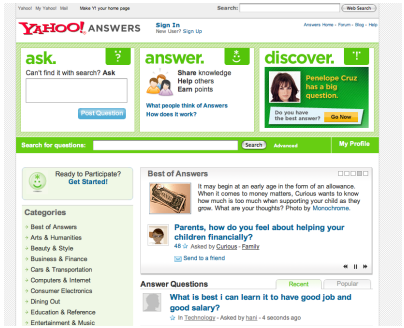
Exploring the representation gap beyond bags-of-words: a pilot study in financial news filtering

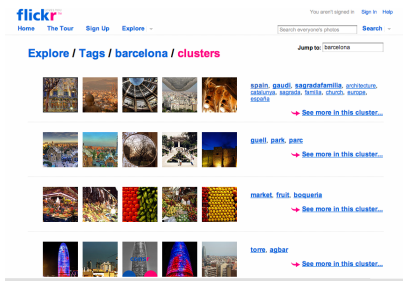
Massimiliano Ciaramita

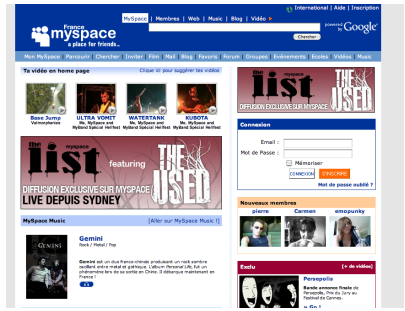
Yahoo! Research Barcelona
Joint work with Jordi Atserias and Bennett Hagedorn

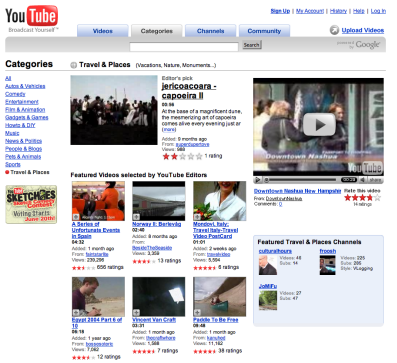
The Future of Web Search - Bertinoro 20/06/2007

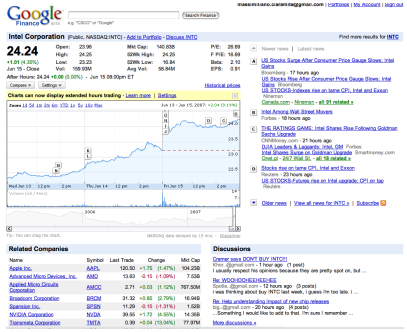
Overview





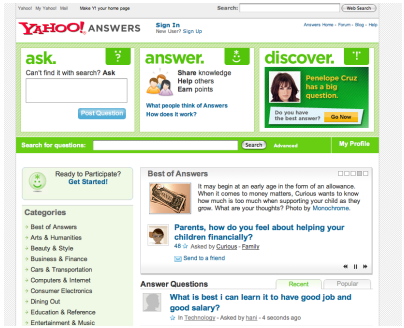


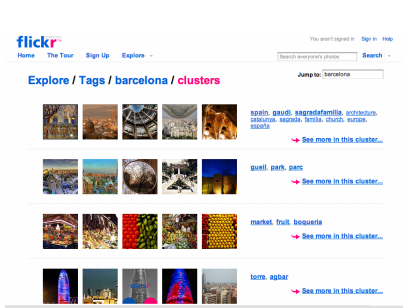


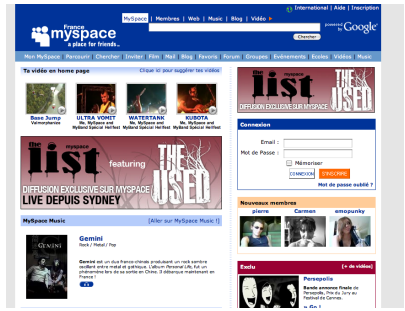


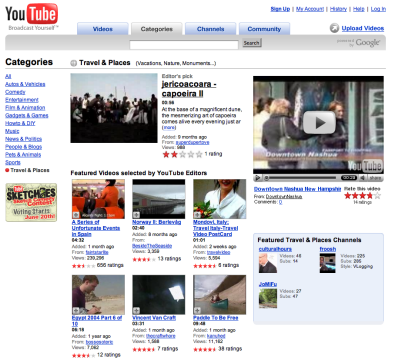
- New types of Web content and user interaction

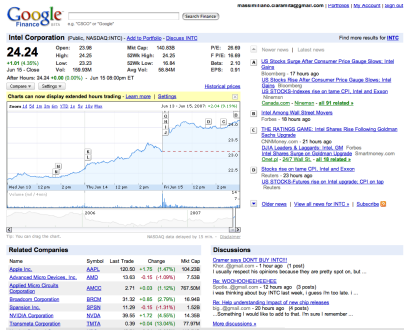
Overview





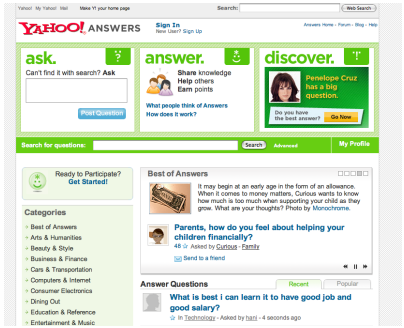


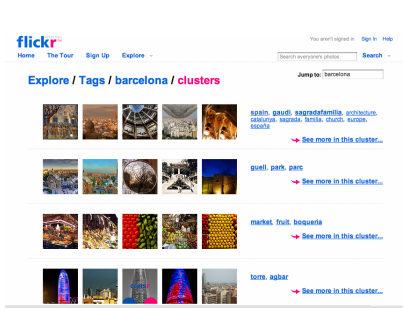


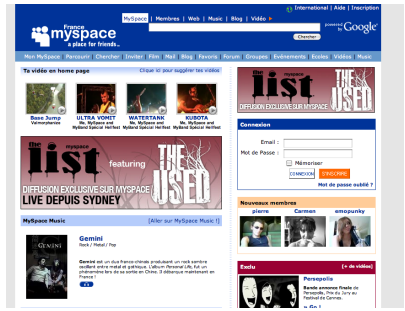


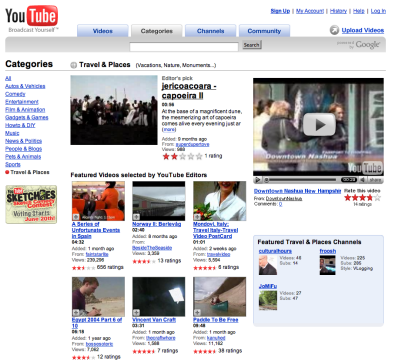
- New types of Web content and user interaction
- Challenging for traditional document-centric IR → opportunity for richer representations and methods

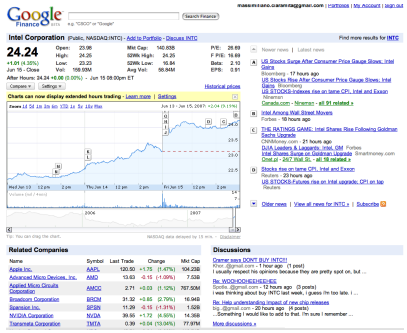
Overview



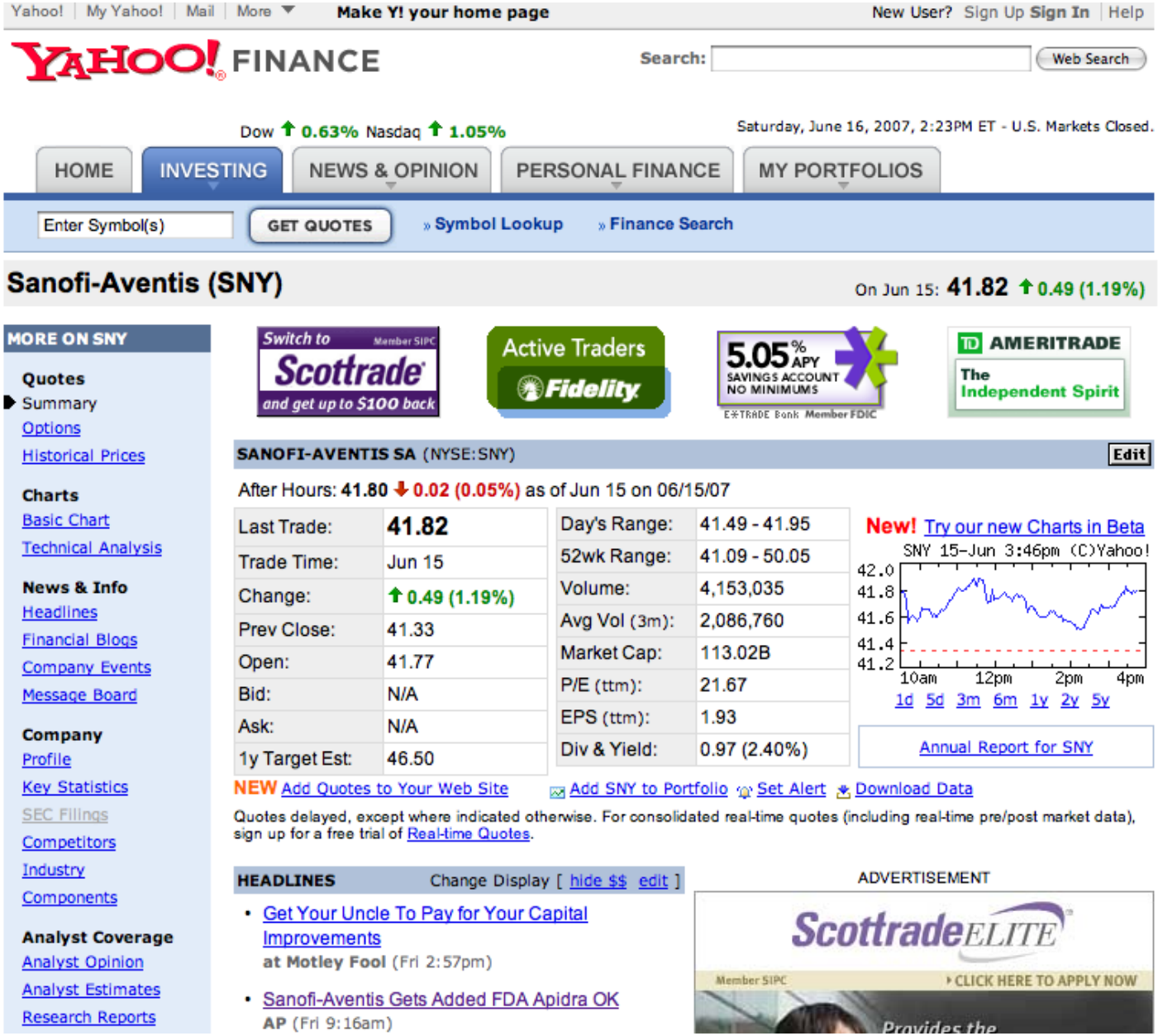








- New types of Web content and user interaction
- Challenging for traditional document-centric IR → opportunity for richer representations and methods
- A pilot study between news filtering and opinion mining to identify crucial components of content-modeling systems.



Yahoo! | My Yahoo! | Mail | More ▼ **Make Y! your home page** New User? Sign Up **Sign In** | Help

YAHOO! FINANCE Search: Web Search

Dow ↑ 0.63% Nasdaq ↑ 1.05% Saturday, June 16, 2007, 2:23PM ET - U.S. Markets Closed.

HOME | INVESTING | NEWS & OPINION | PERSONAL FINANCE | MY PORTFOLIOS

Enter Symbol(s) GET QUOTES » Symbol Lookup » Finance Search

Sanofi-Aventis (SNY)

On Jun 15: **41.82** ↑ 0.49 (1.19%)

MORE ON SNY

Quotes
- Summary
- Options
- Historical Prices

Charts
- Basic Chart
- Technical Analysis

News & Info
- Headlines
- Financial Blogs
- Company Events
- Message Board

Company
- Profile
- Key Statistics
- SEC Filings
- Competitors
- Industry
- Components

Analyst Coverage
- Analyst Opinion
- Analyst Estimates
- Research Reports

Switch to Scottrade Member SIPC and get up to $100 back

Active Traders Fidelity

5.05% APY SAVINGS ACCOUNT NO MINIMUMS E*TRADE Bank Member FDIC

TD AMERITRADE The Independent Spirit

SANOFI-AVENTIS SA (NYSE:SNY) Edit

After Hours: **41.80** ↓ 0.02 (0.05%) as of Jun 15 on 06/15/07

Last Trade:	**41.82**	Day's Range:	41.49 - 41.95
Trade Time:	Jun 15	52wk Range:	41.09 - 50.05
Change:	↑ 0.49 (1.19%)	Volume:	4,153,035
Prev Close:	41.33	Avg Vol (3m):	2,086,760
Open:	41.77	Market Cap:	113.02B
Bid:	N/A	P/E (ttm):	21.67
Ask:	N/A	EPS (ttm):	1.93
1y Target Est:	46.50	Div & Yield:	0.97 (2.40%)

New! Try our new Charts in Beta

1d 5d 3m 6m 1y 2y 5y

Annual Report for SNY

NEW Add Quotes to Your Web Site | Add SNY to Portfolio | Set Alert | Download Data

Quotes delayed, except where indicated otherwise. For consolidated real-time quotes (including real-time pre/post market data), sign up for a free trial of Real-time Quotes.

HEADLINES Change Display [hide $$ edit]

- Get Your Uncle To Pay for Your Capital Improvements
  at Motley Fool (Fri 2:57pm)
- Sanofi-Aventis Gets Added FDA Apidra OK
  AP (Fri 9:16am)

ADVERTISEMENT

ScottradeELITE Member SIPC CLICK HERE TO APPLY NOW

Provides the

HEADLINES Change Display [hide $$ edit]

- Get Your Uncle To Pay for Your Capital Improvements
  at Motley Fool (Fri 2:57pm)
- Sanofi-Aventis Gets Added FDA Apidra OK
  AP (Fri 9:16am)
- Stockpickr Lists: ExxonMobil
  at TheStreet.com (Fri 9:14am)
- [video] 5 Dumbest Things On Wall St. This Week
  at TheStreet.com (Fri 8:56am)
- Sanofi-aventis says FDA OKs new treatment option for Apidra
  at Reuters (Fri 8:25am)
- On the Move: Sanofi, Deutsche Post, Fiat
  at BusinessWeek Online (Fri 8:08am)
- FDA Approves New Route of Administration for Rapid-Acting Apidra(R) Providing an Additional Treatment Option for Hospitalized Diabetes Patients with High Blood Sugar Levels
  PR Newswire (Fri 8:00am)
- The Five Dumbest Things on Wall Street This Week
  at TheStreet.com (Fri 7:30am)
- [$$] Acomplia Setback in U.S. Adds Pressure on Sanofi
  at The Wall Street Journal Online (Thu, Jun 14)
- Sanofi's Tough Swallow
  at Forbes.com (Thu, Jun 14)

Financial news filtering and opinion mining

- News filtering: monitoring a stream of news to identify useful/interesting information

Financial news filtering and opinion mining

- News filtering: monitoring a stream of news to identify useful/interesting information
- Opinion mining: capture the "polarity" of the story: a graded positive/negative opinion with respect to a company

Financial news filtering and opinion mining

- News filtering: monitoring a stream of news to identify useful/interesting information
- Opinion mining: capture the "polarity" of the story: a graded positive/negative opinion with respect to a company
- Motivation (pragmatic): Financial news and stock prices (opinions and trends) tend to be correlated, can be modeled to a certain extent (Lavrenko et al., 2000, Das & Chen, 2001, Seo et al., 2002)

Financial news filtering and opinion mining

- News filtering: monitoring a stream of news to identify useful/interesting information
- Opinion mining: capture the "polarity" of the story: a graded positive/negative opinion with respect to a company
- Motivation (pragmatic): Financial news and stock prices (opinions and trends) tend to be correlated, can be modeled to a certain extent (Lavrenko et al., 2000, Das & Chen, 2001, Seo et al., 2002)
- Motivation (scientific): An opportunity to explore tasks which might require innovative approaches; e.g., deeper analysis on the language side

Setting: Data

- In October-November 2006, we monitored the RSS feeds from Yahoo! Finance (36 sources) for the top 50 company symbols in Standard & Poors index

Setting: Data

- In October-November 2006, we monitored the RSS feeds from Yahoo! Finance (36 sources) for the top 50 company symbols in Standard & Poors index
- We annotated the titles of 7,382 stories using five categories, same as (Seo et al., 2000):

$G^{\Uparrow}$	"GM turnaround lifts bonds to 20-month high" (GM)
$g^{\uparrow}$	"P&G sees better operating environment" (PG)
$U^{\Leftrightarrow}$	"Indonesia seeking $ 12 Billion in capital" (IBM)
$b^{\downarrow}$	"Chinese SUV maker aims to prove itself" (F)
$B^{\Downarrow}$	"AIG units subpoenaed by DOJ and SEC" (AIG)

Why titles?

- Practical advantages:
  - Suitable for manually tagging enough data for a pilot
  - Efficient processing: mixed models titles first, documents if necessary
  - Full docs: more info more noise (Pang et al. 2003)

Why titles?

- Practical advantages:
  - Suitable for manually tagging enough data for a pilot
  - Efficient processing: mixed models titles first, documents if necessary
  - Full docs: more info more noise (Pang et al. 2003)
- Scientific interest:
  - It is a perfectly natural task, even for people with generic backgrounds
  - A good example of short text analysis: SMS, QA, Dialogue, Web advertising, queries ...

Setting: Data

No pre-filtering of stories (improve recall):

- stories which did not explicitly mention the company name/symbol:
  - $<$50% of the stories mention full name/abbreviation of the company the story refers to, the rest do not mention the company name or refer to related entities
  - $\approx$ 40% of the stories which do not mention the company express polarized information.

Setting: Data

No pre-filtering of stories (improve recall):

- stories which did not explicitly mention the company name/symbol:
  - <50% of the stories mention full name/abbreviation of the company the story refers to, the rest do not mention the company name or refer to related entities
  - $\approx$ 40% of the stories which do not mention the company express polarized information.
- stories which did not explicitly mention pre-defined lists of polarized terms:
  - frequent trigger words are relatively infrequent: gain (2.2%), drop (0.9%), growth (0.8%), surge (0.3%), etc.

HEADLINES Change Display [hide $$ edit]

- Get Your Uncle To Pay for Your Capital Improvements
  at Motley Fool (Fri 2:57pm)
- Sanofi-Aventis Gets Added FDA Apidra OK
  AP (Fri 9:16am)
- Stockpickr Lists: ExxonMobil
  at TheStreet.com (Fri 9:14am)
- [video] 5 Dumbest Things On Wall St. This Week
  at TheStreet.com (Fri 8:56am)
- Sanofi-aventis says FDA OKs new treatment option for Apidra
  at Reuters (Fri 8:25am)
- On the Move: Sanofi, Deutsche Post, Fiat
  at BusinessWeek Online (Fri 8:08am)
- FDA Approves New Route of Administration for Rapid-Acting Apidra(R) Providing an Additional Treatment Option for Hospitalized Diabetes Patients with High Blood Sugar Levels
  PR Newswire (Fri 8:00am)
- The Five Dumbest Things on Wall Street This Week
  at TheStreet.com (Fri 7:30am)
- [$$] Acomplia Setback in U.S. Adds Pressure on Sanofi
  at The Wall Street Journal Online (Thu, Jun 14)
- Sanofi's Tough Swallow
  at Forbes.com (Thu, Jun 14)

Setting: Training/Evaluation Data

- Stories partitioned in train, development and test sets:

dev	1,050 titles (October '06)
train	4,513 titles for training a model (November 1-14/06)
test	1,819 from November 15 (752 titles), 16 (811), 17 (256).

Setting: Training/Evaluation Data

- Stories partitioned in train, development and test sets:

  dev 1,050 titles (October '06)

  train 4,513 titles for training a model (November 1-14/06)

  test 1,819 from November 15 (752 titles), 16 (811), 17 (256).

- Splitting by day: same story can appear several time in one day for different companies, or for the same company from different sources

Learning: model, features

- Polarity classifier: regularized multiclass perceptron

Learning: model, features

- Polarity classifier: regularized multiclass perceptron
- Features:
  - Bag of unigrams $\longrightarrow$ Uni
  - Uni plus Bag of bigrams $\longrightarrow$ +Big
  - +Big plus a feature for the company symbol $\longrightarrow$ +Co.

Learning: model, features

- Polarity classifier: regularized multiclass perceptron
- Features:
  - Bag of unigrams $\longrightarrow$ Uni
  - Uni plus Bag of bigrams $\longrightarrow$ +Big
  - +Big plus a feature for the company symbol $\longrightarrow$ +Co.
- Baseline: 3,589 out of 7,382 titles fall in category uncertain ($U^{\Leftrightarrow}$): majority category baseline is correct 48.6% of the time

Results

Model	Uni	+Big	+Co.	P-2	P-3	+15N	+16N
Score	54.6	56.9	58.4	58.2	56.5	59.7	60.5
diff/Base	6.0	8.3	9.8	9.6	7.9	11.1	11.9

- P-2/3: +Co. with polynomial kernel degree two/three
- +15N: +Co. with one additional day of titles for training (752/+16% train)
- +16N: +Co. with two additional days of titles for training (1,566/+35% train)

Results

Model	Uni	+Big	+Co.	P-2	P-3	+15N	+16N
Score	54.6	56.9	58.4	58.2	56.5	59.7	60.5
diff/Base	6.0	8.3	9.8	9.6	7.9	11.1	11.9

- P-2/3: +Co. with polynomial kernel degree two/three
- +15N: +Co. with one additional day of titles for training (752/+16% train)
- +16N: +Co. with two additional days of titles for training (1,566/+35% train)
- Room for improvement!

Mini-discussion

- Classifier: same classifier (+Bi) on TREC question classification data-set in 6 categories (Li & Roth, 2002) achieves accuracy beyond 91%

Mini-discussion

- Classifier: same classifier (+Bi) on TREC question classification data-set in 6 categories (Li & Roth, 2002) achieves accuracy beyond 91%
- Data: quality of our data can be improved (multi annotators, agreement check, etc.)

Mini-discussion

- Classifier: same classifier (+Bi) on TREC question classification data-set in 6 categories (Li & Roth, 2002) achieves accuracy beyond 91%
- Data: quality of our data can be improved (multi annotators, agreement check, etc.)
- Task-1: classifying opinion is harder than classifying by topic (Pang et al., 2003): unstructured bag of words representation is too coarse

Mini-discussion

- Classifier: same classifier (+Bi) on TREC question classification data-set in 6 categories (Li & Roth, 2002) achieves accuracy beyond 91%
- Data: quality of our data can be improved (multi annotators, agreement check, etc.)
- Task-1: classifying opinion is harder than classifying by topic (Pang et al., 2003): unstructured bag of words representation is too coarse
- Task-2:Classifying opinionated sentences is harder than documents (McDonald et al., 2007): $\approx$62.5%, and improves document classification.

Mini-discussion

- Classifier: same classifier (+Bi) on TREC question classification data-set in 6 categories (Li & Roth, 2002) achieves accuracy beyond 91%
- Data: quality of our data can be improved (multi annotators, agreement check, etc.)
- Task-1: classifying opinion is harder than classifying by topic (Pang et al., 2003): unstructured bag of words representation is too coarse
- Task-2:Classifying opinionated sentences is harder than documents (McDonald et al., 2007): $\approx$62.5%, and improves document classification.
- Mini-conclusion: subtler tasks require better models...

A closer look at the results

- Analysis by inspection of the errors of the best system: 256 titles (November 17)
- 101 errors inspected and classified according to the "nature" of the story with respect to the company, main patterns:

K	PROD	PROB	IND	COMP	ECO	PROF	?
#	32	26	17	10	7	5	4

A closer look at the results

PROD: stories about products or properties of a company:
- "Aeromexico chooses GEnx engines" (GE)

PROB: problems, lawsuits, scandals, internal changes:
- "Blog pioneer Calacanis quits AOL" (AOL)

IND: industry sector (general) news:
- "Energy sector shrugs off crude weakness" (VLO)

COMP: news involving competition issues:
- "Long lines greet PlayStation 3 debut" (MSFT)

ECO: general economy news:
- "Yen off low after Japan data" (X)

PROF: news explicitly reporting profit/losses:
- "First Solar rises after IPO" (MS)

Dependencies

Companies are correlated:

- $b^{\downarrow}$ Sanofi-Aventis shares stumble on drug's rejection [SNY]
- $B^{\Downarrow}$ Long lines greet PlayStation 3 debut" (MSFT)
- $b^{\downarrow}$ Investors yawn at IBM's faster chip [IBM]
- $b^{\downarrow}$ Lower chip forecast [INTC]

Dependencies

Companies are correlated:

- $g^{\uparrow}$ Sanofi-Aventis shares stumble on drug's rejection [PFE]
- $G^{\Uparrow}$ Long lines greet PlayStation 3 debut" (SONY)
- $g^{\uparrow}$ Investors yawn at IBM's faster chip [INTC/AMD]
- $b^{\downarrow}$ Lower chip forecast [INTC/AMD/IBM]

Syntax

Polarity of news can be encoded in fine-grained grammatical structures:

- A sues B over patent issue (B) $\longrightarrow b^{\downarrow}$

Syntax

Polarity of news can be encoded in fine-grained grammatical structures:

- A sues B over patent issue (B) $\longrightarrow b^{\downarrow}$
- A sues B over patent issue (A) $\longrightarrow g^{\uparrow}$

Paraphrases

Stories which report same event are correlated:

- 2 AIG units get subpoenas from SEC, DOJ

Paraphrases

Stories which report same event are correlated:

- 2 AIG units get subpoenas from SEC, DOJ
- AIG units subpoenaed by DOJ and SEC

Paraphrases

Stories which report same event are correlated:

- 2 AIG units get subpoenas from SEC, DOJ
- AIG units subpoenaed by DOJ and SEC
- AIG says two subsidiaries receive subpoenas

Paraphrases

- Chevron, USA Petroleum cancel gas station deal

Paraphrases

- Chevron, USA Petroleum cancel gas station deal
- Chevron ends deal to buy 122 stations from USA petroleum

Paraphrases

- Chevron, USA Petroleum cancel gas station deal
- Chevron ends deal to buy 122 stations from USA petroleum
- Chevron and USA Petroleum terminate retail gasoline station deal

Paraphrases

- Chevron, USA Petroleum cancel gas station deal
- Chevron ends deal to buy 122 stations from USA petroleum
- Chevron and USA Petroleum terminate retail gasoline station deal
- Chevron ends deal to buy Calif. stations

Discussion

- Broad-coverage opinion analysis involves a significant amount of, partly domain-independent, world-knowledge: this knowledge must be part of the system at several levels:
  - Model Structure: The topology of the model needs to mirror/exploit important correlations: companies are multi-related (competitor, allied, customer, seller, etc.) at different levels (industry, economy, geography, politics)
  - Story Dependencies: capture the correlations between stories (paraphrases)
  - Representation: extract finer-grained syntactic/semantic patterns, exploit world-knowledge to build better representation
  - Source: should be considered in the loop (analogy with user)
- Similar story might hold for other emerging tasks where content (its semantics) plays a significant role