Automatically Embedding Newsworthy Links to Articles

Hakan Ceylan*
Netflix, Los Gatos
hceylan@netflix.com

Ioannis Arapakis
Yahoo! Labs, Barcelona
arapakis@yahoo-inc.com

Pinar Donmez*
Salesforce.com, San Francisco
pdonmez@salesforce.com

Mounia Lalmas
Yahoo! Labs, Barcelona
mounia@acm.org

ABSTRACT

It is of great interest to news providers such as Yahoo! News to attain higher visitor rates by promoting greater engagement with their content. One aspect of engagement deals with keeping users on the site longer by allowing them to navigate through content with enhanced, click-through experiences. News portals have invested in ways to provide embedded links *within* news stories. So far these links have been manually curated by professional editors, and due to the manual effort involved, the use of such links has been limited. In this paper we propose an automated approach to detecting and linking newsworthy events to associated articles. Our analysis, conducted on Amazon's Mechanical Turk, reveals that our system's performance is comparable to that of professional editors, and that users find the automatically generated highlights interesting and the associated articles worthy of reading.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation

General Terms

Experimentation, Design, Human Factors

Keywords

Newsworthiness, automatic linking, engagement strategy

1. INTRODUCTION

News portals have become a very popular destination for web users who read news online. As there is great potential for online news consumption but also serious competition among news portals, providers must find effective and efficient strategies to engage users longer in their sites. In this paper, we are interested in one type of strategy promoting engagement; *enticing users to browse the site through embedded links within news articles.*

*The work was done while at Yahoo! Labs.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Embedded links, or *hyperlinks*, are a great strategy for prolonging time users spend on a site by sustaining engagement through interactive click-through experiences. Once clicked, users are re-directed to another page showing the referenced content. We studied this phenomenon in the context of news portals, where the embedded links direct users to other pages within the same domain. However, hyperlinks are mostly created by human editors, making it a manual task that is time-consuming and not scalable. We propose an automatic approach to hyperlinking, where for any news article the goal is to identify newsworthy events as a potential source for links. Newsworthy events are more likely to have a related news article already written in the past.

Work in automatically generating hyperlinks has gained new interest, primarily within the context of Wikipedia, looking at cross-referencing documents [8, 10], automatic approaches for hyperlinking [4, 5], and automatically linking documents to encyclopaedic knowledge [6, 9]. Link generation approaches are also used in disambiguation tasks [2, 3, 7]. Many of these works are not directly comparable to ours because of different aims and a reliance on properties specific to Wikipedia. Our focus is to identify the newsworthy events in a given news article and connect them to the appropriate news articles, where it is not likely that all newsworthy events are, or will be, Wikipedia concepts.

In [6] the aim was to generate links from medical reports to Wikipedia pages for explanations or background information. They showed that the approaches in [9, 10] did not yield satisfactory results because medical phrases typically have a more complex syntactic and semantic structure than Wikipedia concepts. They therefore developed their own approaches. Events, as phrased in a news article, have also a specific structure and in addition, very few form concepts in Wikipedia. Similarly, we had to design our own automatic link generation approach.

The goal of our research is two-fold: (i) a fully automated system that constructs hyperlinks in news articles using text processing and understanding techniques; and (ii) assessing the system-embedded links against manually-curated ones by professional editors. This paper focuses on assessing the quality of the system-embedded links and conducted an evaluation using the crowd-sourcing power of Amazon's Mechanical Turk (MTurk).

2. SYSTEM

Our system, called *Linker for Events to Past Articles*, or *LEPA*, has two main components: an indexer and a linker.

2.1 Indexer

The indexer processes articles over a time period by extracting features from each article, and storing them to facilitate faster retrieval. The indexer runs in two stages: the build stage produces an index for a set of articles over a time period, while the update stage is run periodically to add fresh articles to the existing index. For the build stage we constructed an index from articles that spanned over a month, while for the update stage we processed new documents daily and added them to the index.

In both stages, we implemented a simple inverted index approach. The inverted index stores a list of the documents for each word in the vocabulary derived from the corpus, which is formed from the entire set of news articles being indexed. The frequency of each word in the document is stored in the inverted index. These frequencies are calculated during the feature extraction step of the indexer.

Our retrieval goal is to find the article in the index that exactly discusses the corresponding event. Finding a precise matching between the article and the event can be more easily accomplished if both contexts have comparable sizes. Since an event consists of only a few words, only the title and the abstract sections of the news articles are considered and indexed.[1]

2.2 Linker

The goal of the linker is to find newsworthy events in each article and link them to the previously indexed articles. It first identifies sentences that mention newsworthy events, then, for each event matches and retrieves newsworthy articles. Finally, the top ranked article is hyperlinked to the event if it satisfies a certain confidence level criterion.

Selecting the Candidate Sentences We identified three important criteria in selecting the candidate sentences: (i) the sentence must contain a named entity; (ii) the sentence must contain a verb in past tense; and (iii) the verb mentioned in the second criteria must be an *action verb*. Almost all important observed events are regarding one or more important entities that occur as the subject or the object of a sentence. For example, the sentence "A few days ago *Google* announced their acquisition of *Zagat*, the popular publisher of restaurant review guides" contain the named entities *Google* and *Zagat* as part of the events that refer to news in the past. This criterion could be restrictive by ignoring otherwise good candidates such as "The company issued a press letter yesterday regarding the new privacy policies", where e.g. "*The company*" is a co-reference, as it refers to a named entity such as "Google"). Here, we do not employ a co-reference resolution approach due to its complexity and the additional noise it might introduce to the system.

The second criterion is trivial. Since our goal is to hyperlink the events in the current article to previously published content, we ensure that the candidate sentence contains a verb in past tense. Our last criterion stems from the need to eliminate verbs that usually do not specify any event, thus cannot be linked to any previously published content. Examples include *be*, *become*, *seem*, *grow*, etc. We are looking for verbs that describe an action, which are referred to as *action verbs*. Therefore we ensure that all identified past tense verbs are action verbs, otherwise the sentence is eliminated. To filter the sentences based on these criteria we used the *Natural Language Processing Toolkit (NLTK)* [1], a freely available application for research purposes.[2]

Constructing the Query Once the candidate sentences are identified we extract the events from each candidate sentence. An event is contained in a sentence and is determined by a *predicate* of that sentence. We use the term predicate to describe a function over arguments. The function is formed by the verb and its arguments are the noun phrases that are immediately before and after the verb. Thus, it is possible for one sentence to contain more than one event. For instance, the predicate formed from the sentence "Barack Obama announced his candidacy for presidency on Feb. 10." would be "*announced(Barack Obama, his candidacy for presidency)*" as the verb *announce* forms the function, and the immediate noun phrases "*Barack Obama*", and "*his candidacy for presidency*" form the arguments of the predicate. Hence, the event extracted in this example would be "*Barack Obama announced his candidacy for presidency*".

The general pattern being identified are subject-verb-object relationships. This has certain disadvantages, as it assumes that the verbs are normal transitive verbs that take a single direct object. This is not always the case. A verb can be intransitive, i.e. it takes no objects, or it could be transitive but take both a direct and an indirect object, as in the case of *complex transitive* verbs and *ditransitive* verbs *(datives)*. Nevertheless, this leads to an approach that is easily scalable to other languages.

We use NLTK's built-in noun phrase chunker to automatically identify the noun phrases in the sentence, and the verb is identified through the part of speech tags as mentioned previously. Once the event is identified we again use NLTK to remove the stop words and stem each word in the event, including the verb. The resulting phrase forms our query.

Ranking the Results Once the query is formed the matching articles in the index are retrieved and ranked. The inverted index keeps track of each word and the document it appears in together with its frequency. We form vectors of term frequencies, $\vec{q}$ and $\vec{d_i}$, for each query q and document d_i in the corpus, respectively. The dot product of the query vector with a document vector gives us the importance score of that document for the query. We normalize the dot product with the length of the document. Thus, the score of document d_i for the query q is $\vec{q} \cdot \vec{d_i}/|\vec{d_i}|$. The documents are ranked according to this score.

When constructing the vectors for the documents, we give more weight to the frequency of a word appearing in the title of the document. The reason is the more query terms we find in the title, the more confident we are of that document describing the event. For example, consider the event "A magnitude-8.8 earthquake hit Chile", and two matching documents with titles "8.8-Magnitude Quake Hits Chile", and "Millions are Displaced After the Chile Quake". Even though both documents are related to the event, the former matching document is devoted to the event, whereas the latter has only tangential relevance. We set the weight of matching terms in the title to 3, by tuning this parameter on a separate validation set. Finally, if the score of the top re-

[1]The abstract section corresponds to the first paragraph of a news article.

[2]www.nltk.org

sult retrieved from the index is above a predefined threshold, the event is linked to the article in the index. This threshold parameter can be set depending on the application. For example, in a setting where precision is more important than recall one can set the threshold to a high value to make the system very precise while trading-off recall and vice-versa.

3. ASSESSING THE LINKS

In this section we compare the quality of system-embedded against manually-curated links by professional editors.

3.1 Dataset

We used 200 news articles taken from the top-50 most viewed articles from Yahoo! news, on four different dates. We kept our selection random, covering a diversity of topics and a range of document lengths (150-2000 words). The news articles were separately annotated by LEPA and a team of professional editors from Yahoo! news. We repeated this process for our system using four precision settings.

The professional editors, who had no prior knowledge that their work would be evaluated against the proposed system, were asked to read the articles and identify events and entities that were good candidate links. The guidelines indicated this as a routine editing task and instructed them to link articles that were perceived as related and newsworthy and that would provide interesting insights with respect to the main article. The only limitation was that the linked articles had to reside within Yahoo! news. The editors were allowed to embed as many links as they thought appropriate.

Out of the 200 articles we retained 75 after filtering out the articles for which the system did not detect any events. Our system identified a total of 192 links, while the editors identified 211 links. From the latter we excluded 28 links that were embedded in common by LEPA and the editors, as our focus was to compare the quality of system-embedded against the manually-curated links. As common links we treated those cases of anchored text that appeared in the same article, same paragraph/sentence, and shared at least one common word. This resulted in 164 system-embedded and 183 manually-curated links. From the latter we retained a random selection of 164, to have an equal contribution of both types of link.

We examined the performance of our automated approach using standard metrics, quantitative judgments from human evaluators, and compared the system-embedded links against manually-curated ones. We used the Amazon Mechanical Turk crowd-sourcing service.

3.1.1 Design

This study used a between-groups design ("Group A", "Group B") with three independent variables: type of link (two levels: "system-embedded", "manually-curated"), precision configuration (four levels: "0.0", "0.1", "0.2", "0.3") and date of publication (four levels: "19/10/2011", "20/10/2011", "16/11/2011", "17/11/2011"). The type of link was controlled by introducing either system-embedded (*Group A*) or manually-curated (*Group B*) links. The precision configuration was controlled by adjusting accordingly a threshold value in LEPA. The results were filtered based on this confidence level criterion and those that did not make the cut-off were dropped. This allowed some control over the system's levels of precision & recall, producing results that varied between high precision-low recall and low precision-high recall. The date of publication was controlled by building our experimental dataset with news articles crawled on four different dates, thus reducing the dependency of our findings on the temporal factor.

Each participant took part in one condition (one article-link combination) and assessed four different aspects of the system: (i) the main article, (ii) the associated article, (iii) the link, and (iv) the system's performance. With respect to the first category, the dependent variables were: (i) interest, (ii) newsworthiness, and (iii) similarity to other news read online. In terms of the second category, the dependent variables were: (i) type of relation with the main article (five levels: "related to the main topic of the article", "related to a subtopic of the article", "tangentially related", "unrelated", "other"), (ii) newsworthiness, and (iii) interesting insights with respect to the main article. Regarding the third category, the dependent variables were: (i) suitability of the anchored text[3], and (ii) relatedness with the associated article. Finally, the system performance was measured using the standard metrics of precision, recall, and f-measure.

3.1.2 Tasks

We prepared 328 tasks, each a unique combination of article-link, using the 75 news articles and corresponding $164 + 164$ links. Each participant was assigned a single, randomly selected article. While reading the article the participants were instructed to click on the link that appeared in the text and go through the associated article that it pointed to. To mitigate any unwanted effects stemming from the visual saliency of non-relevant elements of the original content, the articles were presented in the *simple html* format. To reduce the subjectivity of individual responses, each task was performed by two different participants. Upon completing the task, the participants were redirected to an online questionnaire.

3.1.3 Questionnaire

A post-task questionnaire was used to elicit information on several aspects of the task such as the main and associated articles, the quality of the embedded links, the participants' reading experience. A demographics section gathered background information and inquired about previous experience with online news reading. All questions were forced-choice type using a 5-point Likert scale. Questions asking for user rating on a unipolar dimension have the positive concept corresponding to the value of five and the negative concept corresponding to the value of one.

3.1.4 Procedure

We designed the task and the questionnaire in such way that completing them accurately and in good faith required approximately the same amount of effort than random or malicious completion. To ensure that the tasks were performed by human participants we employed keyword tagging with respect to the theme of the main and associated articles. In addition, we recorded the times required to complete each step of the task. This allowed us to distinguish automated responders from human participants. A final check was to accept as participants only workers who had gained a high reputation from other requestors, by having at least 90% of their responses to previous tasks accepted, as well as

[3]As *anchored text* we refer to the underlined text (sentence, phrase, etc.) that appeared in each hyperlink.

Table 1: Observed frequencies across the five categories of associated articles.

	System Links (Group A)			Editor Links (Group B)		
	P_A	P_B	$P_A \cup P_B$	P_A	P_B	$P_A \cup P_B$
Related to main theme	80	69	**149**	89	84	**173**
Related to subtopic	34	39	**73**	51	56	**107**
Tangentially related	21	25	**46**	15	19	**34**
Unrelated	25	27	**52**	8	2	**10**
Other	4	4	**8**	1	3	**4**

a number of completed HIT's (Human Intelligence Tasks) greater than, or equal to, 50. The participants were asked to complete the task, including the questionnaire, in a single sitting. They were also informed of their option to opt out from the task at any point without being compensated. The average duration of the reading task was approximated to 10 minutes and the payment for participation was \$0.66.

3.1.5 Participants

Six hundred and sixty-four participants were recruited through MTurk, from which we reached the expected number of approved assignments from 656 different participants, who spent an average of 17.68 minutes on each task and provided a total of 195 hours of labor. The participants were randomly distributed into two even groups, "Group A" (female 45.12%, male 54.87%) and "Group B" (female 38.69%, male 61.3%), and were of mixed ethnicity and educational background. The Mann-Whitney test did not indicate any statistically significant difference between the two groups in terms of age, gender, educational level, proficiency with english, or frequency of reading news.

3.1.6 Results

We evaluate the performance of LEPA against professional editors on the selected 75 news articles, with a total of 328 links. The performance was measured using precision, recall and f-measure, and as ground-truth we used the participants' assessments. Precision was computed as the fraction of links in each article that received, in terms of relatedness, a score equal to, or greater than, 3 on a 5-point Likert scale. The Mann-Whitney test, the Chi-Squared 'Goodness of Fit', and the Chi-Squared Test of Association were used to establish the statistical significance ($p < .05$) of the differences observed in the experimental results as well as isolate the significant pair(s) through pair-wise comparisons. The Mann-Whitney test is a non-parametric test used for testing differences between groups, when there are two conditions and different participants have been used in each condition. The Chi-Squared 'Goodness of Fit' and the Chi-Squared Test of Association are tests for examining the association of different variables when dealing with data frequencies (nominal data). To take an appropriate control of Type I errors we applied a Bonferroni correction, and so all effects are reported at a .005 level of significance.

To evaluate the main article we asked our participants to provide scores for the following questions: (i) "Did you find the article informative?", (ii) "How interesting did you find the article you read?", and (iii) "Was the article you read similar to other news you usually read?". In this evaluation the results are presented across all groups, since we were interested in examining the overall effect of our experimental manipulation on our sample, instead of narrowing it down to specific subgroups. For the first question, the participants reported the main article as somewhat-to-very informative (M=3.5914, SD=0.6164). Regarding the second question the participants felt that the main article was somewhat-to-very interesting (M=3.3978, SD=0.6825), while in the third question they rated it as somewhat similar to other news that they usually read (M=2.8841, SD=0.7832). The results suggest that the news articles we employed was a fair approximation of what users read online. Moreover, the scores assigned to *interest* and *informativeness* indicate that these variables did not suffer from any adverse effects introduced by the manipulation of the independent variables.

Table 1 presents the frequency scores for all five types of links, in relation to the question: "Please indicate if the associated article is related to the overall theme of the main article, related to a subtopic within the main article, tangentially related or unrelated". Since each link was evaluated by two different participants, we present the scores per group ("System Links", "Editor Links") and per participant ("Participant A", "Participant B"). Columns three and six present the sum of counts for both participants. The Chi-Squared 'Goodness of Fit' test was applied and revealed a statistically significant variation in the observed distribution across all types: (1) $\chi^2(4, N{=}164) = 99.354$, $p < .0001$, (2) $\chi^2(4, N{=}164) = 69.293$, $p < .0001$, (3) $\chi^2(4, N{=}164) = 165.634$, $p < .0001$, (4) $\chi^2(4, N{=}164) = 158.134$, $p < .0001$. We, therefore, reject the null hypothesis that the counts are uniformly distributed across the categories.

We also applied the Chi-Squared Test of Association to examine if there is an association between the participant's group and the type of link. Participants from *Group B* were significantly more likely to find the associated articles related to the main theme (52.7%) or a subtopic (32.6%) of the main article, compared to participants from *Group A* (main theme: 45,4%, subtopic: 22.3%). However, the participants from *Group A* were more likely to perceive the articles as tangentially related (14.0%), compared to participants from *Group B* (10.4%). The system's performance was found to be comparable to that of the editors', with only 15.85% of the embedded links having been reported as unrelated. This is a very encouraging finding, suggesting that our system is scalable and efficient in curating the embedded links.

Table 2 shows the means and standard deviations for participants' assessments of the system-embedded and manually-curated links. LEPA's performance is presented across all four precision settings in rows 1 to 4. Four aspects of the links are examined here, namely: (i) if the anchored text was a good location for the link in the article, (ii) if the anchored text was related to the associated article (the one that the link points to), (iii) if the associated article was newsworthy, and (iv) if the associated article provided interesting insight with respect to the main article. As indicated in Table 2 both types of links received average-to-good scores, with variations being more evident in terms of location and relatedness. The editors performed better, especially compared to the versions of the system with lower precision. The Mann-Whitney independent groups test also supports this finding for the differences observed between System@0.0 and the editors. A direct comparison between the remaining systems (System@0.1, System@0.2, System@0.3) and the editors was not possible, since the former incorporated only a

Table 2: Descriptive statistics for editors and across all system configurations.

	Location		Relatedness		Newsworthiness		Interest. Insight	
	M	SD	M	SD	M	SD	M	SD
System@0.0	2.9909	0.8713	3.0427	0.9898	3.314	0.7777	2.747	1.019
System@0.1	2.9692	0.891	3.0548	1.0054	3.3116	0.8012	2.7603	1.0574
System@0.2	3.1012	0.9011	3.2083	0.9946	3.3571	0.805	2.8631	1.0959
System@0.3	**3.1857**	0.9322	**3.4571**	0.9654	**3.4714**	0.7947	**2.9286**	1.119
Editors	3.5792	0.7048	3.8048	0.7232	3.5274	0.7011	3.1859	0.7971

Bold: Highest performance observed across different different system configurations

Table 3: Performance of system across all precision configurations using standard metrics.

	Mean-Average Precision	Average Recall	Average F-measure
System@0.0	0.5468	**1**	**0.306**
System@0.1	0.5562	0.9467	0.3041
System@0.2	0.5892	0.4719	0.2612
System@0.3	**0.692**	0.2303	0.2456

Bold: Highest scores.

subset of the links embedded by System@0.0; thus containing an uneven number of links, compared to the number of links that the editor curated.

The Mann-Whitney test also revealed that the manually-curated links received statistically significant higher scores than the system links, in terms of location (U=8213.5, p=.000, r=-0.35), relatedness (U=7449, p=.000, r=-0.39), newsworthiness (U=11498.5, p=0.02, r=-0.12), and interesting insights (U=9998, p=.000, r=-0.22). However, in all cases the observed differences represent a small effect that accounts for less than 10% of the total variance in our sample. Furthermore, as we increased the precision threshold LEPA's performance drew nearer to that of the editors.

Looking at the performance of the system in terms of precision, recall, and f-measure in Table 3, we notice that the mean-average precision escalates as the threshold value is increased, although this increase has an adverse effect on recall. Apparently there is a trade-off between introducing fewer but more related links and receiving a larger number of links of heterogenous relevance. In terms of performance column three shows the f-measure scores, which indicate System@0.0 as the optimum approach.

4. CONCLUSIONS

We presented the evaluation of LEPA, a fully automated approach to detecting events in news articles and linking them to relevant past articles. One of the main advantages of LEPA over other systems is that it works on sentences as well as events within a sentence in isolation from the main content, whereas other approaches often consider the entire content when generating its related content pool. Moreover, unlike other systems that focus on link detection and disambiguation on Wikipedia articles or concepts, LEPA is less domain-dependent.

We conducted an experimentation via MTurk to evaluate our system-embedded links against links manually-generated by professional editors. When we treat these manual links as a gold standard the results indicate that LEPA is comparable to that across several facets of the news reading experience: relatedness of the anchored text with the associated article, newsworthiness, offering interesting insights, etc. Our evaluation reveals that the editors had an average-to-good performance, whereas our system had an, overall, average performance. When examining the system performance using standard metrics we observe that the f-measure scores decline over high precision values. Apparently, assigning a high value to the precision threshold acts as a trade off for recall and vice-versa. However, the analysis of the open-ended responses indicates that, for our setting (online news reading), a high precision configuration facilitates a better news reading experiences, contrary to the lower thresholds that result in a larger number of links and provide access to a plethora of information. In other words, *less is more*.

The manual creation of hyperlinks is a time-demanding and challenging task, especially for large online providers where the editors are expected to process daily a significant number of news articles. Consequently, any effort towards automating the link curation process could go a long way to improve their efficiency and performance. In a real use-case scenario the final decision might ultimately for the editors to make but our experimental findings indicate that the proposed system, with its massive scalability being its greatest asset, can support the process of identifying interesting links and fulfil its purpose in reducing manual effort. Contrary to other automated systems, LEPA does not require any training data or human intervention, nor is it limited to a specific domain, making it a generalizable and attractive solution for many real-world applications.

5. REFERENCES

[1] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[2] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. *TPDL*, 2011.

[3] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. *EMNLP*, 2007.

[4] J.J. Gardner, and L. Xiong Automatic link detection: a sequence labelling approach. *CIKM*, 2009.

[5] He, J., and de Rijke, M. A ranking approach to target detection for automatic link generation. *SIGIR*, 2010.

[6] J. He, M. de Rijke, M. Sevenster, R.C. van Ommering, and Y. Qian. Generating links to background knowledge: a case study using narrative radiology reports. *CIKM*, 2011.

[7] V. Jijkoun, M.A. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. *AND*, 2008.

[8] Knoth, P., Zilka, L., and Zdrahal, Z. Using explicit semantic analysis for cross-lingual link discovery. *IJC-NLP Workshop on Cross Lingual Information Access*, 2011.

[9] Mihalcea, R., and Csomai, A. Wikify!: Linking documents to encyclopeadic knowledge. *CIKM*, 2007.

[10] Milne, D. N., and Witten, I. H. Learning to link with wikipedia. *CIKM*, 2008.