Multi-stage News Classification System for Predicting Stock Price Changes

주식 가격 변동 예측을 위한 다단계 뉴스 분류시스템

Woojin Paik*, Myoung Hyoun Kyung**, Kyung Soo Min**,
Hye Ran Oh**, Chami Lim**, Moon Sun Shin***

ABSTRACT

It has been known that predicting stock price is very difficult due to a large number of known and unknown factors and their interactions, which could influence the stock price. However, we started with a simple assumption that good news about a particular company will likely to influence its stock price to go up and vice versa. This assumption was verified to be correct by manually analyzing how the stock prices change after the relevant news stories were released. This means that we will be able to predict the stock price change to a certain degree if there is a reliable method to classify news stories as either favorable or unfavorable toward the company mentioned in the news. To classify a large number of news stories consistently and rapidly, we developed and evaluated a natural language processing based multi-stage news classification system, which categorizes news stories into either good or bad. The evaluation result was promising as the automatic classification led to better than chance prediction of the stock price change.

초 록

주시가격을 예측하는 것은 주식 가격 변동에 영향을 미치는 많은 요인과 요인 간의 상호작용에 기인하여 매우 어렵다고 알려져 있다. 이 연구는 어떤 회사에 대한 좋은 기사는 그 회사의 주식가격을 오르도록 영향을 미칠 것이고 나쁜 기사는 그 반대의 작용을 할 것이라는 가정에서 시작했다. 여러 회사들에 대한 기사와 그 회사의 주식가격이 기사가 공개된 후에 어떻게 변했는가에 대한 분석을 통하여 위 가정이 맞는 것을 확인했다. 즉 기사의 내용을 기사에 나온 회사에 대하여 호의적인지 아닌지 신뢰성 있게 분류하는 방법이 있다면 어느 정도의 주식 가격 예측은 가능할 것이다. 많은 기사를 일관적으로 빨리 처리하기 위하여 상장회사에 대한 기사를 자동 분석하는 다단계 뉴스 분류시스템을 개발한 후 성능을 확인하여 자동 시스템이 무작위로 주가 변동을 예측했을 경우보다 높은 정확률을 보이는 것을 확인했다.

Keywords : stock price prediction, text classification, natural language processing, news analysis
주식가격 예측, 문서 분류, 자연언어이해, 뉴스기사 분석

---

* Associate Professor, Dept. of Computer Science, Konkuk University (wjpaik@kku.ac.kr)
** Students, Dept. of Computer Science, Konkuk University (kyoungoon@naver.com, seoulob@hanmail.net, mkandhr@hanmail.net, tonyibi@hanmail.net)
*** Lecturer, Dept. of Computer Science, KonKuk University, Corresponding Author (msshin@kku.ac.kr)
■ Received : 25 May 2007
■ Accepted : 23 June 2007

1. Introduction

Predicting stock price change is known to be a very difficult task. Many methods have been suggested but none has shown to be consistent or reliable in predicting stock price except for very short time in the future. However, there are a few known causal precedents, which make sense in terms of predicting the stock price of a particular company. For example, good news about a particular company, such as a large new contract for a shipbuilding company, will most likely to cause the corresponding stock price to go up. Nevertheless, many people believe buying stocks after seeing good news or selling after bad news is not going to work advantageously for them. The main cause for this skepticism is an idea that any news, which is worth enough to affect the stock price, is already accounted for by the institutionalized or professional brokers representing large investors before the news is released to the public.

The main goal of this research is to find whether the stock price changes after the related news story is released. More specifically, we wanted to determine whether the stock price tends to go up after good news and go down after the bad news on the contrary to the general belief.

The confirmation to this goal will be encouraging as it will allow us to predict the stock price changes to a certain degree given that there are preceding news stories about the companies whose stock prices that we wish to predict and also if we can reliably classify the respective news stories as either good or bad regarding the companies mentioned in the stories.

There are more than two thousand publicly traded companies in Korea. These companies are listed either in Korean Composite Stock Price Index (KOSPI) or Korean Securities Dealers Automated Quotations (KOSDAQ). A large number of news stories about these companies are released to the public around the clock. Furthermore, the news stories should be classified very rapidly if the information is to be used by the day trader type investors as they tend to trade stocks in very short time periods.

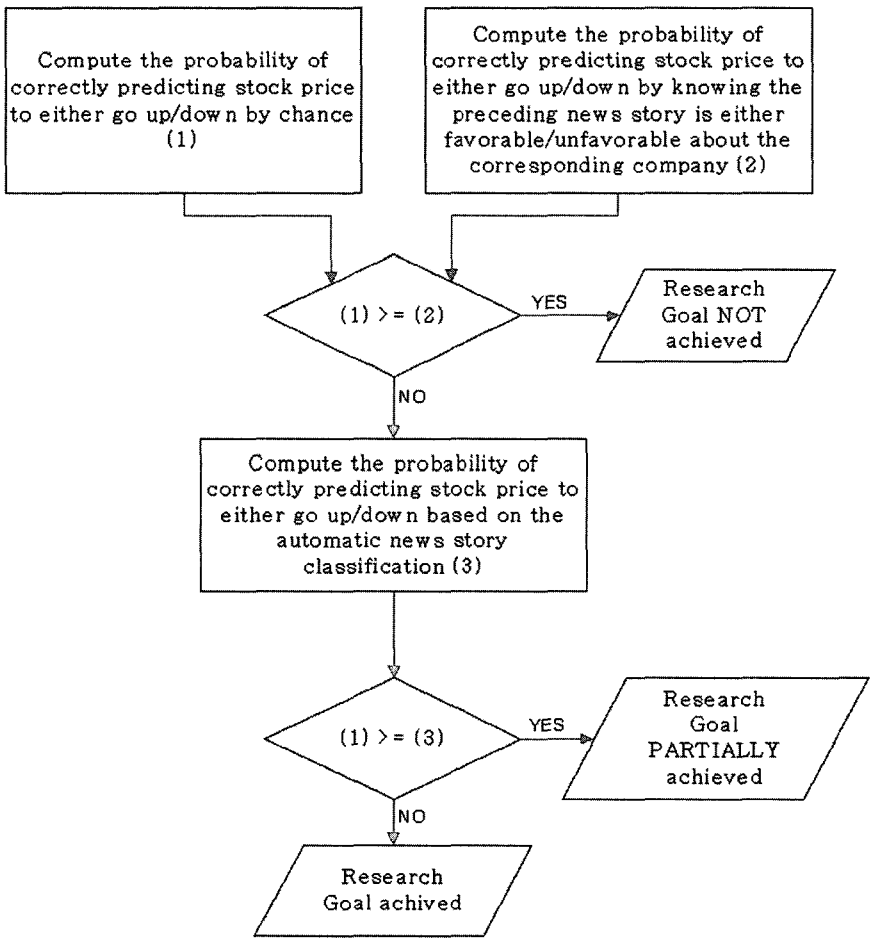
Thus, we developed a natural language processing-based multi-stage news classification system to categorize news stories about the companies. This automatic system was designed to satisfy the processing requirements such as the number of news stories and the classification speed.

Our secondary research goal is to find whether it will be possible to predict the stock price automatically if the accuracy of the automatic news classification is above certain level. This minimum level is the same as the stock price prediction by chance. The maximum possible prediction accuracy of the automatic system will be equal to the correct rate achieved by manually classifying news stories into good and bad categories. Here we assume that careful manual classification will achieve 100% correct categorization rate.

There are many text classification schemes that we can use to classify news stories according to the good and bad categories. Some will be better than the others. However, our research goal is not to compare the effectiveness of various text classification schemes.

Thus, we have chosen to use a simple but easily verifiable classification system where the manually selected patterns, indicating either good or bad news, were used to classify news stories. The patterns were selected from the training data set. The classification system first determines whether an unseen news story mentions at least one KOSPI listed company. The pre-selected patterns were checked to determine whether the story should be classified as good or bad toward the KOSPI listed company mentioned in the news.

The 〈Figure 1〉 shows the overall research process. Basically, three probabilities were



〈Figure 1〉 Research Process

calculated. The probability of correctly predicting the stock price rise or fall randomly was compared with the probability of correctly predicting the stock price rise or fall when we can classify the news stories with the 100% accuracy. If the random based prediction is worse than the probability after knowing the nature of the news story with the 100% accuracy then the random choice based probability is again compared with the probability of correctly predicting the stock price rise or fall based on the automatic classification result, which is less than perfect.

2. Survey of Previous Works

Mittermayer (2004) developed a news categorization and trading system to forecast stock price using text mining techniques. However, the system processed the press releases instead of the general news stories published by news agencies. The press releases are prepared by the companies to disclose information about themselves. The press releases typically include earning figures, acquisition and divestitures of businesses, and appointments, retirement, or dismissal of important company personnels. One of the basic premise of this research is that the positive press releases will cause traders to buy stocks and the bad ones will cause to sell stocks. This is probably true for some cases but many press releases, which are written by the company public relations officers, will not influence the stock price change as the press releases are mainly designed to make most of the company related news look positive. With the exception of the infrequent factual earning related reports, the traders will place less value in the information disclosed in the press releases. This is probably the reason why most of the press release did not affect the stock price changes. Only 5 percent of the press release were the precedents of stocks being sold or bought.

In comparison, our project processed general news stories published by various news agencies to predict the stock price changes. Unlike the press releases, the news agencies try their best to make the story as accurate and objective as possible. We found 76% of the positive news stories preceded the stock price rise for the companies mentioned in the news stories. 62% of the negative news stories preceded the corresponding stock price to fall.

Mittermayer (2004) also used direct approach to train the stock price prediction system. The system used any press releases, which preceded the corresponding stock price of the companies to rise as good news. Similarly, any press releases preceding the stock price fall were considered bad news. All other press releases were considered no movers. However, the good or bad press releases do not have to include any favorable or unfavorable information regarding the companies, who published the press releases. Thus, the good or bad did not refer to the content of the press releases but just how it affected the stock price

changes. This approach has an advantage of easily generating the training and the testing data set as no human judgements will be needed to separate news stories into either good or bad piles. However, the it will not be able to determine why one press release is either good or bad by just reviewing its content alone. This can to lead to less trust in the actual operation of the system.

There is another difference in determining the point-in-time of a stock price change. Mittermayer (2004) considered certain pre-determined level of drop or rise of the stock price within one hour after the press release as the criteria to determine whether the press release is either good, bad, or no mover. However, we used the specific points in time after the news release to determine the stock price change.

Finally, the NewsCATS system (Mittermayer, 2004 and Knolmayer & Mittermayer, 2006) employed an evaluation strategy of stock trade simulation to measure the profit. The result was compared with the a random trader to measure the system's effectiveness. The system showed about 0.9% more profit than the random trading including the transaction costs.

Lee (2003) used logistic regression analysis, Support Vector Machine (SVM), artificial neural network techniques to predict the rise and fall of stock prices by using various economic indicators as the input data. He found that SVM technique was the best in predicting the stock price changes.

Joo (2004) developed a system to generate a linguistic index, which can be a part of the input to the stock price prediction system in conjunction with other variables, by using binary classification of the news articles. His primary goal was to develop a prediction model for stock price change. To develop an experimental system, both similarity measure between the document vectors and the logistic regression method were used to create a binary classifier for the news articles. To develop the system, news articles were preprocessed then the similarity value between two article groups, namely a group of articles, which preceded stock price rise, and the other group, which preceded the stock price fall, were generated. To validate the prediction model, previously obtained similarity values and the stock price change data were analyzed using both single and multiple logistic regression methods. The main research finding was the use of both linguistic index and economic index predicted stock price change better than using each index separately.

The core of our news classification system is categorizing news articles by their respective tone namely good or bad respective to the companies mentioned in the news articles. Similarly, there has been work to categorize Customer Relation Management (CRM) email messages according to the mood of the message creator (Paik et al, 2001). In this work, the email classification system categorized the messages according to the mood of the message senders. The possible mood range was either upset or happy regarding the

services they received or the products that they purchased.

3. News and Stock Price Change Relationship

To test whether the stock price of a particular company indeed goes up after the good news about the company was released and goes down after the bad news release, we used the Korean news stories released in January 2006 as the testing data set.

Many ordinary traders use web-based day trading systems to buy or sell stocks. These trading systems usually provide real-time news feeds from various news agencies for the traders to consider in their buy or sell decisions. Since, it was not possible for us to retrieve news stories stored in these trading systems, we decided to use the easily usable news articles listed in the Naver's stock news section (http://stock.naver.com/news/). Naver is a premier Korean information portal. The news stories, which were made available through Naver, were similar to the ones provided by the trading systems as the news from various agencies are aggregated then provided in the same manner. It is possible for a certain news story to be available through Naver some time after the actual release by the responsible news agency. However, the same type of the temporal delay will likely to occur in the trading system as the same aggregator was used to distributes the news stories from various news agencies.

3.1 Sample Data

In the Naver's stock news section, there are subsections, which are named as 'Good News' and 'Bad News'. Various full-text news articles, which were determined to be either good or bad for the stock market in general by Naver's internal news classification system, are updated in the respective subsections in real time.

Not all news articles in the 'Good News' and 'Bad News' subsections were suitable for our purpose given that our goal was to predict only the stock price change of the companies listed in KOSPI. Based on the manual evaluation, some articles from the 'Good News' section were ignored as they were not about the particular KOSPI companies. These ignored articles were usually about the companies listed in KOSDAQ, which is another Korean stock index, or about foreign stock market trends. Furthermore, some articles were in fact wrongly placed stories namely bad articles about particular KOSPI companies. Others included tabular information, which cannot be judged to be either good or bad. There were 975 articles in the 'Good News' subsection in January 2006. Out of 975 articles, 709 were judged to meet our good criteria. 53 were considered to be bad. The remaining 213 articles were not about the KOSPI listed companies and thus ignored. This shows the correct ratio of Naver's good categorization at 93% by considering only the articles judged to be

either good or bad.

The same analysis of the articles in the Naver's 'Bad News' resulted in 80% correct ratio for Naver's internal classification system. There were 888 articles in the 'Bad News' subsection in January 2006. 153 articles were truly bad and 38 articles were wrongly categorized as bad. The remaining 697 articles were ignored for the same reason applied to the ignored articles from the 'Good News' subsection. The manual evaluation was based on three judges. If two or more judges considered an article to be one particular type then the article was determined to be of that type.

3.2 Factual and Analysis Type News Articles

Many people believe buying stocks after seeing good news or selling after bad news is not going to work advantageously for them as they think any news that is worth affecting the stock price is already accounted for in the stock market or the news stories are intentionally written to influence people to either buy or sell stocks. We considered that this type of problem will be more prevalent in the non-factual news. Thus, we wanted to see whether there is any difference in the prediction power of the news articles based on its factual dimension. This means that we further divided the news articles into two categories. One is factual and the other is analysis. Factual news articles refers to the reporting of the actual events. Analysis type news articles usually refers to the ones forecasting the future stock price change of a company or explaining the reasons for the current stock price of a company by the financial analysts. Our hypothesis is that the factual type news articles will be more likely to have prediction power left to indicate the future stock price change in comparison to the analysis type articles. We came up with this hypothesis by assuming the financial analysts tend to generate analysis type news articles to influence the future stock price sometimes with less than sincere intents. More importantly we thought that most of the readers of the news articles are already aware of their intentions and thus the investors will likely to ignore the information in the analysis type news articles.

We manually categorized only the articles, which were judged to be truly either good or bad by us. The manual categorization according to the factual versus analysis dimension was also based on three judges. If two or more judges considered an article to be as one type then the article was determined to be of that type. Out of 709 truly good news articles, which were originally from the 'Good News' subsection, we came up with 98 factual type articles and 611 analysis articles. Similarly, Out of 153 truly bad news articles, which were originally from the 'Bad News' subsection, we came up with 42 factual type articles and 111 analysis articles.

3.3 Actual Stock Price Change Data

The stock price continuously changes over time. Thus, we sampled the stock price of a particular company at five points-in-time relative to the time of news release. The first point was the time of news release. The second point was one minute after the news release. The third point was three minutes after the new release. The fourth point was five minutes after the news release. The fifth and the last point was the stock price at the market close. The stock price information was from the Trading TIC (TAQ) data available at Koscom DataMall (http://datamall.koscom.co.kr/servlet/infoService/IssueTaqData). The TAQ data shows various trading related information such as the time and price of stock trading for all KOSPI listed companies.

One thing to note is that one news story can have more than one KOSPI listed company names mentioned. We only considered the company, which was mentioned first in the news stories except when the first occurring company was the stock brokerage firm. This exception was used as there were many cases, which referred to the brokerage firms as the source of information at the beginning of the news stories.

To get the baseline stock price change data, we pooled all good and bad news articles then sampled how the stock price changed after the news release. By having this data, we were able to tell how the stock price changed by chance. For example, if the stock price went up 60% of the times regardless of the news article being either good or bad then knowing a news article being good should yield more than 60% of the stock prices to go up to make our work meaningful.

The 〈Table 1〉 shows the baseline stock price change data of the KOSPI listed companies at four points-in-time after the news article release regardless of whether the corresponding news articles were either good or bad. For example, 542 company stock price went up one minute after the news release. This means that there was 62.88% chance of stock price going up one minute after any news release about a company based on the January 2006 data.

The 〈Table 2〉 shows how the stock price changed for a company appeared in the news articles judged to be truly good. On average across four points-in-time, the stock price of companies appeared in the factual type good news went up 81% of the time. Similarly, the stock price of companies appeared in the analysis type good news went up 75% of the time on average across four points-in-time after the news release. By ignoring the factual and analysis distinction in the news articles, the stock price went up 76% of the time on average. This result is encouraging and the finding supports our initial hypothesis that good news articles will positively influence the stock price of the company mentioned in the corresponding news.

The 〈Table 3〉 is analogous to the 〈Table 2〉 except that the 〈Table 3〉 shows how the stock price changed based on the truly bad news. Although the percentage of

stock price going down after the bad news release looks low but it is still higher than the number based on chance as shown in the 〈Table 1〉. Thus, we can claim that our initial hypothesis, that the bad news will negatively influence the stock, holds. One interesting observation is that the analysis type bad news articles strongly affected the stock price in the negative direction. This is opposite from the observation that we had for the good news.

Based on this result, we now have a supporting argument to develop an automatic news classification to predict the stock price change.

In the later section, we will once again compare the correct rate based on the automatic news classification with the

〈Table 1〉 Baseline Stock Price Change Data for the KOSPI listed companies appeared in 862 News Articles

	after 1 minute		after 3 minute		after 5 minute		ending price		averagge
	count		count		count		count		
go up	542	62.88%	592	68.68%	577	66.97%	556	67.50%	66%
same	168	19.49%	87	10.09%	79	9.16%	43	4.99%	11%
go down	152	17.63%	183	21.23%	206	23.90%	263	30.51%	23%
Total	862	100.00%	862	100.00%	862	100.00%	862	100.00%	100%

〈Table 2〉 How the Stock Price Changed based on the Truly Good News Articles after Dividing the News Articles by Factual and Analysis Type

		after 1 minute		after 3 minute		after 5 minute		ending price		averagge
		count		count		count		count		
Factual	go up	78	79.59%	79	80.61%	79	80.61%	82	83.67%	81%
	same	12	12.24%	10	10.20%	7	7.14%	4	4.08%	8%
	go down	8	8.16%	9	9.18%	12	12.24%	12	12.24%	10%
	Total	98	100.00%	98	100.00%	98	100.00%	98	100.00%	100%
Analysis	go up	436	71.36%	479	78.40%	466	76.27%	446	73.00%	75%
	same	107	17.51%	45	7.36%	49	8.02%	31	5.07%	9%
	go down	68	11.13%	87	14.24%	96	15.71%	134	21.93%	16%
	Total	611	100.00%	611	100.00%	611	100.00%	611	100.00%	100%
Total	go up	514	72.50%	558	78.70%	545	76.87%	528	74.47%	76%
	same	119	16.78%	55	7.76%	56	7.90%	35	4.94%	9%
	go down	76	10.72%	96	13.54%	108	15.23%	146	20.59%	15%
	Total	709	100.00%	709	100.00%	709	100.00%	709	100.00%	100%

〈Table 3〉 How the Stock Price Changed based on the Truly Bad News Articles after Dividing the News Articles by Factual and Analysis Type

		after 1 minute		after 3 minute		after 5 minute		ending price		averagge
		count		count		count		count		
Factual	go up	12	28.57%	16	38.10%	15	35.71%	14	33.33%	34%
	same	11	26.19%	8	19.05%	8	19.05%	4	9.52%	18%
	go down	19	45.24%	18	42.86%	19	45.24	24	57.14%	48%
	Total	42	100.00%	42	100.00%	42	100.00%	42	100.00%	100%
Analysis	go up	16	14.41%	18	16.22%	17	15.32%	14	12.61%	15%
	same	38	34.23%	24	21.62%	15	13.51%	4	6.60%	18%
	go down	57	51.35%	69	62.16%	79	71.17%	93	83.78%	67%
	Total	111	100.00%	111	100.00%	111	100.00%	111	100.00%	100%
Total	go up	28	18.30%	34	22.22%	32	20.92%	28	18.30%	20%
	same	49	32.03%	32	20.92%	23	15.03%	8	5.23%	18%
	go down	76	49.67%	87	56.86%	98	64.05%	117	76.47%	62%
	Total	153	100.00%	153	100.00%	153	100.00%	153	100.00%	100%

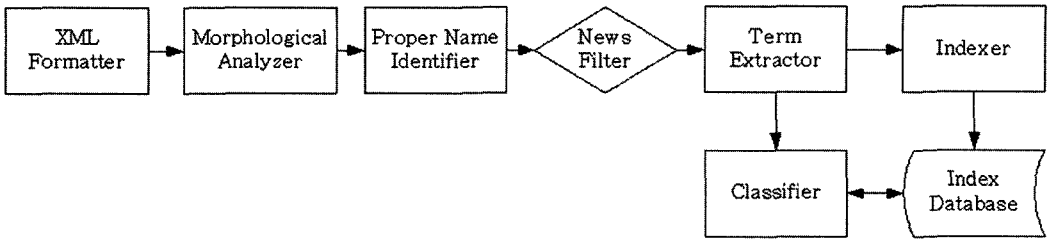
correct rate shown in the 〈Table 1〉 to determine the usefulness of the automatic system. We will consider the automatic system to be useful in predicting the stock price of a particular company if the correct rate is greater than the chance based correct rate.

4. Overall System Architecture

The automatic natural language processing-based multi-stage news classification system was designed to categorize news stories into their respective tones namely either good or bad with respect to the companies appeared in the news.

In terms of automatically categorizing the news articles, we used a simple pattern matching based on the manually extracted target patterns in phrase or clauses from the training data.

The 〈Figure 2〉 shows the overall system architecture of the pattern matching based news classifier. Initially news articles harvested from the web are processed by a XML formatter, which converts the html coded news articles into the XMLized texts. Then, the output of the XML formatter is processed by a morphological analyzer to generate word segmentation and part-of-speech information about each word in the title and the body parts of the news articles. The output from this stage of processing is fed into a proper name

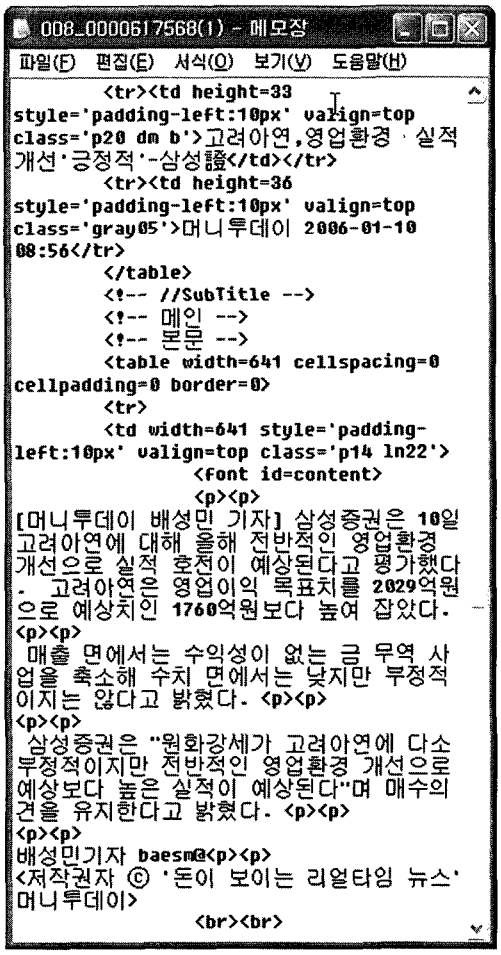


〈Figure 2〉 Pattern Matching based News Classification System

identifier, which identifies the boundary of the KOSPI listed company names and also finds its corresponding numeric company identifier. The news filter takes the output from the proper name identifier then removes the news articles, which did not include any KOSPI listed company names, from further processing. The news filter also removes articles with only the tabular information.

By analyzing the training data, phrases or clauses are manually generated to be either good or bad category indicating patterns. One researcher was responsible for extracting patterns indicating the good category from the good training data. Another researcher was responsible for extracting bad patterns from the bad training data. The researchers did not consult to each other about the patterns they found. The extracted patterns were stored in the pattern database with the corresponding category as the labels. The extracted patterns were based on the output of the XML formatter.

The test data went through the XML formatter, morphological analyzer, proper name identifier, and the news filter to



〈Figure 3〉 html Formatted News Article

remove the news articles without the KOSPI listed company names or the articles, which are mainly filled with the tabular information. After the news stories passed the news filtering criteria, the pattern matcher retrieves the corresponding XML formatted news article then consults the pattern database to determine whether the article should be categorized as either good or bad. If a news article had only bad patterns then the article is categorized as bad. If the news article had either only the good patterns or no matching pattern then the article is categorized as good. If the news article had both good and bad patterns then the article is categorized as good as well.

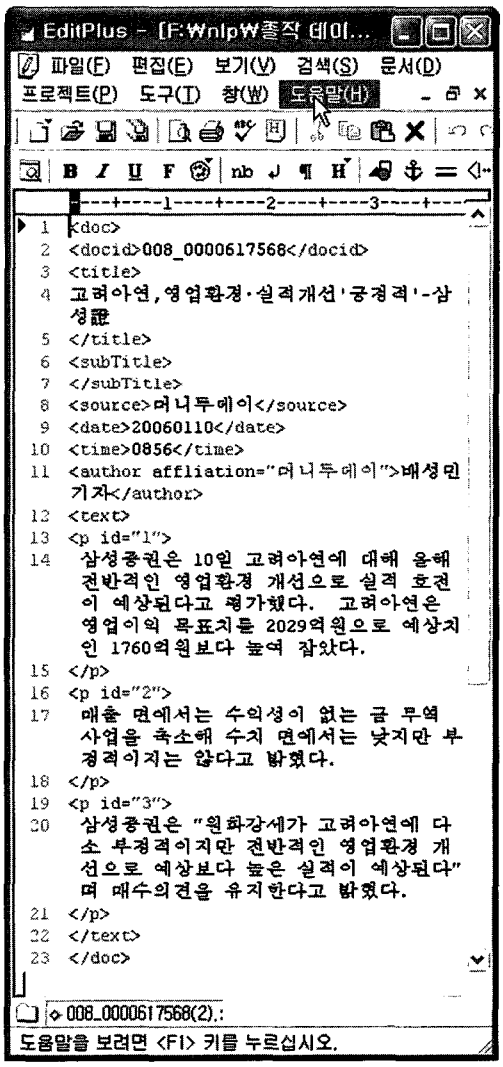
4.1 XML Formatter

The XML formatter is designed to take the html formatted news articles then convert them into XML formatted texts. The XML formatter uses simple pattern action rules to bracket document identifier, title of the news, source of the news, date/time of news release, author, and the body of the news. Each paragraph in the body of the news article is further bracketed with the paragraph identification information.

The 〈Figure 3〉 shows the html formatted news article as harvested from Naver's 'Good News' and 'Bad News' subsections. Some of the preceding, intervening, and following html codes were removed from the figure. The 〈Figure 4〉 shows the output from the XML formatter.

4.2 Morphological Analyzer

We used a KRISTAL Korean Morpho logical Analyzer, which is available from Korea Institute of Science and Technology Information (KISTI), to process the Korean texts in the title and the body part of the news articles.



〈Figure 4〉 XML Formatter Output

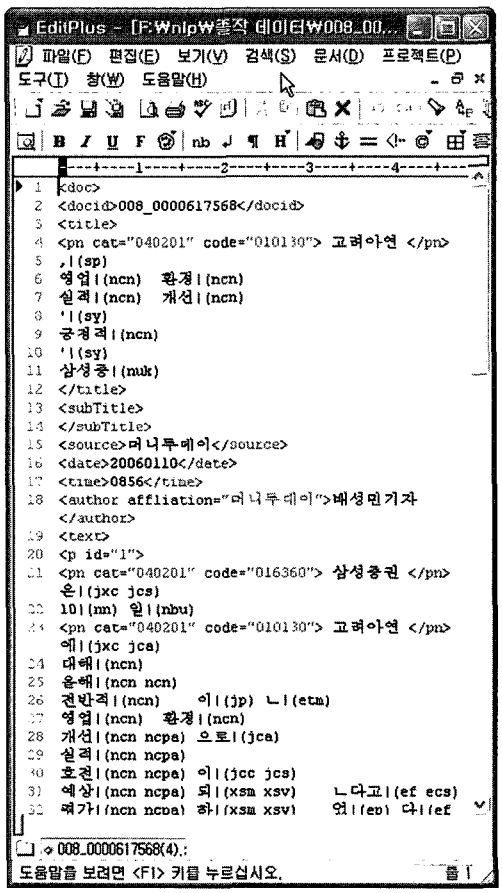
The morphological analyzer produces multiple interpretations of each word in the text. For example, the word such as, '낮지만' will produce three ways to segment the word. Each word segment can be assigned with one or more part-of-speech tags. The first segmentation output is '낮|(ncn ncn) 지만|(jxc)'. The second segmentation produces '낮|(paa) 지|(ecx)'. Finally, the third segmentation generates '낮|(paa) 지만|(ecs)'. The first segment in the first segmentation output shows two possible part-of-speech tags for '낮'. The bar symbol separated the word segment from the following part-of-speech information for the segment. The part-of-speech information is enclosed in the parentheses. The multiple part-of-speech information is separated by comma.

4.3 Proper Name Identifier

The proper name identifier that we developed is limited to finding the KOSPI and KOSDAQ listed company names. Thus, we used a straightforward pattern matching algorithm by having a master list of company names and their variants.

There were 732 KOSPI listed and 1,583 KOSDAQ listed companies including the de-listed companies in the company name pattern database. By testing the company name identification against the news articles from January 2006 data, 9,311 names were correctly recognized. There were 1,071 names, which were unrecognized. 317 company names were wrongly recognized. This resulted in company name recognition precision at 96.7% and recall at 89.7%. The formula that we used to calculate the precision was correctly recognized divided by correctly recognized plus wrongly recognized. The recall formula was correctly recognized divided by correctly recognized plus unrecognized.

The failure analysis revealed a few problems. One of the problems was that certain company names were recognized as the general words. For example, the KOSPI listed company names such as '전방', '유화', '대상', '신흥', '도움', '동원', '유



〈Figure 5〉 Proper Name Identifier Output

화', '신흥', '선진', and '대유' were used more frequently as the general words than the company names in the texts. In addition, the KOSDAQ listed company names such as '다음', '서산', '케이스' also caused the similar problem.

We believe further use of the contextual information will improve the company name recognition accuracy. The 〈Figure 5〉 shows an output from the Proper Name Identifier. Each KOSPI and KOSDAQ listed company name is enclosed in 'pn', which is a XML tag followed by its proper name category information. 040201. The proper name category information is the value for the 'cat' attribute referring to the KOSPI listed companies. The value for the attribute, 'code' refers to the company identifier used in KOSPI. For example, 016360 refers to '삼성증권' and '010130' refers to '고려아연'.

4.4 News Filter

The news filter simply removes the news articles without the KOSPI listed company names from being further processed. In addition, the news filter also removes the news articles, which only include the tabular information. The existence of the tabular information is identified by the embedded html codes, which were carried through the previous processing stages.

4.5 Pattern Database and Pattern Matcher

The pattern database was constructed by manually extracting phrase or clauses, which can characterize the news article to be either good or bad with respect to the KOSPI listed company mentioned in the article. There were 392 training data for the good category and 400 for the bad category. The pattern database consisted of one set of patterns for good category and the other set of patterns for the bad category.

Phrases or clauses were extracted as good patterns if the pattern included two or more words with the positive meaning. For example, the clause, '내수경기 회복 시 수혜가 클 것', includes '회복' and '수혜', which are both positive words. Thus, the clause was selected as one of the patterns for the good category.

Another type of pattern for the good category was a phrase or clause including both positive and negative words given the overall meaning of the phrase or clause was positive. For example, the clause, '하락세를 접고 반등', includes '하락세', which is a negative word and '반등', which is a positive word. The overall emphasis is on the second positive word. Therefore, we consider this clause to be a good pattern.

Similarly, phrases or clauses were extracted as patterns for the bad category if the pattern includes a negative word and another negative word. For example, '영업 손실로 연간 기준 적자로 전환하였다' included '영업 손실' and '적자전환', which are both negative words. Thus, we considered this pattern to be a part of the bad pattern database.

If a phrase or clause containing both

positive and negative words and also if the overall meaning of the phrase or clause is negative then the pattern is considered to be a pattern for the bad category. For example, a clause such as '흑자에서 적자로 전환' included '흑자' which is a positive word but it also included '적자전환', which is negative word. The overall meaning of the clause was negative and thus we considered this pattern as a bad category indicator.

100 patterns, indicating good category, were extracted from the good training data. 103 patterns for the bad category were extracted from the bad training data.

The pattern matching module simply scans the incoming news articles against the pattern database entries. Each incoming news article was then categorized as bad if it only had one or more bad patterns. Similarly, each incoming news article was categorized as good if it had at least one good pattern regardless of having one or more bad patterns. Articles without any matching pattern was also categorized as good.

5. Experiment Results

The pattern matching based news classifier was trained with 392 articles, which belonged to the good category, and 400 articles belonging to the bad category. These training articles were randomly selected from the January and February 2006 data set. The separate testing data consisted of 200 articles judged to be good and 100 articles evaluated to be bad. The testing data was also randomly selected from the January and February 2006 data set.

〈Table 4〉 Pattern Matching based News Classification Experiment Result

		Predicted		Recall
		Good	Bad	
Observed	Good	191	9	95.5% (191/200)
	Bad	17	83	83.0% (83/100)
Precision		91.8% (191/208)	90.2% (83/92)	

The experiment result is presented as a confusion matrix as shown in the 〈Table 4〉. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an observed (or actual) class. If we translated the numbers in the confusion matrix into the typical information retrieval measures such as precision and recall then the precision for correctly categorizing the good news articles is 91.8% by dividing 191, which is the number of correctly categorized good news stories, with the same 191 plus 17, which is the number of wrongly categorized bad news stories. Similarly, the precision for correctly categorizing bad news articles is 90.2% by dividing 83, which is the number of correctly categorized bad news stories, with 9, which is the number of wrongly categorized good news stories, plus 83. The recall for the good

article categorization is 95.5% by dividing 191 with 191 plus 9, which is the number of wrongly categorized good news stories. The recall for categorizing bad news articles is 83.0% by dividing 83 with 17, which is the number of wrongly categorized bad news stories, plus 83.

Not all the stock prices go up after the release of the corresponding truly good news stories. The same is true for the truly bad news stories. The 〈Table 2〉 shows this trend for the good news and the 〈Table 3〉 for the bad news.

These results were computed by assuming that human judgements of the good or bad news stories were 100% correct. 72.5% of the stock prices went up one minute after the corresponding truly good news release. 78.7% went up after three minutes of the good news release, 76.9% after five minutes, and 74.5% at the stock market closing. Similarly, 49.7% of the stock prices went down one minute after the corresponding bad news release. 56.9% went down after three minutes of the bad news release, 64.1% after five minutes, and 76.5% at the stock market closing.

However, the actual percentage of the stock prices correctly going up or down after the good or bad news release will be lower than the ideal case of 100% correct news classification as the correct rate for the automatic news classification is lower than 100%. As we can see from the Table 4, the precision of good news categorization was 91.8% and the bad news categorization was 90.2%. The truly good news stories categorized as bad will wrongly predict the corresponding stock price to go down. The truly bad news stories categorized as good will wrongly predict the corresponding stock price to go up.

〈Table 5〉 Actual Stock Price Change of the Training Data Set

	Good	Bad
go up	141 (70.5%)	27 (27.0%)
same	13 (6.5%)	5 (5.0%)
go down	46 (23.0%)	68 (68.0%)
Total	200	100

The 〈Table 5〉 shows actual stock price change data for the corresponding testing data set. The 〈Table 5〉 shows the stock price changes when the news classification is 100% correct. 141 companies' stock prices went up one minute after 200 news stories, which belong to the good training data set, were released. 68 companies' tock prices went down one minute after 100 news stories, which belong to the good training data set, were released.

〈Table 6〉 Stock Price Change of the Training Data Set based on the Pattern Matching based News Classification

	good	bad
go up	138 (69.0%)	36 (36.0%)
same	0 (0.0%)	0 (0.0%)
go down	62 (31.0%)	64 (64.0%)
Total	200	100

To compute how well does the imperfect pattern matching based news classification

predicts the stock price change, we checked how each news story, which were automatically categorized as either good or bad, led the corresponding stock price to go up or down. The result is shown in the 〈Table 6〉. Since the pattern matching based classification results in the binary categories of good and bad, all news stories led to either the corresponding stock to go up or down. There was no prediction, which led to no change in the stock price as shown in the real cases. However, the actual stock price change data in the 〈Table 5〉 shows that 6.5% of the good news and 5% of the bad news did not influence the stock price to change.

The stock price change prediction correct rate for the pattern matching based classification, which is shown in the 〈Table 6〉, was less than the correct rate shown in the 〈Table 5〉. This was expected as the 〈Table 5〉 is based on the 100% news story classification correct rate.

However, the result in the Table 6 is still better than the chance based figures, which was shown in the 〈Table 1〉. We can expect to achieve 66% correct rate if we considered all stock price to go up no matter what the nature of the preceding news stories were about. Similarly, we can achieve 23% correct rate if we considered all stock price to go down no matter what the nature of the preceding news stories were about.

In comparison, the use of pattern matching based news classification will result in the 69% correct rate for predicting the stock price to go up and the 64% correct for predicting the stock price to go down. In summary, the use of the automatic news classification system will yield better than chance correct rate in predicting the stock price change especially in predicting whether the stock price will go down.

6. Conclusion & Future Works

Our initial hypothesis was that the good news about a particular KOSPI company will influence the stock price to go up and bad news will influence the corresponding stock price to go down. The hypothesis was turned out to be correct based on our empirical analysis of the stock price change data.

We also wanted to test whether the automatic categorization of the news stories according to their tone namely good or bad with respect to the companies mentioned in the news articles can lead to the stock price change prediction, which is better than chance. The pattern matching-based classification system achieved the news story categorization precision at 91.8% for the good news and 90.2% for the bad news.

Consequently, the pattern matching based news classification led to the 69% correct rate when predicting the stock price to go up and the 64% correct rate when predicting the stock price to go down. There correct rates were indeed better than the chance based prediction. Thus, we can claim that our research goals have been achieved.

However, there are many future works to be done. Firstly, we should be able to increase the news story classification correct rate by improving the company name identification accuracy. Although, there are limited number of companies the KOSPI or KOSDAQ listed companies, a simple string matching strategy cannot achieve high accuracy. Thus, we will be required to explore additional use of contextual information to improve the accuracy. In addition, some of the company name identification errors stems from the morphological analysis errors. We should develop an add-on post-processing module for the morphological analysis system to improve its accuracy.

More importantly, we will look into the practicality of using an automatic news classifier. The human classification of news stories might be a better option than the automatic classification if the human can achieve better accuracy in very short time period. Thus, we will compare the human classification performance against the machine classification in the next phase of the research from the accuracy, consistency, processing time perspectives.

In this research, we only considered one news story as the affecting factor in predicting the stock price change. There might be combined effect based on the multiple news articles. Furthermore, we will explore the effect of different news agencies releasing similar news articles about the same event. These almost duplicate stories might overly influence the stock price to go up or down in the automatic news classification. Thus, we will investigate the impact of utilizing multiple news stories in predicting the stock price change in the future research.

References

Joo, Eun Sok. 2004. A study on binary classification of news articles for stock price prediction. Master Thesis, Yonsei University. (In Korean).

KRISTAL Korean Morphological Analyzer 〈http://www.kristalinfo.com/K-Lab/ma/〉.

Lee, Soo Yeon. 2003. A comparative study on stock price prediction using statistical method and artificial intelligence method : KOSPI 200 index and private stock price index. Master Thesis, Yonsei University. (In Korean).

Knolmayer, Gerhard F. and Marc-André Mittermayer. 2006. NewsCATS: A News Categorization and Trading System. Sixth IEEE International Conference on Data Mining (ICDM'06) pp. 1002-1007

Mittermayer, Marc-Andre. 2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques. Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS' 04) - Track 3 - Volume 3

Paik, W., Harwell, S., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., & Amice, C. 2001. " Applying Natural Language Processing Based Metadata Extraction to Automatically Acquire User Preferences." Proceedings of the First International Conference on Knowledge Capture. Victoria, British Columbia, Canada.

Graham-Cumming, John. 2005. "Naive Bayesian Text Classification: Fast, accurate, and easy to implement." Dr. Dobb's Journal.