Erasmus University Rotterdam

Learning Semantic Information Extraction Rules from News

Frederik Hogenboom
fhogenboom@ese.eur.nl

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

In collaboration with:
Flavius Frasincar and Wouter IJntema

The Dutch-Belgian Database Day 2013 (DBDBD 2013)

Introduction (1)

Erasmus University Rotterdam

- 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

(Reuters) - On Wednesday, Silicon Valley legend **Steve Jobs resigned as chief executive of Apple Inc** in a stunning move that ended his 14-year reign at the technology giant he co-founded in a garage.

Apple shares dived as much as 7 percent in after-hours trade after the pancreatic cancer survivor and industry icon, who has been on medical leave for an undisclosed condition since January 17, announced he will be replaced by COO and longtime heir apparent Tim Cook.

Events (1)

Erasmus University Rotterdam

Steve Jobs resigns from Apple, Cook becomes CEO

Wednesday, Silicon Valley legend Steve Jobs resigned as ... move that ended his 14-year reign

Apple stock price falls on news of Steve Jobs's death

(The Guardian) - Apple's stock price has risen more than 9,000% since Steve Jobs returned in 1997, and doubled in the past two years

News of Steve Jobs's death drove the Apple share price down more than 5% in Frankfurt on Thursday morning.

Apple shares are now trading 3.5% lower at €273, after hitting a low of €270 in Frankfurt. The shares are not traded in London. They are expected to open lower when Wall Street opens at 2.30pm London time.

Apple was briefly the most valuable company in the world in the summer, knocking oil giant Exxon Mobil off the top spot. Revenues have soared from $7.1bn (£4.6bn) in 1997 to $65.2bn a year now.

Events (1)

Erasmus University Rotterdam

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:
  - Limited syntax
  - Weak semantics
  - Cumbersome in use
  - Extract entities, but not events

Semantics

Erasmus University Rotterdam

- 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

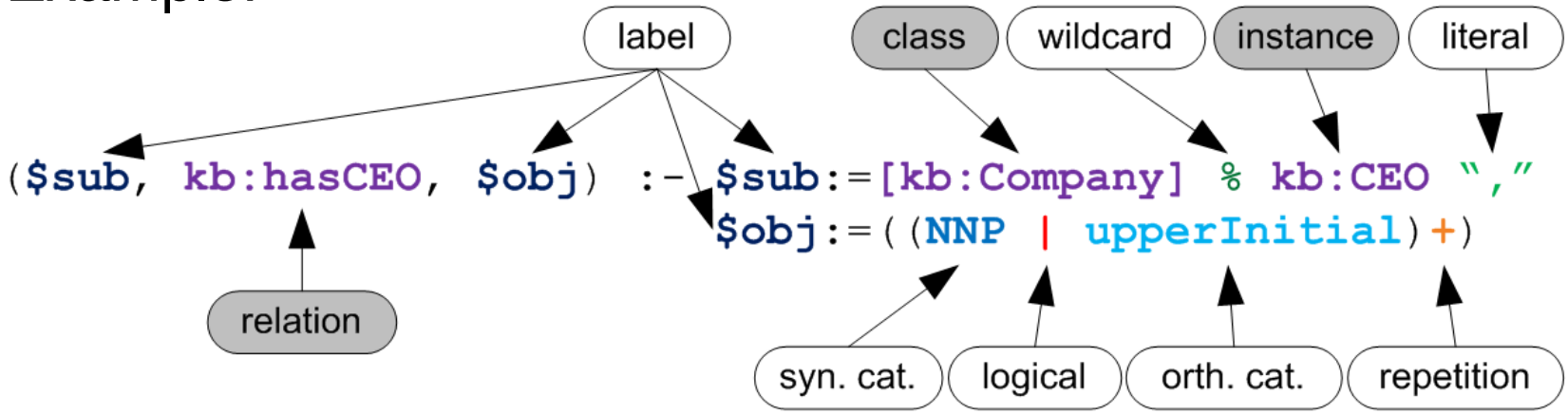
Pattern Language (1)

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)

- Example:

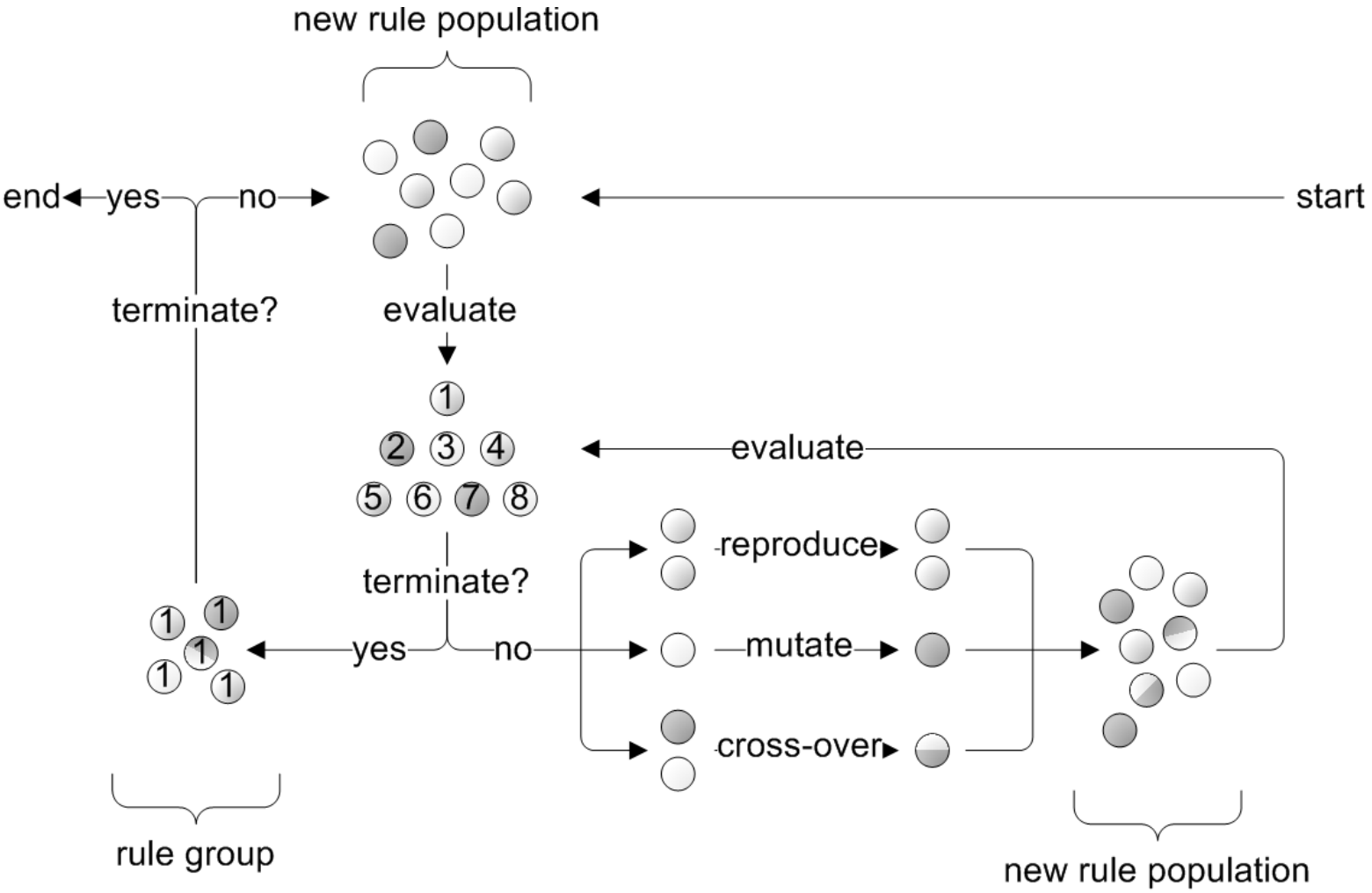


Rule Creation

- Groups of rules extract specific events
- Creating such groups is cumbersome, error-prone and time-consuming
- If the language is implemented using tree structures, a genetic programming approach can be employed for learning rules automatically

Erasmus University Rotterdam

Rule Learning



Erasmus University Rotterdam

Implementation

- 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

Evaluation (1)

- We compare the performance of rule learning versus manually creating rules:
  - Using a data set on economic events (500 news messages):
    - CEO
    - Product
    - Shares
    - Competitor
    - Profit
    - Loss
    - Partner
    - Subsidiary
    - President
    - Revenue
  - By allowing for 5 hours of construction time per rule group (including reading, thinking, writing, …)
  - Based on the Precision, Recall, and F_1-measure

Evaluation (2)

Erasmus University Rotterdam

	Automatic Learning			Manual Creation			
Name	Precision	Recall	F_1	Precision	Recall	F_1	Δ%
Competitor	0.667	0.508	0.577	0.875	0.280	0.424	36.0%
Loss	0.905	0.613	0.731	0.818	0.333	0.474	54.3%
Partner	0.808	0.356	0.494	0.450	0.391	0.419	18.0%
Subsidiary	0.698	0.309	0.429	0.611	0.239	0.344	24.8%
CEO	0.904	0.904	0.904	0.824	0.700	0.757	19.5%
President	0.821	0.793	0.807	0.833	0.455	0.588	37.2%
Product	0.788	0.793	0.791	0.862	0.596	0.704	12.3%
Profit	0.960	0.522	0.676	1.000	0.273	0.429	57.7%
Sales	0.900	0.450	0.600	0.455	0.455	0.455	32.0%
ShareValue	0.939	0.805	0.867	0.530	0.778	0.631	37.5%
Total	0.839	0.605	0.703	0.726	0.450	0.555	26.6%

Conclusions

- We presented HIEL, a lexico-semantic rule language for event extraction
- Rule creation is cumbersome, and hence a genetic programming-based learning approach is proposed
- Lexico-semantic rule learning performs better than the manual alternative in terms of precision, recall, and F_1
- Future work:
  - Evaluate approach for existing lexico-semantic languages
  - Evaluate on other domains
  - Link events to trading algorithms instead of news personalization

Questions

Erasmus University Rotterdam