Commonsense Causal Reasoning Using Millions of Personal Stories

Andrew S. Gordon, Cosmin Adrian Bejan, and Kenji Sagae
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094 USA
gordon@ict.usc.edu, bejan@ict.usc.edu, sagae@ict.usc.edu

Abstract

The personal stories that people write in their Internet weblogs include a substantial amount of information about the causal relationships between everyday events. In this paper we describe our efforts to use millions of these stories for automated commonsense causal reasoning. Casting the commonsense causal reasoning problem as a Choice of Plausible Alternatives, we describe four experiments that compare various statistical and information retrieval approaches to exploit causal information in story corpora. The top performing system in these experiments uses a simple co-occurrence statistic between words in the causal antecedent and consequent, calculated as the Pointwise Mutual Information between words in a corpus of millions of personal stories.

Introduction

Recent advances in open information extraction from the web (Etzioni et al., 2008) have captured the attention of researchers in automated commonsense reasoning, where the failures of formal approaches have been attributed to the lack of sufficiently broad stores of commonsense knowledge with sufficient inferential soundness (Davis & Morgenstern, 2004). There are strong similarities between the products of recent open information extraction systems (e.g. Ritter et al., 2009) and knowledge resources that commonsense reasoning researchers have previously found useful across a wide range of reasoning tasks, e.g. WordNet (Miller et al., 1990). Still, there remains a large gap between these products and the formal axiomatic theories that continue to be the target of commonsense reasoning research. Proponents of these new approaches have advised researchers to eschew elegant theories, and instead "embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data" (Halevy et al., 2009). Still, the development of effective data-driven methods for automated commonsense reasoning remains an open challenge.

Progress can be seen in several recent research efforts. Schubert (2002) presented an approach to acquiring general world knowledge from text corpora based on parsing sentences and mapping syntactic forms into logical forms, then gleaning simple propositional facts from these forms through abstraction. This approach was implemented in the KNEXT system (Schubert & Tong, 2003; Van Durme et al., 2009; J. Gordon et al., 2009), which when given the noun phrase "her washed clothes" derives the knowledge that clothes can be washed and that female individuals can have clothes. Adopting a similar approach, Clark and Harrison (2009) showed how extracted factoids of this type could be used to improve syntactic parsing and the recognition of textual entailment.

Although fact extraction of this sort has proven useful for some tasks, it is still unclear how this knowledge can be applied toward benchmark problems in commonsense reasoning. Problems listed on the Common Sense Problem Page (Morgenstern, 2011) each require reasoning about the details of specific situations, where the causal implications of actions and events are the central concern. Logical formalizations of solutions to these problems (e.g. Shanahan, 2004) have focused on the key causal knowledge required for individual problems, with an emphasis on inferential competency (depth) rather than inferential coverage (breadth). The exact opposite emphasis is pursued in recent efforts to extract causal information from unstructured textual data. For example, Rink et al. (2010) identify lexical-semantic patterns indicative of causal relations between two events in a single sentence by generalizing over graph representations of semantic and syntactic relations. Applied to increasingly large corpora, techniques like this can identify countless numbers of causal relationships between textual clauses. However, even if these approaches succeeded in finding the key causal knowledge required for a problem on the Common Sense Problem Page, it is unreasonable to expect that this text could be utilized to solve the problem as effectively as the hand-crafted axioms of an expert logician. In short, these benchmark problems in commonsense reasoning are not useful tools for gauging the progress of data-driven approaches.

Roemmele et al. (2011) recently addressed this problem by developing a new evaluation for commonsense causal reasoning. Modeled after the question sets of the Recognizing Textual Entailment (RTE) challenges, the Choice of Plausible Alternatives (COPA) evaluation

consists of one thousand multiple-choice questions that require commonsense causal reasoning to answer correctly. As with RTE questions, these questions are posed as English-language sentences, and have been balanced so that the random-guess baseline performance is 50%. With its focus on commonsense causal knowledge, this evaluation retains the central concerns of the previous commonsense challenge problems, while allowing for the straightforward application and evaluation of new data-driven approaches.

With an appropriate evaluation in place, interest now turns to the development of competing approaches. In this paper, we describe our recent efforts to develop an approach that utilizes unstructured web text on a massive scale. This work specifically looks at the unique properties of the genre of the personal story, for which millions of examples are readily available in weblogs. We argue that personal stories from weblogs are ideally suited as a source of commonsense causal information, in that causality is a central component of coherent narrative and weblogs are inherently focused on everyday situations.

In the following sections we present a novel approach to commonsense causal reasoning, where millions of personal stories written in weblogs are used as a knowledge base. We begin by describing the evaluation of commonsense causal reasoning developed by Roemmele et al. (2011), the Choice of Plausible Alternatives. We then describe a large-scale corpus of personal stories created by applying statistical text classification techniques to tens of millions of English-language weblog posts. We then report the results of four experiments to compare methods for using this corpus as a knowledge base for automated commonsense causal reasoning.

Choice of Plausible Alternatives

The Choice of Plausible Alternatives (COPA) evaluation is a collection of 1000 questions that test a system's capability for automated commonsense causal reasoning (Roemmele et al., 2011). Individual questions are composed of three short sentences, namely a premise and two alternatives, where the task is to select which alternative has a more plausible causal relationship with the premise. Questions are written for both forward and backwards causal reasoning, i.e. asking of the premise *What happened as a result?* or *What was the cause of this?* The following are three examples of COPA questions:

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.

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.

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.

Similar to the question sets of the RTE challenges, COPA questions are divided into development and test sets of 500 questions each, where the development set is used to tune the parameters of a competing system without inflating results on the test set due to over-fitting. The correct alternative is randomly balanced, so that the expected performance of a random baseline system is 50%. Human raters validated the correct alternative for each question, achieving high inter-rater agreement (K=0.965).

Roemmele et al. (2011) also provided performance results of several systems based on simple corpus statistics, which are the only published results for this evaluation to date. Their best results were obtained by calculating the Pointwise Mutual Information (PMI) statistic (Church & Hanks, 1990) between words in the premise and each alternative in a large text corpus, and selecting the alternative with a stronger correlation. In this approach, the normalized score between premise p and alternative a is calculated as follows:

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

In this formula, N_p and N_a represent the number of content words in p and a, respectively. As a text corpus for the PMI statistic, Roemmele et al. analyzed every English-language document in Project Gutenberg, a corpus of over 42,000 documents (16GB of text). Frequencies of co-occurrence between words were tabulated for word windows of various lengths (5, 25, and 50) using the method described by Church and Hanks (1990), with the best COPA results found using a window of 5 words. As this method of calculating the PMI statistic is asymmetric with respect to the order of its arguments, the reverse word order was used when answering COPA questions with backwards causality. Table 1 presents the COPA scores for the best performing baseline on the development, test, and combined question sets. For all experiments in this paper, statistical significance is calculated by stratified shuffling (Noreen, 1989).

System	*Dev*	*Test*	*All*
Random baseline	50.0	50.0	50.0
PMI Gutenberg (W=5)	57.8*	58.8**	58.3***

Table 1. COPA baselines results (Roemmele et al., 2011). Performance differences are statistically significant at p<.05 (*), p<.01 (**), and p<.001 (***).

The results presented in table 1 demonstrate that words that are correlated are more likely to impart some causal information, but the performance of this approach over a random baseline is still modest (8.8% above chance on the

test set). Novel techniques are required to bridge the gap between this result and human performance.

Personal Stories as a Knowledge Base

In this work, we explored whether the challenges of open-domain commonsense causal reasoning could be overcome by exploiting the unique properties of personal stories, the narratives that people tell about events in their everyday lives. The narrative genre is particularly interesting because of the role that causal relations play in determining discourse structure. In the field of discourse psychology, Trabasso and van den Broek (1985) developed a highly influential model of the causal structure of goal-based stories, the causal network model. In this model, narratives are viewed as a series of sentences (or clauses) of a particular narrative class, e.g. events, goals, and attempts, which are connected by implicit causal links. By analyzing the underlying causal network of specific narratives, discourse psychologists have been able to predict a wide range of observed memory behaviors, sentence reading times, recognition priming latencies, lexical decision latencies, goodness of fit judgments for story sentences, and the inferences produced during thinking aloud (van den Broek, 1995; Magliano, 1999).

There is also a long history of interest in the genre of the personal story within artificial intelligence, most notably in early work on Case-Based Reasoning. In the Dynamic Memory model articulated by Schank (1982), personal stories serve as a tool to highlight where a knowledge base is faulty, needing revision. Where people's experiences in life are exactly as expected, there is no cognitive utility in storing these experiences away in memory. Faced with the unexpected (an *expectation violation*), the narrative of these experiences is remembered so that the conditions of the situation might weigh in some future knowledge-revision process, helping to tune one's expectations so that they are more in accordance with the way the world actually works. Schank and Abelson (1995) took this idea further, arguing that all knowledge is encoded in stories: one's own collection of personal stories was the knowledge base itself. Although some early software implementations of these ideas existed, a major technical barrier in this line of research was the question of scale (Schank, 1991). Due to the labor of curating and analyzing individual stories, story collections used in these implementations contained less than one thousand narratives: roughly the number of items that could be managed by a handful of researchers over the lifecycle of a typical research project.

In the last 20 years, however, the Internet and statistical natural language processing have drastically changed the way that researchers approach the problem of scale. Gordon and Swanson (2009) estimated that 4.8% of all non-spam weblog posts are personal stories, defined by them as: non-fictional narrative discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the storyteller or a close associate is among the participants. Using supervised machine learning approaches, Gordon and Swanson identified nearly one million English-language personal stories in the ICWSM 2009 Spinn3r Dataset (Burton et al., 2009). This dataset is a corpus of tens of millions of non-spam weblog entries posted in August and September of 2008, provided to researchers by Spinn3r.com, a weblog aggregator.

Our expectation was that this corpus would be particularly well suited as a source for information about causality in the everyday situations. If causal relations were as prevalent in narrative as suggested by discourse psychologists, and if bloggers told stories about all aspects of daily life, then this corpus should contain information relevant to nearly every one of the questions in the COPA evaluation, by virtue of its scale. We obtained the Gordon and Swanson (2009) corpus from the authors, and conducted a series of experiments to directly use it as a knowledge base for answering COPA questions.

In our first experiment, we investigated whether higher accuracy on the COPA evaluation could be achieved simply by swapping this text corpus for the Project Gutenberg corpus used by Roemmele et al. (2011) in their highest-performing baseline. We calculated the PMI statistic between words in the premise and alternatives for variously-sized word windows, and used the same formula for computing the normalized strength of correlation. Table 2 presents the results for this first experiment, and compares these results to the highest-performing baseline.

System	*Dev*	*Test*	*All*
PMI Gutenberg (W=5)	57.8	58.8	58.3
PMI 1M Stories (W=5)	58.8	62.8	60.8
PMI 1M Stories (W=15)	57.6	64.4*	61.0
PMI 1M Stories (W=25)	60.0	65.2**	62.6*
PMI 1M Stories (W=30)	59.4	65.0*	62.2*
PMI 1M Stories (W=50)	57.6	62.8	60.2

Table 2. COPA evaluation results for systems using the PMI statistic on a corpus of nearly one million personal stories. Accuracy improvements over the baseline (PMI Gutenberg) are statistically significant at $p<.05$ (*) and $p<.01$ (**).

The results of our first experiment show that personal stories from weblogs are, indeed, a better source of causal information than Project Gutenberg documents, at least when using simple corpus statistics for the selection of alternatives. It is also encouraging that these improvements were obtained even though the story corpus is substantially smaller than the Project Gutenberg corpus, with 1.9GB of text (compared to 16GB).

In this first experiment, our best results were obtained with a PMI window size of 25 words, substantially larger than the 5-word window used in Roemmele et al.'s best-performing system. This window size suggests that the causal information in word correlations is strongest within the scope of adjacent clauses and sentences. This finding is consistent with the analysis of causal structure in narratives by discourse psychologists (e.g. Trabasso & van den

Broek, 1985), and suggests that attention to this structure could lead to further improvements in COPA scores.

Reasoning With Discourse Relations

Gerber et al. (2010) proposed that the discourse structure of personal stories could be used to support automated commonsense reasoning. Using an automated discourse parser (Sagae, 2009) trained on a corpus annotated with discourse relations of Rhetorical Structure Theory (RST) (Carlson & Marcu, 2001), Gerber et al. identified relations between the elementary discourse units (clauses) in Gordon and Swanson's (2009) corpus of nearly one million stories from weblogs. To aggregate information from fine-grained and symmetric RST relations, all relations related to causality and temporal order were collapsed into these two relations. In total, this analysis identified 2.2 million instances of the cause relation and 220 thousand instances of the temporal order relation.

Gerber et al. used these instances to generate commonsense causal inferences from an input sentence using a simple case-based reasoning approach. Given a sentence describing an arbitrary event, they identified the most similar sentences in their tables of cause relations, using standard text similarity metrics. They then inferred that the causal antecedent or consequent of the input would be the same, i.e. that the sentence causally related to the retrieved sentence also held this relationship with the query. The accuracy of this approach on arbitrary input, based on human judgments, was low (10.19%). However, somewhat better results could be achieved (17.36%) by aggregating evidence from sets of similar sentences.

In our second experiment, we explored the effectiveness of Gerber et al.'s approach by adapting it for the COPA evaluation. First, we obtained the full set of cause relations from the authors, consisting of 2.2 million pairs of text strings for each annotated antecedent and consequent. Second, we generated search indexes for both the antecedents and consequents using the Terrier Information Retrieval Platform and its Divergence From Randomness retrieval model (Ounis et al., 2007). Third, we developed a simple search-based formula for computing the strength of the causal connection between a COPA premise and an alternative. For COPA questions of forward causality, we used the premise as a query to the antecedent index *(a)* and each alternative as a query to the consequent index *(c)*. We then counted the number of times that a causal relation existed between items in the top *N* search results, weighted by the product of their similarity *(w)* to the query.

$$causality(a, c) = \sum_{i=1}^{N} \sum_{j \in X_i} w_i w_j$$

$$X_i = \{ x \mid \exists\, a_x \; causes \; c_i \wedge 1 \leq x \leq N \}$$

The same formula was used for questions of backwards causality, except that we searched the opposite indexes. The alternative with the strongest causal connection was selected as the most plausible. Table 3 presents the results of our second experiment, considering search result sets of various sizes.

System	*Dev*	*Test*	*All*
PMI 1M Stories (W=25)	60.0	65.2*	62.6*
2.2M RST relations (N=100)	51.8	50.6	51.2
2.2M RST relations (N=500)	49.2	53.6	51.4
2.2M RST relations (N=1000)	53.8	55.0	54.4
2.2M RST relations (N=2000)	59.0	58.0	58.5
2.2M RST relations (N=3000)	54.6	57.2	55.9

Table 3. COPA evaluation results for systems using 2.2 million aggregated RST relations, compared with the best result from table 2. The PMI-based system significantly outperforms the best RST-based system (N=2000) at p<.05 (*).

These results cast some doubt on the utility of the causal discourse relationships identified by Gerber et al. (2010) for open-domain causal reasoning. A simple PMI-based approach using the same story corpus significantly outperformed the best RST-based result.

Reasoning With Sentence Proximity

As noted by Gerber et al. (2010), one of the shortfalls of their approach was the poor performance of their RST discourse parser on this genre of text. With low recall, much of the causal information encoded in these stories would be overlooked. With low precision, much of the extracted causal information would be incorrect. We wondered if similar results could be achieved simply by looking for any pair of similar sentences appearing in the same story, regardless of whether we could identify a causal discourse connective between them.

In our third experiment, we developed and evaluated a simpler search-based approach. First, we indexed every sentence in the Gordon and Swanson (2009) story corpus (25.5 million sentences), again using Terrier and the Divergence From Randomness retrieval model. Second, we modified our search-based formula to sum the number of times that the top *N* search results for COPA premises and alternatives appeared in the *same* story, again weighted by the product of their similarity to the query.

We also guessed that sentence pairs in closer proximity to one another should be more likely to be causally related. In this third experiment, we also evaluated a modification to the weighting scheme to favor pairs in close proximity: dividing the product of weights by the distance between sentences in the pair (plus one, to avoid a division by zero error in the case where the premise and alternative both retrieved the same sentence).

Table 4 presents COPA evaluation results for both of these approaches, using search results sets of various sizes. These results are similar to those in table 3, suggesting that there is no advantage gained by trying to identify causal relations in text, given the current state of automated discourse parsers on this genre. Again, a simple PMI-based approach using the same story corpus achieved

significantly better results than our best retrieval-based method on the COPA test set.

System	*Dev*	*Test*	*All*
PMI 1M Stories (W=25)	60.0	65.2	62.6
Sentence Co-occurrence (N=100)	52.4	54.4	53.4
Sentence Co-occurrence (N=500)	55.6	59.0	57.3
Sentence Co-occurrence (N=1000)	54.0	55.6	54.8
Sentence Co-occurrence (N=2000)	55.0	55.2	55.1
Sentence Co-occurrence (N=3000)	54.6	56.0	55.3
Weighted by Distance (N=100)	56.8	53.8	55.3
Weighted by Distance (N=500)	55.8	59.4	57.6
Weighted by Distance (N=1000)	57.8	58.2	58.0
Weighted by Distance (N=2000)	60.6	58.8	59.7
Weighted by Distance (N=3000)	57.4	58.0	57.7

Table 4. COPA evaluation results for systems based on co-occurrence of retrieved sentences in the same story, or weighted by distance, compared with the best result from table 2. The PMI-based system significantly outperforms the two best Weighted by Distance systems on the test set, at $p<.05$ (N=500) and $p<.01$ (N=2000).

Reasoning With Millions of Stories

We were encouraged by the success of the simple PMI-based approach using a corpus of nearly one million personal stories. We reasoned: if one million personal stories from weblogs yielded good results, then scaling up the size of the corpus should be even better. In our fourth experiment, we tested this hypothesis by scaling up the size of the story corpus by an order of magnitude.

We developed a pipeline for automatically identifying personal stories in weblogs, with the aim of collecting every English-language personal story posted to a weblog in the year 2010. This new pipeline closely followed the design used by Gordon & Swanson (2009) to create the corpus of nearly one million stories, where supervised machine learning techniques were used to identify personal stories in a stream of English-language weblog posts. We used a slightly improved version of their story classifier that incorporates lexical, syntactic, and HTML tag features, described in Swanson (2010).

As a source of weblog entries, we partnered directly with Spinn3r.com, the commercial weblog aggregator that provided the dataset for the ICWSM 2009 Spinn3r Dataset Challenge (Burton et al., 2009) used to create Gordon and Swanson's story corpus. Using Spinn3r.com's feed API, our servers ran a daily process to download every English-language weblog entry posted on the previous day, classify the post as either *story* or *non-story*, and save story posts into our repository. This daily process ran successfully for 316 out of 365 days in 2010, with failures largely due to power outages, server maintenance, and changes in the Spinn3r.com API. In total, we processed 621 million English-language weblog posts, and classified 10.4 million posts as personal stories (1.67%).

At 37GB of text, this 2010 story corpus is over twice the size of the English-language documents in Project Gutenberg, and nearly 20 times larger than the corpus used in our experiments. In our fourth experiment, we repeated the PMI-based approach used in our first experiment, but using this new corpus of 10M stories to calculate word co-occurrence statistics. As before, we calculated these statistics using variously sized word windows. Table 5 presents the performance of these systems on the COPA evaluation, and compares these new results to the best-performing PMI-based system using the smaller corpus.

System	*Dev*	*Test*	*All*
PMI 1M Stories (W=25)	60.0	65.2	62.6
PMI 10M Stories (W=5)	60.4	64.4	62.4
PMI 10M Stories (W=15)	62.4	64.8	63.6
PMI 10M Stories (W=25)	62.8	65.4	64.1
PMI 10M Stories (W=30)	61.6	63.8	62.7
PMI 10M Stories (W=50)	61.8	63.2	62.5

Table 5. COPA evaluation results for PMI-based systems using a corpus of 10 million stories, compared to the best result using 1 million stories (W=25). Gains achieved in the best 10M system (W=25) are not statistically significant.

These results show that the larger 10M story corpus yields a very slight gain in performance over 1 million stories, but these gains are not statistically significant. Comparing table 5 to table 2, we again see that a PMI word window size of 25 produces the best results.

In subsequent experiments, we investigated whether our two search-based approaches (table 4) could be improved by using this larger story corpus. We observed modest gains that were not competitive with the PMI approach.

Discussion

We draw three conclusions from the four sets of experiments presented in this paper. First, we have shown that the personal stories that people write in their weblogs are a good source of commonsense causal information. The strong performance of PMI-based techniques with a moderate word window size (W=25) suggests that this causal information exists largely within the scope of adjacent clauses and sentences, which is consistent with analyses by discourse psychologists on other narrative genres.

Second, the relatively low performance of our search-based systems suggests that the causal information in personal stories is best left implicit. Using discourse parsing to explicitly identify causal structure yielded results similar to the sentence-proximity approaches that ignored this structure, and none of the approaches that considered clause or sentence boundaries were competitive with the PMI-based approach.

Third, the approach that used ten million stories yielded the best results, but one million was probably enough. The very slight, insignificant gains obtained by calculating PMI

scores on the larger corpus suggest that the utility of this information had nearly plateaued by the first million stories. The continued creation of larger and larger story corpora will have diminishing returns, at least for approaches that rely on word co-occurrence.

The Choice of Plausible Alternatives (COPA) evaluation was an important enabling tool in this research, but the ability to select between plausible causal relations does not, in itself, constitute a full solution to the larger problem of open-domain commonsense causal reasoning. We believe that future research is needed to develop innovative ways of turning these discriminative approaches into generative reasoning systems, and to integrate them into larger AI architectures. This paper shows that using PMI statistics from a corpus of personal stories is an appropriate first step in this direction.

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

Burton, K., Java, A., & Soboroff, I (2009) The ICWSM 2009 Spinn3r Dataset. International Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA.

Carlson, L. & Marcu, D. (2001) Discourse tagging manual. Technical Report ISI-TR-545, Information Sciences Institute.

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

Clark, P., & Harrison, P. (2009) Large-Scale Extraction and Use of Knowledge From Text. Fifth International Conference on Knowledge Capture (KCAP-09).

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

Etzioni, O., Banko, M., Soderland, S. & Weld, D. (2008) Open Information Extraction from the Web. Communications of the ACM 51(12):68-74.

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

Gordon, A. & 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., & Schubert, L. (2009) Weblogs as a Source for Extracting General World Knowledge. Fifth International Conference on Knowledge Capture (KCAP-09).

Halevy, A., Norvig, P., & Pereira, F. (2009) The Unreasonable Effectiveness of Data. IEEE Intelligent Systems March-April 2009, pp. 8-12.

Magliano, J. (1999) Revealing inference processes during text comprehension. In S. Goldman, A. Graesser, and P. van den Broek (Eds) Narrative Comprehension, Causality, and Coherence: Essays in Honor of Tom Trabasso. Mahwah, NJ: Erlbaum.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990) Introduction to WordNet: An on-line lexical database. International Journal of Lexicography 3(4):235–312.

Morgenstern, L. (2010) Common Sense Problem Page. Retrieved 2/2011 from http://www-formal.stanford.edu/leora/commonsense/

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

Ounis, I., Lioma, C., Macdonald, C., & Plachouras, V. (2007) Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. Upgrade 7(1):49-56.

Rink, B., Bejan, C., & Harabagiu, S. (2010) Learning Textual Graph Patterns to Detect Causal Event Relations. 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS'10), Daytona Beach, FL.

Ritter, R., Soderland, S., & Etzioni, O. (2009) What Is This, Anyway: Automatic Hypernym Discovery. 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read, Stanford, CA.

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

Sagae, S. (2009) Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. 11th International Conference on Parsing Technologies (IWPT' 09), Paris, France.

Schank, R. (1982) Dynamic Memory: A Theory of Reminding and Learning in Computers and People. New York: Cambridge University Press.

Schank, R. (1991) Where's the AI? AI Magazine 12(4):38-48.

Schank, R. & Abelson, R. (1995). Knowledge and memory: The real story. In Wyer, R. S. (Ed.), Knowledge and memory: The real story. Advances in Social Cognition, 8, 1-85.

Schubert, L. (2002) Can we derive general world knowledge from texts? Human Language Technology (HLT-02), San Diego, CA.

Schubert, L. & Tong, M. (2003) Extracting and evaluating general world knowledge from the Brown corpus. HLT/NAACL 2003 Workshop on Text Meaning, Edmonton, Alberta, Canada.

Shanahan, M. (2004). An attempt to formalise a non-trivial benchmark problem in common sense reasoning. Artificial Intelligence 153(1-2):141-165.

Swanson, R. (2010) Enabling open domain interactive storytelling using a data-driven case-based approach. Doctoral dissertation, University of Southern California.

Trabasso, T. & van den Broek, P. (1985) Causal thinking and the representation of narrative events. Journal of Memory and Language 24: 612-630.

van den Broek, P. (1995) Comprehension and memory of narrative text: Inference and coherence. In M. A. Gernsbacher (Ed.), Handbook of psycholinguistics (pp. 539-588). New York: Academic Press.

Van Durme, B., Michalak, P., & Schubert, L. (2009) Deriving generalized knowledge from corpora using WordNet abstraction, European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece.