter. If can differentiate those words with a capital letter are people, organizations or locations can alleviate the misunderstanding between those words.

The constituency parsing is to divide a sentence into sub-phrases, or constituents. Non-terminals in the tree are types of phrases, the terminals are the words in the sentence. The dependency parsing is the grammatical structure of a sentence, establishing relationships between "head" words and words which modify those heads.

This is an example of dependency parsing for a simple sentence in AllenNLP, it can visualize a dependence relationship of each word in a sentence without considering the meaning of words. The root refers to the verb “ate” which is the head word, “James” is a noun subject which has a relation to the verb. The structure shows in figure 3.2, and this box-like structure is less intuitive from the point of demonstrating the dependence relationship compared with Stanford system which will show below.

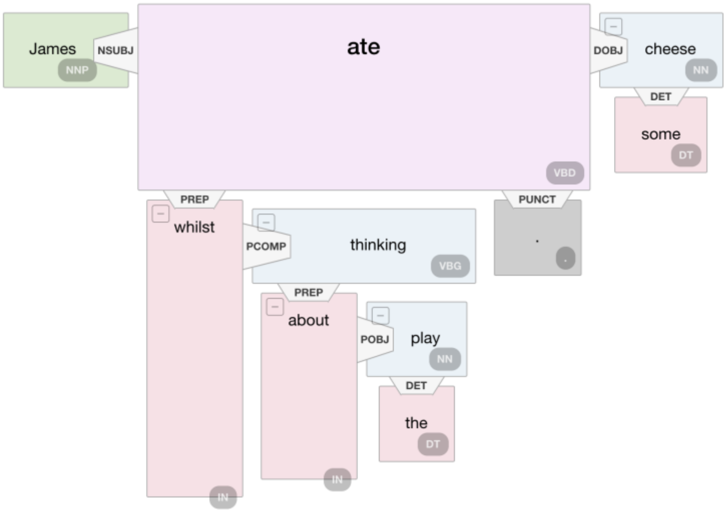
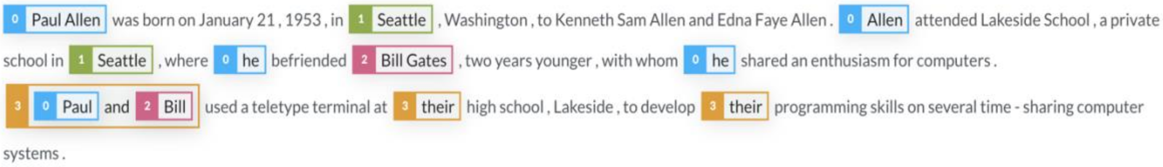


Figure 3.2 Dependency parsing pattern in the AllenNLP.

The coreference resolution is most contributable in the WS problem, because the goal of the coreference resolution is finding all expressions that refer to the same entity in a sentence or passage. It is an end to end neural model which considers all possible spans in the document as potential mentions and learns distributions over possible antecedents for each span, using aggressive, learnt pruning strategies to retain computational efficiency [37]. There is an instance from AllenNLP, as shown in figure 3.3:



0 Paul Allen was born on January 21 , 1953 , in 1 Seattle , Washington , to Kenneth Sam Allen and Edna Faye Allen . 0 Allen attended Lakeside School , a private school in 1 Seattle , where 0 he befriended 2 Bill Gates , two years younger , with whom 0 he shared an enthusiasm for computers . 3 0 Paul and 2 Bill used a teletype terminal at 3 their high school , Lakeside , to develop 3 their programming skills on several time - sharing computer systems .

Figure 3.3 Coreference resolution example in the AllenNLP.

The same colour in the sentence refer to one entity, according to the instance, it is successful to find what is the meaning of the pronoun “he” and “their” in the text. The “he” is Paul Allen, the “their” is Paul and Bill.

By using coreference resolution can have a crude identification of pronouns in the WS sentence which may lead to the answer of the WS questions especially for the pronoun disambiguation problem involved in the WSC.

3.2.2.2 Analysis on the WSC set

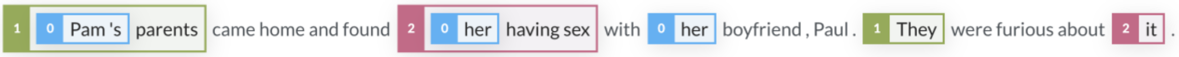
After manually put each sentence from the WSC-285 into the coreference resolution, there are 122 examples out of the 285 returning the correct results that the system successfully identifies related entities in the text, and 7 instances have no decision in the system.

Considering the pair instance instead of the single sentence, for example, “The large ball crashed right through the table because it was made of steel” and “The large ball crashed right through the table because it was made of styrofoam”, those two sentences have the same structure expect the material of pronoun “it” in the text. Only five pairs of sentences correctly point out the meaning of the pronouns in the sentence. The sentence pair which have the similar gramma structure mostly return the same result in the coreference resolution of AllenNLP which means that AllenNLP normally identifies the entity that the pronoun refers to in the sentence according to analysis of the structure of sentence and meaning of a single word rather than understanding the entire sentence.

I find some interesting instances worth to discuss after coreference resolution, there are detail explanation below.

Ex 1 “Pam's parents came home and found her having sex with her boyfriend, Paul. **They** were furious about it.”

The visualisation result is given in figure 3.4:



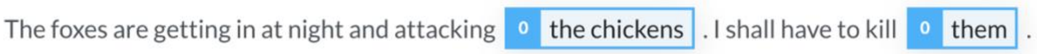
1 0 Pam 's parents came home and found 2 0 her having sex with 0 her boyfriend , Paul . 1 They were furious about 2 it .

Figure 3.4 Named entity recognition in the AllenNLP.

This is an instance that AllenNLP successfully identify all entities in the sentence, “her” in the sentence refers to “Pam”, “they” refers to “parents”, “it” refers to “having sex”. The WS question of this example is “Who are furious”. Given the figure display that “they” have the same colour with “Pam’s parents”, the answer of the WS question is “Pam’s parents” which is the answer we expected.

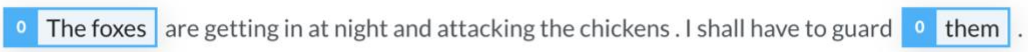
Ex 2 “The foxes are getting in at night and attacking the chickens. I shall have to kill **them**.” and “The foxes are getting in at night and attacking the chickens. I shall have to guard **them**”

The results of the sentence pair as shown in figure 3.5 and 3.6:



The foxes are getting in at night and attacking 0 the chickens . I shall have to kill 0 them .

Figure 3.5 Coreference resolution of “The foxes are getting in at night and attacking the chickens. I shall have to kill them.” In the AllenNLP.



0 The foxes are getting in at night and attacking the chickens . I shall have to guard 0 them .

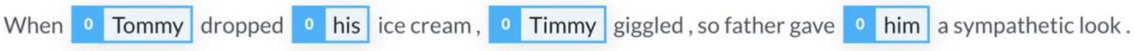
Figure 3.6 Coreference resolution of “The foxes are getting in at night and attacking the chickens. I shall have to guard them.” In the AllenNLP.

This is a sentence pair which have similar structure, the only difference is the penultimate word in second clause. If system identify the pronoun only by the analysis of structure would not get the correct answer. According to most situation, such sentence pair, AllenNLP only can point out one correct answer between those two sentences. However, the AllenNLP can identify different meaning of pronoun due to the situation of different last word in sentence. The interesting point is that although different results are obtained in different situations, the results obtained are the opposite of the correct answers.

As the figure shows above, these sentence pair have the same forward clause, the premise of the problem is that the foxes attack the chickens at night, the difference is the following action should be done by "I" is "kill" or "guard". The question that needs to be answered here is the targets of the implementation of these actions. Under human commonsense reasoning, the foxes attack the chickens, as a human, we should protect the chickens and kill the foxes, because the chicken is a vulnerable group, in this case, the instinct of human is to protect the vulnerable group and fights the group who against vulnerable one. Thus, the answer should be "I kill the foxes" and "I guard the chickens". However, the result in AllenNLP display an opposite situation, the reason for that, in my opinion, the AllenNLP is an end to end neural model which computer all embedding spans in a text as potential mentions that combine context-dependent boundary representations with a head finding attention mechanism [37].

EX 3 "When Tommy dropped his ice cream, Timmy giggled, so father gave **him** a sympathetic look."

The result is given in figure 3.7,



When 0 Tommy dropped 0 his ice cream , 0 Timmy giggled , so father gave 0 him a sympathetic look .

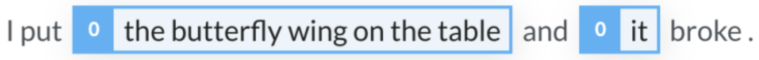
Figure 3.7 coreference resolution of " in the AllenNLP.

There are two human name which are very similar with each other Tommy and Timmy in the sentence. After put the text into system, the returned result shows that the system identify Tommy and Timmy as one entity. The pronoun "his" and "him" both refer to Tommy and Timmy at same time.

Therefore, when the model confuses paraphrase with relevance or similarity, it may tend to predict false positive links, in this case, when system ponder which entities are related, will treat two very similar words as related words.

EX 4 "I put the butterfly wing on the table and **it** broke."

The first result as shown in figure 3.8

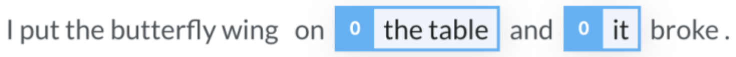


I put 0 the butterfly wing on the table and 0 it broke .

Figure 3.8 Coreference resolution of "I put the butterfly wing on the table and it broke." in the AllenNLP without putting a space in the sentence.

This is a very interesting example in the AllenNLP system. It indicates the space between the words in the sentence can influence the result of coreference resolution. Given the figure above, the system believes that the pronoun "it" refers to "the butterfly wing on the table". This means the system fails to understand the meaning of sentence and fails to separate the butterfly wing and attributive adjunct "on the table". The butterfly wing is a very light and fragile object in human universal thinking, and the table obviously is more solid than the butterfly wing. It is clear to human that put a light and easy to broke object on a sturdy object, the broke thing is the butterfly wing to a greater extent.

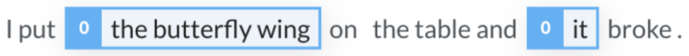
However, as the figure 3.9 shown below, when I put a space in front of word "on" so that using the space separate between "butterfly wing" and "on the table", the system ignores the butterfly wing and identifies the table as what the "it" refers to.



I put the butterfly wing on 0 the table and 0 it broke .

Figure 3.9 Coreference resolution of " I put the butterfly wing on the table and it broke."in the AllenNLP with putting a space between "wing" and "on the table" in the sentence.

In addition, I tried to put the spaces in different places in the sentence. When the space is in front of "the table", the result satisfies the human understanding. As shown in figure 3.10, the system separates the "the butterfly wing on the table" and successfully identify the meaning of pronoun.



I put 0 the butterfly wing on the table and 0 it broke .

Figure 3.10 Coreference resolution of "I put the butterfly wing on the table and it broke." in the AllenNLP with putting a space between "on" and " the table" in the sentence.

3.2.2.3 Stanford CoreNLP

As a more popular and authoritative nature language processing system, Sanford also provides an online API that can visualize the result of resolving coreference problem. I examine all the WSC-285 on the Stanford CoreNLP through the API. The coreference result shows that there are 117 correct instances out of 285, 158 instances return incorrect coreference relationship between entities, and 10 instances have pronoun which is not identified by the system. Given the sentence pairs, there is no correct pairs in the Stanford system, from this point, the performance of the Stanford is not what we expected. Because no correct pairs mean the system cannot detect the meaning of the sentence when substitute the special word with alternative word and charge the meaning of the sentence.

3.2.2.4 Analysis on the WSC set

Compare with the performance of coreference resolution in the AllenNLP, Stanford indicates its strength in named entity recognition, because it can identify more special in the text, for instance, time, number and so on. Those entities cannot be recognized in the AllenNLP that just can simple point out person, location, and organization. Identifying more special entities in the text can help machine understand meaning better. Due to this strength in the Stanford, it can address some pronoun coreference problems which are failure in the AllenNLP.

There is a situation that the performance of Stanford is better than the AllenNLP. That is the “the butterfly wings on the table” case which is introduced above. In the AllenNLP, system believe that “the butterfly wings on the table” as a whole entity without considering the verb “put” is connected to word “on” in front of “the table”. And the space cannot influence the result in the Stanford, no matter put a space in anywhere in the sentence, the result will be the same.

However, on the whole, AllenNLP shows a better score on the coreference. There are some instances which can be solved by the AllenNLP but not in the Stanford.

EX 1 “Since it was raining, I carried the newspaper in my backpack to keep **it** dry.”

There are two pronoun “it” in the sentence, but refer to different objects. “it was raining” means the weather and there is no special object be pointed out by it of “it was raining” in the text. On the contrary, the second “it” has substantive meaning in the text which refers to “the newspaper”.

In the Stanford system, it fails to tell the difference between those two “it” in the sentence. It considers that those two “it” are connected. The result shows in the figure 3.11,

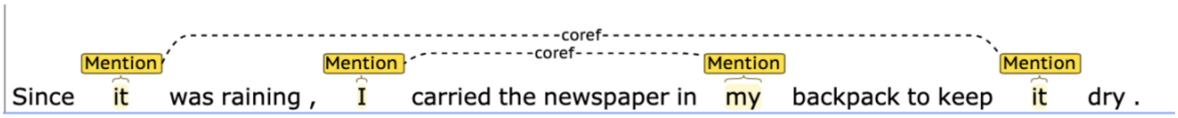
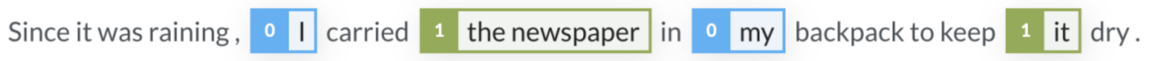

coref
coref
Mention
Mention
Mention
Mention
Since it was raining , I carried the newspaper in my backpack to keep it dry .

Figure 3.11 Coreference resolution of “Since it was raining, I carried the newspaper in my backpack to keep it dry.” in the Stanford.

And the result of the AllenNLP as given in figure 3.12,

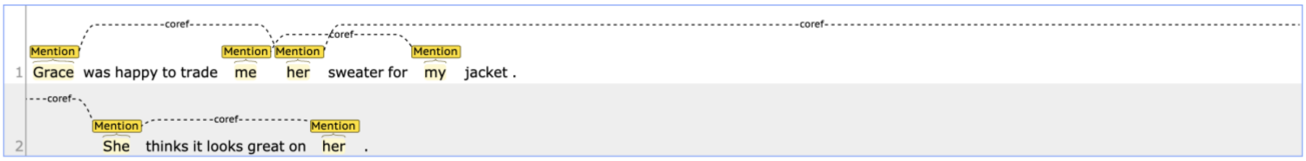


Since it was raining , 0 I carried 1 the newspaper in 0 my backpack to keep 1 it dry .

Figure 3.12 Coreference resolution of “Since it was raining, I carried the newspaper in my backpack to keep it dry.” in the AllenNLP.

EX 2 “Grace was happy to trade me her sweater for my jacket. She thinks **it** looks great on her.”

In the WSC, each example is not composed of simple sentences (one sentence or even no connected words involved), and some examples are composed of two sentences, as in the above example EX2, the second sentence beings with a pronoun “she”, but the problem for this example is not to understand the meaning of pronoun “she” instead of pronoun “it” in the second sentence. Normally, the Stanford system can deal with the situation that the pronoun begin with second sentence is what we try to solve in the WSC. But it cannot resolve the pronoun which is not the first word in the second sentence. As shown in the figure 3.13,



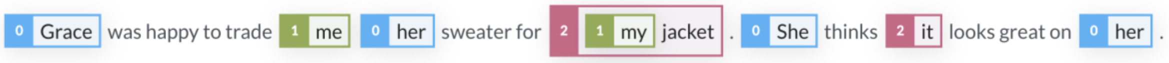
1 Grace was happy to trade me her sweater for my jacket .

2 She thinks it looks great on her .

Figure 3.13 Coreference resolution of “Grace was happy to trade me her sweater for my jacket. She thinks it looks great on her.” in the Stanford.

From the result of coreference, the Sanford system cannot point out which object in the passage is related to the pronoun “it”.

On the other hand, the AllenNLP is successful to determine the meaning of pronoun “it” in the second sentence of the example. As given in the figure 3.14



0 Grace was happy to trade 1 me 0 her sweater for 2 1 my jacket . 0 She thinks 2 it looks great on 0 her .

Figure 3.13 Coreference resolution of “Grace was happy to trade me her sweater for my jacket. She thinks it looks great on her.” in the AllenNLP.

Moreover, there are some problems that AllenNLP and Stanford both show in the coreference resolution. Such as, “the path to the lake” and “chairs in the auditorium”. Those phrases consist of an object and an attribute which describes the state of the object. In this case, both systems cannot separate object and attribute. When the pronoun in the sentence refers to the object in such phrase, both systems identify the phrase as a whole. If the correct answer for the problem of pronoun is involving two objects in the same phrase, the coreference resolution will be unclear. For instance, in the “the path to the lake” case, the WS problem is to distinguish the meaning of the pronoun whether is path or lake in the different situation. As “the path to the lake” identified as a whole, under two different situations, the pronoun in the text both represent “the path to the lake”. It will cause ambiguation about “path” and “lake” which is not what we expect.

Furthermore, as illustrate before, the difficulty of dealing the similarity word in the sentence is both existed in the Stanford and AllenNLP. The “Timmy” and “Tommy” in the sentence, system will determine those two persons as one entity.

	Correct	**Incorrect**	**No decision**	**Correct correspond sentence pairs**
Stanford	41.1% (117/285)	55.4% (158/285)	3.5% (10/285)	0% (0/142)
AllenNLP	42.8% (122/285)	54.7% (156/285)	2.4% (7/285)	3.5% (5/142)

Table 3 Experimental result of Stanford and AllenNLP system on the WSC-285 set.

3.2.3 Dependency parser

Dependency parser can generate a grammatical structure of a sentence. The syntactic structure of a sentence is described only by the word (or term) in the sentence and a set of related directed binary grammatical relationships contained in the word. The dependency parser illustrates relation among the words in a sentence, generally speaking, the dependency syntax analysis identifies the grammatical components of the subject-predicate and the fixed-form complement in the sentence, and analyses the relationship between the components.

The figure 3.14 below shows that Stanford basic dependencies of a simple WS example “The father carried the sleeping boy in his arms”. Comparing with the dependency parsing pattern in AllenNLP as shown above in figure 3.2, the Stanford pattern is clearer and more intuitive.

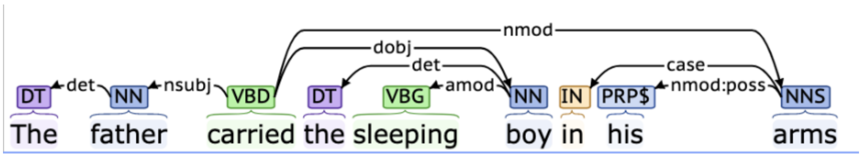


Figure 3.14 Stanford dependency parsing of “The father carried the sleeping boy in his arms”

It is clear that the relation among the words in the example is represented by labelled arcs which directly illustrate from heads to dependents. Those relation labels are abstracted from an existed grammatical relations list. Some grammatical relation descriptions are shown in the following table 4.

And the figure 3.15 shows a dependency style parse of a tree structure which focus on the constituent in the sentence. But in this structure, there is no node linking to phrasal constituents or lexical categories in the dependency parse. The structure of this dependency parse only shows the direct internal relations among the words in the sentence. However, the important information may be ignored by phrase structure tree, for instance, from the figure3.14, the modifier “arms” of the verb “carried” is directly linked to the “carried”, but in the tree structure, there is a distant in the connection from the modifier to the verb.

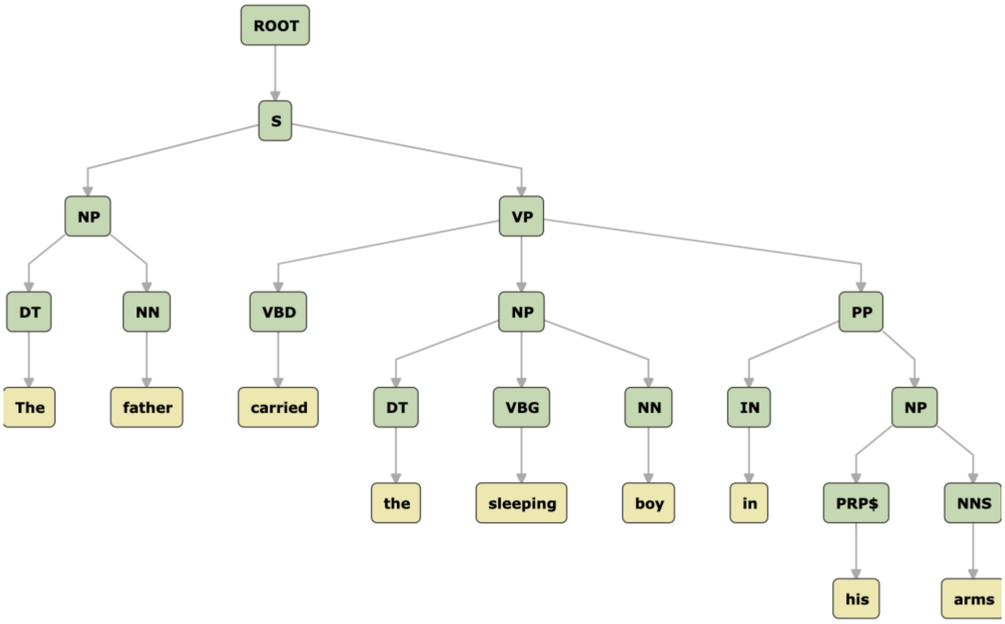


Figure 3.15 Stanford dependency parsing of “The father carried the sleeping boy in his arms”

The general relation among the words can be divided into two main parts, namely, clausal relation and modifier relation. The clausal relation is to describe the syntactic relations, such as, nsubj, dobj or iobj. The modifier relation refers to the words which can modify their heads.

Relation	**Description**
nsubj	nominal subject
xsubj	controlling subject
csubj	clausal subject
nsubjpass	passive nominal subject
case	prepositions, postpositions and other case markers
dobj	direct object
iobj	indirect object
pobj	object of preposition
nmod	nominal modifier
amod	adjectival modifier
nummod	Numeric modifier
det	determiner

Table 4 Example of dependency relations

Using the dependency parse can deal with the language which has more flexible grammar. It does not need the complete structure rule, on the contrary, phrase structure grammar needs a separate rule for each possible place in the parse tree. In the dependency parse, it extracts the information from the words order in the sentence.

3.2.4 Semantic parser

The semantic parser is a tool that can transform the natural language into a machine-understanding logic form of its meaning [38]. A semantic parser can identify the different kinds of events mentioned in a sentence, and generate a tree structure graphical semantic representation beginning with the main verb in the sentence [39]. To distinguish the events mentioned in the text and their environment so that identify the same entities or events by figuring the relation between the events and participants [16].

Sharma [39] introduced a semantic parser named Knowledge Parser which can be used to get the answer to the WS questions by reasoning on the semantic parser graph. The output of Knowledge parser is generated through five procedures, there are generating the syntactic dependency graph from Stanford dependency parser, mapping the semantic relations (represent between events and entities, such as, the cause relation between the certain events, and the relevant relation between the certain entities), determining the superclass of the entities (such as the superclass of “Anna” is “person”, the superclass of “reform” is “act”), correcting the semantic relation according to the superclass of entities, and identifying the semantic role of the entities.

Therefore, the knowledge parser as a useful tool for identifying the event can give a structure of sentence meaning rather only analysing the syntactic. As Sharma mentioned in his report [16], after extracted the semantic relation by the semantic parser, build a knowledge hunting model which can imitate the human action of learning knowledge.

However, the online demo for this parser shut down recently, I cannot examine all instances of the WS sentences and analyse the performance of determining semantic relation. According to evaluation results provided in the Sharma's report [16], the performance of this parser on the aspect of the events, entities and classes are over 0.86(precision) and 0.79(recall).

4 Classification

In this section, I will analyse and classify the WSC-285 sentences through different aspects. Considering the example grammatical and semantical structure, discuss the appropriate methods and tools for different types.

4.1 Connective word

The different connective word in the sentence may affect the result of commonsense reasoning due to the various sentiment expressing of the conjunctions.

Some instance of the WS may have more than one conjunction. As the instance, “Although they ran at about the same speed, Sue beat Sally because she had such a good start.” There are two conjunction involved in the instance, but the main conjunction which can influence the judgment of the meaning of the pronoun is word “because”. Therefore, it is necessary that machine can determine not only the conjunction but also the main conjunction in the multiple conjunctions sentence.

According to the table 5, the majority number of the connective word is “because” followed by the conjunction “so”, those two both represent for a reason. In the logic-based method of the WSC, correlation formula shows positive relation or negative relation that can be identified by the conjunction sentiment. For instance, the word “because” and “since” determine the passive relation, on the other hand, the “although” and “but” normally indicate the negative relation.

Because (77)	That (33)	Although (3)	Though (4)
But (57)	Even though (3)	When (14)	Before (1)
So (67)	Since (8)	As (10)	After (9)
And (39)	If (4)	Then (2)	Until (6)

Table 5 Statistics the number of conjunction words in the WSC-285 set.

In the table 6 shows the sentiment analysis of some conjunctions in both human judgement and AllenNLP system.

Conjunction	**Human judgement**	**AllenNLP**
Because	Positive	Positive
But	Negative	Negative
So	Positive	Positive
And	Neutral	Positive

Although	Negative	Negative
As	Positive	Positive
That	Neutral	Negative
Until	Positive	Negative
Before	Neutral	Negative

Table 6 Sentiment analysis of some conjunctions in both human judgement and AllenNLP system.

It is worth noting that some conjunctions have different meaning, such as, “as”, “since”. When the “as” as a conjunction, it can be used as “because”, “while” refers to “during the time that”, “like” refers to “in the same way” and “although”. Therefore, the polysemous conjunction word increases the difficulty of the machine understanding the meaning of the text. Besides, from the table 6, some conjunctions identified as opposite or wrong sentiment meaning. this also can influence the machine understanding of the natural language.

4.2 Complexity of grammatical structure

In English grammar, sentence structure is the arrangement of words, phrases, and clauses in a sentence. The grammatical meaning of a sentence is dependent on this structural organization.

There are four types of sentence structure, respectively, the simple sentence, the compound sentence, the complex sentence, and the compound-complex sentence. The different complexity of structure can influence the result of understanding the sentence. The compound sentence has at least two independent clauses that have related ideas.

I classify the WSC-285 set into three categories, as shown in the table 7, namely, simple single sentence, multiple clauses in a sentence, and two sentences. Under those three subclasses, the multiple clauses and two sentences can be deeply classified. For instance, the number of clauses, as the example (a) and (b) of multiple clauses in the table, (a) has three clauses and (b) has two clauses. On the other hand, from the aspect of class of two sentences, the position of pronoun which we need to understand in the second sentence is different. As the example (b) shown, the pronoun which should be identified is at the begin of the sentence, on the contrary, the example (a) is not.

There are 98 simple sentence examples, 83 multiple clauses sentence examples and 104 two-sentence type examples among 285 examples of the WSC.

Different type	Example of the WS sentence
Simple-sentence	Jane gave Joan candy because **she** wasn't hungry.
Multiple-clauses	(a) George got free tickets to the play, but he gave them to Eric, even though **he** was particularly eager to see it. (b) As Andrea in the crop duster passed over Susan, **she** could see the landing strip. (c) I'm sure that my map will show this building; **it** is very good.
Two-sentence	(a) Fred was supposed to run the dishwasher, but he put it off, because he wanted to watch TV. But the show turned out to be boring, so he changed his mind and turned **it** on. (b) The scientists are studying three species of fish that have recently been found living in the Indian Ocean. **They** began two years ago.

Table 7 Example of different structure in the WSC-285 set. Bold words are pronouns that need to be understood.

4.3 Complexity of semantical structure

The Winograd schema has not only the complexity of grammar but also the complexity of semantic. As the human, it is also not hundred percentage success accuracy on the whole WSC set, some problems are also difficult to the human.

As the introduction of the study of Bender [40] in “Establishing a Human Baseline for the Winograd Schema Challenge. In their experiment, they build an online question system, and guide the participants answer the question quickly without satisfying the accuracy. After the participants answer each question, the feedback will be given immediately with an updated score. Besides, the participants are asked to provide their age and comment about the question, the questions are whether clear and intuitive or not. There are total 407 people who finish the whole questions. The mean score of accuracy is 92.1% and the average minutes needed to complete the test is 10.2 minutes. The number of people who can answer all 40 questions correctly is 58 out of 407, where is just 14.3 percent.

Meanwhile, in the Bender’s research [39], the WSC set can be divided into easy-WSC and hard-WSC due to that the question whether can be resolved by simple techniques, such as, selectional restrictions, statistical correlations and other syntactic cues, or not. The performance on the easy-WSC is 98% accuracy which is better than the whole WS question. From this experiment, the conclusion is that the result of correctly answered by English-speaking adults is not “presumably close to 100%” [1], and the 92% on the WSC set can be indicated as a reasonable baseline for English-speaking adult performance on the WS questions. In addition, the response time of the question influence the result of accuracy to a large extent.

As a non-native English speaker, some sentence which involve specific culture or life background are hard to have the right answer immediately. For instance,

EX 1 “Jim signaled the barman and gestured toward **his** bathroom key.”

In this example, we are asked to indicate the owner of the bathroom key. the background for this question is that barman normally hold the bathroom key to prevent guests from abusing the bathroom. It is difficult to determine who keep the key of the bathroom without that certain background. Because in my consciousness, the bathroom in the bar is free to use without needing a key.

Therefore, due to the need of certain culture background for resolving problem, the WS set can be classified into easy-understanding and hard-understanding. Although all the problem solutions in the WSC need the commonsense knowledge, the understanding level of commonsense knowledge is different. The knowledge which probably should be learnt in the school or a higher-level way but not only from life, such as, the basic information about Shakespeare, Ovid, and Goethe. Even an English-speaking adult who does not have such knowledge cannot figure the correct answer for the problem involving those persons.

5 Evaluation

In this section, I will evaluate the tools and methods identified in research described above, by comparing and contrasting how they perform on the schema set and relating this to the classification of schema types. Moreover, I also evaluate the aim and objectives of the project.

5.1 Tools and methods evaluation

From the aspect of the connective word, generally, the correlation formula has a better performance on the WC set with a conjunction which shows cause relation, such as, because, so, and although. According to the result of statistics of the conjunction above, there are 147 sentences which have a cause relation conjunction. Among those sentences, there are 128 sentences which can be proved correctly by the correlation formula. The following example is tricky even though it has a "because".

EX 1 "Bill passed the half-empty plate to John because **he** was hungry."

The action "Bill passed the half-empty plate to John" cannot clearly determine who was hungry, because of "the half-empty plate". To the situation of "Bill was hungry", Bill passes the half-empty plate to John that can be explained with the fact "Bill wanted more food in the plate". On the other hand, if the situation is "John was hungry", the half-empty plate passed to John indicates that John needs the food on that plate. Both two candidates are reasonable for the pronoun in the sentence.

Therefore, the correlation's performance is shown in the table 8,

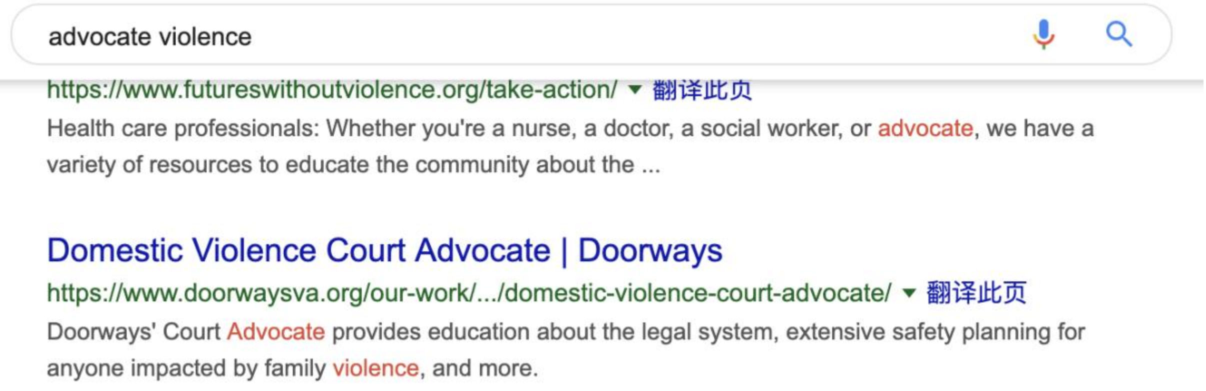
	Cause relation conjunctions	Other relation conjunctions	Total example on the WSC
Correlation's performance	128 (out of 147)	78 (out of 106)	206 (out of 285)

Table 8 the number of the correct answers by using correlation formula method on different sets.

Moreover, the correlation formula is supported by the axiom manually extracted from existing knowledge bases. The performance of correlation method on the complex grammar or semantic example set is better than other methods, especially, on the complex semantic, because using the correlation is to understand the meaning but the other methods normally analyse the syntactic structure.

Besides, the statistic method, the Google hit which introduced above does not shows significant different among various connective words. In fact, when the pronoun in the WS question refers to the certain person's name, the google hit is not available. There are only 84 example involving physical object instead of people name out of 285 examples in the WS set.

As I mentioned above, the statistic way I used in Google hit is using quoted string so that the returned results only show the number of the documents which contain the words in that specific order. The outcoming of using this statistic way is the returned number is small, a lot of example cannot determine a correct answer. But if do not use quoted string, in my opinion, the returned result cannot indicate the relation among the words. Because as long as the document contain the words in the input, it will be counted in the number of results. For example, if I input “advocate violence” without double quote into Google search, one of the returned results is a document in which does not show the meaning of advocating violence rather safety planning for family violence. From the aspect of probability, that document cannot be counted in the probability of advocating violence. Therefore, we should find a more appropriate input string for Google hit.



advocate violence

https://www.futureswithoutviolence.org/take-action/ ▾ 翻译此页

Health care professionals: Whether you're a nurse, a doctor, a social worker, or advocate, we have a variety of resources to educate the community about the ...

Domestic Violence Court Advocate | Doorways

https://www.doorwaysva.org/our-work/.../domestic-violence-court-advocate/ ▾ 翻译此页

Doorways' Court Advocate provides education about the legal system, extensive safety planning for anyone impacted by family violence, and more.

Figure 5.1 The result of Google search on “advocate violence”

In addition, the coreference resolvers, such as AllenNLP, Stanford mentioned above, present a better performance on the simple sentence than multiple clauses and two sentences.

	Number of correct instances	**percentage**
Simple-sentence type	52 (out of 98)	53.1%
Multiple-clauses type	36 (out of 83)	43.4%
Two-sentence type	46 (out of 104)	44.2%

Table 8 the number of the correct instances on the different type of grammar structure of instances.

But there is not significant difference between the easy-understanding WSC and the hard-understanding WSC when using the coreference resolvers.

5.2 Project evaluation

The aim and objective of the project are to (a) study the Winograd schema that is achieved in section 2 and 3, where I do the relevant basic researches on the aspect of Truing test, pronoun disambiguation problem and commonsense reasoning, and analyse and summarise various representative approaches, (b) analyse the tools and methods that is indicated in the section 3, where I use lots of WS examples to explain the principle of different tools and methods, (c) classify the WS instances that are determined in section 4, where I classify the WS from the aspect of conjunction, complexity of grammar and semantic. Also, in this section, (d) evaluate the performance of tools and methods on the classifications. However, the majority of examine and classification are manually achieved in the project.

6 Conclusion

In this section, I will give a personal reflection on the project and future work on the study of the Winograd schema.

6.1 Personal reflection

I enriched the knowledge of natural language processing, particularly the pronoun ambiguation problem, commonsense reasoning and logic understanding of the complex English sentences during the period of working on this project. After I did a number of the previous literature review on the relevant research, I believe the most significant difficulty of resolving the Winograd schema problems is imitating the process that human automatically learns the new knowledge. Meanwhile, the topic about the WSC does not only involve one area of knowledge which needs a comprehensive knowledge background. I constantly encounter new knowledge during the research, the dependency and semantic parsing are both new concepts to me.

Besides, as a non-native English speaker, some concepts take time to fully understand, even some WS sentences are also difficult to understand for the first time. Therefore, it is often happening in the process of researching my project that I know the meaning of every word in the sentence but I do not know what the whole sentence means. This situation can explain why the machine cannot understand the meaning of a whole sentence, currently, it is achievable that makes a machine understand the meaning of the single word but hard to make a machine understand the natural language like a native speaker under the situation that some texts are incomprehensible to a non-native speaker.

During the period of this project experience, I realize the importance of initial preparation work. At the beginning of the preparation, I separately prepared the concepts involved in the project, there is no integration and understanding of these concepts as a whole. Besides, I made a huge mistake during the preparation that I too early wrote the background part of the report. Because with the deepening of the research on the problem, I found that many new knowledge points, or the previous misunderstanding of the problem, which led to the need to modify the content written in advance. When I have a chance to write a thesis in the future, I must have a more complete understanding of the problem before I start to write a report.

Moreover, in the beginning of planning, I did not determine a clear and achievable goal of the project until the progress meeting with my supervisor and assessor who helped me to organize the ideas. I think after this experience I will increase my organizational skill and avoid the situation that unable to determine the clear aim.

To sum up, I learned a lot in this project experience, I believe it is very interesting to research an unfamiliar field even though I faced a lot of difficulties during the period of research. First and most hard is the literature review and understand those theories, it really took me lots of time.

6.2 Challenges

From the aspect of the WSC instances, the future work is to resolve the limitations of the WSC. Currently, there are only 285 instances on the WS set, the limited number of datasets restrict the precise commonsense reasoning. In addition, we should add training and validation set to guarantee the accuracy of commonsense reasoning. Also, some existing instances of the WSC denote that the machine can find the pattern of the answer without the understanding of the sentence. Thus, this will lower the quality of testing the level of AI.

From the aspect of natural language processing, the WSC did not cover the other languages. Even translating the WS sentences into other languages may cause the WS to lose the function of testing anaphora. For example, the paired instances in the WSC, “the I couldn't put the pot on the shelf because it was too tall/high.”, the “tall” and “high” are the same word in Chinese. It may cause confusion when translating into Chinese. Therefore, we should build the independent WS set for the different languages in the future.

List of References

[1] H. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[2] L. Morgenstern, E. Davis, and C. L. Ortiz, "Planning, executing, and evaluating the winograd schema challenge," *AI Magazine,* vol. 37, no. 1, pp. 50-54, 2016.

[3] V. Kocijan, A.-M. Cretu, O.-M. Camburu, Y. Yordanov, and T. Lukasiewicz, "A Surprisingly Robust Trick for Winograd Schema Challenge," *arXiv preprint arXiv:1905.06290,* 2019.

[4] "Computer wins series against Go master", BBC News, 2016. [Online]. Available: https://www.bbc.co.uk/news/technology-35785875. [Accessed: 19- May- 2019].

[5] P. LoBue and A. Yates, "Types of common-sense knowledge needed for recognizing textual entailment," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 329-334.

[6] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *arXiv preprint arXiv:1806.02847,* 2018.

[7] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE,* vol. 88, no. 8, pp. 1270-1278, 2000.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[9] Y.-P. Ruan, X. Zhu, Z.-H. Ling, Z. Shi, Q. Liu, and S. Wei, "Exploring Unsupervised Pretraining and Sentence Structure Modelling for Winograd Schema Challenge," *arXiv preprint arXiv:1904.09705,* 2019.

[10] A. Rahman and V. 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*, 2012: Association for Computational Linguistics, pp. 777-789.

[11] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of ACL-08: HLT*, 2008, pp. 789-797.

[12] P. Schüller, "Tackling winograd schemas by formalizing relevance theory in knowledge graphs," in *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2014.

[13] D. Sperber and D. Wilson, *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA, 1986.

[14] D. Wilson and D. Sperber, "Relevance theory," ed: Blackwell, 2002.

[15] A. Sharma, N. H. Vo, S. Gaur, and C. Baral, "An approach to solve winograd schema challenge using automatically extracted commonsense knowledge," in *2015 AAAI Spring Symposium Series*, 2015.

[16] A. Sharma, N. H. Vo, S. Aditya, and C. Baral, "Towards addressing the Winograd Schema Challenge—building and using a semantic parser and a knowledge hunting module," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[17] A. Sharma, *Solving winograd schema challenge: Using semantic parsing, automatic knowledge acquisition and logical reasoning*. Arizona State University, 2014.

[18] M. Gelfond and V. Lifschitz, "The stable model semantics for logic programming," in *ICLP/SLP*, 1988, vol. 88, pp. 1070-1080.

[19] C. Baral and S. Liang, "From knowledge represented in frame-based languages to declarative representation and reasoning via ASP," in *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[20] D. Bailey, A. J. Harrison, Y. Lierler, V. Lifschitz, and J. Michael, "The winograd schema challenge and reasoning about correlation," in *2015 AAAI Spring Symposium Series*, 2015.

[21] A. Emami, N. De La Cruz, A. Trischler, K. Suleman, and J. C. K. Cheung, "A knowledge hunting framework for common sense reasoning," *arXiv preprint arXiv:1810.01375,* 2018.

[22] A. Emami, A. Trischler, K. Suleman, and J. C. K. Cheung, "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*, 2018, pp. 25-31.

[23] J. Opitz and A. Frank, "Addressing the Winograd Schema Challenge as a Sequence Ranking Task," in *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 2018, pp. 41-52.

[24] E. Davis, "Qualitative spatial reasoning in interpreting text and narrative," *Spatial Cognition & Computation,* vol. 13, no. 4, pp. 264-294, 2013.

[25] V. Ng, "Machine learning for entity coreference resolution: A retrospective look at two decades of research," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research,* vol. 57, pp. 345-420, 2016.

[27] N. Nangia and S. R. Bowman, "Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark," *arXiv preprint arXiv:1905.10425,* 2019.

[28] G. Lample, A. Conneau, L. Denoyer, and M. A. Ranzato, "Unsupervised machine translation using monolingual corpora only," *arXiv preprint arXiv:1711.00043,* 2017.

[29] Q. Liu, H. Jiang, Z.-H. Ling, X. Zhu, S. Wei, and Y. Hu, "Combing context and commonsense knowledge through neural networks for solving winograd schema problems," in *2017 AAAI Spring Symposium Series*, 2017.

[30] J. G. Wolff, "Interpreting Winograd Schemas via the SP Theory of Intelligence and its realisation in the SP Computer Model," *arXiv preprint arXiv:1810.04554,* 2018.

[31] J. G. Wolff, "The SP theory of intelligence: an overview," *Information,* vol. 4, no. 3, pp. 283-341, 2013.

[32] K. Satoh and H. Yamana, "Hit count reliability: how much can we trust hit counts?," in *Asia-Pacific Web Conference*, 2012: Springer, pp. 751-758.

[33] A. Uyar, "Investigation of the accuracy of search engine hit counts," *Journal of information science,* vol. 35, no. 4, pp. 469-480, 2009.

[34] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55-60.

[35] C. Kruengkrai, N. Inoue, J. Sugiura, and K. Inui, "An example-based approach to difficult pronoun resolution," in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, 2014, pp. 358-367.

[36] G. Durrett and D. Klein, "Easy victories and uphill battles in coreference resolution," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1971-1982.

[37] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," *arXiv preprint arXiv:1707.07045,* 2017.

[38] R. Jia and P. Liang, "Data recombination for neural semantic parsing," *arXiv preprint arXiv:1606.03622,* 2016.

[39] A. Sharma, N. Vo, S. Aditya, and C. Baral, "Identifying various kinds of event mentions in k-parser output," in *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 2015, pp. 82-88.

[40] D. Bender, "Establishing a Human Baseline for the Winograd Schema Challenge," in *MAICS*, 2015, pp. 39-45.

Appendix A
External Materials

Appendix B
Ethical Issues Addressed