B. Tech. Project Report

Utkarsh Upadhyay
(Y5488)
Mentor: Prof. R.M.K. Sinha

November 6, 2008

Contents

Real Time Extraction of Quantitative Information from Economic news

1 Introduction & motivation

Extraction of data from natural language text is a very standard problem, be it extraction of links from a web page to extraction of topic from a body of text. A similar attempt is made here to attempt to extract *quantitative information* from Economic news, which is released in real time over the Internet. The importance of this is undoubted, and similarly undoubted are the state-of-the-art techniques in existence. This task of extracting information could have been done on any kind of data but *economic news* has a special significance.

1.1 Interactive Broker's Algorithmic Trading Olympiad

Every year, since 2002, **Interactive Broker** has been running an on-line competition of algorithmic trading over equity markets. It has run for a about three months, i.e., January, February and March of each year since then. Ramnik Arora (MTH) and I took part in the competition last year and wish to do it again this year. According to the guidelines of the competition, we are allowed use of market news in making our decisions for trading, as long as the decisions are taken by a program. I intend to make the program usable as a module till then.

2 The task

As we tried trading using only numeric information in the last year, we clearly realise the importance of Quantitative information (*next Quarter estimates, profits, etc.*) being released in real time over the Internet. We can attempt to perform *Sentiment analysis* over the news, but such information though helpful, is generally inaccurate. *Ankit Soni*[1][2] did his thesis on the topic, and his results were that after extensive testing, he was able to attain about 56% accuracy. His classifier identified each news as positive, negative, or as neutral.

However, if we are able to extract quantitative information from the news, then by comparing it with the previous estimates, we obtain numeric data which

[1] Working under Prof. Harish Karnick, 2005

[2] *Prediction of stock price movements based on concept map information* by Ankit Soni, Nees Jan van Eck, and Uzay Kaymak in 2006

is easier to work with and they generally have a more predictable impact on the market prices of the equity. Also, cultivating information from more than one source would benefit us since if the number of sources is large, then it would no longer remain possible for a human to out-perform the machine in terms of content handled. Also, it would be interesting to note quantitatively different kind of news from different sources and question their reliability.

2.1 Existing work

The academic work in the area is limited to financial ontology[3], and is generally done behind closed doors. Among the better known open implementations is ***FASTUS* (1993)**. Most of the work in the field is closed source now.

2.2 Problem statement

Extract the figures of future estimates of:

1. Revenues
2. Profits/Losses

from natural-language news released over the Internet. Also, to perform this task in real-time, as the markets exhibit extremely quick assimilation of any news.

2.3 Restricting Domain

Also, most commercial software that do the task perform a very detailed level of analysis. We, however, would only need preliminary analysis: Only estimates and Profit/Losses. Hence, limiting the domain.

2.4 First Stage: Condensation

We first skim the news to a find a few sentences that might contain information relevant to us. The task continues to further remove parts of the sentences to arrive at as succinct phrase that we can possibly fit in a template. This is done to limit our search space. For example:

Amid political turmoil in US and quavering price of oil after the recent wars, the estimates for Q3 dropped by 17%.
to
The estimates for Q3 dropped by 17%.

-Yahoo news article.

[3] *WP10: Case study eBanking : Financial Ontology* by Silvestre Losada Alonso, Jose Luis Bas, Sergio Bellido, Jess Contreras, Richard Benjamins and, Jose Manuel Gomez on October 6th, 2005

2.4.1 Non-natural language news

However, it often happens that the news released itself contains quoted tables from the company's reports, etc. which are not parsable. Hence, they need to be treated differently and this stage itself.

2.5 Second Stage: Template Design

There are two primary approaches to template design:

1. **Manually:** A fixed number of templates hand crafted to meet a few sentences.
2. **Automatically:** Analogous to Semantic Role Labeling, using machine learning.[4]

Prior to template filling, we would also need to resolve temporal location of the fact as well as disambiguate the anaphoric references. That would again be accomplished using dependency trees provided by the Stanford Parser, or using the neighbouring sentences.

3 Work done so far

Parsers for two news sources have been made:

1. Yahoo Finance Earnings news
2. MarketWatch (Wall Street Journal)

A RSS reader has been made for Yahoo news and real time fetching of news has been tested, and used to fetch more than 400 items of news meant for back-testing. There were two versions of the parser available:

1. **Factored Lexical Parser**: Better and more accurate parsing using a few English language specific lexical rules (very slow)
2. **PCFG Parser**: Uses only statistical means and, hence, is fairly fast.

As the nature of the application requires execution to be fast, we intend to first parse using the PCFG parser and then parse the shortlisted sentences (received after Stage 1 pruning) using the Lexical Parser. Also, a basic summarizer is also in place, and different patterns are being tested to see which provides maximal coverage of the data. For example, the following is a news article with the sentences chosen in bold.

1. **NEW YORK (AP) – Shares of ITT Corp., a military contractor that provides engineered products, fell to a more than four-year low Friday after it lowered its full-year revenue outlook to below Wall Street 's estimate.**

[4] *Automatic Semantic Role Labeling* by Daniel and Jurafsky in 2002

2. **The White Plains, N.Y., company expects sales of $ 11.5 billion to $ 11.6 billion for 2008, down from a previous estimate for $ 11.6 billion to $ 11.7 billion.**
3. **Analysts polled by Thomson Reuters expect, on average, revenue of $ 11.65 billion for the year.**
4. "Our strategies remain focused on long-term growth and continued operational improvements, however, we are preparing for projected softening of the global economy," Steve Loranger, chief executive, said in a statement.
5. The company also cited currency exchange issues in lowering its guidance.
6. *Shares fell $ 3.91, or 9.5 percent, to close at $ 37.46.*
7. *Earlier in the session, the stock hit a four-year low of $ 36.56.*
8. **Shares had traded between $ 69.73 and $ 39.02 in the past year; the last time shares traded so low was March 2004.**
9. Citi Investment Research analyst Jeffrey T. Sprague said in a client note that ITT's shares are "fully valued," and that he is concerned " that ITT is focused on making acquisitions and would prefer to see the company repurchase shares and raise its dividend even more."
10. *The analyst has a 'Hold' rating on the shares and a $ 63 price target.*
11. The company's "organic growth is no longer exceptional versus the group, yet it trades at a premium and visibility is diminishing," he said.
12. "The stock should trend higher over time with earnings growth, but we do not see much room for additional multiple expansion. "

The lines in **bold** are chosen by a stricter rule than the lines in *italics*. Various different pruning patterns are still being tried for the condensation, and the final choice would definitely depend on the second stage. Initial tagging of text is the next step before we attempt to either make manual templates or machine learn them. Also, we would need to make models for temporal as well as anaphoric disambiguation. The disambiguation and the named entity recognition here can be seen as an intermediately step, as they will be treated next.

4 Time line

The time requirements are in place owing to the competition deadlines. The competition specifications need to be submitted before $31i^{st}$ December. Hence, at least the specifications of the project would be concrete by then. Also, since the deployment takes place on 12th January, the first running prototype should be ready by then. Thereafter, the competition will run to the 6^{th} March, 2008. After that, detailed analysis and possibly investigation of similar application areas would ensue.

5 Tools / References

Tools of the trade:

1. **HTML Parser**: http://htmlparser.sourceforge.net/
2. **The Stanford Parser: A statistical parser**: http://nlp.stanford.edu/software/lex-parser.shtml

Papers being used:

1. *Rule based synonyms for entity extraction from noisy text.* by Rema Ananthanayayanan, Vijil Chenthamarakshan, Prasad M Deshpande, Raghuram Krishnapuram (ACM 2008)
2. *WP10: Case study eBanking : Financial Ontology* by Silvestre Losada Alonso, Jose Luis Bas, Sergio Bellido, Jess Contreras, Richard Benjamins and, Jose Manuel Gomez in 2006
3. *Automatic Semantic Role Labeling* by Daniel and Jurafsky in 2002

6 Acknowledgements

The work is not yet complete, but for the progress made so far, I would like to heartily thank Prof. R.M.K. Sinha and Prof. Harish Karnick for their time and effort in reviewing the strategies and giving me new directions with respect to the problem.

Also, I would like to thank my friend, Mr. Ramnik Arora for his help. I, however, am unable to thank everyone who has helped here, but I would like to express my sincerest thanks to all who have been a part of this project.