Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning

Melissa Roemmele[1], Cosmin Adrian Bejan[2], and Andrew S. Gordon[2]

[1]Department of Linguistics, Indiana University
Memorial Hall 322, Bloomington, IN 47405 USA
[2]Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094 USA
msroemme@indiana.edu, bejan@ict.usc.edu, gordon@ict.usc.edu

Abstract

Research in open-domain commonsense reasoning has been hindered by the lack of evaluation metrics for judging progress and comparing alternative approaches. Taking inspiration from large-scale question sets used in natural language processing research, we authored one thousand English-language questions that directly assess commonsense causal reasoning, called the Choice Of Plausible Alternatives (COPA) evaluation. Using a forced-choice format, each question gives a premise and two plausible causes or effects, where the correct choice is the alternative that is more plausible than the other. This paper describes the authoring methodology that we used to develop a validated question set with sufficient breadth to advance open-domain commonsense reasoning research. We discuss the design decisions made during the authoring process, and explain how these decisions will affect the design of high-scoring systems. We also present the performance of multiple baseline approaches that use statistical natural language processing techniques, establishing initial benchmarks for future systems.

Introduction

The fifty-year history of research in automated commonsense reasoning has seen slow but steady progress (Davis & Morgenstern, 2004). However, measuring this progress is difficult, as there exist few tools that researchers can use to evaluate the performance of their approach, or compare their work to that of other research groups. In lieu of established metrics, logical formalizations of commonsense knowledge have been evaluated using challenge problems of variable complexity. McCarthy (1998) motivated the use of challenge problems, and described how they could be used to judge the quality of a given axiomization of a domain. Using the example of the classic *Missionaries and Cannibals* puzzle, McCarthy argued that a good formalization of the problem domain would not only correctly solve the original problem, but also exhibit a high degree of *elaboration tolerance*. Formalisms are elaboration tolerant to the extent that it is convenient to modify a set of facts expressed in the formalism to take into account new phenomena or changed circumstances. For example, it should be easy to redefine the problem to allow for variable numbers of missionaries and cannibals, or the case where the missionaries are capable of converting an outnumbered cannibal into a missionary. The *Common Sense Problem Page* collects challenge problems of this sort along with their elaborations, and serves as the *de facto* scorecard for progress in the field (Morgenstern, 2011).

As a tool of evaluation, challenge problems of this sort have a number of drawbacks. First, the degree to which a reasoning system successfully solves the problem is a subjective judgment. Researchers typically define their own logical formalization of the problem space, select their own elaborations, and subjectively assess the degree to which their formalizations accommodate these elaborations. Convincingly arguing for the legitimacy of each success typically requires a full conference or journal article, one for each challenge problem attempt (e.g. Lifschitz, 1998; Morgenstern, 2001; Shanahan, 2004; Morgenstern, 2005).

Second, the focus on challenge problems favors research that targets *depth* rather than *breadth* in the pursuit of automated commonsense reasoning systems. That is, successful systems will have the inferential *competency* to solve these challenging problems and their variants, but lack the inferential *coverage* to similarly address problems outside of the narrow problem space. This is particularly problematic for research programs where inferential breadth is the explicit goal, e.g. the CYC project (Lenat, 1995) and ConceptNet (Liu & Singh, 2004).

Dramatically different approaches to evaluation have been successfully employed in other fields of artificial intelligence. The last decade of progress in natural language processing has been dominated by evaluation-driven research, where shared tasks and common test corpora have fueled innovation through competition. Many of these evaluation schemes would be inappropriate for commonsense reasoning research. However, our current work takes inspiration from the approach used in the Recognizing Textual Entailment (RTE) challenges, organized from 2004 to 2007 by the PASCAL Network of Excellence and by the NIST since 2008. In these yearly challenges, research groups compete using a common set of evaluation questions, typically 1600 questions divided equally into development and test sets. Each question consists of two text fragments (a text *T* and hypothesis *H*), where the task is to determine whether the truth of the second is entailed from the first. For example, the following pair is an example of a positive entailment:

T: Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.
H: The Beatles performed at Cavern Club at lunchtime.

As a tool for evaluation, the RTE question sets have number of excellent qualities. First, the inputs and outputs of the evaluation are well defined, with correct answers that have been validated by multiple human raters. Second, the size of the question sets ensures that competitive systems adequately tackle the problem of breadth. Third, splitting the question sets into separate development and test sets enables researchers to tune the parameters of their approaches (during development) without inflating their results due to over-fitting (during final testing). Fourth, the question sets are balanced with 50% positive and 50% negative entailments, so that a system's performance over a random baseline is readily evident.

Although the RTE challenge is itself an evaluation of inferential capability, it does not directly meet the needs of those interested in commonsense inference. Throughout the RTE challenges, a distinction has been made between textual *entailment* and textual *implication*, with only the former being the subject of the task. Although the line between entailment and implication is difficult to define, entailment is meant to include inferences that are necessarily true due to the meaning of the text fragment. In contrast, implications are inferences expected to be true, are likely causes or effects of the text, or are default assumptions. Whereas judgments of entailment between two text segments are strongly positive or negative, implications are judged in degrees of plausibility.

Accordingly, we modified the format of the RTE questions for use as a test of commonsense causal implication. Instead of two text segments, each question has three parts: a given premise and two plausible alternatives for either the cause or the effect of the premise. The following is an example, posed as a choice of plausible alternatives:

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

Each alternative is plausibly true given the premise. The correct answer is the one that is more plausible, in the commonsense view.

This paper describes the Choice Of Plausible Alternatives (COPA) evaluation, a corpus of one thousand questions of this sort to be used as a tool for evaluating progress in open-domain commonsense causal reasoning. We begin with a discussion of causality in the commonsense view, and then describe the authoring methodology we used to generate this question set. We then present performance evaluations of multiple baseline approaches to this task, establishing benchmarks for future systems.

Commonsense Causality

Theoretical investigations of causality have been pursued across many fields, each helping to refine a definition of causality that agrees with our commonsense intuitions. In philosophy, a rigorous test for determining a causal relation between two events is that of "necessity in the circumstances" (Hart & Honore, 1959; Mackie, 1980). According to this criterion, event *A* is necessary for event *B* if the following statement is true: if *A* had not occurred in the circumstances, then *B* would not have occurred (therefore, *A* causes *B*). An alternative view of causality requires "sufficiency in the circumstances" between two events (Mackie, 1980; Trabasso et al., 1984). *A* is said to be sufficient in the circumstances for *B* if it is true that if *A* occurs and things continue normally from there, event *B* will occur (therefore, *A* causes *B*). Necessity and sufficiency do seem to play a role in human reasoning about causality, as demonstrated in experimental settings. When subjects detect a relation between two events in terms of necessity and/or sufficiency, they also deem these events as causally related (Thompson, 1989; Trabasso et al., 1989).

However, the phrase "in the circumstances" in these definitions only hints at the role of background knowledge in causal judgments. Other theories of causality have focused on this knowledge directly. The *mechanism view* of causal reasoning (Salmon, 1984; Harre & Madden, 1975; Shultz, 1982; Ahn et al., 1995) holds that basic theoretical knowledge underlies individuals' conception of causal relations. For instance, in order to recognize the causal relation between the event "the child let go of the string attached to the balloon" and the event "the balloon flew away", one needs the knowledge that balloons naturally rise, for instance. Singer et al. (1992) proposes a role for *causal bridging inferences*, where individuals invoke a statement that bridges the two events into a causal relation, and then validate this bridging statement against commonsense knowledge. For example, the knowledge that "balloons rise" bridges the statements "the child let go of the string attached to the balloon" and "the balloon flew away" into a causal relation, and the validation of this bridging inference against commonsense knowledge affirms the causal relation.

Events in a causal relation always occur within some context, whether explicit or implicit, which some researchers term the *causal field* (Mackie, 1980; Shoham, 1990) or the *causal complex* (Hobbs, 2005). These collections of contributing causal factors are derived from an individual's knowledge about what "usually takes place" in the world (Shoham, 1990). As additional information becomes available, this information may yield different conclusions about causality than were previously made in the absence of that information (nonmonotonic inference). For instance, the following statement is judged a valid causal relation: "the balloon flew away because the child let go of the string attached to the balloon". However, the validity of this statement requires the assumption that the child's balloon was filled with helium and not air, for instance. Explicit knowledge that the balloon contains air rather than helium would render the above statement invalid, since balloons filled with air do not rise. Still, individuals do not require explicit clarification about this factor before accepting the given statement as valid. Here, an inference is *plausible* insomuch as the cost of including "the balloon is filled with helium" in the causal field is relatively low, given the two events.

We used this cost-based view of plausibility to devise a simple question format to test a system's ability to make commonsense causal judgments. A single question in this format consists of a statement (the *premise)* and two choices (the *alternatives)* that both could plausibly have a causal relation with the premise. The correct choice is the alternative that is more plausible, i.e. the cost of including the bridging inferences in the causal field is less than the other, validated by human judgments. This format has two variations, depending on whether the alternatives are to be viewed as plausible effects of the premise (forward causal reasoning) or as plausible causes of the premise (backwards causal reasoning), as in the following two examples.

(forward causal reasoning)
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.

(backwards causal reasoning)
Premise: The man fell unconscious. *What was the cause of this?*
Alternative 1: The assailant struck the man in the head.
Alternative 2: The assailant took the man's wallet.

Authoring Methodology

The Choice of Plausible Alternatives (COPA) evaluation consists of 1000 questions of commonsense causality. The question set was created using a specific authoring methodology that ensured breadth of topics, clarity of the language, and high agreement among human raters. This section describes the authoring methodology, focusing on issues of breadth, clarity and agreement.

The first major concern of the authoring methodology was the breadth of the question set. Our approach was to identify question topics from different sources where a high degree of breadth was already evident, and then elaborate these topics into premises and alternatives through our own creativity. This approach helped balance the analytic and generative aspects of this task, ensuring that the particular topic interests of the author were not over-represented in the question set, but still allowing for the creative design solutions that each of these questions required. Two primary sources of question topics were used to ensure breadth. First, topics were drawn from randomly selected entries in a corpus of one million personal stories written in Internet weblogs in August and September of 2008 (Gordon & Swanson, 2009). We read hundreds of individual stories looking for topics discussed in these daily narratives of people's everyday lives. While diverse, this source tended to focus on social and mental topics, with fewer topics related to natural and physical causality. The opposite was true of our second source of topics, the subject terms of the Library of Congress Thesaurus for Graphic Materials (Library of Congress Prints and Photographs Division, 1995). Developed over the course of decades of library cataloging work, this set of subject terms has broad coverage over the sorts of people, places, and things that appear in photographs and other imagery. We randomly selected hundreds of subject terms from the set to use as question topics, discarding obscure terms or those with no obvious role in causal reasoning.

From each of these question topics, we authored a pair of statements (the premise and the correct alternative) that captured a key causal relationship. This part of the task required subjective creativity, guided by introspective questions about the topic. For instance, from the topic of "unconsciousness" the authors asked themselves "what causes unconsciousness?" and "what does unconsciousness cause?" Answers to these questions were treated as a causal bridging inference, e.g. "injuries to the head cause unconsciousness." From this, a suitable premise and correct alternative could be instantiated as the events of the causal relation, e.g. "the assailant struck the man in the head" and "the man fell unconscious." Either the cause or the effect in this pair could be treated as the premise, depending on whether the question was testing forward or backward causal reasoning.

A challenging part of the authoring task was to establish the incorrect alternative for each question. This statement was intended to be similar in form to the correct alternative, and somewhat related to the premise, but with no obvious direct causal connection. For example, the premise "the assailant struck the man in the head" and correct alternative "the man fell unconscious" both evoke the schema of an assault. "The assailant took the man's wallet" is also a plausible event in this situation, but it is less plausible that this event would be the direct cause of the man falling unconscious. This design is intended to ensure that answering these questions requires causal reasoning, and cannot be answered using purely associative methods.

In authoring premises and alternatives, we took inspiration from Warren et al.'s (1979) taxonomy for propositions participating in causal relations, which includes states ("it was sunny outside"), events ("the car broke down"), actions ("the man went to the doctor"), cognitions ("I forgot to eat breakfast"), displays ("the girl lost her balance"), impulses ("I felt embarrassed"), and goals ("the teenager wanted to rebel"). As much as possible, we tried to ensure that the two alternatives both fell into the same class.

The second major concern of the authoring methodology was the clarity of the language. The natural language representation of each of the question statements followed a number of guidelines to ensure clarity and to reduce the complexity of the natural language processing aspects of evaluated systems. The premise and the alternatives were written in the past tense. They were as brief as possible, omitting words that were not necessary to select the correct alternative. Proper names of people and places were avoided, as were colloquialisms and slang. Personal pronouns and definite determiners were used, which led us to adopt a particular style for co-reference and anaphora. For example, consider the following question:

Premise: The man dropped food on the floor. *What happened as a result?*
Alternative 1. His dog ran over to eat the food.
Alternative 2. His dog jumped up on him.

The alternatives for this question both explicitly reference a dog whose existence must be presumed in the premise. Here the personal and possessive pronouns ("his", "him") must be resolved to "the man", and "the food" must be seen as co-referential with "food" in the premise.

The third major concern of the authoring methodology was that there was agreement among human raters who were asked to answer each question. To validate the set, we enlisted the help of 10 volunteers, all native English speaking adults not affiliated with our project. Each volunteer was given 200 questions, such that two people answered each question. Agreement between authors was high (Cohen's K = 0.965). In all, these volunteers answered 26 questions differently than was intended by the author of the question. These 26 questions were removed from the set, and replacement questions were generated and validated by two additional raters. The final set contained 1000 questions, each validated by two raters who selected the correct alternative intended by the author. The order of the question set was randomized to mitigate the changes in style during the course of the authoring process. The position of the correct alternative was also randomized, ensuring that a random baseline would answer exactly 50% of the questions correctly.

We expect that future automated reasoning systems will see significant performance gains by modifying various system parameters, and hill-climbing over the evaluation results. To facilitate parameter tuning of this sort, we created a set of scripts to automate the evaluation, and divided the question set into equally sized development and test sets so as to avoid over-fitting (500 questions each). Our recommendation is that researchers publish their results on both the development set and the test set to facilitate comparison with competing approaches.

Performance of Baseline Approaches

The COPA evaluation was designed so that a random baseline system, where one of the two alternatives is randomly chosen for each question, would perform at exactly 50%. In addition, we investigated the performance that could be achieved by somewhat stronger baselines based on simple associative methods. While we do not expect these baselines to be competitive with future purpose-built approaches, successful systems must demonstrate improvements over these results that are statistically significant.

Our baseline approaches explore the simple idea that causally related events are often described together in written text. Accordingly, one would expect that correlation statistics between words in large text corpora capture some of this causal information. By computing the combined weight of correlation between words in the premise and each alternative of a COPA question, it may be possible to select the alternative with the stronger causal connection.

The three baselines we explored are simple unsupervised learning algorithms that rely on correlation statistics collected by processing a large corpus of text documents and querying web search engines. In order to decide on the most plausible alternative $a*$ associated with a premise p, each baseline computes a *causality score* that measures the causal relatedness between p and its corresponding alternatives a_1 and a_2, and selects the alternative with the larger score:

$$a^* = argmax_{a \in \{a_1, a_2\}} causality(p, a)$$

Our first baseline performs statistical analysis on all of the English-language text documents contained in Project Gutenberg[1], a corpus of over 42,000 documents (16GB of text). We compute the causality score between a premise p and one of its alternatives a by averaging over all possible correlations holding between content words from p and a:

$$causality(p, a) = \frac{\sum_{w_p \in p} \sum_{w_a \in a} correlation(w_p, w_a)}{N_p N_a}$$

In this formula, N_p and N_a represent the number of content words in p and a, respectively. For the *correlation* measures, we selected the *pointwise mutual information* (PMI) measure as described in (Church and Hanks, 1989) and the *Dice coefficient* (Dice, 1945). As these measures are asymmetric with respect to their arguments, we used

[1] http://www.gutenberg.org

		1. Text collection (word pairs level)			2. Web (word pairs level)			3. Web (word phrase level)		
		W=5	W=25	W=50	Bing	Google	Yahoo	Bing	Google	Yahoo
Test Set (500 questions)	Dice	56.0*	54.6	53.6	47.4	50.6	47.8	57.8**	49.6	57.2*
	PMI	**58.8****	58.6**	55.6*	54.6	51.6	52.8	55.0	48.0	54.8
Dev Set (500 questions)	Dice	53.6	51.8	52.2	52.0	50.4	51.0	55.0	48.0	55.4*
	PMI	**57.8****	**57.8****	56.8*	50.8	52.0	50.6	54.0	47.0	55.0
Dev + Test (1000 questions)	Dice	54.8*	53.2	52.9	49.7	50.5	49.4	56.4**	48.8	56.3**
	PMI	**58.3*****	58.2***	56.2**	52.7	51.8	51.7	54.5*	47.5	54.9*

Table 1: Three baseline results for the COPA evaluation. The results are computed in terms of accuracy and the ones marked with ***, **, and * are statistically significantly better at the 0.001, 0.01, and 0.05 levels, respectively, than the random baseline (50% accuracy).

the reverse word order for questions requiring backwards causal reasoning.

The second baseline employs the same formula for the causality measure as the first baseline, but estimates the correlation measure for each word pair differently. Inspired by the approach of Turney (2001), we replaced the frequency counts computed over the text collection in the first baseline with the number of hits retrieved by a web search engine. For this purpose, we considered three of the most popular search engines: Bing, Google, and Yahoo. Although the mechanisms used to compute the number of hits in these search engines are unknown to us, this baseline has the advantage that the correlation statistics are computed at the scale of the entire web.

Finally, the third baseline tries to determine whether a premise p usually shares the same causal content with one of its alternatives a by computing correlations at the word phrase level and not at the word pair level as performed by the first two baselines. That is, the baseline queries the search engines using the entire textual content from p and a in order to estimate the correlation measures. For example, the formula for selecting the most plausible alternative that employs the PMI measure for this type of approach is reduced to:

$$a^* = argmax_{a\in\{a_1, a_2\}} \frac{hits(p + a)}{hits(a)}$$

In our experiments, we evaluated the baselines on both the development and test set of questions. In addition, since our baselines are unsupervised methods and their decision does not depend on any information extracted from any corpus question, we also performed experiments on the entire collection of questions (development + test).

Table 1 presents the accuracy results obtained by each of these three baselines and for each of the two correlation measures. The best results were obtained using the first baseline, computing correlation statistics over English-language documents in Project Gutenberg. For both the PMI and Dice measures, we computed the correlation of a pair of content words by counting the frequency that the two words occurred together within a specific window size of words (W=5, W=25, and W=50). This baseline performed best using the PMI measure with the smallest window size, suggesting that causally related events are typically described in close proximity to one another.

This conclusion is also supported by the results of the second baseline. In spite of using a much larger collection of documents, this second baseline yields poor results in comparison with the first. Here the window size is not restricted within a web document when computing the correlation measure between word pairs. Therefore, much more non-causal information can be added to the strength of the correlation.

However, the hits-based correlation measures perform better in the third baseline, where the query is restricted by the entire textual context encoded in premises and alternatives. In particular, using the Dice coefficient and query hits from Bing and Yahoo yield results that are significantly better than the random baseline.

Discussion

As a tool for advancing research in commonsense causal reasoning, the COPA evaluation has several desirable characteristics. We have established that human raters can perform extremely well on this task, with near perfect agreement. Conversely, we have established that simple associative techniques based on corpus statistics perform only moderately above the random baseline. The gap between these two scores presents a fertile ground for future research, where the merits of competitive approaches can be measured. The size and breadth of the question set ensures that successful approaches must tackle the problem of coverage as well as competency. The forced choice design of the questions allow for automated scoring of systems, while the split between development and test sets ensures that reported performance results are not inflated due to over fitting of the tuned system parameters.

Considerable challenges must be overcome to develop systems that approach human-level of performance on the COPA evaluation. However, we believe that several existing lines of research could yield results that are substantially above our current baselines. Large-scale logical formalizations of commonsense knowledge, such as the CYC knowledge base (Lenat, 1995), may perform very well on this task. However, applying these knowledge sources will require a robust ability to convert the English text of COPA questions into logical form, although hand-authored translations of the entire question set may be feasible. Approaches that rely on crowdsourcing to collect commonsense knowledge from volunteers on the web, such as ConceptNet (Liu & Singh, 2004) may be particularly well suited for the COPA evaluation. These approaches can specifically target the knowledge necessary to reason about the causal connections between everyday

events, and typically use natural language to represent these events. Finally, approaches involving commonsense knowledge automatically extracted from the web may yield high performance on the COPA evaluation. Promising methods include that of Gerber et al. (2010), where the genre of web content is well chosen for commonsense knowledge extraction, and causal relations are the specific target of the extraction process. It is our hope that the COPA evaluation will be a useful tool for comparing the merits of these and other innovative approaches in the future. Evaluation materials are available online at:
http://ict.usc.edu/~gordon/copa.html

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Ahn, W., Kalish, C., Medin, D., & Gelman, S. (1995) The role of covariation versus mechanism information in causal attribution, Cognition, 54(3): 299-352.

Church, K. & Hanks, P. (1989) Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1):22-29.

Davis, E. & Morgenstern, M. (2004) Introduction: Progress in formal commonsense reasoning. Artificial Intelligence 153(1-2):1-12.

Lenat, D. (1995) Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38(11):33-38, November 1995.

Dice, L. (1945) Measures of the Amount of Ecologic Association Between Species. Journal of Ecology, 26:297-302.

Gerber, M., Gordon, A., & Sagae, K. (2010) Open-domain Commonsense Reasoning Using Discourse Relations from a Corpus of Weblog Stories. 1st International Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR) NAACL 2010 Workshop, Los Angeles, CA, June 6, 2010.

Gordon, A. & Swanson, R. (2009) Identifying Personal Stories in Millions of Weblog Entries. Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA, May 20, 2009.

Harré, R. & Madden, E. (1975) Causal powers: a theory of natural necessity. Rowman & Littlefield, Totowa, NJ.

Hart, M. & Honore, A. (1985) Causation in the law. Oxford, England: Clarendon.

Hobbs, J. (2005) Toward a Useful Concept of Causality for Lexical Semantics. Journal of Semantics 22(2):181-209.

Lenat, D. (1995) Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38(11):33-38.

Library of Congress Prints and Photographs Division. (1995). Thesaurus for graphic materials. Washington, DC: Library of Congress Distribution Service.

Lifschitz, V. (1998) Cracking an Egg: An Exercise in Commonsense Reasoning. Proceedings of Common Sense 98, January 1998.

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

Mackie, J. (1980) The Cement of the Universe - A Study of Causation. Oxford: Oxford University Press.

McCarthy, J. (1998) Elaboration Tolerance. 4th International Symposium on Logical Formalizations of Commonsense Reasoning.

Morgenstern, L. (2001) Mid-Sized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking, Studia Logica, 67(3):333-384, 2001.

Morgenstern, L. (2005). A First-Order Axiomatization of the Surprise Birthday Present Problem: Preliminary Report. 7th International Symposium on Logical Formalizations of Commonsense Reasoning.

Morgenstern, L. (2011) Common Sense Problem Page. http://www-formal.stanford.edu/leora/commonsense/

Salmon, W. (1984) Scientific explanation and the causal structure of the world. Princeton University Press, Princeton, NJ.

Shanahan, M. (2004) An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense Reasoning, Artificial Intelligence, 153(1-2):141-165.

Shoham, Y. (1990) Nonmonotonic reasoning and causation, Cognitive Science, 14(2): 213-252.

Shultz, T (1982) Rules of causal attribution. Monographs of the Society for Research in Child Development 47 (Serial No. 194).

Singer, M., Halldorson, M., Lear, J. & Andrusiak, P. (1992) Validation of causal bridging inferences in discourse understanding, Journal of Memory and Language, 31(4): 507-524.

Thompson, V. (1995) Conditional reasoning: the necessary and sufficient conditions. Canadian Journal of Experimental Psychology 49: 1–60.

Trabasso, T., Broek, P. & Suh, S. (1989) Logical necessity and transitivity of causal relations in stories. Discourse Processes, 12(1): 1-25.

Trabasso, T., Secco, T. & van den Broek, P. (1984) Causal cohesion and story coherence. In: Mandl, H., Stein, N.L. and Trabasso, T., Editors, 1984. Learning and comprehension of text, Erlbaum, Hillsdale, NJ.

Turney, P. (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. 12th European Conference on Machine Learning (ECML), 491-502.

Warren, W., Nicholas, D., & Trabasso, T. (1979) Event chain and inferences in understanding narratives. In: Freedle, RO, Editor, 1979. New directions in discourse processing, Erlbaum, Hillsdale, NJ.