Shout Out: Integrating News and Reader Comments

Lisa Gandy, Nathan Nichols, Kristian Hammond
Northwestern University
2133 Sheridan Rd, Evanston, IL 60208

[lmg-215,ndnichols,khammond]@cs.northwestern.edu

ABSTRACT
A useful approach for enabling computers to automatically create new content is utilizing the text, media, and information already present on the World Wide Web. The newly created content is known as "machine-generated content". For example, a machine-generated content system may create a multimedia news show with two animated anchors presenting a news story; one anchor reads the news story with text taken from an existing news article, and the other anchor regularly interrupts with his or her own opinion about the story. In this paper, we present such a system, and describe in detail its strategy for autonomously extracting and selecting the opinions given by the second anchor.

Categories and Subject Descriptors
I.2 [**Artificial Intelligence**]: Natural Language Processing – *discourse*, *text analysis.*

General Terms Algorithms, Design, Human Factors

Keywords Machine-generated content, cosine similarity, emotional valence detection

1. INTRODUCTION
Building machines and systems that automatically create new content has long been a goal of computer science [1][2][3]. These approaches have typically been hampered by strong domain-dependance and a subsequent lack of scalability. We have been working on a new approach to this problem that we call "machine-generated content."

News at Seven [5] is one such machine-generated content system which creates an automatically generated audio/visual news show complete with animated anchors and text-to-speech generated dialogue. News at Seven consists of several "dynamics" that create a certain type or style of presentation. In this paper, we present and discuss in detail a new dynamic: *Shout Out*. In Shout Out, one anchor presents a news story, while another anchor regularly interrupts with his own short thoughts and opinions about the story.

With the explosion of social media, the average Internet user has more opportunities to comment and express their opinions than ever before. For example, Internet readers often comment directly on the story they just read using a comments section provided by the publisher or website. Because these comments are directly on the page, they are typically more tightly focused on the content of the original than a blog posting that links to the article, for example. It is these reader comments that are ultimately presented by the second, interrupting anchor in the final presentation.

In the remainder of the paper, we discuss the technical details of the Shout Out dynamic and conclude with a user study.

2. THE SHOUT OUT DYNAMIC

2.1 Choosing the Initial News Article
The system begins by choosing the most popular story from an RSS feed. For entertainment news, with its strong focus on celebrities, movies, and other named entities, the system scores the popularity of an article based on the popularity of the named entities in the title of the article. To perform the entity detection, the system uses a locally developed entity extraction service known as WPED (Wikipedia Entity Detector) to extract the entities. The entities are then ranked by popularity using predefined popularity rankings available online; for example, MSN X Rank [4] is used to judge the popularity of actors, musicians, and other entertainers, while TV Guide's Top 100 TV shows [7] is used to score the popularity of a mentioned television show. For other news genres, the system relies on the story's ranking given by Google News.

Once the original article has been chosen, the system follows a two-stage web-scraping strategy to extract the content of the story itself and the reader comments.

2.2 Finding Related News Articles
The original article often does not have enough reader comments to create a Shout Out dynamic. Therefore the system gathers comments from news articles which share the same topic as the original article.

In order to find these related articles, the system uses Google News. It begins with the original article's title. The system then uses the WPED to extract entities from this title as well as all quoted text from within the title itself. So, for example, if the original title is '*For Your Entertainment': Adam Lambert's debut'* single the search term will be *For Your Entertainment Adam Lambert*. We also add a date range to the search that ranges between twelve hours before the publish date of the original article and the publish date itself. The system then uses the Google API to query Google News. Once results have been gathered, the system verifies that at least one entity name is found in the title of the newly retrieved article.

2.3 Discovering and Retrieving Comments
Once related articles have been found, comments are collected from the original and related articles and the system is ready to select which comments will appear in the final presentation.

2.3.1 Filtering Comments

One feature of comments found in comment forums is that they are often agreements or disagreements with other commenters. In order to detect and discard comments which are replies to other commenters we have a simple and effective strategy.

First, as the system gathers reader comments, it keeps a list of commenter names and also extracts all entity names from the news article itself. The system then parses each comment and ensures that another commenter's name is not included in the text. However, if there is an overlap between a commenter's name and an entity name inside the article, then the comment is accepted. Finally, if a comment is ultimately found to be reply then it is discarded and it is not used in the Shout Out dynamic. Additional filtering includes length of the comment and appearance of profanity.

2.3.2 Ranking Comments

Because the Shout Out dynamic calls for a back-and-forth dialog between the news-reading and comment-reading anchors, the system needs to associate each comment with the paragraph to which it is most relevant. The system estimates the semantic relevance between a comment and a news article by measuring the cosine similarity between the original news article and reader comment, after all proper nouns have been removed from both. We remove proper nouns because we observed that if a particular proper noun occurs in a news article and a reader comment frequently, then the cosine similarity score will be high, but the actual content of the comment and the news article might not be similar.

Though the system prefers comments which have a cosine similarity score with the news article that is greater than zero when proper nouns are removed, at times there are too few comments which display this attribute. Therefore if too few comments are found with an appropriate cosine similarity score then the system also performs a second pass without removing proper nouns.

2.4 Creating a Dialog

The system matches mined comments with paragraphs in two passes. At this point, comments have been divided into two groups: group 1 includes those with proper nouns removed and group 2 includes those with proper nouns intact. These groups have been ranked by their cosine similarity to the paragraph. During the first pass the comment the comment from group 1 with the highest cosine similarity, which must also be greater than 0, is chosen. However if no comment meets these requirements, then a comment from group 2 is chosen.

Because all reader comments are presented by the same anchor, it's important that all the comments included in the final dialog have the same sentiment. The sentiment level of a comment measures how positive or negative a comment appears to be. We use a SVM classifier for this task, using a corpus created by Pang and Lee [6]. This corpus is mined from the Internet Movie Database archive of the rec.arts.moviews.reviews newsgroup. Pang and Lee found that using the Support Vector Machine classifier with unigrams and feature presence resulted in a three-fold classification accuracy of 83%; therefore we also follow this strategy and use unigrams and only take into account feature presence.

Once a comment is chosen for a particular paragraph, the sentiment of that comment is recorded and all other comments used have the same sentiment orientation (either positive or negative) thereafter. In the current version of the system, sentiment is only recorded at the sentence level, not at the entity level. Once a dialog is created it is tagged with images and becomes part of a flash presentation.

3. USER SURVEY AND RESULTS

We conducted a survey in which 56 participants were asked to rate how compelling two different Shout Out segments were to them personally. Results are given below in Figure 1.

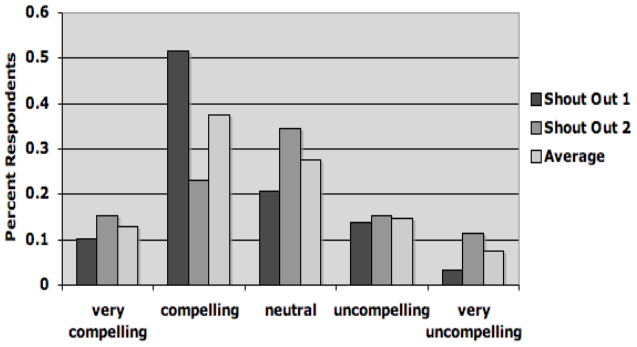

Figure 1. Survey results for final Shout Out presentation.

We do not expect that News at Seven and the Shout Out dynamic will appeal to everyone, or replace traditional written news; rather, we think it has a useful place as available for users who enjoy it, and ignored by users who do not. In this context, we're excited that an average of half of the participants in the survey found the show either compelling or very compelling. In the near future, we are planning to conduct a much larger and more in-depth user-study survey to find the specific areas of strength and weaknesses in the system.

4. REFERENCES

[1] Cohen, Harold. The further exploits of Aaron, painter. Stanford Humanities Review, Volume 4, Issue 2. 1995. Pages 141-158.

[2] Cope, David. Experiments in Music Intelligence. In Proceedings of the International Computer Music Conference, San Francisco: Computer Music Association. 1987.

[3] Meehan, James R. TALE-SPIN, An Interactive Program that Writes Stories. In the Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 1977. Pages 91-98.

[4] MSN X-Rank Live Search. ?http://www.bing.com/xrank FORM=R5FD. Last Retrieved in 8/12/2009.

[5] Nichols, N. and Hammond, K. Machine-Generated Multimedia Content. Proceedings of the Second International Conference on Advances in Computer-Human Interactions, 2009.

[6] Pang, B., Lee, L.A. and Vaithyanathan,S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing (July 6, 2002), 79-86.

[7] TV.com. http://www.tv.com/shows/top-shows/today.html. Last retrieved on 8/10/2009.