A Learning Based Model for Headline Extraction of News Articles to Find Explanatory Sentences for Events

Sandip Debnath [1,3]
Department of Computer Sciences and Engineering [1]
Penn State University
University Park, PA 16802, USA
debnath@acm.org

C. Lee Giles [1,2,3]
School of Information Sciences and Technology[2]
eBusiness Research Center [3]
Penn State University
University Park, PA 16802, USA
giles@ist.psu.edu

ABSTRACT

Metadata information plays a crucial role in augmenting document organising efficiency and archivability. News metadata includes *DateLine*, *ByLine*, *HeadLine* and many others[1]. We found that *HeadLine* information is useful for guessing the theme of the news article. Particularly for financial news articles, we found that *HeadLine* can thus be specially helpful to locate explanatory sentences for any major events such as significant changes in stock prices. In this paper we explore a support vector based learning approach to automatically extract the *HeadLine* metadata. We find that the classification accuracy of finding the *HeadLine*s improves if *DateLine*s are identified first. We then used the extracted *HeadLine*s to initiate a pattern matching of keywords to find the sentences responsible for story theme. Using this theme and a simple language model it is possible to locate any explanatory sentences for any significant price change.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.7 [**Document and Text Processing**]: Electronic Publishing; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords

Metadata Extraction, News Metadata, Explanatory Sentence

General Terms

Algorithms, Experimentation

1. HEADLINE EXTRACTION

In our earlier paper [2], we described in detail, how to use a support vector based learning approach to accurately identify the *DateLine*s of news articles. In short we first devised our own temporal grammar by which we locate all the temporal expressions in an article. Next we create a vector of properties for each of these expressions to train a support vector classifier. Using this classifier we can achieve high accuracy in finding the *Base Time-line* (*BT*) or *DateLine*. Here we show that using the same approach, we can reach higher accuracy of finding the *HeadLine* it we do it step-by-step. It means that the accuracy of finding the *HeadLine* increases if we first identify the *DateLine* location and use this as part of the feature vector to train the *HeadLine* classifier than train the classifier without the *DateLine* information. The classification accuracy of finding the *DateLine* is shown in [2] and so is not shown here. The classification accuracy of *HeadLine* is shown in table 1.

2. FINDING STORY THEME

To the best of our knowledge, finding the theme of a news article based on its *HeadLine* has not been explored yet. News articles consist of one or more paragraphs devoted to the main story which the *HeadLine* refers. After extracting the *HeadLine*, we used Brill's POS tagger [1] to tag each word. Prepositions, conjunctions, interjections, and other low entropy words such as “a”, “and”, “the” etc. are removed from the word-set of *HeadLine*. Next we use Krovetz [3] stemming to get the roots of these filtered words. We then create a language model M_h for the *HeadLine*. It represents a discrete distribution over the words in the vocabulary. We assume that the word distribution inside paragraph P_i in a story is independent of of the distributions of other words. Assuming this and assuming that the distribution of each paragraph generation is also independent, we analyse the paragraphs in the same way as the *HeadLine*. Now we can estimate the maximum likelihood of a paragraph P_i so that it would be generated by the model M_h (for the *HeadLine* h). The sentences of this “theme” paragraph are taken as the theme of the story and to find explanatory sentences. Due to the space constraints, we can not show the details of the result,

[1] Please see NewsML 1.2 DTD specification for more details

K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada.
ACM 1-59593-163-5/05/0010 .

Site	**Number of articles**	Accuracy (w/o) of HeadLine	**Accuracy (w) of HeadLine**
AP	37	82.35	94.11
Briefing.com	132	87.12	90.9
BusinessWeek	128	88.28	93.75
Business Wire	430	84.65	92.3
CBS MarketWatch	606	90.26	94.88
CCBN	92	80.43	85.86
Dow Jones	324	91.97	95.67
EDGAR Online	599	90.65	96.82
Forbes	451	88.91	95.34
Market Wire	158	89.87	94.93
Morningstar	14	71.42	71.42
Motley Fool	162	82.71	89.51
NewsFactor	71	91.5	95.77
PR Newswire	525	94.66	96.19
PrimeZone	64	85.93	90.62
Reuters	641	93.6	96.72
SmartMoney	63	87.3	92.06
StarMine	38	89.47	84.21
TheStreet.com	351	91.16	92.59

Table 1: The columns represent Source Web-site, Total Number of Articles, Accuracy of HeadLine without the use of DateLine information and Accuracy of HeadLine with the use of DateLine information as an extra parameter.

however we can just mention that for a limited number of news stories we achieved over 97% of paragraph-based F-Measure.

3. EXPLANATORY SENTENCES

Explaining a stock price change for a company is a complex and hard problem. To the best of our knowledge, it has not received enough attention. Our work [4] on finding "keyword-based" explanation uses word entropy measures for news articles during the price change (W_A) compared to the past or historical set (W_B) of news articles as shown in figure 1. For artificial market events, it worked pretty well.

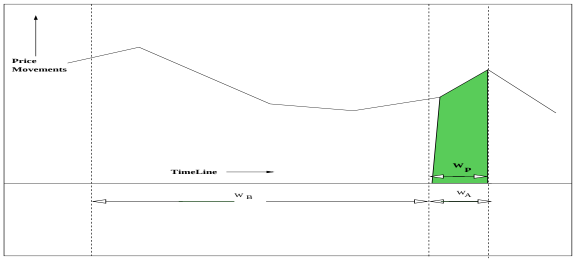

Figure 1: A sample price change curve. W_A and W_B are the time-line windows for considering the document collections to compare (in our previous work[4]) *for possible explanatory keywords*. **Here we consider W_P as the time window to find the** *possible explanatory sentence(s)*.

Company	**Date**	**Sentence**
MSFT	July 1, 2005	IBM Corp. will get $775 million in cash and $75 million worth of software from Microsoft Corp. to settle claims ...
C	July 11, 2005	Merrill Lynch cut its second-quarter earnings estimate for Citigroup, Inc. to $1.04 from $1.09 a share and slashed its fiscal 2005 earnings estimate...

Table 2: Explanatory sentences for companies for a specified date

We wanted to extend the same idea to generate explanatory sentences for any significant price change. However the entropy based model does not work well for sentence generation. We slightly borrowed the concept from Ponte et. al [5] for the use of language models representing the trends. We used the models M_{pt} of price trends for sentences in the "theme" paragraphs of the articles. These articles are selected for the particular price change according to temporal relationship (W_P). That means that the sentence in the following equation is taken as the explanatory sentence:

$$S_{expl} = argmax_{j \in \sum_i^Q |S_i|} Prob(S_j | M_{pt}, \sum_q^Q P^q_{winner})$$

where there are total Q stories selected for the price change, and each has one "theme" paragraph and the total number of sentences in all these paragraphs is $\sum_i^Q |S_i|$. Due to space constraints, we can not show full results; table 2 shows results for two recent price changes of two companies.

4. CONCLUSION

We devised a novel approach to extract news *HeadLine*s using SVM and using them to find story themes to get a sentence based explanation for a stock price change.

5. REFERENCES

[1] E. Brill. A simple rule-based part of speech tagger. In *proceedings of ANLP*, pages 152–155, 1992.

[2] S. Debnath, P. Mitra, and C. L. Giles. Finding base time-line of a news article. In *proceedings of FLAIRS*, pages 142–147, 2005.

[3] R. Krovetz. Viewing morphology as an inference process. In *Artif. Intell.*, volume 118(1-2), pages 277–294, 2000.

[4] D. M. Pennock, S. Debnath, E. J. Glover, and C. L. Giles. Modelling information incorporation in markets, with application to detecting and explaining events. In *proceedings of UAI*, pages 405–413, 2002.

[5] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.