Automated Story Capture From Internet Weblogs

Andrew S. Gordon, Qun Cao, and Reid Swanson
USC Institute for Creative Technologies
Marina del Rey, CA USA
{ gordon | qcao | swansonr }@ict.usc.edu

ABSTRACT

Among the most interesting ways that people share knowledge is through the telling of stories, i.e. first-person narratives about real-life experiences. Millions of these stories appear in Internet weblogs, offering a potentially valuable resource for future knowledge management and training applications. In this paper we describe efforts to automatically capture stories from Internet weblogs by extracting them using statistical text classification techniques. We evaluate the precision and recall performance of competing approaches. We describe the large-scale application of story extraction technology to Internet weblogs, producing a corpus of stories with over a billion words.

Categories and Subject Descriptors

I.2.7. *Natural Language Processing*

General Terms

Algorithms

Keywords

Storytelling, Knowledge Management, Weblogs

STORIES IN INTERNET WEBLOGS

Among the most interesting ways that people share knowledge is through the telling of stories, i.e. first-person narratives about real-life experiences. Few genres of communication are easier to produce or consume, or are more capable of transferring tacit knowledge from experts to novices in any domain. These characteristics of stories have attracted the attention of developers of knowledge management systems and training applications. This has generated interest in computational support for the capture, management, analysis, and telling of stories in innovative ways.

Much of the early work in this area has focused on story collections of modest size, with few knowledge management or training applications incorporating as many as a thousand individual stories, e.g. [3]. Scaling up these technologies by several orders of magnitude will require a qualitative shift in the methods used to manage story content, shifting away from manual collection and analysis toward fully automated story processing.

The future utility of fully automated story processing technology is most obvious when considering the phenomenal rise of Internet weblogging over the last several years. There were an estimated 70 million weblogs in March 2007 (http://www.technorati.com), each with a series of entries containing the thoughts and ramblings of some computer user. We analyzed the entries of a random sample of 100 weblogs (approximately 12,000 sentences or 200,000 words) and found that 17% of weblog text consisted of stories, using the annotations guidelines of previous research [1]. By extrapolation, we can estimate that there are 23.8 billion words of story text available on the web for use in knowledge management and training applications. However, using the web as a story repository will require the development of accurate means of separating story from non-story content, and subsequently integrating story content into story-based applications.

In this paper, we focus on the problem of separating story from non-story content in Internet weblogs. First, we describe our development and evaluation of a set of automated story extraction approaches using machine-learning techniques for text classification. Second, we describe the application of one of these approaches to Internet weblogs on a large scale, producing a story collection of over one billion words.

STATISTICAL STORY EXTRACTION

Gordon & Ganesan [1] first demonstrated the feasibility of automatically extracting stories from text using statistical natural language processing techniques. The aim of their work was to develop technologies for identifying stories in conversational speech data, e.g. the automatically recognized words from audio recordings of interviews with subject-matter experts. To explore similar approaches for written weblog text, we re-implemented the Gordon & Ganesan system. This system is based on a Naïve Bayes machine learning algorithm that assigns a story/non-story classification to a segment of 50 words of weblog text, trained using 100 random weblogs (105 thousand words) annotated with story/non-story labels. The feature set includes unigram and bigram binary features, ignoring case and punctuation, which appear six or more times in the annotated training data. The binary classification is applied iteratively to overlapping segments of the input text, yielding a sequence of story/non-story classifications along with confidence scores. These confidence scores are then smoothed using a

K-CAP'07, October 28–31, 2007, Whistler, British Columbia, Canada.

mean-average function, and sequences of consecutive text segments with positive confidence scores are labeled as story content.

Additionally, we explored a number of enhancements to this previous work, and identified two variations that yielded increased performance. These two variations both made story/non-story classifications at the sentence level, incorporating an automated sentence delimiter [5] as part of the preprocessing stage. Each was provided with roughly twice the amount of training data as used to re-implement the Gordon & Ganesan system, with sentence-level story/non-story annotations. Each is based on a support vector machine (SVM) learning algorithm that uses n-gram features encoded as the log of their normalized frequency in the sentence. A Gaussian filter is used to smooth the classification confidence values across sentences. The second variation also included features for the part-of-speech tags of the words in the sentence, identified using a maximum-entropy algorithm [4].

We evaluated the comparative performance of each of these systems on 12,000 annotated sentences (using 10-fold cross validation to train and test the latter two systems), yielding precision, recall, and equally weighted F-scores as follows:

Table 1. Comparative story extraction performance

System	Precision	Recall	F-score
Gordon & Ganesan	0.302	0.829	0.414
n-grams only	0.464	0.606	0.509
n-grams + part-of-speech	0.497	0.455	0.463

We found that our re-implementation of the original Gordon & Ganesan approach achieved high recall, and that the SVM variation that included part-of-speech features achieved the highest precision. Removing these features improved recall, and achieved the highest F-score performance. In additional experiments, we found no gains could be achieved by including additional syntactic features or by using different smoothing functions.

CAPTURING STORIES FROM WEBLOGS

To explore the application of automated story extraction technology, we developed a large-scale system that automatically discovered weblogs, downloaded their entries, and extracted the stories contained in them.

To acquire a large database of Internet addresses of weblogs, we utilized an application programming interface provided by a major commercial Internet weblog search engine, Technorati.com. To obtain URLs using the API, we submitted thousands of queries using vocabulary from an existing broad-coverage knowledge base of commonsense activities [2]. Search results were then processed to identify unique addresses, resulting in a set of over 390,000 Internet weblogs. Over a period of 373 days, we downloaded and processed the entries contained in 39.9% of these weblogs, with an average of 21.8 entries per weblog (3.4 million entries processed).

Each weblog entry was then processed by our re-implementation of the original Gordon & Ganesan approach to story extraction, which favors high recall performance over precision. On average, 1.32 distinct segments in each weblog entry were labeled as story text, resulting in a corpus of 4.5 million extracted story segments consisting of 1.06 billion words.

We subsequently applied an automated sentence delimiting algorithm [5] to each of the extracted segments. After removing sentence fragments from the beginning of and end of each segment, the corpus contained 3.7 million segments with a total of 66.5 million sentences.

CONCLUSIONS

Scaling up the use of stories in knowledge management and training applications by orders of magnitude will require the application of new techniques that emphasize automation. In this paper we have demonstrated the applicability of machine learning approaches to statistical text classification to the task of automatically extracting stories from Internet weblog text, and applied this technology to create a large-scale corpus of stories. Future work in this area must be directed toward fully automated techniques for analyzing story corpora of this scale, and integrating resources of this size into effective knowledge management and training applications.

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

[1] Gordon, A. & Ganesan, K. 2005. Automated story capture from conversational speech. Third international conference on knowledge capture (KCAP-05), Banff, Canada.

[2] Gordon, A. 2001. Browsing image collections with representations of commonsense activities. Journal of the American Society for Information Science and Technology 52(11):925-929.

[3] Johnson, C., Birnbaum, L., Bareiss, R., & Hinrichs, T. 2000. War stories: Harnessing organizational memories to support task performance. Intelligence 11(1):16-31.

[4] Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. Proceedings EMNLP, University of Pennsylvania.

[5] Reynar, J. & Ratnaparkhi, A. 1997. A maximum entropy approach to identifying sentence boundaries. Proceedings of ANLP, Washington, D.C.