UTDHLT: COPACETIC System for Choosing Plausible Alternatives

Travis Goodwin, Bryan Rink, Kirk Roberts, Sanda M. Harabagiu
Human Language Technology Research Institute
University of Texas Dallas
Richardson TX, 75080
{travis,bryan,kirk,sanda}@hlt.utdallas.edu

Abstract

The Choice of Plausible Alternatives (COPA) task in SemEval-2012 presents a series of forced-choice questions wherein each question provides a premise and two viable cause or effect scenarios. The correct answer is the cause or effect that is the most plausible. This paper describes the COPACETIC system developed by the University of Texas at Dallas (UTD) for this task. We approach this task by casting it as a classification problem and using features derived from bigram co-occurrences, TimeML temporal links between events, single-word polarities from the Harvard General Inquirer, and causal syntactic dependency structures within the gigaword corpus. Additionally, we show that although each of these components improves our score for this evaluation, the difference in accuracy between using all of these features and using bigram co-occurrence information alone is not statistically significant.

1 The Problem

"The surfer caught the wave." This statement, although almost tautological for human understanding, requires a considerable depth of semantic reasoning. What is a surfer? What does it mean to "catch a wave"? How are these concepts related? What if we want to ascertain, given that the surfer caught the wave, whether the most likely next event is that "the wave carried her to the shore" or that "she paddled her board into the ocean"? This type of causal and temporal reasoning requires a breadth of world-knowledge, often called commonsense understanding.

Question 15 (Find the `EFFECT`)
Premise: I poured water on my sleeping friend.
Alternative 1: My friend awoke.
Alternative 2: My friend snored.

Question 379 (Find the `CAUSE`)
Premise: The man closed the umbrella.
Alternative 1: He got out of the car.
Alternative 2: He approached the building.

Figure 1: An example of each type of question, one targeting an effect, and another targeting a cause.

The seventh task of SemEval-2012 evaluates precisely this type of cogitation. COPA: Choice of Plausible Alternatives presents 1,000[1] sets of two-choice questions (presented as a premise and two alternatives) provided in simple English sentences. The goal for each question is to choose the most plausible cause or effect entailed by the premise (the dataset provided an equal distribution of cause and effect targetting questions). Additionally, each question is labeled so as to describe whether the answer should be a cause or an effect, as indicated in Figure 1.

The topics of these questions were drawn from two sources:

1. Randomly selected accounts of personal stories taken from a collection of Internet weblogs (Gordon and Swanson, 2009).
2. Randomly selected subject terms from the Library of Congress Thesaurus for Graphic Materials (of Congress. Prints et al., 1980).

Additionally, the incorrect alternatives were authored

[1]This data set was split into a 500 question development (or training) set and a 500 question test set.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 461–466,
Montréal, Canada, June 7-8, 2012.

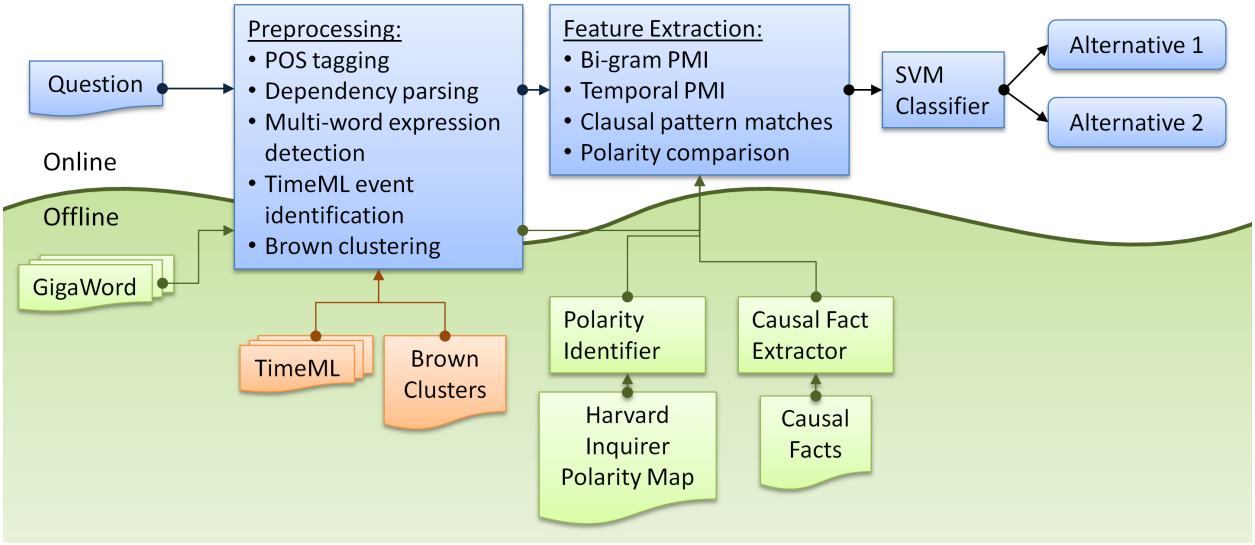

Figure 2: Architecture of the COPACETIC System

with the intent of impeding "purely associative methods" (Roemmele et al., 2011). The task aims to evaluate the state of commonsense causal reasoning (Roemmele et al., 2011).

2 System Architecture

Given a question, such as Question 15 (as shown in Figure 1), our system selects the most plausible alternative by using the output of an SVM classifier, trained on the 500 provided development questions and tested on the 500 provided test questions. The classifier operates with features describing information extracted from the processing of the question's premise and alternatives. As illustrated by Figure 2, the preprocessing involves part of speech (POS) tagging, and syntactic dependency parsing provided by the Stanford parser (Klein and Manning, 2003; Toutanova et al., 2003), multi-word expression detection using Wikipedia, automatic TimeML annotation using TARSQI (Verhagen et al., 2005; Pustejovsky et al., 2003), and Brown clustering as provided in (Turian, 2010).

The architecture of the COPACETIC system is divided into offline (independent of any question) and online (question dependent) processing. The online aspect of our system inspects each question using an SVM and selects the most likely alternative. Our system's offline functions focus on pre-processing resources so that they may be used by components of the online aspect of our system. In the next section, we describe the offline processing upon which our system is built, and in the following section, the online manner in which we evaluate each question.

2.1 Offline Processing

Because the questions presented in this task require a wealth of commonsense knowledge, we first extracted commonsense and temporal facts. This subsection describes the process of mining this information from the fourth edition of the English Gigaword corpus[2] (Parker et al., 2009).

We collected commonsense facts by extracting cause and effect pairs using twenty-four hand-crafted patterns. Rather than lexical patterns, we used patterns over syntactic dependency structures in order to capture the syntactic role each word plays. Figure 3 illuminates two examples of the dependency structures encoded by our causal patterns. Causal Pattern 1 captures all cases of causality indicated by the verb *causes*, while Causal Pattern 2 illustrates a more sophisticated pattern, in which the phrasal verb *brought on* indicates causality.

In order to extract this information, we first parsed the syntactic dependence structure of each sentence using the Stanford parser (Klein and Manning, 2003). Next, we loaded each sentence's dependence tree

[2]The LDC Catalog number of the English Gigaword Fourth Edition corpus is LDC2009T13.

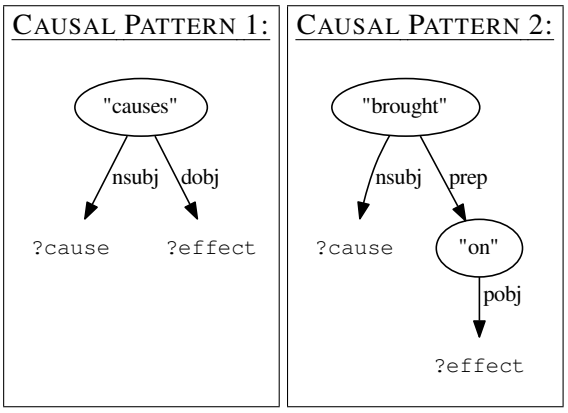

Figure 3: The dependency structures associated with the causal patterns: `?cause` "causes" `?effect`, and `?cause` "brought on" `?effect`.

into the RDF3X (Neumann and Weikum, 2008) implementation of an RDF[3] database. Then, we represented our dependency structures using in the SPARQL[4]query language and extracted cause and effect pairs by issuing SPARQL queries against the RDF3X database. We used SPARQL and RDF representations because they allowed us to easily represent and reason over graphical structures, such as those of our dependency trees.

It has been shown that causality often manifests as a temporal relation (Bethard, 2008; Bethard and Martin, 2008). The questions presented in this task are no exception: many of the alternative-premise pairs necessitate temporal understanding. For example, consider question 63 provided in Figure 4.

Question 63 (Find the `EFFECT`)
Premise: The man removed his coat.
Alternative 1: He entered the house.
Alternative 2: He loosened his tie.

Figure 4: Example question 63, which illustrates the necessity for temporal reasoning.

[3]The Resource Description Framework (RDF) is is a specification from the W3C. Information on RDF is available at http://www.w3.org/RDF/.

[3]The SPARQL Query Language is defined at `http://www.w3.org/TR/rdf-sparql-query/`. An examples of the `WHERE` clause for a SPARQL query associated with the *brought on* pattern from Figure 3 is provided below:

```
{   ?a   <nsubj>      ?cause ;
         <token>   "brought" ;
         <prep>         ?b .
    ?b   <token>      "on" ;
         <pobj>    ?effect .   }
```

In order to extract this temporal information, we automatically annotated our corpus with TimeML annotations using the TARSQI Toolkit (Verhagen et al., 2005). Unfortunately, the events represented in this corpus were too sparse to use directly. To mitigate this sparsity, we clustered events using the 3,200 Brown clusters[5] described in (Turian, 2010).

After all such offline processing has been completed, we incorporate the knowledge encoded by this processing in the online components of our system (online preprocessing, and feature extraction) as described in the following section.

2.2 Online Processing

We cast the task of selecting the most plausible alternative as a classification problem, using a support vector machine (SVM) supervised classifier (using a linear kernel). To this end, we pre-process each question for lexical information. We extract parts of speech (POS) and syntactic dependencies using the Stanford CoreNLP parser (Klein and Manning, 2003; Toutanova et al., 2003). Stopwords are removed using a manually curated list of one hundred and one common stopwords; non-content words (defined as words whose POS is not a noun, verb, or adjective) are also discarded. Additionally, we extract multi-word expressions (noun collocations[6] and phrasal verbs[7]). Finally, in order to utilize our offline TimeML annotations, we extract events using POS. Examples of the retained content words are underlined in Figures 5, 6, 7 and 8.

After preprocessing each question, we convert it into two premise-alternative pairs (PREMISE-ALTERNATIVE1, and PREMISE-ALTERNATIVE2). For each of these pairs, we attempt to form a bridge from the causal sentence to the effect sentence, without distinction over whether the cause or effect originated from the premise or the alternative. This bridge is provided by four measures, or features, described in the following section.

[5]These clusters are available at `http://metaoptimize.com/projects/wordreprs/`.

[6]These were detected using a list of English Wikipedia article titles available at `http://dumps.wikimedia.org/backup-index.html`.

[7]Phrasal verbs were determined using a list available at `http://www.learn-english-today.com/phrasal-verbs/phrasal-verb-list.htm`.

3 The Features of the COPACETIC System

In determining the causal relatedness between a cause and an effect sentence, we utilize four features. Each feature calculates a value indicating the perceived strength of the causal relationship between a cause and an effect using a different measure of causality. The four features used by our COPACETIC system are described in the following subsections.

3.1 Bigram Relatedness

Our first feature measures the degree of relatedness between all pairs of bigrams (at the token level) in the cause and effect pair. We do this by calculating the point-wise mutual Information (PMI) (Fano, 1961) for all bigram combinations between the candidate alternative and its premise in the English Gigaword corpus (Parker et al., 2009) as shown in Equation 1.

$$PMI(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Under the assumption that distance words are unlikely to causally influence each other, we only consider co-occurrences within a window of one hundred tokens when calculating the joint probability of the PMI. Additionally, we allow for up to two tokens to occur within a single bigram's occurrence (e.g. the phrase *pierced her ears* would be considered a match for the bigram *pierced ears*). Although these relaxations skew the values of our calculated PMIs by artificially lowering the joint probability, we are only concerned with how the values compare to each other. Note that because we employ no smoothing, the PMI of an unseen bigram is set to zero. The maximum PMI over all pairs of bigrams is retained as the value for this feature. Figure 5 illustrates this feature for Question 495.

Question 495 (Find the `EFFECT`)
Premise: The girl wanted to wear earrings.
Alternative 1: She got her ears pierced.
Alternative 2: She got a tattoo.

Alternative 1		Alternative 2	
PMI(wear earrings, pierced ears)	= -10.928	PMI(wear earrings, tattoo)	= -12.77
PMI(wanted wear, pierced ears)	= -13.284	PMI(wanted wear, tattoo)	= -14.284
PMI(girl wanted, pierced ears)	= -13.437	PMI(girl wanted, tattoo)	= -14.762
PMI(girl, pierced ears)	= -15.711	PMI(girl, tattoo)	= -14.859
Maximum PMI	= -10.928	Maximum PMI	= -12.77

Figure 5: Example PMI values for bigrams and unigrams (with content words underlined). Alternative 1 is correctly chosen as it has largest maxi mum PMI.

3.2 Temporal Relatedness

Although most of the questions in this task focus on causal relationships, for many questions, the nature of this causal relationship manifests instead as a temporal one (Bethard and Martin, 2008; Bethard, 2008). We use temporal link information from TimeML (Pustejovsky et al., 2005; Pustejovsky et al., 2003) annotations on our corpus to determine how temporally related a given cause and effect sentence are. This is accomplished by using the point-wise mutual information (PMI) between all pairs of events from the cause to the effect (see Equation 1). We define the relevant probabilities as follows:

- The joint probability ($P(x, y)$) of a cause and effect event is defined as the number of times the cause event participates in a temporal link ending with the effect event.
- The probability of a cause event ($P(x)$) is defined as the number of times the cause event precipitates a temporal link to any event.
- The probability of an effect event ($P(y)$) is defined as the number of times the effect event ends a temporal link begun by any event.

We define the PMI to be zero for any unseen pair of events (and for any pairs involving an unseen event). The summation of all pairs of PMIs is used as the value of this feature. Figure 6 shows how this feature behaves.

Question 468 (Find the `CAUSE`)
Premise: The dog barked.
Alternative 1: The cat lounged on the couch.
Alternative 2: A knock sounded at the door.

Alternative 1	Alternative 2
PMI(lounge, bark) = 5.60436	PMI(knock, bark) = 5.77867
	PMI(sound, bark) = 5.26971

Figure 6: Example temporal PMI values (with content words underlined). Alternative 2 is correctly chosen as it has the highest summation.

3.3 Causal Dependency Structures

We attempted to capture the degree of direct causal relatedness between a cause sentence and an effect sentence. To determine the strength of this relationship,

we considered how often phrases from the cause and effect sentences occur within a causal dependency structure. We detect this through the use of twenty-four[8] manually crafted causal patterns (described in Section 2.1). The alternative that has the maximum number of matched dependency structures with the premise is retained as the correct choice. Figure 7 illustrates this feature.

Question 490 (Find the EFFECT)
Premise: The man won the lottery.
Alternative 1: He became rich.
Alternative 2: He owed money.

Alternative 1	Alternative 2
won → rich = 15	won → owed = 5

Figure 7: Example casual dependency matches (with content words underlined). Alternative 1 is correctly selected because more patterns extracted "won" causing "rich" than "won" causing "owed".

3.4 Polarity Comparison

We observed that many of the questions involve the dilemma of determining whether a positive premise is more related to a positive or negative alternative (and vice-versa). This differs from sentiment analysis in that rather than determining if a sentence expresses a negative statement or view, we instead desire the overall sentimental connotation of a sentence (and thus of each word). For example, the premise from Question 494 (Figure 8) is "the woman became famous." Although this sentence makes no positive or negative claims about the woman, the word "famous" – when considered on its own – implies positive connotations.

We capture this information using the Harvard General Inquirer (Stone et al., 1966). Originally developed in 1966, the Harvard General Inquirer provides a mapping from English words to their polarity (POSITIVE, or NEGATIVE). For example, it denotes the word "abandon" as NEGATIVE, and the word "abound" as POSITIVE. We use this information by summing the score for all words in a sentence (assigning POSITIVE words a score of 1.0, NEGATIVE words a score of -1.0, and NEUTRAL or unseen words a score of 0.0). The difference between these scores between the cause sentence and the effect sentence is used as the value of this feature. This feature is illustrated in Figure 8.

[8]Twenty-four patterns was deemed sufficient due to time constraints.

Question 494 (Find the CAUSE)
Premise: The woman became famous.
Alternative 1: Photographers followed her.
Alternative 2: Her family avoided her.

Premise		Alternative 1		Alternative 2	
famous POSITIVE	1.0	follow NEUTRAL	0.0	avoid NEGATIVE	−1.0
		photographer NEUTRAL	0.0	family NEUTRAL	0.0
Sum	1.0	Sum	0.0	Sum	−1.0

Figure 8: Example polarity comparison (with content words underlined). Alternative 1 is correctly chosen as it has the least difference from the score of the premise.

4 Results

The COPA task of SemEval-2012 provided participants with 1,000 causal questions, divided into 500 questions for development or training, and 500 questions for testing. We submitted two systems to the COPA Evaluation for SemEval-2012, both of which are trained on the 500 development questions. Our first system uses only the bigram PMI feature and is denoted as bigram_pmi. Our second system uses all four features and is denoted as svm_combined. The accuracy of our two systems on the 500 provided test questions is provided in Table 1 (Gordon et al., 2012). On this task, accuracy is defined as the quotient of dividing the number of questions for which the correct alternative was chosen by the number of questions. Although multiple groups registered, ours were the only submitted results. Note that the difference in performance between our two systems is not statistically significant ($p = 0.411$) (Gordon et al., 2012).

Team ID	System ID	Score
UTDHLT	bigram_pmi	0.618
UTDHLT	svm_combined	0.634

Table 1: Accuracy of submitted systems

The primary hindrance to our approach is in combining each feature – that is, determining the confidence of each feature's judgement. Because the questions vary significantly in their subject matter and the nature of the causal relationship between given causes and effects, a single approach is unlikely

to satisfy all scenarios. Unfortunately, the problem of determining which feature best applies to a give question requires non-trivial reasoning over implicit semantics between the premise and alternatives.

5 Conclusion

This evaluation has shown that although commonsense causal reasoning is trivial for humans, it belies deep semantic reasoning and necessitates a breadth of world knowledge. Additional progress towards capturing world knowledge by leveraging a large number of cross-domain knowledge resources is necessary. Moreover, distilling information not specific to any domain – that is, a means of inferring basic and fundamental information about the world – is not only necessary but paramount to the success of any future system desiring to build chains of commonsense or causal reasoning. At this point, we are merely approximating such possible distillation.

6 Acknowledgements

We would like to thank the organizers of SemEval-2012 task 7 for their work constructing the dataset and overseeing the task.

References

[Bethard and Martin2008] S. Bethard and J.H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. *Proceedings of the 46th Annual Meeting of the ACL-HLT*.

[Bethard2008] S Bethard. 2008. Building a corpus of temporal-causal structure. *Proceedings of the Sixth LREC*.

[Fano1961] RM Fano. 1961. Transmission of Information: A Statistical Theory of Communication.

[Gordon and Swanson2009] A. Gordon and R. Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*.

[Gordon et al.2012] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. (2012) 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.

[Klein and Manning2003] D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

[Neumann and Weikum2008] Thomas Neumann and Gerhard Weikum. 2008. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*.

[of Congress. Prints et al.1980] Library of Congress. Prints, Photographs Division, and E.B. Parker. 1980. Subject headings used in the library of congress prints and photographs division. Prints and Photographs Division, Library of Congress.

[Parker et al.2009] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword Fourth Edition.

[Pustejovsky et al.2003] J Pustejovsky, J Castano, and R Ingria. 2003. TimeML: Robust specification of event and temporal expressions in text. *AAAI Spring Symposium on New Directions in Question-Answering*.

[Pustejovsky et al.2005] J Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, G. Katz, and I. Mani. 2005. The specification language TimeML. *The Language of Time: A Reader*.

[Roemmele et al.2011] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. *2011 AAAI Spring Symposium Series*.

[Stone et al.1966] P. J. Stone, D.C. Dunphy, and M. S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press.

[Toutanova et al.2003] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of NAACL-HLT*, pages 173–180. Association for Computational Linguistics.

[Turian2010] J Turian. 2010. Word representations: a simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the ACL*, pages 384–394.

[Verhagen et al.2005] M Verhagen, I Mani, and R Sauri. 2005. Automating Temporal Annotation with TARSQI. In *Proceedings of the ACL 2005*, pages 81–84.