

Tell Me More, not just “More of the Same”

Francisco Iacobelli
f-iacobelli@u.northwestern.edu

Larry Birnbaum
birnbaum@cs.northwestern.edu

Kristian J. Hammond
hammond@cs.northwestern.edu

Intelligent Information Lab.
Northwestern University
2133 Sheridan Rd.
Evanston, IL 60208, U.S.

ABSTRACT

The Web makes it possible for news readers to learn more about virtually any story that interests them. Media outlets and search engines typically augment their information with links to similar stories. It is up to the user to determine what new information is added by them, if any. In this paper we present Tell Me More, a system that performs this task automatically: given a seed news story, it mines the web for similar stories reported by different sources and selects snippets of text from those stories which offer new information beyond the seed story. New content may be classified as supplying: *additional quotes*, *additional actors*, *additional figures* and *additional information* depending on the criteria used to select it. In this paper we describe how the system identifies new and informative content with respect to a news story. We also show that providing an explicit categorization of new information is more useful than a binary classification (new/not-new). Lastly, we show encouraging results from a preliminary evaluation of the system that validates our approach and encourages further study.

Author Keywords

New information detection, Information Retrieval, Dimensions of Similarity

ACM Classification Keywords

H.4 Information Systems Applications: Miscellaneous; H.3 Information Storage and Retrieval: Information Search and Retrieval—*Information Filtering*

General Terms

Design, Human Factors, Reliability

INTRODUCTION

The Web offers tremendous opportunities to contextualize information through aggregation and hyperlinks, making it possible for news readers to learn more about virtually any story that interests them. And indeed most online news sources

take advantage of these opportunities by presenting, alongside their stories, lists of “Related Stories” or other media related in some way to those stories. The potential advantages are clear: For users, to create a richer news experience, providing more background, or more detail, than any single story can present. For publishers, to increase the utilization of their content.

However, in many online news outlets the additional information that is presented to audiences is explicitly determined by human editors using their expert judgment. The problem with this approach is that it isn’t scalable. Additionally, when content is automatically generated, the quality of the results often suffers. Yes, readers are presented with “related” information, but too often this information is just a rehash of the story they started with [23] and no guide is provided as to what they add to the main story, if anything.

In this paper we present Tell Me More. A system that doesn’t just present readers with “more of the same.” Instead, it selects stories that go beyond the initial story and presents them in a way that creates a genuinely richer news experience. Tell Me More uses the content of those stories to select and display only stories containing information that is new with respect to the original news story.

In particular, we claim that Tell Me More selects and displays paragraphs from other news stories containing details and background information that are new with respect to the original story. More specifically, the system retrieves new actors, new quotes and new figures which, by themselves, are important kinds of new information.

In terms of the presentation of new information, prior systems aim at detecting new information based on a single score that determines whether a piece of information is new or not. In this paper, however, we provide evidence that presenting this new information in categories that make the selection criteria visible, is more useful than aggregating snippets in one big list with no clues to assess the new information contained in them.

The following sections present the architecture of Tell Me More, a small user study on the value of categorizing information and an initial evaluation of the user experience with Tell Me More. We, then, discuss related work and finish with conclusions and future work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. IUI’10, February 7–10, 2010, Hong Kong, China. Copyright 2010 ACM 978-1-60558-515-4/10/02...\$10.00.

Figure 1. Screenshot of Tell Me More showing three kinds of new information

ARCHITECTURE

The goal of our system, given a seed news article, is to find other sources reporting on the same situation and then to present only the paragraphs that contain supplementary information reported by these sources that is not present in the seed article. Figure 1 shows Tell Me More’s interface with three of four possible kinds of new information: additional actors, additional figures and additional quotes.

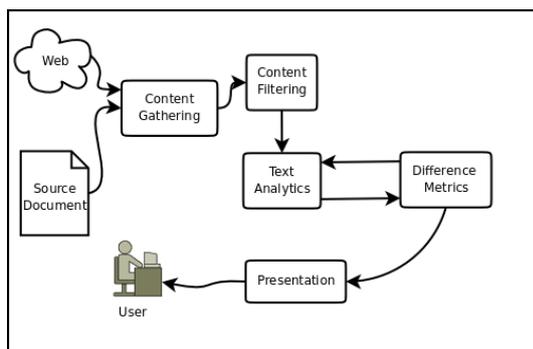


Figure 2. System architecture

Core Modules

Tell Me More employs five core modules to extract and present new information to the user. Given a source document, the

system uses it as a seed for retrieval of similar stories and comparison. At the end of the cycle the user is presented with a categorization of the new information discovered, if any. The five modules (see Figure 2) are described next.

1. **Content Gathering:** This module can gather documents from search engines as well as from user defined sources. The implementation presented in this paper develops a basic model of the source article through text analytics (see module 3 below) and uses this to form a query that is sent to the Google News API. When Google News finds a news story, it usually returns a cluster of very similar stories embedded in the result set. The module then scrapes the text from the highest ranked story and the cluster of stories associated with it. The system has a number of strategies for forming queries so that if the first query does not return results, it is always possible to form a second and a third query to this effect. These techniques are based on our previous work [2, 7] and are described in some detail in the next section.
2. **Content Filtering:** This module filters results and eliminates documents that are exactly the same as the seed document or documents that are too different, and hence probably concern different situations. Content filtering also takes care of discarding articles that look like compendia of articles or spam using simple heuristics.

3. **Text Analytics:** This module is used to develop a statistical and heuristic model both at the paragraph and at the document level on the source article, as well as analytics on every incoming new article. For every new source that it examines, it detects entities, quotes, figures and also computes a vector representation of each paragraph.
4. **Difference Metrics** After the text analytics are gathered for a new document, this module compares them with the previously seen text and determines which paragraphs contain new information. In addition, the Difference Metrics module keeps track of the kind of new information contained on each selected paragraph.
5. **Presentation** This module categorizes and ranks the new information based on the output of the Difference Measures module.

The following sections treat each of these modules in detail.

Content Gathering

Once a seed document is provided, this module uses text analytics [2] to extract key features from it, specifically, the title, named entities and a histogram of content bearing words. This module, as implemented for this paper, tries to search Google News (at this point the system is able to also use Google, Yahoo News and user specified websites) using the title of the news story as a query. If Google News has indexed the given story, it is likely that its URL will be at the top of the result set and, associated with it, Google returns a cluster of URLs of the most similar stories.

It may be the case that Google News has not indexed many similar stories and therefore it will not return the cluster of stories associated with the search query. The content gathering module detects this situation and then performs, in order, the following additional queries:

1. A query with the most frequently occurring words from the histogram plus the named entity most frequently mentioned. Once Google returns a result set, the module looks for the highest ranked result and its cluster, if present.
2. If a cluster is still not present, the module retains the top five URLs returned by Google.

After the URLs of these additional news resources have been gathered, they are processed by a webpage scraper which obtains the text of the news by looking into the DOM tree of the webpage and selects the `div` tag with most readable text in it (i.e. not scripts nor tags). These methods are largely based on our previous work [2, 7]. The scraped content is then stored and handed on to the next module: Content Filtering.

Content Filtering

Because a web search can produce sets with documents that are not too similar, Tell Me More takes a further step by filtering documents based on their similarity to the seed story. We compute this similarity using Latent Semantic Analysis[12] (LSA). LSA¹ extracts semantic information of words

¹LSA is based on singular value decomposition (SVD) of the term x document matrix. SVD is an algebraic transformation on a matrix

based on contextual information provided by the training data. Therefore, we trained LSA on actual news stories – 1000 news stories about popular topics from the web [7]– to provide a good context for semantic comparison of news stories. Content Filtering, then, creates a multi dimensional vector representation of the seed document, and of each additional document gathered in the previous module, using this LSA space. Each vector is computed by averaging the LSA vectors of the individual words of the story excluding stop-words. With this corpus and our implementation of LSA, a document similar enough that it is likely to be referring to the same situation as the original seed article, but that contains meaningful differences, can have a similarity score of about 0.8 when compared to the seed article. Therefore, Content Filtering will only allow documents with similarity score x where $0.8 \leq x < 1.0$. In addition, this module has simple heuristics, based on [7], to exclude articles that look like news summaries and articles that may be spam.

Text Analytics

Tell Me More is based on the premise that the presence of new named entities, new quantifiers, and additional quotes are important kinds of new information. It is necessary, then, to have a module that analyzes the textual information accordingly. In addition, this module computes a semantic representation of each paragraph using LSA.

Thus, each paragraph in each document retrieved is represented as a vector with four features (a) a latent semantic representation of the paragraph; (b) a list of entities; (c) a list of quantifiers and (d) a list of quotes present in the paragraph. Each of these features is discussed in further detail below:

A Latent Semantic Representation of The Paragraph

Sometimes, new information may be present at a semantic level. Therefore, this module computes a multi dimensional vector representation of the paragraph using LSA and tags each new paragraph with this information.

List of Entities

When named entities are introduced in a paragraph of a news stories they usually signal important information such as evidence supporting claims in the news story[1]. Based on this idea, our system tracks entities mentioned in each paragraph within a few categories: persons, places, cities and organizations. These are computed for each paragraph. Entities that refer to the same base entity (e.g. “O.A.S.” and “Organization of American States”) are normalized. That is, mapped to a single entity. We detect and categorize entities using a boosting approach to entity detection based on two named entity recognizers (NERs) described below. Finally, this approach allows the system to tag some entities as “well-known.”

such that a term-document matrix T can be decomposed as follows: $T = USV^T$. Then the dimensions of S, U and V are reduced to produce an approximation of T . Finally, the matrix TT^T is used as a term co-occurrence matrix[5].

Boosting an entity detection system

Due to its robust support for entity recognition, we used OpenCalais², an off the shelf commercial entity detection system that works very well for detecting instances of common categories such as people, places and organizations. However, because OpenCalais lacks good normalization capabilities, we augmented it with a modified version of WPED, an in-house entity detection system based on Wikipedia entries that does a good job at detecting different instances of the same entity. WPED has been used previously in other intelligent systems [15].

OpenCalais is a Thomson-Reuters free web service that performs named entity recognition and extracts relationships and events from text. OpenCalais uses natural language processing techniques and machine learning to recognize instances of named entities. Because OpenCalais is not based solely on hand-crafted databases of entities it can recognize new entities that may not yet be incorporated in any database, but that can be detected based on surface features of the text such as capitalization. The main weakness of OpenCalais is that it doesn't always normalize instances of entities. For example, "*Ted Kennedy*" and "*Senator Kennedy*" are detected as two different entities by OpenCalais when in reality, they are two instances of the same entity.

WPED, on the other hand, matches text to Wikipedia entries and therefore can normalize entity names with multiple instances —as long as they are pointing to the same wikipedia article. In a departure from other wikipedia named entity recognizers[10] or "wikifiers" [14, 3], WPED maps entities to overarching categories such as "person," "location," "organization," etc. thus, making uniform the, far-from-standard, default classifications provided by Wikipedia users. This allows compatibility with the classifications used by OpenCalais.

In sum, OpenCalais can detect people's names and organizations that may not appear in Wikipedia. Therefore it is the main entity detection method used. We, then, boosted the entity recognition process using WPED as follows: if any of the entities from OpenCalais can be mapped to an instance of an entity detected by WPED, we use the WPED base entity for annotation in the Text Analytics module.

Lastly, we considered the entities detected by both OpenCalais and WPED to be "well known" compared with those detected only by OpenCalais. This becomes relevant for a later discussion ranking paragraphs.

List of Quantifiers

Bell [1] argues that numbers in a new story are a marker of evidence and relevance. In particular, numbers that serve as quantifiers are a strong indication of relevant information.

For example: the number 100 could be a quantifier of time, money, or people involved in an accident or it could carry a different semantic content altogether as in the phrase "100

percent successful" where it serves as a synonym for "totally" and carries little information.

By discriminating and choosing only quantifiers, the system ensures it has detected a true piece of evidence or a fact related to the story.

For the present version of the system, we decided that any number that was in the proximity of a proper noun was a quantifier. It is usually the case that in news stories quantifiers and the object they quantify are in the same sentence. Using Montylingua, a part of speech tagger (POS) developed at MIT [13] we detect whether plural proper nouns occur in the same sentence together with a number. When this happens we add the number to the list of quantifiers. This gives us the desired effect: In the case of "100 percent successful", the system will ignore the number "100" as a quantifier.

List Of Quotes

Additional quotes usually indicate the opinion of experts, witnesses or relevant officials with respect to an event. Therefore, one of the analytics we collect are quotes. The system collects a list of text between quotation marks or **"e;**HTML markers per each paragraph.

To summarize, Text Analytics computes LSA vectors of the paragraphs, and collects quantifiers, entities and quotes. This conforms the four features of a vector of meta information about the paragraph that is then handed to the difference metrics module.

Difference Metrics

The difference metrics module compares the vector for each paragraph to the vectors of all previously selected paragraphs, including all the paragraphs of the seed story, and determines which ones contain new information based on meaningful differences.

The comparison of paragraphs with the previously seen documents occurs at two levels: At the paragraph level and at the document level. At the paragraph level, the system compares each new paragraph with each of the previously collected paragraphs using LSA and determines whether the new paragraph is sufficiently semantically different, below a threshold, to be considered new information. At the document level, each new paragraph's entities, quantifiers and quotes are compared to those detected in the previously seen text as a whole.

Each of the features of a paragraph (LSA representation, entities, quantifiers and quotes) has its own difference metric that will be computed and stored. Later on, the presentation module will use these scores to categorize the difference and present the paragraphs in the correct category. Each difference metric is now discussed in detail.

Semantic difference with LSA

To determine semantic difference we compare each new paragraph with all the previously seen ones. The most different paragraphs will naturally be least similar. Therefore, if

²<http://www.opencalais.com>

paragraphs score below a threshold in similarity, they are considered semantically different. For each paragraph we keep track of the highest similarity score obtained as a proxy for how similar or different the paragraph is with respect to other paragraphs. The highest similarity score is obtained by a cross-paragraph comparison between any new paragraph and those previously retrieved. The maximum similarity score is computed as follows:

$$Score_i = \max_{k \in P} (sim(v_i, v_k)); P = \{0 \dots i - 1\}$$

Where v_i is the LSA vector representation of paragraph i , v_k is a vector of any of the previously seen paragraphs. The similarity function sim is the cosine similarity between the two LSA representations.

This score allows us to determine a baseline similarity between a new paragraph and the previously seen text. If the similarity is below the threshold of 0.3, the paragraph is considered to be semantically different from any other paragraph and therefore a candidate for containing new information. Conversely, high similarity scores indicate little semantic difference between the new paragraph and the previously seen text and, therefore, it is not a candidate for new information by this metric. This threshold has been pre-tested and works well in practice with our LSA space.

Different Entities

When a new paragraph is processed, its entities are compared to all entities associated with previously seen paragraphs. This is a straight string comparison except for one caveat. If an entity detected in the new paragraph is a substring of an entity already seen, both entities are considered to be the same. The rationale for this is that because the articles being compared are very similar, it is very probable that a substring of an entity is a reference to that entity. For example, if an article mentions “*Barack Obama*” and later on mentions “*Obama*” it is likely that it is referring to the same entity.

If a paragraph mentions at least one new entity, the paragraph is considered sufficiently different and therefore, containing new information. The score for this metric is equal to the number of new entities detected.

Different Quantifiers

This algorithm is similar to that for entities. The system compares the quantifiers of the new paragraphs with those previously collected from the processed documents. Again, if a paragraph includes a new quantifier, it is considered to be different than the rest due to this new information. The score of this metric is equal to the number of new quantifiers detected. Although previous systems have used dollar amounts as a measure of novelty [18], Ours is, to the best of our knowledge, the first to use quantifiers in a more general sense.

Different Quotes

Again, a similar algorithm compares the quotes present in the new paragraph with quotes previously seen in the collected text. However, for this feature one cannot rely on straight string comparison of the quotes under consideration and the previously collected ones.

When journalists transcribe quotes obtained orally, they make editorial choices and it may be the case that transcripts come out slightly differently or that the quotes are edited at slightly different points. For example, one quote may start with “You know, when I was (...)” and another source may quote the same person omitting the initial “You know.”

Because we compare quotes in the current paragraph against all previously contributed quotes we have to take into account these minor editorial choices. Therefore, straight string comparison is not enough. Comparisons need to be somewhat flexible. To address this, we scored string similarity using the Smith Waterman algorithm [11] commonly used in bio-informatics to align DNA sequences. Smith Waterman is a variant of the longest common sub-sequence algorithm that detects the longest common alignment of letters between two strings. The similarity score consisted in the percentage of aligned letters with respect to the longest string. We set a threshold of 70% which, empirically, was the lowest score at which we considered both quotes to be the same. Therefore, a quote is considered different if it matches less than 70% of any pre existing quote in previously seen documents.

When quotes are found to be different, they are scored by dividing the length of the quote with respect to the length of the paragraph in terms of letters. Because paragraphs are usually complete units of thought, this score is a good proxy of the prominence of the quote in that context. Therefore, more prominent and less edited quotes should obtain higher scores.

To the best of our knowledge, novelty detection systems have not considered quotes as a unit of differentiation for new paragraphs and we believe they are a valuable source of new information in the detection of novel information with respect to a news stories.

In sum, the Difference Metrics module compares the meta information about paragraphs obtained in the Text Analytics module. Namely, the LSA representations from previously seen paragraphs and the previously seen entities, figures and quotes. It assigns a score to each of these comparisons that establishes how different the new paragraphs are compared with to previously seen text, and, more importantly, along which dimensions.

Presentation: More than single scores

In our view, presenting new information based on a binary classification (new/not-new) provided by a single, all encompassing score lacks the usefulness that comes from making the criteria for novelty selection visible. A binary classification assumes that the new information should be obviously new to all users. However, this is not necessarily the

case. People may not be in agreement of what constitutes novel information when they compare news stories.

Individual differences in the perception of novelty vary greatly. This is evident when researchers have looked at human assessments of what counts as new information. For the TREC conferences that hosted a novelty track, inter rater reliability scores, that incorporate agreement as well as correction for disagreements, measured by Cohen’s Kappa, were 0.54³ for relevance and even lower for novelty[20]. Schiffmann [17] tried to build a corpus of pairs of documents for the detection of new information and obtained a Kappa of 0.24 on judgments of novelty. Other methods to improve reliability have been explored, such as evaluating agreement on the answers to questions about the different texts (fact-focused novelty detection) but the Kappa scores are still low[16].

Because people do not always agree on judgments of information novelty, we think it is particularly important to make the selection criteria for new information visible to users. Research on presenting search results suggests that categorizing the results in meaningful ways helps user navigate them more easily [4, 8]. Thus, Tell Me More recommends paragraphs in one of four sections based on the difference metrics obtained earlier: (a) “additional actors” are new proper names, countries, cities and organizations; (b) “new numbers” which in this first iteration are any numbers not appearing in the story that the user is reading; (c) “additional quotes” are quotes not appearing in the source document and (d) “supplementary information” is text that is semantically different from previously seen text. Later in this paper we present a user study in which we compare our approach to a binary classification version of our system.

Currently, the system presents at most two paragraphs with new information per category on the front page (see Figure 1); however, it gives the users the opportunity to explore other new paragraphs within each category by clicking on “more additional < category_name >.”

To be able to know the category for a paragraph, the system looks at the difference metrics and determines its score in each category. Currently, the system has one rule for classifying paragraphs: if the new paragraph contains a new quote of length greater than 80% of the paragraph, then the paragraph should be classified as “additional quotes.” Otherwise, the paragraph should be classified under the difference metric that is most distinctive of the paragraph –that is, the one with the highest score.

Because only two paragraphs show on the main webpage it is necessary to have a ranking mechanism that attempts to put the most relevant new paragraphs there.

Ranking new information

When presenting paragraphs with new information in categories, we rank these paragraphs according to the counts of new information detected by the Difference Metrics module in descending order. There are two exceptions:

³Usually, a Kappa above 0.65 is considered decent reliability.

1. **additional actors:** Because research suggests that there are entities that may matter more to the average reader, and that these entities tend to be the most well known or influential [1] we think that having a wikipedia page (i.e. WPED finds them) is a good proxy to measure reader interest in the entity. Therefore, we rank first the paragraphs by number of popular actors present and then by the number of non-popular actors.
2. **Supplementary Information:** Because low scores indicate greater semantic difference, supplementary information is ranked by its scores in ascending order.

In the next sections we present a user studies to evaluate the validity of our approach in terms of categorizing results and a survey on the user experience with Tell Me More.

USER STUDY. BINARY CLASSIFICATION VERSUS MULTIPLE DIMENSIONS

When evaluating new information, most systems judge novelty with a single score and determine whether the information is new or not with respect to some seed document. As explained in Section , We believe that a single score is not informative enough for the task of reading news stories and that systems should make the dimensions of novelty visible to users.

Methodology

To test this hypothesis we conducted a small user study where participants saw a Tell Me More webpage with new information on the side. Users were told to read the new information and find two names that did not appear in the main article and at least one new fact that did not appear in the main article. Then they were told to look at the same news story with a different version of the interface. Users were then asked: “Do you think this format would have made finding the previous information easier or harder?” The two interfaces were: (a) MH: information of the sidebar is categorized under multiple headings: additional actors, additional figures and additional quotes and bolded entities, quotes and figures when applicable; and (b) OH: the new information was presented without any formatting or categorization and the only heading of the sidebar read “New Information.” We counter balanced the order in which users saw each interface. 11 adults participated in this study.

Results and discussion

All but one participant preferred the interface with multiple headings (MH). This difference is significant ($\chi^2(1) = 11.63; P < 0.001$). Despite the small sample size, due to the huge difference in percentages (9% preferred OH versus 91% who preferred MH), the statistical power, considering a significance level of 0.05, is 92%, which is a very strong suggestion that no matter how big the sample size, the results are bound to suggest that users prefer the information when it is presented in a categorized manner and when it highlights the new information contained.

In other words, we find strong support for the hypothesis that users prefer an interface that makes visible the criteria for

selecting new information, when the task consists of finding names or other supplementary facts. This is a departure from previous approaches to the detection and presentation of new information which do not make the selection criteria visible to users in the presentation [6, 17].

EVALUATION OF THE INTERFACE

At this point we believe that comparing Tell Me More with standard news readers would not be the as informative as we would like. The reason for this is the lack of a truly comparable news reading interface free of, or with few, confounds and that can provide ecological validity to our results.

Therefore, for this evaluation we are concerned with the subjective user experience and trust on Tell Me More. In particular, we are interested in users' initial reaction to the following questions: does Tell Me More help users understand news better? Does it provide a trusted source of news? Does it provide relevant detail or background information with respect to the main story? And, do people like the ability to see new information side by side with the news?

Methodology

To evaluate the user interface of Tell Me More more generally, we asked participants to read a news story and respond to a questionnaire with respect to their experience. The stories were chosen at random from among 25 stories that were, themselves, chosen randomly from stories utilized by the system. Their only requirement is that they had at least one piece of new information in them. There were at least two stories in each of the following topics: politics, entertainment, world news, business, technology, health, sports and crime. The questionnaire asked people to evaluate their experience in terms of organization of new information, ease or difficulty of finding new information and trust in the information presented (as a whole and separately for the seed article and new information snippets). The questionnaire also asked people to rate the truth of statements about relevant background information in the snippets and relevant details. It then asked the users to rate their understanding of the topic after reading new information snippets and whether they would like to have an interface with new information in their news readers. Finally we collected some demographic information such as how often did users read news online, from how many sources, participant's age, occupation and gender. 24 adults responded the survey and all, except one, read news online at least once a day. 79% of the respondents actually consult more than 1 source. There were 13 females and 11 males and approximately 54% were between 25 and 35 years of age, 29% were younger than 25 and 17% older than 35.

Results

To answer the question of whether Tell Me More helps users understand news better, we analyzed the response to the item "I have a better understanding of the topic of the news story thanks to snippets of new information presented." which was responded to using a 5 point Likert scale that went from "Strongly disagree" to "Strongly Agree." We grouped the

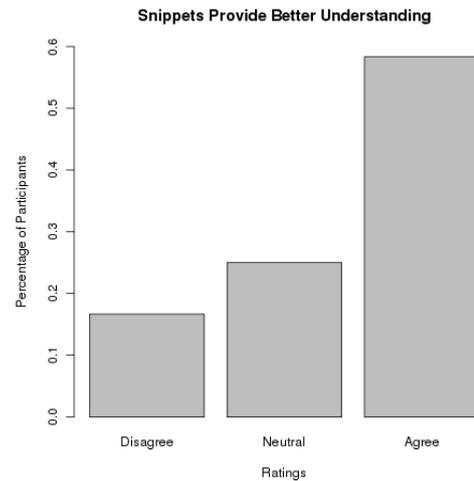


Figure 3. Level of agreement over having a better understanding of the news thanks to the snippets of new information. Difference between agreeing and other responses is significant to the $p < 0.05$ level.

entries in three groups that can be seen in Figure 3: (a) people who disagree with the statement, that is, scores of 1 or 2; (b) people that neither agree nor disagree, that is responses of score 3; and (c) people that agreed with the statement, that is, scores of 4 or 5. A test for equality of proportions $\chi^2(2) = 10.5; p < 0.01$ reveals that the difference between the percentages of the three groups are significant. Post hoc analysis, using a test for equality of proportions, shows that the difference between (a) and (b) was not significant, but the differences between those groups and those who agree (c) was significant at least at the $p < 0.05$ level.

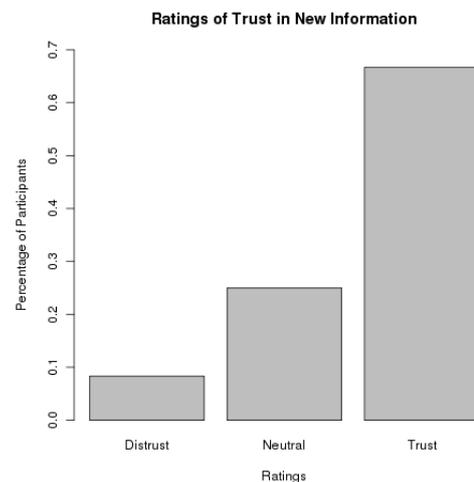


Figure 4. Ratings of trust of the new information presented. All differences are significant to the $p < 0.005$ level, except for Neutral > Distrust

Approximately 66.7% of the respondents rated the item "Rate your trust with respect to the webpage as a whole" with a 4 or a 5. A linear regression shows, not surprisingly, that trusting the new information sections is the main driver for people to

trust the webpage as a whole ($F(2, 21) = 10.37; p < 0.01$). Therefore, we analyzed the responses to the item “Rate your level of trust with regards to the new information presented.” The item had ratings on a 5 point Likert scale that went from “totally distrust” to “totally trust.” Again, for the purpose of analysis we grouped the responses in three groups. (a) Those who distrusted the information (scores 1 and 2); (b) those that remained neutral, that is neither trusted nor distrusted (score 3); and (c) those who trusted the information (scores 4 and 5). Figure 4 shows these ratings. A test for equality of proportions ($\chi^2(2) = 19.5; p < 0.001$) shows that the differences in ratings are significant. A more detailed analysis shows that all differences are significant at the $p < 0.005$ level, except for the difference between group (a) and (b) which were not significant.

Our next question had to do with the relevance of the information presented in terms of background information and additional details. Participants were asked whether they agreed or disagreed, on a five point Likert scale, with the following statements: “Relevant background information was contained in the new information snippets” and “Relevant additional details were contained in the new information snippets.” By analyzing the data using the same methodology of grouping the ratings in three, we found that most respondents agreed with those two statements: 62.5% agreed with the statement about relevant background information ($\chi^2(2) = 13.85; p < 0.005$) and 75% agreed with the statement about relevant details ($\chi^2(2) = 28.5; p < 0.001$). See Figure 5

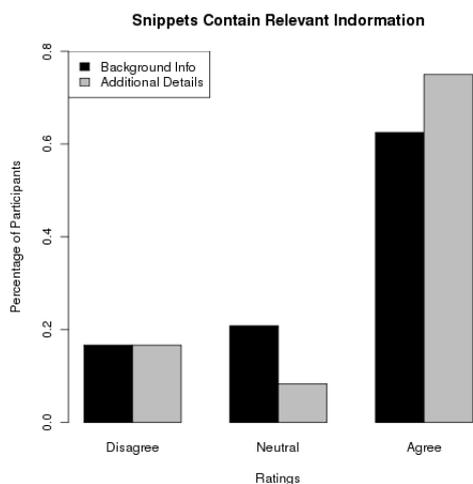


Figure 5. Agreement over the relevancy of background and details contained in the new information.

Our last question was to see whether people would like to have an interface like this one in their regular news reading experiences. Participants had to rate their agreement with the item “I would like to see an interface like this in my online news sources” on a five point Likert scale from “Strongly disagree” to “Strongly agree.” Again, for the analysis we grouped them in three: (a) disagree (scores 1 and 2); (b) neutral or neither agree nor disagree (score 3); and (c) agree (scores 4 and 5). Approximately 58% of the re-

spondents agreed with the statement. A test for equality of proportions show that there is a significant difference in the scores ($\chi^2(2) = 10.5; p < 0.05$) and detailed analysis shows that the difference between group (c) and (a) is significant to the $p < 0.005$; the difference between groups (c) and (b) is slightly significant ($p < 0.07$) and the difference between groups (a) and (b) was not significant. See Figure 6

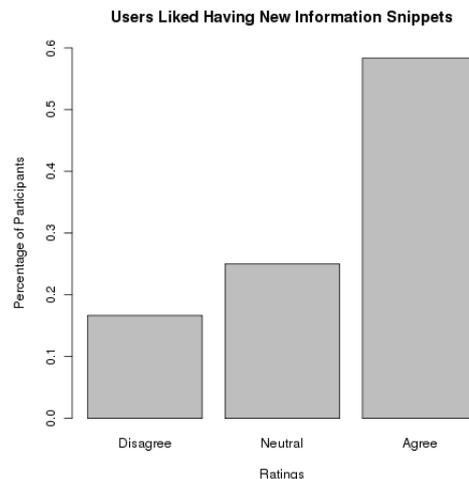


Figure 6. Users liked the ability to have new information snippets along with the news

Users were allowed to freely write feedback on the project. Most feedback had to do with the names of the headings, how little attention they pay to headings on a side bar, how the highlighting of a quote and a name were misplaced under “Additional Actors.” Other feedback included suggestions on clustering and ranking. Positive feedback had to do with users liking the “Additional Quotes” section and the overall utility of the system.

Discussion

The evaluation of the interface suggests that users trust the new information presented. Because 96% of the participants read news online, and 76% of them consult more than one source, it is plausible to assume that when they responded the trust questions, they may have had their usual news sources in mind for comparison. Respondents reported that Tell Me More contains relevant details and background information and that the interface is one that they would like to see in their news reading experience. This suggests that Tell Me More is a feasible and potentially interesting news-reading interface and that it is worth optimizing its individual components and studying their effects on the overall news reading experience.

Relevance is a hard concept to define in terms of user experience, so it would be hard to quantify the “amount” of relevant information. However, the fact that the new information helped people get a better understanding of the news story, suggests that whatever relevant information is contained in the paragraphs is easy to access and augments the main news story in useful ways.

Additionally, the good feedback on the “Additional Quotes” section suggests that this is a meaningful criteria for selecting new information. To the best of our knowledge, Tell Me More is the first system to use quotes, and quantifiers, as indicators of new information.

Further analysis is needed to explore which topics are better served with the current implementation of Tell Me More. Lastly, given the user feedback, better wording and, at times, different organization can make the experience even more useful and pleasant.

RELATED WORK

Systems that detect new information based on one score have produced low to moderate precision rates. Significant research on detecting new information has been carried out as part of the TREC novelty track competitions which ran from 2002 to 2004. The novelty track is comprised of several tasks related to new and relevant sentence retrieval. At TREC 2002 and 2003, the first task was: “given an ordered set of 25 relevant documents, systems should return the relevant and new sentences from this set”[19]. In this task, the highest precision and recall measures were around 0.55 and 0.78.[19] Among the techniques used, the team from the University of Maryland Baltimore[9] used a similar technique to our LSA representation. They used a vector-space model based on SVD to create a matrix of word co-occurrence. Then they used it to compare sentences from each new document to the sentences previously contributed by other documents. On the same task at TREC 2004 the team from The University of Massachusetts scored sentences based on a combination of vector-space model using TFIDF and the mentions of new named entities with respect to previously mentioned ones. Their F measure scores was 0.61. By considering named entities, they consider some of the context of the previously retrieved documents. All this signals the difficulty of systems that use traditional, vector based, similarity measures to judge what is considered new information.

Schiffman[17] argues that, to detect novel information it is necessary to consider both sentence-level information and contextual information. At TREC 2004 Schiffman did not use a vector-space model. Instead he incorporated the detection of named entities, cash amounts, nouns and verbs and the notion that new information usually comes in consecutive phrases or is contained within a phrase in a few words. He employed hill climbing algorithms to detect thresholds that indicated new content. The best precision scores obtained by his team were around 0.6 [18]. Our system, however, makes use of quotations and quantifiers as additional units of information.

New information detection research has largely been used for news summarization software[22, 17]. However, because summarization systems aggregate information from various sources, it is hard to tell why the system includes the texts it does. In the realm of news, researchers have detected new information comparing the word distribution of different documents. Swan [21] used this technique to build time lines of events. Kuo [24] detected new information not only by com-

paring the distribution, but by assigning different weights to different kinds of terms, such as named entities, dates, etc. skewing the document vector representations. However, many of these approaches have been used in hand processed corpora and have selected new information based on a single score model.

The system that is most similar to Tell Me More is NewsJunkie [6]. NewsJunkie utilizes vector representations and entity detection to judge novel content in news, however the novelty detection is used to provide readers with updates, developments and recaps of news stories. In contrast to our system, Newsjunkie does not specify what exactly is new information in the articles presented. Another difference is that Newsjunkie operates on the document as a whole, ranking it according to “how different” it is to the seed story. Tell Me More ranks paragraphs, thus pointing out specific new information contained within additional news stories. Lastly, in a user study of NewsJunkie users expressed the opinion that it was hard to judge the novelty of articles because of their relevance with respect to the seed story. We believe that making the selection criteria visible to users can help bridge that obstacle.

In sum, finding relevant new information is not a trivial task. Previous systems have used different criteria than we do, and none have made their selection criteria visible to users. We believe this is an essential component of a usable system that presents new information about news articles to users.

CONCLUSIONS

In this paper we presented Tell Me More, a rich news reading system that displays new information alongside a news story. We showed that Tell Me More, as designed, selects and displays paragraphs from other news stories containing information that is new with respect to the original story. New actors, new quotes and new figures are important kinds of new information retrieved and presented by the system. An initial evaluation of the system validates our approach and encourages us to continue development and research.

Our second claim, and the point of user study 1, is narrower yet: to provide evidence that presenting this new information in categories is more useful than aggregating snippets as one big list with no indication as to the nature of the new information contained in them.

In this paper we also propose that quotations and quantifiers are valuable kinds of new information in news reading.

Tell Me More aims to realize the promise of the Web by delivering a truly richer news-reading experience in a scalable and economical way.

REFERENCES

1. A. Bell. *The Language of News Media*. Language in Society. Wiley-Blackwell, September 1991.
2. J. Budzik, K. J. Hammond, and L. Birnbaum. Information access in context. *Knowledge-Based Systems*, 14(1-2):37–53, March 2001.

3. S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
4. S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284, New York, NY, USA, 2001. ACM.
5. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480, New York, NY, USA, 1988. ACM Press.
6. E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 482–490, New York, NY, USA, 2004. ACM Press.
7. F. Iacobelli, K. Hammond, and L. Birnbaum. Makemypage: Social media meets automatic content generation. In *ICWSM 2009*, 2009.
8. M. Käki. Findex: search result categories help users when document ranking fails. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 131–140, New York, NY, USA, 2005. ACM.
9. S. Kallurkar, Y. Shi, R. S. Cost, C. Nicholas, A. Java, C. James, S. Rajavaram, V. Shanbhag, S. Bhatkar, and D. Ogle. Umbc at trec 12. In *TREC Notebook Proceedings*, 2003.
10. J. Kazama and K. Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *EMNLP-CoNLL*, 2007.
11. D. E. Krane and M. L. Raymer. *Fundamental Concepts of Bioinformatics (The Genetics Place Series)*. Benjamin Cummings, 1 edition, September 2002.
12. T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, April 1997.
13. H. Liu. Montylingua: An end-to-end natural language processor with common sense. Available at: <http://web.media.mit.edu/hugo/montylingua>, 2004.
14. R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.
15. N. Nichols and K. Hammond. Machine-generated multimedia content. In *ACHI '09: Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 336–341, Washington, DC, USA, 2009. IEEE Computer Society.
16. J. Otterbacher and D. Radev. Exploring fact-focused relevance and novelty detection. *Journal of Documentation*, 64(4):496–510, 2008.
17. B. Schiffman. *Learning to identify new information*. PhD thesis, Columbia University, 2005.
18. B. Schiffman and K. R. Mckeown. Columbia university in the novelty track at trec 2004. In *Proceedings of the TREC 2004*, 2004.
19. I. Soboroff and D. Harman. Overview of the TREC 2003 novelty track. In *Proceedings of TREC-2003*. Citeseer, 2003.
20. I. Soboroff and D. Harman. Novelty detection: the trec experience. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
21. R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80, 2000.
22. S. Sweeney, F. Crestani, and D. Losada. 'show me more': Incremental length summarisation using novelty detection. *Information Processing & Management*, 44(2):663–686, March 2008.
23. The Associated Press and The Context-Based Research Group. A new model for news: Studying the deep structure of young-adult news consumption. Technical report, 2008.
24. K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 215–222, New York, NY, USA, 2007. ACM.