Commonsense Causal Reasoning by Causal Relation Extraction from the Web

Abstract

This paper studies the problem of commonsense causal reasoning. We propose a framework to automatically harvest a network of causal-effect terms from a web corpus. We then encode a causal direction and strength on the network based on lexico-syntactic analysis of sentences. Based on this network, we propose an algorithm to detect events from sentences and compute the causal strength between events. We validate our framework using SEMEVAL benchmark called COPA, outperforming all the reported results of existing state-of-the-arts.

1 Introduction

Commonsense causal reasoning aims at understanding the causal dependency between concepts or events in our daily life. To illustrate the problem, we present a question from Choice of Plausible Alternatives (COPA) evaluation [12], which consists of one thousand multiple-choice questions requiring common causal reasoning to answer correctly. Specifically, each question is composed of a premise and two alternatives, where the task is to select the more plausible alternative as a cause (or effect) of the premise.

> Premise: *I knocked on my neighbor's door.* What happened as an effect?
>
> Alternative 1: *My neighbor invited me in.*
>
> Alternative 2: *My neighbor left her house.*

Commonsense causal reasoning has been actively studied, as such understanding is crucial in text understanding, natural language processing, artificial intelligence and other fields. From the above example, we can observe that a key challenge is harvesting common sense causal knowledge that the action of knocking causes that of invitation.

Existing work can be categorized by how such knowledge is harvested. First category is data-driven approach of harvesting causality from web corpus. Best known results in this category leverage Pointwise Mutual Information (PMI) statistic [21] between words in the premise and alternative, to identify the pairs with high correlation. In our example, we can expect that two words *knock* and *invite* co-occur frequently in web documents, which indicate a potential causality.

Though PMI has been an effective indicator in prior literature, it suffers from the following limitations: First, lexical co-occurrence can be a false alarm. In our example, *door* and *house* are also observed frequently together, but identifying this pair as causality leads to falsely identifying the second sentence as a result. This observation suggests that term causality from lexical co-occurrence alone is somewhat noisy and harvesting from causal lexico-syntactic patterns would avoid collecting house and neighbor as a causal pair. Second, co-occurrence is undirectional, while direction is crucial in causality. COPA task is directional such that our example question can be asked in another direction, that is, asking what is a cause of knocking. In this direction, *call* can be a strong cause, but it cannot be a result, though PMI statistic would model the two directions equally likely.

Second category, pursuing the opposite emphasis of depth in understanding sentences, seeks to overcome the limitation of the first approach. These approaches build on deeper lexico-syntactic analysis of sentences, to identify knocking and inviting in our examples as *events*, and determine whether causality between two events hold. Alternatively, ConceptNet [13] leverages human efforts to encode causal events as common sense knowledge. However, these approaches, building on human and heavy analysis, inherently lack coverage, compared to the first category, which is reported to outperform the second [12].

In contrast, our goal is to pursue both breadth and depth in modeling commonsense causality. To pursue breadth, we propose a data-driven approach of harvesting *term causality network* from a large corpus. To pursue depth, we conduct lexico-syntactic analysis of the sentences to extract events and identify events that are strongly causal using the causality network and other semantic resources such as WordNet [22]. Our approach overcomes the two limitations of existing data-driven approaches, with the following novel contributions:

- We harvest term causality network, selectively from causal lexico-syntactic patterns, effectively pruning out false causality observed from lexical co-occurrence. This network encodes both causal direction and strength between terms.
- We redefine causal strength $u \rightarrow v$ to reflect directions, by combining conditional probability of u being the cause of the pairwise causality and v being its effect.

- To quantify causality between phrases, we aggregate term causality leveraging both syntactic and semantic understanding on the premise and the alternatives. For syntactic understanding, we parse sentences to extract *event* from premise and alternatives, consisting of head words in verb and objects. For semantic understanding, we leverage semantic knowledge on each term in the event obtained from WordNet, to properly discount causality from ambiguous terms.

We evaluate the strength of our proposed approach using COPA task, from which ours outperformed all existing results of state-of-the-arts by achieving 68.8% in accuracy. In addition, we validate the accuracy of our causality detection using manually labeled causal relations from ConceptNet as ground truth.

2 Approach

To identify commonsense causality between two statements, our framework includes i) a network of causal relations between words that is extracted from a large web corpus; ii) a metric to compute causal strength between any two words using this network; iii) a heuristic method to extract intra-sentence events that contain causality; and iv) a simple algorithm for aggregating the causal strengths between words and events to compute the overall causality score between two sentences. Next, we describe these components.

2.1 Causal Network

Causality exists in natural language sentence and can be identified by linguistic patterns known as *causal cues* [2]. For example, *"A cause B"* is an intra-sentence causal cue where *A* is a text span that represents the cause and *B* is a span that represents the effect. Table 1 shows all 53 intra-sentence and inter-sentence causal cues used in this work. We extract all such patterns from a large web corpus, and after lemmatization, pair each word in A with each word in B to form a list of *causal pairs*. These pairs form a *directed* network of causal relations. Each node in this network is a lemmatized word, while a directed edge between two words u and v indicates a causal relation, e.g., $u \rightarrow v$. In this process, only pairs involving nouns, verbs, adjectives and adverbs from WordNet are included in the network. A fragment of the causal network with three words in the network is shown in Figure 1. Each edge is annotated with the *causal strength*, which will be defined next.

We choose to extract word pairs in a rather simplistic way, without deeper syntactic analysis, because i) we opt for breadth in the causal knowledge hence the input corpus is extremely large (around 10TB), and consequently deep parsing of the text becomes prohibitive; and ii) the sheer quantity of the word pairs thus obtained provides excellent statistics for us to distinguish true causal pairs against false ones.

2.2 Causal Strength Computation

Our causal strength metric can be seen as a variant of PMI (Pointwise Mutual Information), computed over a causal pair $u \rightarrow v$ and their frequencies extracted from the web corpus. We can omit the usual logarithm for ranking word pairs [27]

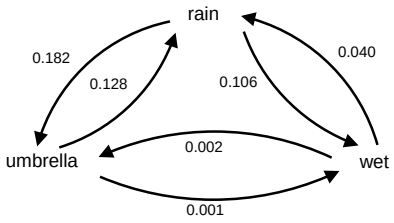

Figure 1: A fragment of causal network

To distinguish between a cause word and an effect word, we write u_c to denote u appeared in the cause span in the cue patterns, and u_e to denote u appeared in the effect span in the cue patterns. We derive the *causal strength* between two words as follows. First we define the joint probability of a causal pair as:

$$\begin{aligned} P(u \rightarrow v) &= P(u_c \mid v_e)P(v_e) \\ &= P(v_e \mid u_c)P(u_c) \end{aligned} \tag{1}$$

Then, we define a basic causal strength score CS_0 over the causal pairs as:

$$\begin{aligned} CS_0(u,v) &= \frac{P(u \rightarrow v)^2}{P(u_c)^2 P(v_e)^2} \\ &= \frac{P(u_c \mid v_e)P(v_e) \times P(v_e \mid u_c)P(u_c)}{P(u_c)^2 P(v_e)^2} \\ &= \frac{P(u_c \mid v_e)P(v_e|u_c)}{P(u_c)P(v_e)} \end{aligned} \tag{2}$$

The intuition for CS_0 is to take advantage of the conditional probability in both directions in Equation (1). We further generalize Equation (2) to include two tunable parameters α and β that have been effectively adopted for PMI to penalize high-frequency terms [3].

$$CS(u,v) = \frac{P(u_c \mid v_e)P(v_e \mid u_c)}{P^{\alpha}(u_c)P^{\beta}(v_e)} \tag{3}$$

where $P(u_c \mid v_e)$ and $P(v_e \mid u_c)$ can be computed as follows:

$$P(u_c \mid v_e) = \frac{f(u \rightarrow v)}{\sum_{w \in W} f(w \rightarrow v)} \tag{4}$$

$$P(v_e \mid u_c) = \frac{f(u \rightarrow v)}{\sum_{w \in W} f(u \rightarrow w)} \tag{5}$$

Here, $f(u \rightarrow v)$ is frequency of observing the causal pair from the corpus; W is the set of all words in the causal network. We compute the causal strength between every pair of words in the causal network according to Equation (3). Where an edge is missing in the network, we assign a causal strength of zero.

2.3 Intra-sentence Event Enhancement

Causality, in reality, exists between events, which are often expressed in more than one word. If we can detect events in a sentence, which are more likely to be involved in causal relations, we can reduce the noises produced by other unimportant words which are outside the events in that sentence

Table 1: 53 Causal cues. *A* is a cause span, and *B* is an effect span. DET stands for a/an/the/one. BE stands for is/are/was/were.

intra-sentence			inter-sentence		
A lead to B	A leads to B	A led to B	If A, then B	If A, B	B, because A
A leading to B	A give rise to B	A gave rise to B	B because A	B because of A	Because A, B
A given rise to B	A giving rise to B	A induce B	A, thus B	A, therefore B	B, A as a consequence
A inducing B	A induces B	A induced B	Inasmuch as A, B	B, inasmuch as A	In consequence of A, B
A cause B	A causing B	A causes B	B due to A	Due to A, B	B in consequence of A
A caused B	B caused by A	A bring on B	B owing to A	B as a result of A	As a consequence of A, B
A brought on B	A bringing on B	A brings on B	A and hence B	Owing to A, B	B as a consequence of A
B result from A	B resulting from A	B results from A	A, hence B	A, consequently B	A and consequently B
B resulted from A	the reason(s) for/of B BE A		A, for this reason alone , B		
DET effect of A BE B	A BE DET reason(s) of/for B				

and increase the accuracy of the causal reasoning. Moreover, it is observed that an event that contains causality between the words within itself appears to be more probable to cause other events in a "ripple effect". Consider these two examples.

- They *cut* the hamburger in *half*.
- The man *grew old*.

Events *cut* $\rightarrow$ *half* and *grow* $\rightarrow$ *old* exhibit positive causality. These events usually contain a strong action verb, which may lead to other consequences. In this subsection, we seek to identify events in the input sentences of commonsense causal reasoning task, and boost their weights in the computation of causal strength between two sentences.

To extract the events, which we treat as a set of words here, we parse the input sentence into dependency tree, and identify verbs, their direct objects as well as some of their significant modifiers. For example, from "I knocked on my neighbor's door," we can extract an event *knock-neighbor-door*. Only nouns, verbs, adjectives and adverbs are extracted as part of an event. It is possible to extract multiple events from a sentence, in which case we keep the one closest to the root of dependency tree. Algorithm 1 gives the details of this process, while $rel(u)$ denotes dependency relation of node u, the distance between an event to the root is calculated as:

$$\mathrm{RootDist(e)} = \frac{\sum_{\mathrm{w}\in\mathrm{e}} \mathrm{dist(Root, w)}}{|\mathrm{e}|} \quad (6)$$

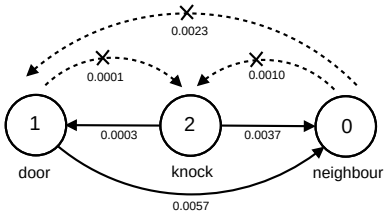

Figure 2: Weights in an event in a cause sentence

For the words in the extracted events, we perform the intra-sentence enhancement to strengthen the causal signal of some important words. Recall that in commonsense causal reasoning, we need to compute the overall causal strength between the premise and an alternative. It is known *a priori* whether the premise is a cause or an effect. Therefore, when we extract event from an input sentence, which can be either the

Algorithm 1 Events Extraction

Parse sentence S to obtain dependency parse tree T
Define dependency relation set:
    $R \leftarrow \{dobj, pobj, amod, nn, acomp, ccomp\}$
$EventSet \leftarrow \{\}$
$event \leftarrow \{\}$
for node $u \in T \ \wedge \ rel(u) \in R$ **do**
    $event \leftarrow event \cup \{u\}$
    while u has a dependency head v **do**
        $event \leftarrow event \cup \{v\}$
        if v is a verb **then**
            $EventSet \leftarrow EventSet \cup \{event\}$
            $event \leftarrow \{\}$
            break;
        else $u \leftarrow v$
for $e \in EventSet$ **do**
    if $\exists e^{'} \in EventSet, e^{'} \subseteq e$ **then**
        remove $e^{'}$ from $EventSet$
    if $\exists e^{'} \cap e \neq \emptyset$ **then**
        $e^{*} = \arg\min_{e} RootDist(e)$
        remove $e^{'}, e$ from $EventSet$
        $EventSet \leftarrow EventSet \cup \{e^{*}\}$

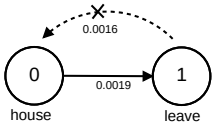

Figure 3: Weights of event in an effect sentence

premise or the alternative, we know whether the event is to be a cause, which we call *cause event*, or an effect, which we call *effect event*. We formulate events as light-weight patterns to refine the causality score between two sentences. In simple terms, it is a process of boosting weights of cause words in cause events and effect words in effect events, by assigning weights to words according to the the number of outgoing and incoming edges. In Figure 2, which shows an event *knock-neighbor-door* extracted from a cause sentence, there are originally six edges among the three words, each carries a *CS* score between the two adjacent words. For each pair of words in this sub-graph, we remove the edge with a smaller score, and effectively retain three edges (marked by solid lines) as a result. Then for each word, we count the number of *outgoing* edges (since this is a cause event), which acts as

the weight for that word. In this case, *knock* has weight 2, *neighbor* has weight 0, while *door* has a weight of 1. Similarly weights can be computed by counting the number of *incoming* edges for the effect event *leave-house* in Figure 3.

In the next subsection, when we compute the causal strength score between two sentences, the cause strength between any two words will be boosted by their respective weights which are computed here.

2.4 Commonsense Causal Reasoning

To compute whether alternative a_1 or a_2 is more plausible w.r.t. the premise p, we need to compare the overall causal strength $CS_T(p, a_1)$ and $CS_T(p, a_2)$, assuming p is asking for an effect. The overall causal strength score from text T_1 to text T_2 is computed as:

$$CS_T(T_1, T_2) = \frac{1}{|T_1| + |T_2|} \sum_{u \in T_1} \sum_{v \in T_2} (1 + \omega(u))(1 + \omega(v)) \delta(u, v) CS(u, v) \quad (7)$$

where $\omega(w)$ is the weight given by event enhancement for word w, and $\delta(u, v)$ is a penalty factor for the semantic ambiguity of u and v, defined as:

$$\delta(u, v) = \frac{1}{\#senses(u) + \#senses(v)} \quad (8)$$

where $\#senses(u)$ denotes number of WordNet synsets word u belongs to, which indicates how ambiguous u can be. We penalize ambiguous words in the same spirit as the inverse document frequency (IDF) in information retrieval. The reason is when an ambiguous word u is paired with every other word in another sentence, the causal strengths calculated for each pair may be due to different context and different senses of u and thus produce unreliable overall causal strength. Suppose we change the premise in Section 1 to

Premise: *My neighbor is a doctor*.

When computing the causal strength with Alternative 2, since the word *leave* is ambiguous, it means "depart" when paired with *neighbor*, whereas it means "absence from duty (medical)" when paired with *doctor*.

In our causality model, we treat each cause word as a *trigger* and each effect word as a *response*. We modeled one-to-many relationship between triggers and responses, which means one trigger can cause many responses, similarly one response can be caused by many triggers. That's the reason we normalize causality score by $|T_1| + |T_2|$ not $|T_1| \times |T_2|$ presented in previous papers.

3 Experimental Results

In this section, we first give some statistics of our corpus and the extracted causal network, and then evaluate the quantity and quality of the cue patterns used in the extraction. We further compared the end-to-end results on COPA task with several previously published results. Finally, we evaluate our commonsense reasoning ability on two additional tasks using data from ConceptNet 4 to further showcase the power of our framework. A demo of our network as well as the test sets we used in this section at `http://202.120.38.146/causal`.

3.1 Data Set and Extraction of Causal Network

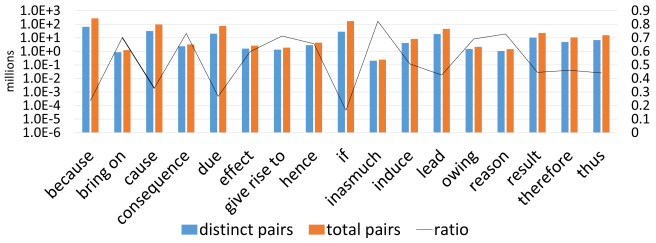

Figure 4: Number of (distinct) pairs extracted by Cues

We extracted our term causality network, which we call "CausalNet" for convenience in this section, from 1/10 of a commercial search engine web snapshot [1]. The snapshot was generated in February, 2013 and contains about 1.6 billion web pages with a combined of nearly 10TB. We extract 68,217,404 distinct word pairs from this corpus, which amounts to roughly 8GB. The number of unique lemmatized words in these pairs is 64,436, covering 41.49% (64,436/155,287) of the words in WordNet.

The 53 causal cues we used can be grouped into 17 sets, each containing cues of the same meaning or lemma form but with different tenses. Word pair distribution over these sets is shown in Figure 4. The blue bars (left) are the number of distinct pairs and the orange one (right) show the total number of pairs. Inter-sentence cues like "if" and "because" harvested the largest number of pairs. But more specific patterns such as "reason" and "inasmuch" find more diverse pairs, since the number of distinct pairs is relatively large compared to the total pairs extracted.

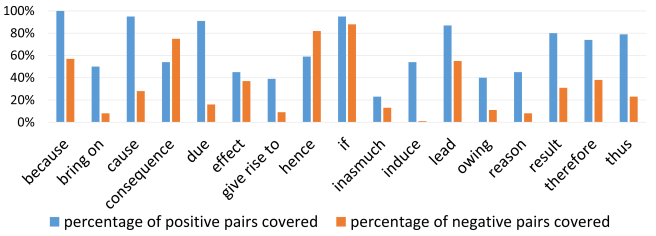

Figure 5: Positive vs. negative ConceptNet causal pairs covered by cues

To evaluate the quality of the causal cues, we make use of the manually labeled causal events in ConceptNet [20] as ground truth. ConceptNet 4 contains 74,336 unlemmatized English words, forming 375,135 unique concepts, which are connected by 610,397 relations. It is significantly smaller than our CausalNet in scale, especially in terms of number of edges. In otherwise, relations are more sparse in ConceptNet. We randomly collect 100 positive and 100 negative causal event pairs, based on feedbacks from human volunteers of OMCS project, casting positive votes for each *Causes* relationship that is causal, and negative for that is not. Since the pairs from ConceptNet contain phrases and not just words,

[1] Anonymized to honor double blind policy

we consider a pair is covered by a causal cue, if it extracted at least one pair of words from the web corpus, where the cause word appears in the cause event and effect word appears in the effect event in the ConceptNet pairs. Figure 5 shows that in general, our cues can effectively distinguish between positive and negative causal pairs, with the exception of "hence" and "consequence", both of which represent relatively coarse-grained entailment relation. Particularly good cues to distinguish the positive and negative pairs are "due" and "induce".

3.2 End-to-end Evaluation on COPA

COPA task consists of 1000 causal reasoning questions, divided into development question set and test question set of 500 each. We pruned the parameter α and β on the development set by attempting all combinations of values from 0.1 to 1.0 with a step of 0.1. The best combination turns out to be $\alpha = 0.4$, $\beta = 0.3$. All competing systems were assessed based on their accuracy on the 500 questions in the COPA test split [12].

Table 2 shows the results.

Table 2: COPA results comparison

Methods	Accuracy(%)
PMI Gutenberg (W=5)[24]	58.8%
UTDHLT Bigram PMI[10]	61.8%
UTDHLT SVM Combined[10]	63.4%
PMI 10M Stories (W=25)[11]	65.4%
CausalNet w/o events	67.6%
CausalNet w/ events	**68.8** %

PMI Gutenberg uses PMI statistic calculating from data in Project Gutenberg (16GB of English-language text). They pair the words from premise and alternative and choose the alternative with higher PMI. Their result is the best with a window size of 5. UTDHLT is the result of SemEval-2012 Task 7 systems. The team proposes two approaches for the task. The first one uses PMI over bigrams as a feature. For the second one, they treat it as a classification problem and combine the features of approach one with some other features to train an SVM model. Their PMI statistic is calculated from the LDC Gigaword corpus (8.4 million documents). The last PMI method, which was also the best performing method in the last 4 years, uses a larger corpus of personal stories (37GB of text) with a window of 25. There are two variants from our framework; the one without event detection does not detect any events or boost the strength of any word in the input sentence but merely use the causal network and the causal strength scores between words. Observe that our system with the event detection and boosting in Section 2.3, shown in bold, achieves 68.8% and outperforms all existing approaches.

3.3 Causality Detection

Another task is to investigate the following two research questions on our proposed network, using data from ConceptNet4.

- **RQ1:** For arbitrary event pair manually labeled as *causal* (positive) or *not causal* (negative), we investigate whether our proposed causality score clearly separates the two.
- **RQ2:** Inspired by COPA, we investigate positive and negative pair sharing the same premise, and investigate the accuracy of our selection of positive alternative.

For **RQ1,** we randomly collect 100 positive and 100 negative ground truth of causal Figure 6 shows the causality score (y-axis) of 100 positive and negative pairs indexed randomly (x-axis). We can observe that scores of positive and negative pairs are accurately distinguished by a linear function, such as $y = 10^{-2}$, indicated by the green line, with little overlap.

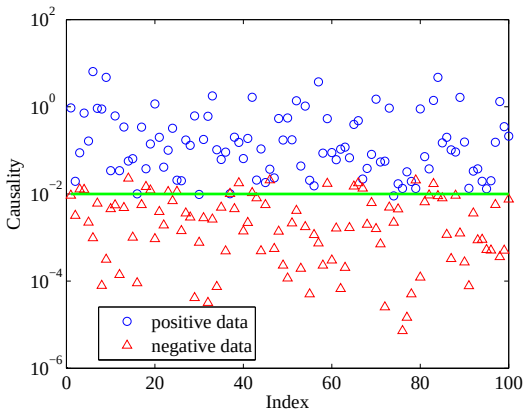

Figure 6: Distinguishing causality on ConceptNet

For **RQ2,** we test our network in a COPA-like setting of classifying between positive and negative pairs sharing the same premise. Due to sparsity of such pairs, we use *pseudo-disambiguation task* in [5]. In particular, we follow [5] to use *Causes* relationship (u, v) with positive votes, such that u is the shared premise and v is a positive alternative. We then generate a negative alternative by randomly selecting v' without *Causes* relationship with u. This approach is widely adopted in many tasks, as a large scale test cases can be generated, but as ConceptNet does not exhaustively label all possible causal relationships, randomly selected v' can be actually causal, or *false negatives* exist. In such situation, we removed the question involving such false negative, and consequently obtained a dataset of 412 questions in which 259 looks for an effect while 153 looks for a cause. Table 3 shows the result using our framework with event detection outperforming the other.

Table 3: Result of ConceptNet RQ2

Methods	Accuracy(%)
CausalNet w/o events	78.4%
CausalNet w/ events	**80.1%**

4 Related Work

We start by discussing previous work to extract causal relation term pairs from text and briefly mention the general task of relation extraction. Then we present various past attempts

to solve the commonsense causal reasoning problem. Common ingredients in these approaches are word association or similarity measures, which are discussed last.

4.1 Causal Relation Extraction

Previous work on causal relation extraction is relatively sparse. The existing approaches use hand-coded and domain-specific patterns to extract causal knowledge. Girju et al. [8] were the first to work on casual relation discovery between nominals. They semi-automatically extracted causal cues, but only extracted noun category features for the head noun. Chang et al. [2] developed an unsupervised method and utilized lexical pairs and cues contained in noun phrases as features to identify causality between them. Both of them ignored how the remaining causal text span between noun phrases affects the semantics. We proposed numeric features based on that, and get better results. Blanco et al. [1] used different patterns to detect the causation in long sentences that contain clauses. And most recently, Do et al. [4] introduced a form of association metric into causal relation extraction. They used discourse connectives and similarity distribution to identify event causality between predicate, not noun phrases, but achieved a F1-score around 0.47.

4.2 General Relation Extraction

Besides causal relations, much work has been done on extracting many other types of relations from text, e.g., hyponymy (isA) [6; 29]; meronymy (part-whole) [9], metaphor [19], relatedness [16] as well as general relations [30; 26; 14; 7; 23]. Relation extraction generally involves identifying the target terms or entities in text and then annotating the relations properly. Previous approaches are either supervised or semi-supervised.

Supervised approaches usually treat the extraction as a classification problem, where the input is the sentence with marked target entities/terms, and the output is the classification into one of the predefined relations or none. Marking the entities often relies on syntactic patterns or named entity recognition. These approaches require labeled data and hence cannot be easily extended to new types of relations. They also make heavy use of NLP tools such as POS tagger and dependency parser which are all error-prone. Semi-supervised method often starts with a seed set of entity pairs, and uses a bootstrapping strategy to accumulate more pairs either by gradually discovering contextual patterns that represent the target relation [6], or by using a fixed set of strong patterns and some logical rules to determine the plausibility of a pair in each iteration [29]. Our extraction of causal pairs is completely unsupervised. enables allows us to harness a web-scale evidences though with noises, which we eliminate using statistical evidences.

4.3 Commonsense Causal Reasoning

Commonsense causal reasoning is a grand challenge in artificial intelligence. Earlier attempts on the problem were largely linguistic, for example, developing formal theories to capture temporal or logical properties in causal entailment [18; 17]. These approaches were not effective due to the difficulty in handcrafting the theories for board-ranging open domain reasoning.

Recently, the NLP community has explored knowledge based approaches and show substantial potential. One approach toward this goal is to accrue common sense knowledge through crowdsourcing. A prominent example along this line is the Open Mind Common Sense (OMCS) project by MIT [25]. Some of the knowledge such as "effect of" relation in the ConceptNet [20] which is a sub-project under OMCS can be used to identify causal discourse in COPA task. However, the scale of such human curated knowledge suffers from scalability bottleneck. In fact, the ConceptNet is only a fraction of our causal network by size after 15 years of community efforts.

More successful efforts are centered around using correlational statistics [12] such as pointwise mutual information (PMI) between unigrams (words) or bigrams from large text corpora [21]. Corpora attempted include LDC gigaword news corpus [10], Gutenberg e-books [24], personal stories from Weblogs [11] and Wikipedia text [15]. Previous research show that the type of information source has significant impact on the accuracy of such knowledge based approach. This paper falls into this category of research, but instead proposed to compute a generalized PMI measure [27] not from the plain text corpus but from a causal relation graph induced from large web text. In addition, instead of fixing the target language units in the discourse sentences to either word or n-gram, we dynamically construct events which contain internal causality information and make use of these multi-word events in the computation of the final causal strength between two sentences.

4.4 Word Association Metrics

Our generalized causality is inspired from association strength measure proposed by Wettler et al. [28] and Washtell [27], which introduced parameter α being 0.66 and 0.5 respectively. Our causality strength considers both directions of the causality relation. Causality strength is similar to association strength to some degree, since association between term pair (u, v) which also asymmetrical treated u and v. We introduced an generalized formula by introducing α and β following the same intuition for discounting high frequency causes and effects respectively.

5 Conclusion

This paper proposes a novel framework of deducing commonsense causality by automatically harvesting a network of causal-effect terms extracted from a large web corpus. Such a network can achieve a high coverage including long-tailed causality relations. We then implemented an algorithm for detecting events in two input sentences and computing a causal strength between them. Evaluation shows that such a framework is capable of outperforming the previous best approach for solving the competitive SEMEVAL task known as COPA, and also shows a great potential in solving other related causality reasoning tasks.

References

[1] E. Blanco, N. Castell, and D. I. Moldovan. Causal relation extraction. In *LREC*, 2008.

[2] D. Chang and K. Choi. Causal relation extraction using cue phrase and lexical pair probabilities. In *IJCNLP*, pages 61–70, 2004.

[3] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990.

[4] Q. X. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification. In *EMNLP*, pages 294–303, 2011.

[5] K. Erk. A simple, similarity-based model for selectional preference. In *ACL*, 2007.

[6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall. In *WWW'04*, 2004.

[7] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545, 2011.

[8] R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 76–83, 2003.

[9] R. Girju, A. Badulescu, and D. I. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, 2006.

[10] T. Goodwin, B. Rink, K. Roberts, and S. M. Harabagiu. UTDHLT: COPACETIC system for choosing plausible alternatives. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 461–466, 2012.

[11] A. S. Gordon, C. A. Bejan, and K. Sagae. Commonsense causal reasoning using millions of personal stories. In *AAAI*, 2011.

[12] A. S. Gordon, Z. Kozareva, and M. Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, 2012.

[13] C. Havasi, R. Speer, K. C. Arnold, H. Lieberman, J. B. Alonso, and J. Moeller. Open mind common sense: Crowd-sourcing for common sense. In *AAAI Workshop*, 2010.

[14] J. Hoffart, F. M. Suchanek, K. Berberich, E. L. Kelham, G. de Melo, and G. Weikum. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *WWW'11*, 2011.

[15] S. Jabeen. Exploiting wikipedia semantics for computing word associations. 2014.

[16] K. Q. Z. Keyang Zhang and S.-W. Hwang. An association network for computing semantic relatedness. In *AAAI*, 2015.

[17] A. Lascarides and N. Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

[18] A. Lascarides, N. Asher, and J. Oberlander. Interferring discourse relations in context. In *ACL*, pages 1–8, 1992.

[19] H. Li, K. Q. Zhu, and H. Wang. Data-driven metaphor recognition and explanation. *TACL*, 1:379–390, 2013.

[20] H. Liu and P. Singh. Commonsense reasoning in and over natural language. In *Knowledge-based intelligent information and engineering systems*, pages 293–306. Springer, 2004.

[21] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 775–780, 2006.

[22] G. A. Miller. WordNet: A lexical database for english. *Commun. ACM*, 38(11), 1995.

[23] N. Nakashole, G. Weikum, and F. M. Suchanek. PATTY: A taxonomy of relational patterns with semantic types. In *EMNLP-CoNLL*, pages 1135–1145, 2012.

[24] M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[25] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002.

[26] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of the semantic knowledge. In *WWW'07*, 2007.

[27] J. Washtell and K. Markert. A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *EMNLP*, pages 628–637, 2009.

[28] R. R. Wettler, M. Computation of word associations based on the co-occurrences of words in large corpora. In *the 1st Workshop on Very Large Corpora*, 1993.

[29] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD'12*, 2012.

[30] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland. Textrunner: Open information extraction on the web. In *NAACL HLT Demonstration Program*, 2007.