Open-domain Commonsense Reasoning Using Discourse Relations from a Corpus of Weblog Stories

Matt Gerber
Department of Computer Science
Michigan State University
gerberm2@msu.edu

Andrew S. Gordon and **Kenji Sagae**
Institute for Creative Technologies
University of Southern California
{gordon,sagae}@ict.usc.edu

Abstract

We present a method of extracting open-domain commonsense knowledge by applying discourse parsing to a large corpus of personal stories written by Internet authors. We demonstrate the use of a linear-time, joint syntax/discourse dependency parser for this purpose, and we show how the extracted discourse relations can be used to generate open-domain textual inferences. Our evaluations of the discourse parser and inference models show some success, but also identify a number of interesting directions for future work.

1 Introduction

The acquisition of open-domain knowledge in support of commonsense reasoning has long been a bottleneck within artificial intelligence. Such reasoning supports fundamental tasks such as textual entailment (Giampiccolo et al., 2008), automated question answering (Clark et al., 2008), and narrative comprehension (Graesser et al., 1994). These tasks, when conducted in open domains, require vast amounts of commonsense knowledge pertaining to states, events, and their causal and temporal relationships. Manually created resources such as FrameNet (Baker et al., 1998), WordNet (Fellbaum, 1998), and Cyc (Lenat, 1995) encode many aspects of commonsense knowledge; however, coverage of causal and temporal relationships remains low for many domains.

Gordon and Swanson (2008) argued that the commonsense tasks of prediction, explanation, and imagination (collectively called *envisionment*) can be supported by knowledge mined from a large corpus of personal stories written by Internet weblog authors.[1] Gordon and Swanson (2008) identified three primary obstacles to such an approach. First, stories must be distinguished from other weblog content (e.g., lists, recipes, and reviews). Second, stories must be analyzed in order to extract the implicit commonsense knowledge that they contain. Third, inference mechanisms must be developed that use the extracted knowledge to perform the core envisionment tasks listed above.

In the current paper, we present an approach to open-domain commonsense inference that addresses each of the three obstacles identified by Gordon and Swanson (2008). We built on the work of Gordon and Swanson (2009), who describe a classification-based approach to the task of story identification. The authors' system produced a corpus of approximately one million personal stories, which we used as a starting point. We applied efficient discourse parsing techniques to this corpus as a means of extracting causal and temporal relationships. Furthermore, we developed methods that use the extracted knowledge to generate textual inferences for descriptions of states and events. This work resulted in an end-to-end prototype system capable of generating open-domain, commonsense inferences using a repository of knowledge extracted from unstructured weblog text. We focused on identifying

[1] We follow Gordon and Swanson (2009) in defining a story to be a "textual 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 author or a close associate is among the participants."

strengths and weaknesses of the system in an effort to guide future work.

We structure our presentation as follows: in Section 2, we present previous research that has investigated the use of large web corpora for natural language processing (NLP) tasks. In Section 3, we describe an efficient method of automatically parsing weblog stories for discourse structure. In Section 4, we present a set of inference mechanisms that use the extracted discourse relations to generate open-domain textual inferences. We conclude, in Section 5, with insights into story-based envisionment that we hope will guide future work in this area.

2 Related work

Researchers have made many attempts to use the massive amount of linguistic content created by users of the World Wide Web. Progress and challenges in this area have spawned multiple workshops (e.g., those described by Gurevych and Zesch (2009) and Evert et al. (2008)) that specifically target the use of content that is collaboratively created by Internet users. Of particular relevance to the present work is the weblog corpus developed by Burton et al. (2009), which was used for the data challenge portion of the International Conference on Weblogs and Social Media (ICWSM). The ICWSM weblog corpus (referred to here as Spinn3r) is freely available and comprises tens of millions of weblog entries posted between August 1st, 2008 and October 1st, 2008.

Gordon et al. (2009) describe an approach to knowledge extraction over the Spinn3r corpus using techniques described by Schubert and Tong (2003). In this approach, logical propositions (known as *factoids*) are constructed via approximate interpretation of syntactic analyses. As an example, the system identified a factoid glossed as "doors to a room may be opened". Gordon et al. (2009) found that the extracted factoids cover roughly half of the factoids present in the corresponding Wikipedia[2] articles. We used a subset of the Spinn3r corpus in our work, but focused on discourse analyses of entire texts instead of syntactic analyses of single sentences. Our goal was to extract general causal and temporal propositions instead of the fine-grained properties expressed by many factoids extracted by Gordon et al. (2009).

Clark and Harrison (2009) pursued large-scale extraction of knowledge from text using a syntax-based approach that was also inspired by the work of Schubert and Tong (2003). The authors showed how the extracted knowledge tuples can be used to improve syntactic parsing and textual entailment recognition. Bar-Haim et al. (2009) present an efficient method of performing inference with such knowledge.

Our work is also related to the work of Persing and Ng (2009), in which the authors developed a semi-supervised method of identifying the causes of events described in aviation safety reports. Similarly, our system extracts causal (as well as temporal) knowledge; however, it does this in an open domain and does not place limitations on the types of causes to be identified. This greatly increases the complexity of the inference task, and our results exhibit a corresponding degradation; however, our evaluations provide important insights into the task.

3 Discourse parsing a corpus of stories

Gordon and Swanson (2009) developed a supervised classification-based approach for identifying personal stories within the Spinn3r corpus. Their method achieved 75% precision on the binary task of predicting *story* versus *non-story* on a held-out subset of the Spinn3r corpus. The extracted "story corpus" comprises 960,098 personal stories written by weblog users. Due to its large size and broad domain coverage, the story corpus offers unique opportunities to NLP researchers. For example, Swanson and Gordon (2008) showed how the corpus can be used to support open-domain collaborative story writing.[3]

As described by Gordon and Swanson (2008), story identification is just the first step towards commonsense reasoning using personal stories. We addressed the second step - knowledge extraction - by parsing the corpus using a Rhetorical Structure Theory (Carlson and Marcu, 2001) parser based on the one described by Sagae (2009). The parser performs joint syntactic and discourse dependency

[2] http://en.wikipedia.org

[3] The system (called SayAnything) is available at http://sayanything.ict.usc.edu

parsing using a stack-based, shift-reduce algorithm with runtime that is linear in the input length. This lightweight approach is very efficient; however, it may not be quite as accurate as more complex, chart-based approaches (e.g., the approach of Charniak and Johnson (2005) for syntactic parsing).

We trained the discourse parser over the causal and temporal relations contained in the RST corpus. Examples of these relations are shown below:

(1) [$_{cause}$ Packages often get buried in the load] [$_{result}$ and are delivered late.]

(2) [$_{before}$ Three months after she arrived in L.A.] [$_{after}$ she spent $120 she didn't have.]

The RST corpus defines many fine-grained relations that capture causal and temporal properties. For example, the corpus differentiates between *result* and *reason* for causation and *temporal-after* and *temporal-before* for temporal order. In order to increase the amount of available training data, we collapsed all causal and temporal relations into two general relations *causes* and *precedes*. This step required normalization of asymmetric relations such as *temporal-before* and *temporal-after*.

To evaluate the discourse parser described above, we manually annotated 100 randomly selected weblog stories from the story corpus produced by Gordon and Swanson (2009). For increased efficiency, we limited our annotation to the generalized *causes* and *precedes* relations described above. We attempted to keep our definitions of these relations in line with those used by RST. Following previous discourse annotation efforts, we annotated relations over clause-level discourse units, permitting relations between adjacent sentences. In total, we annotated 770 instances of *causes* and 1,009 instances of *precedes*.

We experimented with two versions of the RST parser, one trained on the fine-grained RST relations and the other trained on the collapsed relations. At testing time, we automatically mapped the fine-grained relations to their corresponding *causes* or *precedes* relation. We computed the following accuracy statistics:

Discourse segmentation accuracy For each predicted discourse unit, we located the reference discourse unit with the highest overlap. Accuracy for the predicted discourse unit is equal to the percentage word overlap between the reference and predicted discourse units.

Argument identification accuracy For each discourse unit of a predicted discourse relation, we located the reference discourse unit with the highest overlap. Accuracy is equal to the percentage of times that a reference discourse relation (of any type) holds between the reference discourse units that overlap most with the predicted discourse units.

Argument classification accuracy For the subset of instances in which a reference discourse relation holds between the units that overlap most with the predicted discourse units, accuracy is equal to the percentage of times that the predicted discourse relation matches the reference discourse relation.

Complete accuracy For each predicted discourse relation, accuracy is equal to the percentage word overlap with a reference discourse relation of the same type.

Table 1 shows the accuracy results for the fine-grained and collapsed versions of the RST discourse parser. As shown in Table 1, the collapsed version of the discourse parser exhibits higher overall accuracy. Both parsers predicted the *causes* relation much more often than the *precedes* relation, so the overall scores are biased toward the scores for the *causes* relation. For comparison, Sagae (2009) evaluated a similar RST parser over the test section of the RST corpus, obtaining precision of 42.9% and recall of 46.2% ($F_1 = 44.5\%$).

In addition to the automatic evaluation described above, we also manually assessed the output of the discourse parsers. One of the authors judged the correctness of each extracted discourse relation, and we found that the fine-grained and collapsed versions of the parser performed equally well with a precision near 33%; however, throughout our experiments, we observed more desirable discourse segmentation when working with the collapsed version of the discourse parser. This fact, combined with the results of the automatic evaluation presented above,

	Fine-grained RST parser			Collapsed RST parser		
Accuracy metric	causes	precedes	overall	causes	precedes	overall
Segmentation	36.08	44.20	36.67	44.36	30.13	43.10
Argument identification	25.00	33.33	25.86	26.15	23.08	25.87
Argument classification	66.15	50.00	64.00	79.41	83.33	79.23
Complete	22.20	28.88	22.68	31.26	21.21	30.37

Table 1: RST parser evaluation. All values are percentages.

led us to use the collapsed version of the parser in all subsequent experiments.

Having developed and evaluated the discourse parser, we conducted a full discourse parse of the story corpus, which comprises more than 25 million sentences split into nearly 1 million weblog entries. The discourse parser extracted 2.2 million instances of the *causes* relation and 220,000 instances of the *precedes* relation. As a final step, we indexed the extracted discourse relations with the Lucene information retrieval engine.[4] Each discourse unit (two per discourse relation) is treated as a single document, allowing us to query the extracted relations using information retrieval techniques implemented in the Lucene toolkit.

4 Generating textual inferences

As mentioned previously, Gordon and Swanson (2008) cite three obstacles to performing commonsense reasoning using weblog stories. Gordon and Swanson (2009) addressed the first (story collection). We addressed the second (story analysis) by developing a discourse parser capable of extracting causal and temporal relations from weblog text (Section 3). In this section, we present a preliminary solution to the third problem - reasoning with the extracted knowledge.

4.1 Inference method

In general, we require an inference method that takes as input the following things:

1. A description of the state or event of interest. This is a free-text description of any length.
2. The type of inference to perform, either causal or temporal.
3. The inference direction, either forward or backward. Forward causal inference produces the effects of the given state or event. Backward causal inference produces causes of the given state or event. Similarly, forward and backward temporal inferences produce subsequent and preceding states and events, respectively.

As a simple baseline approach, we implemented the following procedure. First, given a textual input description d, we query the extracted discourse units using Lucene's modified version of the vector space model over TF-IDF term weights. This produces a ranked list R_d of discourse units matching the input description d. We then filter R_d, removing discourse units that are not linked to other discourse units by the given relation and in the given direction. Each element of the filtered R_d is thus linked to a discourse unit that could potentially satisfy the inference request.

To demonstrate, we perform forward causal inference using the following input description d:

(3) John traveled the world.

Below, we list the three top-ranked discourse units that matched d (left-hand side) and their associated consequents (right-hand side):

1. traveling the world → to murder
2. traveling from around the world to be there → even though this crowd was international
3. traveled across the world → to experience it

In a naïve way, one might simply choose the top-ranked clause in R_d and select its associated clause as the answer to the inference request; however, in the example above, this would incorrectly generate "to murder" as the effect of John's traveling (this is

[4] Available at http://lucene.apache.org

more appropriately viewed as the *purpose* of traveling). The other effect clauses also appear to be incorrect. This should not come as much of a surprise because the ranking was generated soley from the match score between the input description and the causes in R_d, which are quite relevant.

One potential problem with the naïve selection method is that it ignores information contained in the ranked list R'_d of clauses that are associated with the clauses in R_d. In our experiments, we often observed redundancies in R'_d that captured general properties of the desired inference. Intuitively, content that is shared across elements of R'_d could represent the core meaning of the desired inference result. In what follows, we describe various re-rankings of R'_d using this shared content. For each model described, the final inference prediction is the top-ranked element of R'_d.

Centroid similarity To approximate the shared content of discourse units in R'_d, we treat each discourse unit as a vector of TF scores. We then compute the average vector and re-rank all discourse units in R'_d based on their cosine similarity with the average vector. This favors inference results that "agree" with many alternative hypotheses.

Description score scaling In this approach, we incorporate the score from R_d into the centroid similarity score, multiplying the two and giving equal weight to each. This captures the intuition that the top-ranked element of R'_d should represent the general content of the list but should also be linked to an element of R_d that bears high similarity to the given state or event description d.

Log-length scaling When working with the centroid similarity score, we often observed top-ranked elements of R'_d that were only a few words in length. This was typically the case when components from sparse TF vectors in R'_d matched well with components from the centroid vector. Ideally, we would like more lengthy (but not too long) descriptions. To achieve this, we multiplied the centroid similarity score by the logarithm of the word length of the discourse unit in R'_d.

Description score/log-length scaling In this approach, we combine the description score scaling and log-length scaling, multiplying the centroid similarity by both and giving equal weight to all three factors.

4.2 Evaluating the generated textual inferences

To evaluate the inference re-ranking models described above, we automatically generated forward/backward causal and temporal inferences for five documents (265 sentences) drawn randomly from the story corpus. For simplicity, we generated an inference for each sentence in each document. Each inference re-ranking model is able to generate four textual inferences (forward/backward causal/temporal) for each sentence. In our experiments, we only kept the highest-scoring of the four inferences generated by a model. One of the authors then manually evaluated the final predictions for correctness. This was a subjective process, but it was guided by the following requirements:

1. The generated inference must increase the local coherence of the document. As described by Graesser et al. (1994), readers are typically required to make inferences about the text that lead to a coherent understanding thereof. We required the generated inferences to aid in this task.

2. The generated inferences must be globally valid. To demonstrate global validity, consider the following actual output:

   (4) I didn't even need a jacket (until I got there).

   In Example 4, the system-generated forward temporal inference is shown in parentheses. The inference makes sense given its local context; however, it is clear from the surrounding discourse (not shown) that a jacket was not needed at any point in time (it happened to be a warm day). As a result, this prediction was tagged as incorrect.

Table 2 presents the results of the evaluation. As shown in the table, the top-performing models are those that combine centroid similarity with one or both of the other re-ranking heuristics.

Re-ranking model	Inference accuracy (%)
None	10.19
Centroid similarity	12.83
Description score scaling	17.36
Log-length scaling	12.83
Description score/log-length scaling	16.60

Table 2: Inference generation evaluation results.

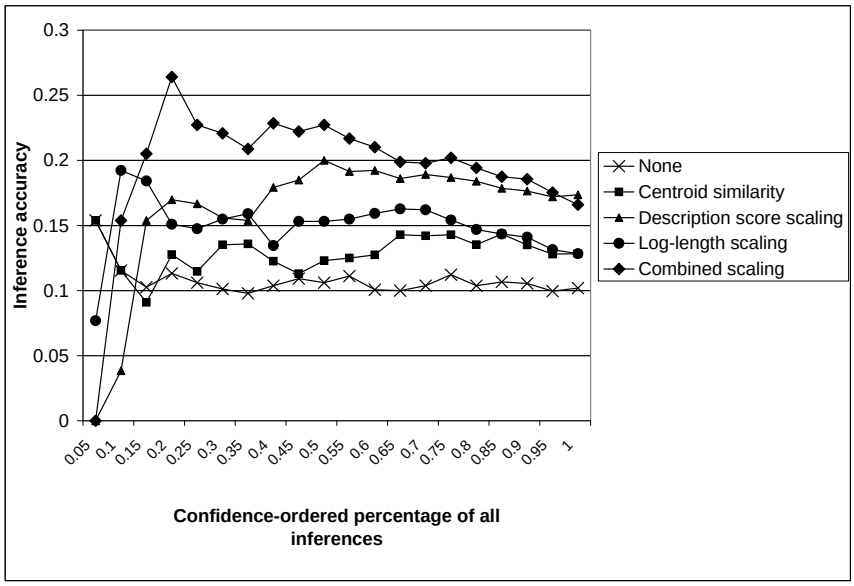

Figure 1: Inference rate versus accuracy. Values along the *x*-axis indicate that the top-scoring *x*% of all inferences were evaluated. Values along the *y*-axis indicate the prediction accuracy.

The analysis above demonstrates the relative performance of the models when making inferences for all sentences; however it is probably the case that many generated inferences should be rejected due to their low score. Because the output scores of a single model can be meaningfully compared across predictions, it is possible to impose a threshold on the inference generation process such that any prediction scoring at or below the threshold is withheld. We varied the prediction threshold from zero to a value sufficiently large that it excluded all predictions for a model. Doing so demonstrates the trade-off between making a large number of textual inferences and making accurate textual inferences. Figure 1 shows the effects of this variable on the re-ranking models. As shown in Figure 1, the highest inference accuracy is reached by the re-ranker that combines description score and log-length scaling with the centroid similarity measure. This accuracy is attained by keeping the top 25% most confident inferences.

5 Conclusions

We have presented an approach to commonsense reasoning that relies on (1) the availability of a large corpus of personal weblog stories and (2) the ability to analyze and perform inference with these stories. Our current results, although preliminary, suggest novel and important areas of future exploration. We group our observations according to the last two problems identified by Gordon and Swanson (2008): story analysis and envisioning with the analysis results.

5.1 Story analysis

As in other NLP tasks, we observed significant performance degradation when moving from the training genre (newswire) to the testing genre (Internet

weblog stories). Because our discourse parser relies heavily on lexical and syntactic features for classification, and because the distribution of the feature values varies widely between the two genres, the performance degradation is to be expected. Recent techniques in parser adaptation for the Brown corpus (McClosky et al., 2006) might be usefully applied to the weblog genre as well.

Our supervised classification-based approach to discourse parsing could also be improved with additional training data. Causal and temporal relations are instantiated a combined 2,840 times in the RST corpus, with a large majority of these being causal. In contrast, the Penn Discourse TreeBank (Prasad et al., 2008) contains 7,448 training instances of causal relations and 2,763 training instances of temporal relations. This represents a significant increase in the amount of training data over the RST corpus. It would be informative to compare our current results with those obtained using a discourse parser trained on the Penn Discourse TreeBank.

One might also extract causal and temporal relations using traditional semantic role analysis based on FrameNet (Baker et al., 1998) or PropBank (Kingsbury and Palmer, 2003). The former defines a number of frames related to causation and temporal order, and roles within the latter could be mapped to standard thematic roles (e.g., cause) via SemLink.[5]

5.2 Envisioning with the analysis results

We believe commonsense reasoning based on weblog stories can also be improved through more sophisticated uses of the extracted discourse relations. As a first step, it would be beneficial to explore alternate input descriptions. As presented in Section 4.2, we make textual inferences at the sentence level for simplicity; however, it might be more reasonable to make inferences at the clause level, since clauses are the basis for RST and Penn Discourse TreeBank annotation. This could result in the generation of significantly more inferences due to multi-clause sentences; thus, more intelligent inference filtering will be required.

Our models use prediction scores for the tasks of rejecting inferences and selecting between multiple candidate inferences (i.e., forward/backward causal/temporal). Instead of relying on prediction scores for these tasks, it might be advantageous to first identify whether or not envisionment should be performed for a clause, and, if it should, what type and direction of envisionment would be best. For example, consider the following sentence:

(5) [$clause_1$ John went to the store] [$clause_2$ because he was hungry].

It would be better - from a local coherence perspective - to infer the cause of the second clause instead of the cause of the first. This is due to the fact that a cause for the first clause is explicitly stated, whereas a cause for the second clause is not. Inferences made about the first clause (e.g., that John went to the store because his dog was hungry), are likely to be uninformative or in conflict with explicitly stated information.

Example 5 raises the important issue of context, which we believe needs to be investigated further. Here, context refers to the discourse that surrounds the clause or sentence for which the system is attempting to generate a textual inference. The context places a number of constraints on allowable inferences. For example, in addition to content-based constraints demonstrated in Example 5, the context limits pronoun usage, entity references, and tense. Violations of these constraints will reduce local coherence.

Finally, the story corpus, with its vast size, is likely to contain a significant amount of redundancy for common events and states. Our centroid-based re-ranking heuristics are inspired by this redundancy, and we expect that aggregation techniques such as clustering might be of some use when applied to the corpus as a whole. Having identified coherent clusters of causes, it might be easier to find a consequence for a previously unseen cause.

In summary, we have presented preliminary research into the task of using a large, collaboratively constructed corpus as a commonsense knowledge repository. Rather than relying on hand-coded ontologies and event schemas, our approach relies on the implicit knowledge contained in written natural language. We have demonstrated the feasibility of obtaining the discourse structure of such a corpus via linear-time parsing models. Furthermore,

[5] Available at http://verbs.colorado.edu/semlink

we have introduced inference procedures that are capable of generating open-domain textual inferences from the extracted knowledge. Our evaluation results suggest many opportunities for future work in this area.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. 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

Collin Baker, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California. Morgan Kaufmann Publishers.

Roy Bar-Haim, Jonathan Berant, and Ido Dagan. 2009. A compact forest for scalable inference over entailment and paraphrase rules. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1065, Singapore, August. Association for Computational Linguistics.

K. Burton, A. Java, and I. Soboroff. 2009. The icwsm 2009 spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media*.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging manual. Technical Report ISI-TR-545, ISI, July.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

Peter Clark and Phil Harrison. 2009. Large-scale extraction and use of knowledge from text. In *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*, pages 153–160, New York, NY, USA. ACM.

Peter Clark, Christiane Fellbaum, Jerry R. Hobbs, Phil Harrison, William R. Murray, and John Thompson. 2008. Augmenting WordNet for Deep Understanding of Text. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 45–57. College Publications.

Stefan Evert, Adam Kilgarriff, and Serge Sharoff, editors. 2008. *4th Web as Corpus Workshop Can we beat Google?*

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2008. The fourth fascal recognizing textual entailment challenge. In *Proceedings of the First Text Analysis Conference*.

Andrew Gordon and Reid Swanson. 2008. Envisioning with weblogs. In *International Conference on New Media Technology*.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media*.

Jonathan Gordon, Benjamin Van Durme, and Lenhart Schubert. 2009. Weblogs as a source for extracting general world knowledge. In *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*, pages 185–186, New York, NY, USA. ACM.

A. C. Graesser, M. Singer, and T. Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*, 101:371–395.

Iryna Gurevych and Torsten Zesch, editors. 2009. *The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources*.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and Lexical Theories*.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344, Morristown, NJ, USA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2009. Semi-supervised cause identification from aviation safety reports. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 843–851, Suntec, Singapore, August. Association for Computational Linguistics.

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie

Webber. 2008. Penn discourse treebank version 2.0. Linguistic Data Consortium, February.

Kenji Sagae. 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'09)*, pages 81–84, Paris, France, October. Association for Computational Linguistics.

Lenhart Schubert and Matthew Tong. 2003. Extracting and evaluating general world knowledge from the brown corpus. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*, pages 7–13, Morristown, NJ, USA. Association for Computational Linguistics.

Reid Swanson and Andrew Gordon. 2008. Say anything: A massively collaborative open domain story writing companion. In *First International Conference on Interactive Digital Storytelling*.