Extraction of Quantitative information from economic news

Utkarsh Upadhyay
Mentor: Prof. R.M.K. Sinha

Indian Institute of Technology, Kanpur

November 8, 2008

Outline

1 Introduction

2 The Attempted Solution

3 Work completed so far

4 Further work

Motivation for the problem

- *Why?* Information is money, literally
- *Why news?* Fast, free and fresh information
- *Why quantitative?* Numbers are easier to work with
- *Why automatic?* Multiple sources of news and efficient aggregation

Why Economic?

- A real time arena.
- Excitement
- *Interactive Broker's* International Algorithmic Trading Olympiad.

What already exists?

- News sources
- Natural Language Processing tools
- Ontology of the financial world
- A real time application exists *(Stock Market)*

Why not already then?

- *Not true:* **FASTUS(1993)** for mergers and acquisitions news
- *But almost true:* Most work under closed doors
- Academic interest is more general *(Semantic Role Labeling)*

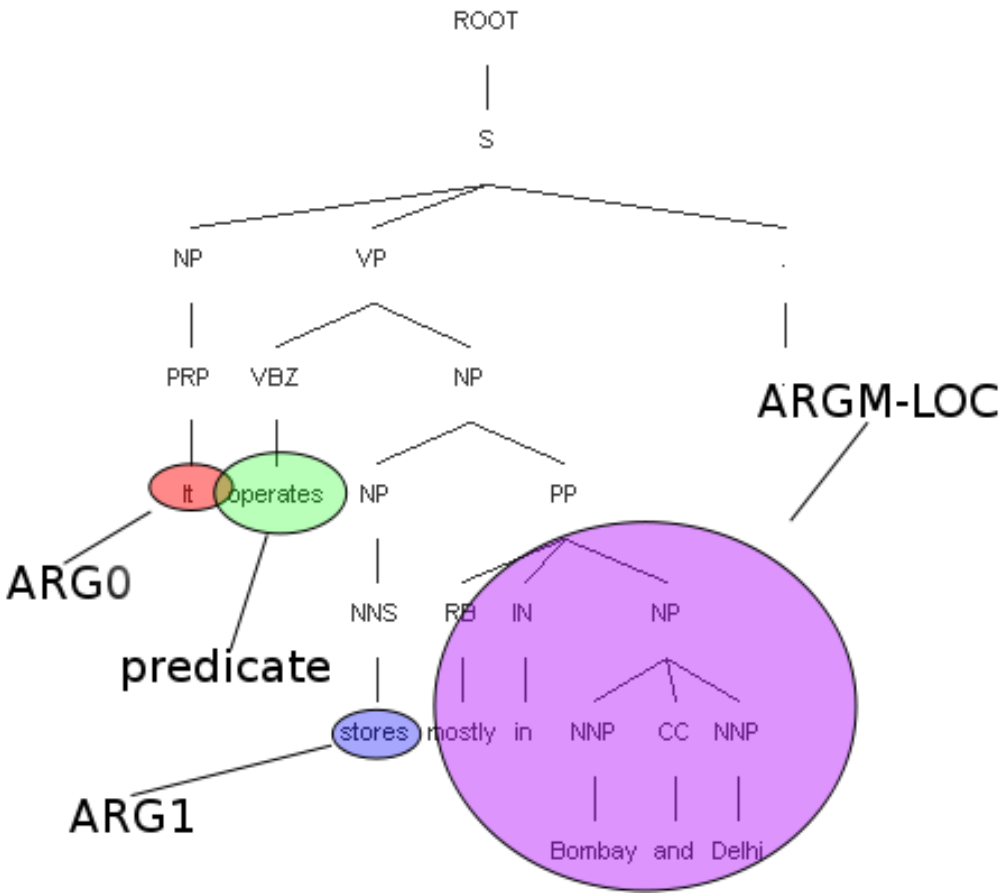
Problem Statement

Determine in real time from economic news for a group of companies:

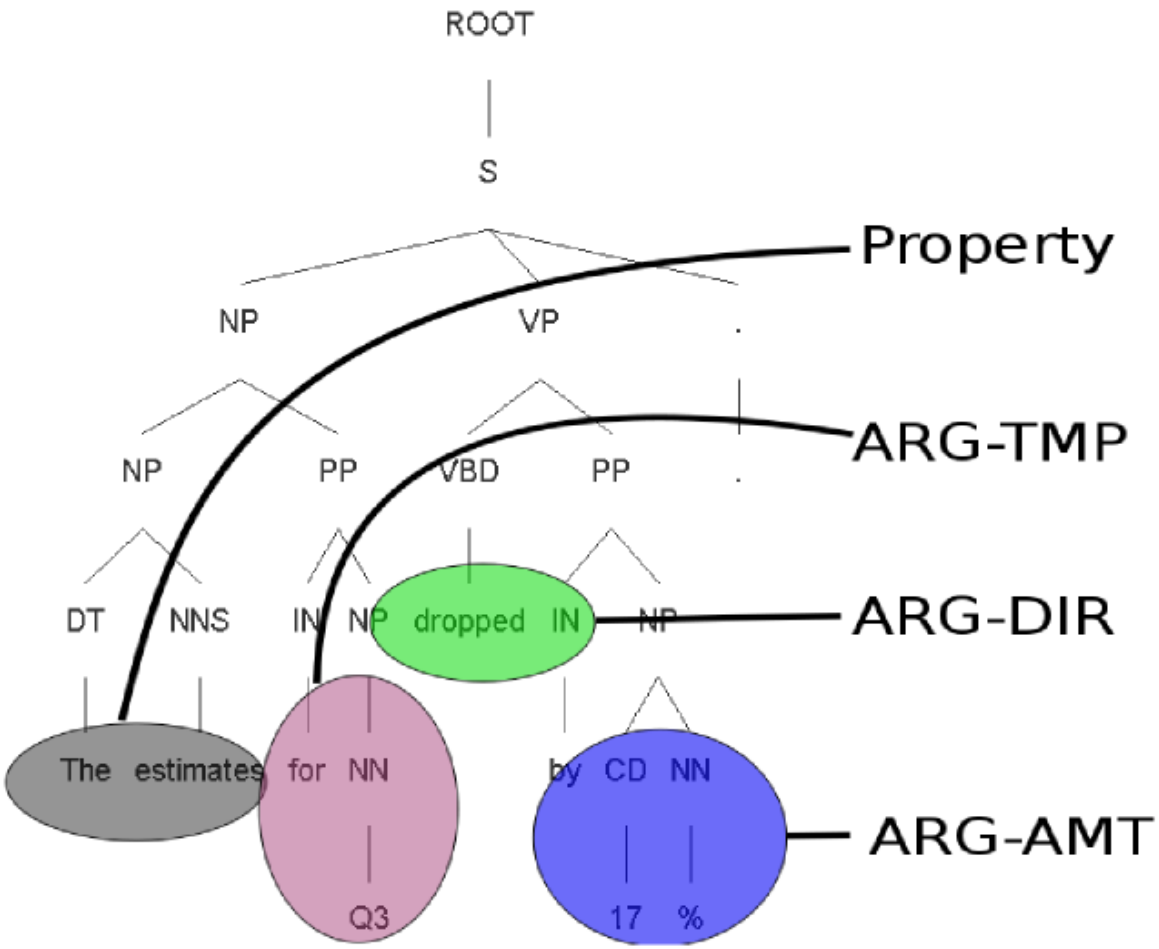
1. Recent or Estimated Profit/Loss
2. Revenues Estimated

Can be seen as a very restricted case of a very general problem.

The related academic problem (Semantic Role Labelling)



The exact Problem



Outline

1 Introduction

2 The Attempted Solution

3 Work completed so far

4 Further work

Two stage architecture

1. Condense the news
2. Extract information from the pith
    1. Resolve anaphoric references using neighboring sentences
    2. Retrieve temporal information
    3. Fit into a template

Where's the news?

The news is fetched from the web:

1. *When?*: Updates real time RSS sources
   - Yahoo Finance
   - MarketWatch (WSJ)
2. *What form?*: News in human readable form, in HTML
3. *HTML or Text?*
   - A manually designed filter extracts textual news
   - However, tables are present in the some (17%) articles

What's in a summary?

The aim:

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.

Done by:

1. Using a parse of the sentences, and,
2. Looking for Quantitative information

Template Design

- Designed manually
- Assignation could be:
  1. Using hand made rules
  2. Using machine learning (not unlike SRL)

Outline

1 Introduction

2 The Attempted Solution

3 Work completed so far

4 Further work

Pre-condensation-processing

- The HTML parsers for:
  1. MarketWatch (WSJ)
  2. Yahoo Earnings feed
- A Multipurpose RSS reader : Can work with any RSS feed
- Integration of RSS reader and HTML parser for Yahoo!
  1. Works in real time (successfully checks for updates every 15 sec)
  2. Has fetched more than 400 items of news.

Integration with the Parser

- The Stanford Parser is being used. (Not Charniak, or Collins)
- Tuning the parser:
  - Choosing **PCFG** or **Lexical** Parsing
    1. Accuracy
    2. Speed
  - Maximum sentence length to parse (80), etc. ...
- Groping in the parse-trees : Tgrep2 *(Stanford NLP group)*
- Deciding **what** to look for

An example summarization is provided in the report.

Outline

1 Introduction

2 The Attempted Solution

3 Work completed so far

4 Further work

Next rung...

With the parsing and summarization nearing completion, the second stage:

- Fixing temporal and anaphoric references.
- Manual tagging of data to recognize templates.

Working on a tight schedule.

Specification submission	**31st Dec '07**
Contest starts	**12th Jan '08**
Contest ends	**6th Mar '08**

Thank you

Utkarsh Upadhyay
utkarshu@iitk.ac.in