Erasmus University Rotterdam

Lexico-Semantic Patterns for Learning Ontology Instances from Text

Frederik Hogenboom
fhogenboom@ese.eur.nl

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

February 7, 2013
1st International Symposium on Semantic Web Information Systems (IS-SWIS 2013)

ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus University Rotterdam

Introduction (1)

- Increasing amount of (digital) data
- Problem: utilizing extracted information in decision making processes becomes increasingly urgent and difficult:
  - Too much data for manual extraction
  - Yet most data is initially unstructured
  - Data often contains natural language
- Solution: automatically process and interpret information, yet automation is a non-trivial task

Introduction (2)

Erasmus University Rotterdam

- Information Extraction (IE)
  - Multiple sources:
    - News messages
    - Blogs
    - Papers
    - …
  - Text Mining (TM):
    - Natural Language Processing (NLP)
    - Statistics
    - …
  - Specific type of information that can be extracted: events

Events (1)

Erasmus University Rotterdam

Steve Jobs resigns from Apple, Cook becomes CEO

Wednesday, Silicon Valley legend Steve Jobs

Apple stock price falls on news of Ste

(The Guardian) - Apple's
Jobs returned

Google buys Motorola Mobility for $12.5B

(VentureBeat) - This morning, Google announced that it will buy Motorola Mobility — Moto's mobile device arm — for $12.5 billion. Google will acquire Motorola Mobility for $40 per share in cash, a 63 percent premium over the company's Friday closing price. Google says it will run Motorola Mobility as a separate business. Motorola spun off its business into two divisions last year, Mobility and Solutions (the data and telecom portion), as a response to declining profits.

Google shares were down around 1.5 percent, while Motorola Mobility's stock jumped 57 percent. The company says Motorola Android phones won't be receiving any special treatment as a consequence of the deal — but that's a tough nut to swallow, since Google often plays favorites.

Events (2)

Erasmus University Rotterdam

- Event:
  - Complex combination of relations linked to a set of empirical observations from texts
  - Can be defined as:
    - <subject> <predicate> e.g., <Person> <Resigns>
    - <subject> <predicate> <object> e.g., <Company> <Buys> <Company>
- Event extraction could be beneficial to IE systems:
  - Personalized news
  - Risk analysis
  - Monitoring
  - Decision making support

Events (3)

- Common event domains:
  - Medical
  - Finance
  - Politics
  - Environment

Event Extraction

Erasmus University Rotterdam

- In analogy with the classic distinction within the field of modeling, we distinguish 3 main approaches:
  - Data-driven event extraction:
    - Statistics
    - Machine learning
    - Linear algebra
    - …
  - Expert knowledge-driven event extraction:
    - Representation & exploitation of expert knowledge
    - Patterns
  - Hybrid event extraction:
    - Combine knowledge and data-driven methods
- Our focus: expert knowledge-driven event extraction through the usage of pattern languages

Existing Approaches

- Various pattern-languages for:
  - News processing frameworks (e.g., PlanetOnto)
  - General purpose frameworks (e.g., CAFETIERE, KIM, etc.)
- Language types:
  - Lexico-syntactic
  - Lexico-semantic
- However:
  - Insufficient use of semantics (if any)
  - Limited syntax
  - Cumbersome in use

Semantics

- Semantic Web:
  - Collection of technologies that express content meta-data
  - Offers means to help machines understand human-created data on the Web
- Ontologies:
  - Can be used to store domain-specific knowledge in the form of concepts (classes + instances)
  - Also contain inter-concept relations

Pattern Language (1)

Erasmus University Rotterdam

- Basic syntax:
  - LHS :- RHS
  - LHS: subject, predicate, object (optional)
  - RHS: pattern in which subject and object are assigned:
    - Literals (text strings)
    - Lexical categories (nouns, prepositions, verbs, etc.)
    - Orthographic categories (capitalization)
    - Labels (assigning subject and object)
    - Logical operators (and, or, not)
    - Repetition (≥0, ≥1, 0-1, {min,max})
    - Wildcards (skip ≥0 or exactly 1 word)
    - Ontological concepts

Pattern Language (2)

- Provocation example (lexico-syntactic):

```
($sub, kb:provokes, $obj) :-
    $sub:=(((JJ | NNS | NNP | NNPS | NN) &
            (upperInitial | allCaps | mixedCaps)
           )('.' NNP '.'?)?)+
    (!'.' & !'(' & !')' & !'-'){0,3}
    ('angers' | 'angered' | 'accuses' | 'accused' |
     'insult' | 'insulted' | 'provokes' | 'provoked' |
     'threatens' | 'threatened')
    (!'.' & !'(' & !')' & !'-'){0,3}
    $obj:=(((JJ | NNS | NNP | NNPS | NN) &
            (upperInitial | allCaps | mixedCaps)
           )('.' NNP '.'?)?)+
```

Pattern Language (3)

- Provocation example (lexico-semantic):

```
($sub, kb:provokes, $obj) :-
    $sub:=([kb:Country] | [kb:Continent] | [kb:Union])
    (!'.' & !'(' & !')' & !'-'){0,3}
    (kb:toAnger | kb:toAccuse | kb:toInsult |
     kb:toProvoke | kb:toThreaten)
    (!'.' & !'(' & !')' & !'-'){0,3}
    $obj:=([kb:Country] | [kb:Continent] | [kb:Union])
```

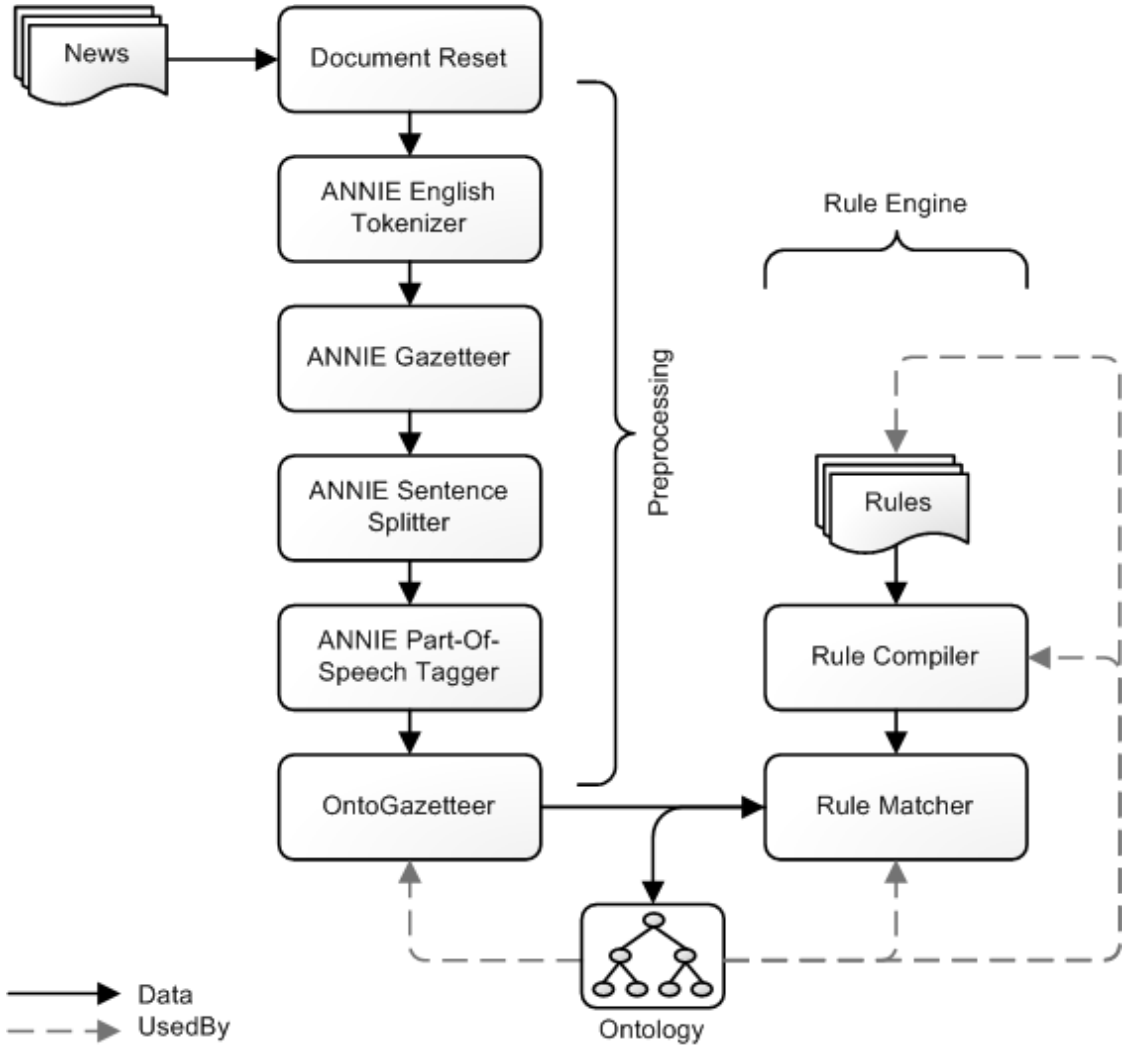
Erasmus University Rotterdam

Implementation (1)

- The Hermes News Portal (HNP) is a stand-alone Java-based news personalization tool
- We have implemented the Hermes Information Extraction Engine (HIEE) within the HNP
- Pipeline-architecture is based on GATE components

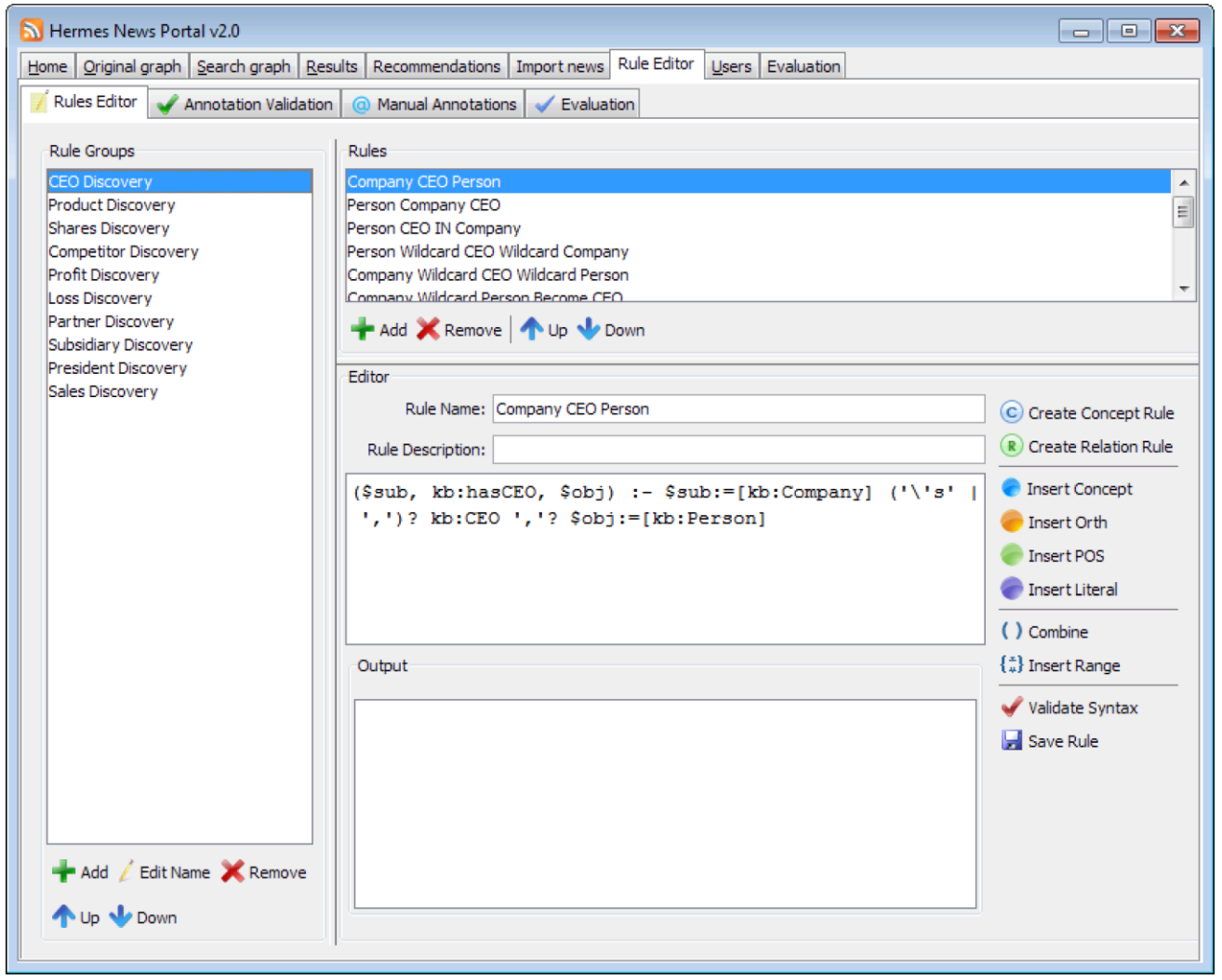
Implementation (2)

Erasmus University Rotterdam



Implementation (3)

Erasmus University Rotterdam



Implementation (4)

Erasmus University Rotterdam

Hermes News Portal v2.0

Home | Original graph | Search graph | Results | Recommendations | Import news | Rule Editor | Users | Evaluation

Rules Editor | Annotation Validation | Manual Annotations | Evaluation

Run! | Reclassify | Keep History for Evaluation | Save Annotations | Clear All Annotations

Patterns Found

- Subsidiary Discovery
- Shares Discovery
- Sales Discovery
- Profit Discovery
- Product Discovery
- President Discovery
- Partner Discovery
- Loss Discovery
- Competitor Discovery
- CEO Discovery

Subject	Relation	Object	Times Found	Valid
kb:Google	kb:hasCEO	kb:Eric_Schmidt	2	
kb:Disney	kb:hasCEO	kb:Robert_A._Iger	1	
kb:Gameloft	kb:hasCEO	kb:Michel_Guillemot	1	
kb:BP	kb:hasCEO	kb:Tony_Hayward	1	
kb:Alcatel-Lucent	kb:hasCEO	kb:Ben_Verwaayen	1	
kb:Viacom	kb:hasCEO	kb:Philippe_P._Dauman	1	
kb:NTR	kb:hasCEO	kb:Jim_Barry	1	
kb:Coke	kb:hasCEO	kb:Muhtar_Kent	1	
kb:Picarro	kb:hasCEO	kb:Woelk	1	
kb:Microsoft	kb:hasCEO	kb:Steven_A._Ballmer	1	
kb:National_Arts_Council	kb:hasCEO	kb:Benson_Puah	1	
kb:British_Airways	kb:hasCEO	kb:Willie_Walsh	1	
kb:Royal_Dutch_Shell	kb:hasCEO	kb:Peter_Voser	1	
kb:Ford	kb:hasCEO	kb:Jacques_Nasser	1	
kb:Merck	kb:hasCEO	kb:Richard_T._Clark	1	
kb:BP	kb:hasCEO	kb:Anthony_B._Hayward	1	

News Items

Past Clouds Future of Europe's New
Wed Mar 03 02:26:22 CET 2010

Past Clouds Future of Europe's New Antitrust Chief

clouding his new role, in which he must keep companies and countries from breaching rules on market-rigging and government subsidies. Mr. Almunia's predecessor as antitrust commissioner, Neelie Kroes of the Netherlands, had a reputation for bargaining long and hard with powerful people like Steven A. Ballmer , chief executive of Microsoft . At the height of the financial crisis, she insisted that rules on competition be respected as governments bailed out banks, auto companies and other businesses. Giants like the Dutch bank ING and the Royal Bank of Scotland were required to make significant divestitures. Now, the question is whether Mr. Almunia will be up to his new job's challenges, including the protectionist tendencies rising in conjunction with unemployment in many European countries and the potential for trans-Atlantic tensions, especially as companies like Google continue to expand globally. His spokeswoman, Amelia Torres, said Mr. Almunia would not grant interviews so early in his nearly five-year term as competition commissioner. Last Monday, in his first public appearance in the role, Mr. Almunia

Evaluation (1)

Erasmus University Rotterdam

- We compare the performance of lexico-syntactic rules with lexico-semantic rules on two data sets:
  - Economic events (500 news messages):
    - CEO
    - Product
    - Shares
    - Competitor
    - Profit
    - Loss
    - Partner
    - Subsidiary
    - President
    - Revenue
  - Political events (100 news messages):
    - Election
    - Visit
    - Sanction
    - Join
    - Resignation
    - Investment
    - Riots
    - Collaboration
    - Provocation
    - Help
- Performance is evaluated based on rule creation times and precision, recall, and F_1-measure.

Evaluation (2)

Erasmus University Rotterdam

- Creation times for $F_1 \geq 0.5$: 90% improvement

Name	Lex-Syn	Lex-Sem
CEO	8424	281
Product	9428	132
Shares	2403	648
Competitor	9116	133
Profit	1923	416
Loss	5991	313
Partner	4924	185
Subsidiary	6620	776
President	4239	179
Revenue	5317	498
Overall	5839	356

Name	Lex-Syn	Lex-Sem
Election	1517	232
Visit	4238	543
Sanction	4013	419
Join	3986	297
Resignation	1259	366
Investment	5162	781
Riots	1734	306
Collaboration	1103	137
Provocation	1428	530
Help	1987	211
Overall	2643	382

Evaluation (3)

- Scores after maximum time (finance)

	Lex-Syn			Lex-Sem		
Name	Precision	Recall	F_1	Precision	Recall	F_1
CEO	0.5217	0.6000	0.5581	0.8966	0.8667	0.8814
Product	0.6667	0.4118	0.5091	0.8607	0.7721	0.8140
Shares	0.4286	0.6667	0.5218	0.9000	0.8000	0.8471
Competitor	0.5333	0.4800	0.5052	0.7600	0.7600	0.7600
Profit	0.7500	0.4545	0.5660	0.8800	0.6667	0.7586
Loss	0.6471	0.4074	0.5000	0.8125	0.4815	0.6047
Partner	0.3864	0.7391	0.5075	0.8000	0.8696	0.8333
Subsidiary	0.7500	0.3913	0.5143	0.9063	0.6304	0.7436
President	0.4333	0.5909	0.5000	0.6667	0.6364	0.6512
Revenue	0.6429	0.4091	0.5000	0.7143	0.6818	0.6977
Overall	0.5492	0.4935	0.5199	0.8390	0.7414	0.7872

Evaluation (4)

Erasmus University Rotterdam

- Scores after maximum time (politics)

	Lex-Syn			Lex-Sem		
Name	Precision	Recall	F_1	Precision	Recall	F_1
Election	0.4615	0.5455	0.5000	1.0000	0.8182	0.9000
Visit	0.6774	0.4038	0.5060	0.7027	0.5000	0.5843
Sanction	0.5263	0.4918	0.5085	0.7857	0.7213	0.7521
Join	0.7222	0.4063	0.5200	0.8125	0.8125	0.8125
Resignation	0.9091	0.3846	0.5405	0.8000	0.9231	0.8572
Investment	0.5208	0.4808	0.5000	0.7778	0.6731	0.7217
Riots	0.5200	0.5200	0.5200	0.7069	0.8200	0.7593
Collaboration	0.4250	0.6538	0.5151	0.7143	0.7692	0.7407
Provocation	0.7727	0.4250	0.5484	0.7857	0.5714	0.6616
Help	0.5510	0.4737	0.5094	0.7541	0.8070	0.7797
Overall	0.5664	0.4694	0.5134	0.7630	0.7164	0.7390

Conclusions

- Compared to lexico-syntactic alternatives, our lexico-semantic patterns perform better in terms of precision, recall, F_1, and creation times
- Also, our rules are more expressive and easier to use than their lexico-syntactic and lexico-semantic alternatives
- Future work:
  - Implement language in new IE pipeline
  - Evaluate on big data set
  - Link events to trading algorithms instead of news personalization

Questions

Erasmus University Rotterdam