ELEC 896 report

Semeval 2020 Task 5 Modelling Causal Reasoning in Language: Detecting Counterfactuals

by

Jiaqi Li

A thesis submitted to the
School of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Engineering

Queen's University
Kingston, Ontario, Canada
August 2019

Abstract

In this report, we build a new dataset for natural language inference based on counterfactual. Counterfactual is a concept that can show logical relationship and causal reasoning, which can be helpful to find whether a model gives out results based on statistics or based on the meaning between words. Also, it determines to let the computer tell why it makes a choice like this.

Thus, we have organized this task with two parts. The first one is just judging whether a sentence is counterfactual or not, while the other one is finding the antecedent and consequent part of the counterfactual sentences.

To build up such dataset that can show causal reasoning, we first crawled news in specific domains for later use. And then, with the help of crowdsource platform, these selected sentences will be labelled following our rules. Finally, a dataset is built after we carefully select positive examples and negative examples in the results we have. Some baseline models have been tested and it shows that the dataset can be good for natural language inference and it still has some room to improve the results.

Acknowledgments

Thanks for the supervise and opportunity provided by Prof. Zhu. He gave me the chance to explore this interesting topic and many in-time helps can't be ignored to finish this thesis. Also, thanks for the patience from MASC student Xiaoyu Yang, who is the leader of this project. Everytime we have problems she will help to solve the problems carefully and give out some interesting ideas. Besides, thanks for the contribution of other members in our team from Stephen Obadinma, Qianyu Zhang and Rohan Bhambhoria. Without your hard working, the task wouldn't be carried out. The last, thanks for the accompany of my girlfriend Liyi Xue, who provides me a lot of caring in life which can be a part never be ignored.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The topics of natural language processing (NLP) have gained more popularity recently, as an intelligent system or the artificial intelligence can't be intelligent without it. Thus, as one of the part that will never be ignored, a lot of efforts [2, 9, 12, 13] have been made in natural language inference with semantics analysis or common sense. However, we focus on modelling causal reasoning in this task.

Causal reasoning, together with semantics knowledge or just common sense, can show the inference, which we hope the model to learn, in different directions. On the other hand, causal reasoning is the most basic and important one as it can show the logical relationship just in the sentences provided. While causal reasoning is a wide problem, as mentioned in the work [7], counterfactual is the top-level expression that human show causal relationship. Also, in different areas, counterfactual is regarded as an interesting and important part to understand human's reaction to the world [3, 6, 8, 11]. In NLP, it is a tendency to let the model tell us why it makes the decision instead of just observing many corpuses and giving out answers based on statistics

results. Therefore, counterfactual is a choice as it can not only show the way how human react to a problem but also contain the antecedents and consequents, which means models can explain their results or even show the causal relationship implied in the text. The true inference in natural language should have the models infer by themselves based on the corpus but not only find the similarity between sentence and sentence and give a result.

For the reason of this, we release this task in natural language inference on SemEval 2020 by counterfactual detection. The task includes two parts. The first part, subtask 1, participants need to build a model to figure out whether a sentence is counterfactual or not, where the model is expected to show an ability in understanding what counterfactual is and the differences between counterfactual sentences and non-counterfactual sentences. While, the second part of the task, subtask 2, needs the built model to select antecedents and consequents from the counterfactual sentences given, where some inferences are supposed to be made based on the corpus and some basic relationships implied in the sentence are believed to be found by the model.

1.2 Problem

As counterfactual shows the basic logical relationship in the sentence with the antecedent and consequent, it is worth studying to bring some progress in natural language inference. Our work is inspired and based on work [10], and we found that counterfactual is simple in definition but difficult in recognition. In other word, counterfactual is something easy to understand while hard to be identified. Another problem is that counterfactual actually doesn't appear frequently in normal conversations or articles which brings more difficulty to our collection.

To make it clear that what is counterfactual, we will have several examples in the following part. The most common one is shown below:

- If I hadn't broken my arm, we would have won.

Counterfactual should have two parts, one being counter and another being fact. It needs to describe something that is counter to the fact [5, 7]. Thus, the only way to figure out whether a sentence is counterfactual or not is to find if there is something that contradicts the fact. So, the key point is the fact and we just need to make out if the assumption and the fact are the same or different to find the counterfactual sentences. Usually, the facts include four aspects: events that have already happened, facts about the present, objective truths and common sense. In the example above, because *I* did break my arm and the assumption in the sentence is a contradiction to the fact that already happened, it is a counterfactual and from it we can indicate that my broken arm leads to our failure. In this degree, counterfactual is easy to be made out from other sentences as the rule is simple, but when we get deeper, it is another story.

As mentioned in [4], counterfactual is the problem between antecedent and consequent. Due to different kinds of antecedent, counterfactuals are divided into three kinds: counteridenticals, countercomparatives and counterlegals.

- If I were Julius Caesar, I wouldn't be alive in the twentieth century.
  (**counteridenticals**)

- If I had more money,...
  (**countercomparatices**)

- If triangles were squares,...

  (**counterlegals**)

The above examples agree with the concept we gave out about the fact. But further in this work, it mentioned that counterfactual isn't only the simple contradictions between assumption and fact. The following examples will show what the problem is.

- If match M had been scratched, it would have lighted.
- If match M had been scratched, it would not have been dry.

It is obvious that these two sentences all meet the need to be counterfactual, but the first one is affirmed while the second one is denied. The reason is that in the second sentence, the antecedent is not "cotenable" with the consequent. That is to say, the antecedent and the consequent are not related to each other. Therefore, to be counterfactual, the consequent should not only be compatible with the antecedent, but also "joint cotenable" or "cotenable" with the antecedent. However, it leads to a chicken-egg problem because if we want to make sure whether it is a counterfactual or not, we need to find a consequent which is cotenable with antecedent. But if we want to find whether the consequent is cotenable with antecedent, we need to determine whether the counterfactual "If antecedent were true, then consequent would not be true" is true itself. That means you need to find another consequent that is cotenable with antecedent that leads to anti-consequent. Another problem is that language is changing all the time based on human usage, which means human decide how to use languages but not languages decide how human use languages. Maybe someone thinks it is a counterfactual in his direction but in another's point it is not.

In another work [1], it proposed another interesting problem of counterfactual. Let's see an example first:

- If Jones had taken arsenic, he would have shown just exactly those symptoms.
- If Jones had taken arsenic, he would have shown just exactly those symptoms, which he does in fact show.

These two sentences are similar with each other but the first one is counterfactual while the second one is not. It is a reminder that a contradiction between assumption and fact determines counterfactual. In the first sentence, it is easy to understand why it is counterfactual but in the second one, just with an addition of a clause that makes sure the assumption, it transfers into a non-counterfactual. In fact, the border between counterfactual and non-counterfactual is so weak and only a clause can change its meaning. Although this kind of sentences can be tricky and hard to find, it can be super useful to find whether a model makes decisions based on inference or based on statistics results.

Therefore, counterfactual can be hard to select due to the relationship between antecedent and consequent and the weak border. Also, the low appearance in normal text adds the difficulty. To solve the problems, we crawled a large number of articles online and used some specific rules to filter the sentences. After we got the sentences, they were put onto the crowdsource platform, Amazon Mechanical Turk, to be labelled by the workers. Then we got the vote results for each of the sentences. From all the results we got, a balanced dataset was created.

1.3 Organization of Thesis

In the following Chapter 2, we will get into details about the dataset, talking about the corpus, the way we used to get the sentences and how we get the labelled dataset. And then some baseline models for this task will be presented in the Chapter 3. Finally, in the Chapter 4 there will be a summary about the whole work.

Chapter 2

Corpus and Dataset

2.1 Corpus

As the appearance of counterfactual is rare in the normal scenario of conversations and articles, hundreds of thousands of articles are crawled online and specific domains have been chosen to better show relationship between antecedents and consequents. Later, some rules have been used to filter out sentences that are possible to be counterfactual. Finally, these selected sentences are ready to be annotated on the crowdsource platform.

In the pioneer work [10], they did collect many of counterfactuals, but there are so many problems about their dataset. Firstly, they used twitter data to find counterfactual and as tweets are so oral and there are often some sentences that don't make sense or even with no meaning, the dataset won't be a great help in improving the performance of inference model. Also, some symbols without meaning appear usually which can only be noises to the results. Besides, the dataset is quite small, with only 500 samples for training, 1266 samples for supplement and 1137 samples for testing. In total, there are only about 3000 sentences available. Finally, the distribution of the

data isn't balanced, as the rare appearance of counterfactual, there are only about 10% counterfactual in the dataset while the others are non-counterfactual.

Table 2.1: Dataset Details in Work [10]

Dataset	**Counterfactual**	**Non-Counterfactual**	**Total**
Train	49	451	500
+ Supplement	768	498	1266
Test	104	1033	1137

Therefore, based on [10] and with improvements, our dataset is far more powerful than it. To gain more useful and less noisy sentences, we choose to focus on three popular domains that can show counterfactual better: finance, health and politics. All of the domains use formal writing language and assumptions often appear in these circumstances as well as useful information can be derived from these domains. Thus, we crawled more 900,000 articles in all the domains and details are showed in Appendix A Table A.1 on page 28. Similarly, derived from [10], all the selected articles are processed by a rule to find specific sentences that are potential to be counterfactual. The rule consists of 2 parts, one being pattern while another being POS tag. For the pattern part, we have 15 different patterns, which are shown in Appendix A Table A.2 on page 29. They are summarized from many ways, some from subjunctive which agrees with the results in [1] while others from the sentences we came across when collecting data. But not all the sentences got by the patterns are all counterfactual, they are just candidates for counterfactual labelling later. After the filter of the patterns, we find most of sentences are not counterfactual because some of the patterns are not strict enough such as "If.... then..." while some other patterns

perform so well, namely the inversion of "were" and "had". Then, a second filter, the POS tag mentioned above, is introduced. For the reason that normally subjunctive conditional can show counterfactual, we hope to extract the structure of the good representation of the conditional sentences. After selection, several structures are chosen as the POS tag test in Appendix A Table A.3 on page 29. Here, we use the basic toolbox, Natural Language Toolkit to parse the sentences we have now to get the POS tags, and then only those sentences that pass the POS tests will be kept. The sentences left are ready for the manually annotation on the crowdsource platform. The statistics details are shown in Table 2.2.

Table 2.2: Stactistics for Selected Sentences

	Finance	**Health**	**Politics**
article	543,541	208,251	171,258
sentence	141,254	90,197	34,160

We plan to collect about 15,000 counterfactuals and 15,000 non-counterfactuals from the candidate sentences, with an even distribution in three domains. Therefore, comparing with [10], our work is about 10 times larger, and more balanced, which can be more representative and useful as a dataset. Besides, based on the rules in [10], we proposed a better version with the combination of patterns and POS tags, which improves the counterfactual rate from about 9% to 17 - 30%.

2.2 Dataset

2.2.1 Worker Selection

After we got the candidate sentences for labelling, we didn't put it online at once. As the detection for the counterfactual is not so easy as some other tasks like figuring out how many furniture in the picture or choosing who is on the bike, and it is unfamiliar for most people that what is counterfactual. We need to make sure that the workers understand what counterfactual is and how to find counterfactual. Thus, the challenge is how to teach the workers properly and efficiently as on crowdsource platform, time is money and no worker will spend much time on a super difficult task. Therefore, the problem is how to select qualified workers in a quick and convenient way.

Here, we use a pre-test to select skillful worker. At first on Amazon Mechanical Turk, we published a simple test about counterfactual including a clarification on what is this test for, a definition on what is counterfactual, some tips on how to figure out counterfactual, several examples about counterfactual and non-counterfactual and the reasons why it is or not, and the last part is a 22-sentence test. In the clarification part, we declared that this is a pre-test of the coming work and if workers' performances are good, they will be invited into a group for the further work. Then following is the definition part, where worker can learn the concept on what is counterfactual with one basic example. Following are some tips that we conclude and induct from the definition of the counterfactual and the experience when we seek for counterfactual. After that, there are several interesting examples covering all kinds of facts that can be counter to and in the negative example part, there are some tricky examples. The example part is shown like below:

Positive Examples:

- I wouldn't have done that **if I were you**. (We all know that I can't be you, the assumption contradicts **Facts About the Present**.)
- They are discussing what our clothes would say **if they could talk**. (We all know that the clothes can't speak, the assumption contradicts **Common Sense**.)
- Would the prisoner be alive **if rifleman A had not shot**? (It's a counterfactual question. It can be inferred that rifleman A had shot, and as a result the question is asking what would happen if rifleman A had not shot. Thus, it is an assumption that contradicts **Events that Have Already Happened**.)

Negative Examples:

- "As far as the market is concerned, **you couldn't have scripted it any better**," said JJ Kinahan, chief market strategist at TD Ameritrade. (There is **no assumption here that is counter to a fact**. Instead, it is just making a comment on market behavior that has transpired. Hence you cannot say that it is counterfactual.)
- "**If Jones had taken arsenic**, he would have shown just exactly those symptoms which he does in fact show." Rouhani said. (We can find that Jones does show the symptoms of taking arsenic and so it isn't counter to the sentence" If Jones had taken arsenic", which means there is **no assumption here that is counter to a fact**.)

The last part is the test with 22 sentences. Each of the sentence is gathered online in the three domains. The workers are then asked to decide whether the sentence

is counterfactual or non-counterfactual. To check whether the workers are carefully going through the tutorial part, 2 of 22 sentences are from the example parts. These two sentences help to filter out unqualified workers in a more efficient way.

To evaluate the performance of workers, we set a threshold that among 22 sentences, those who make mistakes no more than 5 times are included in the later work group. However, if a worker fails in a sentence that appears in the example part, he will be directly removed from the group. Then, notification emails are sent to inform them that they are invited in the group. Also, the answers of the test and the analysis on the sentences in the test are attached in the email. Two days are given to workers to review the test and then are ready to the formal work.

2.2.2 Subtask 1: Counterfactual Recognition

In this section, we will talk about the first task, counterfactual recognition. In the previous chapter, counterfactual is known as an important way to realize the language inference and from the definition of counterfactual, it can be learnt that counterfactual describes something that is counter to the fact, which can show the relationship between antecedent and consequent. Thus, the first task is set to let the model detect the counterfactual from the candidate sentences. As the appearance of counterfactual is low, although we have used some patterns and rules to filter the sentences, there are still many sentences that are not counterfactuals and it can be hard and tricky for the model to learn. So, to get a better result, the model must know the differences between the counterfactual sentences and non-counterfactual sentences as the statistics features can be similar in these two kinds. We hope the first task can help to tackle the problems that the model giving results relies more on the statistics

patterns. Also, solving this problem is the foundation for all other counterfactual related causal inference analysis in natural language. Therefore, we believe this task can be important.

As discussed above, counterfactual isn't an obvious thing for human to detect. Many reasons can affect the results given. During our labelling work, workers are found to have about 75% accuracy in labelling counterfactual. That is to say that even several workers vote for the result, it can still be an error. To deal with this problem, we combine the subtask 1 labelling together with the subtask 2 labelling because these tasks are related to each other. As the sentences are labelled in subtask 1, it will come to subtask 2 for workers to have a re-labelling. Then combining these two parts together, we can filter out some counterfactual sentences labelled wrongly. The details about subtask 2 will be presented in the following section. Here, for the first part, five workers vote together for the results. The potential counterfactual sentences are filtered by the patterns and POS tags. Then, in theory, the accuracy rate about the vote result can be about 89%. And these voted counterfactuals will be ready for the subtask 2. After subtask 2, the results are sent back to subtask 1 to correct results that aren't correctly labelled. That can make sure the correctness of the labelled results. In the labelling of sentences, there are some different degrees. All the results that agree to each other are strong agreement while if there is one different answer, it is medium agreement. The rest are border line, also known as weak agreement. In the dataset, we only include those counterfactual sentences that are strong and medium agreement to make sure the accuracy of the dataset. All the workers are presented tasks with 20 sentences the same with the test. Each sentence is following the sample below:

- **Sentence 1**

  Had Russia possessed such warships in 2008, boasted its naval chief, Admiral Vladimir Vysotsky, it would have won its war against Georgia in 40 minutes instead of 26 hours.

- It is a counterfactual.

- It isn't a counterfactual.

Comparing with the results in [10], our results are more reliable as the results are voted by several workers instead of one. Besides, our sentences are more formal and informative than the data collected from twitter which are mostly oral English with many meaningless symbols. That also means there are less noises and more things that are worth being mined. The sample results are displayed in Table 2.3 on the next page.

2.2.3 Subtask 2: Detecting Antecedent and Consequent

In this section, we will talk about subtask 2. For this task, the motivation is to let computer find the antecedent and consequent from the counterfactual sentences. It is important and interesting because making the computer find the antecedent and consequent of the counterfactual is, to some degree, to let the computer tell us why they make this choice. As the logical relationship between antecedent and consequent decides why it is a counterfactual and what does this sentence imply. For example, if a sentence is given like: *Her post-traumatic stress could have been avoided if a combination of paroxetine and exposure therapy has been prescribed two months earlier*, we can learn that the antecedent part is *if a combination of paroxetine*

Table 2.3: Subtask 1 Sample Examples

Sentence	Domain	L1	L2	L3	L4	L5	gold_label
If Michael Vascitelli ($64m) hadn't been running Vornado Realty Trust, somebody else would have.	Politics	1	1	1	1	1	1
If they were rational, they would work hardest on the good days (to maximise their take) but give up early when fares are few and far between.	Finance	1	1	0	0	0	0
For one thing, Democrats would have probably paid a much smaller political price if their effort wasn't billed as an extravagant government grab to take over the nation's health care system.	Health	1	1	1	0	1	1

and exposure therapy has been prescribed two months earlier and the consequent part is *Her post-traumatic stress could have been avoided*. More in-depth, we can find that this counterfactual means the combination of paroxetine and exposure therapy haven't been taken and her post-traumatic stress still exists. We believe that only when a computer learns these information can we say the model is smart enough to do some inferences. Therefore, this task can be a little more difficult than the previous task while it can tell us why the computer gives out the answers and what the basis is to make the judgement.

After the sentences being processed by the workers in subtask 1, the selected sentences that are counterfactual with strong and medium agreement are here ready for a re-annotation for the subtask 2. In this part, the workers are asked to find the antecedent part and consequent part. There is no pre-test in this part but we still use the workers who pass the previous test. Then, a brief instruction on what is antecedent and consequent is given at the beginning of this task. As not all the sentences in subtask 2 are counterfactual, but most of them are, we provide an option that the workers can choose if the candidate sentence isn't a counterfactual. In the annotation, sentences may have both antecedent and consequent or just the antecedent and if the workers think it is a counterfactual, the antecedent part must be given as the antecedent is the key part that makes a sentence into a counterfactual. Thus, the results can be varied from sentences to sentences. After that, the results received from online platform are checked manually. More attention is paid on those sentences that the workers think aren't counterfactual. With a double-check on them, some of those can be corrected into counterfactual, namely finding out the antecedent and consequent, while others are left unchanged. Finally, the sentences for subtask 2 are ready and only those sentences that are regarded as counterfactual that will be included in the dataset. Moreover, from the results of re-labelling of the sentences selected by subtask 1 in subtask 2, some corrections are introduced in the subtask 1 dataset. Those sentences which are regarded as counterfactual in subtask 1 but refused in the subtask 2 will be changed into a borderline non-counterfactual example in the dataset. In all, for each task we all have at least labelled twice and the accuracy of the sentences in the dataset can be valid.

This task can be unique as usually the counterfactual dataset won't get into such

Table 2.4: Subtask 2 Sample Examples

Sentence	Domain	Antecedent	Consequent
But my view is that we should have focused on China.	Politics	we should have focused on China	**N/A**
Had the president-elect made a swift promise to pursue his protectionist agenda, however, bond yields might have fallen, since a tariff war would hurt economic growth.	Finance	Had the president -elect made a swift promise to pursue his protectionist agenda	bond yields might have fallen
"These children, who would have died very young without treatment, are participating in life as fully as their brothers and sisters," researcher H. Bobby Gaspar, MD, PhD, tells WebMD."	Health	without treatment	These children, who would have died very young

detail about the structure of counterfactual and the logic relationship inside. Besides, these two tasks are comprehensive for the basic understanding for counterfactual, which is a typical kind in natural language inference. In subtask 1, we want the computer to learn the counterfactual's features and differences between counterfactual and normal sentences while in the subtask 2, the understanding in the reason why it is counterfactual and what the relationship between antecedent and consequent are needed. The sample examples are shown in Table 2.4

2.2.4 Rewritten Counterfactual

In this section, we will talk about rewritten counterfactual. As counterfactual sentences are hard to detect, and most of those found are simple in logical relationship, we decide to produce some tricky counterfactual manually to add the diversity of the dataset. Several benefits can be brought by rewritten sentences. First is that the appearance of counterfactual is rare, so rewritten can improve much efficiency. Besides, in the tasks that worker finished, the tricky counterfactual examples might be missed as they aren't obvious and to make up for this part, some rewritten sentences can be added in to make the whole dataset stronger. Finally, counterfactual like *If Jones had taken arsenic, he would have shown just exactly those symptoms which he does in fact show." Rouhani said.* can be super useful examples for the dataset and it can be a good test for whether the model understand the concept. Thus, here we have the third part of the dataset, the rewritten sentences part and this part will be processed and added to the previous two subtasks. All the sentences are in the three domains and these sentences are corrected from the sentences that are similar with counterfactual but not counterfactual.

For the rewritten sentences, they are from the sentences left out when we select the candidate sentences. Those which have passed the pattern filter but are filtered out by POS tags are the material for rewriting. Also, we keep a balanced distribution in three domains and the amount is small. When rewriting, some specific rules should be obeyed such as pattern word must be kept or *if* should be removed. After finishing the rewriting, they are further ready for the antecedent and consequent labelling. All the labels of rewritten sentences in the dataset are N/A. Below you can find some examples of rewritten sentences in Table 2.5 on the next page. The last step is to give

Table 2.5: Rewritten Sentences Sample Examples

Original Sentence	Domain	Modified Sentence
So if someone wants to seek care from a practitioner or a discipline that may work for them, they should have the freedom to do so.	Politics	So if someone wants to seek care from a practitioner or a discipline that may work for them moving forward after a bad experience, they should have possessed the freedom to do so.
If quotas are filled first-come , first-served, then exporters rush to get in quickly, and imports surge.	Finance	If not for the quotas being filled first-come, first-served, the exporters would not have had a chance to get in quickly.
"He doesn't know if she had registered to be a donor, or if her family decided to donate her organs."	Health	"There's no record of her registering to be a donor before she died, yet assuming that she decided to do so while she was alive, someone being saved as a result was a real possibility."

each sentence a unique sentence id. In subtask1, the id of sentences in the domain of finance, health and politics start with 4, 5 and 6 respectively. In the same way, the sentences in subtask 2 start with 7, 8 and 9 respectively for finance, health and politics domain.

Chapter 3

Baseline Model

3.1 Subtask 1 Baseline Model

In this section, we test our dataset with different baseline models. For the subtask 1, recognizing counterfactual, we follow the previous work [10], using SVM to classify the sentences based on the frequency of key words in the sentences. First, we do some preprocessing on the sentences, such as making lowercase and tokenizing. Then the whole dataset is divided into training part and testing part in a ratio of 80% to 20%. After that, word frequency is chosen as features and SVM is used to do classification on the training part and test part. The results are show in Table 3.1 below.

Table 3.1: SVM Confusion Matrix

	Precision	**Recall**	**F1-score**
Counterfactual	72.30%	71.33%	71.81%
Non-counterfactual	72.96%	73.89%	73.42%
SVM Accuracy Score		72.64%	

The results shown are just based on the dataset we have now as we haven't finished

all the dataset and also the dataset haven't been smoothed. Thus, we believe the performance of the baseline results can be more challenging for the participants of this task.

3.2 Subtask 2 Baseline Model

Also, similar to the previous task, here we just use a heuristic method as the baseline model. The sequence labelling model based on bag-of-words features is introduced. This baseline model annotates the antecedents and consequents using the B/I/O scheme, which can be similar to name entity recognition, determining if a word is at the Beginning, Inside, or Outside of the antecedent and consequent part. Specifically, we tag each token in the sentence with **B-Ant**, **I-Ant**, **B-Con**, **I-Con**, or **O**. The preprocessing part can be the same, making word lowercase and tokenizing. Then different labels are attached to the tokens. The dataset is divided into training part and test by 80% to 20% too and there are 200 epochs for the training. After the model is built, the test part will be process with the model trained. The following part shows the results in Table 3.2.

Table 3.2: Sequence Labelling Confusion Matrix

	Precision	**Recall**	**F1-score**
I-Con	62.78%	64.06%	63.41%
B-Con	69.46%	61.54%	65.26%
I-Ant	60.98%	61.16%	61.07%
B-Ant	56.89%	50.00%	53.22%
O	57.26%	56.92%	57.08%

This part is also run on the dataset we have now, and after the dataset is finished,

the task can be harder for computer than it is now. Besides, subtask 2 is based on subtask 1 but more difficult. We hope it will be more useful and interesting than subtask 1.

Chapter 4

Summary and Conclusions

4.1 Summary

In total, we build a dataset ourselves for natural language inference in counterfactual. In this dataset, the sample sentences are first selected from articles crawled from websites and then some specific patterns and POS tags are used to filter out potential counterfactual. Next some qualified workers are ready to label these processed sentences. In the process of these sentences, each of them for each task will be annotated at least twice to ensure the accuracy. Besides, we rewrite some sentences which are similar with candidate counterfactual sentences and make them into positive examples following some rules. After that, positive examples and negative examples are selected carefully from the labelled results. Comparing with the pilot work [10], our dataset is much larger and more representative as we have about 30,000 sentences found in the authoritative news websites instead of some twitter information. The most important thing is each sentence from our dataset has been labelled at least twice which guarantees the candidates in dataset are counterfactual.

4.2 Future Work

This task is just a beginning of counterfactual recognition. It just involves the detection of basic structure of counterfactual. As counterfactual can show much information and logical relationship, namely causal reasoning, there are still many useful aspects that are ready to be mined. For example, besides the tasks mentioned above, when presented with a counterfactual sentence, you can judge what direction is the assumption towards, such as whether this counterfactual is "upward" or "downward". Also, based on subtask 2, the antecedent and the consequent can give us a lot of useful information like from a example above *Her post-traumatic stress could have been avoided if a combination of paroxetine and exposure therapy has been prescribed two months earlier.*, we can induct that a combination of paroxetine and exposure therapy can be useful to her post-traumatic stress. For this reason, in the domain of counterfactual, there are still many things waiting for us to explore.

4.3 Conclusion

In this report, a more powerful dataset built by us is introduced to provide another approach to natural language inference by causal reasoning. We hope this dataset can be helpful in improving the inference of the computer. Besides, we hope it can be a foundation for the future more powerful dataset and model.

Bibliography

[1] Alan Ross Anderson. A note on subjunctive and counterfactual conditionals. *Analysis*, 12(2):35–38, 1951.

[2] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[3] Anneke Buffone, Shira Gabriel, and Michael Poulin. There but for the grace of god: Counterfactuals influence religious belief and images of the divine. *Social Psychological and Personality Science*, 7(3):256–263, 2016.

[4] Nelson Goodman. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128, 1947.

[5] Lauri Karttunen. Counterfactual conditionals. *Linguistic Inquiry*, 2(4):566–569, 1971.

[6] Laura J Kray, Linda G George, Katie A Liljenquist, Adam D Galinsky, Philip E Tetlock, and Neal J Roese. From what might have been to what must have been: Counterfactual thinking creates meaning. *Journal of personality and social psychology*, 98(1):106, 2010.

[7] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[8] Lawrence J Sanna and Kandi Jo Turley. Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin*, 22(9):906–919, 1996.

[9] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019.

[10] Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658, 2017.

[11] Jan P Vandenbroucke, Alex Broadbent, and Neil Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International journal of epidemiology*, 45(6):1776–1786, 2016.

[12] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[13] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.

Appendix A

Corpus Parsing Information

A.1 Websites for Article Crawling

Table A.1: Websites for Article Crawling

Finance	**Politcs**	
Investing	Reuters	breitbart
BusinessInsider	BusinessInsider	Daily KOS
Cnbc	ThisIsInsider	
The Economist	The Economist	
Financial Times	NYtimes	
Reuters	washington	
Motley Fool	Politico	
Health		
MNT(MedicalNewsToday)	WebMD	NYtimes
Healthline	ftimes	medcity
healthcareit	mobihealth	mayoclinic
kaiser	medscape	medical daily
healthcareblog	cfhj	betterhealth
healthcare_economist	healthbusinessgroup	nih
psychology today	psychCentral	ScientificAmerican

A.2 Patterns

Table A.2: Pattern List

Index	Pattern
Pattern #1	If...then...
Pattern #2	If..had/hadn't...
Pattern #3	could/may/might/should/would/wouldn't/couldn't/shouldn't have/haven't
Pattern #4	What if...
Pattern #5	Even if...
Pattern #6	If I/there/he/she/you were...
Pattern #7	Wish...could/may/might/would/should/wouldn't/couldn't/shouldn't have/haven't...
Pattern #8	Wish...were/weren't/had/had't...
Pattern #9	Wish...
Pattern #10	But for...could/might/would/should/wouldn't/couldn't/shouldn't have/haven't...
Pattern #11	If only...
Pattern #12	Had/Were...
Pattern #13	If...
Pattern #14	...could/might/would/should/wouldn't/couldn't/shouldn't have/haven't...without
Pattern #15	...could/might/would/should/wouldn't/couldn't/shouldn't have/haven't...had/were I...

A.3 POS Tag

Table A.3: POS Tag Rules

Index	Structure
POS Tag #1	If+VBD/MD/VBN+MD
POS Tag #2	MD+VB/VBN+if+MD/VBD/VBN
POS Tag #3	MD+VBN+MD
POS Tag #4	wish+VBD/VBN
POS Tag #5a	Had/had/were/Were+VBD/VBN+MD+VB/VBP/VBD/VBZ
POS Tag #5b	MD+VB/VBP/VBD/VBZ+Had/had/were/Were+VBD/VBN