Extraction of Quantitative information from economic news

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

Indian Institute of Technology, Kanpur

April 14, 2009

Outline

- Introduction
- The Solution
- Meanwhile . . .
- The Problem
- Current state
- Further work

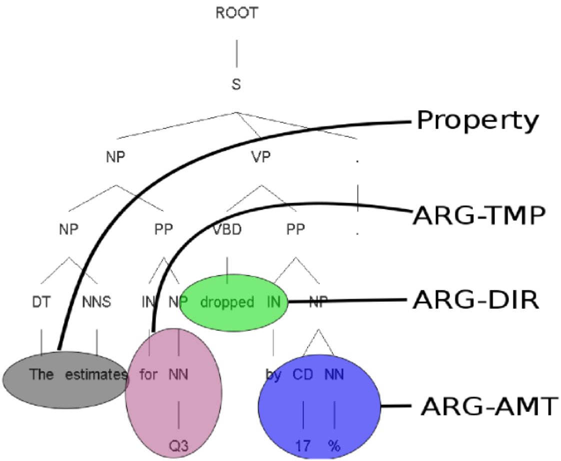
What was the problem, again?

Determine in real time from economic news:

1. Turn Over
2. Profit/Loss
3. Revenues

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

The exact Problem



Outline

Introduction

The Solution

Meanwhile . . .

The Problem

Current state

Further work

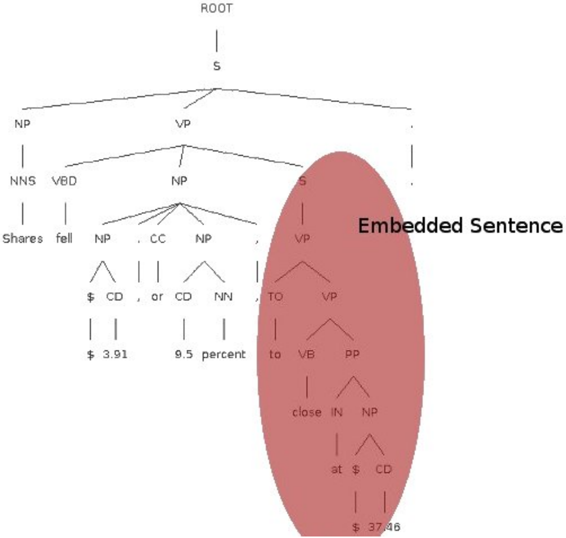
Where's the news?

The news is fetched from the web:

1. *When?*: Updates real time RSS sources
2. Yahoo Finance News
3. *What form?*: Human readable form, HTML
   - Manually designed text filter
   - However, (17%) articles have tables

Keeping it small

Breaking a sentence into simplest form.



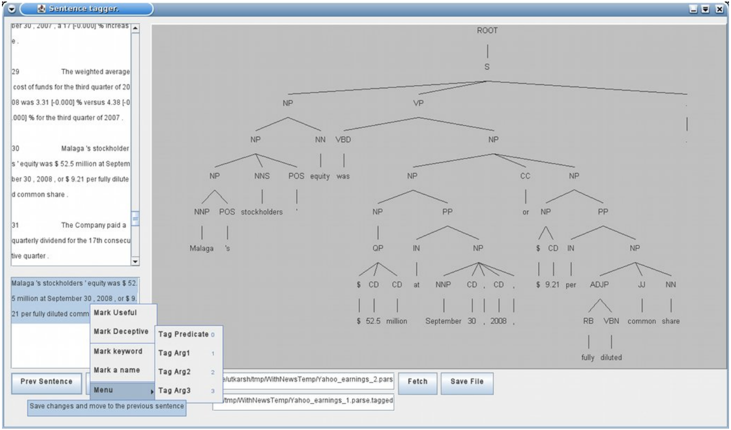
The right path

Frequency	Path	Usual Description
14.2%	$VB \uparrow VP \downarrow PP$	PP argument/adjunct
11.8%	$VB \uparrow VP \uparrow S \downarrow NP$	Subject (ARG0/ARG1)
10.1%	$VB \uparrow VP \uparrow NP$	Object (ARG0/ARG1)
...		

Table: Path from the predicate to the frame

Read/Tag/Repeat

Stanford Manual Annotator?



ber 30 , 2007 , a 17 [-0.000] % increas
e .

29 The weighted average cost of funds for the third quarter of 2008 was 3.31 [-0.000] % versus 4.38 [-0.000] % for the third quarter of 2007 .

30 Malaga 's stockholders ' equity was $ 52.5 million at September 30 , 2008 , or $ 9.21 per fully diluted common share .

31 The Company paid a quarterly dividend for the 17th consecutive quarter .

Malaga 's stockholders ' equity was $ 52.5 million at September 30 , 2008 , or $ 9.21 per fully diluted comm

e/utkarsh/tmp/WithNewsTemp/Yahoo_earnings_2.pars

/tmp/WithNewsTemp/Yahoo_earnings_1.parse.tagged

Save changes and move to the previous sentence

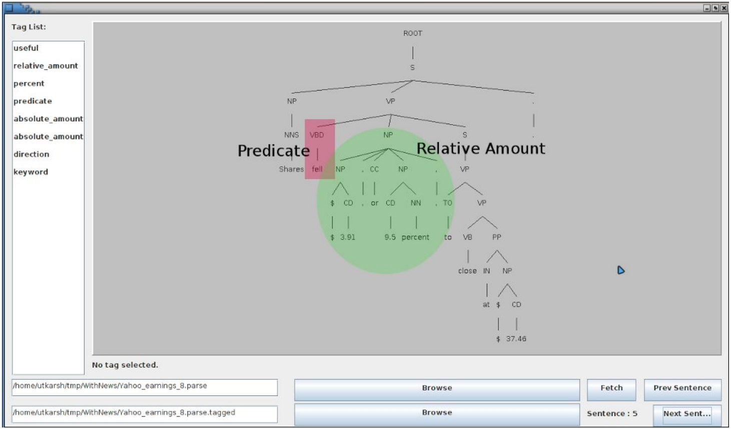
Figure: How to do it right.

What we get

```
<useful>
        <keyword>Shares</keyword>
        <direction>
                 <predicate>fell </predicate>
        </direction>
        <relative_amount>
                 <absolute_amount>$ 3.91</absolute_amount>
                 , or
                 <percent>9.5 percent</percent> ,
        </relative_amount>
</useful>
```

Putting them together

- Tregex
- Tsurgeon



No tag selected.

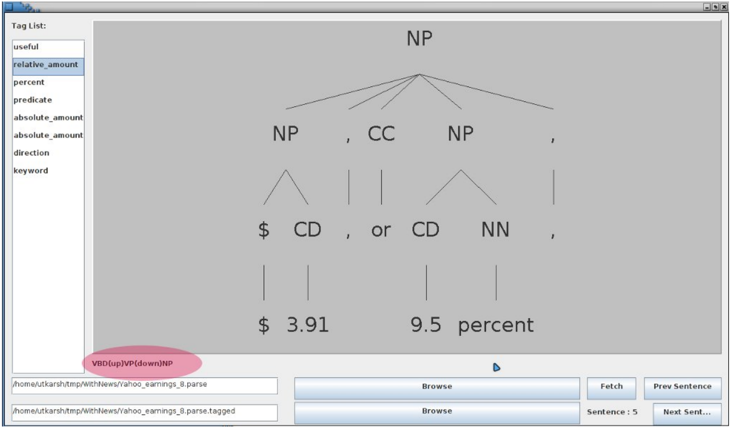
/home/utkarsh/tmp/WithNews/Yahoo_earnings_8.parse

/home/utkarsh/tmp/WithNews/Yahoo_earnings_8.parse.tagged

Sentence : 5

Figure: The whole sentence

The path



Tag List:

- useful
- relative_amount
- percent
- predicate
- absolute_amount
- absolute_amount
- direction
- keyword

VBD(up)VP(down)NP

/home/utkarsh/tmp/WithNews/Yahoo_earnings_8.parse

/home/utkarsh/tmp/WithNews/Yahoo_earnings_8.parse.tagged

Sentence : 5

Figure: A path from the *predicate* to the tag (*rel amt*)

Outline

Introduction

The Solution

Meanwhile . . .

The Problem

Current state

Further work

IB trading Olympiad

1. Direct Connection to the Internet
2. Ptolemy
3. Sectoral Arbitrage strategy

Paper in Advanced Data Analysis, Business Analytics and Intelligence conference in IIM-A.

Outline

Introduction

The Solution

Meanwhile . . .

The Problem

Current state

Further work

The Sparsity

Very sparse data:

Example

For the relation:be -> absolute_amount

`VBD(up)VP(down)NP(down)NP`: 2

`VBD(up)VP(down)PP(down)NP(down)PP(down)NP`: 1

Adding more features?

Changing stance

- Some tags discarded
- Dictionary based name extractor
- RegEx based time extractor

Outline

Introduction

The Solution

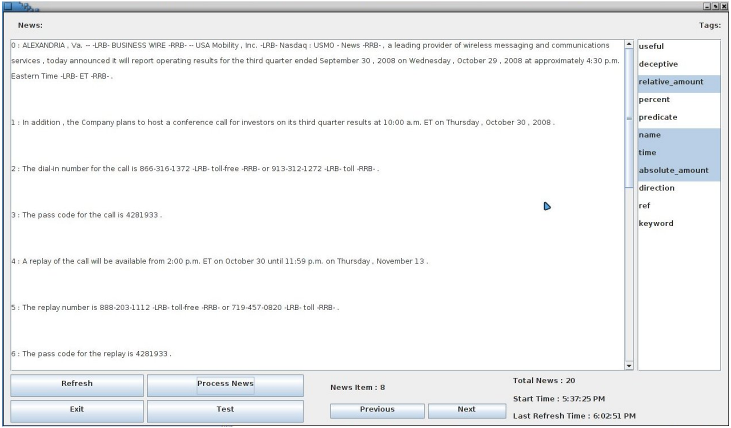
Meanwhile . . .

The Problem

Current state

Further work

Useless news



News:

0 : ALEXANDRIA , Va. -- -LRB- BUSINESS WIRE -RRB- -- USA Mobility , Inc. -LRB- Nasdaq : USMO - News -RRB- , a leading provider of wireless messaging and communications services , today announced it will report operating results for the third quarter ended September 30 , 2008 on Wednesday , October 29 , 2008 at approximately 4:30 p.m. Eastern Time -LRB- ET -RRB- .

1 : In addition , the Company plans to host a conference call for investors on its third quarter results at 10:00 a.m. ET on Thursday , October 30 , 2008 .

2 : The dial-in number for the call is 866-316-1372 -LRB- toll-free -RRB- or 913-312-1272 -LRB- toll -RRB- .

3 : The pass code for the call is 4281933 .

4 : A replay of the call will be available from 2:00 p.m. ET on October 30 until 11:59 p.m. on Thursday , November 13 .

5 : The replay number is 888-203-1112 -LRB- toll-free -RRB- or 719-457-0820 -LRB- toll -RRB- .

6 : The pass code for the replay is 4281933 .

Tags:

useful
deceptive
relative_amount
percent
predicate
name
time
absolute_amount
direction
ref
keyword

News Item : 8

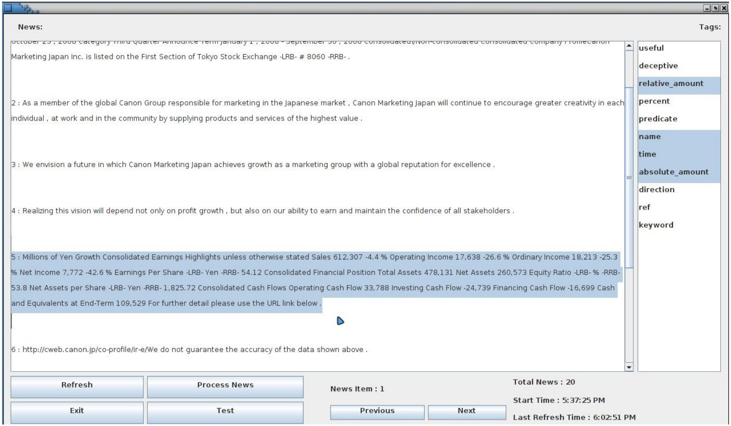
Total News : 20

Start Time : 5:37:25 PM

Last Refresh Time : 6:02:51 PM

Figure: No useless news

Unparsed news



October 23 , 2008 Category Third Quarter Announce Term January 1 , 2008 - September 30 , 2008 Consolidated/Non-consolidated Consolidated Company ProfileCanon Marketing Japan Inc. is listed on the First Section of Tokyo Stock Exchange -LRB- # 8060 -RRB- .

2 : As a member of the global Canon Group responsible for marketing in the Japanese market , Canon Marketing Japan will continue to encourage greater creativity in each individual , at work and in the community by supplying products and services of the highest value .

3 : We envision a future in which Canon Marketing Japan achieves growth as a marketing group with a global reputation for excellence .

4 : Realizing this vision will depend not only on profit growth , but also on our ability to earn and maintain the confidence of all stakeholders .

5 : Millions of Yen Growth Consolidated Earnings Highlights unless otherwise stated Sales 612,307 -4.4 % Operating Income 17,638 -26.6 % Ordinary Income 18,213 -25.3 % Net Income 7,772 -42.6 % Earnings Per Share -LRB- Yen -RRB- 54.12 Consolidated Financial Position Total Assets 478,131 Net Assets 260,573 Equity Ratio -LRB- % -RRB- 53.8 Net Assets per Share -LRB- Yen -RRB- 1,825.72 Consolidated Cash Flows Operating Cash Flow 33,788 Investing Cash Flow -24,739 Financing Cash Flow -16,699 Cash and Equivalents at End-Term 109,529 For further detail please use the URL link below .

6 : http://cweb.canon.jp/co-profile/ir-e/We do not guarantee the accuracy of the data shown above .

News Item : 1

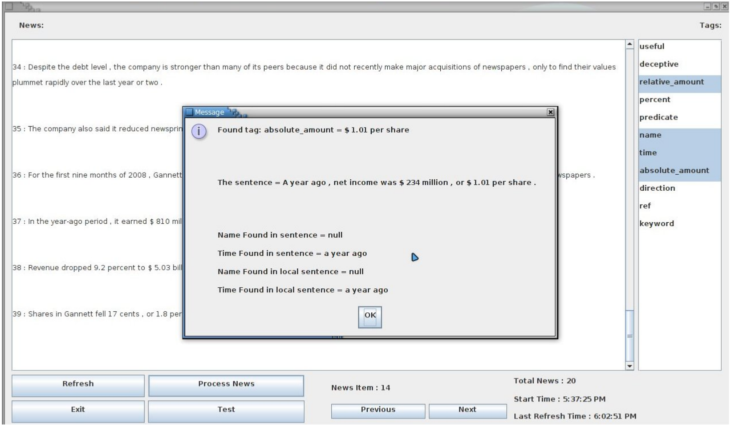
Total News : 20

Start Time : 5:37:25 PM

Last Refresh Time : 6:02:51 PM

Figure: No unparsed news

Absolute amount



News:

34 : Despite the debt level , the company is stronger than many of its peers because it did not recently make major acquisitions of newspapers , only to find their values plummet rapidly over the last year or two .

35 : The company also said it reduced newsprin

36 : For the first nine months of 2008 , Gannett … wspapers .

37 : In the year-ago period , it earned $ 810 mil

38 : Revenue dropped 9.2 percent to $ 5.03 bill

39 : Shares in Gannett fell 17 cents , or 1.8 per

Message

Found tag: absolute_amount = $ 1.01 per share

The sentence = A year ago , net income was $ 234 million , or $ 1.01 per share .

Name Found in sentence = null

Time Found in sentence = a year ago

Name Found in local sentence = null

Time Found in local sentence = a year ago

News Item : 14

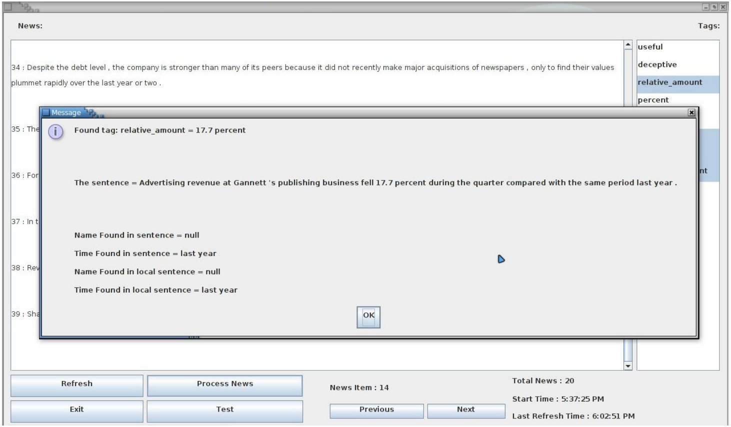
Total News : 20

Start Time : 5:37:25 PM

Last Refresh Time : 6:02:51 PM

Figure: Absolute Amount

Relative amount



34 : Despite the debt level , the company is stronger than many of its peers because it did not recently make major acquisitions of newspapers , only to find their values plummet rapidly over the last year or two .

35 : The

36 : For

37 : In t

38 : Rev

39 : Sha

Found tag: relative_amount = 17.7 percent

The sentence = Advertising revenue at Gannett 's publishing business fell 17.7 percent during the quarter compared with the same period last year .

Name Found in sentence = null

Time Found in sentence = last year

Name Found in local sentence = null

Time Found in local sentence = last year

News Item : 14

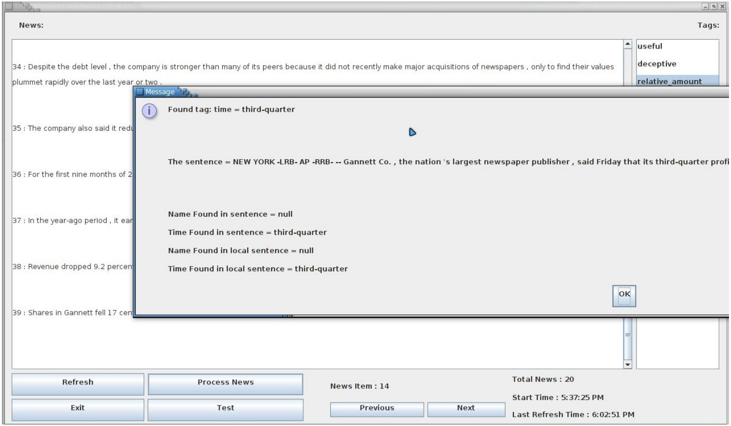
Total News : 20

Start Time : 5:37:25 PM

Last Refresh Time : 6:02:51 PM

Figure: Relative Amount

Not as helpless



News:

Tags:

34 : Despite the debt level , the company is stronger than many of its peers because it did not recently make major acquisitions of newspapers , only to find their values plummet rapidly over the last year or two .

35 : The company also said it redu

36 : For the first nine months of 2

37 : In the year-ago period , it ear

38 : Revenue dropped 9.2 percen

39 : Shares in Gannett fell 17 cen

useful

deceptive

relative_amount

Message

Found tag: time = third-quarter

The sentence = NEW YORK -LRB- AP -RRB- -- Gannett Co. , the nation 's largest newspaper publisher , said Friday that its third-quarter profi

Name Found in sentence = null

Time Found in sentence = third-quarter

Name Found in local sentence = null

Time Found in local sentence = third-quarter

OK

Refresh | Process News

Exit | Test

News Item : 14

Previous | Next

Total News : 20

Start Time : 5:37:25 PM

Last Refresh Time : 6:02:51 PM

Figure: Same time references

Maintainers are psychopathic killers

- NetBeans
- GUI
- JavaDoc
- JUnit testing

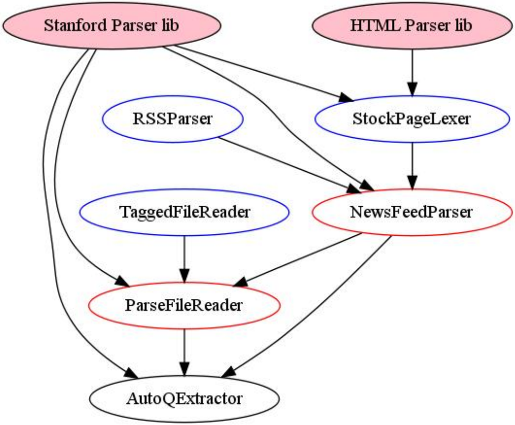


Figure: Dependency Tree of the project

Outline

Introduction

The Solution

Meanwhile . . .

The Problem

Current state

Further work

Next rung. . .

1. More data
2. Fixing temporal references, Grammatical dependency
3. HTML text extraction
4. Knowledgebase
5. Finer classification, discarded information

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

Introduction
Problem Statement
The Solution
The collection
The simplification
The aim
The manual tagging
The combination
Meanwhile . . .
The Problem
Modifications
Current state
Functioning
Code
Further work

Thank you

Utkarsh Upadhyay
utkarshu@iitk.ac.in