Identifying Personal Narratives in Chinese Weblog Posts

Andrew Gordon, Luwen Huangfu, Kenji Sagae, Wenji Mao, and Wen Chen

Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
Institute of Automation, Chinese Academy of Science, Beijing, China
gordon@ict.usc.edu, luwen.huangfu@ia.ac.cn, sagae@ict.usc.edu, wenji.mao@ia.ac.cn, chenwen@usc.edu

Abstract

Automated text classification technologies have enabled researchers to amass enormous collections of personal narratives posted to English-language weblogs. In this paper, we explore analogous approaches to identify personal narratives in Chinese weblog posts as a precursor to the future empirical studies of cross-cultural differences in narrative structure. We describe the collection of over half a million posts from a popular Chinese weblog hosting service, and the manual annotation of story and nonstory content in sampled posts. Using supervised machine learning methods, we developed an automated text classifier for personal narratives in Chinese posts, achieving classification accuracy comparable to previous work in English. Using this classifier, we automatically identify over sixty-four thousand personal narratives for use in future cross-cultural analyses and Chinese-language applications of narrative corpora.

Introduction

In his book *Human Universals*, anthropologist Donald Brown listed narrative among 63 features of culture, society, language, behavior, and psyche for which there are no known exceptions (Brown, 1991). People in every culture incorporate narrative as part of their everyday discourse, from the casual recounting of the day's events among family members to the more formal presentations of testimony and life experiences across different social contexts. We are all experts at telling narratives, and therefore may neglect to ask the important cross-cultural question: Are there differences in the way that various cultural groups employ narrative as a means of communication?

At a superficial level, the answer is most certainly yes. People tell narratives in their local languages, and the events described in these narratives reflect cultural differences in daily life. More interestingly, we can ask if narrative exhibits *structural* differences across cultures, indicating that the composition of everyday narrative is subject to the influence of cultural norms and conventions. These differences may appear as tendencies toward certain rhetorical structures (Mann & Thompson, 1988), toward the inclusion or exclusion of expressions of private mental states (Wiebe et al., 2004), or toward statements at particular narrative levels (Genette, 1980), among others.

To empirically identify structural differences in narrative across cultures, we are immediately faced with the problem of data acquisition. Statistically significant differences can only be identified if we begin our analyses with very large samples of naturally occurring narratives gathered from two or more cultures, using equivalent collection methodologies. Previously, various social and cultural efforts have been successful in harvesting large narrative corpora at national levels, e.g. the Federal Writer's Project of the Works Progress Administration (Mangione, 1972) and the StoryCorps Project (Isay, 2007). Although projects like these could be replicated across different cultures, the interview methods used in these efforts may systematically influence how people tell narratives, such that they are not representative of narratives used in everyday discourse.

The phenomenal rise of weblogs and social media creates new opportunities to study everyday narrative across cultures. While blogging is popularly associated with high-profile celebrities and political commentators, the typical weblog takes the form of a personal journal, read by a small number of friends and family (Munson & Resnick, 2011). As a personal journal, bloggers include personal narratives of their life experiences among other posts in their weblogs, although the fraction of narratives among all posts is small. Swanson (2011) estimated that only 5.4% of all non-spam English-language weblog posts are personal stories, defined as non-fictional narrative discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the storyteller or a close associate is among the participants. Using supervised machine learning

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

techniques, Swanson constructed a text classifier to identify these personal stories in streams of weblog posts. Applying the classifier to the 25 million English-language weblog posts in the ICWSM 2009 Spinn3r Dataset (Burton et al., 2009), Swanson identified nearly one million personal stories.

We hypothesize that narratives appearing in personal weblogs would exhibit structural differences endemic to particular cultures, if indeed these differences exist. In this paper, we focus our attention on the technical challenge prerequisite to cross-cultural comparisons, namely the acquisition of a suitably large corpus of weblog narratives from distinct cultures. Here, we describe our efforts to extract a large corpus of personal narratives from Chinese-language weblog posts on a popular commercial weblog-hosting service in China. Applying supervised machine learning approaches, we trained, tested, and applied a narrative classifier to hundreds of thousands of Chinese weblog posts to identify tens of thousands of personal narratives. As a precursor to cross-cultural studies of narrative structure, we conclude this paper with a comparison of lexical features that best distinguish between narrative and non-narrative weblog posts in English and Chinese.

Chinese Weblog Posts

The People's Republic of China has the largest number of Internet users, crossing the half-billion mark at the end of November 2011 (Lee, 2012). Online media conglomerates such as Sina, Tencent, and Baidu compete for users from Chinese communities worldwide, offering a range of online services including news, email, search, games, blogging, and microblogging.

As a source of weblog posts, we selected the weblog hosting service provided by Sina (blog.sina.com.cn). Topics discussed by users on this hosting service span the full breadth of Chinese society, with increased attention on Chinese popular culture and the entertainment industry. Most posts appearing on Sina's weblog hosting service are written in Standard Chinese (Mandarin) using the simplified Chinese character set.

We chose Sina Blogs primarily because they provide a public directory listing to the blogs of over ten thousand of their users. By crawling this directory, we gathered the URLs for each user's weblog. Unfortunately, the XML RSS feeds provided by Sina for each of these weblogs contained only a subset (the latest five) of users' posts. Accordingly, we downloaded the textual content of all publically accessible posts of each user as HTML files, and extracted the user-authored content of each page automatically. Using this approach, we collected 572,208 posts from 10,398 weblogs.

Annotation of Weblog Posts

Despite the journal-like quality of weblogs, the majority of weblog posts are not narratives. By analyzing a sample of English-language weblog posts, Swanson (2011) estimated that only 5.4% were personal narratives of authors' life experiences. Instead, bloggers posted a variety of other content to their weblogs, including extracts from news articles, personal or popular photographs, comments on current events, links to other websites, culinary recipes, schedules, and "to do" lists. Although non-textual content can be removed heuristically, the identification of narrative text among other textual content requires human judgment. As in Swanson's previous work, we modeled this judgment using supervised machine learning methods and large amounts of hand-annotated training data.

To collect training data, we developed a simple web-based annotation interface that presented to the annotator the textual content of a selected weblog post, along with a hyperlink to the post as it appears on the web. For each post, annotators were instructed to read each post, and then select among three labels the one that best characterized the content of the post. These labels were defined to annotators in the following way:

1. *STORY*: "Non-fiction narrative discourse describing a specific series of events in the past, spanning minutes, hours, or days, where the storyteller or close associate is a participant. Examples: A blogger writes about yesterday's doctor's appointment, dinner at a restaurant, a conversation they had with a friend, a trip they took on a train, an accident they had while riding their bicycle, or something funny that their child did in the morning."

2. *NONSTORY*: "Posts that are not primarily personal stories. Examples: Posts with only a few words on them, posts that only include pictures and no text, newspaper articles or excerpts from newspaper articles, links to videos or other web pages, sports scores, cooking recipes, and any written text that is primarily opinion, or primarily discusses things that are going to happen in the future, rather than the past."

3. *SKIP*: "Posts that you do not want to annotate for any reason. There is a lot of objectionable content on the web, and I would prefer that you "skip" posts that you would rather not read or view. Also "skip" any posts that are not being displayed correctly by the annotation interface. All posts that are "skipped" will not be used to train or test our automated story classification technology - they will be ignored. You should not use this label when the post is difficult to judge, when you are uncertain whether or not it is a story. However, there are some cases that will arise where the judgment is clear, and you still don't want the post to be included in the training data."

We enlisted the aid of six annotators, each a native of China and a graduate of a Chinese college or university.

These annotators used our web-based interface intermittently over a two-week period to make 13,288 judgments, 12,901 of which were *story* or *nonstory* labels for 6,946 unique weblog posts. To calculate inter-rater agreement, multiple annotations were collected for 5,708 of these posts. Table 1 compares the annotation results for our work and the English-language results of Swanson (2011).

	Chinese	*English*
Posts annotated	6,946	4,985
Narratives	11.9%	5.4%
Cohen's Kappa	0.46	0.68

Table 1. Chinese annotation results compared to previous work by Swanson (2011).

Compared to Swanson's English-language results, we observed a much higher percentage of personal narratives posted to Sina Blogs (11.9%) than to the aggregated English-language weblog stream provided to Swanson by Spinn3r.com (5.4%). We expect this reflects the intentions of the self-selected users of this particular weblog hosting service, rather than a cross-cultural difference in the propensity for narrative.

Mean pairwise inter-rater agreement (Cohen's Kappa) for our six Chinese annotators was substantially lower (0.46) than reported by Swanson for English weblog posts (0.68). Unlike the annotators in Swanson's work, the six individuals we enlisted in our research were not researchers in the area of narrative analysis, nor had any special training in linguistic analysis. The lower inter-rater agreement suggests that the specification of annotation task could be improved, with additional guidance provided for common borderline cases.

For this purpose, we conducted a qualitative analysis of posts assigned conflicting annotations. In a small percentage of cases, disagreements appeared to be simply due to annotator error. However, the majority of cases were due to genuine disagreements about whether the posts should be viewed as narratives, given that they exhibited unusual characteristics. Among the large variety of unusual posts were those that consisted almost entirely of personal photographs of some trip or event, where a narrative could be constructed from the textual captions that appeared below each one. Other posts were sampled from the weblogs of professional and semi-professional entertainers, and contained narrative-like press releases and promotional pieces. Other posts were extremely long and unfocused, containing a large number of disconnected narratives mixed with commentary. Other posts contained narrative content that spanned the breadth of the author's life, mixing reflections and feelings with aspirations and plans for the future. Collectively, these posts reflect the great diversity in ways that narrative is used in communication. Future efforts to improve inter-rater agreement could provide specific guidance to annotators on how to label instances from a broad catalog of post types, although theoretical or practical principles are needed to determine the appropriate label.

Preprocessing Chinese Weblog Posts

In machine learning approaches to automated text classification, it is necessary to preprocess each document in order to identify the distinct lexical features that will be used to make label predictions. Electronic text documents are commonly represented as ordered sequences of sentences comprised of ordered sequences of tokens (words and punctuation marks). Automatically encoding documents into this representation has been a major focus of natural language processing research over the last several decades, but much of this research has been aimed at handling the complexities of the English language. The unique characteristics of Chinese text both simplify and complicate different aspects of the preprocessing pipeline, compared to English text. Here, we focus on issues related to sentence delimiting, word segmentation and part-of-speech tagging.

First, delimiting sentences in Chinese can be performed similarly to English, with some slight modifications. The Chinese language uses a different set of punctuation marks than Western languages in order to maintain a consistent character width. Chinese punctuation marks are orthographically similar to those used in Western languages, but are assigned different Unicode code points than their Western counterparts. In Chinese, the punctuation marks of period (。), exclamation (！) and question (？) are obvious sentence separators, although the corresponding Western punctuation characters are also occasionally used in online text. Chinese has fewer exceptions and special cases than English, e.g. the non-delimiting periods used in "e.g." and "Mr."

Second, segmenting sentences into sequences of words is much different in Chinese. Whereas it is easy for Western languages to tokenize sentences by dividing contractions and breaking on spaces and punctuation marks, word segmentation is a significant challenge in Chinese text processing (Jin, 2008). Chinese words consist of one or more characters, and are written consecutively without word boundaries. To complicate matters, the correct word segmentation for a sequence of characters is context dependent. In some cases, the dependency is on the lexical context. In the sentence "我／对／他／有／意见" (*I have a suggestion to him*), the two characters "意见" should be combined together. However, in the sentence "我／有意／见／他" (*I intend to meet him*), the two character "意见" should be separated. In other cases the dependency is on the semantic context. For example, “门把手弄坏了。” can be segmented into “门/ 把/ 手/ 弄/ 坏/ 了/ 。” (*The door breaks (my) hand*) or into “门把手/ 弄/ 坏/ 了/ 。” (*The doorknob is broken*). The complexities of the Chinese word segmentation problem have led to a variety of technical solutions, including purely dictionary-based approaches like maximum matching (Liu et al., 1994), purely statistical approaches (Sproat & Shih, 1990), statistical dictionary-based approaches (Sproat et al., 1996) and transformation-based error-driven algorithms (Brill, 1993).

Third, the assignment of grammatical part-of-speech tags to Chinese words differs from English in the role of morphological features. English part-of-speech categories are strongly indicated by morphological features and spelling cues such as capitalization (Toutanova et al., 2003) and derivational and inflectional affixes (Brants, 2000). An English word ending in the suffix “-able” is likely to be an adjective, an English word ending in the suffix “-ly” is likely to be an adverb, and an English word ending in the suffix “-er” is likely to be a noun. Likewise, verb tenses are largely distinguished by regular inflections. In contrast, Chinese affixes are ambiguous and not strongly correlated with grammatical part-of-speech labels (Tseng, 2005). Instead, successful approaches to automated Chinese part-of-speech tagging have not relied on morphological features. These approaches include the application of rule-based systems (Greene & Rubin, 1971), stochastic systems that model contextual dependencies (Church, 1988), transformation-based learning approaches (Brill, 1995), and artificial neural networks (Nakamura et al., 1990).

Recent interest in automated Chinese text processing has led to the development of several end-to-end preprocessing pipelines that implement successful approaches. In our research, we utilized the Language Technology Platform (LTP) developed by Harbin Institute of Technology, which implements lexical analysis, syntactic parsing, and semantic parsing (Che et al., 2010). In this platform, a conditional random field model (Lafferty et al., 2001) is used to segment Chinese words, a discriminative classifier is adopted for Chinese part-of-speech tagging, and a maximum entropy model (Berger et al., 1996) is utilized for named-entity recognition.

Using LTP, we identified sentence boundaries, word segments, and part-of-speech tags for each document in our Sina Blogs corpus. In order to run LTP smoothly, we removed HTML tags and un-escaped HTML encoded characters. We removed a small number of text fragments written in languages other than Chinese, e.g. English and Russian. We replaced consecutive occurrences of identical punctuation marks with a single instance. Even with these accommodations, LTP would periodically fail to process certain posts, particularly those that were excessively long. We found that processing large documents in blocks of 90,000 characters was most successful, but doing so required us to take special care not to break a sentence across two blocks. In all, we successfully processed 478,550 posts from our original collection, of which 5,868 had been labeled as “story” or “nonstory” during our annotation exercise.

Automated Classification

To leverage our manual annotation of 13,288 story vs. non-story judgments toward mining our corpus of half a million Chinese weblog posts for personal narratives, we used the annotated portion of our corpus as training data for a text classifier following a straightforward machine learning framework. Because judgments provided by multiple annotators regarding individual posts were not always in agreement, and because story posts are relatively rare and might easily be missed, we considered any post that was labeled by at least one annotator as “story” to be a personal narrative, even if another marked it as a “nonstory” post. With this strategy, the percentage of posts labeled in our annotated corpus as stories is 17.2%, which is considerably higher than our estimate of 11.9% for how often any single annotator judges a post to be a story. It is possible that our lenient criterion for inclusion of posts in the set of stories results in an increased level of noise in our labels. However, given the rate at which nonstory posts outnumber story posts even with our lenient selection criterion, we decided that the cost of missed story posts is substantially higher than that of a small number of false alarms.

With story/nonstory labels fixed for each post in our annotated corpus, we proceeded with supervised training and evaluation of a classifier for personal narratives in Chinese weblog posts. To estimate the quality of our classifier, we set aside 25% of the posts in our manually annotated corpus as a test set (1467 posts), and used the remaining 75% as a training set (4401 posts). Following Swanson (2011), we used a variant of the perceptron algorithm for our binary classification task. The specific approach we adopted, the averaged perceptron (Freund & Schapire, 1999), is a large margin classification approach that has been shown to be efficient, straightforward to implement, and very effective in a number of language processing tasks. Like the standard perceptron, the averaged perceptron is an online learning algorithm that updates model parameters based on the model's errors on the training data. The key difference is that final parameter values are averaged over training iterations, which reduces overfitting, especially with data that are not linearly separable. The algorithm works as follows. We start with a vector $\boldsymbol{w_0}$ of feature weights, and a function **F** that maps an input instance x into a vector of feature counts. Given a set of training data (x_i, y_i), where each x is an input (in our case, a weblog post) each y is a label (in our case, +1 for story or -1 for non-story) corresponding to one of our training examples, we loop over each example, each time trying to classify it with the current model parameters: the predicted label y_i' is +1 if $\mathbf{F}(x_i) \cdot \boldsymbol{w}_t > 0$, and y_i' is -1 otherwise. If the predicted y_i' matches y_i, we simply move on to the next training example. Otherwise, we update our model parameters: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_i\boldsymbol{x}_i$. We go over all training examples T times. At the end of each of the T iterations, we save a snapshot of the weight vector $\boldsymbol{w}$. The final set of feature weights $\boldsymbol{w}_{avg}$ is an average of all T snapshots. The value of T and the features included in $\mathbf{F}$ were determined empirically using 10-fold cross-validation on the training set. Following Swanson, we experimented with unigram features (bag-of-words), where each word type is a feature, bigram features (two adjacent words), and trigram features (a contiguous sequence of three words). However, we did not experiment with any of Swanson's advanced features (e.g. syntactic information, entity grids, and features derived from the HTML in which the text is contained).

Although machine learning approaches are often applied to text classification in the context of topic categorization (e.g. determining if a newspaper article is about politics or sports), Swanson's work on classification of stories showed that some of the same techniques can be effective in distinguishing specific language genres, rather than topics. The success of automatic classification approaches is sometimes evaluated using a simple accuracy metric, which is calculated by dividing the number of correct decisions by the total number of decisions made by the classifier. Success in our task, where we aim to identify examples of a relatively infrequent class, is well measured in terms of *precision* and *recall* of personal stories. Precision of our story classifier is defined as the number of correctly identified stories divided by the total number of posts our classifier identifies as stories (how frequently the classifier is correct when it thinks it has found a story). Recall is defined as the number of posts identified correctly as stories divided by the total number of actual stories (what fraction of the stories in the dataset our classifier is successful in identifying). The harmonic mean of precision and recall is frequently reported in precision and recall evaluations, and it is referred to as F-score.

After experimenting with unigram, bigram and trigram features, and different numbers of training iterations, our best F-score in 10-fold cross-validation of the training set was obtained with unigram and bigram features. In contrast to Swanson's classification of English stories, the use of trigrams did not improve our results. Values of T between 10 and 20 made little difference in classification accuracy in our cross-validation. We trained our final model on all of the training data using unigram and bigram features, and 20 training iterations. Our best precision, recall, and F-score for the 10-fold cross-validation on the training set and our test set precision, recall, and F-score for the final model trained on the entire training set are shown in Table 2. It is interesting that our final precision, recall, and F-score is at a similar level achieved by Swanson (2011) for classification of personal narratives in English-language weblog posts.

	Precision	*Recall*	*F-score*
Cross-validation	0.625	0.464	0.533
Test set	0.596	0.464	0.521
Swanson (2011)	0.591	0.414	0.487

Table 2. Summary of results obtained with automatic story vs. non-story classification of Chinese weblog posts, compared with Swanson's (2011) best English-language performance.

Comparison to English Classification

Given the similarity between the performance of our classifier and that achieved by Swanson (2011), we were curious to know whether similar lexical features were predictive of class labels across the two languages. To make this comparison, we obtained Swanson's original annotated data and used it to train a new English-language classifier with the same averaged perceptron framework used for our Chinese classifier. The performance of this new English-language classifier was comparable to that reported by Swanson. We then examined feature weights learned by our models in each of the two languages, focusing on the unigrams (words) and bigrams (two-word sequences) that are most indicative of personal narratives in English and Chinese.. In our classification framework, these features are those with the highest positive weight for the "story" class label. Both in English and in Chinese, features reflecting past tense and first person pronouns are highly indicative of stories. While features related to past tense appear prominently in the top story features for both languages, the features reflect grammatical differences in the two languages. In English, this is expressed most clearly in verbs such as *went*, *had* and *was* (each of these was among the top ten story features for English), while in Chinese this is expressed through the temporal expressions *yesterday* and *that time* were (each in the top ten story features for Chinese).

Figures 1 and 2 depict the top 150 story features in English and (translated) Chinese, respectively. In each figure, the size of the font is proportional to the unigram or bigram's weight in the story classification model.

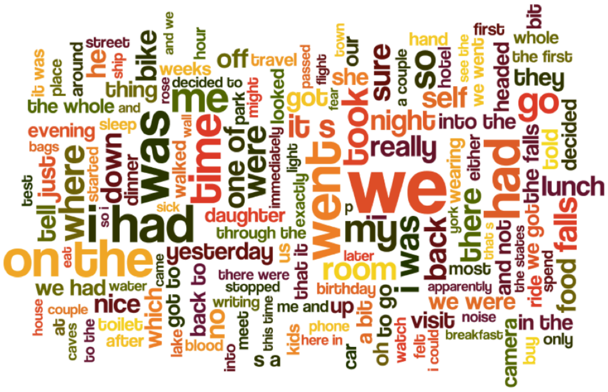

Figure 1. Word cloud based on the 150 highest-weighted features in the English story model

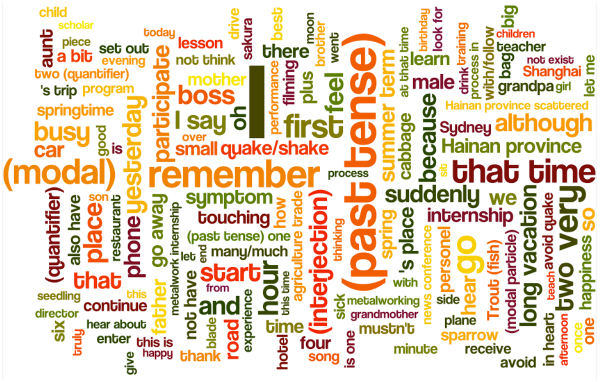

Figure 2. Word cloud based on English translations of the 150 highest-weighted features in the Chinese story model

Corpus Creation

As a final step in this research, we applied our classifier to the full set of 478,550 posts from Sina Blogs. In order to maximize the performance of our final classifier, we trained a new version that pooled both the training and testing annotation data. The expectation was that the performance of this version would be slightly higher than that presented in Table 2, of indeterminate magnitude.

From the full set of 478,550 posts, 64,231 posts were classified as personal narratives (13.4%).

Discussion

We were successful in our efforts to automatically create a corpus of tens of thousands of Chinese personal narratives extracted from public weblogs using supervised machine learning techniques. In comparison to Swanson's (2011) work on narratives in English weblogs, we found that comparable amounts of training data and comparable text classification methodologies yielded comparable classification accuracy. Our comparison of highly weighted lexical features across English and Chinese exhibit many similarities, despite significant differences in grammatical characteristics. These similarities provide us with a solid foundation to begin to look beyond superficial differences in narrative across cultures, examining whether there are structural differences in the way that different cultures compose personal narratives. In order to take advantage of the scale afforded by large collections such as ours, future cross-cultural studies will require automated means of analyzing narrative structure across languages using common analytical schemes.

The strong performance of our text classifier encourages us to apply this technology to larger collections of Chinese weblog posts, where they are accessible, in order to amass narrative collections of a much greater scale. As seen in the subsequent application of Swanson's work on English-language narratives, very large corpora of personal narratives have been used in studies of health-related population studies (Gordon et al., 2012), activity-based narrative retrieval (Wienberg & Gordon, 2012), automated commonsense reasoning (Gordon et al., 2011), and in Swanson's (2011) own research on interactive narrative. In developing an automated approach to the identification of personal narratives in Chinese weblogs, it is now possible to explore the direct application of these technologies to the world's largest population of Internet users.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency, grant number D12AP00069. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Berger, A., Della Pietra, V., and Della Pietra, S. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics 22*, 1 (1996): 39-71.

Brants, T. 2000. TnT: A statistical part-of-speech tagger. *Sixth conference on applied natural language processing*. Association for Computational Linguistics.

Brill, E. 1993. A corpus-based approach to language learning. Dissertation. University of Pennsylvania.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics 21*, 4 (1995): 543-565.

Brown, D. 1991 *Human Universals*. Philadelphia, PA: Temple University Press.

Burton, K., Java, A., and Soboroff, I. 2009. The ICWSM 2009 Spinn3r Dataset. *Third Annual Conference on Weblogs and Social Media*. AAAI Press.

Che, W., Li, Z., and Liu, T. 2010. LTP: A Chinese language technology platform. *Twenty-third international conference on computational linguistics: Demonstrations*. Association for Computational Linguistics.

Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Second conference on applied natural language processing*. Association for Computational Linguistics.

Freund, Y. and Schapire, E. 1999. A short introduction to boosting. *Journal of Japanese society for artificial intelligence, 14*, 5(September):771-780.

Genette, G. 1980. *Narrative Discourse: An Essay in Method*. Ithica, NY: Cornell University Press.

Gordon, A., Bejan, C., and Sagae, K. (2011) Commonsense Causal Reasoning Using Millions of Personal Stories. *Twenty-fifth conference on artificial intelligence.* AAAI Press.

Gordon, A., Wienberg, C., and Sood, S. (2012) Different Strokes of Different Folks: Searching for Health Narratives in Weblogs. *ASE/IEEE International Conference on Social Computing*.

Greene, B., and Rubin, G. 1971. Automatic grammatical tagging of English. Technical Report. Department of Linguistics, Brown University, 1971.

Isay, D. (ed.) (2007) *Listening is an act of love: A celebration of American lives from the StoryCorps Project*. New York: The Penguin Press.

Jin, G., Chen, X. The fourth international chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging. *Sixth SIGHAN Workshop on Chinese Language Processing. 2008.*

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Eighteenth International Conference on Machine Learning* (ICML 2001).

Lee, M. 2012. China's Internet users breach half billion mark. *Reuters.com*, Jan 11, 2012.

Liu, Y., Tan, Q., and Shen, K. 1994. *Contemporary Chinese word segmentation standard used for information processing and automatic word segmentation methods*. Tsinghua University Press. (In Chinese).

Mangione, J., 1996. *The Dream and the Deal: The Federal Writers' Project, 1935-1943*, Syracuse University Press.

Mann, W. and Thompson, S. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8:243–281.

Munson, S. and Resnick, P. 2011. The prevalence of political discourse in non-political blogs. *Fifth International AAAI Conference on Weblogs and Social Media*. AAAI Press.

Nakamura, M., Maruyama, K., Kawabata, T., and Shikano, K. 1990. Neural network approach to word category prediction for English texts. 13th conference on Computational linguistics-Volume 3. Association for Computational Linguistics.

Sproat, R. and Shih, C. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages* 4.4 (1990):336–351.

Sproat, R., Shih, C., Gale, W., and Chang, N. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics* 22.3 (1996): 377-404.

Swanson, R. 2011. Enabling open domain interactive storytelling using a data-drive case-based approach. Dissertation. University of Southern California.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics.

Tseng, H., Jurafsky, D., and Manning, C. 2005. Morphological features help POS tagging of unknown words across language varieties. *Proceedings of the fourth SIGHAN workshop on Chinese language processing.*

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wienberg, C. and Gordon, A. (2012) PhotoFall: Discovering Weblog Stories Through Photographs. *21st ACM Conference on Information and Knowledge Management (CIKM-2012)*, Association for Computational Machinery.