World Knowledge in Broad-Coverage Information Filtering

Bennett A. Hagedorn
Indiana University
bhagedor@indiana.edu

Massimiliano Ciaramita
Yahoo Research Barcelona
massi@yahoo-inc.com

Jordi Atserias
Yahoo Research Barcelona
jordi@yahoo-inc.com

Categories and Subject Descriptors: H.3.3 Information filtering: Information Search and Retrieval

General Terms: Experimentation.

Keywords: Financial news filtering, world knowledge, NLP, machine learning.

1. INTRODUCTION

Document retrieval is a well-understood problem, as a consequence search technology has been able to support growth and innovation in scientific and industrial domains. As the Web evolves new types of content emerge: blogs and other types of communities, often based on multimedia content sharing, feeds for information browsing and delivery, vertical domains for shopping and performing other transactions, Web advertising. New types of content are challenging for traditional IR approaches based on the idea that the unit of information is a full "document", whose content can be reasonably approximated with a bag-of-words representation, and whose ranking or matching can be assessed in isolation, within the vector space model paradigm.

News filtering tasks involve monitoring a stream of news to identify stories belonging to a predefined set of categories. In the context of *opinion mining* we define five categories which capture the "polarity" of the story, a graded positive/negative opinion, with respect to a company. Financial news services are an important component of major Internet content providers (e.g., Google Finance, Yahoo! Finance). Financial news and stock prices tend to be correlated, and opinions and trends of financial stories can be modeled to a certain extent [2, 3, 7]. Tools for automated monitoring of such news could be valuable to users, financial analysts or small investors. Previous work has focused on clues for polarity such as selected keywords (plunge, surge, etc.), instead we consider the task of classifying all news stories that are relevant to a set of companies, coming through a stream.

We present an exploratory study on the problem of classifying financial news stories streamed through RSS feeds. In particular, we focus on news stories *titles*. The reason for this is threefold. First, processing titles is faster than full documents and allows monitoring of larger numbers of sources efficiently. Secondly, and more interestingly, humans seems to be capable of performing such a task effortlessly on the basis of the little information provided by the title, even with no specific domain expertise. Furthermore, processing titles, e.g., short sentences, is a kind of short text processing and retrieval task, an area of research which is becoming increasingly relevant [4]. We implement a system for classifying news titles based on traditional document representation concepts and machine learning. We analyze the output of the system to identify the causes of its limitations and discuss desirable properties of more advanced systems. We conclude that to improve accuracy in broad-coverage settings a substantial amount of world knowledge is necessary.

2. STUDY

Over a month period, October-November 2006, we monitored the RSS feeds from Yahoo! Finance (36 sources) relative to the top 50 company symbols in the Standard & Poors index. The first author, a senior student in economics, annotated the titles of 7,382 stories using the following five categories, the same as [7]: "bad" (B), "almost bad" (b), "uncertain" (U), "almost good" (g), and "good" (G). Our data collection method differs from [7] in that we do not automatically assign to category "uncertain" all news which do not mention explicitly the company name. Also, we do not limit our attention to news which contain predefined trigger words, but consider all news stories in the stream. The motivation for this is valuable recall: the most frequent trigger words are relatively infrequent; e.g., "gain" occurs in 2.2% of the titles, "drop" in the 0.9%, "growth" and "gain" in the 0.8%, "surge" in the 0.3%, etc. The following table illustrates a few example titles:

C	Title	Co.
B	"AIG units subpoenaed by DOJ and SEC"	AIG
b	"Chinese SUV maker aims to prove itself"	F
U	"Indonesia seeking $ 12 Billion in capital"	IBM
g	"P&G sees better operating environment"	PG
G	"Valero energy 3Q profit surges"	VLO

As an illustration, the first title mentions a legal action against the company (B), while the second (b), although not explicitly negative, implies the possibility of more competition on a strategic market for an auto manufacturer, etc.

We partition the stories in train, development and test sets. The development set contains 1,050 titles collected in October 2006, the training set consists of 4,513 titles from the first 14 days of November 2006, and the test portion contains 1,819 news from November 15 (752 titles), 16 (811) and 17 (256). Splitting by day is necessary since the same news can appear several time in one day for different companies, or for the same company from different sources. As a classifier we used a regularized multiclass perceptron of our

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

implementation based on the algorithm introduced in [1], we set the free parameters of the algorithm on the development set. We tried three different types of binary feature encodings: bag of unigrams (Uni), bag of unigrams and bigrams (+Big), the latter plus a feature for the company stock symbol (+Co.). Considering all 7,382 titles, 3,589 of them fall into the category "uncertain". Thus the simplest baselines which chooses the majority vote category would be correct roughly 48.6% of the time. The following table summarizes the results of our experiments:

Uni	+Big	+Co.	P-2	P-3	+15Nov	+16Nov
54.6	56.9	58.4	58.2	56.5	59.7	60.5

The Uni model has an accuracy of 54.6%, the best model is obtained adding bigrams and the company symbol (58.4%). Richer feature representations in the dual space with polynomial kernel functions of degree two (58.2%) and three (56.5%), were not useful. To estimate the impact of more training data we added to the linear model one day of stories (November 15, 752 titles), and two days of stories (15 and 16, 1566 titles) and testing on the remaining day(s). This amounts to adding respectively 16% and 35% more training data, but results improved only slightly (59.7% and 60.5%).

3. DISCUSSION

While the model outperforms the baseline, its accuracy is poor. By comparison, the same classifier with similar features achieves accuracies above 91% in classifying a TREC question classification data-set in 6 categories. Possibly, with higher quality data, e.g., larger in size and produced by several annotators, results might improve. However, it is known that classifying documents by opinion is a harder problem than classifying by topic, typically because the unstructured bag of words representation is not expressive enough a task which can involve sophisticated inferences [6].

To identify peculiar features of this task we analyzed the errors made by the system in the last experiment, comprising 256 news from November 17. The systems makes 101 errors which we inspected and classified according to the nature of the story with respect to the company. We propose that there are six main patterns which emerge from the mistakes, listed below:

PROD	INT	IND	COMP	ECO	PROF	?
32	26	17	10	7	5	4

The most frequent mistakes concern news about products or properties of a company ("PROD"); e.g.,"Aeromexico chooses GEnx engines" (GE), to classify this correctly it might be necessary to know that GEnx engines are products of GE and that "being chosen" is a positive event. The category "INT" concerns issues internal to the company about lawsuits, scandals or company's components; e.g., "DELL wrestles with its accounting", here some knowledge of a company's structure and internal dynamics could help. "IND" concerns remarks about the company's industry sector; e.g., "Energy sector shrugs off crude weakness". "COMP" news involve the company's competitors; e.g., "HP poised to unseat IBM" (DELL), "Long lines greet PlayStation 3 debut" (MSFT); these are typically very polarized stories. Some misclassified stories concern the general economy news ("ECO"); e.g., "Yen off low after Japan data". A few mistakes ("PROF") concern explicit mentions of profit or loss for the company ("First Solar rises after IPO")

This analysis suggests that this type of task, in a broad-coverage setting, involves a significant amount of, partly domain-independent, world-knowledge which needs to be available to the system. Companies are not independent from each other: complex dependency structures are defined by relations such as competitor, allied, customer, sellers. Probably a structured learning approach would be beneficial. More NLP, e.g., syntactic analysis and detailed tree comparison [5], might be needed to capture the role of entities involved in a story; e.g., the same story "A sues B over X" can be good news for A (the subject) and bad for B (the object). We noticed also that several stories repeats in more or less the same form for different companies or from different sources or both. The classification of such stories might be best modeled as an ensemble and not in isolation since their classification mutually affect each other. More world knowledge needs to be used, to improve the representation and uncover structure in the data. Such knowledge should be automatically generated in order to have enough coverage and be up-to-date.

There is valuable information to be gained in broad coverage settings. Less than 50% of the stories mention the full name or an abbreviation of the company the story refers to, the rest do not mention the company name or refer to related entities. Stories in the category uncertain have a considerable amount of cases with no explicit mention of the company. However almost 40% of the stories do not mention the company although they express polarized information. Finally, we suggest that the title polarity task, because it relies on rather transparent world knowledge-grounded inferences, could serve as a benchmark to evaluate the capability of systems to integrate and use world-knowledge.

4. REFERENCES

[1] K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, Jan 2003.

[2] S. R. Das and M. Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*, 2001.

[3] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 389–396, 2000.

[4] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, 2007.

[5] A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the 17th European Conference on Machine Learning, ICML 2006*, 2006.

[6] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[7] Y.-W. Seo, J. A. Giampapa, and K. P. Sycara. Text classification for intelligent agent portfolio management. In *International Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2002.