UNIVERSITÄT
KOBLENZ · LANDAU
Faculty 4: Computer Science

WeST
People and Knowledge Networks
Institute for Web Science
and Technologies

Commonsense reasoning using path analysis on semantic networks

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Web Science

submitted by
Adam Mtarji

First supervisor: Prof. Dr. Steffen Staab
Institute for Web Science and Technologies

Second supervisor: Dr. Claudia Schon
Institute for Web Science and Technologies

Koblenz, Oktober 2019

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	□	□
I agree to have this thesis published on the Web.	□	□
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	□	□
The source code is available under a GNU General Public License (GPLv3).	□	□
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	□	□

....

(Place, Date) (Signature)

Note

- If you would like us to contact you for the graduation ceremony,
  please provide your personal E-mail address: . . . .
- If you would like us to send you an invite to join the WeST Alumni
  and Members group on LinkedIn, please provide your LinkedIn ID : . . . .

Abstract

Commonsense reasoning can be seen as a process of identifying dependencies amongst events and actions. Understanding the circumstances surrounding these events requires background knowledge with sufficient breadth to cover a wide variety of domains. In the recent decades, there has been a lot of work in extracting commonsense knowledge, a number of these projects provide their collected data as semantic networks such as ConceptNet [10] and CausalNet [11]. In this thesis, we attempt to undertake the Choice Of Plausible Alternatives (COPA)[1] [19] [4] challenge, a problem set with 1000 questions written in multiple-choice format with a premise and two alternative choices for each question. Our approach differs from previous work by using shortest paths between concepts in a causal graph with the edge weight as causality metric. We use CausalNet as primary network and implement a few design choices to explore the strengths and drawbacks of this approach, and propose an extension using ConceptNet [10] by leveraging its commonsense knowledge base.

[1] `http://people.ict.usc.edu/~gordon/copa.html`

Contents

1 Introduction 1
1.1 Commonsense reasoning 1
1.2 COPA challenge 1
1.3 Problem description and approach 2

2 Related work 3
2.1 PMI approach 3
2.2 CausalNet 3
2.3 Feature classification approach 3
2.4 Unsupervised learning approach 4

3 Initial Analysis using shortest paths 5
3.1 CausalNet and its structure 6
3.2 Methodology 9
3.3 Implementation 10
3.4 Results and evaluation 13
3.5 Analyzing different design decisions 15
3.5.1 The shortest path between short text 15
3.5.2 CausalNet structure and path relevance 16
3.6 Alternative approaches 18
3.6.1 Measuring causality between short texts 18
3.6.2 Pruning CausalNet 20

4 Extending CausalNet with ConceptNet 26
4.1 ConceptNet and ConceptNet Numberbach structures 26
4.2 Motivation and assumptions 27
4.3 Extracting context-relevant subgraph 27
4.3.1 Semantic relatedness of a node to a context 27
4.3.2 Edge relevance 28
4.3.3 Path length penalty 28
4.4 Results and evaluation 29

5 Conclusion 32

6 Future work 34
6.1 Annotated CausalNet 34
6.2 Leveraging additional data sources and corpora 34

7 References 36

8 Acknowledgments 38

1 Introduction

1.1 Commonsense reasoning

In the artificial intelligence field, open-domain commonsense reasoning has always been one of its greatest challenges since its beginning. Work on this topic was mainly dominated by hand-crafted logical formalizations of commonsense reasoning theories made by experts [8]. However, this approach showed slow progress over the decades due to the inherent difficulty of modelling commonsense reasoning theories by hand that span a sufficiently broad variety of domains, and the lack of a common metric for evaluation that can be used to assess progress and perform comparisons [4].

In the past decade, research in natural language processing has been exploring novel reasoning approaches to commonsense that attempts to extract background knowledge from a mixture of large text corpora and other crowd-sourced knowledge bases. These new approaches to acquiring commonsense background knowledge showed a lot of potential in solving the knowledge acquisition issue that faces open-domain commonsense reasoning. Many projects in this field provide the result of their work as semantic networks containing vast amounts of background knowledge such as CausalNet [11] and ConceptNet [22], [10]. For measuring progress, there are currently multiple benchmarks used for evaluating commonsense reasoning approaches. In this thesis proposal, we are going to focus on one particular benchmark, the choice of plausible alternative (COPA) [4] challenge.

1.2 COPA challenge

The COPA challenge is used in research to assess progress in open-domain commonsense causal reasoning. It consists of a set of questions each composed of a premise and two alternatives. The aim is to answer the question by choosing the alternative that is most plausible, in other words, the alternative with the strongest causal relation with the premise. A causal relationship is defined as one event causing another, which can also be interpreted as the later event been the effect of the first.

An example of COPA question is like the following:

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1. He got a hole in his sock.
Alternative 2. He dropped a hammer on his foot.

The evaluated system has to answer such a question by choosing the most plausible alternative based on if it is asking for CAUSE or EFFECT.

Other questions ask for the EFFECT, like in the following:

Premise: I knocked on my neighbor's door. What happened as a RESULT??
Alternative 1. My neighbor invited me in.
Alternative 2. My neighbor left his house.

The COPA set is organised in two sets of 500 questions each, the first 500 are the Dev set, only used for training and tuning, while the second set is used for final evaluation.

1.3 Problem description and approach

In this thesis, we aim to introduce a shortest path approach to inferring commonsense causality. We use CausalNet as base causal network for this task, and explore different design choices including an extension of this network using ConceptNet.

This work is organised in two parts, in the first part we explore using CausalNet as a causal graph, and various different design choices for measuring causality between concepts and short texts. In the second part we explore using a combination of CausalNet and ConceptNet to achieve better accuracy and improve on some of the drawbacks discovered in the first part. We also discuss various results and assumptions made throughout the work, and propose some improvment that can be considered in future work.

2 Related work

2.1 PMI approach

In [11], the system achieves 65.4% on the COPA test set, using the Pointwise Mutual Information between words in a corpus of millions of personal stories with different window sizes and reasoning with discourse relations. The assumption they used is that weblogs include a substantial amount of information about the causal relationships between everyday events, they run four experiments that compare various statistical and information retrieval approaches to exploit causal information in these stories. The result of their experiment shows that causal knowledge is represented largely in these stories and that using a sentence proximity approach where sentences closer to one another are more likely to be causally related have shown the best performance.

2.2 CausalNet

In [11], the system achieves 70.2% on the test set, using a framework that automatically harvests a network of causal-effects terms from a large corpus text, going by the intuition that narrations typically describe a series of events ordered by time, and this can be exploited to extract causal cues and model causality with causality co-occurrences. Using this framework they implemented a metric to properly model the causality strength between terms, then aggregated this measure for causality reasoning between short texts.

In [20], the system achieves 71.2% on COPA test set, it focuses on the proper treatment of multi-word expressions by additionally considering them a single event or word. Their approach is based on the previous work in causal reasoning between short tests and attempts to improve it by considering multi-word expression in their framework.

2.3 Feature classification approach

In the SemEval 2012 Task 7 of the 6th international workshop on Semantic Evaluation where the COPA challenge was a shared task (task 7), the winning system was the UT-DHLT: COPACETIC System for Choosing Plausible Alternatives [2]. The best system achieves 63.4% on the test set and uses classification based on features derived from bigram co-occurrences and other components. The system uses four features to determine the causal relatedness between a cause and an effect, each feature computes a value that indicates the perceived strength of the causal relation using a different measure of causality. The first feature computes the degree of relatedness between all pairs of bigrams between the premise and each alternative, the second feature computes the temporal relatedness between a cause and effect, the third feature attempts to capture the degree of direct causal relatedness based on how often phrases from the premise and alternative occur within a causal dependency, and the last feature compares the polarity of the premise and the alternative, by mapping each word to its polarity and deducing the overall polarity of the sentence.

2.4 Unsupervised learning approach

In [18], they propose a task-agnostic system that combines transformers [23] and unsupervised pre-training [1]. The system works in two stages, a language model is trained in an unsupervised manner on a large amount of unlabeled data in the first stage using a transformer decoder. Then in the second stage, task specific training is done to fine tune the model for the task at hand, such as the COPA challenge, by performing supervised fine tuning using a labeled dataset. The system achieves currently the best score on COPA test with an accuracy of 78.6%.

Using machine learning approaches such as unsupervised learning has some important drawbacks however. First, a common drawback that most machine learning methods suffer from is bias. Learning about the world from textual data exposes the system to any bias in such data, which can easily happen since, especially in large corpora, textual data is not fool proof reliable, and it is difficult to expect it to represent general knowledge fairly across all its text. Another drawback of machine learning is its poor performance on data it has never seen before, in fact it generally performs closely to a random system when faced with situations it hasn't seen before.

3 Initial Analysis using shortest paths

In this section, I will explain in detail our proposed approach and the methodology adopted throughout the work done on this thesis. Our approach explores using causality and background knowledge extracted from semantic networks to tackle the COPA challenge. We build an inference system that explores paths between concepts in these semantic networks and infer the causal strength between them. The inferred metric we get is then used to compute the plausibility of each alternative to the premise and decide which answer is most plausible.

This thesis is split into two parts, in the first part, we apply our approach to CausalNet only and study the results. We will go through a couple of design choices based on feedback we get from the initial reults of part 1 which will lead us to to second part. In the part two, we explore the use of ConceptNet to tackle or attenuate some of the drawbacks that come with using CausalNet. In Figure 1 we show an overview of the proposed approach. We start by preparing CausalNet and the COPA questions as input for our inference system, which in turn will run the shortest path algorithm and compute the causality strength between the premise and each of the alternatives. As output, the system returns the COPA questions with scored alternatives, the alternative with the highest score identified as the most plausible by the system. Additional changes are explored as we go further, we use the results we get in the first part to identify some areas of improvement we can work on in the scope of this thesis.

As an example of how asingle COPA question is processed, let us consider the following question:

P: The pond froze over for the winter = [pond,freeze,winter]
A_1: People skated on the pond = ['people,skate,pond]
A_2: People brought boats to the pond = ['people,bring,boat,pond]
Asks for: EFFECT

As output we get the follwoing:

A_1 score $= S_1$
A_2 score $= S_2$
if $= S_1 > S_2$, the most plausible alternative is $M_A = A_1$, other wise $M_A = A_2$

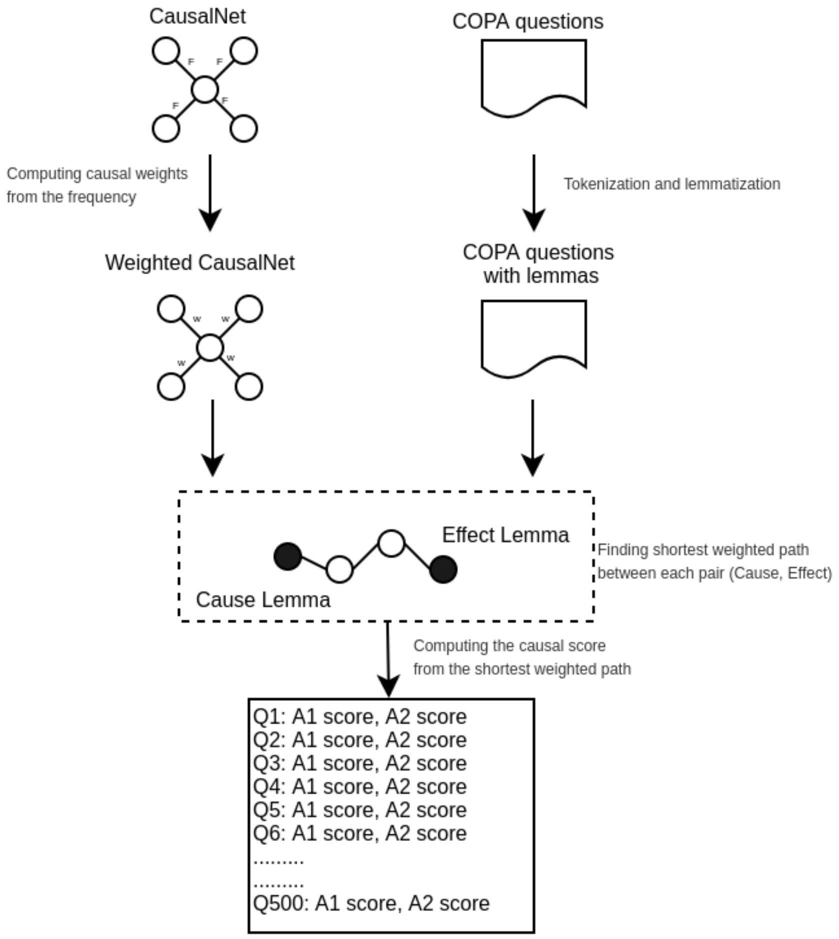


Figure 1: Overview of the proposed approach

3.1 CausalNet and its structure

CausalNet [11] is a network of causal relations weighted with causality co-occurrences between concepts extracted from natural language texts. Causal cues such as "caused by", "induce", "leads to", are used to extract these relations from a large corpus. For example, a causal relation between the concepts "Winter" and "Ice" is represented as a directed edge from "Winter" to "Ice", with a value 468 representing the frequency of which these two concepts were in two different text spans with a causal cue linking them, such as "The water turned into ice because winter is upon us" 1. A fragment of the network can be seen in Figure 2.

In their paper [11], the causal strength of two concepts is computed based on their co-occurrences as well as their individual occurrences in the corpus. They measure causal strength between two terms with the insight that the connotation of causality integrates necessity causality with sufficiency causality. Considering a causal pair (i_c, j_e), necessity causality encoded by $CS_{nec}(i_c, j_e)$ represents that the cause i_c must be present in order for the effect j_e to take place, while sufficiency causality encoded by $CS_{suf}(i_c, j_e)$represents that cause i_c is all it takes to bring about the effect j_e. The causal strength is computed by combining these two as:

Cause	**Effect**	**Causal co-occurrences (frequency)**
Winter	Ice	468
Winter	Freeze	244
Ice	Skate	231
Water	Ice	2202
Freeze	Ice	917
Pond	Water	359
Pond	Winter	90

Table 1: Example pairs in CausalNet of cause and effect and their causal co-occurences in the text corpora

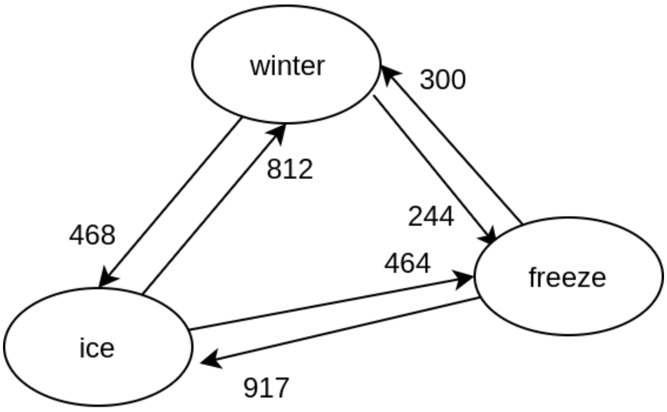


Figure 2: A fragment of CausalNet

$$CS(i_c, j_e) = CS_{nec}(i_c, j_e)^{\lambda} CS_{suf}(i_c, j_e)^{1-\lambda}$$

Their best result is achieved by setting $\lambda = 1$, which effectively means that using only the sufficency component CS_{nec} is the most efficient model. Knowing this, we opt to use CS_{nec} for the rest of our work as the causality metric for each pair (i_c, j_e) in CausalNet.

An initial exploration of CausalNet shown in Table 2 describes a large number of edges with a relatively low number of nodes. Additionally, there is a big gap between the highest frequency of 282927 compared to the lowest been 1. The average frequency is at 10.53 which indicates a large number of very low frequency, in fact, nearly 45.3% of CausalNet edges have a frequency of 1.

The first major difference between our approach and the one used by the CausalNet paper [11] is that in our case, we see CausalNet as a causal graph with weights representing the cost of traveling a specific edge, meaning that higher causality equals lower cost or weight. We then use this network to compute the causal strength of two terms based on the shortest path between them. In the original paper, they only consider direct relations, meaning that the causal strength is dependent on the edge between the two terms instead of the shortest path. Since we plan to use the shortest path instead, we need weights inversely proportional to the causal strength, this implies computing a new causality metric in CausalNet to reflect this paradigm. This leads us to do some

Number of edges	62 675 002
Highest frequency	282 927
Lowest frequency	1
Average frequency	10.53
Frequency = 1	28 429 664
Average out degree	997.2 (max 32066)
Average in degree	997.2 (max 33567)

Table 2: CausalNet anatomy

pre-processing of CausalNet before we apply our approach. For the rest of the work, we will be using the inverse of the causal metric from the original work as the weight w_c of the edges.

$$w_c = \frac{1}{CS_{nec}}$$

As a result, we have CausalNet as a graph with concepts as vertices and causal weights as edges. A fragment of this graph can be seen in Figure 3, Table 3 is a subset of the final data we will be working with.

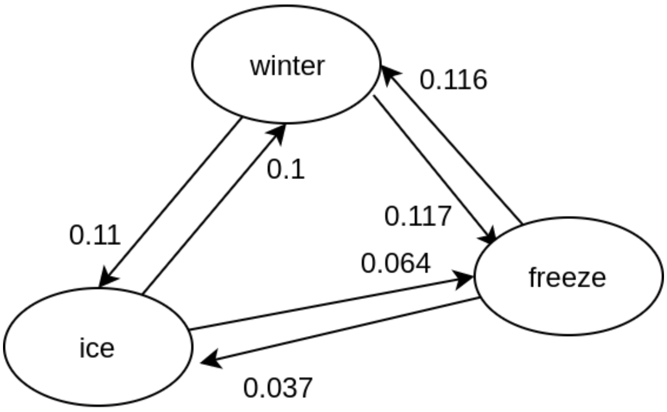


Figure 3: A fragment of CausalNet with final computed wights

Cause	Effect	Frequency	CS_{nec}	w_c
Winter	Ice	468	8.462	0.118
Winter	Freeze	244	8.513	0.117
Ice	Skate	231	40.998	0.0243
Water	Ice	2202	8.462	0.118
Freeze	Ice	917	26.365	0.037
Freeze	Winter	300	8.59	0.116
Pond	Water	359	4.066	0.245
Pond	Winter	90	1.014	0.985
Ice	Winter	812	9.92	0.1
Ice	Freeze	464	15.46	64

Table 3: A subset of data from CausalNet with final computed wights

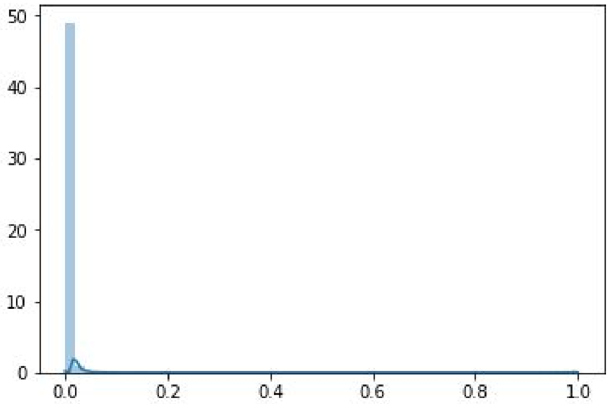


Figure 4: Distribution of computed weights

3.2 Methodology

In the first approach, we build an inference system that takes as input a COPA question, which has a premise and two alternatives or choices and asks for either the cause or the effect. The system returns as output the alternative with the highest plausibility to the premise based on the direction (cause or effect). We will use the examples found in Figures 6 & 7 for illustration throughout the rest of the thesis.

In order to find out the most plausible answer, we need a reliable metric to measure causality between the premise and each alternative. Since CausalNet only provides causality between single terms such as 'skate' and 'winter', we need to leverage this information to get the causality between two sentences such as 'The pond froze over for the winter' and 'People skated on the pond'. In the related work done by CausalNet, they used a simple straight forward approach where they aggregate all the causal links between terms from both sentences with the intuition that each pair of terms contributes to the overall causal strength between the two sentences. In this first approach, we are going to use a similar method with the difference been, that in our case, we will be using the shortest path between the terms instead of the direct causal link used by CausalNet.

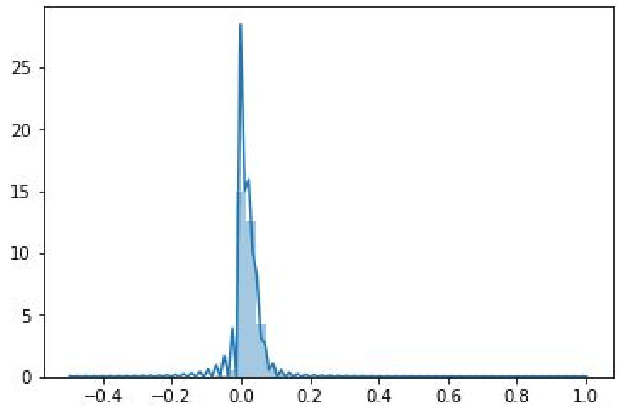


Figure 5: Applying the log function on the causal weights

Premise: My body cast a shadow over the grass. What was the CAUSE of this?
Alternative 1: The sun was rising.
Alternative 2: The grass was cut.

Figure 6: Backward causal reasoning example from the COPA set

3.3 Implementation

Pre-processing

Each of the premises and the alternatives should be pre-processed in order to make their content mappable to CausalNet. We refer to lexical analysis methods such as tokenization and lemmatization [2] to achieve this.

We start with standard tokenization of the sentences which involves breaking down the sentence into an array of words called tokens, for example, the previous premise can be broken down into:

The pond froze over for the winter = [The, pond, froze, over, for, the, winter]

The next step is lemmatization which consists of reducing each token to its common base form based on its grammatical sense, this form is called a lemma. The reason we need lemmatization is that often these sentences are using different forms of the same word such as 'freeze', 'froze', 'frozen' and 'freezing', while in CausalNet we usually only have the base form of the word available except rare cases where we can have more than one form. With lemmatization, we can break down a sentence into lemmas that can be mapped directly to CausalNet without any additional processing or matching. The lemmatization of the previous sentences would then output the following:

P: The pond froze over for the winter = [pond, freeze, winter]
A_1: People skated on the pond = [people, skate, pond]
A_2: People brought boats to the pond = [people, bring, boat, pond]

[2] https://maelfabien.github.io/machinelearning

Premise: The pond froze over for the winter. What happened as a RESULT?
Alternative 1: People skated on the pond.
Alternative 2: People brought boats to the pond.

Figure 7: Forward causal reasoning example from the COPA set

The final step in the pre-processing phase is balancing out each of the three sentences by removing any word that occurs at least once in two different sentences, thus keeping only unique words in each one relative to all three of them. For example, in the previous case, the word 'pond' occurs in all three sentences and the word 'people' occurs in both alternatives, these two words will be removed from all three sentences giving us the final lemmas:

P: The pond froze over for the winter = [freeze, winter]
A_1: People skated on the pond = [skate]
A_2: People brought boats to the pond = [bring, boat]

Basically this means we remove any intersections between all three sentences, avoiding any redundant information that may impact or dilute the score.

Inferring causality

The inference system will be using the shortest path between terms or concepts as a measure of causality between them. In order to compute the causality between the premise and an alternative, the system takes as input the pairs of concepts from both texts, then computes the shortest path for each pair, returning the path and the weight of each edge. The score of each pair is computed as the inverse of the sum of the shortest path weights. The causality score of the premise to the given alternative is then computed as the average causal score, the sum of all the pair causalities divided by the number of pairs. Contrary to the previous paper, we choose to use a more straightforward scoring by taking the average score instead of dividing the sum of the scores by the number of active agent [11]

Preparing the input

The input to the inference system will be structured based on the result of the pre-processing phase. We take the lemmas of the premise and the alternative and we pair them together as a pair of concepts (cause, effect). Deciding if a concept should be in the cause slot or the effect slot depends on the direction of causality defined based on the question in COPA, if it is asking for the cause, then this means that the premise is the effect of the alternative, meaning that the premises concepts should be put in the effect slot, while the alternatives concepts should be put in the cause slot. If the question is asking for the effect, it means the premise is the cause of the alternatives, in that case, we reverse the slots.

This allows us to simulate looking for the cause (backward causal reasoning) and for the effect (forward causal reasoning) by specifying the direction of the shortest path. CausalNet is organized as cause → effect, meaning that in the shortest path, the source is always the cause, and the target is always the effect, so by switching the source and target of the path based on if we are looking for the cause or effect, we can simulate forward and backward causal reasoning in CausalNet.

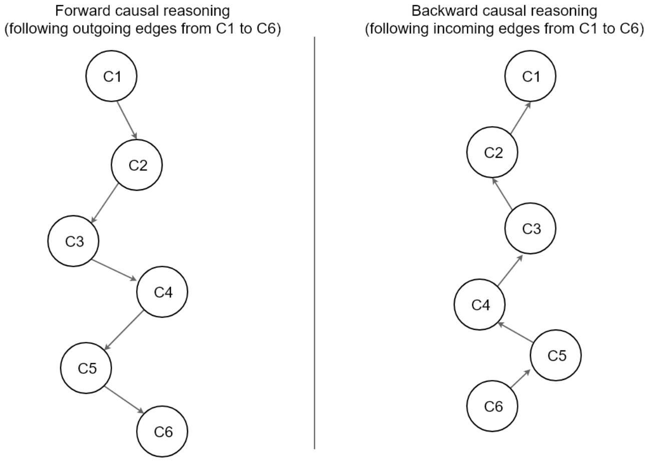

Forward causal reasoning
(following outgoing edges from C1 to C6)
Backward causal reasoning
(following incoming edges from C1 to C6)
C1
C2
C3
C4
C5
C6
C1
C2
C3
C4
C5
C6

Figure 8: Forward and backward causal reasoning using shortest path algorithm

In the previous example, the input would look like the following:

P: The pond froze over for the winter = [freeze, winter]
A_1: People skated on the pond = [skate]
A_2: People brought boats to the pond = [bring, boat]
Asks for: effect

Computing causality between P and A_1:

Context: [freeze,winter,skate]
Cause slots: [freeze,winter]
Effect slots: [skate]

Causal pairs to lookup: (freeze, skate), (winter, skate)

Computing causality between P and A_2:
Context: [freeze, winter, bring, boat]
Cause slots: [freeze, winter]
Effect slots: [bring, boat]
Causal pairs to lookup: (freeze, bring), (freeze, boat), (winter, bring), (winter, boat)

Computing the causal score

In the previous example, the system receives as input a set of causal pairs, for which it needs to fetch the shortest path. The first output is then a path and its edge weights for each pair.

For example, for P and A_1:

Data source	Method	Dev Accuracy %	Test Accuracy %
Gutenberg	PMI (W=5)	57.8	58.8
CausalNet	$CS_{\lambda=1.0}$	62.6	69.8
CausalNet weighted	Shortest path	61	67
CausalNet weighted	Shortest unweighted path	54	54.2

Table 4: Results of the first approach using the new weighted CausalNet and the shortest path algorithm

(freeze, skate): [freeze, ice, hockey, skate]→$[w_{freeze \to ice}, w_{ice \to hockey}, w_{hockey \to skate}]$
(winter, skate): [winter, skate]→ $[w_{winter \to skate}]$

The weights are used to compute the causality score of the pair of cause and effect. The system uses the inverse of the sum of the weights as causality score since the weights are inversely indicative of the causal strength between two concepts.

$$CS(cause, effect) = \frac{1}{\sum_{n=1}^{l} w_n}$$

For example, the causal score for freeze and skate can be computed as the following:

$$CS(freeze, skate) = 1/(w_{freeze \to ice} + w_{ice \to hockey} + w_{hockey \to skate})$$

Inferring the plausibility

We compute a plausibility score based on the causal score of each pair of concepts. The plausibility score PS of an alternative to the relative premise is given by the sum of all the causal scores CS_n divided by the number of pairs (or paths) N.

$$PS = \frac{\sum_{n=1}^{p} CS_n}{N}$$

The alternative with the highest plausibility score is then returned by the inference system as the most plausible answer to the COPA question.

3.4 Results and evaluation

In Table 4 we present results obtained using both the shortest path and the shortest unweighted path. We also perform statistical significance test with the CausalNet system to check the null hypothesis [3] between our proposed approach and the one used in the previous paper.

As shown in the Table 4, using the shortest path outputs lower accuracy on COPA then the original method used by CausalNet. Ignoring the edge weight results in a higher score then if using the shortest weighted path, this raises a number of questions regarding the method used and the structure of CausalNet. We computed the average path length, and found that using the shortest unweighted path method, we get an average path lenght of 0.85, while we get and average length of 10.39 using the shortest weighted path method. This high path length using the shortest weighted

[3] https://en.wikipedia.org/wiki/Null_hypothesis

System 1	System 2	p-value (dev)	p-value (test)	p-value (all)
CausalNet Weighted	CausalNet	0.0071	0.0001	0.0001
CausalNet Weighted	PMIgutenbergW5	0.2368	0.1461	0.0553

Table 5: Statistical significance test between our system (CausalNet Weighted) and previous work

path poses question relative to the transitivity of causality from one node to another, and the potential for the path to go out of the context of the question itself.

We can think of a number of problem areas we could focus on for this thesis. First the method used, one of the drawbacks of using the simple aggregation of the causal scores of the cause and effect pairs is the fact that each pair has no information about the existence of the other terms, this approach relies on the intuition that each terms or concept is an active agent in the causality between the two short texts and that it is sufficient without taking into consideration any contextual information. Relying on disjoint scores means that we lose information about the context when computing the causality. The context consists of all the concepts, their order and sense in a short text.

Problem areas we can identify from the method used in the first approach:

- Loss of information about other concepts in the same text.
- Loss of information about the order of the concepts.
- Loss of information about the weight of contribution of each concepts
- Loss of information about the proper grammatical sense of each concept

In Figure 9, we give an example of two paths using the weighted shortest path algorithm. As you can see, there is a number of out of context nodes that the paths goes through, due to the shear density of CausalNet, every node has on average about a thousand connection 2, meaning that a path can easily deviate from the context of the premise and the alternative after couple of hops. In a sense, a lot of nodes in CausalNet are hubs connected to a large number of other nodes, which causes an issue for the shortest weighted path algorithm resulting in long paths due to the algorithm relying on the sum of weights of the edge and not the number of hops.

Problem areas we can identify from the data source used for the first approach:

- High density of CausalNet impacting the context of shortest weighted path algorithm
- Long paths and transitivity issue
- Existence of named entities in CausalNet

Before we go further, we ran a statistical significance test to check the null hypothesis and if our proposed approach is statistically significant from the previous approach mentioned in the paper [11]. Table 3.4 shows the results of our statistical significance test, as you can see, our system (CausalNet Weighted) is statistically significant from the previous work. The smaller the p-value is the more significant the statistics are, with any p-value bellow 0.1 considered significant. Our system seems to be less statistically significant with the PMI Gutenburg [19] system.

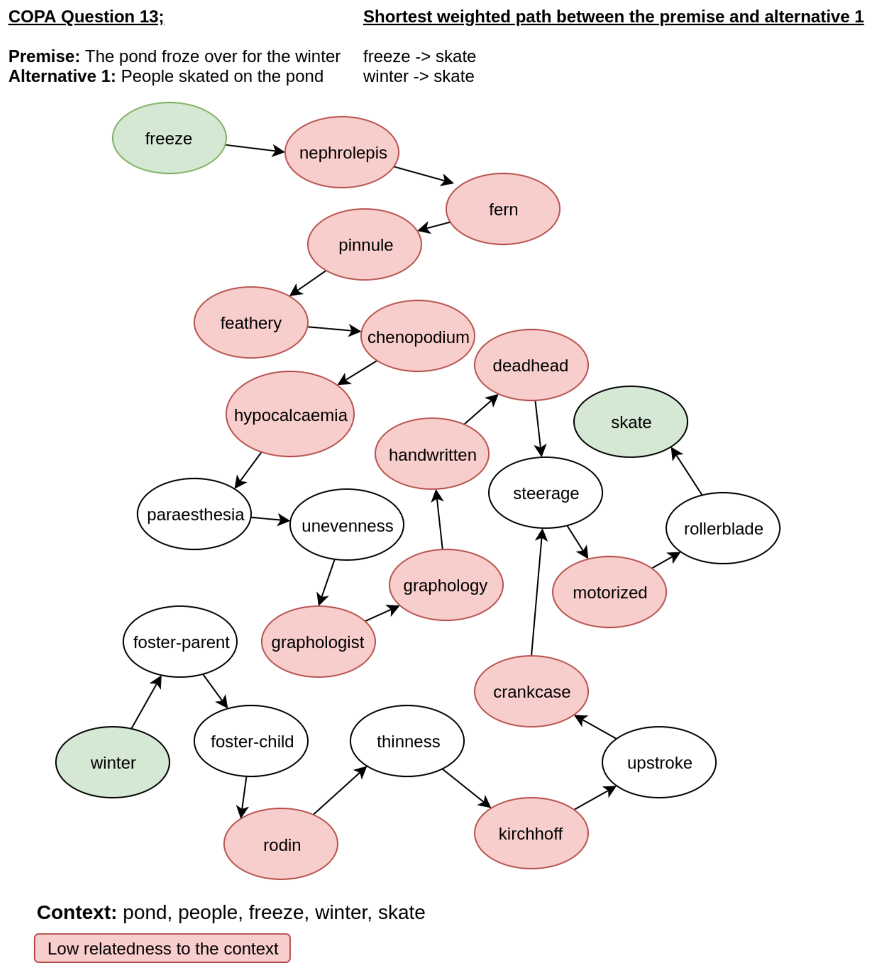


Figure 9: Fragment of CausalNet containing weighted shortest paths from the premise to the first alternative of question 13 of COPA

3.5 Analyzing different design decisions

This first round of results gave us insight about two sides of the problem, on one side we have inherited weaknesses in computing causality between short text and on the other side we have issues with the quality of the paths resulting from the complicated and dense structure of CausalNet. These findings lead us to propose improvements on each side with the aim of making the shortest path method more reliable for this problem.

3.5.1 The shortest path between short text

The initial approach uses a straight forward simple method to estimate causality between short text based on the shortest path between each pair of concepts contained in both short texts. As mentioned before, there are weaknesses to this method, the concepts don't have any information about other concepts in the short text, the order of the concept is not considered either, and finally, the sense of the concepts is also ignored. It uses a direct mapping from the short text lemmas to CausalNet, which in theory, the intuition isthat each concepts is an active agent in the short text causality context, and therefore computing the causal score of each pair between the two short text should be enough to represent their causality to one another. While the results reached by CausalNet using this intuition are proving that it is a reliable method, we would like to take it further by proposing some possible solutions that can help tackle the issues mentioned before.

The first issue would be the word sense in the short text, the main problem is that CausalNet doesn't store nodes with their different senses, in other semantic networks such as WordNet [13] or BabelNet [15], the nodes of the graph are synsets, a collection of words with the same sense. By using NLP [4] and WordNet, it is possible to get a pretty accurate sense of a word in a short text, this could have been easily mappable to CausalNet if it was built with synsets as nodes. Integrating synsets into CausalNet would require us to annotate the nodes with WordNet synsets, this would require rebuilding CausalNet from scratch which would go outside the scope of this thesis, so we won't be implementing any suggestions regarding this issues, but a brief proposal will be discussed in the future work section.

The second issue concerns the order of the concept in the short text, conserving such information while mapping to every single concept to CausalNet nodes is very challenging. We could imagine some sort of algorithm that constructs the shortest path between two short texts that starts from the first concept in the first short text and goes through every other concepts in the exact order it is until it reaches the last concepts in the second short text. But this might have some merit only if we assume that the order in the short text is equivalent to its order in the timeline of the causal event, which is not always the case in natural language. Another approach would be to store ngrams of different sizes as nodes in CausalNet, integrating different expressions as nodes made of different ngrams, this would require rebuilding CausalNet. Since this would also go outside the scope of this thesis, we are not going to implement it but we will discuss this suggestion in the future work section.

The last issue that we are going to look at regards awareness of each concept of the other concepts in the same short text (regardless of the order). We can imagine multiple ways to somehow preserve at least part of this information when computing the causal score, and we would like to focus on one possible approach specifically. We opt to leverage the shortest path algorithm for this problem, so we propose an approach where we build a new temporary node from all the concepts in a short text, this node would contain every lemma, and will have all the edges of the nodes combined. We will then temporary add a node for each short text in CausalNet and run the shortest path algorithm to find the shortest path between these temporary nodes, the intuition here is that the shortest path algorithm will have information about all the concepts in the short text and their edges in the starting node, and would choose the best path based on that information, which we assume should at least preserve part of the information encoded in the short text. Further detail on the implementation and its results are discussed in the next section.

3.5.2 CausalNet structure and path relevance

CausalNet is comprised of millions of edges, with almost half of them of frequency 1, this isn't an issue in the original method implemented by the CausalNet team because they only use edge between two concepts to compute the causal score. However, our

[4] https://en.wikipedia.org/wiki/Natural_language_processing

methodology uses the shortest path between concepts to compute the score, meaning that the density of the graph has an important impact on the path found. In our previous results, after exploring multiple paths, we notice many out of context edges that are not relevant to the question such as in the example shown in Figure **??**. In order to reduce the chances that the shortest path algorithm would pick these edges, we try to reduce the density of CausalNet by pruning it and reducing it to its most relevant edges. We provide further detail on its implementation and the results found in the next section.

3.6 Alternative approaches

3.6.1 Measuring causality between short texts

As mentioned before, we have a number of issues with the initial methodology used to compute the causal score between two shortest texts. Improvement can be made in theory if we take additional steps when building CausalNet, but since this thesis focuses on using CausalNet as is and extending it, we will focus on the third approach mentioned previously in the discussion about the shortest path between two short texts.

The approach consists of extending CausalNet with new temporary compound nodes that each represent a short text. The idea is to pack as much information extracted from a short text in a single node, then leverage the shortest path algorithm to find the best path between two short texts. The intuition is that having a single node represents a short text, and be aware of its composing concepts, should be able to partially preserve context information when looking up the shortest path.

We build these new temporary compound nodes by attaching to it all the edges of its composing concepts, if an edge is present in two concepts, we simply pick the best edge (edge with the lowest weight).

For example, let us consider the first question in COPA:

P: The sun is rising
A_1: My body casts a shadow over the grass
A_2: The grass was cut

After preprocessing we get the following lemmas for each:

P: The sun is rising = sun, rise
A_1: My body casts a shadow over the grass = bod, cast, shadow
A_2: The grass was cut = cut

In order to find the shortest path between the premise P and the first alternative A_1, we first create a new compound node for each.

The sun is rising = sun_rise
My body casts a shadow over the grass = bod_cast_shadow

These nodes will have the edges of its composing concepts, for example, sun_rise will have the edge of both sun and rise, if an edge to a concept C is present for both sun → C and rise → C, we pick the edge with the lowest weight. Fig 10 represents an illustration of this process.

We then simply run the shortest path algorithm from sun_rise to body_cast_shadow like we would for any two concepts in CausalNet. Our previous methodology will then change to using a single pair causal score instead, with this pair being the two temporary compound nodes. The rest of the approach is identical, meaning that this becomes a new proposed approach is a special case of the first approach, with a single pair instead of multiple pairs.

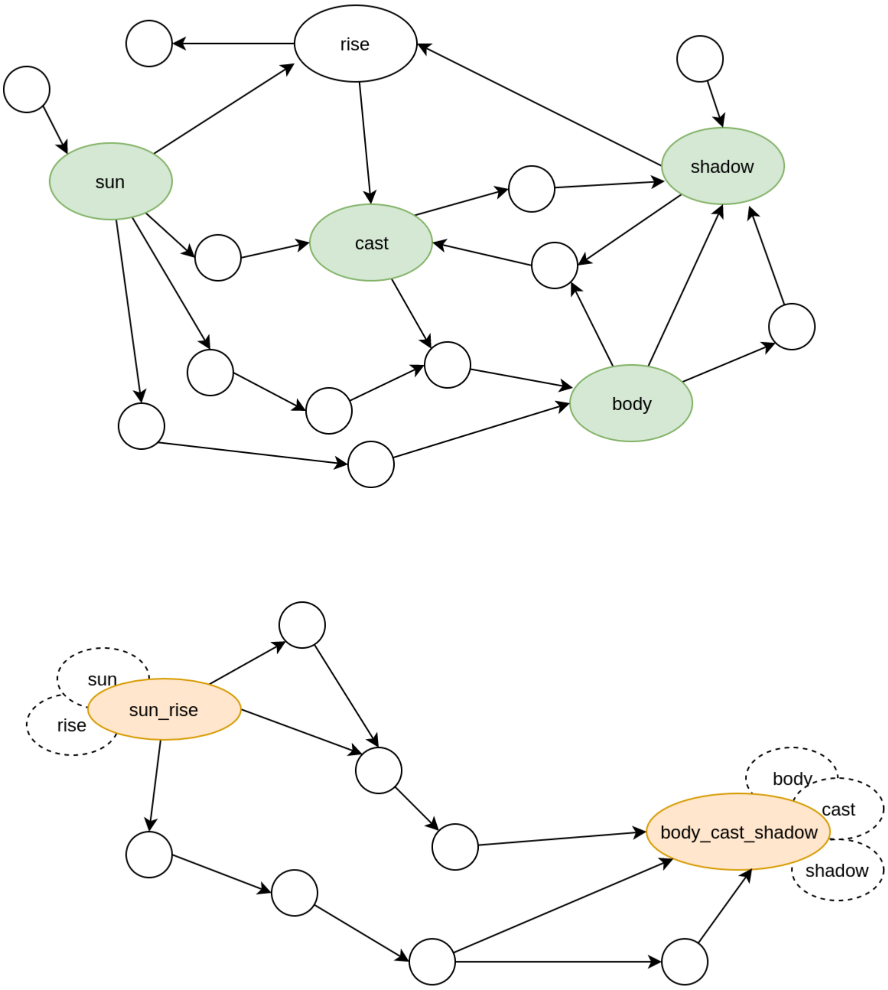


Figure 10: Example of building new compound temporary nodes into CausalNet

Results and evaluation

When using the simple shortest path (unweighted) we get very close results to our initial approach. On the other hand, if we use the weighted shortest path algorithm, we notice an improvement on the test set of COPA at 56%, and slight under-performance on the dev set at 53%, with an overall increase in the average score by about 0,5 at 54,5%.

Next, we take a look at the paths returned by this new approach, and overall we tend to have similar comments as previously with the difference that now we have only one path, intuitively it should be the best representative path that CausalNet can return for each. Meaning that if we had to pick one path to represent each computation between the premise and an alternative from the initial approach, we will end up with something very close to what we get in this second approach.

A statistical signficance test between this approach and the previous one shows p-values of 0.58650 (Dev), 0.34480 (Test) and 0.79450 (All). Since the p-values are > 0.1,

Threshold	Dev %	Test %	avg path length
Top 20%	53.8	52	10.39
Top 40%	53	52	10.39
Top 60%	53.8	52	10.39

Table 6: Results of pruning CausalNet to the top 20%, 40% and 60% edges

this means the output of this approach is not significantly different from the previous approach.

Looking at the average path lenght, we see very comparable results with an average path lenght of 10.9.

What we can conclude from these results is that this approach is at least very comparable to the initial single concept node approach, with slightly improved performance on the test score, and it might be more relevant than the previous method because it reduces the problem to the most important path between the two short texts. The single path between two text is also more relatable for humans since we tend to reason in a contextual manner.

3.6.2 Pruning CausalNet

We have seen so far the density of CausalNet, the number of edges relative to the number of nodes is very high, this may not be the best conditions to use the shortest path algorithm. While exploring the outputted paths, we find multiple examples of path going out of context in order to reach the target, we believe this increases the randomness of the results due to non-relevant edge been prevalent in a lot of the paths. The most straightforward approach we can use to reduce this is pruning, by simply applying some filter criteria on the edges, and reducing the graph to keep as much of the relevant edge as possible.

Pruning CausalNet can be done in various ways, we will start simply by pruning edges based on its causal scores, we apply a threshold to all edges and remove low causality edges. We try a second approach also by applying a dynamic threshold based on each node, we pick a threshold based on the frequency of all the incident edges of a node. In our last approach, we go even further by taking a different perspective, we consider the problem as a link prediction problem, basically we score each edge by finding how probable is its existence.

Pruning by applying a global threshold

The first try is a straight forward pruning by removing the edges with lowest causality, by taking the top 20%, 40%, 60% edges of CausalNet. Results can be found in Table 3.6.2.

As you can see there is little difference if we prune CausalNet this way, it doesn't impact any of the issues mentioned before. The scores are very close and the average path length is unchanged. We take a look at the previous path example to see if there is any noticable changes in terms of out of context edges, and as you can see in Figure **??**, while the first path changed a little bit, the fact is we still have out of context edges all along the path.

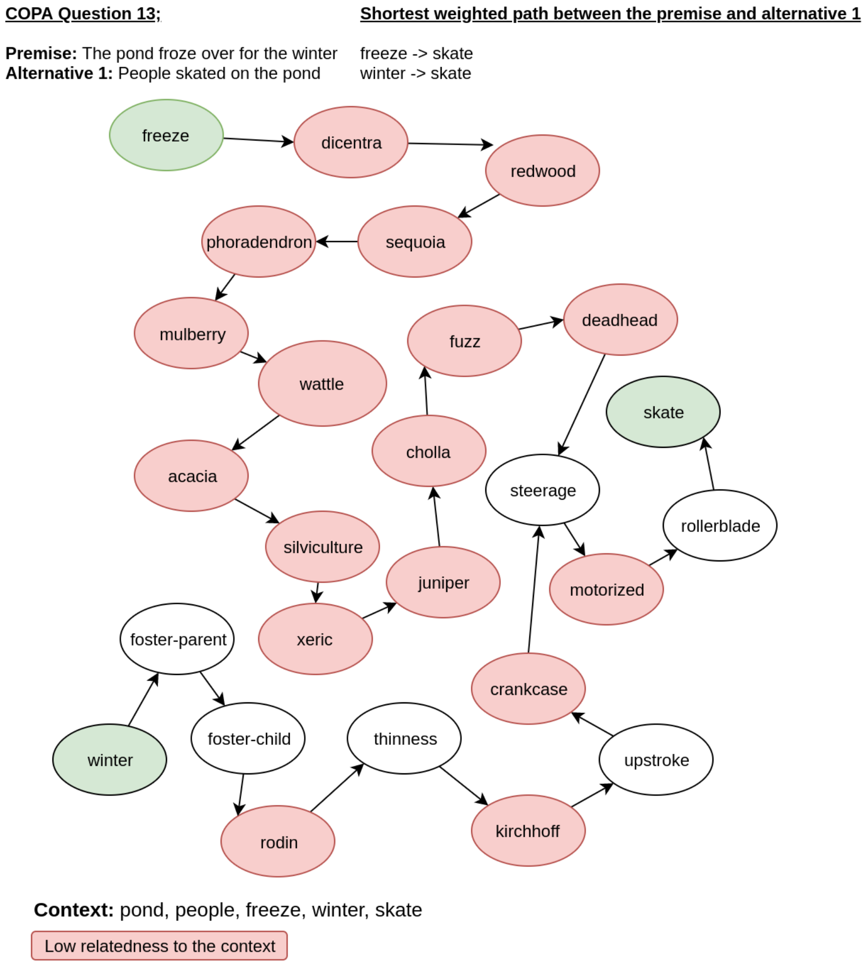


Figure 11: Example of building new compound temporary nodes into CausalNet

A statistical significant test between the pruned (using top 40% of the edges) shows p-values of 1 (Dev), 0.48 (Test) and 0.58 (All). This indicates little to no statistical significance, in fact we get a null hypothesis on the Dev set with a p-value of 1.

Pruning by applying a local or dynamic threshold

Similar to the previous pruning approach, we apply a threshold but this time in a dynamic way, by considering a local threshold based on the target node and the frequency of all the incident edges. For each node, we reduce its incident edges to its top 20%, 40%, 60%, and 80%.

Table 3.6.2 shows the results of this pruning approach, overall we have the same comments as in the global threshold version, with little noticeable change or impact on the problem areas we are looking at.

A statistical significant test between the pruned (using top 40% of the incident edges) shows p-values of 1 (Dev), 0.70 (Test) and 0.83 (All). This indicates little to no statistical significance, we get a null hypothesis on the Dev set with a p-value of 1, and a very low significance on the test set.

Threshold	Dev %	Test %	avg path length
Top 20%	48.2	54.6	10.92
Top 40%	54.2	53	10.65
Top 60%	54.6	50.6	10.54

Table 7: Results of pruning CausalNet to the top 20%, 40% and 60% of each node incident edges

Pruning by removing lowest frequency

Alternatively we tried to prune CausalNet base on the frequency alone, the intuition here is that since a lot of the edges (45%) have a frequency of 1, and many of these edges end up with decent causal strength due to the fact that they generally have low occurrence in the corpus (like some rare words or scientific terms). We can at least try with the minimum possible threshold of > 1.

This threshold returned an accuracy of 52.4 % (Dev) and 52.6 % (Test) and very little impact on the overall average path length. Overall the results were very similar to the global threshold pruning method.

pruning using link prediction and scoring the edges

The idea is to score the edges using a link prediction algorithm [7] which helps determine how probable is a link between two nodes . We use this new score to filter out edges that may be not contributing much to CausalNet. The assumption here is that, since CausalNet was not built manually, it must have collected many odd links. The overall structure of CausalNet should be more leaning towards true causal links then false ones, so by using link prediction on CausalNet, we can score each edge by how probable its existence in CausalNet is and then apply a score threshold to prune it.

Link prediction algorithm:

We will be using a Global link prediction algorithm, meaning that we take into consideration the whole topology of the graph.

The method we will be using is based on identifying a kernel function (or Objective function) for the graph which is capable of predicting if there is an edge between two nodes. In case of a weighted graph like CausalNet, it predicts the potential weight that the edge has.

In a typical link prediction scenario, the goal is to predict future edges or missing links, but in our case, we want to use it differently, we want to score the existing edges and find a way to filter out “noisy” links.

In the normal case, we would need to start by identifying a kernel function that fits our graph. For this, we chose the exponential function with a decay factor Alpha, which provides a score based on all paths between two nodes with longer paths weighing less. Another parameter to consider is the K largest eigenvalues of the eigendecomposition of the adjacency matrix, since in practice it wouldn’t be possible to compute all the Eigenvalues for such a large matrix, and we believe that smaller Ks should perform well in predicting the scores. Since computation of the eigendecomposition takes a long time and rquires a lot of machine power, testing various possible values for K is very difficult, we did try with values of 25, 100 and 300 and we didn't notice any significant difference, so we are fixing K at 100 for the rest of the experiment for convenience.

We tried to test how well the exponential function (our kernel) fits the graph and what values to give α. The function in question is as follows:

$$F(A) = UF(\lambda)U^{-1}$$

With:

$$F(x) = \alpha e^{x} + \beta$$

With α been the decay factor and β a constant used to adjust the curve.

We split the edges into training and testing and ran a few iterations to find good a and β values. We ended up with $\alpha = 0.1$ and $\beta = -1$, the curve fitting problem can be seen in the plot in Figure 12

Before computing the kernel function on the eigenvalues, we first normalize them by dividing each value by the largest eigenvalue, we do the same on the target diagonal. The reason we do this is to scale them both on the same level so we can compare and plot.

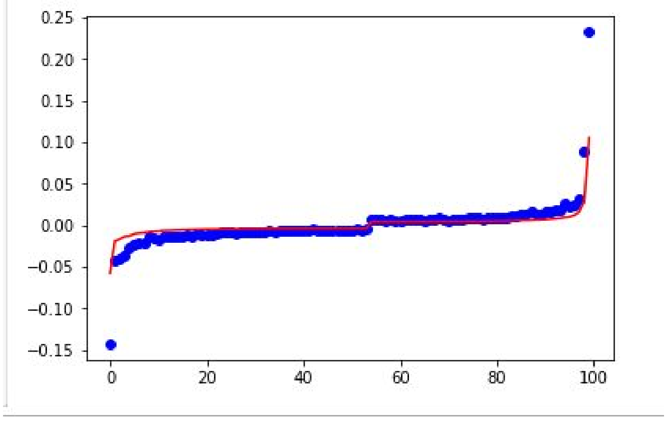


Figure 12: Fitting the kernel function

As you can see in Figure 12, the exponential function with decay factor $\alpha = 0.1$ and a $\beta = -1$ has a very good fit. We also consider a different set of data just in case CausalNet had a lot of noise by default. We take the Dev results of our first approach, from the first 500 COPA questions, and we go through each path and collect edges that have been present only on true positives. We call this new subset of CausalNet edges the good edges and consider them a gold standard. We fit our kernel function on these good edges to evaluate it and as you can see in Figure 13, the kernel function is still a good fit.

After applying this function on CausalNet graph, we get a prediction score for every pair of nodes, resulting in nearly 4 billion scores (Matrix of 62Million X 62Millions). Since we are using link prediction to filter existing edges, this means we have to slice this new matrix to only keep the scores of the existing edges.

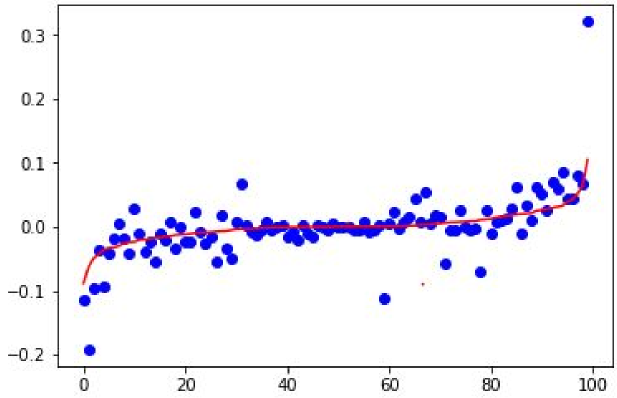


Figure 13: Fitting the kernel function using only confirmed good edges

Score	> 0	> 0.01	> 0.05	> 0.1	> 0.15
N edges pruned	7 472 134 11.9%	24 562 563 39.1%	55 181 194 88%	61 916 821 98.8%	62 512 990 99.7%
COPA-Dev %	53.80%	51.80%	55.20%	49.80%	50.20%
COPA-Test %	57.20%	56.80%	56.40%	53%	53.60%
COPA-All %	55.50%	54.30%	55.80%	51.40%	51.90%

Table 8: Pruning CausalNet using the scores returned by link prediction

Table 8 represents our results applying various filters on the computed scores, notice the similarity this has with the previous pruning methods, performing any filter (e.g prediction score > x) behaves very similarly to applying the filter on the CausalNet causal scores. This may probably be because what we are doing is using a kernel that fits all of CausalNet, meaning that noise is already implicitly included. You can see that we can get away with pruning a lot of the graph without losing that much. After an initial loss of accuracy by pruning around 10% of the graph, there isn't much difference between pruning 11.9% of the graph and 88% of the graph, a small difference of 0.3% in favor of 88%. The initial 11.9% pruned edges resulted in a loss of 6.9% in the accuracy.

Conclusion regarding pruning CausalNet

These results leads us to believe that our understanding of the issue is still lacking, we should probably not aim to improve the score by simply filtering out the edges, the original paper only used direct links (no paths) so they didn't't have to deal with noise or transitivity of causality between nodes. In our case, a global approach to pruning has little impact on our methods accuracy, we simply can't reach and prune the problematic edges this way.

At this point, we should think of the problem differently, we can see it as a pruning problem where we aim to maximize the number of edges pruned while minimizing accuracy loss. But, unless we find ways to contain the context of each question, we will be missing the problematic edges.

From another perspective, intuitively we can say that if we use the shortest weighted path, then edges with very low frequency will not be prioritized by the algorithm and that in the end, they won't have that much impact. The issue is that low frequency in CausalNet does not mean low causality, the causal score computed by the paper takes into consideration the frequency of the edge relative to the overall frequency of both extremities of the edge. In the end, we end up with a good number of these edges with high causality due to the rarity of the concepts in the edge. This does not mean that their causal score is bloated and not relevant, in fact, it should be seen more as simple out of context edges relative to the COPA question. Taking this perspective into consideration, we can think of ways to extract relevant information to a specific COPA question while minimizing any out of context information. For this purpose, we need to turn to other sources of knowledge such as ConceptNet, a semantic network built from different sources and contains different types of relations representing all kinds of knowledge. In the next section, we will propose an extension of CausalNet using ConceptNet and ConceptNet numberbach with the aim of extracting subgraphs for each question in COPA that contain as much relevant information as possible relative to the context of the question.

4 Extending CausalNet with ConceptNet

4.1 ConceptNet and ConceptNet Numberbach structures

ConceptNet

ConceptNet is a semantic network of commonsense knowledge, it was originally built within the Open Mind Common Sense (OMCS) project from the MIT labs, and as of version 5.5, it expanded to other different sources such as Wikitionary, "Games with a purpose", Open multilingual WordNet, DBpedia, and OpenCyc.

ConceptNet is a knowledge graph that connects words and sentences in natural language with labeled weighted edges. It can represent assertions such as:

A net is used for catching fish

This knowledge is represented as triples where concepts can be compound words or sentences:

net UsedFor catching_fish

ConceptNet offers many different types of relations such as "RelatedTo", "TypeOf", "Entails", "Causes", "HasProperty", "UsedFor" and more. An example of a small subgraph from ConceptNet with the word "thunder" as main subject can be seen in Figure 6.

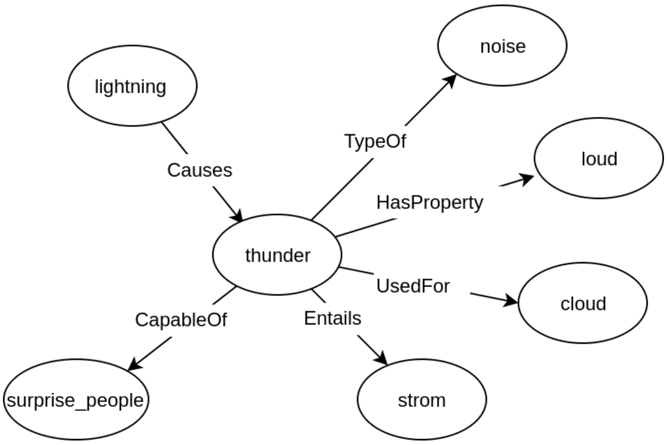


Figure 14: A fragment of ConceptNet

ConceptNet Numberbach

ConceptNet Numberbach is a set of semantic vectors (also known as word embeddings) than can be used directly as a representation of word meanings. It is a snapshot of the word embeddings provided by the ConceptNet open data project. These embeddings have the benefit of having semi-structured, common sense knowledge from ConceptNet along with data from word"vecm GloVem and OpenSubtitles 2016.

The set contains concept vectors of 300 dimensions, each concept can be a term or even an expression. It is also multilingual and shares the same semantic space across all languages, although in our case we will only use the English subset. In Table **??**, we describe a brief view of ConceptNet (English only) anatomy, as you can see, the network is composed of 632 641 edges and 397 925 nodes. The large number of nodes is due to the prevalance of compound concepts or expressions, they constitute 238 024 of those nodes. There are also 46 type of relations currently in ConceptNet, with some of them specific such as "Entails", "Desires" and "UsedFor". While other relations are vague, mainly the "RelatedTo" relation.

4.2 Motivation and assumptions

Our first approach uses CausalNet solely to tackle COPA using the shortest paths instead of direct relations between concepts. We have seen the main weaknesses of this approach, particularly the density issue of CausalNet and the prevalence of out of context edges. One of the strengths of ConceptNet is its semi-structure common sense knowledge base and the fact it captures more than causality making it great for background knowledge discovery between nodes.

Due to CausalNets structure, it would be difficult to simply extend it with new nodes and edges from ConceptNet. The problem is in the difference of the number of nodes and edges between the two networks. On one hand, CausalNet has 62Million edges with around 61 000 nodes, while ConceptNet has 397 925 nodes and around 632 641 edges (English only). This difference in anatomy would mean that simply adding them together will have little impact because of the sheer dominance of CausaltNet due to its high connectivity, especially since a considerable amount of ConceptNet nodes are compound nodes, meaning concepts such as catch_fish, which would be difficult to map to correctly.

We propose then a different approach where we use ConceptNet and ConceptNet Numberbach [5] to try and reduce the issue we get concerning the out of context edges and nodes in the intial analysis. ConceptNet numberbach can enable our system to filter out nodes that are irrelevant to the current context, by extracting a subgraph that only contains nodes that have a certain minimum relatedness score to the context of the question. ConceptNet graph can then be used to filter out the edges that are unlikely to be relevant, our assumption here is that causality should at least imply some correlation in ConceptNet, so if two nodes in ConceptNet don't have any path between them, we consider the edge between them in CausalNet to be irrelevant so we filter it out, essentially doing pruning based on which question, premise and alternative we are processing.

4.3 Extracting context-relevant subgraph

4.3.1 Semantic relatedness of a node to a context

Example of context:

Premise: my body cast a shadow over the grass.
Alternative: the sun is rising.
Context : [body, cast, sun, rise, grass, shadow]

[5] https://github.com/commonsense/conceptnet-numberbatch

We compute the semantic relatedness of a node in the graph to the context of the premise and the alternative by calculating the dot products of the word vectors in ConceptNet Numberbach of the node to each word in the context. If the node has a relatedness score greater then a tunable threshold t to at least one word in the context, we keep it in the subgraph.

If a word is not in Numberbach, we average out all vectors of the concepts that contains that word. For example, the word 'feel' is not in Numberbach but the words feeling and feelings are, so we add a new word to numberbach with a vector that is the average of the other two vectors (this is what the numberbach team recommends when dealing with missing vocabulary).

The relatdness measure R used with with Numberbach is computed as the following dot product:

$$R(concept_1, concept_2) = V_1.V_2$$

With V_1 and V_2 the Numberbach vectors of $concept_1$ and $concept_2$ respectively. The threshold t for R is to be tuned, for our next experiments we will be using 2 thresholds, $R > 0.25$ and $R > 0.5$ with $-1 < R < 1$. Identifying the best threshold can take a long time computing, but as we go beyond 0.5, the extracted nodes are usually just the context nodes themselves and nothing else, which does give a decent score in theory because it is equivalent to falling back to the direct relation method used in the previous work [11], but we would like to focus less on the accuracy score and more on the path quality w.r.t the current context of the question. We set $R > 0.25$ because it gives a decent number of related nodes without falling back to very small context subgraphs of only the nodes in the context. This threshold gives us a good enough number of nodes to both see the use of shortest path and limit the out of context issue.

4.3.2 Edge relevance

Given a subgraph of only relevant nodes to the context, We go through each edge and check it against ConceptNet. The assumption is that if there is an edge in CausalNet between two nodes, there should be at least a path between them in ConceptNet, we assume that causation implies some correlation, meaning some form of common sense connecting them in ConceptNet. We then only keep edges that have a corresponding path in ConceptNet.

Extracted subgraph based on the question context

After extracting a first subgraph made of only the nodes that have some relatedness to the context, we extract a second subgraph from it that contains only edges that are related in ConceptNet, essentially creating a final subgraph which has a strong relation to the context of a specifc premise and alternativein a question. Figure 16 provides an illustration of this extraction process.

4.3.3 Path length penalty

It is common to find a penalty factor in a number of path based relatedness approaches [17], the intuition is that the longer the path, the more penalised it should be. This effectively adds a factor that takes into consideration the number of hops between the

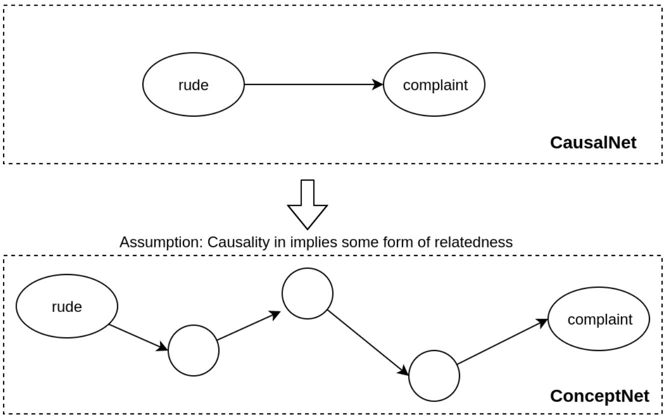


Figure 15: An example of edge filtering

	Dev %	Test %
$\beta = 1$	58.6	60.2
$\beta = 0.9$	59.8	63.6
$\beta = 0.8$	60.2	64.2
$\beta = 0.7$	58.6	64.8
$\beta = 0.6$	60	64.2
$\beta = 0.5$	59.8	63
$\beta = 0.4$	59.2	63.8
$\beta = 0.3$	59	64.6
$\beta = 0.2$	59.4	64.4
$\beta = 0.1$	59.4	63.8

Table 9: Using ConceptNet (no edge filtering) with different β

source and target, instead of just relying on the sum of the path weights. This factor can be seen as a way to simulate causal transitivity, the longer the path, the more penalised the score should be because of loss of causal transitivity from one node to another. A penalty factor β can be added to causality score computation, a factor that scales based on the length n of the path, the score equation becomes the following:

$$CS(cause, effect) = \frac{\beta^n}{\sum_{n=1}^{l} w_n}$$

4.4 Results and evaluation

We run a first experiment with only extracting relevant nodes and not filtering any edges, we also try different values for the penalty factor in order to tune it to a stable value.

In Table 4.4 we show the results of the first experiment for β values between 1 and 0.1. As you can see, for $\beta < 1$, there is an improvement especially on the test set. We can say that any penalty value has some positive impact on the scores, however, it is difficult to decide on a specific β since the improved scores are not consistent as β goes lower. We decide to fix $\beta = 0.1$ for the rest of the experiments for convenience, we can

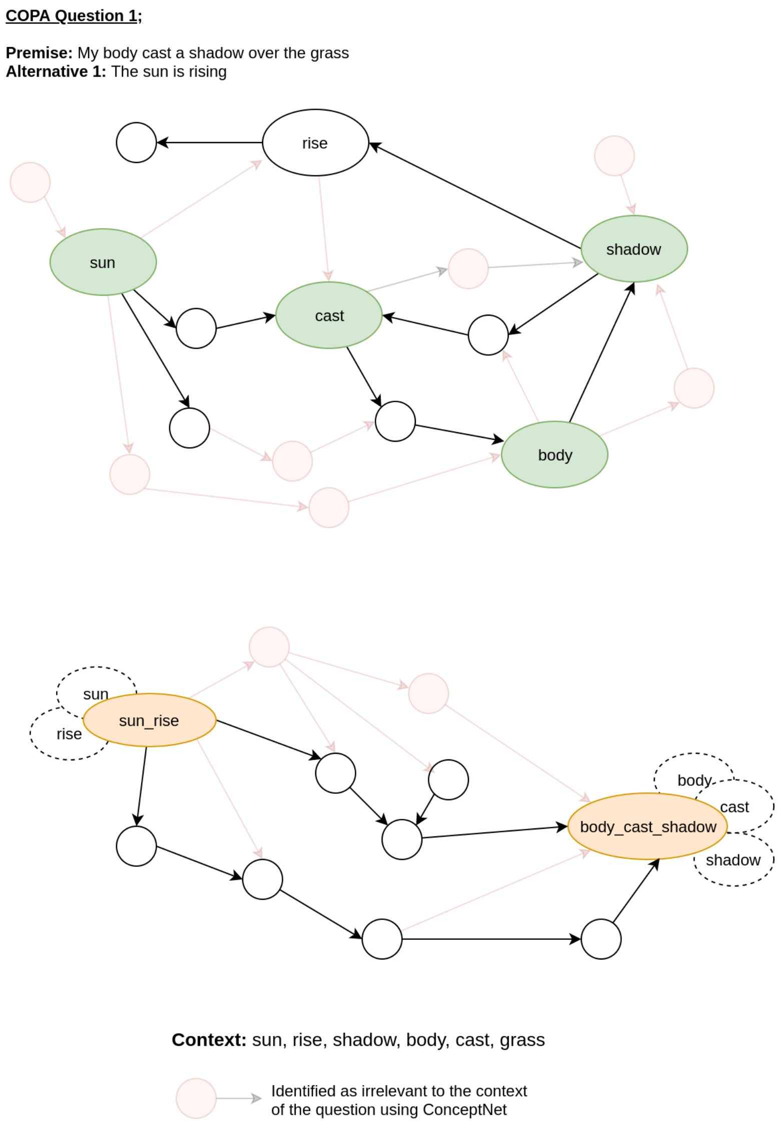

COPA Question 1;
Premise: My body cast a shadow over the grass
Alternative 1: The sun is rising
rise
sun
cast
shadow
body
sun
rise
sun_rise
body
cast
body_cast_shadow
shadow
Context: sun, rise, shadow, body, cast, grass
Identified as irrelevant to the context of the question using ConceptNet

Figure 16: An example of subgraph extraction using ConceptNet and ConceptNet Numberbach

think of further future work that may be able to define a stable β, but for our thesis scope, this value should be fine.

In Tables 4.4 and 4.4 we show the results of multiple methods using ConceptNet and ConceptNet Numberbach. Overall, there is an increase in accuracy across the board, and a fair decrease in the average path length. The edge filtering method however seems to be a little bit underwhelming, although during the experiment, it usually filter out about 20% of the edges, it seems they had little impact on average on the system. We do notice however that the edge filtering method helps the compound method (add compound nodes to CausalNet) increase the accuracy on the Dev set, but it comes with a slight decrease on the Test bringing the average score to about the same range as non edge filtering methods.

In Figure 17 we take a look at the previous example from the COPA question 13, we notice much more relevant paths between the concepts with no out of context edges.

Method	Dev %	Test %	Avg path length
Subgraph with top 1000 context related nodes No edge filtering	55	63.2	4.35
Subgraph with R > 0.25 No edge filtering	59.4	63.8	3.32
Subgraph with R > 0.25 with edge filtering	56.2	60.8	3.17
Subgraph with R > 0.5 with edge filtering	56.8	64	1.71

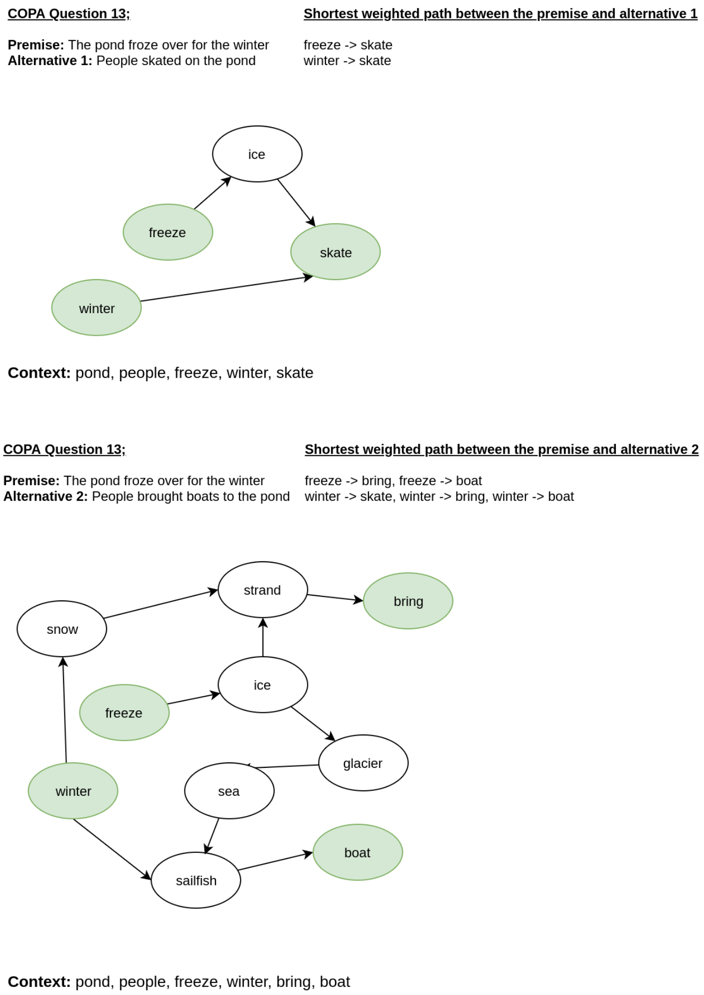
Table 10: Extension using ConceptNet and ConceptNet numberbach

Method (compound nodes)	Dev %	Test %	Avg path length
Subgraph with top 1000 context related nodes No edge filtering	56.2	60.8	3.79
Subgraph with R > 0.25 No edge filtering	57.8	65	2.79
Subgraph with R > 0.25 with edge filtering	60	62,8	2.67
Subgraph with R > 0.5 with edge filtering	58.2	64.8	1.53

Table 11: Extension using ConceptNet and ConceptNet numberbach (compound nodes)

Running a statistical significance test between the ConceptNet extended system and the initial CausalNet system returns p-values of 0.0258 (Dev), 0.0254 (Test) and 0.0013 (All) (using compound nodes, edge filtering and $R > 0.25$). This shows statistical significance between the method used in the first phase and the one used in the second, which indicates that the ConceptNet extension is bringing value to CausalNet by altering it to be usable to handle questions that failed before, this also means that some question that didnt fail before are now failing. Runnig the same test but this time on the original CausalNet system returns p-values of 0.0343 (Dev), 0.019(Test) and 0.0018(All), which also indicates statistical signficance w.r.t the first CausalNet approach.

Overall the ConceptNet extension seems to bring significant statistical difference, which indicates potential for future work in extending or combining both networks to overlap true positives between the two methods.



COPA Question 13;

Premise: The pond froze over for the winter
Alternative 1: People skated on the pond

Shortest weighted path between the premise and alternative 1

freeze -> skate
winter -> skate

Context: pond, people, freeze, winter, skate

COPA Question 13;

Premise: The pond froze over for the winter
Alternative 2: People brought boats to the pond

Shortest weighted path between the premise and alternative 2

freeze -> bring, freeze -> boat
winter -> skate, winter -> bring, winter -> boat

Context: pond, people, freeze, winter, bring, boat

Figure 17: A view on the previous paths for COPA question 13, now with better and more relevant paths

5 Conclusion

In this work, a description of the current challenges in the field of commonsense reasoning was given, along with an introduction to the COPA challenge, a common evaluation tasks used by researchers in this problem area to evaluate and compare their systems on a common problem. An brief overview of current related work showed a number of varied methodologies were used to tackle the commonsense reasoning problem using COPA as evaluation ground. A proposal to use semantic networks and shortest path algorithm was described, which would use a dataset built in a previous work done by the CausalNet team. The thesis work was split into two phases, an initial analysis using the shortest path on CausalNet showed different results and issues regarding using shortest path on CausalNet. Such issues concerned CausalNet density and a number of out of context edges found in returned shortest weighted path, as well as indifference to the contextual information of a node in the initial approach. Some proposed solution and design choices were given, such as CausalNet pruning, and introducing a different method to compute the shortest path between two short text. Results showed slight improvment, but nothing significant using such methodologies, a realisation that moving to a question by question context approach would probably have better results. The second phase consisted of extracting context relevant subgraphs for each COPA question using ConceptNet and ConceptNet Numberbach, the idea is to extract nodes and edges that are the most relevant to the current premise and alternative the system is processing. Initial results showed a good improvement in accuracy, and especially good improvement in path quality w.r.t the context of the question itself.

6 Future work

The scope of this thesis is limited to studying different approaches that leverages shortest path in a causal graph to determine causality between different concepts. While exploring different methodologies we came across some potential areas of research that are outside our scope and would fit better in the future work and perspective section.

6.1 Annotated CausalNet

We focused on the context issue we had to deal with quiet a bit, and as mentioned before, one of the difficulties reside in CausalNet context-less structure where the context of each node is unknown. In fact we can think of it as if each node is the merger of all of its context into a single point, with all the edges converging to it. This of course causes issues for shortest path algorithms since it makes it easy for paths to go outside the context, in fact, the paths end up indifferent and would mix and merge all kinds of contexts between two nodes.

One of the current standard methods to encode contextual information on a graph node are synsets (or synonym sets) were a node is identified by an ID and a set of synonyms or concepts that have very similar meaning grammatically. Some of the best examples currently available are WordNet and BabelNet which see a wide area of use across multiple fields especially NLP and AI.

Theoretically, it would be possible to leverage WordNet and BabelNet along with NLP tool kits such as the Stanford CoreNLP kit or Spacy during the building phase. The idea is to introduce synsets once a pair of cause and effect are identified by CausalNet algorithm (based on some causal cue in a corpus). Instead of integrating the edge between a pair of words representing cause and effect, we would integrate into CausalNet an edge between their respective synsets instead (taken either from WordNet or BabelNet). NLP toolkits allows us to easily annotate a short text with grammatical position, which would then make it easier to map to WordNet or BabelNet.With nodes representing synsets in CausalNet, we can have easier time to map a concept in a short text to its corresponding node based on its context. It is also a lot easier to compute relatedness between sysntest using either WordNet or BabelNet, other common relatedness approaches also exist in previous word that have fairly good success rate.

6.2 Leveraging additional data sources and corpora

During our work on this thesis, we came across different data sources that seem to have a lot of potential for this research area. One particular corpora are movie scripts, they have a particularity that is very hard to find in any other corpus. In a movie script, scenes are described in writing, what the spectator is supposed to see, the meaning tat it should convey and so on, such information is usually non existent in other natural text due to it been implicit, part of the common sense. For example, a person playing music, would be described in writing in a scene as someone holding an instrument, in a bar, with people listening to him.

Such information would look different if it was told in a a blog story for example, it may be described as " someone playing music for an audience", information that playing music requires holding an instrument, and playing for an audience can be done in bars, that people listen to the music, all these information is implicit in our natural day to day talk and is considered part of the common sense realm. since movie scripts are designed in a way that puts everyone, actor and movie staff alike, on the same page, it has to be very precise in describing these scenes. This corpora can be leveraged, maybe even use CausalNet algorithm with some small tweaks on movie scripts instead of short stories.

Other data sources that could interesting to use either for pruning CausalNet, or simply testing purposes, are the likes of SWAG and SNLI. SWAG is similar to COPA, as in its a multiple choice questionaire but with four alternatives ins tead of two. The premise and alternative are automatically extracted from movie subtitles. The assumption is that given te nature of movie dialogs, a lot of common sense knowledge can be extracted from the subtitles. The intuition is that humans can anticipate the situation given a partial description of a scene, for example: "she opened the hood of the car," we can reason that something like "then, she examined the engine" would come next. This is prevalent in movie subtitles, the dataset contains about 113k multiple choice questions about grounded situations, these question can be used either for testing different approaches using CausalNet, or even used as gold standard to prune bad edges for example. SNLI is another common sense dataset that consists of texts and a hypothesis surrounding it. similar to SWAG, this can also be used to either test CausalNet approaches, or prune noisy unreliable edges.

7 References

References

[1] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning, 2015.

[2] Travis Goodwin, Bryan Rink, Kirk Roberts, and Sanda M. Harabagiu. Utdhlt: Copacetic system for choosing plausible alternatives. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 461–466, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Andrew S. Gordon, Cosmin Adrian Bejan, and Kenji Sagae. Commonsense Causal Reasoning Using Millions of Personal Stories. In *25th Conference on Artificial Intelligence (AAAI-11)*, San Francisco, CA, 2011.

[4] Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada, June 2012.

[5] Thad Hughes and Daniel Ramage. Lexical Semantic Relatedness with Random Graph Walks. *Computational Linguistics*, 7(June):581–589, 2007.

[6] Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard H. Hovy. Detecting and explaining causes from text for a time series event. *CoRR*, abs/1707.08852, 2017.

[7] Jérôme Kunegis and Andreas Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 561–568, New York, NY, USA, 2009. ACM.

[8] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November 1995.

[9] Hongyu Lin, Le Sun, and Xianpei Han. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2032–2043. Association for Computational Linguistics, 2017.

[10] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, October 2004.

[11] Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. Commonsense causal reasoning between short texts. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'16, pages 421–430. AAAI Press, 2016.

[12] Nicole Maslan, Melissa Roemmele, and Andrew S. Gordon. One Hundred Challenge Problems for Logical Formalizations of Commonsense Psychology. In *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense-2015)*, Stanford, CA, March 2015.

[13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[14] Leora Morgenstern. Common sense problem page. `http://www-formal.stanford.edu/leora/commonsense/`. Accessed on 20/08/2018.

[15] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[16] E.J. Pedhazur. *Multiple Regression in Behavioral Research: Explanation and Prediction*. Harcourt Brace College Publishers, 1997.

[17] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data*, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[18] Alec Radford. Improving language understanding by generative pre-training. 2018.

[19] Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. 01 2011.

[20] Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. Handling multiword expressions in causality estimation. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*, 2017.

[21] Satinder P. Singh and Shaul Markovitch, editors. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 2017.

[22] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[24] Sewall Wright. The method of path coefficients. *Ann. Math. Statist.*, 5(3):161–215, 09 1934.

8 Acknowledgments

I would like thank my supervisors Prof. Dr. Steffen Staab and Dr. Claudia Schon for providing me with the opportunity to work on this subject. I appreciate their help and support during my thesis, their guidance was crucial to the completion of this work. I would also like to give special thanks to Jun Sun for his help and insight on using Link Prediction algorithm for this problem, and the special care he gave to my inquiries.