Automated Story Capture From Conversational Speech

Andrew S. Gordon
USC Institute for Creative Technologies
Marina del Rey, CA USA
gordon@ict.usc.edu

Kavita Ganesan
University of Southern California.
Los Angeles, CA USA
ganesan@usc.edu

ABSTRACT
While storytelling has long been recognized as an important part of effective knowledge management in organizations, knowledge management technologies have generally not distinguished between stories and other types of discourse. In this paper we describe a new type of technological support for storytelling that involves automatically capturing the stories that people tell to each other in conversations. We describe our first attempt at constructing an automated story extraction system using statistical text classification and a simple voting scheme. We evaluate the performance of this system and demonstrate that useful levels of precision and recall can be obtained when analyzing transcripts of interviews, but that performance on speech recognition data is not above what can be expected by chance. This paper establishes the level of performance that can be obtained using a straightforward approach to story extraction, and outlines ways in which future systems can improve on these results and enable a wide range of knowledge socialization applications.

Categories and Subject Descriptors
I.2.7 *Natural Language Processing*

General Terms
Algorithms.

Keywords
Storytelling, Knowledge Management.

KNOWLEDGE SOCIALIZATION
Much of the knowledge that is shared among members of communities and organizations is exhibited only in the telling of stories in spoken conversations. While support for storytelling in organizations has been long recognized as important to effective knowledge management [2][12], few attempts have been made to specifically support storytelling through technology. Instead of creating specific technologies for automatically capturing, analyzing, and routing stories that are naturally told in conversations, knowledge management technology development has targeted the more general problem of supporting computer-mediated communication, without much regard to the genre of the content [9]. In not distinguishing between storytelling and other types of human-human communication, today's knowledge management technologies fail to exploit the value of stories in packaging and transmitting tacit knowledge [13], understanding organizational change [10], and driving the development of professional training applications [5]. Furthermore, knowledge management technologies have had difficulty breaking out of the mold of traditional networked groupware applications, which limits their applicability to the fraction of people who spend their days working at computer terminals.

A different vision for knowledge management technology is one that is specifically targeted to the capture and use of the stories told in communities and organizations in the context of normal, spoken conversations. The role of technology would be to support the capture of stories from spoken conversations, perform a task-directed analysis of its content, and present stories or analysis of stories to people in service of their organizational tasks.

As a hypothetical example, consider the utility of story management technology for militaries with active-duty soldiers who are hospitalized due to injuries sustained in battle. These soldiers may volunteer to have a story management system monitor the conversations that they conduct with hospital staff, visitors, and other patients. Audio clips of stories would be extracted and analyzed for content, and connections would be drawn between the experiences of these soldiers and people who are developing new military doctrine and in need of their first-hand subject-matter expertise. While they are waiting in hospitals for their injuries to heal, they could then be helping to ensure that other soldiers learn from their experiences.

In this paper, we focus on a key enabling technology to support this vision of story-based knowledge management, namely the automated identification and extraction of stories in conversational speech. First we discuss issues surrounding the design of story capture devices, and motivate the need for high-accuracy story extraction techniques. We then describe our first attempt at constructing an automated story extraction system using statistical text classification and a simple voting scheme. An evaluation of the performance of this system is described, along with a review of methods that could be used to improve the levels of precision and recall that were obtained. We conclude with a discussion of applications that are enabled by our approach to story extraction.

K-CAP'05, October 2-5, 2005, Banff, Canada.

DESIGN OF STORY CAPTURE DEVICES

Possible story capture systems can be broadly classified into two categories. First are *active story elicitation systems*, in which users interact with a system that is actively supporting and engaging users in storytelling behavior. This approach is best exemplified by recent work on Story Listening Systems [3], where children's storytelling is encouraged through interaction with virtual characters in support of literacy development. A second category of story capture is passive story extraction systems, where the conversations that people have among themselves are monitored for the presence of stories to be automatically extracted. Passive story extraction systems have the advantage that extracted stories more accurately reflect an organization's natural storytelling behavior, but the disadvantage of relying solely on the quality of automated story recognition processes.

In our own work, we have constructed prototypes of both active story elicitation and passive story extraction systems. Our first exploration of this area was in the development of a *Telephony Story Exchange*, an example of an active story elicitation system constructed in the context of the IBM Research Knowledge Socialization project in 2001. The prototype was an interactive telephony application built upon a speaker-independent continuous speech recognition engine for telephone speech. Users would call the Telephony Exchange Server and be prompted to tell their story. At the completion of a users' story, the audio recording of the story would be archived along with the set of words that could automatically be recognized from the users' speech. The system would then use the recognized words to retrieve the most similar story (via word overlap metrics) in the archive, and play back the audio of the retrieved story to the user ("Your story reminds me of something that someone else told me. Here's what they had to say…").

One of the lessons that we learned in prototyping the Telephony Story Exchange concerned the costs and rewards associated with voluntary user story contributions. Our prototype sought to balance the costs associated with user story contributions (time, privacy, phone charges) with the prospect of hearing an entertaining story that is related to the user's interests. Either reducing the costs or increasing the rewards seemed to be necessary if the prototype system was to be adopted.

Our second exploration of this area was to prototype a passive story extraction system, called the *Personal Story Monitor*, which was also based on the use of a telephone. Instead of calling a special story exchange number, however, the Personal Story Monitor would simply listen to the words that the user was speaking in the course of the users normal telephone conversations with other people, and extract and archive the audio segments of those conversations that were automatically judged to be stories (using the story extraction technology described in later sections of this paper). The implementation of this system involved a speaker-dependent speech recognition system that was connected directly to an additional microphone embedded in the user's telephone handset.

While the Personal Story Monitor greatly reduced the time costs associated with users' story contribution, the cost in loss of privacy is substantially greater than in the Telephony Story Exchange. One solution is to ask users to review and approve the audio clips associated with extracted stories before they are archived (after each phone call, or at the end of each day), although this would incur a time cost. The solution that we developed in prototyping the Personal Story Monitor was to enable users to approve the archiving of their stories in real time, as they are telling them, using an interactive visual indicator. When the system judges that the user is telling a story, an LED lamp/switch positioned next to the phone begins flashing, prompting the user to make an approval decision. If the user feels that they are actually telling a story and they want it to be contributed to an archive, they press on the lamp/switch, and the audio associated with the story they are telling (or have just finished telling) is extracted and archived.

Our experiences in prototyping these systems lead us to believe that passive story extraction systems (with interactive approval) may offer an effective means of story capture with minimal costs for users. However, the success of these systems will depend heavily on the quality of the automated story recognition and extraction processes. A story recognition process with low recall performance will fail to prompt the user for approval when stories are actually being told, while poor precision performance will lead to inappropriate prompts that will quickly become annoying to users. A poor extraction process will fail to find suitable start and end points of extracted audio segments, preventing the full automation of a story capture pipeline.

STATISTICAL STORY CLASSIFICATION

The core technical component of passive story capture devices is a story recognition process that accepts speech as input (which is processed by a speech recognition algorithm) and produces as out a classification of the text (story / not-story) and a confidence score. This text processing task is similar in many respects to automated text segmentation by topic boundaries [4][6], and related tasks in topic detection and tracking. However, the aim in story classification is to separate story content from surrounding context regardless of whether a topic shift occurs. Likewise, the task is related to automated genre classification of whole documents, however the aim is to find the specific regions in a text that are appropriately labeled as stories.

This section describes our attempt at creating a story recognition process using statistical text classification techniques, where training data was provided to machine learning algorithms in order to identify the features (words) that are most indicative of story and non-story text classifications. Developing the story recognition process involved the preparation of training data, identifying the number of words to be evaluated for each classification, and trimming the feature set to reduce the time required for each classification.

Preparing the Training Data

To provide training data for the machine-learning algorithm it was necessary to collect a corpus of conversational speech data and have human raters separate story from non-story segments. For this purpose, we began with transcripts of interviews conducted in the Summer of 2004 with 10 U.S. Army officers (rank of Captain), originally conducted as part of a related project to author prototype leadership development applications [7]. These interviews consisted of five sessions with pairs of interviewees, for a total of 292 minutes and 58,222 transcribed words.

These transcripts were given to five human raters tasked with marking the continuous segments of story content in each text. These raters, who were graduate students and staff members at the University of Southern California, were given the following definition of a story:

"The definition of a story is somewhat ambiguous. Generally, the stories that people tell are about events that have happened in the past. Accordingly, people use a lot of past tense verbs (e.g. said, went, gave) when telling stories. However, not all descriptions of events that happened in the past count as stories. Stories give descriptions of specific events that actually occurred, not generalizations over multiple events or times. Stories generally have a sequential structure to them, providing a description of events that happened one after another. Collectively, these events are composed to create a complete narrative. Finally, stories usually have some point to them: the reason that the person is telling the story in the first place. Sometimes stories are truly pointless, though, but some message is usually still conveyed."

Each of the five transcripts was segmented by two of the five raters (a different pair for each transcript), so that inter-rater agreement could be computed between pairs of raters. Agreement was computed using Cohen's Kappa statistic, where raters agreed if they both thought that a given line of text on a printed page included text that was part of a story. Kappa scores were 0.66, 0.80, 0.85, 0.72, and 0.40, for an average score of 0.68.

In order to create the training corpus, two separate files were created of story and non-story content by sorting the transcript texts based on the judgments of the human raters. For each transcript, the judgments of only one of the two raters were used for sorting (all five raters contributed judgments for exactly one of the transcripts). A total of 61 continuous story segments (41.1% of the 58,222 transcribed words in the transcripts) were sorted into the file consisting of story content. An analysis of the 61 continuous story segments identified in these transcripts indicated that the mean length of stories was 390 words ($\sigma = 399$).

Selecting the Word Count

Our approach to statistical text classification is to partition the corpus of story and non-story text into equal-sized segments and encode them as feature vectors to be passed as training data to a machine learning algorithm. To use the classifier, unlabelled input text of the same size would then be encoded as a feature vector and passed to the algorithm for class assignment. Our first development concern is to identify the most appropriate size (word count) for segments of training and input data. A tradeoff exists between classification performance and story size, where setting the word count to a low number ensures that very short stories can be recognized, but at the expense of overall classifier accuracy. Likewise, setting the word count very high will lead to better classification performance on long stories, but risks passing over stories that are very short.

To select the smallest word count that achieves acceptable classification performance, a performance study was conducted for word counts between 25 words and 350 words at 25 word increments. For each increment, the training data was partitioned into equal-size segments of text with the given word count (without regard to sentence or segment boundaries). Both story and non-story text segments were then encoded as feature vectors of single words (unigrams) and pairs of words (bigrams), ignoring punctuation and case (just over 30,000 features). Then each set of feature vectors was used to train a Naïve Bayes machine learning algorithm. Stratified cross-validation (10-fold) was then used to estimate each classifier's performance on unlabeled data. Figure 1 graphs the relative performance of each classifier, both in terms of percent-correct classification and using Cohen's Kappa statistic (accounting for chance agreement).

The Kappa statistic results were used to select the most appropriate word count. A Kappa score of 0.6 is generally considered acceptable in information retrieval research that requires inter-rater agreement between human raters. Naïve Bayes machine learning first achieves this level of agreement (between the classifier and human judgments) at 175 words (Kappa = 0.6152), correctly classifying 80.92% of the 325 training examples in stratified cross validation. If we view the stories in our interviews as representative examples of the length of stories (μ=390 words, σ=399 words), then we expect that 70.50% of stories will be longer than 175 words.

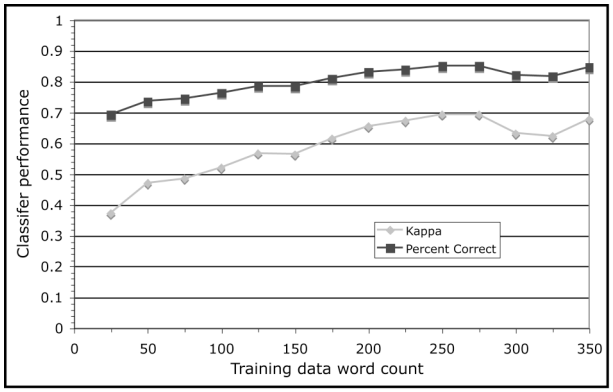

Figure 1. Word count study

The selection of a 175-word count was based on the criteria of minimum classification performance of a classifier operating on the training data, using cross-validation. We recognize that this selection may not yield the best overall performance

when the classifier is integrated into the story extraction process. However, the analysis necessary to determine the best performing word count has not yet been completed.

Reducing the Feature Set

For many applications of story recognition (e.g. real-time interactive approval), it is necessary to have the story/non-story classification operate in a reasonable amount of time. However, the 175-word classifier that we selected did not exhibit reasonable speed, largely due to the amount of time required to construct the feature vector for unlabeled input text (hashing on unigrams and bigrams to identify feature locations). Therefore, our second major challenge was to determine if the size of the feature set could be reduced without compromising classification performance. Our approach was to set minimum values on the number of times that features had to appear in the training data to be included as a feature in the final feature vector. A study was conducted to determine largest value for the minimum feature count such that the classification performance was not degraded. Minimum feature counts from one (all 30,619 features) to 15 (686 features) were evaluated for the 175 word-count classifier using stratified cross validation. Figure 2 graphs the relative performance that was obtained using these different encodings of the training data.

Results indicate that acceptable Kappa scores (above 0.6) can still be obtained by requiring that each feature in the encoded feature vector appears in the training data at least 6 times. A minimum feature count of 6 for the 175 word classifier yields a Kappa value of 0.6086 and correctly classifies 80.62% of the training data in stratified cross validation, while reducing the size of the feature set from 30,619 features to just 2,064.

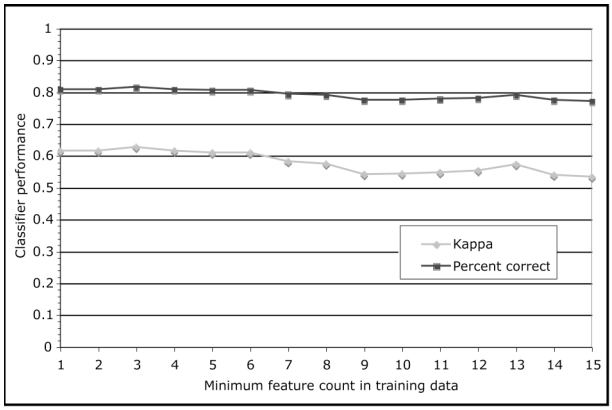

Figure 2. Minimum feature count study

STORY EXTRACTION BY VOTING

Our next challenge was to develop a method of using the statistical story/non-story classifier to automatically extract story segments from streams of text (either transcribed or recognized from speech audio data). Our general strategy is to iteratively apply the story/non-story classifier to successive portions of the input text stream, with the heuristic that adjacent portions of text with the same classification are part of the same story or non-story text segment.

The extraction process begins by classifying the first 175 words in the text stream and recording the confidence level that the classifier assigns to a story or non-story classification. Then the process advances the search ahead by a single word, classifying the second through the 176th word in the stream and recording the confidence level. This process continues ahead one word at a time until the window of 175 words includes the last word in the text stream, for a total number of classifications equal to the number of words in the text stream minus 174.

The second step in the extraction process is to smooth the resulting set of classifications. We selected the simplest method of smoothing classification data, namely to aggregate classification evidence using a voting scheme. The method is to elect either a story or non-story label for each of the words in the input stream, where each classifier that included a given word in its input span casts one vote. In this manner, each of the words in the document receive 175 votes (either story or non-story), except the first and last 175 words in the stream. After each of the classifiers casts its votes, the final label assignment for a word (story/non-story) is then determined via a simple majority.

The third step is to extract sequences of words identified as contiguous stories. The application of this simple voting scheme yields a set of story/non-story label assignments for words with clear transitions between story and non-story word spans. Story boundaries are identified simply as consecutive words with different label assignments.

Figure 3 and 4 show the results of applying the first two steps of this process on a segment of transcribed interview data collected as part of our evaluation (next section). Figure 3 shows a series of 2000 story/non-story classifications obtained by successively applying the 175-word classifier to a segment of 2175 words of text. Figure 4 shows the corresponding tally of votes that each of these 2175 words were given, graphed as how many more (or less) votes were cast for story over non-story labels. Story segments are identified in this data as continuous segments of words where there is one or more greater number of votes for story (5 segments total in Figure 4).

While the advantage of this voting scheme is its simplicity, several disadvantages are apparent. First, the scheme is more susceptible to noise at the very beginning and very end of the input stream, where fewer than 175 classifiers cast votes for class labels. In the extreme, the first (and last) word of the document receives only one vote, as it is included as a word in the input span of only one classifier. Second, the scheme is susceptible to over-smoothing. Spans of story or non-story text that are less than 175 words are ignored because they are outvoted by classifications of text before and after the span.

A straightforward improvement would be to apply a more sophisticated smoothing technique than voting. In signal processing terms, our voting scheme is equivalent to a mean filter with a 175-value kernel width. Smaller sized kernels or

different filters (e.g. Gaussian) may better preserve information that is present in the classification data. In addition, many of these disadvantages would be mediated if the word count of the story/non-story classifier could be made smaller (less than 175 words) while retaining a high level of performance.

For this current study, we opted to evaluate the performance of story extraction using the simple voting scheme, and leave the exploration of different smoothing techniques for future research in this area.

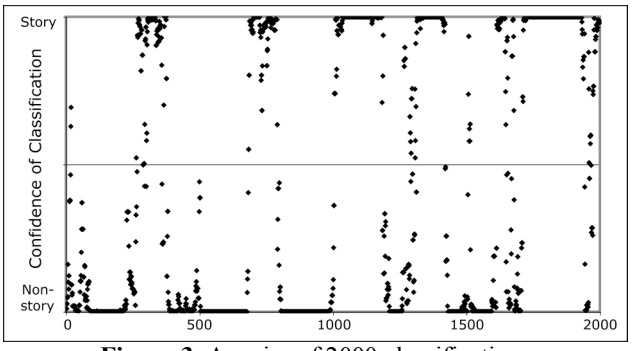

Figure 3. A series of 2000 classifications

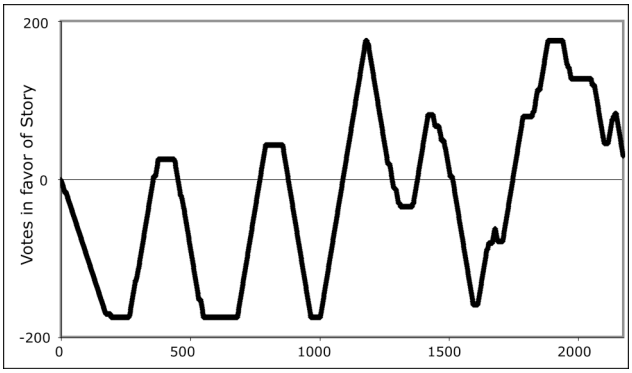

Figure 4. Votes in favor of story for the 2175 words covered by the classifications in Figure 3.

PERFORMANCE EVALUTION

Although our story extraction system was developed using training data consisting of transcripts of audio interviews, the primary aim of our evaluation was to determine the performance of the system on extracting stories from speech data.

Collecting the Test Data

Our test data was collected by conducting interviews with three staff members employed by our research group. These interviews were informal, but where the intent of the interviewer was to encourage the interviewee to tell stories, largely about their occupation and their recreational activities. The duration of these interviews was 28, 39 and 30 minutes, for a total of 97 minutes.

Before their interview, each interviewee participated in a speech-recognition training session using a commercial system (IBM ViaVoice for Windows, Release 10). During the interview, the interviewee wore the same headset microphone used during the speech-recognition training session, and each recognized word spoken by the interviewee was time-stamped and recorded to a log file. The number of words recognized in each interview was 1280, 3438 and 3340, for a total of 8058 recognized words.

In addition, audio from both the interviewer and interviewee was recorded and time-synced to the speech recognition system, enabling us to identify the audio segment associated with each recognized word. These audio recordings were transcribed by hand after each interview. The number of transcribed words for each interviewee was 3424, 4785 and 4663, for a total of 12872 transcribed words. The ratio of recognized words to transcribed words was 0.626:1.

An analysis was conducted to determine the accuracy of the speech-recognition system. We used an error-rate calculation based on minimum edit distance (also known as the Levenshtein distance). The minimum edit distance between the transcript and the recognized words is divided by the total number of words in each transcript. Using this equation, we calculate the accuracy of speech recognition for the three interviews as 0.706, 0.698, and 0.856 (mean = 0.753). These performances levels are strikingly worse than levels that can be achieved for the task of open-vocabulary dictation (the task for which the system was trained), and underscores the differences between the ways that people speak in these two speech contexts. Still, we were interested in determining if the performance of the speech recognition system on conversational speech was sufficient enough to support the story extraction task for conversational speech.

Three raters (from the pool of five raters used in preparing the training data) were each given printed copies of the transcripts and asked to mark story segments. These raters were given the same instructions and definition of a story as they had been given during the markup of the training data, where they had achieved an average inter-rater agreement of Kappa=0.68.

The transcripts were then filtered by removing the lines that were corresponded to utterances made by the interviewer, leaving only the text corresponding to utterances of the interviewee. Each line of printed text in these filtered transcripts was then numbered and labeled with the human judgment of either story or non-story text, which served as the basis for judging the performance of the automated story extraction system. The number of lines in each test set was 249, 377, and 339, for a total of 965 lines. 383 of these lines (39.7%) were labeled as containing story text by human raters.

Analyzing the Test Data

The performance of the story extraction system was evaluated on both the filtered transcript data and the words recognized by the speech recognition software. In each case the input streams were encoded as an ordered set of words, ignoring punctuation and case (as with the original training

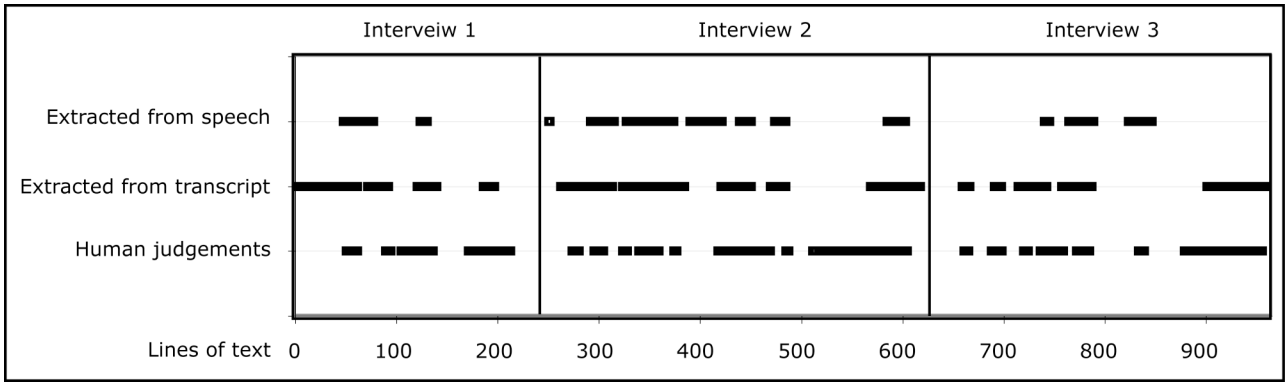

Figure 5. Segments of stories identified in 965 lines of transcribed interview text

data). Story segments were identified as spans of consecutive words that received a majority of votes for story labels using our voting scheme.

The line numbers in the test set corresponding to the start and end words of extracted story segments were identified for both types of input text, where the time codes of words and audio recordings were used to assist in this task for the speech recognition data. The system was judged as having made a correct judgment exactly when both the human rater and the system agreed that some or all of the text on the transcribed line of text was part of a story, or when both agreed that no part of the transcribed line was part of a story.

Figure 5 charts the story segments that were identified in 965 lines of transcript text by human raters, our system operating on the transcript data, and our system operating on the voice recognition data.

With 39.7% of the lines in the test set labeled as story by human raters, some baseline performance scores can be estimated. Randomly selecting a line from the test set as a positive example of a story would be correct 39.7% of the time (baseline precision), while selecting a random 39.7% of the lines in the test set would retrieve 15.8% of the actual story lines (baseline recall). An equally weighted harmonic mean of baseline precision and recall (F-score) would be 0.226. If we instead consider the case where the baseline system simply selects all of the lines as story content (100% recall), then the F-score would be 0.568.

Analyzing the transcript data, the story extraction system labels 455 of the 965 lines in the test set as part of a story, agreeing with the human rater on 241 lines. The precision of the story extraction system for transcript data is 53.0% (241 correct out of 455 lines judged as story), and the recall is 62.9% (241 story lines found out of 383 possible). The F-score for transcription data is 0.575. A calculation of inter-rater agreement between the system and human raters is Kappa=0.253.

Analyzing the voice-recognition data, the story extraction system labels 245 of the 965 lines in the test set as part of a story, agreeing with human raters on 97 lines. The precision of the story extraction system for voice-recognition data is 39.6% (97 correct out of 245 lines judged as story), and the recall is 25.3% (97 story lines found out of 383 possible). The F-score for speech recognition data is 0.309. A calculation of inter-rater agreement between the system and human raters is Kappa=-0.001 (all agreement would be expected by chance).

The poor precision and recall results of story extraction process on speech recognition data is disappointing, particularly because there are many cases where the overlap between human annotations and extracted segments are remarkably close. Figure 6 offers an example of a case where a story just over 100 words was located both through an analysis of the transcript as well as the speech recognition data.

In summary of these results, the performance of our story extraction system is very different when analyzing transcript data than when analyzing speech recognition data. With transcript data, precision and recall levels are high enough to provide technological support for some story management applications. However, when analyzing speech recognition data, only the recall performance of the system is above baseline levels that would be expected by random chance. Our approach to story extraction from speech recognition data does not achieve levels that would realistically support automated story collection from audio only.

Improving the system

There are a number of promising ways to improve the story extraction system, both on transcript and speech data, which we consider to be avenues for future research in this area.

First and foremost, better speech recognition performance would greatly improve story extraction performance. Using speech recognition engines that are tailored specifically for conversational speech (or even interviews) would help a great deal, as well as customizing the language model for words that are found to best discriminate between story and non-story segments. Better speech recognition performance would definitely give us precision and recall levels closer to that of the transcripts of interviews

Second, the quality of the classifier that is used to determine whether a span of text is part of a story could be greatly im-

Human judgment:

740 And one day the boss calls him into his office and says
741 "You cant hang out with her anymore"..and he said "why?".."well you just can't".
742 Now that is really immature…I mean…I understand if you are in third grade you are in
743 the back of the room and you are like chatting with your friend while the teacher is
744 talking, but not in a work environment…you don't tell people who to have lunch with
745 and who not to have lunch with …and I think this boss was systematically telling people
746 not to hang out with me and I don't know why…

Extracted from transcript:

738 just friends from the beginning. there are people you are just friends with
739 and you know you are gonna like and you get along with. And he and I would go to lunch
740 all the time um and hang out. And one day the boss calls him into his office and says
741 "You cant hang out with her anymore"..and he said "why?".."well you just can't".
742 Now that is really immature…I mean…I understand if you are in third grade you are
743 in the back of the room and you are like chatting with your friend while the teacher is
744 talking, but not in a work environment…you don't tell people who to have lunch with
745 and who not to have lunch with …and I think this boss was systematically telling people
746 not to hang out with me and

Extracted from speech recognition data:

744 don't tell people who have lunch with and
745 not to have lunch with and I think his boss was systematically telling people
746 not said how with me and I know new-line
747 never figured out that my other friends would come in the same thing to really
748 stranger he go to them and

Figure 6. A successful story recognition and extraction from transcript and speech recognition data

proved. The challenge is not to increase its accuracy (which already approaches human levels of agreement), but to decrease the word count size. The current size of 175 words corresponds to an average duration of 53 seconds, based on the average rate of speech in the training data. Intuitively, it seems that people are at least an order of magnitude faster at accurately judging when someone has started telling a story to them. While collecting additional training data is an obvious direction, it would also be productive to consider using different types of features besides simple unigrams and bigrams. In particular, the preponderance of past tense and pronouns in stories suggest that grammatical features would be helpful, from simple part-of-speech tags to syntactic parses of phrases. Also, the presence of certain cue phrases [8] may be particularly useful to textual boundary detection, as in seen in other text segmentation research [1]. Non-lexical features, including changes in spoken word rate, inflection, and turn taking in dialogue may each reliably signal the presence of a story boundary as well, as in other speech-based information extraction tasks [11][14].

Third, the voting scheme for story extraction could be improved using the various methods described in the earlier section of this paper, or by replacing the extraction scheme altogether with a different approach. One promising direction is to combine story detection and story extraction into a single search for story boundaries (the beginnings and ends of stories), using statistical techniques analogous to those used for identifying sentence boundaries in speech.

DISCUSSION

While some of the technical challenges of story extraction from interviews may be overcome by pursing the improvements listed in the previous section, the full utility of automated story extraction will not be realized until reliable speaker-independent speech recognition technology is pervasive. Until then, the direct application of this technology may be limited to a number of special cases.

First, there are cases where reliable speaker-dependent speech-recognition models may be available. Insofar as these models achieve error rates similar to human transcriptions on conversational speech, then reasonable recall and precision performance can be achieved using the approach described in this paper. Second, there are many applications of this technology to corpora of interview transcripts that were created for purposes other than story extraction. Three examples are transcripts of police interrogations, transcriptions of courtroom testimony, and closed-captioning of televised interviews for the hearing impaired. The increasing availability of these materials in electronic format make possible today a number of applications involving the automated extraction of stories from these sources.

In the near term, the broadest application of story-extraction technology may be in its application to written text (rather than transcripts of speech), particularly when applied to personal accounts of the events in people's lives (e.g. weblogs). The approach outlined in this paper should work as well (if not better) on written text as on transcriptions of interview text, and could easily be enhanced by incorporating gram-

matical, case, and punctuation features, which were not identified in the speech data used in our current work.

In this paper we have motivated the need for technological support for storytelling, and the capacity for automated story extraction, in particular. We have outlined the design of a number of story capture devices that would be enabled by the development of a reliable story-extraction algorithm that operates on speech data. This paper has described our first attempt at constructing an automated story extraction system using statistical text classification and a simple voting scheme. Our evaluation of this system has demonstrated that useful levels of precision and recall can be obtained when analyzing transcripts of interviews, but that performance on speech recognition data is not above what can be expected by chance. This paper establishes a level of performance for a straightforward approach to story extraction, and outlines ways in which future systems can improve on these results and enable a wide range of knowledge socialization applications.

ACKNOWLEDGMENTS

Thanks to John C. Thomas (IBM TJ Watson Research) for his vision of knowledge socialization and technological support for storytelling. Regina Cabrera, Julia Kim, Stewart King, Joel Maas, Scott Rocher, and Reid Swanson contributed to this research. 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] Beeferman, D., Berger, A., Lafferty, J. (1999) Statistical Models for Text Segmentation. Machine Learning 34(1-3) :177-210.

[2] Boyce, M. E. (1996) Organizational story and storytelling: A critical review. Journal of Organizational Change Management 9(5):5-26.

[3] Cassell, J. (2004) Towards a model of technology and literacy development: Story listening systems. Applied Developmental Psychology 25:75-105.

[4] Choi, F. (2000) Advances in domain independent linear text segmentation. The North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, USA, 2000.

[5] Gordon, A. (2004) Authoring Branching Storylines for Training Applications. Proceedings of the Sixth International Conference of the Learning Sciences (ICLS-04). Santa Monica, CA, June 22-26.

[6] Hearst, M. (1997) TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics , 23 (1), pp. 33-64, March 1997.

[7] Hill, R., Gordon, A., and Kim, J. (2004) Learning the lessons of leadership experience: Tools for interactive case method analysis. Proceedings of the 24th Army Science Conference, Orlando Florida.

[8] Litman, D. (1996) Cue phrase classification using machine learning. Journal of Artificial Intelligence Research 5(1996):53-94.

[9] Marwick, A. (2001) Knowledge Management Technology. IBM Systems Journal 40(4):814-830.

[10] McCormack, C., & Milne, P. (2003) Stories create space for understanding organizational change. Qualitative Research Journal 3(2):46-59.

[11] Shriberg, E. & Stolcke, A. (2000) Prosidy-based automatic segmentation of speech into sentences and topics. Speech Communications 32 (1-2), September 2000.

[12] Snowden, D. (1999) Story telling for the capture and communication of tacit knowledge. Thesis, Indiana University, Bloomington, IN (1995).

[13] Sternberg, R., Forsythe, G., Hedlund, J., Horvath, J., Wagner, R., Williams, W., Snook, S., and Grigorenko, E. (2000) Practical Intelligence in Everyday Life. New York: Cambridge University Press.

[14] Wang, D. & Narayanan, S. (2004) A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues. In Proc. ICASSP, Montreal, Canada, May 2004.