Relation extraction and similarity for commonsense causal reasoning

Ben Hixon
CSE 573
University of Washington
bhixon@cs.washington.edu

Abstract

Causal reasoning is the process of recognizing that two propositions are connected by cause-and-effect. Humans make these judgments easily but machines are less adept. Keyword co-occurrence in text corpora has been shown to provide limited support for causal reasoning. We use open information extraction to construct a knowledge base of frequently co-occurring relations, and use relation pair similarity scores to map propositions from the Choice of Plausible Alternatives task to relations in the knowledge base. Our hypothesis is that co-occurring relations are more likely than random relations to be causally connected. Our best performing method mines a small subset of the Gigaword corpus to answer a small number of 13 COPA questions with 62% accuracy. This is close to the highest reported accuracy but at a trade-off of very small coverage over the COPA questions. Our methods may scale to larger corpora for improved coverage.

1 Introduction

The goal of a causal commonsense reasoning system is to infer a causal connection between two everyday-language propositions. The causal commonsense reasoning problem has been formalized in the Choice of Plausible Alternatives task (Roemmele, Bejan, & Gordon, 2011), a set of 1000 cause-and-effect questions in which the objective is to choose between two alternative propositions given a premise. The premise can either be a *cause* in which case the choice is between alternative effects, or else the premise is an *effect* in which case the choice is between alternative causes. Both types of questions are evenly represented.

Premise: My favorite song came on the radio. *What happened as a result?*
Alternative 1: I covered my ears.
Alternative 2: **I sang along to it.**

Table 1. Example COPA question.

Humans achieve 99% accuracy on this task, but the state-of-the-art algorithm achieves only 65% accuracy while random choice achieves 50% accuracy. The state-of-the-art algorithm (Gordon, Bejan, & Sagae, 2011) uses pointwise mutual information between keywords in a large corpus of personal stories. That method relies on unstructured keyword knowledge. Structured knowledge may do better. The thesis of this paper is that causal information is embedded in *relations,* and that given a large enough text corpus, frequent co-occurrence of two relations is evidence that the propositions from which the relations were extracted are causally connected.

We use OLLIE (Mausam et al., 2012), a state-of-the-art open information extraction (Open IE) system, to extract relations from both the set of COPA questions and from the Gigaword corpus[1]. Relation extraction finds the semantic relations between entities in text. Open IE finds relations in open-domain free text. OLLIE finds only binary relations, but does not require a fixed ontology. Our hypothesis is that OLLIE extractions that more frequently co-occur in a large text corpus will be more causally connected. For example, if the rela-

[1] Gigaword relation extractions were kindly provided by Niranjan Balasubramanian of the University of Washington

tions (I; poured; coffee), (I; added; milk) occur close together in a corpus more often than the relations (I; poured; coffee), (I; voted for; Obama), then the system should infer that (I; added; milk) has the closer causal connection and is the more plausible outcome to (I; poured; coffee).

Open IE has high precision but low recall. We are unlikely to find many instances of relation pairs in Gigaword that both occur close together and that also exactly match a COPA (premise, alternative) relation pair. We solve this problem with a *relation pair similarity score.* Our hypothesis then becomes that a pair of propositions are causally connected given the frequent co-occurrence of relation pairs *similar* to the pair extracted from the propositions. For example, if the relations (Princess Di; was; famous) and (The press; chased her) co-occur in a large corpus, then it should lend support to the claim of a causal connection between relations in the *similar* relation pair (the woman; became; famous) and (photographers; followed; her) which may not occur in the corpus.

To find calculate relation pair similarity, we consider several similarity metrics. We calculate *coverage*, the number of COPA questions the system attempt to answer, and *accuracy*, the percent of covered questions answered correctly.

2 Methodology

The question we want to answer is whether the frequency of co-occurrence of relation pairs similar to COPA relation pairs can be used to choose the alternative that correctly follows from a premise. We investigate the performance of different semantic similarity scores on this task.

2.1 Relation extraction

Each COPA question consists of three sentences. I considered only the 500 questions in which the premise is a cause and the choice is between two alternative effects. OLLIE can extract multiple relations from a single sentence with an accompanying confidence measure. We consider all extracted relations from each sentence to be of equal weight. We restrict our attention only to COPA questions in which OLLIE successfully extracted relations from all three propositions. We refer to these relations as *Copa,* and a pair of (premise, alternative) relations as a COPA relation pair. Relation extraction with OLLIE on the Gigaword corpus produced a set of Y relations. For time constraints, we restricted the Gigaword corpus to only 2500 articles, from which OLLIE extracted 142,374 relations. We refer to this smaller set of relations as *Giga*.

2.2 Relation pre-processing

Each relation needs to be processed in order to reduce noise. My goal in processing relations is to map each relation argument onto its smallest meaningful semantic unit so that relations with identical or similar arguments can be grouped together. For example, consider the relations (Princess Di; drank; her coffee) and (her highness Diana; is drinking from; her coffee cup). We want to match their arguments together: the computer needs to learn that "Princess Di" and "her highness Diana" are semantically 'close enough.'

I pre-processed each relation in *Giga* and *Copa* in the following ways. I split the words on whitespace into tokens, removed any capitalization, punctuation, and personal pronouns, and dropped all articles and prepositions.

I also expanded male and female named entities. Gigaword contains many named entities and the particular entity often makes no difference to the argument's semantic role, so I wanted to expand named entities with their semantic class. Two relations with arguments Princess Di and Princess Elizabeth are semantically indistinguishable in causal commonsense use; without *context*, the particular name is irrelevant and the only feature of semantic importance is the fact that it is a female name. The expansion I used here was simply to find male and female names and append 'male' to the argument that contains a male name and 'female' to the argument that contains a female name. For example, the relations (Princess Di; drank; her coffee) becomes (princess di female; drank; coffee), which can now be mapped to other relations with female names expanded in their arguments. I compare results with and without name expansion. I used the 1219 most frequent male first names and 4275 most frequent female first names in the United States Census. This idea is similar to query expansion (Mitra, Singhal, & Buckley,

1998) in which a keyword query is expanded to include synonyms or other relevant terms. Future work will perform more sophisticated category expansions. For example, a relation argument that contains the word 'Microsoft' can be expanded to contain the words 'company' or 'corporation.'

After processing Giga relations, I found every co-occurring Giga relation pair semantically similar to a COPA (premise, alternative) relation pair. *Semantic similarity* of relation pairs is defined in the next section. I defined a Giga relation pair to be co-occurring if each relation in the pair was extracted within two sentences of each other in the Gigaword corpus. This very inclusive definition of co-occurrence was designed to combat low-recall, as more stringent definitions yielded few relation pairs similar to a COPA relation pair.

2.3 Relation Pair Semantic Similarity Scores

Open information extraction frequently exhibits low recall. Low recall and a finite corpus together make it unlikely that we will find many relation pairs in Gigaword that exactly match a COPA (premise, alternative) relation pair. We therefore find relation pairs *similar* to COPA relation pairs, and infer that high frequency of co-occurrence of relation pairs similar to a target COPA relation pair implies that the target COPA relation pair is causally connected. Two relation pairs are similar if the corresponding component relations are each similar. We denote relation semantic similarity between relations R and S by R ~ S, where relations R and S each have the form `(arg1; predicate; arg2)`. We consider multiple measures of semantic similarity between relations R and S.

With *Partial Argument Identity (PAI)*, R ~ S if both predicates share a common word and at least one pair of corresponding arguments also share a common word. Because of pre-processing, common words between arguments are expected to be semantically meaningful. *Full-argument identity*, for which R ~ S if all three corresponding component pairs share a common word, was not used because it produced zero coverage. In the 2500-article subset of Gigaword, only 1202 of the 142,374 relations were similar by full-argument identity to a single COPA relation. Of these, there were no co-occurring relations R and S for which R was similar to some COPA premise and S similar to a valid alternative to that premise.

With *Wordnet Partial Argument Similarity (WPAS),* R ~ S if in each corresponding argument pair there exists a word pair with a high enough Wordnet similarity. For every pair of words in a corresponding argument pair, I found the WUP similarity (Wu & Palmer, 1994) between their wordnet synsets if they had any synsets. For words with multiple synsets I took the highest WUP distance over all synset comparisons. The arguments were considered similar if they each had a word pair with WUP > 0.5, where 1.0 indicates word identity. A further elaboration on this score could POS-tag the sentence to identify the word's part of speech and use that information to choose between the appropriate synset. I used the Python NLTK Wordnet module's implementation of WUP.

Finally, with *Wordnet Full Argument Similarity (WFAS)*, R ~ S whenever all three pairs of corresponding components contain a common word pair with a WUP score > 0.5.

2.4 Results

For each semantic similarity score we report on *accuracy*, the percent of COPA questions successfully predicted, and *coverage*, the number of COPA questions on which an attempt could be made. A COPA question consists of two relation pairs, (premise, alternative1) and (premise, alternative2). If *Giga* contains more co-occuring relation pairs similar to one of these relation pairs than to the other, then the COPA question is covered.

	PAI
-name exp, 55K rels	Coverage: 17 Accuracy: 53%
+name exp, 55K rels	Coverage: 26 Accuracy: 46%
-name exp, 142K rels	Coverage: 22 Accuracy: 59%

Table 2. Results for Partial Argument Identity (PAI) with and without name expansion on corpora of 55748 and 142374 *Giga* relations.

I tried partial argument identity (PAI) on two subsets of Giga: the 1000-article corpus and the 2500-article corpus. On the 1000-article corpus, containing 55,748 distinct relations, *PAI* found enough co-occurrences to answer 17 questions. It achieved 53% accuracy on these. Using name expansion increased coverage to 26 questions but lowers accuracy to 46% (worse than random).

On the 2500-article subcorpus, with 142,374 distinct Giga relations, *PAI* found enough co-occurrence information to answer 22 COPA questions and achieved 59% accuracy.

For Wordnet similarity, I further restricted the corpus to the first 25,000 relations extracted in Giga. The Python NLTK WUP function ran slowly; to account for this I pre-calculated the WUP similarity between every pair of distinct words ahead of time and writing the dictionary to file (resulting in a 600 megabyte record of the WUP similarities for all distinct word pairs in the first 25K relations). Without name expansion and using wordnet partial argument similarity (*WPAS*), with WUP>0.5, the coverage over the COPA questions was significant even with such a small subset of Gigaword. We found 20,630 distinct relations similar to a COPA relation. Among these, 66,517 pairs (of the 20630^2 possible pairs) were extracted within two sentences of each other and so were considered co-occurring. Of these co-occurring relation pairs, 37,772 mapped to a COPA (premise, alternative) pair and were therefore helpful in making a decision. 88 questions were answerable, with accuracy of 53%.

Requiring that the WUP similarity score exceed 0.75 significantly reduced coverage without an increase in accuracy. Only 15 questions were answerable (that is, co-occurring relation pairs in Giga only mapped to 15 COPA questions), and the system was able to answer 8 of these correctly for a 53% accuracy. Name expansion also fared poorly. For both WPAS and WFAS, name expansion increased coverage but reduced performance.

	WPAS	WFAS
WUP> 0.5, -name exp	Coverage: 88 Accuracy: 53%	Coverage: 13 Accuracy: 62%
WUP>0.5, +name exp	Not performed	Coverage: 29 Accuracy: 52%
WUP>0.75, -name exp	Coverage: 15 Accuracy: 53%	Not performed

Table 3. Wordnet similarity results with and without name expansion and with varying strictness of Wu-Palmer score on 25K relations. Partial argument similarity (WPAS) always has greater coverage because it only requires a predicate and one argument to be similar.

Wordnet full argument similarity with WUP > 0.5 and without name expansion used the first 25K relations to answer 13 COPA questions. Of these, 8 were correct for an accuracy of 62%. Fewer COPA questions were answerable because we restricted the coverage by mandating all three components (both arguments as well as the predicate) each contain a word pair with WUP>0.5. Scaling to larger subsets of Gigaword or to other corpora such as the Gutenberg story corpus may improve coverage.

3 Conclusion and Future Work

The methods did not work as well as keyword co-occurrence. Only one of our methods performed competitively, Wordnet Full Argument similarity without name expansion on 25 thousand Giga relations. Its trade-off is very small coverage, only answering a total of 13 COPA questions and getting 8 correct. Larger corpus size may improve coverage. Contrary to my expectations, name expansion decreased performance in all cases. I conjecture that not all names are semantically equivalent to the class they represent, and that the expansion of names to their classes introduced extra noise.

Lengthy running time made debugging difficult and prevented data collection on all runs. Another technical difficulty, which I don't feel I adequately overcame, is *argument matching:* how to map relation arguments to semantic units in order to match them together. Relation extraction often simply segments a sentence, leaving non-contributing tokens such as prepositions in place. This makes it difficult to identify relations with similar arguments, such as "the princess" and "her highness". Lexical noise that is not semantically meaningful needs to be removed from each argument. A classifier could be built to classify matching arguments.

Future work will consider improved methods to map Giga relations onto similar Copa relations, better named entity expansion, and faster algorithms to scale to larger corpora. A better mapping from relation argument to semantic unit is also required, perhaps using lambda calculus or other sophisticated semantic parsing. A clustering approach such as k-means may be effective at finding clusters of semantically similar relations, so that co-occuring relations can be found within each cluster. However, this would require the onerous computation of a similarity weight between every pair of relations in a large corpus.

This task can objectively compare the performance of different relation extraction engines. The more accurate relation extraction engine produces a knowledge base of co-occurring relations that more accurately answers the COPA questions.

4 Related Work

The Choice of Plausible Alternatives task was proposed by Roemmele, Bejan, and Gordon (2011). The best performing algorithm, which uses pointwise mutual information between keywords drawn from a large corpus of personal story blogs, was presented by Gordon, Bejan, and Sagae (2011). They found that keyword PMI outperformed more a sophisticated method that identifies temporally related clauses using rhetorical discourse theory, and that a personal story corpus generated from weblogs worked better than a corpus derived from Project Gutenberg. COPA was also a SemEval 2012 task (Gordon, Kozareva, & Roemmele, 2012). Only one team completed the task in the two-week period (Goodwin et al., 2012). They built an SVM to classify a COPA (premise, alternative) pair as causally connected or not. Its features include parts of speech and syntactic dependencies found by the Stanford parser, as well as event extractions and mutual information between bigrams. In spite of this complexity, their classifier achieved only 62% accuracy.

Several methods have considered relation similarity. Turney (2006) describes Latent Relational Analysis that finds similarity between analogical relations such as *mason:stone::carpenter:wood.* Nakov and Hearst (2008) also solve these types of analogical relation similarities; given an example relation such as *mason:stone,* they collect a web corpus from Google search results on that pair of nouns and create feature vectors with lexical information extracted from the search result corpus.

Talukdar, Wijaya, and Mitchell (2012) find temporal constraints between relations. They consider a *domain*, a set of relations with a matching argument *type*. For example, *actedIn(person, film)* and *wonPrize(film, award)* belong to a 'film' domain. Their graph-based algorithm, *GraphOrder*, finds all temporal constraints between relations in a given domain. Temporal constraints are estimated values for two weights, TBefore and TSimultaneous, which respectively represent a before-after temporal relationship and a simultaneous temporal relationship between two relations in a domain.

Rel-grams (Balasubramanian et al., 2012) are n-gram frequencies of relations over large corpora and could conceivably be applied to this problem. However, they would also be subject to the same problem of argument matching, and they also do not account for similarity between relation pairs.

References

Balasubramanian, N., Soderland, S., Mausam, & Etzioni, O. (2012). Paper presented at the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX '12).

Goodwin, T., Rink, B., Roberts, K., & Harabagiu, S. M. (2012). *UTDHLT: COPACETIC system for choosing plausible alternatives*. Paper presented at the Proceedings of the First Joint Conference on Lexical and Computational Semantics.

Gordon, Bejan, & Sagae. (2011). *Commonsense causal reasoning using millions of personal stories*. Paper presented at the Twenty-Fifth Conference on Artificial Intelligence (AAAI '11).

Gordon, Kozareva, & Roemmele. (2012). *SemEval-2012 task 7: choice of plausible alternatives*. Paper presented at the Proceedings of the First Joint Conference on Lexical and Computational Semantics.

Mausam, Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). *Open language learning for information extraction*. Paper presented at EMNLP-CoNLL '12.

Mitra, M., Singhal, A., & Buckley, C. (1998). *Improving automatic query expansion.* Paper presented at the 21st annual international ACM SIGIR conference on Research and development in information retrieval.

Nakov, P., & Hearst, M. A. (2008). *Solving Relational Similarity Problems Using the Web as a Corpus*. Paper presented at the Proceedings of ACL-08: HLT.

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). *Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning*. Paper presented at the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning.

Talukdar, P. P., Wijaya, D., & Mitchell, T. (2012). *Acquiring temporal constraints between relations*. Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Turney, P. D. (2006). Similarity of Semantic Relations. *Computational Linguistics, 32*(3), 379--416. doi: 10.1162/coli.2006.32.3.379

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection*. Paper presented at the 32nd annual meeting on Association for Computational Linguistics (ACL '94).