Extracting Information from CNN Financial News

Luis Pérez-Breva, Giorgos Zacharia & Osamu Yoshimi

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

Objective: Define a representation and a learning algorithm to learn from financial news.

Motivation: Extracting automatically relevant information from wire news is of increasing importance in a variety of areas. We are developing state of the art learning techniques to extract information from on-line CNN financial news.

Previous Work: Difficulties in extracting quantitative knowledge from text have traditionally prevented that field from evolving beyond particular applications. McCallum [2] introduces a set of tools based on bag of words and information gain to extract symbolic information from text documents. Freitag et al [1] present a framework that combines bag of words with web structure information for classification purposes. Despite the significant experimental results they obtain, their approach is restricted to a very specific type of web pages.
Attempts to provide a link between Machine Learning and linguistics to take advantage of semantic and syntactic information have been made (for example [3]) although no direct application has been implemented yet.

Approach: In our initial approach we plan to use techniques such as "bag of words" [2], and benchmark it with focus on different parts of the information (Title, body, HTML tags). Revision of String Kernel approaches and application of alternate hierarchical structures is a second step we are planning to undertake.

The ultimate goal is to quantify information in financial text news, possibly incorporating knowledge from the field of linguistics, for feature extraction and classification purposes.

Difficulty: Despite the large amount of news available, processing of text corpora with machine learning is still an open field. Extracting meaningful features from text corpora, summarizing semantic and syntactic information, yet reducing redundancy and ambiguity is the key item to be addressed by most of the applications in text recognition.

Impact: A framework able to quantify text news would dramatically decrease complexity of any classification task requiring up to date information.

Data:

- Collection of articles from the CNN web site (1996 → 2001).
- Reuters Corpus (August 1996 → August 1997).

Preliminary results: We have carried out preliminary experiments to benchmark bag of words as input to Naive Bayes and Maximum entropy classifiers. The goal of these experiments was determining how far should a classifier should read in an article before fully determining its class. For the task of determining whether news from Reuters are financial or not, table 3 shows sample results using naive Bayes and Maximum Entropy classifiers over a "bag of words" input.

These results show that it is not necessary to read the entire article, and confirm the hypothesis that (within Reuters corpus) the first few paragraphs contain sufficient information to perform high level classification. Furthermore, a decrease in accuracy of less powerful classifiers (naive bayes) might occur when reading the entire article. It is also interesting to note that the size of the vocabulary is not critical to increase accuracy.

Naive Bayes										
	words in vocabulary									
Paragraphs read	100	200	500	1000	2000	5000	10000	20000	50000	all
headline only	83.66	85.87	87.88	89.01	90.14	90.86	91.34	91.44	-	91.3
1	85.03	87.08	89.48	91.1	91.66	92.42	92.49	92.66	-	92.7
2	85.2	87.65	89.74	90.9	91.77	92.08	92.09	92.33	92.39	92.47
3	86.02	87.98	90	91.16	91.49	91.9	92.05	92.08	92.23	92.32
4	86.21	88.12	90.15	91.03	91.54	91.64	91.76	92	92.1	92.13
5	85.96	88.15	90.29	91.02	91.22	91.7	91.74	91.86	91.93	91.8
7	86.06	88.49	89.96	90.68	90.75	91.04	91.36	91.5	91.53	91.68
all	85.77	88.54	89.33	89.09	89.41	89.48	89.62	89.78	89.74	90.01

Maximum Entropy										
	words in vocabulary									
Paragraphs read	100	200	500	1000	2000	5000	10000	20000	50000	all
headline only	57.02	64.03	76.22	83.39	88.64	90.95	91.25	91.32	-	91.32
1	70.43	80.07	89.16	92.01	93.05	93.58	93.76	93.79	-	93.96
2	77.56	84.32	91.12	92.76	93.48	93.85	94.01	94.12	94.18	94.38
3	79.56	86.87	91.69	93.22	93.79	94.23	94.3	94.41	94.48	94.46
4	81.18	88.22	91.94	93.13	93.92	94.21	94.37	94.5	94.63	94.57
5	80.51	88.81	92.04	93.29	93.95	94.21	94.25	94.47	94.51	94.63
7	82.62	89.39	91.9	93.17	93.83	94.16	94.36	94.46	94.54	94.66
all	84.14	89.62	91.9	92.85	93.24	93.76	93.95	94.11	94.3	94.37

Table 3: Classify financial vs non-financial news from Reuters. Naive Bayes and Maximum Entropy accuracy using Bag of words as an input. Number of sentences read and number of words in the vocabulary (sorted by information gain [1]) are analyzed.

Future Work: Those results encourage further research in feature extraction from text documents to complement "bag of words" with document structure and perhaps syntactic information.

Research Support: Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] Mark Craven, Dan DiPasquo, Dayne Fre-itag, Andrew McCallum, Tom Mitchell, Kamal Nigam, , and Sean Slattery. Learning to extract symbolic knowledge from the world-wide web. In *In Proceedings of the AAAI Fifteenth National Conference on Artificial Intelligence*, 1998.

[2] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[3] Dan Roth. Learning to resolve natural language ambiguities: a unified approach. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 806–813, Madison, US, 1998. AAAI Press, Menlo Park, US.