Identifying Personal Stories in Millions of Weblog Entries

Andrew S. Gordon and Reid Swanson

Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
gordon@ict.usc.edu, swansonr@ict.usc.edu

Abstract

Stories of people's everyday experiences have long been the focus of psychology and sociology research, and are increasingly being used in innovative knowledge-based technologies. However, continued research in this area is hindered by the lack of standard corpora of sufficient size and by the costs of creating one from scratch. In this paper, we describe our efforts to develop a standard corpus for researchers in this area by identifying personal stories in the tens of millions of blog posts in the ICWSM 2009 Spinn3r Dataset. Our approach was to employ statistical text classification technology on the content of blog entries, which required the creation of a sufficiently large set of annotated training examples. We describe the development and evaluation of this classification technology and how it was applied to the dataset in order to identify nearly a million personal stories.

Weblog Stories as Data

In a telephone survey of a nationally representative sample of bloggers conducted by the Pew Internet & American Life Project (Lenhart & Fox, 2006), American bloggers most frequently cited "my life and experiences" as a primary topic of their blog (37%). Nearly one million new blog posts are made each day on the web (Technorati.com, 2008), raising the possibilities for new quantitative approaches to the study and analysis of human life and experiences where weblog text is treated as data. In pursuing these new quantitative approaches, it is important to understand how the content that bloggers characterize as "my life and experiences" relates to the text that they actually compose. One particularly interesting presentation of life experiences in weblogs is in the genre of the personal story, consisting of the non-fiction narratives and anecdotes that people tell about their lives. In our research we define personal stories as textual discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the author or a close associate is among the participants. For example, the following text is an excerpt from a personal story that appears on an anonymous Internet weblog.

> "*I cracked the egg into the bowl and then I saw it, yes a baby chicken was in the egg that was going to be our breakfast. I felt like I might be sick, but the rest of my family found this to be very interesting.*"

One of millions of stories that appear in weblogs, this excerpt illustrates some of the interesting characteristics of anecdotal evidence as data in qualitative analyses. If the events in this narration are to be believed, it provides some evidence concerning the frequency of errors in our food-supply chain. More directly, this passage tells us about the sorts of events that people find surprising enough to write about (baby chickens in breakfast eggs). Even if this story is completely fictitious, its narration provides us with evidence about the psychology of its author, and how they perceive the people who constitute the imagined audience. Weblog stories are data points in a composite model of how the world is, how people perceive the world, and how people narrate these perceptions to others.

Although some researchers might prefer that the psychological and sociological components of this model were separable from the cold hard facts of the world, the production of this data is inextricably tied to human experience and the need to share it. Langellier & Peterson (2004) consider the characteristics of storytelling in weblogs and argue that this new technological medium retains the performance aspects that characterize storytelling in other social situations. Although the form of the genre most closely resembles the private journal or daily diary, the public nature and rapid feedback of weblog stories move this style of discourse closer to that of storytelling in oral conversation, where issues of identity, morality, and authenticity are brought to the foreground. In particular, they note the balance between the sincerity of the story and the sincerity of the storyteller, a combination they refer to as the *creative double bind*:

> [...] only by being insincere in making up a "good telling" can a sincere or "good story" be told. At the same time, a good or "tellable" story is one that is about *an* experience - something out of the ordinary - in which the sincerity of the storyteller is established through the ordinariness of the telling rather than the extraordinary events of the story (p. 179).

Rather than working to separate the world, its perception, and its narration in weblog stories, the most immediate applications of datasets of weblog stories will use this mix directly to achieve their aims. As of yet, this mix has only been exploited in the context of interactive arts and entertainment technologies. Owsley et al. (2006) describe a system called *Buzz*, a digital theater installation where animated virtual characters delivered emotionally-evocative stories extracted from weblogs. In this work, candidate weblog entries are retrieved from a commercial weblog search engine, using the day's most popular Internet searches and a list of controversial topics as weblog search queries. Swanson & Gordon (2008) describe a system called *Say Anything*, an interactive storytelling application where a user and the computer take turns writing sentences in an emerging fictional narrative. In this work, the computer adds sentences to the unfolding story by finding the most similar events in a large corpus of weblog stories, returning the next event in the retrieved story as the next event in the unfolding narrative. In both of these cases, the perceptions of the weblog authors and their sincerity as storytellers are at least as important as the sincerity of the narrated events to achieve the goals of the application.

Continued innovation will certainly produce other applications with similar dataset requirements. However, the full potential of weblog stories as a source of data about the world, human psychology, and social interaction will require a deeper understanding of the balance between these three factors in a blogger's work. At the very least, some quantitative information about the practice of weblog storytelling is sorely needed. What percentage of all weblog stories is completely fictitious? What is the likelihood that a blogger will tell a story about any one of their experiences? What percentage of bloggers who participate in shared experiences will narrate them on their blog? What types of events are most likely to be omitted in the narration of these experiences? More basically, what is the average length of a weblog story, when are they typically written, and on what weblog services do they most often appear?

Answering most of these questions will require future innovation in the research methods used by psychologists and sociologists to analyze weblogs stories. In this paper, we focus on one particular barrier to research innovation: the lack of comprehensive, large-scale corpora of weblog stories that can be shared and analyzed as a standard dataset for researchers in this area. Standard datasets have long proven their effectiveness in a wide range of research areas, e.g. the Wall Street Journal sections of the Penn Treebank for use in computational linguistics research (Marcus et al., 1993). To be effective in answering the primary research questions about weblog storytelling, a comprehensive corpus is among the greatest needs. Ideally, this corpus should contain *every* story posted by *every* weblog author over some known duration of time. To be truly useful, such a corpus would also need to have clear usage agreements in place that allow for different research groups to freely share data results without violating privacy or copyright standards.

This paper describes our efforts to create a comprehensive corpus of weblog stories with these ideals in mind. In our work, we capitalize on the availability of the ICWSM 2009 Spinn3r Dataset (ICWSM, 2009), a collection of millions of weblog entries written between August 1 and October 1, 2008. To identify weblog posts that can best be characterized as personal stories, we developed a new automated story classifier using supervised machine learning techniques. Applying this classifier to the full corpus, we identified nearly a million weblog entries that contain personal stories. These entries are shared publically as indexes to the original corpus, allowing any researcher with the appropriate access privileges to the ICWSM 2009 Spinn3r Dataset to reconstruct this subset with the application of a simple filtering script. Although the resulting corpus fails to identify *every* story from *every* blogger, and is currently limited to weblogs written in English that appear in the dataset, the identified corpus can be readily used to support innovation in psychology and sociology research methods that were previously impossible.

Previous Story Collection Research

The central technical challenge in automatically creating a large corpus of weblog stories is discriminating between story content and non-story content in weblog entries. Non-story weblog content appears in many forms, and commonly includes excerpts from news articles written by journalists, commentary on events in the news, lists of all sorts, cooking recipes, and analyses of sporting events. The actual percentage of content that can be judged as story-like varies according to the methods used to collect candidate weblog posts, but no previous efforts have seen percentages higher than 17%.

Much of the work on automated story collection from weblogs grew out of our previous work on extracting personal stories from audio-recordings of directed face-to-face interviews with experts in various domains of practice. In our first attempt in this area, we explored the use of machine-learning techniques for identifying stories in segments of conversational speech, using the words recognized with commercial speech-recognition software (Gordon & Ganesan, 2005). We followed a traditional text classification approach, where a corpus of transcribed conversational speech was first hand-annotated (story / non-story) for use as training and testing data. By developing a clear definition of what counted as a story, our annotators were able to achieve reasonably high inter-rater agreement

(K=0.68). Segments of training data were then encoded as high-dimensional feature vectors (word-level unigram and bigram frequency counts) and used to train a naïve Bayes binary classifier. To apply this classifier to test data, overlapping consecutive segments of test data were individually assigned to either the story or non-story class, with confidence values smoothed across segments using a simple mean-average function. Performance evaluations of our approach yielded low precision (39.6%) and low recall (25.3%), which was equal to random chance performance on this task. However, we observed substantially higher performance when using transcribed test data (as opposed to the output of a speech recognition system), with reasonable results (precision = 53.0%, recall = 62.9%, F-score = 0.575).

Given the low performance of story capture from speech data we decided to shift our focus to written electronic discourse, specifically weblogs (Gordon et al., 2007). To acquire a large database of Internet address 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 a vocabulary from an existing broad-coverage knowledge base of commonsense activities (Gordon, 2001). Search results were then processed to identify unique addresses, resulting in a set of over 390,000 weblogs. By randomly sampling weblog entries from this set, we found that 17% of the text in weblog entries consisted of stories. To create a story classification technology for Internet weblog entries, we created a new hand-annotated (story / non-story) corpus of the entries in 100 randomly sampled weblogs, and used these annotations to train and evaluate our classification approach on weblog text (precision = 30.2%, recall = 80.9%, F-score = 0.414). We then applied this classifier to the 3.4 million entries in our weblog set over the course of 393 days, producing a corpus of 4.5 million extracted story segments consisting of 1.06 billion words.

Subsequently, we explored whether the performance of our story extraction technology could be improved through the use of more sophisticated text classification techniques (Gordon et al., 2007). By incorporating techniques for automatically detecting sentence boundaries in the test data, utilizing a contemporary Support Vector Machine learning algorithm, and using a Gaussian function to smooth the confidence values, we were able to significantly improve the overall performance of this approach (precision = 46.4%, recall = 60.6%, F-score = 0.509).

Although this previous work has been successful in producing a story collection of considerable size, there are several factors that limit the utility of this collection for use in new research. This corpus does not have high enough levels of precision for some potential applications. The method of selecting candidate weblogs for analysis is not sufficiently random for many statistical analyses of weblog storytelling. In hindsight, our method of extracting only portions of weblogs entries as story content was unfortunate, as it produced a splintered collection of text segments that obscure the natural cohesiveness seen in whole weblog posts. Most problematic, the legalities of distributing this corpus to other researchers and collaborators were unclear, particularly given the number of copyright holders of the text therein.

The ICWSM 2009 Spinn3r Dataset

In this paper, we describe our efforts to overcome the limitations of our previous story collection research using new technologies and by capitalizing on the availability of a new weblog dataset. In 2009, the 3rd International AAAI Conference on Weblogs and Social Media sponsored the ICWSM 2009 Data Challenge to spur new research in the area of weblog analysis. A large dataset was released as part of this challenge, the ICWSM 2009 Spinn3r Dataset (ICWSM, 2009), consisting of tens of millions of weblog entries collected and processed by Spinn3r.com, a company that indexes, interprets, filters, and cleanses weblog entries for use in downstream applications. Available to all researchers who agree to a dataset license, this corpus consists of a comprehensive snapshot of weblog activity between August 1, 2008 and October 1, 2008. Although this dataset was described as containing 44 million weblog entries when it was originally released, the final release of this dataset actually consists of 62 million entries in Spinn3r.com's XML format.

In addition to the RSS and ATOM information that are published by the original weblog hosting sites, Spinn3r.com's format includes several meta-data tags that aid in processing a dataset of this size. Spinn3r.com assigns a tag indicating the language of the written entry, using a proprietary mathematical model. Roughly a third of the entries are assigned to one of thirteen "tiergroups" using a calculation of blog influence, computed in part based on a count of other sites that link to the blog.

There are also some problematic characteristics of this dataset. Since the items in the dataset originate from RSS and ATOM feeds, a portion of these items contain only the first few hundred characters rather than the entire text of the weblog entry. This collection also appears to contain a large portion of items that are not typically considered to be weblog posts, e.g. posts in threaded online discussion forums, descriptions of products from online retailers, and articles that appear in the online version of professional news organizations.

The substantial size of the ICWSM 2009 Spinn3r Dataset, coupled with a clear licensing agreement, make it particularly well suited as a source for personal stories. If they could be reliably identified in this dataset, then it would be possible to create a new corpus of nearly all of the personal stories posted in weblogs between August 1 and October 1, 2008.

Annotation of the Training Data

As in previous work on automated story extraction from weblogs (Gordon et al., 2007), we pursued a machine learning approach to text classification. This required the development of a sufficiently large set of annotated training examples, where individual blog entries are assigned "story" or "non-story" labels. Unlike previous work where segments of text within individual blog posts were labeled, we opted instead for a whole-entry labeling scheme. Blog entries would be labeled as "story" if an annotator judged that the content of a blog post primarily consisted of story content, even if some non-story text was included, and vice versa.

In our initial studies of the dataset, we found that blog posts consisting primarily of story content were much less frequent than in our previous work. As a practical matter, this meant that the size of the training corpus would need to be comparatively large in order to include enough positive examples of stories to train the classifier. Accordingly, our first challenge was to determine how these annotations could be acquired in a cost-effective manner.

We briefly explored the use of anonymous non-expert paid annotators, using Amazon's Mechanical Turk to solicit and distribute the annotation labor over the Internet. Although this service has recently been touted as an inexpensive and effective alternative to expert annotation for natural language processing tasks (Snow et al., 2008), this was not true in our experience. We found that the vast majority of annotations produced through this service were either completely random or were generated by automated response engines, yielding unusable results. The labor required to separate legitimate from illegitimate annotations (for the purpose of awarding micropayments) was far more expensive than the annotations themselves.

Our solution was to develop a simple annotation tool, designed with speed and accuracy in mind, and to annotate the training data ourselves. We found that the single most important factor in reducing annotation costs (time) was how the content of the entry was presented to the annotator. It is far easier to judge whether or not a weblog entry is primarily a story if it is viewed in its original form through a web browser, rather than in the XML format of the dataset. Accordingly, our simple annotation tool was built as a web application that displayed the webpage of a weblog entry alongside a simple interface for assigning "story" or "non-story" annotations.

The weblog entries in the dataset each provide the location of its corresponding webpage (identified by the <*link*> XML tag). However, a large percentage of the weblog entries in the dataset have links that are no longer active, i.e. the link is broken or it points to a webpage indicating that the post is no longer available. To identify a set of weblog entries for annotation, we built an automated process to identify all entries in the dataset that met the following criteria:

1. The <*dc:lang*> field was *"en"*, indicating that the weblog was written in English.

2. The <*description*> field contained at least 250 characters after un-escaping XML and HTML characters and removing HTML tags, indicating that the content of the post was sufficiently long and had not been truncated in its original RSS feed.

3. The <*link*> field contains the URL of a webpage that actually exists.

4. Nearly all (90%) of the words in the <*description*> field appear in the HTML source of the webpage in the <*link*> field, indicating that the content of the weblog entry is still available.

We applied these criteria to the 196,503 entries in the "*Aug/01*" folder of the "*tiergroup-1*" section of the dataset using our automated process, producing a filtered set (35,275 entries) that could be annotated using our simple annotation tool. We chose to limit our annotation to this one section of the dataset in order to avoid the complexities involved in randomly sampling the entire corpus. However, this decision would later require us to conduct a separate evaluation of classification performance when applied to entries outside of our set of annotations.

Using this tool, the first author of this paper annotated 5002 weblog entries, assigning the label "story" or "non-story" to each. The label "story" was assigned to 240 of these entries (4.8%). No attempt was made to compute an inter-rater agreement score with other annotators. Although modest compared to the size of annotated training datasets used in other natural language processing tasks, these 5002 annotations are substantially greater than those used in previous story classification research, e.g. Gordon et al. (2007), where the entries in 100 random weblogs were annotated for use as training data.

Development of the Story Classifier

As in our previous work, we treated story identification as a simple binary classification task to be executed by a trained model. We chose to use a confidence-weighted linear classifier (Dredze et al., 2008) as our machine learning algorithm. This linear classifier is similar to the Perceptron (Rosenblatt, 1958), but adds additional information to each feature weight to estimate the confidence of its current assignment. This value allows the weights to be adjusted more accurately with each new training instance and helps to avoid over-fitting. Because this classifier is still a simple linear classifier, it is extremely efficient in both training and application. Training is several orders of magnitude faster than a linear kernel Support Vector Machine and only slightly slower than the standard Perceptron algorithm. Despite the large reduction in training time, the confidence-weighted linear classifier is competitive with other contemporary machine learning algorithms, and often outperforms state-of-the-art classifiers such as Support Vector Machines and Maximum Entropy. For our experiments we used a Java port of the Google Code project, Online Learning Library (Okanohara, 2008).

Selecting appropriate features for any classifier is challenging and often requires rich knowledge of the do-

main and a reliable methodology for extracting features from raw data. Selecting effective features for story classification is no different. The defining characteristics of personal stories are particularly abstract and difficult to automatically extract from natural language text. In our research we define personal stories as textual discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the author or a close associate is among the participants. Unfortunately, these features are difficult to extract directly using currently available tools for natural language processing, particularly given the scarcity of annotated training data in the genre of weblog text. However, our previous work has shown that simple lexical features can often provide surprisingly high levels of performance on the story classification task (Gordon et al., 2007). Given our definition of a personal story, one would expect to more frequently see first person pronouns (i.e. *I*, *me*, *my*) along with a greater proportion of past tense verbs. Additionally, some simple connective phrases for events may appear more frequently in personal stories (i.e. *and then*, *finally*, or *suddenly*). Accordingly, we focused our efforts on learning the lexical features that are most predictive of story content.

We attempted to classify individual weblog entries using the text that appears in the <*description*> field of the XML record. To accomplish this, we investigated several variations of n-gram features (e.g. unigrams and bigrams) that have worked relatively well for us in the past. However, there were a few differences in this dataset that led us to slightly alter the design of our features. In particular, since we were looking at entire entries rather than spans of sentences, it was important to encode the frequency of terms in the entry and not simply that one had appeared. There are at least two obvious ways to encode this frequency information. The first is to create a new binary feature that simply appends the frequency to the existing token. For example, if the token *me* has been seen 3 times, it would be transformed into the new token *me-3*. Alternatively, the feature token could remain unchanged and the feature value could be weighted to reflect the frequency. Although we tried both, the best results were obtained using the latter approach. The feature values were normalized to be between 0 and 1 and a maximum frequency value of 7 was imposed to simplify the normalization process.

Like many other classifiers, the confidence-weighted linear classier has a regularization and a bias parameter that affect the learning process. Typically, the optimal values for these parameters are not known in advance and must be learned from data. We created a development set of 750 weblog entries that we used for finding appropriate values for these two parameters. A simple grid search over a range of values was performed and the best pair of values was used for the experiments. We then performed a 10-fold cross validation on the remaining 4252 entries for each variation of our n-gram feature sets. Our best results were obtained using simple unigram features (precision = 66%, recall = 48%, F-score = 0.55).

In previous work we have found that combining both lexical and part-of-speech information yields the best performing classifiers for this task. Similarly, in this work we also found that combining these types of features produced a classifier with the highest F-score. However, the improvement was only slightly higher than simple lexical features alone and was not statistically significant. Considering the large volume of data in the ICWSM 2009 Spinn3r Dataset, which represents only a small fraction of weblog entries on the Internet, we judged that the overhead penalty of the additional processing was not worth the minor improvement in accuracy.

The ICWSM 2009 Spinn3r Dataset includes other information in the XML records that could potentially be used to improve classification without significant amounts of additional processing. We ran several experiments that incorporated some of these additional features, such as the text in the <*title*> and <*category*> fields. Unfortunately, these features did not contribute significant improvements over the best performing system.

Unlike previous work, unigram features resulted in a higher performing classifier than either bigrams or trigram features. The relatively poor performance of bigram and trigram features is likely due to over-fitting caused by disproportionately high weights given to rare n-grams in the training data. Despite the simplicity of unigram features and their own potential for over-fitting, our new classifier outperforms the best F-score result of Gordon et al. (2007) by 4 percent.

Our current approach identifies stories at the weblog entry level, unlike previous systems that operated at the segment or sentence level. We believe that this approach produces a more useful story corpus, free of the discontinuities and noise introduced by the misclassification of story fragments. However, the entries in the ICWSM 2009 Spinn3r Blog Dataset introduce a new complication. The weblog entries in this corpus were aggregated from a variety of services that do not all share a common policy for the information provided in their XML feed. In particular, the <*description*> field, where the text of the weblog entry is provided, does not always include the entire post. For many items in the corpus, only the first 250 characters are given. To estimate the impact of this problem on our classifier, we ran an experiment that applied our best-trained classifier to test data that was artificially limited to 250 characters. The performance on this data set was dramatically reduced (precision = 47%, recall = 3%, F-score = 0.06). Although this performance is terrible, the precision is high enough to ensure that relatively few story fragments will be introduced to the resulting corpus.

One attractive feature of confidence weighted linear classifiers is that the learned models are easily interpretable, as the learned weights of features can be sorted to see which are most predictive of category labels. Table 1 shows a list of the top features indicative of stories and of non-stories using our best performing unigram feature set. It is interesting that many of the highest weighted features support the intuition that past tense verbs and personal

pronouns would be highly predictive. However, the top verbs that it learns are not specific to particular genres of stories, such as fishing or traveling, but are rather generic, such as going and doing. Similarly many intuitive features were also learned corresponding to non-story text, such as future and present tense verbs.

Story Features	*Non-Story Features*
went	will
send	1
took	/
back	years
i	blog
had	has
evening	team
down	many
comments	can
friend	are
was	love
art	being
got	use
did	before
headed	:

Table 1. The top 15 features for each class

Application to the Dataset

We applied our story classifier to each of the English-language weblog posts in the ICWSM 2009 Spinn3r Dataset, identified by the <*dc:lang*> field of each entry. Entries with fewer than 250 characters in their <*description*> field were included, but flagged in consideration of the lower accuracy of our classifier on entries. A total of 960,098 entries were labeled "story" by our classifier, where 937,994 of these assignments were made on entries with <*description*> fields containing more than 250 characters.

We conducted a second evaluation to determine the overall precision of our classifier when applied to the entire dataset. Since our classifier was trained using 5002 blog entries from the Aug/01 section of tiergroup-1, there was some concern that its performance would decrease when applied to entries in other sections. We randomly sampled 300 of the weblog entries that our classifier identified as stories. Each of these entries was then hand-annotated as "story" or "non-story" by the same annotator as in the training set. As before, the annotator made assignments by following the <*link*> field in the entry, reading the original post, and making a judgment. Of these 300 entries, 47 had <*link*> fields that no longer addressed the original weblog post on the Internet. Of the remaining 253 entries, 191 were assigned the label "story" and 62 were assigned the label "non-story" (precision = 75%). This precision score is slightly higher than in our original cross-validation evaluations (precision = 66%), which may be partly due to the inclusion of the training data in all cross-validation folds for the final version of our classifier. However, the difference in scores is within the standard deviation (13%) of these cross validation results. We conclude that there is no evidence that the precision of our classifier is degraded when applied to data from different sections.

During this second evaluation, we made special note of the characteristics of the 62 non-story entries that were mislabeled by our classifier. In nearly all cases these entries appeared in personal weblogs, as opposed to the commercial websites and web forums that are frequent in the whole dataset. The most frequent mislabeling occurred with posts that consisted of fictional creative writing, such as chapters of a novel being written by the author of the weblog. Other mislabeled posts included plans for the coming day, one-line summaries of each day in the previous week, jokes, generalized summaries of recurring events, and the fictional weblogs written by players of online fantasy role-playing games. Although we do not consider these posts to be personal stories of people's real-life experiences, it is clear that distinguishing them from stories using our current approach would be difficult given the similarity in vocabulary.

We conducted a number of analyses on the set of entries labeled as "story" by our classifier, focusing specifically on the 937,994 entries with <*description*> fields containing more than 250 characters. First, we examined the relationship between Spinn3r.com's tiergroups and the frequency of stories. Figure 1 plots the percentage of items labeled as "story" for each of the 14 tiergroups. These results indicate that the ratio of stories to weblog posts varied widely across tiergroups, with the highest rate seen in tiergroup 1.

Second, we analyzed the story set to determine the weblog hosting services that were most frequent, as indicated by the root of the domain name in the <*link*> fields. Table 2 lists the number of stories that were posted on the top five most frequent weblog hosting services. These results indicate that the vast majority of stories were written using a handful of weblog hosting services, with over half appearing on a single service (livejournal.com).

Third, we conducted a simple analysis to determine the day of the week that weblog stories are most frequently written. Figure 2 plots the average number of stories written per day of the week as indicated by the <*pubDate*> field in the entry, normalized by the number of times that day is included in the corpus. These results indicate that stories are most frequently written on Tuesday, and least frequently written on Thursday.

In order to support further analyses and applications of this story corpus, we generated a data file that identifies the location of stories in the ICWSM 2009 Spinn3r Dataset. Each of the 960,098 entries labeled "story" by our classifier is listed, represented by the source file name, the start and end line where the XML entry can be found, the raw confidence value of the story classification, and an indication of whether the <*description*> field contained less than 250 characters. Researchers who have acquired the full

dataset can use this data file and a simple filtering script to generate a comprehensive corpus of stories for use in their own research. This data file is available from the authors of this paper via email request.

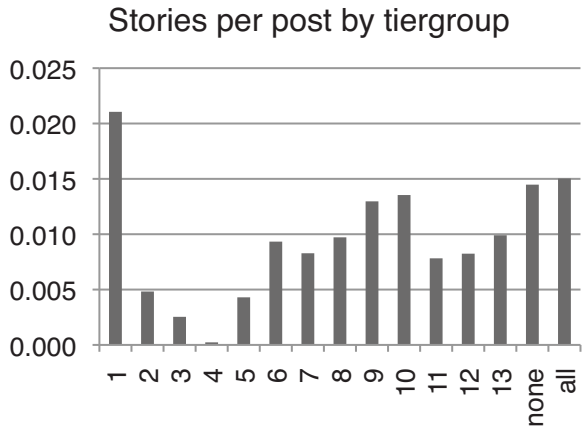

Figure 1. The ratio of stories written in English to all posts for each of the 14 Spinn3r.com tiergroups

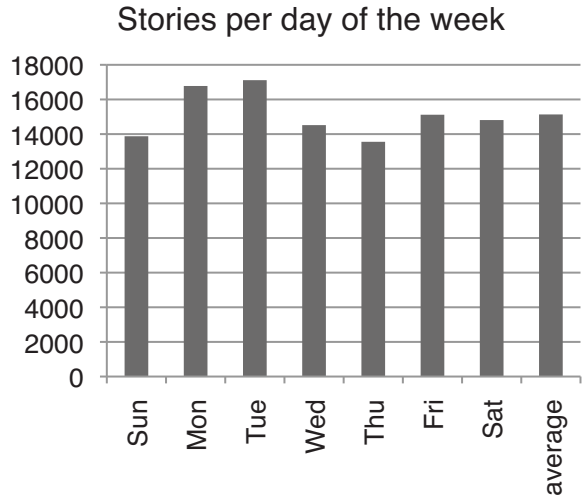

Figure 2. The average number of stories written in English on individual days of the week

source	*stories*	*percent*
livejournal.com	543682	57.9%
wordpress.com	124468	13.3%
blogspot.com	57226	06.1%
typepad.com	36300	03.9%
vox.com	23744	02.5%
top 5	785420	83.7%

Table 2. The top 5 sources of stories written in English

Discussion

Despite the importance of storytelling in human interaction, it has been difficult for psychology and sociology researchers to empirically study this genre of communication on a large scale. With the rise of Internet weblogs, a new channel for everyday storytelling has emerged that is more amenable to quantitative analysis. In this paper, we describe our efforts to further enable this research by creating a large-scale comprehensive corpus of personal stories found in English-language weblogs, a corpus of stories that is among the largest ever produced. This current effort improves on our previous work (Gordon et al., 2007) in a number of ways that enhance the utility of this story corpus as a research resource. First, by classifying personal stories at the level of the blog post rather than at the segment or sentence level, the stories in this corpus can be analyzed in their entirety along with the details of their provenance. Second, we have improved the precision of automated story classification by employing a confidence-weighted linear classifier and by training this classifier on larger hand-annotated training sets. Third, by using the ICWSM 2009 Spinn3r Dataset as a source of weblogs posts, we have established a means for researchers to freely obtain a large story corpus with a clearly defined usage agreement.

There are at least two directions for future work that address limitations of this corpus. First, this corpus includes only weblog posts written in English. In order to contrast English weblog storytelling with that of other frequent languages, e.g. Japanese, it will be necessary to develop new classifiers. We believe that a confidence-weighted linear classifier will work equally well in other languages given the right training data, but we expect that the linguistic features that best indicate story content will vary significantly across languages.

Second, the size of this story corpus may be too small for certain types of analyses or applications. For example, studies of long-term storytelling behavior of individuals may require corpora that span durations longer than two months. Increasing the size of the story corpus will require access to new sources of weblog data, e.g. the data that Spinn3r.com provides to its customers as part of their business practices. Using a confidence-weighted linear classifier to identify stories provides the efficiency needed to identify stories in all weblog posts in real time. However, the use of new sources would require additional consideration of usage and licensing agreements.

Despite these two limitations, the story corpus that we have identified in the ICWSM 2009 Spinn3r Dataset should be a useful resource for a wide range of future research efforts in psychology and sociology. In this paper we have only hinted at the sorts of quantitative analyses of storytelling that are enabled by a comprehensive corpus of this size. While it may seem trivial to discover that more English-language stories are written in weblogs on Tuesdays than on Thursdays, it is significant that the scale of this corpus allows us to definitively conclude this fact. Further analyses of this corpus will allow researchers to study the genre of weblog stories, and answer some of the most

fundamental questions in storytelling behavior in quantitative terms.

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

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.

ICWSM (2009) ICWSM 2009 Spinn3r Dataset. Proceedings of the Third International Conference on Weblogs and Social Media, San Jose, CA, May 2009.

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.

Gordon, A. & Ganesan, K. (2005) Automated Story Extraction From Conversational Speech. Third International Conference on Knowledge Capture (K-CAP 05), October 2-5, Banff, Canada.

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

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

Lenhart, A. & Fox, S. (2006) Bloggers: A Portrait of the Internet's New Storytellers. Pew Interent & American Life Project. Available at http://www.pewinterent.org

Marcus, M., Marcinkiewicz, M., and Santorini, B. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics 19(2):313-330.

Okanohara, D. (2008) Online Learning Library. Available at http://code.google.com/p/oll/

Owsley, S., Hammond, K., Shamma, D., Sood S. (2006) Buzz: Telling Compelling Stories. ACM Multimedia, Interactive Arts program, Santa Barbara, CA.

Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386-408.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. (2008) Cheap and Fast - But is it Good? Evaluating non-expert annotations for natural lanugage tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08), Honolulu, HI.

Swanson, R. & 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/