

**School of Computing**

FACULTY OF ENGINEERING



**UNIVERSITY OF LEEDS**

---

**Practice, analysis, and evaluation of methods solving Winograd  
Schema Challenge**

**Xiangru Cui**

**Submitted in accordance with the requirements for the degree of  
<MSc Computer Science>**

**<2018/2019>**

The candidate confirms that the following have been submitted:

Items	Format	Recipient(s) and Date
<i>Deliverables 1, 2, 3, 4</i>	<i>Report</i>	<i>SSO (11/09/2019)</i>

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**

This project is about the problems of the Winograd Schema Challenge, and will introduce several current possible solutions for solving Winograd Schema Challenge problems. To introduce these methods, the idea of every solution algorithm will be shown in details; then each solution will be tested, and the result will be analysed; finally, the limitation will be summarised.

## **Acknowledgements**

I want to thank my supervisor, Dr Brandon Bennett, who helped me to understand the Winograd Schema Challenge very patiently and gave me ideas of my project. Besides, I would like to thank my Assessor, Dr Kevin McEvoy, for the helpful feedback and gave me the right direction of the project in the progress meeting. Finally, I want to thank my friend, Zihao Li, who supported me with a high-performance laptop and help me in experiments.

# Table of Contents

<b>Summary .....</b>	<b>iii</b>
<b>Acknowledgements.....</b>	<b>iv</b>
<b>Table of Contents .....</b>	<b>v</b>
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 Overview .....	1
1.2 Problems Statement .....	1
1.3 Project Aim .....	2
1.4 Objectives .....	2
1.5 Deliverable .....	3
1.6 Methodology .....	3
1.7 Project Planning.....	3
<b>Chapter 2 Background Research.....</b>	<b>5</b>
2.1 Turing Test.....	5
2.1.1 Imitation game .....	5
2.1.2 Turing Test.....	6
2.1.3 Standard of Pass .....	7
2.1.4 Limitation of the Turing Test .....	7
2.1.4.1 Problems of deception .....	8
2.1.4.2 Not rigorous enough .....	8
2.2 Winograd Schema Challenge .....	9
2.2.1 The Basic rule of WSC.....	9
2.2.2 WSC Competition .....	11
2.2.3 The PDPs in Competition.....	11
2.2.4 Rules of advancing .....	12
2.2.5 Advantages .....	13
<b>Chapter 3 Methods and Approaches .....</b>	<b>14</b>
3.1 First-Order Logic Reasoning.....	14
3.1.1 Framework for WSC .....	14
3.1.2 Example 1 schema 41 & 42 .....	14
3.1.3 Example 2 schema 1 & 2 .....	17
3.1.4 Reasoning with Correlation.....	17

3.1.5 Limitation.....	19
3.2 Co-reference resolution methods .....	20
3.2.1 Co-reference resolution .....	20
3.2.2 Experiments and Results .....	22
3.2.3 Limitation.....	27
3.3 Google Hit Counts .....	30
3.3.1 The algorithm of GSC .....	30
3.3.2 A Better Algorithm.....	32
3.3.3 Result and Limitations.....	34
3.4 Methods for WSC .....	38
3.4.1 How to train.....	38
3.4.1.1 KEE method.....	38
3.4.1.2 FBC method.....	40
3.4.2 Result and Limitation .....	43
3.5 Discussion .....	46
<b>Chapter 4 Conclusion and Evaluation.....</b>	<b>48</b>
4.1 Achievement.....	48
4.2 Personal Reflection.....	49
4.3 Future Work .....	49
4.3.1 Complete the FBC and KEE method .....	49
4.3.2 Localisation .....	50
<b>List of References .....</b>	<b>52</b>
<b>Appendix A External Materials.....</b>	<b>55</b>
A.1 Code and Online Tools .....	55
<b>Appendix B Ethical Issues .....</b>	<b>56</b>

# Chapter 1

## Introduction

### 1.1 Overview

Nowadays, the development of various fields requires the support of artificial intelligence technology. Natural language processing (NLP), as an essential sub-area of artificial intelligence, has gradually penetrated into people's lives. NLP is also used to test whether the computer is intelligent. The well-known Turing test, which has been the most recognized test standard for many years, has been questioned in recent years, because in order to pass the Turing test, the computer has to pretend to be a human being and “fooling” the tester, rather than thinking like human beings. Meanwhile, since Winograd Schema Challenges was proposed as an upgraded test, many researchers were starting to solve the challenges, but the result is not quite excellent.

### 1.2 Problems Statement

Levesque(2012)[1] tried to improve the traditional Turing test and proposed Winograd Schema Challenges (WSC) as a new criterion for judging whether a computer is intelligent. WSC provides this set of cleverly predesigned questions to test whether a computer can “understand” text in natural language. WSC questions come in pairs, and all pairs are similar to the following example:

***“The city councilmen refused the demonstrators a permit because **they** [feared/advocated] violence.”***

*Answer: The city councilmen/ the demonstrators.*

The task of computers is to match the pronoun mentioned in the question to a corresponding participant, an event, or an object. For each pair of question, the only difference is a word or phrase, but the correct answer becomes different. For each question, 95 percent of human beings can easily find out the correct answer, but it is a difficult task for computers.

## 1.3 Project Aim

The main aim of the project is to study and analyse the WSC problems, and according to relevant approaches, find out a method or tool that can be applied to solving the Winograd Schema challenge problems. Through comparing four different types of methods, understanding and summarise basic principles of them, discuss their limitation, and then suggest the best one in different situations.

## 1.4 Objectives

- To provide the definition of the Turing Test and Winograd Schema Challenge, and explain the reason why the WSC was proposed.
- To find methods and tools to solve Winograd schema challenges from different sources.
- To understand and explain how these methods and tools works, then summarize their limitations.
- Discuss and try to find the conditions required to solve the WSC problem, identify what is the key and what is not desirable.

Approaches:

1. Logic-based knowledge representation and reasoning;
  2. Semantic method;
  3. Statistic methods;
  4. Methods using Commonsense.
- To identify several publicly available resources that can be used to either directly perform anaphora resolution or produce output that indicates structural properties of sentences that can be used for pronoun resolution.

Methods or tools:

1. First-order logic reasoning;
2. Co-reference resolution (via AllenNLP and Stanford CoreNLP);
3. Hits on Google search;



4. Commonsense Knowledge Enhanced Embeddings and BERT-commonsense.

- To identify the rule of classifying the type of Winograd Schema Challenge according to different methods.
- To evaluate the tools and methods mentioned above, including how they perform on whole set or some of the types of WSC problem.
- To identify which types of WSC problems cannot be solved well, and try to carry out an appropriate method.
- To give suggestions for appropriate optimization or improvement for some methods.

## **1.5 Deliverable**

The report includes but not limited to:









1. Background research: Description of Turing Test and Winograd Schema Challenges;
2. Methods and Tools: Description of methods mentioned above, includes the main idea of methods, what problems type they can solve, performance;
3. Evaluation: Limitation and possible improvement for methods

## **1.6 Methodology**

The methodology of this project will involve:

1. Studying academic paper about Winograd schemas;
2. Analysing the main methods and tools;
3. Understanding of various approaches to solving Winograd schemas.
4. Testing the methods or tools if necessary

## **1.7 Project Planning**

 Title	T/M	Start	End		%
<input type="radio"/> prepare for Background	 T	10/06/2019	27/06/2019	14 days	%
<input type="radio"/> First-order logic Reasoning	 T	28/06/2019	08/07/2019	7 days	%
<input type="radio"/> Co-reference resolution	 T	08/07/2019	25/07/2019	14 days	%
<input type="radio"/> Google Search Counts	 T	26/07/2019	06/08/2019	8 days	%
<input type="radio"/> Commonsense Methods	 T	07/08/2019	22/08/2019	12 days	%
<input type="radio"/> Conclusion & Evaluation	 T	23/08/2019	02/09/2019	7 days	%



## **Chapter 2**

### **Background Research**

Since the birth of the computer, it has gradually embedded in people's lives. From the beginning of the vast, slow computing computer, to the current small size computer, from high-power, large computing servers, to portable and low-power mobile phones, tablets and laptops, computers become essential tools can improve and help humans in many ways. As the concept of artificial intelligence is gradually known, the goal of computer service for human beings is gradually manifest. However, so far, to determine whether a computer is intelligent, there is still not clear enough and precautionary rule. In this part of the project, the basic idea of Turing Test and Winograd Schema Challenges will be introduced, and there will be a brief comparison of them

### **2.1 Turing Test**

The Turing Test is well-known as a rule to consider whether a computer can be intelligent; it was proposed by Alan Turing in 1950 [2]. Before he came up with this test, people had been getting caught up in a debate whether the machine can think as intelligently as human beings. However, Turing thought that was a meaningless argument, and suggested that it is supposed to be tested for machines (Levesque, Davis, and Morgenstern) [1]. To test, machines should be observed and analysed, and if a machine meets some particular conditions, then it can be defined as an intelligent machine.

#### **2.1.1 Imitation game**

The principle of the Turing test is based on the imitation game (Figure. 2.1). In the first version of his paper, Turing proposed the imitation game: This game is a group game requires three participates, at least one male and one female participate play as interrogees, and the last one plays as an interrogator. The interrogator, player X is not allowed to see and chat two interrogees face to face, but communicate with them through written notes, and the interrogator needs to ask two interrogees question, and the question can be any topic. Through considering the answers from two interrogees, player X tries to determine who is the male participant or the female participant. For one of two interrogees, Player Y, his / her task is to and hinder X from correctly determining his / her gender. Conversely, another

interrogee, player Z, he/she needs to assist the interrogator (player X) in determining his / her gender.

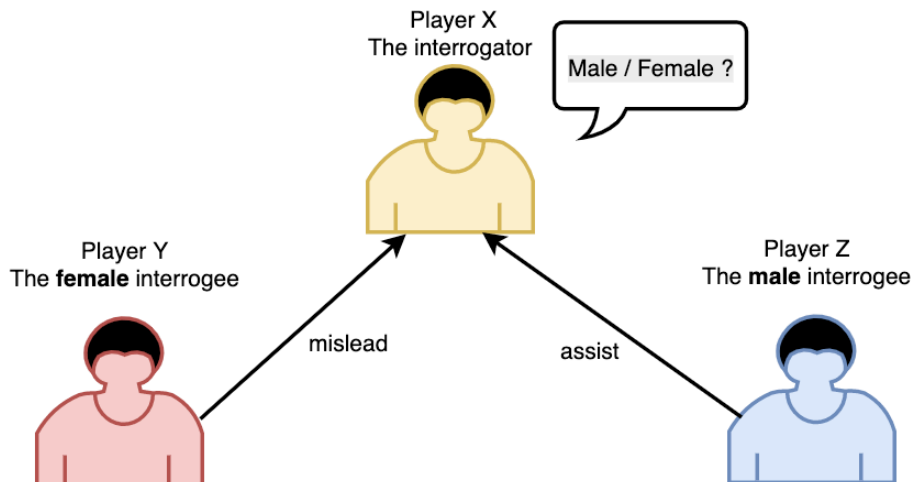


Figure. 2.1 The Imitation Game

### 2.1.2 Turing Test

The Turing Test was proposed in the second version of Turing's Paper. Similar to the imitation game, this time, the role of player Y is replaced by a computer to be tested, and both two interrogees need to pretend a female interrogees and mislead the interrogator, make the interrogator believe they are female (As shown in Figure.2.2). Also, player X is allowed to ask any question to both the computer and the man.

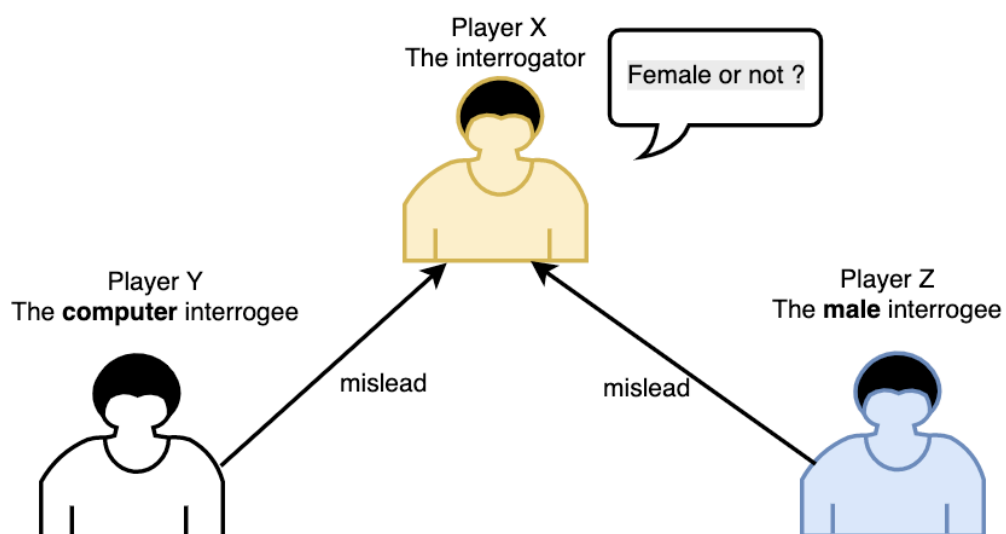


Figure. 2.2 The Turing Test

Though in Turing's paper, that situation was the only one mentioned, as it is generally understood, what Turing Test basically tried to consider is **not** whether a computer is able to simulate just a woman, but human beings (Saygin, Cicekli, and Akman, 2000) [3]. Therefore, the generally accepted rule is, to let a computer act one of interrogees, and imitate human being to communicate with the interrogator.

To test, the interrogator (player X ) must do not know it is possible a computer acting one of the interrogees. Otherwise, the player X can ask some strange question to identify a player whether it be a computer. For example, if interrogator repeat asks: "Are you a computer?", the computer may repeat answers "No, I am not", but the man may feel impatient and reply: "No, I have told you for many times."

### **2.1.3 Standard of Pass**

About the standard of passing the Turing Test, Turing did not mention any relevant information, but a prediction of future computer level:

*"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about  $10^9$ , to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification after five minutes of questioning."*

Turing believed that in the future (about in 2000), there would be a computer programmed can play imitation game and make interrogator believe that at least 30% of its answer is from a human being. Though the technologies of Artificial Intelligence is far from meeting this prediction, the "30%" standard is recognized by many people. Therefore, many Artificial Intelligence scholars use "30%" as the standard for passing the Turing test.

For instance, in 2014, at a competition marking the 60th anniversary of Turing's death, a computer programme Eugene Goostman (University of Reading 2014) [4] was considered passing the Turing Test, because it successfully made 33% of 30 human judges believe it is a 13-year-old boy.

### **2.1.4 Limitation of the Turing Test**

It has been nearly 70 years since Turing Test first proposed in 1950 [2], but there are few computers can pass the so-called Turing Test standard. Hayes and Ford (1995) [5] thought the Turing Test maybe not useful anymore.

#### **2.1.4.1 Problems of deception**

There is no doubt that the purpose of the Turing test is positive and can promote the development of artificial intelligence. But in order to pass the Turing test, the computer will deceive the judge. However, deception is not what the Turing test wants the computer to do. For example, to pass the test, computers need to not only imitate humans but give judges impressions that they are human. It is not enough to be able to communicate with judges in natural language, which means that they need fake human identities. That's why Eugene Goostman made judges to believe it is a 13-year-old boy. Computers need some basic information, such as name age. In addition, it also needs some ideas such as hobbies or attitude towards things. Therefore, to pass the Turing Test, a computer needs to lie to the judge and even needs to make a mistake deliberately, to let the judge believe that he behaves more like a human being. For example, some of the winners of the Loebner competition [6] (a well-known Artificial Intelligence competition that the format is the Turing Test), occasionally misspelled or slowed down the speed of replying (Hayes and Ford, 2014). In a word, Turing Test can be a competition of deception, at the same time, Turing Test can also be a test of the human ability to consider deception (Morgenstern and Ortiz, 2015) [7].

#### **2.1.4.2 Not rigorous enough**

The chat software Eugene Goostman is set to be a 13-year-old child. From the its dialogues with some judges [8], it can be found that when Eugene meets a question it is not able to answer, it will ignore the question and try to shift the topic, even though sometimes judges repeated a question ignored again. Moreover, the innocence and ignorance of a 13-year-old child can also be used to deceive a judge. Even it will pretend not to understand and give an answer not asked. Which is not a real intelligence, imagine that if we set a chat system as a baby boy, and reply with some garbled characters, e.g., "lafjh", or some simple words, e.g., "papa" or "mama", which will be more possible to make a judge to believe the chat system is a baby. So the standard passed through the Turing test should be the intelligence of a normal human being, not a child or a mentally handicapped person.

## 2.2 Winograd Schema Challenge

Winograd Schema Challenge (WSC) was proposed as a new standard of Artificial Intelligence (Levesque, Davis, and Morgenstern, 2012) [1], it is an updated or improved format of the Turing test. The purpose of WSC focuses more on understanding but deceptions. Through providing questions in natural languages, the WSC can test the ability of an Intelligent computer or programme to digest and understand the information. WSC is based on “Anaphora Resolution” (AR). AR is the problem of matching the pronoun to a corresponding item earlier in a natural language text (Sayed, 2003). WSC includes more than 100 schema, 285 sentences, and most of them come in pairs. The first paradigm WSC was presented by Terry Winograd in 1972 [9]:

*The town councillors refused to give the angry demonstrators a permit because they feared violence. Who feared violence?*

In this example, the question is shown: “Who feared violence?”, which means “who are ‘they’”, what does “they” refer to?; And it is provided that two answers that occur in the text:

Answer 0: the town councillors

Answer 1: the angry demonstrators

The natural answer of the question is provided by Answer 0: “the town councillors”. If the “special” word in the text is alternated to a different one, then this schema becomes a similar one, and the answer will be another one:

*The town councillors refused to give the angry demonstrators a permit because they advocated violence. Who advocated violence?*

The alternative answers of this schema are the same as the previous one, but the natural response becomes Answer 1: the angry demonstrators. In this pair of schemas, the alternative words are “feared” and “advocated”. Though it is the only difference between schemas in one pair that one word or phrase, the answers of the questions in schemas are absolutely different. This is the basic standard format of WSC.

### 2.2.1 The Basic rule of WSC

A WSC problem must be consist of:

At least one descriptive sentence must be included. In the descriptive sentence(s), there must be:

- 1 Two participants of noun phrases. Participants can be two humans names, objects, groups. But they must be the same semantic type: If participants A is male, then participants B cannot be female or an object.
- 2 One pronoun can refer to both of participants. That is why both participants need to be the same semantic type. First two rules prevent the machine from choosing the correct pronoun by judging the participants' attributes.
- 3 One pair of phrases, one of them is a special phrase, and another one is the alternate phrase. If the special one is replaced by the alternate one, then the participant that the pronoun refers to becomes another one. This prevents the machine from choosing the correct pronoun by semantics methods.

In addition, a question asking which participant is pronouns referring to, and two alternative answers that are corresponding to two participants mentioned in the descriptive text.

According to the WSC paper (Levesque, Davis, and Morgenstern, 2012) [1], a paradigm Winograd schema is the following:

**Joan** made sure to thank **Susan** for all the help **she** had [ \_\_\_\_ ]. Who had [ \_\_\_\_ ]?

Special word: received

Alternate word: given

Answer 0: Joan

Answer 1: Susan

The [ \_\_\_\_ ] in Winograd schema is the blank for special word and alternate word. A Winograd schema provides two alternative answers for the question. When the blank is filled with the special word "received", the correct answer is Answer 0 "Joan". However, when it is filled with the alternate word "given", the correct becomes Answer 1 "Susan".

In this schema, the special word "received" and the "given" expressed two absolutely opposite meanings. However, it is not necessary that two words must be antonyms for each other. For instance:

The sculpture rolled off the shelf because it wasn't [ \_\_\_\_ ]. What wasn't [ \_\_\_\_ ]?

Special word: anchored



Alternate word: level

Answer 0: the sculpture

Answer 1: the shelf

In this schema, the special word “anchored” and the “level” is not the opposite, but the answer of the two questions are still different.

### 2.2.2 WSC Competition

Four years later since WSC proposed, sponsored by Nuance Communication, the WSC competition was first held in 2016 [10]. The competition includes two rounds: the first round is to solve the Pronoun Disambiguation Problem (PDP), and the second round is to solve WSC problems. There were four entrants participating, though the highest accuracy of 58% by Quan Liu [11], the competition did not process continually because of the strict competition requirements. Hence no one did win the reward.

### 2.2.3 The PDPs in Competition

As mentioned above, the first and second rounds of WSC competition are respectively addressing the Pronoun Disambiguation Problem (PDP) and WSC problems. PDPs are similar to WSC problems, PDP is the problem of identifying referents for pronouns occurs in the natural text, but PDP has no restrictions like WSC. Morgenstern, Davis, and Ortiz (2016) [12] introduced that, in the beginning, it was considered as an annual WSC completion, and each problem was meant to be disposable. However, it is difficult to propose new WS problems. Due to the contradiction between them, and the burden of creation, the PDP was set as the first round to filter more entrants.

There were some examples mentioned in their paper:

1. Tom handed over the blueprints he had grabbed and, while his companion spread them out on his knee, walked toward the yard.

His knee: Tom / companion

2. Mariano fell with a crash and lay stunned on the ground. Castello instantly kneeled by his side and raised his head.

His head: Mariano / Castello

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

She dressed: Mrs. March / the mother

As if it had been: tea / gruel / baby

4. One chilly May evening, the English tutor invited Marjorie and myself into her room.

Her room: the English tutor / Marjorie

In their paper, it was explained that the rule of PDPs is simpler than WSC's:

Firstly, PDPs can be extracted from natural text in a wide variety of genres, modified or not, and they can be found in a text whatever it is novels, autobiography, or other genres. All instances above were extracted from different sources. In addition, they were modified very slightly except for the first instance.

Secondly, there can be more than one but rarely more than three sentences in one PDP problems. Considering the complexity of PDPs, it will not be too many sentences in one instance, as in instance 2.

Thirdly, in one PDP problem, it is possible that there are more than two pronouns to be identified, and the alternative answers for each pronoun can be multiple but binary. As shown in instance 3, there are two pronouns to be identified: "she" and "it", and three alternative answers for "it": "tea", "gruel" and "baby".

PDPs can be extracted from different genres. Hence they are easy to be constructed and can cover a variety of scenario. For an intelligent system, the basic task of both PDPs and WSC problems is to select matching a corresponding item in text to a pronoun, that is why PDP is in the first round, and why an approach designed to focus on PDPs will be mentioned later in this paper.

#### **2.2.4 Rules of advancing**

For both PDP round and WSC round, there is no difference between the rules of advancing and winning. First, the quantity of question in PDP and WSC round must be sufficient (at least 60 questions). For questions in PDP set and WSC set, they must have been tested by at least three adult human judges, and they must be answered with high accuracy [12]:

1. Most of the questions in one round (not less than 90%) should be able to be answered correctly by all judges.

2. The rest of the questions (not more than 10%) should be able to answer correctly with an accuracy of more than 50% accuracy (not less than two judges are able to answer correctly).

An intelligent system must be able to answer the questions set with an accuracy better than a threshold, and then it can advance to the next or win the competition:

$$Accuracy_{IS} \geq threshold$$

$$threshold = \text{Max} (90\%, JA - 3\%)$$

*JA*: Agreement set by all judges.

Hence Quan Liu cannot advance to the next round with 58%, though it is the highest of the competition in 2016

### 2.2.5 Advantages

From the introduction of the WSC, the advantages can be summarised:

First of all, WSC problems try to make computer focus more on “understanding” and provide questions in natural language. Hence, they can be solved by, but not exclusively, semantic methods. It is far from being able to completely solve the WSC problem that just simply analyses the structure of the text. In this way, WSC can motivate intelligent systems to understand natural language better, thinking, and solving problems like humans.

Second, the WSC has more rigorous rules, such as high-level advancing and winning requirements, more detailed criteria of screening questions in each round. The standard of an intelligent system is supposed to be expected to be as smart as an ordinary “adult human”, not any level like a “13-year-old boy” or lower.

In addition, the questions of WSC competition can cover different common-sense because the question might be extracted from a variety of genres. Through studying and analysing the intelligent systems will be exposed to more knowledge systems of different backgrounds, not limited to daily communication.

## Chapter 3

### Methods and Approaches

In this chapter, a total of four approaches will be introduced from simple to complicated, from logically reasoning, semantic to statistical prediction.

### 3.1 First-Order Logic Reasoning

#### 3.1.1 Framework for WSC

In chapter 2, WSC problem was introduced; in short, WSC problems are binary decision problems. To solve WSC, firstly, it is necessary to extract the information involved in the context. The information includes information that can be extracted directly from the text and background knowledge that needs to be supplemented based on the scenes in the sentence. Secondly, using some information, facts, and background knowledge to reason and verify that the answer is correct.

Bova and Rovatsos(2015) [13] introduce a framework for solving WSC. They divided the task into three parts:

1. Extract information in the sentences of WSC problems, and express information using First-Order Logic(FOL).
2. In addition to this, the relevant facts, truths, and knowledge base are supplemented according to the scene of the sentence and constructed as external information. Convert this external information one by one into the expression of FOL.
3. The output mentioned above is used as the input to the automatic theorem prover(ATP). If ATP can verify the correct answer by reasoning, WSC can be solved by ATP. In this part, we use Prover9 [14], an automated theorem prover for first-order.

#### 3.1.2 Example 1 schema 41 & 42

According to the rules, we tested schema number 41 and 42.

Sentence: The older students were bullying the younger ones, so we rescued them.

Pronoun: them

Candidates: The older students/the younger students

Correct answer: the younger students

Firstly, translate the sentence to FOL expression: for two candidates, the older students and the younger students are two groups that belong to the human class. Therefore, they can be expressed:

$$human(c\_older\_student). \text{ and } human(c\_younger\_student).$$

Then for the events “bullying”, we ignore the tense:

$$bully(c\_older\_student, c\_younger\_student).$$

Finally, for the clause with the pronoun “them”, since the subject is “we” that is useless, it was expressed in the passive voice. For the convenience of the experiment, we use A, B to represent two different situations of a pair of schemas, and use the same method when inputting in Goal later:

$$A \rightarrow beRescued(c\_them). \text{ } B \rightarrow bePunished(c\_them).$$

In this way, the FOL conversion for this sentence is complete. Next, we need to supplement the background, facts, rules, and truths of the sentence. Firstly, we declare that bullying is an event that causes harm. Therefore, "we" must punish the perpetrators and protect the victims:

$$all\ x\ all\ y((bully(x, y)) \rightarrow doHarmTo(x, y)).$$

and

$$all\ x\ all\ y((human(x) \& human(y) \& doHarmTo(x, y) \& x \neq y) \\ \rightarrow (bePunished(x) \& beProtected(y))).$$

Then we declare that rescue is a kind of protection:

$$\forall x((beRescued(x)) \rightarrow beProtected(x)).$$

Finally, we declare that "them" can be one of two candidates:

$$(c\_them = c\_younger\_student \& c\_them \neq c\_older\_student) \mid (c\_them = c\_older\_student \& c\_them \neq c\_younger\_student).$$

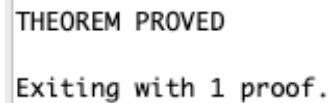
We put all the above FOL expressions into the "Assumption" of Prover9 software as input, and put following into the "Goal", then run:

$$A \rightarrow c\_them = c\_younger\_student.$$

OR,

$$B \rightarrow c\_them = c\_older\_student.$$

After click "start", both "goal"s can be proved:



```
THEOREM PROVED
Exiting with 1 proof.
```

Figure 3.1.1 The output of successful proof

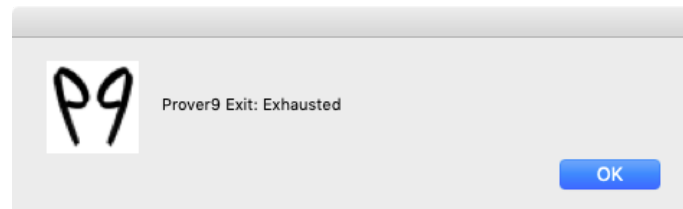
However, if we change them to:

$$B \rightarrow c\_them = c\_younger\_student.$$

OR,

$$A \rightarrow c\_them = c\_older\_student.$$

They cannot be proved, and the software will give an alert and failure output:



```
SEARCH FAILED
Exiting with failure.
```

Figure 3.1.2 The output and alert of failed proof

### 3.1.3 Example 2 schema 1 & 2

The schemas in Example 1 are relatively simple. Because there is not a lot of background knowledge to be added in these two schemas, basically only one or two supplements are needed. However, not every schema needs to be supplemented with so little background knowledge. For example, schema 1&2:

**Sentence:** The city councilmen refused the demonstrators a permit because they [feared/advocated] violence. Who [feared/advocated] violence?

**Answers:** Sue/Sally.

In this example, we need to add the knowledge that can be extracted from the sentence:

1. If A is a city councilmen, then A will fear things that are harmful to the city and the citizens.
2. If A is a city councilman, B is a demonstrator. A will pass or refuse B a permit.
3. If A will pass or refuse B a permit. Then B has a proposal, and A pass or refuses B's proposal.
4. If the city councilmen refuse the demonstrators' proposal a permit, then the city councilmen fear the content of the proposal.
5. Violence is a thing that harmful to cities and citizens
6. The content of the demonstrators' proposal is what they advocate.

These schemas need more external information as supplements than the schemas in the first Example, and the information is more complex.

### 3.1.4 Reasoning with Correlation

Bailey, Harrison, and Lierler (2015) [15] introduced correlation calculus to help to solve WSC problems. They examined the first hundred WSC problems, and 72 of them can be solved correctly, which proved that correctly combined with axioms, correlation calculus can solve many WSC problems.

In short, the correlation method is based on FOL expression. The basic expression is:

$$F \oplus G$$

Here,  $F$  and  $G$  are FOL formulas, and it means  $F$  and  $G$  are positively correlated. Similarly,  $F \ominus G$  means  $F$  and  $G$  are negatively correlated.

In their paper, it is specified that there are four basic inference rules:

1. If  $F$  implies  $G$ , then  $F$  and  $G$  are correlated:

$$\frac{\forall(F \rightarrow G)}{F \oplus G} \quad \frac{\forall(F \rightarrow \neg G)}{F \ominus G}$$

2. and the correlation between them is symmetry

$$\frac{F \oplus G}{G \oplus F} \quad \frac{F \ominus G}{G \ominus F}$$

3. The correlation is transitive:

$$\frac{\forall(F \leftrightarrow G) \quad F \oplus H}{G \oplus H} \quad \frac{\forall(F \leftrightarrow G) \quad H \oplus F}{H \oplus G}$$

4. The negation rule:

$$\frac{F \oplus G}{\neg F \oplus \neg G}$$

Based on the rules above, the correlation calculus can simplify the FOL reasoning. Three reasoning examples were given in their paper, the following one of three:

**Sentence:** The trophy doesn't fit into the brown suitcase because it is too small.

**Answers:** the trophy/the suitcase

**Correct:** the suitcase.

First step is also translate the information in sentence into FOL expression:

$$trophy(T); \text{ suitcase}(S); \neg fit\_into(T, S); small(it);$$



Secondly, we replace the pronoun with the one answer as final goal and justify whether it is correct:

$$\neg fit\_into(T, S) \oplus small(S); \text{ or } \neg fit\_into(T, S) \oplus small(T);$$

Then we supplement some background knowledges:

$$\forall x(suitcase(x) \rightarrow phycical\_object(x))$$

$$\forall x(phycical\_object(x) \rightarrow (small(x) \leftrightarrow \neg large(x)))$$

$$\forall x \forall y(fit\_into(x, y) \rightarrow large(y))$$

then reasoning process is the following:

- 1  $small(S) \leftrightarrow \neg large(S)$
- 2  $fit\_into(T, S) \rightarrow large(S)$
- 3  $\neg fit\_into(T, S) \rightarrow \neg large(S)$
- 4  $\neg fit\_into(T, S) \rightarrow small(S)$

from the last axiom above, we can make sure that “the suitcase” is correct answer.

### 3.1.5 Limitation

In this part, it is introduced that how it is used that FOL and ATP (prover9) for reasoning WSC problems. Whether it is a simple or complex schema, by translating sentences and supplemented background knowledge into FOL expressions. Take the FOL expression as input and submit it to the computer for reasoning using ATP (prover9). And it is also introduced that Correlation Calculus as a method simplifies the FOL. From the examples previous, it can be proved that the FOL method can solve the WSC problems. However, it is difficult to automatic FOL expressions of both information in sentences and external background knowledge. There is no technology that can automatically extract the FOL expression of information and supplement the background knowledge. If the FOL method depends on manual translation to FOL and input of background knowledge, it will be a very huge and impossibility completed workload. So we should find an automatic and simpler method for solving WSC problems.

## 3.2 Co-reference resolution methods

In this part, two automatic tools for solving Co-reference Resolution problems will be examined.

### 3.2.1 Co-reference resolution

Co-reference (also coreference), is a relationship between at least two expressions in a natural language text, but these expressions refer to the same referent ---- person, object or event (Sukthanker, Poria and Cambria. 2018) [16]. Co-reference Resolution (CRR) is defined by the Stanford NLP group [] as the task of clustering all these words or phrases with the coreference relation. CRR is based on not only AR (Anaphora Resolution), but also Cataphora Resolution (CR). The former (AR) is to find the referent that appears before the pronoun in a natural language text, while the latter (CR) is to find the z that appears after the pronoun. An AR example was given by Sayed (2003) [28]:

“John found the love of **his** life.”

Where “his” in the sentence refers to “John”. A CR example excerpted from Wikipedia:

“When **he** arrived home, John went to sleep.”

Where “he” refers to “John” As they are defined, each “John” occurs respectively after and before the pronouns. CRR combines the two, does not limit the location where the referent appears and does not limit the number of clusters (one or more). For instance:

“ ‘I voted for Nader because he was most aligned with **my** value.’ **she** said.”

In this sentence, “I”, “my” and “she” are in the same cluster, and “he” refers to “Nader”.

From the perspective of the problem format, both the CRR problems and the WSC problems are to find the referent that the pronoun refers to. But unlike WSC, the CRR problem does not limit the number of pronouns and participants in a piece of text. And CRR simply clusters all the people, things or others that appear in the text. For example (Figure 3.2.1), excerpted from the AllenNLP website [17]:

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 3.2.1 Example of CRR problems

The former is to find the referent that appears before the pronoun. The answer to this example is shown in the figure. There are more than two entities identified: “Paul Allen”, “Seattle”, “Bill Gates” and “Paul and Bill”, and these pronouns are not marked in advance. From this perspective, CRR problems are more difficult. On the other hand, it makes WSC more difficult that except for special words, the structures of two WSC schemas in one pair are exactly the same. Therefore, the difficulties of CRR and WSC are generally similar.

Currently, there are two open-source methods that can be found. AllenNLP coreference resolution and Stanford coreference resolution [31], Each of both methods has an online demo and WSC schemas can be tested. These methods are respectively implemented in End-to-end Neural (Lee et al., 2017) [19] and Deep Reinforcement Learning (Clark and Manning, 2016) [20]. They find a possible solution through analyse the semantic structure of sentences.

For example:

Schema 3: The trophy doesn't fit into the brown suitcase because it is too large.

The results of the two methods online are shown in the following figures (Figure 3.2.2):

0 The trophy does n't fit into the brown suitcase because 0 it is too large .

### Coreference:

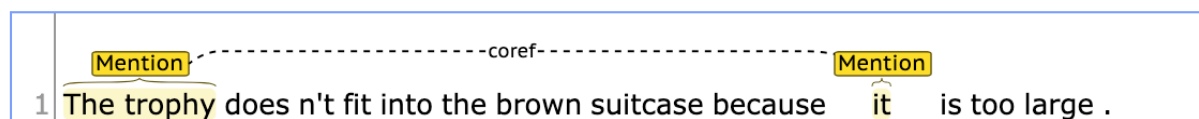


Figure 3.2.2 Result from AllenNLP and Stanford CoreNLP

It can be found that both methods are able to solve this schema correctly: “it” refers to “the trophy”. Therefore, CRR methods may be able to solve the WSC problems.

### 3.2.2 Experiments and Results

In this part, I used the AllenNLP python package, which version is 0.8.5. I also tested with the Stanford CoreNLP 3.9.2 updated on 5<sup>th</sup> October 2018, but the online demo is the version 3.9.2 updated on 29<sup>th</sup> November 2018. The results from two methods are obtained through tests on **16<sup>th</sup> July 2019** and **15<sup>th</sup> August 2019**. Therefore they may not be the same with that using the newest version or online demo.

Before showing the results, it must be declared that, in the 285 WSC set, all the schemas come in pairs except for schema No.253, 254 and 255. All of these three schemas come in one group.

For the result from AllenNLP (Table 3.2.1):

	Correct(%)	Wrong(%)	Half correct(%)	Not recognised(%)
For single	122(42.8%)	163(57.2%)	—	43(15.1%)
For pair (both)	5(3.55%)	25(17.7%)	111(78.7%)	14(9.93%)
For Group of 3	1	2	—	2

Table 3.2.1 The result from AllenNLP

For the result from Stanford CoreNLP (Table 3.2.2):

	Correct(%)	Wrong(%)	Half correct(%)	Not recognised(%)
For single	123(43.2%)	162(56.8%)	—	42(14.7%)
For pair (both)	2(1.42%)	22(17.0%)	117(81.6%)	17(13.5%)
For group of 3	2	1	—	0

Table 3.2.2 The result from Stanford CoreNLP

In two tables, “Group of 3” means schemas No.253, 253 and 255. In the experimental results, the results of "Not recognised" included that there were no clusters in the results, for example, schema 258:

Document

I tried to paint a picture of an orchard, with lemons in the lemon trees, but they came out looking more like light bulbs.

Run

I tried to paint a picture of an orchard , with lemons in the lemon trees , but they came out looking more like light bulbs .

Figure 3.2.3 Result of schema 258 from AllenNLP

— Text to annotate —

I tried to paint a picture of an orchard, with lemons in the lemon trees, but they came out looking more like light bulbs.

— Annotations —

coreference ✕

### Coreference:

1 I tried to paint a picture of an orchard , with lemons in the lemon trees , but they came out looking more like light bulbs .

Figure 3.2.4 Result of schema 258 from Stanford CoreNLP

and the results were not related to the WSC answer. For example, schema 274:

Document

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

0 This book introduced Shakespeare to Ovid ; 0 it was a major influence on his writing .

Figure 3.2.5 Result of schema 274 from AllenNLP

The "Not recognised " class is a subclass of the error class. The subclass of the error class also has a "normal error class" and the result is the opposite of the correct answer.

The result from AllenNLP can be seen from Table 3.2.1: From the perspective of a single schema, 122 schemas were correctly answered, with a correct rate of 42.8%, less than half. If we only focus on the schema that can be recognized normally, the "Not recognised" is not counted, the correct is only 50.4%, slightly higher than the random selection of 50%. From the perspective of a pair of schemas, there are only five pairs (accounted for 3.55%) that both schemas in each pair were correctly answered. they are, respectively, schema 17 & 18, 35 & 36, 97 & 98, 179 & 180 and 272 & 273. Result of one pair of them:

☐ Jim yelled at Kevin because ☐ he was so upset .

Jim comforted ☐ Kevin because ☐ he was so upset .

Figure 3.2.6 Result of schema 35 & 36 from AllenNLP

There are 110 pairs (78%), only one of each pair is wrong. Because AllenNLP gave the same answer to both of each pairs. For example:

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

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

Figure 3.2.7 Result of schema 1 & 2 from AllenNLP

There are a few (15 pairs) that are unable to identify one of them and cause a wrong result. Due to the online demo and the version of this experiment, the results of the AllenNLP are given here.

```
CAN'T BE ANALYSED
('Bill', 'PERSON');('John', 'PERSON');

Bill,= his; ('Bill', 'PERSON');('John', 'PERSON');
```

Figure 3.2.8 Result of schema 229 & 230 from AllenNLP

There are 14 pairs (9.93%) of "Not recognised" in 25 pairs (17.7%) of "errors". There remain 11 pairs whose results are one "wrong" and one "Not recognised":

```
The sack of potatoes , the bag of flour , = it;
The sack of potatoes , = it;
```

Figure 3.2.9 Result of schema 37 & 38 from AllenNLP

(They can both be answered correctly by the newest version of AllenNLP)

here are also some "normal errors class ":

```
the chickens , = them;

The foxes , = them;
```

Figure 3.2.10 Result of schema 155 & 156 from AllenNLP

From the above results, the solution of AllenNLP cannot be considered ideal. After all, it just achieved a random level with "not recognised" not accounted. In addition, there are too many pairs of results that are "half-correct", which means that AllenNLP does not really understand the information mentioned in the text.

For the results from Stanford CoreNLP, from the perspective of a single schema, There were 123 (43.2%) answered correctly, only one more than that from AllenNLP. There are 43 "Not recognised", this is the same as the number of AllenNLP result. If we only focus on the schema that can be recognized normally, the "Not recognised" is not counted, the accuracy is very close to the result of AllenNLP: 50.8%. From the perspective of a pair of schemas, The total number of pairs is three fewer than AllenNLP. Only two pairs are both correct: schema 260 & 261 and 266 & 267. One of the results:

**Coreference:**

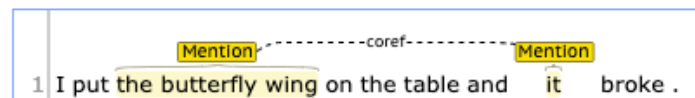


Figure 3.2.11 Results of schema 266 & 267 from Stanford CoreNLP

**Coreference:**

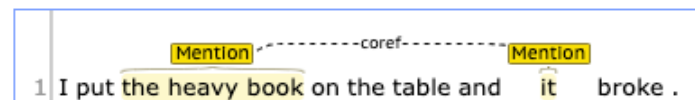


Figure 3.2.12 Results of schema 266 & 267 from Stanford CoreNLP

Similarly, similar to Allen's results, the result of 117 pairs is "half-correct". Among them, all the results mentioned in AllenNLP are also included. Most of the reasons for getting "half-correct" results are also "one correct and one wrong". This shows that Stanford CoreNLP is also unable to understand the information in the text. It is worth mentioning that 92 of the answers from the two methods are different. Of these, 42 results are the opposite of the answers AllenNLP gets, and other results are different in their ability to identify entities. Such as:

0 Mary took out 0 her flute and played one of 0 her favourite pieces . 0 She has had it since 0 she was a child

**Coreference:**

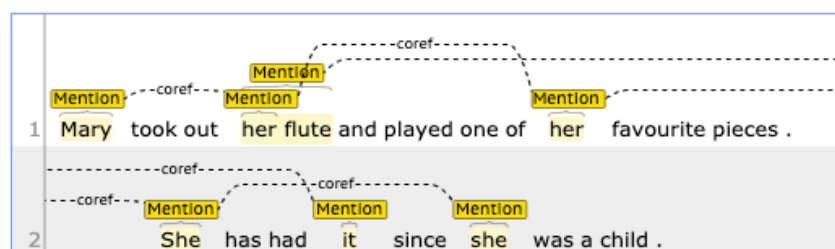


Figure 3.2.13 Results of schema 121 from AllenNLP and Stanford CoreNLP



### 3.2.3 Limitation

From the results of the two methods, the accuracy of the CRR methods is not high, and the main reasons are two:

First, the ability to identify entities is low. About 15% of the sentences in the sentence are wrong for the entity to identify. And it was found strange: for schema 10,

Sentence: The lawyer asked the witness a question, but **he** was reluctant to answer it.

answer: the lawyer/the witness

correct: the witness

the results of AllenNLP and Stanford CoreNLP are both wrong:

0 The lawyer asked the witness 1 a question , but 0 he was reluctant to answer 1 it .

Coreference:

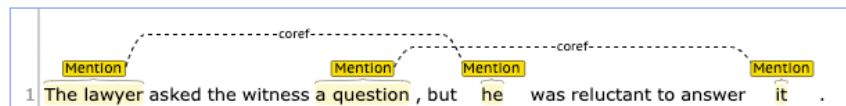


Figure 3.2.14 Results of schema 10 from AllenNLP and Stanford CoreNLP

But if we replace “the witness” with a name “Bob”:

The lawyer asked 0 Bob a question , but 0 he was reluctant to answer it .

Coreference:

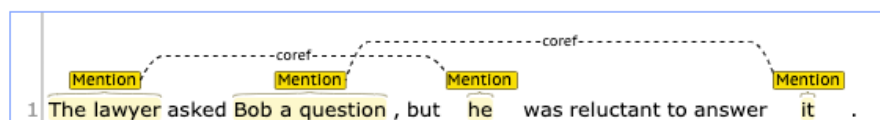


Figure 3.2.15 Results of schema “Bob” from AllenNLP and Stanford CoreNLP

The result from AllenNLP is correct and can recognise the “Bob” correctly. But Stanford CoreNLP cannot answer correctly. If the name is “John”, the results are the same. But if the name is “Jack” or “Kevin”, the results from both NLPs is as same as the original schema:

0 The lawyer asked Jack a question , but 0 he was reluctant to answer it .

### Coreference:

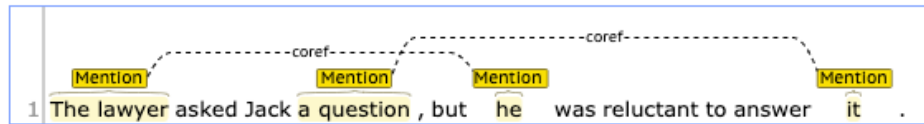
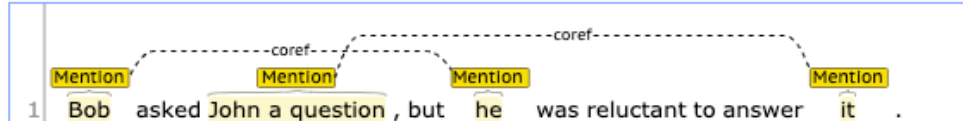


Figure 3.2.16 Results of schema "Jack" from AllenNLP and Stanford CoreNLP

Now, we replace both "lawyer" and "witness" with two names can be recognised: "John" and "Bob":

Bob asked 0 John a question , but 0 he was reluctant to answer it .

### Coreference:



0 John asked Bob a question , but 0 he was reluctant to answer it .

### Coreference:

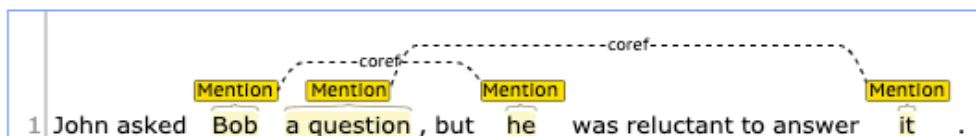


Figure 3.2.17 Results of schema "Bob" and "John" from AllenNLP and Stanford CoreNLP

It seems that for AllenNLP, "John" has a higher priority than "Bob". However, for Stanford CoreNLP, the priority is just the opposite. Therefore, for both NLPs, "John" and "Bob" may appear more often in their respective training sets.

Secondly, there is no process for analysing the information in the text in two methods. Thus, even if their recognition success rate can be improved, the result is obtained by a semantic analysis method, this does not meet the requirements of the computer to understand natural language text. This makes it impossible for the computer to choose the correct alternative answer through the knowledge background. This makes it impossible for the computer to choose the correct alternative answer through the knowledge background. Overall, the correctness of pairs of schemas two methods, and it is possible that the training set used by the method contains them. Solving the WSC problem still requires the extraction and understanding of the knowledge and information of the text.

## 3.3 Google Hit Counts

Through calculating the numbers of Google search results, Google Searching Counts (GSC) judges the answers to the WSC problems based on the results of the calculations. This method can use Google's search results for an item to introduce commonsense.

### 3.3.1 The algorithm of GSC

There is no clear or best definition of the GSC algorithm. The easiest solution is the following.

Algorithm 0:

- i. Replace the question word with the pronoun, and convert the question of WSC to declarative sentence from interrogative sentence;
- ii. Replace the pronouns in the sentence from step (i) with one of the alternative answers;
- iii. Search for the sentence from step (ii) and record the results;
- iv. Compare the results from step (iii), and the larger one is the answer.

For example:

Schema: The trophy doesn't fit into the brown suitcase because it is too small.

Question: What is too small

Answer 0: the trophy

Answer 1: the suitcase

To solve this schema:

- i. "It is too small".
- ii. "The trophy is too small" and "The suitcase is too small."
- iii. The result of "The trophy is too small" is 75,900,000,  
The result of "The suitcase is too small." is 65,000,000.
- iv. So the answer is Answer 0.

However, it should be Answer 1 that the correct nature response of this schema.

In algorithm 0, its idea seems to be no big problem: the more an item Google can search, the more likely it is that this item is the correct answer. But an important detail of this

algorithm is ignored: the relationship, or the correlation between two answers and the question. So this algorithm can be improved, it is supposed to be compared to the relative proportion but absolutely counting.

Improved algorithm 0 ---- algorithm 1:

- i. Replace the question word with the pronoun, and convert the question of WSC to declarative sentence from interrogative sentence;
- ii. Replace the pronouns in the sentence from step (i) with one of the alternative answers;
- iii. Search for the answers and record the results  $R^{A_0}$  and  $R^{A_1}$
- iv. Search for the sentence from step (ii) and record the results  $R^{S_0}$  and  $R^{S_1}$ . Calculate  $P_0 = S_0/R^{A_0}$  and  $P_1 = S_1/R^{A_1}$ .
- v. Compare the results from step (iii), and the larger one is the answer.

In this schema, the result of “the trophy” is 654,000,000, the result of “the suitcase” is 256,000,000, so

$$P_0 = \frac{75,900,000}{654,000,000} \approx 11.6\% < P_1 = \frac{65,000,000}{256,000,000} \approx 25.4\%.$$

According to Algorithm 1, Answer 1 with a larger relative proportion  $P_1$  is the answer is Answer 1, which is the same as the correct nature response of this schema.

In conclusion, Algorithm 1 is more reasonable and more appropriate than the previous Algorithm 0, whether it is logically explained or according to the results of relative proportions calculation.

For this schema, if the special word “small” is replaced by the alternate word “big”, does this algorithm still work?

The result of both full sentences are 133,000,000 and 70,600,000, so:

$$P_0 = \frac{133,000,000}{654,000,000} \approx 20.3\% < P_1 = \frac{70,600,000}{256,000,000} \approx 27.6\%.$$

due to the slight difference, the answer is still Answer 1, while the correct nature response changes to answer 0 “the trophy is too big”.

### 3.3.2 A Better Algorithm

Rahman and Ng (2012) [21] proposed an algorithm using the count results returned from Google search, and it is more logic. Inspired by their idea, a better algorithm will be proposed in this part later.

Rahman and Ng introduced their algorithm in the paper. Initially, they define different part with a capital letter. In a descriptive sentence  $S$  in a schema:  $V$  represents the verb comes with the main pronoun;  $W$  is the series of words with the verb  $V$ ;  $C_1$  and  $C_2$  are respectively two alternative answers. If there is an adjective after the verb  $V$  in  $W$ , let  $J$  be the adjective.

There will be at least 4 query task:

$$\begin{aligned} Q_1: C_1V; & \quad Q_2: C_2V; \\ Q_3: C_1VW; & \quad Q_4: C_2VW. \end{aligned}$$

If  $J$  exists, then

$$Q_5: JC_1; \quad Q_6: JC_2.$$

For the schema, "The trophy doesn't fit into the brown suitcase because it is too **big**."  $Q_1$  to  $Q_6$  will be shown in Table 3.3.1. :

$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$
"the trophy"	"the suitcase"	"the trophy is big"	"the suitcase is big"	"the big trophy"	"the big suitcase"

Table 3.3.1 The query tasks

The rules of judgment are relatively simple:

Let  $Cnt_1$  to  $Cnt_6$  be the count results of  $Q_1$  to  $Q_6$ , Group two adjacent results into a total of three groups:  $(Cnt_1, Cnt_2)$ ,  $(Cnt_3, Cnt_4)$  and  $(Cnt_5, Cnt_6)$ . Compare result in the first group, if one result count is greater than another at least 20%, then the preferred answer is the greater one; if not, continue to compare the next group until the last one is compared. If no group meet the condition "20%", then the answer of this schema cannot be considered.

For both schema:

*"The trophy doesn't fit into the brown suitcase because **it** is too small."*

and

*“The trophy doesn't fit into the brown suitcase because it is too big.”*

the results of query tasks are the following: Table 3.3.2 and Table 3.3.3

	$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$
Result	630,000,000	292,000,000	276,000,000	125,000,000	265,000,000	221,000,000
Rate	215.8%		220.8%		119.9%	

Table 3.3.2 Result of “small” one

	$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$
Result	630,000,000	292,000,000	372,000,000	135,000,000	327,000,000	157,000,000
Rate	215.8%		275.6%		208.3%	

Table 3.3.2 Result of “big” one

Regarding the rule, for both sentences, the process ended after the first round comparison and then generated answers. However, both generated are the same: “the trophy is too big / small”. Obviously, one of the answers is wrong. From algorithm, it takes into account the impact of the control group more comprehensively, while ignores the relative proportions. To improve this algorithm, we stay the search tasks be the same without changing the rules of judging answers:

If there are no  $Q_5$  and  $Q_6$  of a sentence, we calculate the  $\frac{Q_3}{Q_1}$  and  $\frac{Q_4}{Q_2}$ , and let the greater one be the answer. If there are  $Q_5$  and  $Q_6$ , we first calculate  $\frac{Q_3}{Q_1}$  and  $\frac{Q_4}{Q_2}$ , if any result is greater another at least x% ( $\frac{Q_3}{Q_1} - \frac{Q_4}{Q_2} \geq X\%$ ) or is X times greater, then let the greater one being the answer; if not, we calculate  $\frac{Q_5}{Q_1}$  and  $\frac{Q_6}{Q_2}$ , and the greater be the answer. X% is the threshold that ensures the big enough difference; in the experiments, we set X as 10.

In this way, there will be an answer for every schema, and the algorithm tries to consider every parameter. It will be more logical and reasonable algorithm:

I examined the algorithm with two schemas above (trophy and suitcase). The result is shown in Table 3.3.3 and 3.3.4. For the first one, the difference between  $\frac{Q_3}{Q_1}$  and  $\frac{Q_4}{Q_2}$  is not enough,

so after comparing  $\frac{Q_5}{Q_1}$  and  $\frac{Q_6}{Q_2}$ , the answer is “the suitcase”. The same reason, after comparing  $\frac{Q_3}{Q_1}$  and  $\frac{Q_4}{Q_2}$ , the answer is “the trophy”.

“trophy”	$Q_1$ “the trophy is”	$Q_3$ “the trophy is small”	$Q_5$ “the small trophy”	$\frac{Q_3}{Q_1}$	$\frac{Q_5}{Q_1}$
Result	630,000,000	276,000,000	265,000,000	43.8%	42.1%
“suitcase”	$Q_2$ “the suitcase is”	$Q_4$ “the suitcase is small”	$Q_6$ “the small suitcase”	$\frac{Q_4}{Q_2}$	$\frac{Q_6}{Q_2}$
Result	292,000,000	125,000,000	221,000,000	42.8%	75.7%

Table 3.3.3 Proportions of “small” one

“trophy”	$Q_1$ “the trophy is”	$Q_3$ “the trophy is big”	$Q_5$ “the big trophy”	$\frac{Q_3}{Q_1}$	$\frac{Q_5}{Q_1}$
Result	630,000,000	372,000,000	327,000,000	58.9%	51.9%
“suitcase”	$Q_2$ “the suitcase is”	$Q_4$ “the suitcase is big”	$Q_6$ “the big suitcase”	$\frac{Q_4}{Q_2}$	$\frac{Q_6}{Q_2}$
Result	292,000,000	135,000,000	157,000,000	46.2%	53.8%

Table 3.3.4 Proportions of “big” one

### 3.3.3 Result and Limitations

There are 285 schemas in WSC, but there were only about 120 schemas can be examined. In the beginning, the accuracy was high, but with the number of schemas tested grows, the accuracy became lower. 56 of 120 schemas can be answered correctly. Accuracy is 46.7%, below 50%. Note that this is not at all what we want!

Through experimentation, there are some limitations that can be drawn.

Firstly, This algorithm cannot cover all schemas. There are just 120 schemas can be examined. In some schema of the WSC set, there exists a question of identifying names. For example:

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



The natural response of this schema is “George”, but if we exchange the position of two names, the answer will be “Paul”. Even though they can be processed by the algorithm, this is a meaningless try.

Secondly, the algorithm succeeds some simple examples, such as the “trophy and suitcase” one above, and this pair:

The sculpture rolled off the shelf because **it** wasn't **anchored**.

The sculpture rolled off the shelf because **it** wasn't **level**.

For each of these two pairs, the algorithm works correctly. However, if the sentence structure is complex, the algorithm can only answer one or even none of one pair. For example:

Sam pulled up a chair to the piano, but **it** was broken, so he had to **stand** instead.

Sam pulled up a chair to the piano, but **it** was broken, so he had to **sing** instead.

For both schemas, the search tasks of them are the same, because the special word does not come with the pronoun. This makes the algorithm aware of the inability to special words, which makes the algorithm unable to receive the commonsense in the sentence. So even if the algorithm can answer one of them, it cannot answer another correctly. This is why the test accuracy rate in the first part (40 schemas) is considerable. For the first 40 schemas, there were 24 schemas answered correctly (60% accuracy). And 7 of 20 pairs were full-correct (both schemas in one pair were answered correctly). However, for the high-complexity sentences, 80 schemas were tested. Thirty-two of them were correctly answered (40% accuracy). 11 of 40 pairs were full-wrong (both schemas in one pair were not answered correctly), but only two pairs were full-correct.

It can be a potential issue that the existence of ambiguous words and phrases, for example:

I couldn't put the pot on the shelf because it was too **tall / high**.

In schema 73 and 74, the “pot” can represent the kitchenware used for cooking, while “pot” can also represent a kind of drug, marijuana. There are 13 different means listed by Longman Dictionary. Obviously, “the pot” does not represent a toilet or marijuana in this schema. It is more likely to represent a container than these unreliable meanings. Therefore, there could be a lot of useless results that contain the keywords, which does not represent

the same meanings as they represent in the schemas. However, Google search cannot filter the results with a different sense from that in schemas

Also, in Davis' report (2015), he mentioned some bugs of Google Search. He tried to search "it is you who are mad", the result was "About 2,840,000 results", but there were only four pages in a total of 40 results. As of now (19th August 2019), this error still exists (shown in Figure3.3.1 and 3.3.2). There are six pages and 49,000 results. But on the last page, the result changes to 51 results (shown in Figure 3.3.3).

In summary, this algorithm is low-coverage and does not solve complex sentences well. Due to the tools used, Google search, the results of some searches are not very accurate, and there are errors, the algorithm is not very good overall.

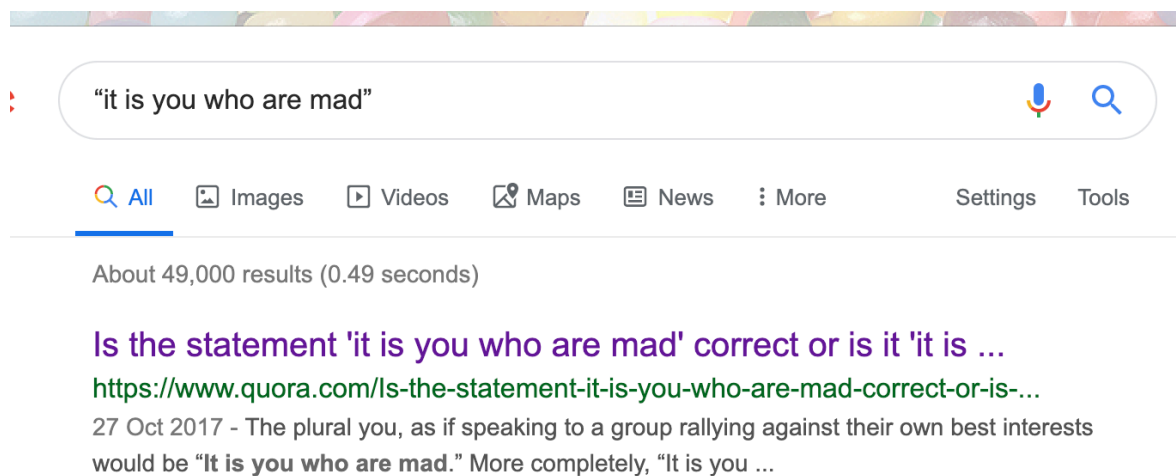


Figure3.3.1 Result at the top



### Searches related to "it is you who are mad"


mad at meaning	mad at something
i m getting mad at you	how to ask are you angry with me
i am mad at you meaning in hindi	mad in american english
are you angry at me	are you home yet





Figure3.3.1 Result at the bottom


"it is you who are mad"





 All

 Images

 Videos

 Maps

 News

 More

Settings

Tools

Page 6 of about 51 results (0.38 seconds)

**QE Madness - The New York Times**

<https://krugman.blogs.nytimes.com/2010/11/05/qe-madness/> ▼

5 Nov 2010 - It has been really interesting to watch some of the commentary over quantitative easing by the Fed: while people like me see the Fed's actions ...

Figure 3.3.3 Result on Page.6

## 3.4 Methods for WSC

There were three different methods introduced in the previous three sections. They use different methods, but the correct rate is not ideal. In this section, two methods of using commonsense will be introduced. These two methods use the fine-tuned BERT-commonsense (FBC) (Kocijian et al., 2019) [29] and commonsense knowledge enhanced Embeddings (KEE) methods (Liu et al., 2016) [24]. The KEE method is for the PDP problem. In the background section, it has been introduced, and the difficulty of PDP and WSC is the same, so KEE is feasible as a method for solving PDP problems to solve WSC problems.

### 3.4.1 How to train

Both methods use unsupervised learning (USL), and a large-scale training set is required for machine learning. However, the available PDP problems and WSC problems are not sufficient for machine learning, because there are too few examples in PDP and WSC questions sets, and WSC and PDP problem sets are set to default as the test set and validation set, so they cannot be set as training sets. Since the existing problems set cannot be used as a training set, why not create a new training set? In the background section, it was also introduced, it is hard to create WSC problems and PDP problems. This process of creation is inseparable from creativity and inspiration. And currently, there is no recognized problem set similar to the WSC problem model with a large-scale. The biggest challenge in using WSR to solve WSC is to find a set of training.

#### 3.4.1.1 KEE method

KEE uses two training sets: one is the corpus of books from Project Gutenberg, it is called CBTest; and another one is the corpus of Wikipedia, which is similar to Wikimedia. In addition, KEE also used various of Commonsense Knowledge bases (CSKB) to train the word similarity model. Combined Corpus with CSKBs to analyse the similarity relationship of words in each sentence, KEE generates an analytical model and use this model to speculate on the candidate of the pronoun.

In KEE method, there were 3 different CSKBs used: ConceptNet[24], WordNet[25], and CauseCom[26].

First, ConceptNet is a knowledge base that contains a lot of relationships between words or phrases. For every two different words or phrases, the relationship between them is denoted by a triple:  $(word_1, relationship, word_2)$ . For example, let the word "suitcase" be an input, a

part of the results from ConceptNet online demo is the following Figure 3.4.2. The first one, “Suitcase is used for carrying clothes”, it can be denoted as:

*(suitcase, isUsedFor, carrying clothes).*

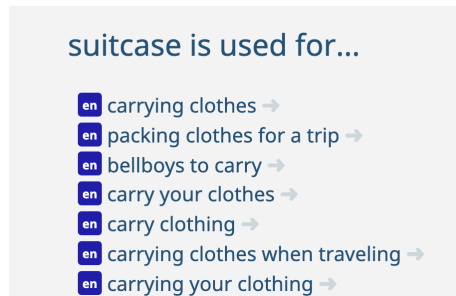


Figure 3.4.2 part of the results of “suitcase”

Second, WordNet is a huge dataset of English like an English dictionary. For every word or phrase, it provides a concise definition, an example sentence, the synonyms of the word, and the category of the word. The result of “suitcase” from WordNet is shown in Figure 3.4.3. It can be found that there is a number at the beginning of the result. This number is the location the words be in the dataset. The word and its synonyms are in the same location. The location is used to consider the similarity between two words, which will be introduced hereafter.

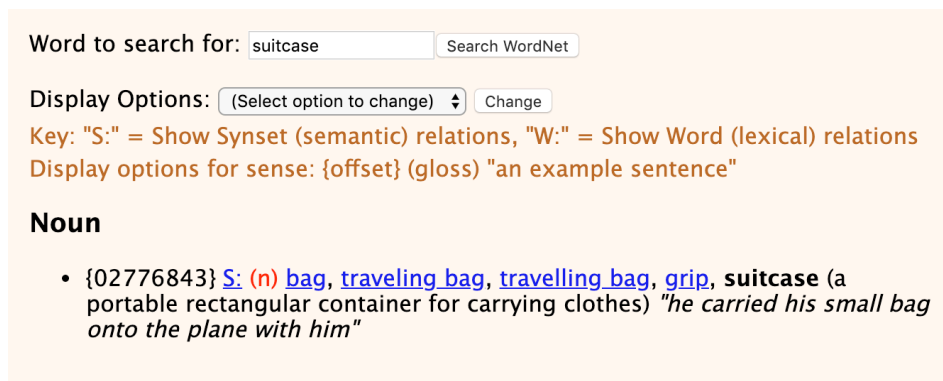


Figure 3.4.3 The result of “suitcase” from WordNet

The last one is CauseCom is a dataset consist of cause-effect pairs(Liu et al., 2016) [26]. Each pair is extracted automatically from natural language text (Figure 3.4.4). For example, the sentence, “Playing basketball’ can cause ‘win’”, is denoted as *(play basketball, win)*.

To obtain appropriate similarity results for every words pair, there must be reasonable rules of calculation for all the CSKBs:

- 1 For ConceptNet, the similarity between two words/phrases with at least one relationship must be larger than that without relationship;

- 2 For WordNet, the similarity between two words/phrases with a shorter location distance must be larger than that with a longer different location distance. The location is the number at the beginning aforementioned;
- 3 For WordNet, the similarity between words/phrases with the same semantic type must be larger than that with a different semantic type;
- 4 For CauseCom, the similarity between words/phrases with a cause-effect relation is larger than that without cause-effect relation.

Based on the rules above, the KEE method uses the similarity produced to train the solver.



Figure 3.4.4 Example of cause-effect events [26]

### 3.4.1.2 FBC method

The KEE uses machine learning to train computers with original sentences in datasets, so that predict the candidate that the pronoun refers to. However, facing the same situation of a training model without a training set of WSC problems, The FBC method chooses to use a pre-trained language model, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) [23]. The pre-trained model, the BERT is now available and referenced from Pytorch. Masked Language Model in BERT can predict “masked” words in a sentence based on context. By using the training set to train the BERT, the language model is tuned to produce a language model suitable for WSC problems. To solve the problem of no training set, Kocijjan et al. tried to create a data set with a large number of "WSC problems", enough to serve as a training set. How did they create a training set? First, sentences are extracted and analysed from the English Wikimedia (EW) dataset [28], which

is different from Wikipedia that KEE used. There are a large number of English texts in EW, including many genres. They use the WSC rules to filter sentences from this large text set that can be used as WSC questions.

In a sentence extracted, there must be a noun or a person appear twice. And in addition to the restrictions on nouns and names, there is at least one word with the same part of speech (noun or person name), and this (these) word(s) appears only once in the sentence, then the sentence will be extracted. For example:

*“The syllables are pronounced strongly by **Gaga** in syncopation while her vibrato complemented Bennett's characteristic jazz vocals and swing. Olivier added, '**Gaga**'s voice, when stripped of its bells and whistles, showcases a timelessness that lends itself well to the genre.”*

This sentence is taken from the Wikipedia entry "Cheek to Cheek (album)" [27]. In this sentence, "Gaga" as a name appears two words (in bold), "Bennett" and "Olivier" as names appear once (underlined).

This screening process utilizes the Stanford POS-tagger (SPOST) implementation. E.g., input the following sentence:

The trophy doesn't fit into the brown suitcase because it is too large.

text	xops	text	xops
The	DT	Joan	NNP
trophy	NN	made	VBD
does	VBZ	sure	JJ
n't	RB	to	TO
fit	VB	thank	VB
into	IN	Susan	NNP
the	DT	for	IN
brown	JJ	all	PDT
suitcase	NN	the	DT
because	IN	help	NN
it	PRP	she	PRP
is	VBZ	had	VBD
too	RB	received	VRN
large	JJ	.	.
.	.		

Figure 3.4.1 Results of two sentences from SPOST

The output from SPOST is shown in Figure 3.4.1. The "trophy" and "travel box" in the sentence have been marked as "NN" as a noun. If the noun in the sentence is plural, the noun will be marked as "NNS". The name of the person appearing in the sentence will be Marked as 'NNP' for example:

Joan made sure to thank Susan for all the help she had received.

The result is shown in Figure 3.4.1.

Through such a method, the qualified sentences are screened out. Although SPOST may not be able to identify certain special names correctly, it does not affect the quality of the selected sentences. Through such a method, the qualified sentences are screened out. Although SPOST may not be able to identify certain special names correctly, it does not affect the selected sentences. By manual recognition, in a random sampling of 200 sizes, 90.5% of the sentences are identified as solvable. There are finally eight groups that meet the criteria, each with 130 million sentences. Each set of data is awaiting processing as a formal training set.

In order to adjust the BERT language model and make BERT perform better in WSC problems, it is also necessary to use the test set to test the resulting BERT language model and adjust it. So, in addition to the training set, a test set is needed to adjust the BERT language model. The test set used in the FBC method is Definite Pronoun Resolution Dataset (DPR) (Rahman and Ng., 2012) [21]. According to Rahman and Ng, the WSC standard is relatively strict, so the DPR data set as a lowered-standard WSC is proposed. Kocijjan et al. used the test set in DPR as the FBC and removed the six sentences from the WSC data set to prevent overfitting, as the WSC data set would be used in later validation sets.

Finally, the FBC uses the WSC dataset and WNLI (Wang et al., 2019) training set, WNLI is one of the GLUE benchmark tasks [], WNLI dataset and WSC issues very similar. The WNLI data set consists of three sets of training sets, test machines, and verification sets, with a total of 852 sentences. The sentences in the test set are not duplicated with WSC. This set of 146 sentences is used for the FBC's validation set.

Before the FBC experiment, the data set needs to be converted into a usable model. According to the requirements accepted by Masked LM in the BERT language model, the pronouns in the sentence need to be replaced with "[MASK]" and the candidate is taken as part of the input. e.g.



**Sentence:** The trophy doesn't fit into the brown suitcase because [MASK] is too large.

**Candidates:** The trophy/ suitcase

BERT will calculate the likelihood of the candidate and choose the one with the most likely answer:  $C_0$  and  $C_1$  represent two candidates in a sentence  $S$ , respectively.

$$\mathbb{P}(C_0|S) \text{ and } \mathbb{P}(C_1|S)$$

The average of the possibilities is represented by  $\log \mathbb{P}(C_1|S)$  during training and testing. If it is the correct answer for  $C_0$ ,  $C_1$  is the wrong answer, and the loss equation is expressed as the following:

$$L = -\log \mathbb{P}(C_0|S) + \alpha \times \max(0, \log \mathbb{P}(C_1|S) - \log \mathbb{P}(C_0|S) + \beta)$$

Both  $\alpha$  and  $\beta$  are Hyperparameters.

### 3.4.2 Result and Limitation

For the KEE method, it cannot be tested because it is not an open-source programme, but the result was provided with the KEE method paper. Due to KEE method focused on the PDP problems, they used the PDP question set (60 questions in total) as the validation set. Using PDP question set is reasonable because PDP challenge is the first round of the WSC competition.

The correct rate of the model obtained through the CBTest training set is 65.0%, while the correct rate of the model obtained through the Wikipedia training set is 66.7%, which means that the difference in the numbers of correct answers from both two methods is only 1. As mentioned in Chapter 2, the alternative answers to PDP problems can be more than Two, so the accuracy of random selection will be less than 50%. As predicted, the result of random selection obtained in KEE paper is 45%. Therefore, if the difficulties of WSC and PDP are the same, KEE can be an effective method for solving WSC problems.

For the FBC method, there was an open-source programme on GitHub, and it was downloaded and tested. On the GitHub web page of the FBC method, it is provided that several instructions for testing the code, they are respectively for evaluation and training models. However, I cannot re-train the model using the code, because the performance of my computer cannot fully meet the requirements of running all the programme. Therefore, I

can only test the performance of the model they (\*) pre-trained and provided. The results are shown in Table 3.4.1:

	BERT	BERT_wiki	BERT_DPR	BERT_wiki_DPR
WSC	61.9%	61.9%	71.4%	72.5%
WNLI	65.8%	71.2%	71.9%	74.7%

Table 3.4.1 The results from FBC

The results in Table 3.4.1 are absolutely the same with that provided in FBC paper because the models I used and tested are absolutely the same with that used in FBC paper. In this table, BERT is the model provided by the Pytorch package; BERT\_wiki is the new model BERT retrained with the Wikimedia training set; BERT\_DPR is the model BERT fine-tuned with the DPR test set; BERT\_wiki\_DPR is the model BERT\_wiki fine-tuned with the DPR test set. It can be seen that BERT model in the FBC method does extract and utilise the commonsense, though it does not mention that.

On the one hand, for the result of the WSC validation set, it can be found that result of the BERT and BERT\_wiki are the same accuracy, and there is slight improvement between BERT\_DPR and BERT\_wiki\_DPR. However, for both BERT and BERT\_wiki model, after fine-tuned, the accuracy increased by about 10%. Therefore, the automatically collected Wikimedia dataset may not be an ideal training set for the WSC validation set. In contrast, the fine-tuning of the DPR test set has a greater improvement on the results of the WSC validation set. On the other hand, for the result of WNLI validation set, the performances of all models on all WNLI are better than that on WSC. Both fine-tuned DPR and Wikimedia set can improve the accuracy of BERT model, but DPR can only improve the accuracy of BERT\_wiki a little. From the information above, it can be found that the formats of Wikimedia dataset and WNLI set are not very similar to WSC schema, but DPR dataset is better. Generally, the performance BERT\_wiki\_DPR model is the best one to solve WSC problems, and the accuracy is the best of all method in this paper.

Although both methods performed better than other methods, there are still some areas for improvement in some parts of the algorithm. For KEE method, PDP problems were used as a validation set for evaluating the performance. Though the difference of the PDP problems and WSC problems are the same, WSC problems can be used to check whether a method

is too dependent on analysing the sentence structure. Therefore, if KEE method has been tested on WSC question set and make a comparison, the result will be more convincing. For FBC method, it used WSC and WNLI as a validation set, and it trained three models and used the BERT model as a control group to make a complete comparison. But from the result, we can find that the WSC-like dataset cannot represent the format of WSC because the accuracy almost stays the same after training model train with WSC-like dataset. Therefore, it is supposed to be improved that the poor performed WSC-like dataset.

## 3.5 Discussion

In the third chapter, four types of, a total of seven methods were introduced: FOL reasoning, CRR method, GSC method, and methods for WSC problems. From the results of all the methods tested, there are some facts that can be learned.

For the FOL reasoning, It requires humans to inform the computer in advance of the knowledge needed. For each WSC problem, depending on the complexity of the problem, it is also unequal that the amount of background information that needs to be inputted. If every time, solving problem, human input the knowledge background that the computer's required for reasoning, which is not an intelligent solution because the final goal is not solving WSC but understanding like a human. However, if all the knowledge is inputted in advance, even if we don't consider the storage space required by the computer and the time required to solve the problem, it will be a huge workload for human beings to prepare all the knowledge. Not only that, but it also requires human to help the computer identify the scenes in the sentence, such as the emotions involved in the sentence, the location of the event, and different cultures. So it is not a wise method that "teaching" computers and make computers reasoning to solve WSC problems.

For the CRR method, both AllenNLP and Stanford CoreNLP were tested, and the results can be quickly derived (a few seconds per schema). However, the accuracy is slightly higher than 40%, because, for both methods, they cannot recognise pronouns and candidates of about 15% questions. Even though the "not recognised" is not counted, the accuracy is slightly higher than the 50% of random selection. From the fact that the small number of "both correct" in the results and the "half-correct" of over 80%, it can be found that the CRR method mainly relies on sentence structure and semantic analysis to judge the answer of WSC. Therefore, a purely semantic approach like the CRR method is not able to understand, analyse, think, and answer the WSC questions with high accuracy like a human being.

For the GSC, it is a limited and low accuracy method. First, less than half of the WSCs can be tried to answer, because GSC can't be solved for someone with a name in the sentence. Secondly, for the problem that can be solved, due to the existence of complex sentences, the correct rate of GSC is still lower than 50% of the random selection. Moreover, this method cannot address the issue of ambiguous words and phrases, which will lead to inaccurate results from Google search.

For the methods using commonsense, these two methods are the best in this project. Although there are some areas that can be improved, they have a big breakthrough for both WSC and PDP issues. However, their training sets are not perfect enough.

It can be seen that to solve the WSC problem, it is necessary to solve the following conditions:

- 1) Need to introduce commonsense
- 2) Need to introduce semantic analysis, but not rely entirely on semantic analysis
- 3) Require good ability of entity recognition
- 4) Need to solve the problem of word ambiguity
- 5) Need to solve the introduction of commonsense by solving the training set

## **Chapter 4**

### **Conclusion and Evaluation**

For a better summary, this chapter will evaluate the entire project based on whether or not the aim of this project is completed. Then to summarise what I have learned from this project, the following part will be a personal reflection which obtained based on the experience. And the final part will be some perspectives on future work, because this time the project is not complete and perfect, it still needs to continue the experiment and research.

#### **4.1 Achievement**

To evaluate, this part will check whether the objectives of this project proposed in chapter 3 were achieved.

- i. This project has introduced the basic definition and history of the Turing Test and Winograd Schema Challenge, and concisely described the difference between them;
- ii. In chapter 3, there are four different types of, a total of seven methods for solving WSC problems has been found;
- iii. And in every section of chapter 3, the specific algorithms and principles of methods were introduced in detail. By thoroughly analysing the results of methods, the limitation of methods has been summarised respectively.
- iv. In the Discussion part, this project has analysed the principles and results of all the methods and found out what is the key point to solve WSC problems

In general, basically all tasks have been completed. In general, all tasks have been completed. All In the last two methods, only used the results given in the two papers for analysis, while the complete verification of the results was not achieved. However, it does not affect the fact that they are the best performers of all methods, nor does it affect the results of the final analysis.

## 4.2 Personal Reflection

Finally, through the learning and experience of this project, I have a deeper understanding of natural language processing and Winograd Schema challenge. I first learned that the WSC problem was in the Module "Knowledge Representation and Reasoning" (COMP5450M) taught by Dr Brandon Bennett in the first semester. And since that time, I was interested in the WSC problems, but I didn't think this is a tricky issue. Later, after learning the module "Data Mining and Text Analytics" (COMP5840M) taught by Dr Eric Atwell, I learned that natural language processing is not a simple question, and I understand the difficulty of WSC problems. After selecting this Project, I tried to solve the WSC problem with code. But under the guidance of my supervisor, Dr Brandon Bennett, I gradually found the difficulty of solving WSC. Regardless of the design of the algorithms, the implementation of the code, or the requirements of the computer hardware level, the WSC problem is very difficult to solve. So in the end, analysing existing artefacts is seen as the basis for solving WSC problems in the future.

Moreover, in the three months of completing the project, I also learned the ability to manage time. Although each subtask of this project can basically be completed within the planned time, in fact, this does not prove that my plan is reasonable, because some tasks are completed under overtime. When I made the plan, I ignored the extra time required for the experiment and overestimated my English writing speed. Therefore, when the actual execution of the plan, some tasks are completed over time. After the plan is made, I will extend a certain task time.

## 4.3 Future Work

### 4.3.1 Complete the FBC and KEE method

Due to being lack of the open source KEE code, I did not test the KEE with the WSC question set. So I will contact the authors of KEE method, and then finish the experiment and get results. For the FBC method, the open source code requires a computer with a GPU, I tried the code on my laptop which equipped with NVIDIA 1060 Ti (memory of 6 Gigabyte memory), but it shut down halfway and alter "out of memory". Therefore, I also used the result in the author's paper for analysis. If it is possible to have a computer with excellent equipment, I will try to train the model in person and test on both WSC and PDP question set in the future. Moreover, I will try KEE with another CSKB, Event2Mind, to replace the

CauseCom. They are very similar, but Event2Mind is the latest one. Event2Mind is not open for now, if they post the code, I will try as soon as possible.

### 4.3.2 Localisation

For later development, WSC problems need to be localised because current WSC problems may not be suitable for all language and cultural contexts. In order to solve these issues, it is necessary to create WSCs in different languages. For Chinese, there are some WSC issues that are directly translated, and you can get the available Chinese WSC, but some can't. For example:

Sentence: I couldn't put the pot on the shelf because it was too **tall / high**.

Answers: The pot/The shelf

If we translate Schema 73 and 74 into Chinese (Table 4.1 and 4.2, Figure 4.1 and 4.2):

I couldn't	Put the pot	On the shelf	Because	It is too	Tall
我不能	将锅放在	架子上	因为	它太	高（了）

Table 4.1 Schema 73

I couldn't	Put the pot	On the shelf	Because	It is too	High
我不能	将锅放在	架子上	因为	它太	高（了）

Table 4.1 Schema 74



From the result, we can see that “high” and “tall” in sentences are both translated to the same word “高”. Though in Chinese, the real meanings of “高” in both sentences are different, they expressed in the same word. Therefore, for this pair, there is no special or alternate word in Chinese, and they will be useless schema.

After finish the experiment of KEE and FBC, I will try to focus on the translate and filter available WSC questions in Chinese.



ENGLISH ↔ CHINESE (SIMPLIFIED)

I couldn't put the pot on the shelf because it was too tall.



我不能把锅放在架子上，因为它太高了。

Wǒ bùnéng bǎ guō fàng zài jiàzi shàng, yīnwèi tā tài gāole.

Figure 4.1 Schema 73 translate result from Google translate

ENGLISH ↔ CHINESE (SIMPLIFIED)

I couldn't put the pot on the shelf because it was too high.

我不能把锅放在架子上，因为它太高了。

Wǒ bùnéng bǎ guō fàng zài jiàzi shàng, yīnwèi tā tài gāole.

Figure 4.2 Schema 74 translate result from Google translate

## List of References

- 1 Levesque, H., Davis, E. and Morgenstern, L., 2012, May. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- 2 Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind* 59(236), pp. 433–460.
- 3 Saygin, A.P., Cicekli, I. and Akman, V., 2000. Turing test: 50 years later. *Minds and machines*, 10(4), pp.463-518.
- 4 Warwick, K., 2014. Turing Test success marks milestone in computing history. *University of Reading Press Release*, 8.
- 5 Hayes, P. and Ford, K., 1995, August. Turing test considered harmful. In *IJCAI (1)* (pp. 972-977).
- 6 <http://www.loebner.net/Prizetf/loebner-prize.html>
- 7 Morgenstern, L. and Ortiz, C., 2015, March. The winograd schema challenge: evaluating progress in commonsense reasoning. In *Twenty-Seventh IAAI Conference*.
- 8 2015, Eugene the Turing test-beating 'human computer' – in 'his' own words., <https://www.theguardian.com/technology/2014/jun/09/eugene-person-human-computer-robot-chat-turing-test>
- 9 Winograd, T. 1972. *Understanding Natural Language*. New York: Academic Press.
- 10 <http://commonsensereasoning.org/winograd.html>
- 11 <https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html>
- 12 Morgenstern, L., Davis, E. and Ortiz, C.L., 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1), pp.50-54.
- 13 Bova, N. and Rovatsos, M., Towards a Framework for Winograd Schemas Resolution.
- 14 W. McCune, "Prover9 and Mace4," 2005–2010. [Online]. Available: <http://www.cs.unm.edu/~mccune/prover9>

- 15 Bailey, D., Harrison, A.J., Lierler, Y., Lifschitz, V. and Michael, J., 2015, March. The winograd schema challenge and reasoning about correlation. In *2015 AAAI Spring Symposium Series*.
- 16 Sukthanker, R., Poria, S., Cambria, E. and Thirunavukarasu, R., 2018. Anaphora and Coreference Resolution: A Review. *arXiv preprint arXiv:1805.11824*.
- 17 <https://demo.allennlp.org/coreference-resolution>
- 18 Sayed, I., 2003, Issues in Anaphora Resolution
- 19 Lee, K., He, L., Lewis, M. and Zettlemoyer, L., 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.
- 20 Clark, K. and Manning, C.D., 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.
- 21 Rahman, A. and Ng, V., 2012, July. 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* (pp. 777-789). Association for Computational Linguistics.
- 22 Davis, E., 2015. A difference of a factor of 70,000 between hit counts and results returned in Google. *Unpublished technical note*.
- 23 Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 24 Liu, Q., Jiang, H., Ling, Z.H., Zhu, X., Wei, S. and Hu, Y., 2016. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in winograd schema challenge. *arXiv preprint arXiv:1611.04146*.
- 25 Miller, G.A., 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp.39-41.
- 26 Liu, Q., Jiang, H., Evdokimov, A., Ling, Z.H., Zhu, X., Wei, S. and Hu, Y., 2016. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.
- 27 Cheek to Cheek (album)., [https://en.wikipedia.org/wiki/Cheek\\_to\\_Cheek\\_\(album\)](https://en.wikipedia.org/wiki/Cheek_to_Cheek_(album)), August 29, 2011)
- 28 <https://dumps.wikimedia.org/enwiki/dump> id: enwiki-20181201

- 29 Kocijan, V., Cretu, A.M., Camburu, O.M., Yordanov, Y. and Lukasiewicz, T., 2019. A Surprisingly Robust Trick for Winograd Schema Challenge. *arXiv preprint arXiv:1905.06290*.
- 30 Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R., 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- 31 Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

# Appendix A

## External Materials

### A.1 Code and Online Tools

AllenNLP package: <https://github.com/allenai/allennlp>

AllenNLP online Demo :<https://demo.allennlp.org/coreference-resolution>

Stanford CoreNLP package: <https://stanfordnlp.github.io/CoreNLP/download.html>

Stanford CoreNLP online Demo<http://corenlp.run/>

CBTest set: <https://www.gutenberg.org/>

Pytorch package: ([https://pytorch.org/hub/huggingface\\_pytorch-pretrained-bert\\_bert/](https://pytorch.org/hub/huggingface_pytorch-pretrained-bert_bert/))

BERT package in Pytorch: [https://pytorch.org/hub/huggingface\\_pytorch-pretrained-bert\\_bert/EW](https://pytorch.org/hub/huggingface_pytorch-pretrained-bert_bert/EW)

Wikimedia: <https://dumps.wikimedia.org/enwiki/>

WNLI data set: <https://gluebenchmark.com/tasks/>

ConceptNet: <http://conceptnet.io/>

WordNet: <http://wordnetweb.princeton.edu/perl/webwn>

SPOST: <https://nlp.stanford.edu/static/software/tagger.shtml>

PDP dataset:  
<https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/PDPChallenge2016.xml>

FBC method: <https://github.com/vid-koci/bert-commonsense>

## **Appendix B**

### **Ethical Issues**

During the FBC experiment, I have tried to test the results of my experiments on my laptop. However, due to the poor performance of your own computer, it is impossible to support the completion of the experiment. The school's virtual machine also cannot support. After that, I turned to my friend for help, although his computer could not train a new language model, but it was enough to get the key result of the pre-trained model.