Mining Commonsense Knowledge From Personal Stories in Internet Weblogs

Andrew S. Gordon GORDON@ICT.USC.EDU
Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Rey, CA 90292 USA

Abstract

Recent advances in automated knowledge base construction have created new opportunities to address one of the hardest challenges in Artificial Intelligence: automated commonsense reasoning. In this paper, we describe our recent efforts in mining commonsense knowledge from the personal stories that people write about their lives in their Internet weblogs. We summarize three preliminary investigations that involve the application of statistical natural language processing techniques to corpora of millions of weblog stories, and outline our current approach to solving a number of outstanding technical challenges.

1. Introduction

The utility of information extraction in specialized domains with specialized reasoning tasks is clearly evident, e.g. in the automated creation of databases of protein interactions by mining biomedical journal articles (Yang et al., 2010). However, recent advances in *open information extraction* (Etzioni et al., 2008) have also captured the attention of researchers in automated commonsense reasoning, whose failures over the last fifty years 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 (e.g. Hobbs & Gordon, 2010). 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) Skeptical but open-minded, the commonsense reasoning researcher is forced to ponder: Is there an automated data-driven solution for the construction of commonsense knowledge bases?

Several recent research efforts have taken up this challenge. 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 factoids 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 textual entailment recognition.

Although factoid extraction of this sort has proven useful for some tasks, benchmark problems in commonsense reasoning (Morgenstern, 2010) also require knowledge about the way situations progress over time, i.e. causal and temporal inference over states and events. Accordingly, a number of recent efforts have looked at automatically extracting causal and temporal knowledge from unstructured textual data. 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. Chambers and Jurafsky (2009) induce temporal schemas from news articles using unsupervised methods, capitalizing on co-reference of arguments across sentences and the temporal ordering of narrated events that is common in this genre.

Assessing this line of research overall, the results are mixed. Each of these previous efforts provides some encouraging results, but a robust automated data-driven solution for the construction of commonsense knowledge bases remains elusive. However, this research has given us some insights into what issues need to be addressed. The first issue is *structure*; each of these approaches begins by identifying the implicit structure of textual data, either the syntactic or co-reference structure, over which patterns are identified that capture semantic relationships. Some sorts of textual structure are easier to identify than others, particularly for certain genres. Conversely, different sorts of commonsense inferences may capitalize on different sorts of textual structure. Successful solutions will therefore be found in the overlap, where the structure that can be accurately identified in text is the same that can be utilized by the inference process.

Because of this, the second critical issue is *genre*. Robust parsers for identifying syntactic and semantic structure currently are largely limited to the textual domain of the newspaper article, a direct consequence of availability of existing treebanks (e.g. Marcus et al., 1993). However, we would not expect that all genres are equally good for the identification of commonsense knowledge about the everyday world. Indeed, we would expect that only those genres that constitute discourse about the everyday world are going to be useful. Successful solutions will therefore be applied to textual resources that are both easy to parse and are used to relay information about everyday life. Additionally, this genre will need to be abundantly available as electronic text on a massive scale.

In this paper, we describe our recent efforts toward the automated creation of commonsense knowledge bases, exploring the relationship between issues of genre and structure in this task. This work specifically looks at the unique properties of the genre of the personal story, for which millions of examples are readily available in Internet weblogs. We argue that the personal story is ideally suited as a target for commonsense knowledge extraction, and summarize three of our previous preliminary efforts in extracting knowledge from large story corpora. Motivated by lessons learned from these previous efforts, we outline our current research directions aimed at overcoming the central technical challenges.

2. Personal Stories in Internet Weblogs

There is a long history of interest in the genre of the personal story within computer science, particularly in the area of Artificial Intelligence. Most associated with this interest is Roger Schank, who led decades of research on the role of stories in cognition and who advised scores of graduate students pursuing theses on this topic. Schank's central idea about stories, best articulated by Schank (1982) and Schank and Abelson (1995), was that all knowledge is story-based. Specifically, 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 (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.

Ironically, a major criticism of Schank's technical approach in this area is the same one that he directed toward his contemporaries (Schank, 1991): it did not scale. Although his students managed to create computational cognitive models that demonstrated different aspects of this dynamic memory, none of them got much past the *1000-item barrier*. In short, if all of the knowledge in a system needs to be hand-authored by some smart knowledge engineer, then the maximum bits of knowledge in the end will be that which a single PhD student can author over the course of a 3-5 year thesis project, i.e. 1000 items. In the 20 years since this argument was made, the Internet and statistical natural language processing have drastically changed the way that researchers approach the problem of scale.

Over the last several years, our research group has undertaken several large-scale story collection efforts, focusing on the personal stories that people occasionally write in their Internet weblogs. In our most recent estimate, roughly 5.4% of all non-spam weblog posts are personal stories, using the following definition for this genre: 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. With nearly one million new weblog posts made each day (Technorati, 2008), millions of personal stories about the everyday lives of people have been posted on the web over the last year.

In each of our attempts to identify large story collections from weblog entries, we have employed supervised machine learning approaches to text classification, applying a trained story classifier to the textual data in large corpora of weblog entries (Gordon et al., 2007; Gordon & Swanson, 2009). These attempts have differed primarily in the corpus of weblog entries analyzed, the text classification approach employed, and the size of the annotated training data. In the second of these attempts (Gordon & Swanson, 2009), we 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 by a commercial weblog aggregator (Spinn3r.com). To classify each entry as story or non-story, we trained a confidence-weighted linear classifier (Dredze et al., 2008) with unigram lexical features obtained from five thousand annotated training examples (precision = 66%, recall = 48%, F-score = 0.55). Although these results are rather modest compared to some other applications of statistical text classification in other natural language processing tasks, they demonstrate that at least some of characteristics that are particular to this genre can be recognized in the surface-level features of the text, but still be independent of any topic domain or narrative context. In short, people predictably use different words with different frequencies when they are telling stories compared to other forms of discourse.

In pursuing this line of research, we have often reflected on Schank's original ideas about the role of stories in cognition, as experiences that are remembered when they violate one's expectations. Aggregated over millions of people, the theory predicts that the stories in our collection would define the *fringe of expectation*, the

experiences of humankind that are novel enough to be remembered and which fall outside of the commonsense knowledge that people share. If this were true, then the stories themselves would not constitute a commonsense knowledge base, but rather would only outline the bodies of commonsense expectations that are violated in the everyday world.

Looking at the data, however, suggests that a more nuanced theory of storytelling is needed as it relates to personal weblogs, cf. Langellier and Peterson (2004). Although our collections are certainly ripe with amazing stories of unusual and unexpected events, the more typical story might best be described as the *narration of the mundane*. Personal weblogs lend themselves to stories about what happened during a day at work or school, stories of casual conversations with friends, and even stories about sitting around doing nothing. While something about these events was interesting enough to be offered by the storyteller, it is often not evident in the narration itself. As a consequence, our corpora of personal stories collected from weblogs constitute an enormous source of knowledge about the everyday lives of bloggers. Instead of defining the fringe of expectation, it elaborates what is actually expected.

The weblog personal story, as a narration of the mundane, is ideally suited as a target for commonsense knowledge extraction. Having settled the issue of *genre*, the main concern of our ongoing research has been the issue of *structure*. The remainder of this paper summarizes our previous work, and outlines our current research approach.

3. Three Preliminary Investigations

We conducted three preliminary investigations into the mining of commonsense knowledge from large-scale corpora of weblog stories. In these attempts, our goal was not to generate the sorts of formal axioms that are typical of commonsense reasoning research. Instead, we focused on the problem of generating valid commonsense inferences in the open domain, i.e. given an arbitrary antecedent as input. Each approach employed a different inference mechanism, and each was evaluated using different metrics.

3.1 Language Modeling With Verb-Patient Pairs

In our first attempt (Manshadi et al., 2008) we sought to develop a probabilistic model of event sequences, with the assumption that the order of events in narratives has some relationship to the order of events in the real world. Our approach was to represent weblog stories as ordered sequences of events, one event per sentence, where each event was encoded as a lexical bigram composed of the main verb of the sentence and the head word of the patient argument of that verb. By treating whole stories as a "sentence" of verb-patient pairs, we applied standard statistical language modeling methods to induce a probabilistic model of event sequences. To illustrate this idea, the following five sentences are shown with their main verbs and patient head words indicated with single and double underlines, respectively.

1. I got a flat tire coming back from the concert.
2. I thought it was just the bad pavement at first.
3. The guy next to me pointed to my tire, so I pulled over.
4. I wish I had a better towing plan, though.
5. I blew a few hundred dollars to get the car home.

Our aim was to capture information about story events, but do so in a way that facilitated judgments of similarity between different events in different stories. For example, these five events were compactly encoded as five verb-patient pairs: got/tire, thought/was, pointed/tire, wish/had, and blew/dollars.

We developed an approach to automatically identifying main verbs and patient head words using only part-of-speech tags, without requiring full syntactic parses of sentences. We applied this approach to 66 million sentences from one of our early weblog story collections (Gordon et al., 2007), yielding hundreds of thousands of ordered verb-patient pairs. To induce a probabilistic model of event sequences, we treated each pair as a single lexical token, and the ordering of pairs in a single story as a sentence. This allowed us to apply existing language modeling tools (Stolcke, 2002) to the entire corpus, which we used to calculate the probability of any given event sequence.

We devised two simple evaluations of the utility of our model, each using event sequences identified from held-out story data. First, we showed that this model could correctly distinguish between a real narrative event sequence and a random ordering of the same events 63% of the time, significantly above random chance (50%). Second, we showed that this model could correctly distinguish between the actual event that follows a given event sequence and a random event only 53% of the time, barely above random chance (50%).

Neither of these results was particularly encouraging. However, this investigation raised a number of important questions for us to address in our subsequent efforts. In particular, we noted that the representation of story events was critical step, and we had last a lot of information by equating all sentences that shared the same main verb and patient head word. Also, we were dissatisfied with our evaluation setup, primarily because it focused on predicting events in other narratives, rather than using information obtained from narratives to make commonsense inferences about the real world. We feel that this is also a significant drawback of related work in this area, e.g. the narrative cloze evaluation used by Chambers and Jurafsky (2008, 2009).

3.2 Similarity-Based Narrative Event Prediction

In our second attempt (Swanson & Gordon, 2008) we again held the assumption that the order of events in a narrative has some relationship to the order of events in the real world, but took a new approach to the problem of representing and comparing narrative events. Instead of attempting to equate compact representations of narrative events, we treated the text of each sentences as its own best representation, and used traditional textual similarity metrics to identify events that are most similar. In this model, the story corpus is represented in its original form, as a large number of sets of ordered sentences. Event prediction is then cast a retrieval problem. Any arbitrary event can serve as an input antecedent, and is itself represented as a sentence of arbitrary narrative text. This antecedent is then used a as query to the entire story corpus, identifying the sentence that is most similar to the antecedent from some story in the collection. The inferred consequence is then simply the *next* sentence that appears in that story.

The technical implementation of this approach was straightforward. We stored each of the 66 million sentences from one of our early weblog story collections (Gordon et al., 2007) into a relational database along with their ordering information. We then indexed the corpus using the Apache Lucene search engine (Gospodnetic & Hatcher, 2004), treating each sentence as a separate document. Input sentences are treated as Lucene queries, identifying the most similar sentence in the corpus, and then the database is consulted to identify the sentence that followed after it. Despite the simplicity of the approach, the resulting inferences are sometimes remarkable, as in the following input-output pair:

I: The weather broke, so we sailed out of the harbor.
O: As Victoria grew nearer, the waves grew larger and we furled some foresail and turned to run.

As a commonsense inference about *what happens next*, this result is not ideal. In this one example, we can already see a number of problems that will limit the utility of this approach for open-domain commonsense reasoning. The output is not an *expectation* about the next event, but rather an *instance* of a next event in a specific narrative context. Still, the inference tells you that when sailing out of a harbor, storms (Victoria), waves, and furling the foresail are going to be task-relevant.

Although not ideal for commonsense expectation, Swanson and Gordon (2008) saw the direct utility of this approach as an interactive storytelling application. In the *Say Anything* interactive storytelling system, a user and the computer take turns contributing sentences to an emerging fictional storyline. The user's contributions are limited only by their creativity. The computer's contributions are produced using this simple retrieval-based approach, using only the most recent sentence contributed by the user as the query. For example, the input-output pair presented above was the first of a longer exchange from one user of this system. The contributions of the user (I) and the system (O) for subsequent turns are listed below:

I: We sailed at about 9 knots with good trim, but the storm eventually caught up with us.
O: With its big open cockpit and heavy nose, I didn't like its chances in the kind of sea you get out there almost continuously that time of year.

I: Sure enough the boat was completely inadequate, and we were tossed into the cold ocean.
O: Everyone in our group of seven tourists -- five locals and a Japanese couple -- was pretty excited about the experience.

I: The Japanese couple were the ones that saved us though, with their expert swimming abilities.
O: as far as that goes it was just the four of us.

I: The last tourist was lost at sea, never to be found.
O: Drowned or murdered, the bloated, stinking bodies that turn up by the hundreds will look much the same.

The *Say Anything* interactive storytelling system solves a critical content-authoring problem that has debilitated research in interactive storytelling systems throughout the history of this area of research (Spierling & Szilas, 2009). The solution, to crowdsource the authoring task to millions of weblog authors, shares much in common with the work of Hays and Efros (2007) for visual scene completion using millions of photographs. Here missing or removed sections of digital photographs are automatically replaced by finding a close match using the remaining photographic content as a query, and stitching the content from the retrieved photograph into the query. The coherence of the resulting image is poor when the size of the corpus is small (e.g. 1000 photographs), but is remarkably believable when web-scale collections are employed. In both cases, the massive scale of the corpora reduces the hard problem of aligning the critical features of two instances (unification) to one of matching surface features of the data.

The *Say Anything* interactive storytelling system forced us to question a basic assumption of both of our first two attempts to extract commonsense knowledge from stories, that the order of events in a narrative has some relationship to the order of events in the real world. While this may still be true, the interpretation and narration of experienced events contributes at least as much to the way that an author orders the sentences in narrative discourse. To identify only those relationships in stories that pertain to the way the world works, which would serve as data for inducing commonsense knowledge, a deeper interpretation of the *structure* of narrative text was needed.

3.3 Following Discourse Relations

In our third attempt (Gerber et al., 2010) we continued to represent events as the text of narrative sentences, but we abandoned the linear ordering of narrated events as a proxy for the relationship between these events in the real world. Instead, we attempted to identify the underlying discourse structure of each of the narratives in our collection, and use the semantics of these discourse relations to support specific types of commonsense inferences.

In particular, our focus was on the causal and temporal relationships that exist between textual clauses. Consider, for example, the nine discourse elements in the following excerpt from a weblog story:

1. I arrived at 5pm on Monday evening and
2. quickly made my way to the hotel.
3. My electric razor is now in the hands of whoever took it out of my checked suitcase, but
4. the joke is on them because
5. I left the plug at home so
6. they might get about a weeks worth of shaves out of it, not to mention
7. the whole voltage difference.
8. I was able to purchase a disposable razor at an "Auto Service".
9. These are little convenient stores mostly run by Koreans down in Paraguay.

Here the narrative ordering of events differs from the temporal ordering in the world, which might be:

$$7 \rightarrow 9 \rightarrow 5 \rightarrow 1 \rightarrow 2 \rightarrow 8 \rightarrow 3 \rightarrow 4 \rightarrow 6$$

In addition, there are causal relationships that exist between these nine events, including the following three:

$$3 \Rightarrow 8$$
$$3 \wedge 5 \wedge 7 \Rightarrow 4$$
$$3 \wedge 5 \wedge 7 \Rightarrow 6$$

The automatic identification of discourse relations such as these has been a focus of research since the creation of large annotated corpora that could serve as training data for supervised approaches. Two such corpora are the Rhetorical Structure Theory corpus (Carlson et al., 2001) and the Penn Discourse Treebank (Miltsakaki et al., 2004), both of which provide annotations for various causal and temporal discourse relations (among others) between discourse elements in news article text. Sagae (2009) describes an automated approach to discourse parsing in the style of the Rhetorical Structure Theory corpus, employing a stack-based linear-time shift-reduce algorithm that parallels the method used in shift-reduce parsing for syntactic dependencies.

We applied this discourse parser directly to the story corpus of Gordon and Swanson (2009), containing just fewer than one million weblog stories with more than 25 million sentences. The discourse parser identified 2.2 million causal relations and 220 thousand temporal relations between discourse elements in this corpus.

As in our second effort, we again treated the inference task as a retrieval problem, but this time used the identified discourse relations instead of the narrative ordering of sentences to support inference. We loaded each causal and temporal relation into a database table, and indexed the text of each discourse unit as a separate document using the Apache Lucene search engine. This approach supported inferences of four types: forward and backward inferences for both causal and temporal relationships. As with our previous effort, this approach sometimes yielded very encouraging results, as in the following input-output pair for a causal relationship:

I: John traveled the world
O: to experience it.

We devised a new scheme for evaluating the effectiveness of this inference approach on arbitrary narrative text. We selected five random stories from our collection, and passed each sentence in these stories as an antecedent query to our inference engine (256 sentences x 4 inference types = 1024 consequent inferences). We then had a human rater manually evaluate the reasonableness of each generated inference. Using the default Lucene retrieval method, only 10.19% of the generated inferences were judged as valid by our human rater. Although 10% is much better than the 0% obtained by all previous open-domain causal/temporal reasoning systems (there are none), there remained lots of room for improvement.

In trying a number of different schemes, we identified two strategies that greatly improved the accuracy of this inference engine. First, we noted that the most similar discourse unit to the antecedent is often related with an atypical consequence, when compared to the consequences that would have been inferred from other highly-ranked discourse unit matches. We devised a method for pooling consequences for the n-best matches, and then selecting the most typical item in this list as the inferred consequence. Here the most typical item was judged as the one with the shortest cosine distance to the centroid of the set, where each discourse element is represented as a vector of term frequency scores. Balancing antecedent similarity and consequence centrality, we raised the accuracy of inferences to 17.36%. The lesson we learned here was that aggregating evidence improves the accuracy of inferences.

Second, we noted that a poor match between the input antecedent and the closest discourse unit in the corpus invariably led to a poor inferred consequence. We explored whether incorporating a simple threshold on the retrieval score would prevent spurious inferences. We found that different retrieval schemes reacted differently to the use of a threshold, with the highest accuracy achieved by accepting only the top 25% of confidence-ordered inferences, yielding an inference accuracy of

27%. The lesson we learned here was that filtering low-confidence inferences improves accuracy.

4. Current Work

Our three previous attempts to mine commonsense knowledge from personal stories in weblogs helped us formulate a new research approach to this problem, and highlighted a number of key technical challenges that must be overcome. In this section, we outline the current strategy of our ongoing research, which is informed by the key lessons learned from our previous work. We describe our strategy in reference to four central issues that must be addressed: structure, aggregation, confidence, and evaluation.

First, our last attempt demonstrated to us that it is necessary to identify the implicit discourse structure of stories in our collection. While we still feel that there is much to be gained by identifying causal and temporal links between discourse elements, our last attempt highlighted some inadequacies of existing methods. The discourse parser that we employed, described in Sagae (2009), achieves state-of-the-art performance on the news text with which it was trained. Still, the overall performance on document-level discourse parsing of news text is quite low (0.45 F-score). In Gerber et al. (2010), we estimated that the actual accuracy of correctly identified and linked causal and temporal relations in weblog story text was around 30%, at best. Barring some major advancement in both discourse parsing and domain adaptation, the utility of general-purpose discourse processors will be unacceptably low.

Instead, much of our current effort is devoted to the development of a high-precision discourse parser that is specifically trained to identify causal and temporal links in weblog story text. Primarily, the challenge is obtaining a suitably large corpus of annotated training data, with acceptably high levels of inter-rater agreement. In favoring high precision at the expense of recall performance, we see an increasing need for even larger corpora of weblog stories. Accordingly, we are developing technologies for identifying all of the English-language personal stories that exist in all weblogs on the Internet, facilitated by partnerships with commercial weblog aggregators.

Second, we have seen that the aggregation of evidence across multiple discourse relations improves the accuracy of inferences. In our third attempt, we saw accuracy improve significantly when we selected the more prototypical consequences among a large pool. We have been investigating alternative techniques that would allow us to pool evidence from multiple antecedents as well.

One insight we had was that this aggregation of evidence could be achieved by clustering all of the discourse elements that participate in a causal or temporal relationship across the entire story corpus. That is, instead of working with millions of individual discourse relationships in a massive database table, the data would be represented as hundreds of thousands of text clusters, each with hundreds of similar discourse elements. The individual discourse relations would then be aggregated at the cluster-level, created a giant graph of millions of directed links between hundreds of thousands of clusters. As in our previous work, we expect that standard textual similarity metrics will serve as sufficient distance measures for use in the clustering process, although efficiency becomes a major concern at this scale.

Third, we believe that this clustering approach affords a novel method of assessing the confidence of any given inference. As in our previous attempts, an arbitrary antecedent would be provided as input to retrieve the closest of all discourse elements in the corpus using standard text retrieval methods. The cluster governing the closest match would serve as the proxy for the antecedent, however, and all of the clusters linked to it would be considered as potential consequents. The most *likely* consequent cluster is assumed to be the one that is most *frequent*, i.e. has the most links directly to elements in the cluster from the antecedent cluster. Similarly, the relative likelihood of any two arbitrary consequences could be approximated using the relative frequencies of links from the antecedent cluster. Although mathematically dubious, this approach would provide future Artificial Intelligence systems with some basis for making arbitrary judgments of relative likelihood, based on the collective experiences narrated by millions of weblog authors.

Fourth, each of our previous attempts has wrestled with the problem of evaluation. As open-domain commonsense inference is not a challenge that is widely pursued, there exists no accepted set of benchmark problems to use to measure the relative strengths of competing approaches. Needed are gold-standard test sets and automated evaluation tools, which would enable a level of performance tuning that is impossible when results must be evaluated by hand. A focus of our current work is the development of a test set similar to those used in the Recognizing Textual Entailment (RTE) challenges (Dagan et al., 2006). In doing so, we aim to encourage researchers who have been successful in those challenges to explore the commonalities between the problems of textual entailment and commonsense inference.

5. Conclusions

The automated creation of web-scale knowledge bases is fast becoming a mature technology, with enormous potential for immediate, commercially viable applications. Equally important, however, is the potential for these technologies to address long-standing problems in the basic research of Artificial Intelligence. The problem of commonsense reasoning has been resistant to progress since it was first identified over 50 years ago (McCarthy, 1958), and continues to struggle with the

question: Who is going to author all of the commonsense knowledge required for human-like intelligence? The prospect of open information extraction from web text suggests a novel answer: It will be authored by millions of people who publish on the web.

In this paper we have argued that the genre of the personal story, as written in Internet weblogs, is ideally suited as a target for the automated construction of commonsense knowledge bases. Abundant and collectable, personal stories are unique among other genres of web text, not on the *fringe of expectation*, but rather as *narrations of the mundane*. We summarized three of our preliminary efforts to mine commonsense knowledge from large corpora of weblog stories. Each of these efforts yielded mixed results, but helped us identify the key issues that must be addressed in order to make progress in this area. We outlined our current research efforts, focusing on the four issues of structure, aggregation, confidence, and evaluation.

Our aim in presenting this overview of ongoing work is to encourage the cross-fertilization of ideas among members of this research community, and further afar. Our hope is that others see opportunities to apply their methods and technologies to the unique genre of the personal story, and to improve upon the approaches we have pursued to exploit web-scale personal story collections.

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. In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA, May 2009.

Carlson, L., Marcu, D., and Okurowski, M. (2001) Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Chambers, N. & Jurafsky, D. (2009) Unsupervised Learning of Narrative Schemas and their Participants. Association for Computational Linguistics 2009, Singapore.

Chambers, N. and Jurafsky, D. (2008) Unsupervised Learning of Narrative Event Chains. Annual Meeting of the Association for Computational Linguistics 2008, Ohio.

Clark, P., and Harrison, P. (2009) Large-Scale Extraction and Use of Knowledge From Text. In Proc Fifth Int Conf on Knowledge Capture (KCap'09).

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

Davis, E. & Morgenstern, L. (2004) Introduction: Progress in formal commonsense reasoning, Artificial Intelligence, Volume 153, Issues 1-2, Logical Formalizations and Commonsense Reasoning, March 2004, Pages 1-12.

Dredze, M., Crammer, K., and Pereira, F. (2008) Confidence-weighted linear classification. In Proceedings of the 25th international Conference on Machine Learning, July 5-9, Helsinki, Finland.

Etzioni, O., Banko, M., Soderland, S. and 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. Proceedings of the 2010 NAACL Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR).

Gordon, A. and 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.

Gordon, A., Cao, Q., and Swanson, R. (2007) Automated Story Capture From Internet Weblogs. Proceedings of the Fourth International Conference on Knowledge Capture, October 28-31, 2007, Whistler, BC.

Gordon, J., Van Durme, B., Schubert, L. (2009) Weblogs as a Source for Extracting General World Knowledge. In the Proc. of the Fifth International Conference on Knowl-edge Capture (K-CAP 2009).

Gospodnetic, O. and Hatcher, E. (2004) Lucene in Action. Greenwich, CT: Manning Publications.

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

Hays, J. & Efros, A. (2007) Scene Completion Using Millions of Photographs, ACM Transactions on Graphics (SIGGRAPH 2007), August 2007, vol. 26, No. 3.

Hobbs, J. & Gordon, A. (2010) Goals in a Formal Theory of Commonsense Psychology. Proceedings of the 6th International Conference on Formal Ontology in

Information Systems (FOIS-2010), Toronto, Canada, May 11-14, 2010.

Langellier, K. & Peterson, E. (2004) Storytelling in Daily Life. Philadelphia, PA: Temple University Press.

Manshadi, M., Swanson, R., and Gordon, A. (2008) Learning a Probabilistic Model of Event Sequences From Internet Weblog Stories. Twenty-first International Conference of the Florida AI Society, Applied Natural Language Processing track, May 15-17, 2008, Coconut Grove, FL.

Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993) Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics, 19(2), June 1993.

McCarthy, J. (1958) Programs with common sense. In Proceedings of the Symposium on the Mechanization of Thought Processes, Teddington, England: HMSO.

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

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004) In Proceedings of the Language Resources and Evaluation Conference. Lisbon, Portugal.

Morgenstern, L. (2010) Common Sense Problem Page. Retrieved May 3, 2010 from http://www-formal.stanford.edu/leora/commonsense/

Rink, B., Bejan, C., & Harabagiu, S. (2010) Learning Textual Graph Patterns to Detect Causal Event Relations. In Proceedings of the 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS'10), Applied Natural Language Processing track, Daytona Beach, FL, USA, May 2010.

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

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

Schank, R. (1982) Dynamic Memory: A theory of reminding and learning in computers and people. New York, NY: 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? Proceedings of Human Language Technology 2002, March 24-27, San Diego, CA, pp. 84-87.

Schubert, L. and Tong, M. (2003) Extracting and evaluating general world knowledge from the Brown corpus, Proc. of the HLT/NAACL 2003 Workshop on Text Meaning, May 31, Edmonton, Alberta, Canada.

Spierling, U. and Szilas, N. (2009) Authoring Issues beyond Tools. Proceedings of the International Conference on Interactive Digital Storytelling 2009, pp. 50-61.

Stolcke, A. (2002) SRILM: An Extensible Language Modeling Toolkit, International Conference on Spoken Language Processing, Denver, Colorado.

Swanson, R. and Gordon, A. (2008) Say Anything: A Massively Collaborative Open Domain Story Writing Companion. First International Conference on Interactive Digital Storytelling, Erfurt, Germany, November 26-29, 2008.

Technorati.com (2008) State of the Blogosphere 2008. Available at http://technorati.com/blogging/state-of-the-blogosphere/

Van Durme, B., Michalak, P., and Schubert, L. (2009) Deriving generalized knowledge from corpora using WordNet abstraction, 12th Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL-09), Mar. 30 - Apr. 3, 2009, Athens, Greece.

Yang, Z., Lin, H., and Li, Y. (2010) BioPPISVMExtractor: A protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. Journal of Biomedical Informatics 43(2010):88-96.