Supervised Project - Semester Report

Having a share in SemEval-2020 Task 5:
Detecting Counterfactuals

Farnaz Ghassemi, Tuan Anh Vo,
Léo Jacqmin

A project supervised by:
Pierre Ludmann and Philippe de Groote
Loria - Inria

M1 Natural Language Processing
IDMC - Université de Lorraine
2019-2020

1 Introduction

How can we represent knowledge about the world in a way that computers can understand it? Knowledge representation and reasoning are some of the key concepts in the field of artificial intelligence. To solve this problem would mean giving computers the ability to perform complex tasks such as diagnosing a medical condition or having a conversation in a natural language. Automatic reasoning is an area of cognitive science dedicated to giving computers the ability to reason. The study of knowledge representation is closely related to automatic reasoning since the underlying goal of representing knowledge is to reason, to make inferences about that knowledge in order to build new one.

Along this line of thinking, shared task organizers at SemEval-2020 published a series of shared tasks that include three tasks dealing with knowledge and reasoning. This project, conducted under the supervision of Pierre Ludmann and Philippe de Groote, aims at tackling one of these tasks, i.e. Task 5: Modelling Causal Reasoning in Language: Detecting Counterfactuals.

Counterfactual thinking is one of the highest level of reasoning. It is the basis of all scientific thought. When conducting an experiment, a scientist asks himself: "What would have occurred if things had been different?" Using this reasoning allows him to infer causality, X implies Y. But so far, machines do not share our intuition about cause and effect. They reason by association from large sets of data. Many consider causal reasoning to be the missing piece needed to build truly intelligent systems.[3]

This shared task is divided into two subtasks: detecting counterfactual sentences and detecting antecedents and consequents in counterfactual sentences. Tackling the first subtask serves as the basis for further analysis of causal inference in natural language. The second subtask's goal is to help model this causal knowledge. Labeled datasets are available for participants to solve both subtasks.

The main problem is, counterfactuals are difficult to define and there is no clear consensus about what makes a counterfactual. Even a supposedly simpler aspect that is its syntactic structure has been still very controversial and debatable. It is partly because of the definition of counterfactual itself, partly because of the divergence of different languages when expressing this universal cognitive process of us. Take this sentence as an example:

(1) "If Mr. Trump nominates a member of Senate Democrats as a candidate, Senate Democrats will definitely deny the pledge they made before."

Any of us who is informed of American politics knows that Mr. Trump, a Republican, would never nominate any Democrat. We are dealing with a classic counterfactual sentence, but its syntax tells a different story. Here, the indicative mood, which normally points to a fact, is used to express an hypothetical situation. This poses a real problem for us to recognize the causal relationship of counterfactuals, with all of its ambiguities, using a computational approach.

The core puzzle of counterfactuals, as seen above, is their conceptual appa-

ratus, which remains peculiar to human intelligence. The procedure is always up to human to determine better grammatical models, before finding the most suitable mathematical interpretation so that the machine can learn and detect counterfactuals. Because of the inherent complexity regarding the definition of counterfactuals, our approach aims at describing what are the building blocks of a counterfactual by analysing data. We use a bottom-up approach with regard to how one can deconstruct language. Starting from a lexical analysis, we make our way up the main components of language to end with semantics and pragmatics. While many would consider using a machine learning model to solve this task, our analysis of counterfactual sentences has led us to believe that this symbolic approach could potentially be more effective. As we have observed during our lexical and syntactic analysis, counterfactual sentences follow common patterns that can be defined.

Counterfactuals are used to express how a hypothetical statement could have changed the present situation. In English, expressing such states of unreality is done through the use of the of modal verbs and keywords.

(2) "I was thinking that I might write a book if Hillary Clinton had won."

Thus our first approach consisted in analysing the lexicon of counterfactual sentences. We identified the most common keywords and modal verbs used in counterfactuals by counting the occurences in the dataset. We then established a statistical significance for the usage of these keywords in counterfactual sentences.

Most counterfactual sentences follow common syntactic forms. Our next approach was to study syntax and the way counterfactual sentences are constructed. Using school grammars, we identified the various English tenses. Then, we classified each counterfactual sentence from the dataset using the Stanford Parser. From this classification, we were able to establish what are the most commonly used verb tenses in counterfactual sentences. Under the premise that counterfactual structures are transposable to positive antecedent and consequent [12], we identified the most common grammatical combinations for antecedents and consequents.

During this first phase of the project, we focused on giving a clearer view of what makes a counterfactual sentence, summarized in the following points:

- We identified keywords and modal verbs that are correlated with counterfactuality
- We identified common syntactic forms for counterfactual sentences
- We laid the foundation for the semantic analysis of counterfactuals

2 Context

2.1 Sémagramme

This project is being conducted under the supervision of Pierre Ludmann and Philippe de Groote, respectively PhD student and senior researcher at Inria - Loria and part of the Sémagramme team. Researchers at Sémagramme intend to develop logic-based models, methods and tools for the semantic analysis of natural language. One of their research axis is to process discourse dynamics, such as counterfactuals, by using Montague semantics. Following the works of Chomsky that brought the use of mathematical methods into syntax and formalized generative grammar[11], Montague applied the same methods to semantics by setting truth-value and entailment to sentences[15]. However, since Montague semantics only dealt with isolated sentences, discourse and dynamic phenomena were considered to be beyond its bounds. More recently, team members at Sémagramme have tackled this problem by using continuation semantics and the theories of functional control operators[13].

2.2 SemEval-2020

Following this line of research, this project is Sémagramme's informal entry to the SemEval-2020 Task 5: Detecting Counterfactuals. This is Sémagramme's first entry to a shared task, the aim is to explore this format. Shared tasks have become popular in various research areas as a way to tackle specific problems and compare systems from different research groups on shared data. One famous example is the 2012 ImageNet Large Scale Visual Recogniton Challenge where a deep convolutional neural network called AlexNet created a major breakthrough in the field of computer vision by outperforming other systems by a large margin; this marked the start of a surge in artificial intelligence. As far as natural language processing is concerned, acquiring new data is a considerable investment of time and/or money and remains one of the obstacles to progress. Shared tasks thus provide a convenient solution to this issue. Organizers introduce set challenges to competitors along with relevant data. Submissions are then evaluated according to a baseline and research teams may gather afterwards for a workshop to share their results.

SemEval, as the name suggests, is concerned with the evaluation of computational systems for semantic analysis and more broadly with the exploration of the nature of meaning in language. The SemEval-2020 Task 5 deals with the modelling of counterfactual semantics and causal reasoning in natural language. This shared task aims to provide a benchmark for two problems related to counterfactuals, namely subtask1 and subtask2. For subtask1, participants are to build a system that determines if a given statement is counterfactual or not. Counterfactual statements include imaginary outcomes and are thus relevant to research concerned with the representation of knowledge and reasoning. This subtask serves as the basis for further analysis of causal inference related to counterfactuals. To help model this causal knowledge, subtask2 aims to iden-

tify antecedent and consequent in counterfactuals. According to Goodman[12], a counterfactual statement can be converted to a conditional with a positive antecedent and consequent. Consider this example:

(3) "If Mr. Trump nominates a member of Senate Democrats as a candidate, Senate Democrats will definitely deny the pledge they made before."

Here, the antecedent refers to the if-clause and the consequent to the second half of this hypothetical situation.

This shared task is hosted on CodaLab. CodaLab is an open-source web-based platform for running collaborative competitions in the field of data-driven research. One of its goals is to ensure reproducibility, a major concern in this field.

2.3 Datasets

Train datasets have been provided for subtask1 and subtask2. Participants are encouraged to upload the results they have achieved with the training data to confirm the submission format. Test datasets will be released during the evaluation phase for participants to upload their final submission. The datasets are provided in the CSV file format. To build these datasets, sentences have been extracted from news articles in the domains of health, politics and economics. By looking up the provided sentences on a search engine, we identified that they have been extracted from various online news sources such as The Economist, Business Insider, The New York Times and the Financial Times. Subtask1's train dataset contains 13,000 rows. Each row is split into the following cells: the sentence ID number, the gold label (0 for non-counterfacual sentences, 1 for counterfacual sentences) and the sentence itself. About 89% of the sentences are labeled as non-counterfactual and 11% as counterfactual. Subtask2's train dataset contains 3551 rows. Each row is composed of the sentence ID number, a counterfactual sentence, its antecedent and its consequent followed by their respective start and end string indexes. Some counterfactual sentences have no consequent.

2.4 Evaluation

The scope of this project is to participate informally to this shared task. However, as the evaluation phase has been scheduled from 19 February to 11 March 2020, we might be able to submit our final results officially depending on the progress we make during the second semester. Participants are free to use any resources they want to build their system. Baselines have been made available for participants to assess their results: a SVM model has been used to set the baseline for subtask1 and a Sequence Labeling model for subtask2, whose scripts are publicly available. Competitors have to participate in both subtasks and submit a CSV file for each subtask. Subtask1's file should contain the sentences ID along with their predicted labels.

sentence ID	predicted label
322893	1
322892	0
...	...

Table 1: Submission format for subtask1

As for subtask2, each row should include the sentence ID followed by its antecedent's and its consequent's respective start and end string indexes.

sentence ID	ante. start ID	ante. end ID	cons. start ID	cons. end ID
104975	15	72	88	100
104976	18	38	-1	-1
...	...	...	...	...

Table 2: Submission format for subtask2

For subtask1, the evaluation script will verify if the predicted binary labels match the expected labels and will then compute the precision, recall and F1 score. Subtask2's evaluation script will assess the percentage of predicted antecedents and consequents that match the expected results and will compute the precision, recall and F1 score. After the evaluation phase, participants' submissions will be ranked. Additionally, the evaluation script and gold labels will be released for those willing to self-assess their system.

3 Literature review

3.1 Syntactic Investigation of Counterfactuals

[5] This is the most detailed article covering the syntactic problem of counterfactuals we came across in our literature review. In this paper, author Sabine Iatridou from MIT highlights a diverse ecosystem of syntactic structures that can be considered indicators of counterfactual thinking. At the same time, by comparing different contexts as well as different languages, she shed light upon those structures in terms of whether or not they represent counterfactuals. The article, therefore, set a stricter and more accurate standard, from a syntactic point of view, to thoroughly investigate the conceptual and terminological apparatus of counterfactuals.

As we will discuss later in section 7, the concept of counterfactual, as opposed to common belief, is much broader than just being considered as conditionals with contrary-to-fact antecedents. In many cases, a statement can be considered as a counterfactual even if it contains an antecedent with truth-value.

That's why we have subjunctive conditionals, with the subjunctive mood at the antecedent, i.e. with a syntactic model that feature verbs in the past perfect (or "pluperfect") with a modal would in the consequent. Thus there is a

classic comparison, mainly in linguistic or syntactic form, between subjunctive and indicative conditionals, with the latter featuring verbs in the simple past tense form, and no modal auxiliary in the consequent. From here it is possible to interpret and deduce the opposite of the semantic behavior with the starting point being the difference in syntax.

However, the author points out that when looking at other non-English languages, the subjunctive is not necessary to create a counterfactual conditional. The simplest example including Dutch and plenty of other languages that do not have a subjunctive at all and still have counterfactual conditionals. Another good example in a different case is French, when this language still has a subjunctive mood, but does not even use it in counterfactual. Furthermore, languages like Icelandic do not consider subjunctive alone as being sufficient to make a conditional counterfactual. These cases bring up the issue of indicative, and as a correlate, the term "indicative conditionals' is inappropriate for non-counterfactual conditionals, as there are plenty of languages where counterfactual conditionals are in fact in the indicative mood: those that do not have a past subjunctive, and those that do not have a subjunctive at all.

Another typical element in the counterfactual analysis that the author points out is the counterfactual marker, which is an indicator for counterfactuals in languages like Hungarian. However, other languages have absolutely no such signs, but listeners are still aware of counterfactual implication. This phenomenon is explained by Iatridou in the sense that counterfactual signs are morphologically marked by elements that are pooled from other parts of the grammar, that is, by elements that have uses/meaning other than those signs. Finally, the author tries to debunk some myths about what are considered as signature structures of counterfactual like "if", "if... then", or inversion of a tensed verb, to show that these signs are not always a clear indication for conditional interpretation.

4 Lexical Analysis

To build a language model that could detect counterfactual sentences and identify antecedents and consequents, many would consider using neural networks. In contrast, we decided to take a symbolic approach. We would first start from the core components of language and then make our way up the different levels with which one can study language (from morphology to semantics/pragmatics). We assumed morphology would not be of any help for this problem and chose to first focus on a lexical analysis. Looking at the lexicon used in conditional clauses, could we infer a dependence between a specific lexicon and counterfactuality? In other words, could specific keywords help us determine whether a sentence is counterfactual?

4.1 Keyword Search

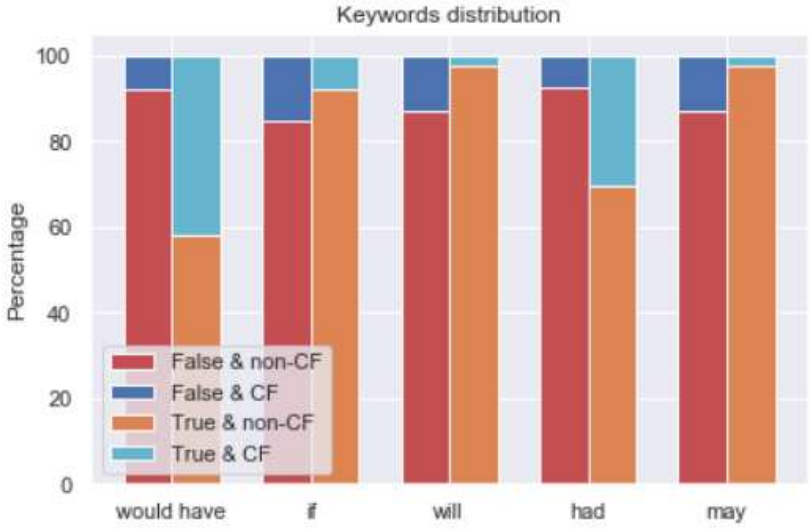
For this lexical approach, we used the dataset from subtask1, containing sentences labeled as counterfactual or non-counterfactual. The first step of this

query consisted in counting the number of occurrences of a specific keyword in all the examples. The keywords were picked from a set of modal verbs (would, should, might,...), based on our intuition concerning the way counterfactuals are constructed. From these values, we established the following two percentage distributions to get a better understanding of the lexicon used in counterfactual conditionals:

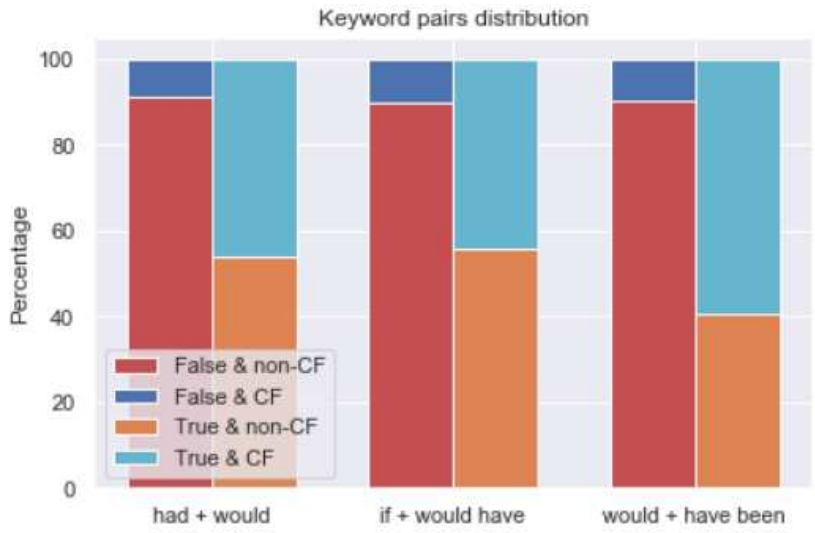
- Of all the examples containing a keyword, what percentage of them are labeled as counterfactual?
- For all examples labeled as counterfactual, what percentage of them contain that keyword?

The results were encouraging; this seemed to be a promising and yet simple path to pursue. Most notably, 57.63% of the counterfactual examples contained the keyword "would", and the presence of modal verbs such as "should" and "might" seemed to be a good indicator that a sentence is not counterfactual.

Consequently, we chose to refine this keyword search. First, we listed the most common words that appeared in the given examples. From this list, we then identified some of the keywords that could help us detect a counterfactual sentence (i.e. mostly modal verbs). What helped us select relevant keywords was to look at the dataset manually, examining counterfactual examples and trying to find common patterns among these examples. We then compared the percentage distribution for the chosen keywords with regard to two variables: presence of the keyword (True/False) and counterfactuality (CF/non-CF). Some of these keywords such as "might" or "could" appeared to be evenly distributed. However, other keywords such as "will", "would", "had", "may" and "if" showed an important disparity in their distribution and seemed to be good indicators as to whether an example contains a counterfactual conditional or not.



Next, we specified our query to observe the distribution for multiple keyword strings in an example. We found an even greater disparity for string pairs such as "had" + "would", "if" + "would have" and "would" + "have been".



We then further specified the query to take word order into account. However, this yielded similar values to those provided by the multiple keyword strings search. Nonetheless, we found that if a sentence starts with "had" (1), it almost certainly is counterfactual: 59 of them were counterfactual sentences and only two were non-counterfactual.

(4) "Had he been a House committee chair, he would have had to step aside."

This corresponds to an inversion between "had" and the subject, which is a typical way to express counterfactuality in the antecedent. This lexical search only accounts for the very first word of a string. However, as we have observed in section 3, a counterfactual sentence does not necessarily start with the antecedent (2). This marks the limitation of the lexical analysis, and this is where the syntactic analysis will prove to be useful in combination with our improved understanding of the lexical aspect of counterfactuals.

(5) "How much easier it would have been had I made the choice for her."

4.2 Statistical Hypothesis Testing

From looking at these figures, we could sense that these keywords were indeed good indicators of a counterfactual sentence. The next step was to test our assumption with statistical hypothesis testing in order to establish if there exists a significant relationship between the two categorical variables "presence of a keyword" and "counterfactuality". The null hypothesis for our test was that

the two variables are independent. We used a Pearson's Chi-square test (χ^2)(1) to compare the observed frequencies to the expected probability distribution.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{1}$$

Using the 2x2 contingency tables we had built for specific keywords, we performed the test with the help of the chi2_contingency method from the SciPy library. We found that, for these keywords, most of the resulting p-values were well below the significance level of 0.05. We thus reject the null hypothesis and can assert that the presence of specific keywords is statistically significant in regard to counterfactuality. The presence of some keywords such as "had", "would" and "been" is strongly correlated with a sentence being counterfactual. Other keywords such as "will", "may" and "can" have an inverse correlation and could help us classify a sentence as non counterfactual.

Keyword / Gold label	False	True	All
0	10117	1429	11546
1	829	625	1454
All	10964	2054	13000
P-value: 1.52e-195			

Table 3: "Contingency table for "had"

Not surprisingly, the presence of some string pairs such as "if" + "would have", "had" + "would" and "should" + "have been" is even more strongly correlated with counterfactuality.

Keyword / Gold label	False	True	All
0	11114	432	11546
1	1083	371	1454
All	12197	803	13000
P-value: 6.25e-231			

Table 4: "Contingency table for "had" + "would"

A lexical analysis of the data showed that some keywords can indeed help us detect counterfactual sentences. Lexicon thus remains a useful tool that we will take into account when building our detection system.

5 Syntactic Analysis

Counterfactuals seem to follow common structures and one way of analyzing them is through syntactic rules. In this research, we first started by studying

how counterfactuals conformed to conditional rules found in formal grammars. Then we classified counterfactual sentences into different forms of verb tenses.

5.1 Conditional Forms

Conditionals in English have dedicated forms and normally come with specific structure. As a preliminary level of this syntactic analysis, the closeness of counterfactual structures to conditionals can be studied.

According to school grammars, there are four types of conditionals[1]: First, Second, Third and Real. First conditionals are used when we want to talk about the imagined future which can be quite likely. Second conditionals are used to talk about the required conditions in order to have a different situation in the present or future. Third conditionals are used for imagining a different past which did not happen and consequently imagining a different outcome. On the other hand, real conditions are used mainly for expressing real events which normally happen or are very likely to happen. Besides these four forms, conditionals can also appear in the mixed form. The different conditional forms are presented in table 5.

Conditionals		
Type	Antecedent	Consequent
First	Present Simple	modal verb with future meaning
Second	Past Simple	modal verb with future-in-the-past meaning
Third	Past Perfect	modal verb with future-in-the-past meaning
Real (1)	Present Simple	Present Simple
Real (2)	Present Continues	Present Simple
Real (3)	Present Continues	Present Continues
Real (4)	Past Simple	Past Simple
Real (5)	Past Simple	Past Continues
Mixed	Past Perfect	modal verb with future-in-the-past meaning

Table 5: Conditional forms in English

Our aim is to study the similarity of each counterfactual input to conditional forms. To do that, the verb phrase forms of each conditional for both antecedent and consequent were first extracted. The same process was done for each input sentence in counterfactual form. Then, the accordance of input sentence with each conditional rule, based on verb phrases, was evaluated. The result of this study, which was performed using the positive train dataset provided on the shared task (mentioned in section 2.3), can be seen in table 6.

From the result provided in table 6, we can observe that nearly 70 percent of the input sentences are not recognized as conforming to any of the existing conditional rules. The second and third highest rates correspond to the third and second conditional rules.

Train data	
Conditional Type	Rate (out of 3551 lines)
First	20
Second	273
Third	469
Real (1)	12
Real (4)	31
First or Real	153
Second or Real	52
Third or Mixed	20
Sentences with no verb phrase	138
None	2383

Table 6: Different conditional forms rate on train data

5.2 Verb Tenses

Sentences can be composed of a variety of verb tenses. Verb tenses can be considered as the most basic properties of a sentence that provide valuable syntactic information. Counterfactuals are no exception to this and one of the main ways to analyze them syntactically is through constituent verb phrases. In English, verb tenses are divided into two main classes, the present and the passive form. These two classes can be further divided into three categories: present, past and future. Various forms of English verb tenses with their respective part-of-speech (POS) tag can be found in table 7.

Verb tenses of counterfactual input sentences were studied using the subtask2 data provided on the shared task (mentioned in section 2.3). Each input sentence is composed of an antecedent and a consequent with their specific verb phrase forms that had to be extracted. Therefore, occurrence rates of different verb tenses for both antecedent and consequent was counted. The result is shown in figure 1.

5.2.1 Most frequent forms

As shown in figure 1, some tense combinations for antecedents and consequents are observed at a higher rate than other modes. Following in this section are some examples of the most frequent forms we identified.

1. Form 1 (rate of 469)

Verb Tenses	
Verb tense	POS form
present simple	VBP/VBZ
negative present simple	VBP/VBZ + VB
passive present simple	VBP/VBZ + VBN
present continuous	VBP/VBZ + VBG
passive present continuous	VBZ/VBP + VBG + VBN
present perfect simple	VBP/VBZ + VBN
passive present perfect simple	VBZ/VBP + VBN + VBN
present perfect continuous	VBP/VBZ + VBN + VBG
past simple	VBD
negative past simple	VBD + VB
passive past simple	VBD + VBN
past continuous	VBD + VBG
passive past continuous	VBD + VBG + VBN
past perfect simple	VBD + VBN
passive past perfect simple	VBD + VBN + VBN
past perfect continuous	VBD + VBN + VBG
future simple	MD + VB
passive future simple	MD + VB + VBN
future continuous	MD + VB + VBG
passive future continuous	MD + VB + VBG + VBN
future perfect simple	MD + VB + VBN
future perfect continuous	MD + VB + VBN + VBG
future past perfect	MD + VB + VBN + VBN

Table 7: Verb tenses with their corresponding POS form

The antecedent consists of a conditional conjunction[1] followed by past perfect simple or passive past simple verb form. The consequent consists of future perfect or passive future simple verb tense.

> "If our ancestors had given up to the cynics, we couldn't have gotten through war, through depression."

2. Form 2 (rate of 278)
   The antecedent consists of a conditional conjunction followed by past simple verb form. The consequent consists of future perfect or passive future simple verb tense.

   > "If it was down to the Bundesbank then we wouldn't have bought any euro zone debt," he said.

3. Form 3 (rate of 273)
   The antecedent consists of a conditional conjunction followed by past sim-

[1] Conditional conjunction describes that something will happen or happened, if the condition is satisfied. Some examples include "if", "if only", "provided" and "unless".

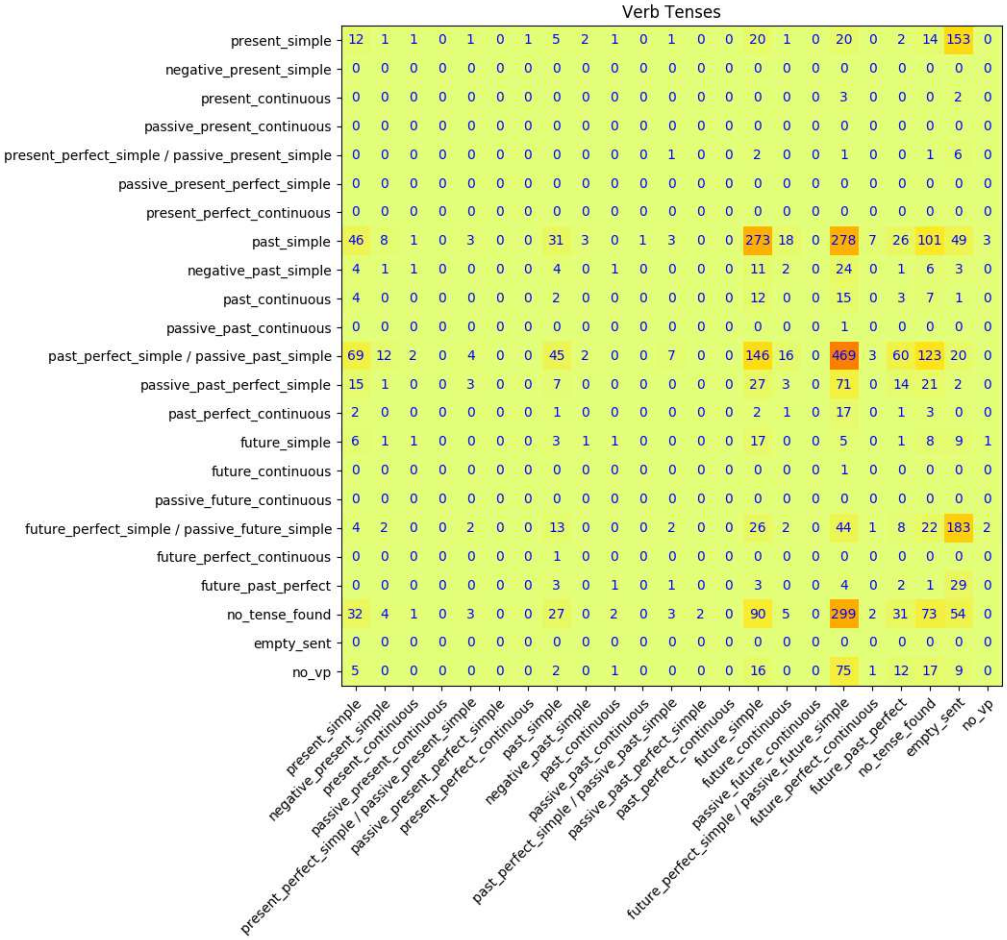
ple verb form. The consequent consists of modal verb with future-in-the-past meaning verb tense.

"Oh if only students knew how easy this stuff is to spot, the entire practice would cease worldwide!."

4. Form 4 (rate of 183)
   This form only includes an antecedent and no consequent. The antecedent consists of a future perfect simple or passive future simple verb tense.

   "They should have turned Afghanistan into a perfect Jeffersonian democracy, ideally with a special focus on gender equality."

Figure 1: Number of different verb tenses on train data. Rows and columns correspond verb tense forms of antecedent and consequent respectively



5.3 Parser

Languages follow a structure and there exists methods to formalize the ordering and arrangement of words in sentences. Stanford CoreNLP [9] is a fast and integrated NLP toolkit with a broad range of grammatical analysis tools. Among some of the provided linguistic analysis tools, constituency parsing can be mentioned. For the syntactic analysis of counterfactuals, this tool enabled us to extract constituency-based parse trees from input sentences that represent their syntactic structure. Constituency parsing used in this project is based on accurate unlexicalized PCFG method. In the following subsections, we will briefly outline how this model works and how it differs from conventional methods.

5.3.1 Context Free Grammar (CFG)

In 1950, Noam Chomsky developed the formalism of context free grammars [4]. Context free grammars are a type of formal grammar composed of a set of production rules. Production rules describe the approach of forming phrases in formal language from smaller units by the simple action of replacement. In other words, the recursive formation of sentences has been captured by these grammars. A CFG consists of a set of terminal symbols, a set of non-terminal symbols, a starting symbol and a set of rewrite rules. Languages generated from context free grammar are called context free languages.

5.3.2 Probabilistic Context Free Grammar (PCFG)

Probabilistic context free grammar is a CFG except that each rule has an associated probability score which indicates the probability of rewriting. A PCFG consists of a set of terminals, a set of non-terminals, a selected start symbol, a set of rules and a set of probabilities for each rule in the grammar. Using PCFGs, we can build up probabilistic models for languages and capture their syntactic structure more accurately. Parsing with a PCFG is finding the most probable derivation for a given sentence. This can be done with the assist of dynamic programming or more specifically CKY algorithm. However, PCFG models come with some limitations. The main deficiency is that PCFGs fail to make use of the lexical context and just take into account the structural factors; probabilities are estimated purely based on syntactical parameters.[10]

5.3.3 Lexicalized PCFG

As it was stated, one of the main weaknesses of PCFGs is the lack of sensitivity to lexical information. This issue has been taken into consideration by introducing a new model called lexicalized PCFG. This model makes PCFG aware of words. The main idea is based on adding head words to each non-terminal and the fact that each constituent has one word which captures its essence. In lexicalized PCFG in Chomsky normal form, rules take one of three forms:

$$X(h) \rightarrow Y_1(h)Y_2(w) \quad \text{for} X \in N, \text{ and } Y_1, Y_2 \in N, and h, w \in \text{terminal symbols}$$
$$X(h) \rightarrow Y_1(w)Y_2(h) \quad \text{for} X \in N, \text{ and } Y_1, Y_2 \in N, and h, w \in \text{terminal symbols}$$
$$X(h) \rightarrow h \quad \text{for} X \in N, \text{ and } h \in \text{terminal symbols}$$

Adding lexical heads to the rules made them more sensitive to lexical properties and significantly improved the performance of PCFGs [8], although this action expands the space of non-terminals and increase the time complexity.

5.3.4 Accurate Unlexicalized Parsing

In this model, the attempt is to focus on unlexicalized parsing to capture structural context rather than too detailed lexical information as in lexicalized PCFG. The method includes applying annotation, markovization of rules, keeping track of node history and tag splitting. The aim is to create PCFGs without lexical information just by using linguistic insights to modify the structure of grammar. PCFGs lack the sensitivity to structural information and their independence assumption is too strong resulting in the loss of information. The motivation behind unlexicalized PCFG is to weaken independence assumption by encoding dependencies into model and annotating each node by its parent category. Unlexicalized PCFGs are time and space sufficient and facilitate interpretation and optimization.[6]

6 Self-examination of our Methodology

During the syntactic analysis, we found some sentences for which we could not detect any antecedent. Therefore, these sentences would fall between the cracks of our syntactic filtering. Some example of these sentences are:

- "The democratic backing of the noble class should have provided the Polish crown with unrivaled consensus and strength, but as time went on the elections were increasingly monopolized by the wealthiest and most powerful magnates, and the trip to Warsaw ceased to seem worth it for anyone else."
- "In a perfect world, the AHA says, people shouldn't have more than 1,500 milligrams (1.5 grams) of sodium per day."
- "Outstanding IAM options will be exchanged for IAM Shares next week based on the in-the-money amount of such options and cash in an amount equal to the special dividend, if any, that would otherwise have been paid on such IAM Shares."

The reason for this result is that we have incorporated in our approach the grammatical definition of conditionals in English, written by the Cambridge

dictionary mostly for educational purposes. So far it is the most programmable product rules for us to process a large corpus with many sentences.

As we analyzed before, the dictionary traditionally classifies the different types of conditional sentences in English. About the semantic interpretation of them, some types are possible or likely, others are unlikely and others are impossible. There are three basic types of conditional sentences, namely type 1, type 2 and type 3, in addition to real and mixed conditionals.

When using these definitions, we cannot locate special sentences, with antecedents that do not have an if structure, or in an inversion of verbs (which are also included in our code). These sentences, after being carefully examined, have some keywords that make their antecedents being contrary-to-fact. These keywords, some even extend to be phrases, which we have not filtered include 'without', 'in a perfect world', 'if not for', 'even with', 'for instance', 'excluding' , 'in that case', or conjunctions such as 'but', 'nevertheless', 'otherwise', 'though'. In addition, as pointed out in the literature review section of the author Sabine Iatridou, even traditional structures such as 'if' or inversion are unlikely to prove a counterfactuals statement.

So while general rules bring us the convenience to program as well as encouraging results, they are not enough to fully cover syntactically ambiguous linguistic phenomena like counterfactuals. We also learned that by applying these rules we had followed an approach that closely resembles a method called prescriptive linguistics, which is in fact applied to construct the dictionary itself and other standard dictionaries in general. Regarding this method, with a biggest goal in mind of creating a set of definitive rules for a particular language, or more precisely, accept this debatable ideology of a standard language, linguistic prescriptivism set the foundation for all the aspects of this language including spelling, pronunciation, vocabulary, syntax, and semantics, to teach what a speech community perceives as a correct form. Some authors even define the term as the concept where a certain language variety is promoted as linguistically superior to others, thus establishing the prescriptive attitude as an approach for norm-formulating and to codification that involves imposing arbitrary rulings of this certain speech community.

This tendency to formulate norm and codify, therefore, explains why we overlook those unconventional keywords that turn out to be crucial in detecting counterfactuals. It raises an issue with prescriptivism is that it tends to explicitly devalue non-standard dialects. In a large corpus that typically assembling many different dialects from various sources, with diversity in linguistic devices, a different form of register and variety, these dialects are easily and mistakenly ignored, although they may fully represent the semantic goal of the analysis. In addition, languages constantly change, therefore standardization tools are quickly outdated and subject to inflexibility.

However, by analyzing a large corpus as well as using verb phrases as a standard for detecting counterfactuals, we also incorporate a method more widely used in academia, which is description linguistics. In the center of descriptivism is the structural approach, which involves collecting a corpus of utterances and then attempting to classify all of the elements of the corpus at their different

linguistic levels: the phonemes, morphemes, lexical categories, noun phrases, verb phrases, and sentence types. It is not an exaggeration to say that it also lies at the heart of many core techniques and methodologies that define Natural Language Processing nowadays. It is therefore a more substantial foundation to investigating counterfactual than traditional tools like prescriptivism. Although based on the foundation of corpus analysis, the fact that we also used a prescriptive set of rules from the Cambridge dictionary (which is much more convenient when programming) has caused undesirable results.

7 Semantic Analysis

The semantic aspect is the key issue that lies behind the nature of the concept of counterfactuals and makes it one of the most thrilling subjects for investigation in philosophy as well as modern science. Recognizing its importance, we took the time to research on this issue, partly to have better insights on a debatable topic that is counterfactuals, partly in the hope of being able to find the semantic features which are applicable to an NLP technique.

The reason counterfactuals carry such weight in fields such as philosophy, artificial intelligence or cognitive science, is because it plays a pioneering role in the theory of rational agencies. For human agencies, counterfactual thinking is an intuitive expression of our free will as well as our rational imagination. For example, people use this form of thinking to think about a specific event after it happens, to gain experience and plan for further actions. Psychological experiments also showed that there is a strong correlation between the belief in free will, thus belief in the meaning of life/ego embedded in experiences, and counterfactual thinking [18]

We can count on the applicability of counterfactual thinking because it reduces central scientific problems to the acts of agency: an agent seeks to learn and make decisions on a given task despite its uncertain nature. The compatibility of this problem with Bayesian epistemology has allowed the application of mathematical models such as probability distribution, or more broadly, probability calculus to binary facts or variables to interpret the world through the lens of counterfactuals. However, it was not until recent powerful connection made by mathematicians [16] between probability calculus and other fields such as Bayesian networks, structural equations and causal models that we actually witnessed the entry of computer science in the field. It is now possible for computer scientists to develop algorithms to perform a semantic interpretation on counterfactuals.

Another important point in the semantic analysis of counterfactual is the failure in traditional logic models to use Boolean truth-functional connectives and truth tables in evaluating the value of a counterfactual statement. This is partly because there are many counterfactuals that have false antecedents and consequents. But the truth of these statements is not entirely based on the relationship between the correctness of the components. Rather, it is decisively dependent on these statements' context. This led to authors like Lewis

[7]coming up with a theory of the context-sensitive of truth-conditions, or Goodman, indicating that adding information to the antecedent could turn a true counterfactual into false, as well the decisive dependent of the truth value of a counterfactual on its background facts, conditions, and laws.

It was Goodman's [12] findings that made him a pioneer in modern analysis of counterfactual. Therefore current semantic models have always been trying to capture logically valid inferences involving counterfactuals, while still holding the premises from Goodman's problem such as the non-truth-functional and context-sensitive of antecedents. Two common models can be referred to as strict and similarity analyses. Both methods have been improved by adding a robust concept of "possible world", and developing logical relationships (based on set theory) between antecedent and consequent in this world. Going into the details of the differences between these methods as well as their sub-methods is beyond the scope of this paper. However, their feasibility shows us the pathways to address the semantic puzzles with formally explicit logical models, thereby possibly help address these puzzles with the power of computation, the same way as what we are doing with probability calculus.

8 Future Work

Some different approaches have been left for future work due to a lack of time. So far, our work has been mainly focused on the use of keyword, or lexical analysis, and establishing a statistical significance for the usage of these keywords. However, keyword search is, as we know, not the most advanced method. Its preliminary result, though encouraging, does not seem to be flawless. And therefore further study is still required in order to improve the efficiency of our current method, and also to try other approaches.

The key to improvement for keyword search is to come up with a set of comprehensive keywords that fully represent the high variability of natural language counterfactuals. Therefore we can expect to improve the method by defining several sub-types of counterfactuals, allowing better coverage of rarer sub-types. Some critical criteria will be set for this process of defining counterfactuals. As we analyzed the division between descriptive and prescriptive approach, a more corpus-based, semantically-oriented categorization of keywords will be implemented. This can be done by extensive literature reading and thoroughly examining real-life counterfactual examples.

While it is possible to view detecting counterfactuals as a task in discourse relation classification, our goal for the near future is to focus on critical features of counterfactuals and provide a more accurate demarcation of each argument of the relation. This will allow us to be more precise in syntactic analysis. So far, our main syntactic concern for this work are constituent verb phrases. Along with connectives, from these starting points as boundaries, other counterfactual arguments will be explored. For example, for one argument detection, we will demarcate from a set of cue phrases to the end of sentence. For two arguments, we will demarcate from conditional connectives to the end of statement or before

the start of the second verb phrase.

As we contemplated before starting the project, the use of machine learning can be a robust way to boost our performance. More advanced supervised statistical approaches, therefore, promise to yield the best results, especially when combining with our previous rule-based approach. Especially, we will apply statistical approaches to counterfactual forms that cannot be easily differentiated by their peculiar patterns. Some of the best models to identify tricky false-positive counterfactuals could be linear support vectors machines (Linear SVM) [2].

Finally, it could be interesting to consider semantic approaches for our future works. The SVM model will be a great start as it is primarily a base method for semantic role labeling (or shallow semantic parsing), a technique that assigns labels to words or phrases in a sentence that indicate their semantic role in the sentence, such as that of an agent, goal, or result. There are also the tools combining Bayesian networks, structural equations, and causal models, developed by Spirtes, Glymour, and Scheines (1993, 2000) [17] and Pearl (2000, 2009) [14]. They address the limitation of detecting approaches only utilizing probability calculus, by affording simple algorithms for causal and counterfactual reasoning, among other cognitive processes. This mechanism would have great potential to aid to tackle very complex problems we encounter in our work, such as those sentences that are syntactically not counterfactual but clearly possess semantic implications.

9 Conclusion

The goal of this project is to detect and model counterfactuals in natural language as part of the shared task "SemEval-2020 Task 5: Detecting Counterfactual". For the bibliographic part, we focused on the concept of counterfactuals and its inherent complexity. We explored how they can be defined and methods to identify them.

In order to get a better insight of counterfactuals on a lexical level, we started by counting the occurrence of some specific lexicon such as modal verbs, which showed that some keywords like "would" appear in close to 58 percent of counterfactual sentences. Then, we compared the distribution of some keywords relevant to counterfactuals. Using the two variables "presence of a keyword" and "counterfactuality", we found that some keywords occured more frequently in counterfactual sentences. We then went on to test these results with statistical hypothesis testing and were able to establish that the presence of these keywords is strongly correlated with counterfactuality.

As a second approach, we examined the syntactic structure of counterfactuals and their correspondence to different formal conditional rules. It was observed that the majority of counterfactual sentences did not fit any conditional forms. However, it is worth noting that third and second conditional rules were more prevalent than other rules. The analysis of the counterfactuals' verb tenses also showed that counterfactuals can appear in diverse forms, but that some

forms are more frequent than others. The fact that a majority of counterfactual sentences follow the same syntactic patterns could turn out to be a solid basis on which to build our detection model.

The results we obtained reflect the controversial nature of counterfactuals. Counterfactual sentences cover a wider linguistic scope than conditional rules defined in school grammars. While syntactic analysis remains a helpful tool to model counterfactual sentences, it is not a sufficient indicator. Therefore, we need to consider semantic approaches which examine counterfactuals more broadly.

References

[1] Ronald Carter, Michael McCarthy, Geraldine Mark, and Anne O'Keeffe. Conditionals and wishes. In *English Grammar Today: An A–Z of Spoken and Written Grammar*. Cambridge University Press, Oxford, 2016.

[2] Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. Semeval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, 2015.

[3] Marcus G. and Davis E. No, a.i. won't solve the fake news problem. *NY Times*.

[4] Ullman J. D. Hopcroft J. E., Motwani R. *Intro To Automata Theory, Languages And Computation*. Addison-Wesley, 1979.

[5] Sabine Iatridou. Grammar matters. *Conditionals, probability, and paradox: Themes from the philosophy of Dorothy Edgington. Oxford University Press. http://web. mit. edu/linguistics/people/faculty/iatridou/Edgington-3. pdf*, 2014.

[6] Manning C. D. Klein D. Accurate unlexicalized parsing. *Association for Computational Linguistics*, page 423–430, 2003.

[7] David Lewis. K.(1973a), counterfactuals.

[8] Collins M. Lexicalized probabilistic context-free grammars. *Lecture Notes*, 2013.

[9] Bauer J. Finkel J. Bethard S. J. McClosky D. Manning C. D., Surdeanu M. The stanford corenlp natural language processing toolkit. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[10] Schutze H. Manning C. D. *Foundations of Statistical Natural Language Processing*. Schutze, 1999.

[11] Chomsky N. On certain formal properties of grammars. *Information and Control, Vol. 2*, pages 137–167, 1959.

[12] Goodman N. The problem of counterfactual conditonals. *The Journal ofPhilosophy 44.5*, pages 113–128, 1947.

[13] De Groote P. Towards a montagovian account of dynamics. *Semantics and Linguistic Theory Vol. 16*, pages 1–16, 2006.

[14] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[15] Montague R. The proper treatment of quantification in ordinary english. *Philosophy, language, and artificial intelligence*, pages 141–162, 1973. Springer Netherlands.

[16] James M Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

[17] Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000.

[18] William Starr. Counterfactuals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.