SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning

Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
Los Angeles, CA
gordon@ict.usc.edu

Zornitsa Kozareva
Information Sciences Institute
University of Southern California
Marina del Rey, CA
kozareva@isi.edu

Melissa Roemmele
Department of Linguistics
Indiana University
Bloomington, IN
msroemme@gmail.com

Abstract

SemEval-2012 Task 7 presented a deceptively simple challenge: given an English sentence as a premise, select the sentence amongst two alternatives that more plausibly has a causal relation to the premise. In this paper, we describe the development of this task and its motivation. We describe the two systems that competed in this task as part of SemEval-2012, and compare their results to those achieved in previously published research. We discuss the characteristics that make this task so difficult, and offer our thoughts on how progress can be made in the future.

1 Motivation

Open-domain commonsense reasoning is one of the grand challenges of artificial intelligence, and has been the subject of research since the inception of the field. Until recently, this research history has been dominated by formal approaches (e.g. Lenat, 1995), where logical formalizations of commonsense theories were hand-authored by expert logicians and evaluated using a handful of commonsense challenge problems (Morgenstern, 2012). Progress via this approach has been slow, both because of the inherent difficulties in authoring suitably broad-coverage formal theories of the commonsense world and the lack of evaluation metrics for comparing systems from different labs and research traditions.

Radically different approaches to the commonsense reasoning problem have recently been explored by natural language processing researchers. Speer et al. (2008) describe a novel reasoning approach that applies dimensionality reduction to the space of millions of English-language commonsense facts in a crowd-sourced knowledge base (Liu & Singh, 2004). Gordon et al., (2010) describe a method for extracting millions of commonsense facts from parse trees of English sentences. Jung et al. (2010) describe a novel approach to the extraction of commonsense knowledge about activities by mining online how-to articles. We believe that these new NLP-based approaches hold enormous potential for overcoming the knowledge acquisition bottleneck that has limited progress in commonsense reasoning in previous decades.

Given the growth and enthusiasm for these new approaches, there is increasing need for a common metric for evaluation. A common evaluation suite would allow researchers to gauge the performance of new versions of their own systems, and to compare their approaches with those of other research groups. Evaluations for these new NLP-based approaches should themselves be based in natural language, and must be suitably large to truly evaluate the breadth of different reasoning approaches. Still, each evaluation should be focused on one dimension of the overall commonsense reasoning task, so as not to create a new challenge that no single research group could hope to succeed.

In SemEval-2012 Task 7, we presented a new evaluation for open-domain commonsense reason-

ing, focusing specifically on commonsense causal reasoning about everyday events.

2 Choice of Plausible Alternatives

Consider the following English sentence, describing a hypothetical state of the world:

The man lost his balance on the ladder.

In addition to parsing this sentence, resolving ambiguities, and constructing a semantic interpretation, human readers also imagine the causal antecedents and consequents that would follow if the statement were true. With such a brief description, readers are left with many questions. How high up on the ladder was this man? What was he doing on the ladder in the first place? How much experience does he have using ladders? Was he intoxicated? The answers to these questions help readers formulate hypotheses for the two central concerns when reasoning about events: *What was the cause of this?* and *What happened as a result?*

As computational linguists, we imagine that our automated natural language processing algorithms will also, eventually, need to engage in similar reasoning processes in order to achieve human-like performance on text understanding tasks. Progress toward the goal of deep semantic interpretation of text has been slow. However, the last decade of natural language processing research has shown that enormous gains can be achieved when there is a clear evaluation metric. A shared task with an automated scoring mechanism allows researchers to compare different approaches, tune system parameters to maximize performance, and assess progress toward broader research objectives. Developing an evaluation metric for causal reasoning poses a number of challenges. It is necessary to formulate a question with answers that can be automatically graded, but can still serve as a proxy for the complex, generative imagination of readers.

Roemmele et al. (2011) offered a solution in the form of a simple binary-choice question. Presented with an English sentence describing a premise, systems must select between two alternatives (also sentences) the one that more plausibly has a causal relation to the premise, as in the following example:

Premise: The man lost his balance on the ladder. *What happened as a result?*
Alternative 1: He fell off the ladder.
Alternative 2: He climbed up the ladder.

Both of these alternatives are conceivable, and neither is entailed by the premise. However, human readers have no difficulty selecting the alternative that is the more plausible of the two. This question asks about a causal consequent, and a complimentary formulation asks for the causal antecedent, as in the following example:

Premise: The man fell unconscious. *What was the cause of this?*
Alternative 1: The assailant struck the man on the head.
Alternative 2: The assailant took the man's wallet.

Roemmele et al. describe their efforts to author a collection of 1000 questions of these two types to create a new causal reasoning evaluation tool: the Choice of Plausible Alternatives (COPA). When presented to humans to select the correct alternative, the inter-rater agreement was extremely high (Cohen's kappa = 0.965). Where disagreements between two raters were found (in 26 of 1000 items), questions were removed and replaced with new ones with perfect agreement.

To develop an automated evaluation tool, the 1000 questions were randomly ordered and sorted into two equally sized sets of 500 questions to serve as development and test sets. The order of the correct alternative was also randomized, such that the expected accuracy of a random baseline would be 50%. Gold-standard answers for each split are used to automatically evaluate a given system's performance.

The distribution of the COPA evaluation includes an automated test of statistical significance of differences seen between two competing systems. This software tool implements a compute-intensive randomized test of statistical significance using stratified shuffling, as described by Noreen (1989). By randomly sorting answers between two systems over thousands of trials, this test computes the likelihood that differences as great as observed differences could be obtained by random chance.

The COPA evaluation is most similar in style to the Recognizing Textual Entailment challenge (Degan et al., 2006), but differs in its focus on causal implication rather than entailment. Instead of asking whether the interpretation of a sentence necessitates the truth of another, COPA concerns

the defeasible inferences that can be drawn from the interpretation of a sentence. In this respect, COPA overlaps in its aims with the task of recognizing causal relations in text through automated discourse processing (e.g. Marcu, 1999). Some progress in automated discourse processing has been made using supervised machine learning methods, where system learn the lexical-syntactic patterns that are most correlated with causal relations from a large annotated corpus (Sagae, 2009). Lacking a dedicated training corpus, the COPA evaluation encourages competitors to capture commonsense causal knowledge from any available corpus or existing knowledge repository.

3 SemEval-2012 Systems and Results

The COPA evaluation was accepted as Task 7 of the 6th International Workshop on Semantic Evaluation (SemEval-2012). In several respects, the COPA evaluation was different than the typical shared task offered as part of this series of workshops. First, the task materials were available and distributed long before the evaluation period began, and there were published results of previous systems using this evaluation.[1] Second, the task included no training data, only sets of development and test questions (500 each). Participants were encouraged to use any available text corpus or knowledge repositories in the construction of their systems. Success on the task would not be possible simply through the selection of machine learning algorithms and feature encodings. Instead, some creativity and ingenuity was needed to find a suitable source of commonsense causal information, and determine an automated mechanism for applying this information to COPA questions.

Only one team successfully completed the task and submitted results during the official two-week SemEval-2012 evaluation period. This team was Travis Goodwin, Bryan Rink, Kirk Roberts, and Sanda M. Harabagiu from the University of Texas at Dallas, Human Language Technology Research Institute. This team submitted results from two different systems (Goodwin et al., 2012), which they described to us as follows:

UTDHLT Bigram PMI: The team's first approach selects the alternative with the maximum Pointwise Mutual Information (PMI) statistic (Church & Hanks, 1990) over all pairs of bigrams (at the token level) between the candidate alternative and the premise. PMI statistics were collected using 8.4 million documents from the LDC Gigaword corpus (Graff & Cieri, 2003). A window of 100 terms was used for finding pairs of co-occurring bigrams, and a window/slop size of 2 for the bigram itself.

UTDHLT SVM Combined: The team's second approach augments the first by combining it with several other features and casting the task as a classification problem. To this end, they consider the PMI between events participating in a temporal link on a Time-ML annotated Gigaword corpus. That is, events that occur together frequently will have a higher PMI. They also consider the difference between the number of positive and negative polarity words between an alternative and premise using information from the Harvard Inquisitor. In addition, they used the count of matching cause-effect pairs extracted using patterns on dependency structures from the Gigaword corpus. Combining all of these sources of information, they trained a support vector machine (SVM) learning algorithm to classify the alternative that is most causally related to the premise.

These systems were assessed based on their accuracy on the 500 questions in the test split of the COPA evaluation, presented in Table 1. Both systems significantly outperformed the random baseline (50% accuracy), but the gains seen in the second approach were not significantly different than those of the first.

System	*Accuracy*
UTDHLT Bigram PMI	61.8%
UTDHLT SVM Combined	63.4%

Table 1. SemEval-2012 Task 7 system accuracy on 500 questions in the COPA test split

4 Comparison to Previous Results

In order to better evaluate the success of these two systems, we compared these results with the published results of other systems that have used the COPA evaluation. Three other systems were considered.

PMI Gutenberg (W=5): Described in Roemmele et al. (2011), this approach calculated the PMI between words (unigrams) in the premise and

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

each alternative, and selected the alternative with the stronger correlation. The PMI statistic was calculated using every English-language document in Project Gutenberg (16GB of text), using a window of 5 words.

PMI Story 1M (W=25): Described in Gordon et al. (2011), this approach was identical to that of Roemmele et al. (2011) except that the PMI statistic was calculated using a corpus of nearly one million personal stories extracted from Internet weblogs (Gordon & Swanson, 2009), with 1.9 GB of text. Using this corpus instead of Project Gutenberg, the best results were obtained by using a window of 25 words for the PMI statistic.

PMI Story 10M (W=25): Also described in Gordon et al. (2011), this approach explores the gains that can be achieved by calculating the PMI statistic using a much larger corpus of weblog stories. The story extraction technology used by Gordon and Swanson (2009) was applied to 621 million English-language weblog entries posted to the Internet in 2010 to create a corpus of 10.4 million personal stories (37GB of text). Again, the best results were obtained by using a window of 25 words for the PMI statistic.

Table 2 compares the results of these three previous systems with the two SemEval-2012 systems. Although the last two of these three previous systems achieved higher scores than both of the SemEval-2012 submissions, the differences are not statistically significant.

System	*Accuracy*
PMI Gutenberg (W=5)	58.8%
UTDHLT Bigram PMI	**61.8%**
UTDHLT SVM Combined	**63.4%**
PMI Story 1M (W=25)	65.2%
PMI Story 10M (W=25)	65.4%

Table 2. Comparison of SemEval-2012 Task 7 systems (in bold) with previously published results on the 500 questions in the COPA test split

5 Discussion

The two systems from the University of Texas at Dallas make an important contribution to progress on open-domain commonsense reasoning. Some lessons are evident from the short descriptions of their systems that they provided to us.

As in each of the previously successful systems, this team focused their efforts on calculating correlational statistics between words in COPA questions using very large text corpora. In this case, the Gigaword corpus is used, and the calculation is based on bigrams rather than unigrams. We believe that the content of the news articles that comprise the Gigaword corpus is a step further away from the concerns of COPA questions than both the Project Gutenberg corpus and the weblog story corpora used in previous efforts. Indeed, the gains achieved by Gordon et al. (2011) appear to be entirely due to the relationship between COPA questions and the personal stories that people write about in their public weblogs. However, the use of a large news corpus affords the use of more sophisticated analysis techniques that have been developed for this genre. Here, the Gigaword corpus is annotated using Time-ML relationships, which in turn are used to modify the PMI strength between words.

The use of bigrams is an additional enhancement explored by this team, as is the casting of COPA questions as a classification task using a diverse set of lexical and discourse features. Such an approach can facilitate the combining of diverse systems in the future, where correlational statistics are gathered from a diverse set of text corpora, each suited for specific domains of COPA questions or yielding complimentary feature sets.

Still, the modest COPA performance seen from all existing systems is somewhat discouraging. With the best systems performing in the 60-65% range, we remain much closer to random performance (50%) than human performance (99%). These results cast some doubt that the information necessary to answer COPA questions can be readily obtained from large text corpora. Certainly the use of simple correlational statistics between nearby words is not enough. In the best case, we might wish for perfect identification of causal relationships between events in an extremely large text corpus of narratives similar in content to COPA questions. Semantic similarity between these events and COPA sentences could be computed to gather evidence to select the best alternative. Even if it were possible to achieve this ideal, it is difficult to imagine that such an approach could mirror human performance on this task.

To move closer to human performance, systems may need to stretch beyond corpus statistics into

the realm of automated reasoning. Just as human readers do when hearing that "the man lost his balance on the ladder," successful systems may need to treat COPA premises as novel world states, and imagine a broad range of interconnected causal antecedents and consequents. Useful knowledge bases will be those that have adequate *coverage* over commonsense concerns, but also adequate *competency* to support generative inference of the sort more commonly seen in deductive and abductive automated reasoning frameworks. This knowledge may or may not be represented as text, but any successful system must have the capacity to apply this knowledge to the understanding of COPA's textual premises and alternatives. We consider the successful application of commonsense inference to text understanding to be one of the grand challenges of natural language processing, and hope that the COPA evaluation continues to be a useful tool for benchmarking progress toward this goal.

Acknowledgments

The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Church, K. and Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1):22-29.

Dagan, I., Glickman, O., and Magnini, B. (2006) The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.), Machine Learning Challenges. Lecture Notes in Computer Science, Vol. 3944, pp. 177-190, Springer, 2006.

Goodwin, T., Rink, B., Roberts, K., and Harabagiu, S. (2012) UTDHLT: COPACETIC System for Choosing Plausible Alternatives. Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), June 7-8, 2012, Montreal, Canada.

Gordon, A., Bejan, C., and Sagae, K. (2011) Commonsense Causal Reasoning Using Millions of Personal Stories. Twenty-Fifth Conference on Artificial Intelligence (AAAI-11), August 7–11, 2011, San Francisco, CA.

Gordon, A. and Swanson, R. (2009) Identifying Personal Stories in Millions of Weblog Entries. International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA.

Gordon, J., Van Durme, B., and K. Schubert, L. (2010) Learning from the Web: Extracting General World Knowledge from Noisy Text. Proceedings of the AAAI 2010 Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence (WikiAI 2010).

Graff, D. and Cieri, C. (2003) English Gigaword. Linguistic Data Consortium, Philadelphia.

Jung, Y., Ryu, J., Kim., K. and Myaeng, S.(2010). Automatic Construction of a Large-Scale Situation Ontology by Mining How-to Instructions from the Web. Journal of Web Semantics 8(2-3):110-124.

Lenat, D. (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure, Communications of the ACM 38:33-38.

Liu, H. and Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal 22(4):211-226.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. The 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), pages 365-372, Maryland, June 1999.

Morgenstern, L. (2012) Common Sense Problem Page. Retrieved April 2012 at http://www-formal.stanford.edu/leora/commonsense/

Noreen, E. (1989) Computer-Intensive Methods for Testing Hypotheses: An Introduction. New York: John Wiley & Sons.

Roemmele, M., Bejan, C., and Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 21-23, 2011.

Sagae, K. (2009) Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 81--84. 2009.

Speer, R., Havasi, C. and Lieberman, H. (2008) AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. Proceedings of AAAI 2008.