Introduction to AI:

Topics in natural language processing

Roman Yangarber
University of Helsinki
Spring 2017
Lecture 2

- **Motivation: high-level application**
  - Introduction to Information Extraction
  - Finding facts from text

- **Requires many lower-level NLP components**
Demos at the end of lecture

- **PULS**
  - [newsweb.cs.helsinki.fi](http://newsweb.cs.helsinki.fi) (login via LinkedIn)
  - Project focus: Web-scale surveillance of news
  - Current topic: business (other topics: epidemics, security, etc.)
  - Tracking news from thousands of news sites about business activities
  - Among hundreds of thousands of entities:
    - Companies, persons, etc.
  - Tracking many kinds of activities:
    - Mergers, buyouts, bankruptcies, layoffs, product launches and recalls, etc.
Demos at the end of lecture

- **PULS**
  - Research challenges:
  - need methods for finding useful patterns in the data:
    - is an entity getting attention from many sources?
    - is it positive or negative attention?
    - what entities are similar to given entity, and how?
    - ...
Demos at the end of lecture

- Revita
  - revita.cs.helsinki.fi
  - Academy Project: original plan:
    - revitalization of endangered languages
    - focus on Finno-Ugric languages (related to Finnish)
  - new broader focus:
    - language learning (also some non-endangered languages, including German, Swedish, Russian, etc.)
  - fully automatic: infinite variety of exercises, based on user’s own texts—to assure the user is interested
Demos at the end of lecture

- Revita
  - Research challenges:
    - assessment of user's level of competence
    - very important, since the system should simulate a good teacher:
      - examples must not be too difficult (else learner becomes discouraged)
      - nor too simple (else learner becomes bored)
NLP vs. CL dichotomy

- Much information exists that is “trapped” in text
  - Task: “release” the information from the trap
  - natural language processing

- Language is very good at conveying/encoding information
  - How does language do it?
  - What mechanisms in language allow it to convey so much information so well?
  - computational linguistics
Goals

- Get information out of text
  - How?
- Using IR/keyword search alone is not sufficient
- Main goal = understanding text
  - Can support many high-level applications
- Understanding is hard
  - Because (in part) of ambiguity in language
  - Will require many lower-level components
Applications

- Information extraction
- Question answering
- Text summarization
- Machine translation
Ambiguity

- Morphological ambiguity
- (Phonetic ambiguity)
- Lexical ambiguity
- Word-sense ambiguity
- Syntactic ambiguity
- Semantic ambiguity (quantifier scope...)
- Lexical + syntactic ambiguity
- Reference resolution as a kind of “ambiguity” resolution
  - in discourse
- Discourse/pragmatic ambiguity
Resolving Ambiguity

- In many cases, to resolve ambiguity, we will need to develop methods for reasoning about which of several alternatives or which interpretation is more plausible/probable
  - Develop systems of rules to guide the decisions
  - Use language of statistical/probabilistic reasoning
    - “more likely”, “less likely”, “impossible”, etc.
    - (**Especially language of conditional probability \( \rightarrow \) ...)
  - To make decisions, will use information from context

- Examples:
  - I like your work. The man sued the bank\(^1\)
  - Eye lie cure work. The man fished on the bank\(^2\)
Theoretical point of view

- Levels of language & of text analysis:
  - Phonology
    - How sounds are put together
  - Morphology
    - How words are put together
  - Syntax
    - How sentences are put together
  - Semantics
    - Meaning of words and sentences
  - Discourse
    - How texts are put together, meaning of text/document
Levels of analysis

• On the way to understanding (for text:)
  – Sentence splitting
  – Word segmentation
  – Morphological analysis
  – Part-of-speech tagging
  – Syntactic analysis: partial (shallow) or full parsing
  – Reference resolution
  – Discourse analysis

• Need to incorporate pragmatic world knowledge
  … (AI)
Understanding is hard. However...

- The engineering point of view: ... top-down
- We may be able to do well on some applications even *without* full understanding
- “Partial” understanding (and partial analysis!) may be enough

**Compromise:**
- Succeeding *without full* understanding
- Lower-level processing may support applications to a *sufficient* degree
Evaluation

- Key ideas:
  - For many tasks can tolerate some imprecision
    - (A measure of urgency!!)
  - For many tasks/applications, no single, “correct” answer
- Evaluation not straightforward, becomes especially important
  - Developing guidelines, resources, “gold standards”
  - Competitions
Finding facts: Information extraction
Finding facts: Information extraction

- **Introduction: the task and general principles**
- **How is it done:**
  - Fundamental techniques
  - Overall architecture
  - Back-end IE engine
- **Assignment**
  - Annotate documents and cross-validate
- **Demonstration**
  - Front-end: browse database
  - Front-end: customization environment
- **Project**
Finding facts: Information extraction

- Process arbitrary free, unstructured text, e.g.,
  - newspaper/magazine articles
  - transcribed TV/radio broadcasts
  - web pages, etc.
- Find specific facts reported in the text,
  - e.g., Corporate Acquisitions and Mergers
- Reduce discovered facts to structured form
  - such as table (in database)
  - spreadsheet
Information extraction

- More tractable task than **full understanding**:
  - Closed domain!
  - Know in advance what type of facts we are looking for
- More restricting/limiting
- But still quite realistic:
  - Typical of *analytic* work
- Easier to evaluate (though still not easy!)

- Demo:
- puls.cs.helsinki.fi/jrc
Information Extraction: introduction

- Objectives of IE
- Definition of key terms/concepts
- Examples of tasks

- Architecture overview
- Demonstrations

- Performance and evaluation
- Customization
- Advanced:
  - Learning for customization
Foundation

- Information Explosion
- Preponderance of Text
- Importance of Focused Search
Information Explosion

- Size of digital archives growing \textit{hyper}-exponentially
- Estimate that information on Internet is currently doubling every 8 months
- Successful management of information requires the ability to locate all \textit{relevant} information correctly and efficiently
Preponderance of Text

- Estimated over 80% of available information is in the form of **natural language text**: 
  - on the Internet
  - in electronic archives
  - in electronic newspapers and magazines
  - from news agencies and other sources
- This is information in natural language text, unstructured, as opposed to tabular data – stock prices, financial statements, game scores, ...
Importance of **Focused** Search

- Contrast to *spontaneous, random* search
- Users of textual information spend a significant amount of time on *persistent, focused* search—repeated pursuit of particular pieces of information that is important in their analysis/research
- User places higher *value* on information relating to long-standing interest, to which he has a long-standing commitment, than on info relating to one-time interest
Objective of IE

- **Text → Table (as in a database)**
  - From plain NL text to structured representation
- **Each record = event / fact**
- **Relations with multiple arguments**
- **Always a given, fixed, pre-determined task**
- **Thoroughly researched**
  - in MUC program (late 80's - 90's)
  - in ACE (1999-present)
IE: Definition

- Extract meaning from free text in a specific subject domain.
- “Meaning” is defined in terms of a set of specified types of objects
  - **Entities**: persons, locations, artifacts, organizations, ...
  - **Relationships**: organization/location, parent/subsidiary, ...
  - **Events**: acquisition: buyer/seller/company/price
- Use this information to fill in a data base (template)
Example IE Task

- **Domain:** Business News
- **Scenario:** Management Succession
  - Tested in MUC-6, 1996
- **Entities:** Person, Company, Location
- **Events:** Hirings and Firings of High-level Corporate Executives
Example: Executive Search

- George Garrick, 60 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.
Example: Executive Search

- George Garrick, 60 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.
Scenario Patterns

- \( \text{np(Company)} \text{ vg(V-Appoint) np(C-Person) as? np(Post)} \)
  - matches text:
  ```
  \`
  Nielsen has appointed George Garrick CEO.```

- Pattern = trigger + action

- Epidemic domain:
  - \( \text{np(Disease)} \text{ vg(V-Afflict) np(C-Person)} \)
    - matches text:
    ```
    \`
    [...since Tuesday] Ebola has killed 5 people [in Uganda...]
    ```

**Succession:**
- Company
- Post
- Person
- Status
- Date
Epidemics scenario

- **Domain:** epidemiological reports
  - Plain-text documents
  - Sources: ProMED mailing list

- **Extract facts**

- **An incident:** atomic event = database record
  - *Disease*
  - *Location*
  - *Date*
  - *Victims*
    - Descriptor: \{people | animals | plants\}
    - Number
    - Status: \{affected | dead\}
    - ...
Example: Epidemics

Viet Nam: 2 additional deaths confirmed; total now 50

Asia's [human] death toll from avian influenza rose to 50 on Wed 6 Apr 2005, when Vietnamese health officials and a hospital doctor confirmed 2 additional deaths in Viet Nam. A 10-year-old girl, who tested positive for the H5N1 virus, died of lung failure hours after she was admitted to St. Paul's Hospital in Hanoi on 27 Mar 2005, a hospital doctor said on

Rule/Pattern: * confirm N death [in Loc]
Terminology

- Domain
- Scenario
- Event / Fact
- Template / schema
- Slot / Field

- Focusing on scenario makes IE task easier than general text understanding
- Makes evaluation possible
Demo systems

- Management succession (MUC-6)
- Corporate Acquisitions
- Mergers

- Epidemic Outbreaks
System

Volume of data:

<table>
<thead>
<tr>
<th>Source</th>
<th>Documents</th>
<th>Events</th>
<th>Outbreaks</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web/JRC</td>
<td>291 216</td>
<td>104 064 (22 155)</td>
<td>2 296</td>
<td>2006—</td>
</tr>
<tr>
<td>ProMED-Mail</td>
<td>29 713</td>
<td>58 095 (12 074)</td>
<td>5 283</td>
<td>1998—</td>
</tr>
</tbody>
</table>

Performance
- 71.16F = 67% recall, 74% precision

Room for improvement

Extracted data is imperfect

Input data is highly redundant
- News events evolve
IE and IR

IR: keywords
“Hire / fire / executive…”

IE
“Who works where?”

Additional processing

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>European Information Services Inc.</td>
<td>London</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
<tr>
<td>CEO</td>
<td>Nielsen Marketing Research</td>
<td>USA</td>
<td>George Garrick</td>
<td>In</td>
</tr>
<tr>
<td>VP</td>
<td>Sony</td>
<td>Japan</td>
<td>John Smith</td>
<td>In</td>
</tr>
<tr>
<td>CFO</td>
<td>Hewlett-Packard</td>
<td>USA</td>
<td>Bill Jones</td>
<td>Out</td>
</tr>
</tbody>
</table>
IE vs Full Text Understanding

- IE is a more **focused** task than full text understanding
- We restrict in advance the type of information we want to get out of the text
- Makes the task much more **tractable**
- Makes **evaluation** possible
  - in terms of Recall/Precision measures
  - MUC
MUC and ACE

- US Government has conducted a series of Message Understanding Conferences to evaluate and promote research

- Participants prepare an extraction system according to task specifications, and are then evaluated on their system’s performance on a test corpus
MUC topics

1. 1987: Messages about naval operations
2. 1989:"
3. 1991: Terrorism in Latin America
4. 1992:"
5. 1993: Corporate Joint Ventures
6. 1995: Corporate Management Succession
7. 1998: Rocket/Missile Launches
8. 1999: ACE: names and (binary) relations
9. ...
Performance evaluation

- Borrow ideas from IR
  - Recall
  - Precision

- Tools: MUC scoring

- Not always well-defined
  - Kehler, Bear & Appelt, j. Comp. Ling, June 2001
Recall & Precision

- Recall: proportion of correct items the system found from all correct items
- Precision: proportion of correct items the system found relative to all answers it found
- \( \text{Rec} = \frac{\text{cor}}{\text{cor} + \text{mis}} \)
- \( \text{Pre} = \frac{\text{cor}}{\text{cor} + \text{spu}} \)
Scoring a response

System performance can be scored by comparing the system response to the answer key.

The most common metrics are:

\[
\text{recall} = \frac{\text{number of slots correctly filled}}{\text{number of filled slots in key}}
\]

\[
\text{precision} = \frac{\text{number of slots correctly filled}}{\text{number of slots filled by system}}
\]

\[
F = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}
\]
Lecture 2b

- Continuation: high-level application
  - Introduction to Information Extraction
  - Finding facts from text

- Requires many lower-level NLP components

- Assignment
- Project
Why the Interest in IE?

- the interest reflects an overlap of need and ability

- like much natural language engineering, it reflects a realization that we can do a great deal with a limited technology
Information Extraction: Need

- Too much text to read and analyze
- Information retrieval, at best, finds relevant documents or passages
  - must still read each one to get information
- Information extraction can obtain facts
  - (from newspaper) who was hired?
  - (from newspaper) which company sold which product?
  - (from scientific journals) which compounds were synthesized?
  - (from hospital reports) what were the symptoms and diagnosis?
Information Extraction: Ability

- Simple text analysis methods can be quite effective for information extraction
  - although they will miss some of the facts
  - technology / performance slowly improving
- Less-than-perfect extraction can be useful in some applications
  - can extract information for manual verification
  - in situations where we don’t have time to read the text, some facts are better than no facts
  (analogy to situation in MT)
How is IE done?

- Combine lower-level, better-understood technologies
Task Specification

Step 1: “user” should provide

- Template structure
  - set of slots to be filled in database
- Fill rules/guidelines
  - narrative description of what constitutes an interesting event, and what goes into each slot
- Training corpus
  - set of sample texts, with hand-filled answers
How are IE Systems built?

Basic idea:
write patterns to match events of interest
___ Corp. fired Mr. ___ ___

But ...
- it is hard to state patterns at level of words
- may need to combine facts from several sentences

So ...
- combine multiple stages of text analysis before event pattern matching
- need to analyze discourse to combine facts
Customization environment

Knowledge bases: Lexicon/Ontology Patterns
(Inference rules)

Core IE Engine

Extracted facts

User query
Response

Extracted facts

DB server

Web server

Text documents

Data collection

Data correction
Noise reduction
Cross-validation
Trend detection

Un/Supervised learning

Other corpora

User

Candidates knowledge

Publishers

Core IE Engine

(1)

Knowledge bases: Lexicon/Ontology Patterns
(Inference rules)

(2)

Customization environment

(3a)

Un/Supervised learning

(3b)

Trend detection
Cross-validation
Noise reduction
Data correction

(4a)

Web server

(4b)

DB server

(4c)

Data collection

Text documents

Other corpora

Publishers
How is IE done?

- **IE system structure overview**
  - Cascade of modules
  - Knowledge bases
  - Knowledge is stratified across several levels
    - much of it task-specific

- **Stratification / modularization allows customization across scenarios**
Core Engine Architecture

Lexical Analysis

Name Recognition

Partial Syntax

Scenario Patterns

Reference Resolution

Discourse Analyzer

Extracted Information

Output Generation

Input Text

sentence
discourse
IE Pipeline & knowledge bases

Input text

Lexical Analysis (1a)

Name Tagging (1b)

Partial Syntax (1c)

Pattern Matching (1d)

Reference Resolution (1e)

Discourse Analysis (1f)

Extracted facts

Lexicon

Semantic Ontology

Pattern Base

Inference Rules

IE Pipeline & knowledge bases

Extracted facts
Walk-through Example

- Extraction task: “management succession”
  - identifying people who were hired/fired by companies

- Sample text:

  Next week, Henry Jones will retire as executive vice president of the famous fast food restaurant, Hottest Burgers Inc.
Stage 1: Dictionary Look-up

- Use a broad-coverage English dictionary (e.g., Comlex Syntax)
- + specialized dictionaries for:
  - company names
  - government organizations
  - locations (countries, states, major cities)
  - common first names
  - scenario-specific terms ...
Next week, **Henry** Jones will retire as *first name* executive vice president of the famous *fast food* restaurant, **Hottest Burgers Inc.** *noun (idiom)* *company suffix*
Pattern Matching

- Subsequent stages of text analysis use regular expression pattern matching (in effect, bottom-up deterministic analysis, or FSA)

- Simple and quite fast
  - process ~1 document per second on a PC, with Lisp
Stage 2: Name recognition

- Recognizes named entities
  - person names,
  - organization names,
  - locations,
  - dates,
  - currencies,
  - percentages,
  - etc.

- Relates different forms of a name: aliasing
  - Sam Schwartz and Mr. Schwartz
  - International Business Machines and IBM
Next week, Henry Jones will retire as executive vice president of the famous fast food restaurant, Hottest Burgers Inc.
Proper Name Classification

- **Person**

- **Location**

- **Organization**
  - “IBM”, “Sony, Ltd.”, “Calvin Klein &Co”, “Calvin Klein”

- **Products/Artifacts/Works of Art**
  - “DC- 10”, “SCUD”, “Barbie”, “Barney”, “Gone with the Wind”, “Mona Lisa”

- **Other groups**

- **Laws, Regulations, Legal Cases**

- **Major Events, political, meteorological, etc.**
  - “Hurricane George”, “El Niño”, “Million Man March”, “Great Depression”
Stage 3: Noun and Verb Groups

- recognizes nouns with their left modifiers
  - “the manufacturer”
  - “the famous manufacturer”
  - “the electronics manufacturer”
  (decisions can be made locally based on syntactic information alone, no semantics needed, yet)

- recognizes tensed verb groups, active and passive, with and without auxiliaries
  - “retired”
  - “will retire”
  - “has been fired”

- domain independent
Noun and Verb Groups

- Next week, Henry Jones will retire as executive vice president of the famous fast food restaurant, Hottest Burgers Inc.
Stage 4: Noun Phrase Building

- **selectively** build larger noun phrases which satisfy **domain-specific semantic** constraints, such as:
  - `person, age, \rightarrow person`
    - John Smith, 42,
  - `post of company \rightarrow post`
    - CEO of Sony, Inc.
  - `company-name, company-description, \rightarrow company`
    - Ben & Jerry, the deluxe ice-cream maker,
  - ...
Next week, Henry Jones will retire as executive vice president of the famous fast food restaurant, Hottest Burgers Inc.

,np(post)
post( executive- VP, company(HBI) )
Stage 5: Event Recognition

- final set of patterns recognize events to be extracted
  
  person retires as position
  company appoints person as position
  person succeeds person as position

- separate patterns for syntactic variants of patterns
  - active clauses: the cat ate the mouse
  - Passives: the mouse was eaten by the cat
  - relative clauses: the cat that ate the mouse
  - relative passives: the mouse that was eaten by the cat
  - reduced relatives: the mouse eaten by the cat
  - conjoined verb phrases: the cat jumped up and ate the mouse
  - nominalized forms: the consumption of the mouse by the cat

- create “event” structures for subsequent processing
Event Recognition

- Next week, Henry Jones will retire as executive vice president of the famous fast food restaurant, Hottest Burgers Inc.

(event:
  retire(person(Henry Jones),
  post(executive vice president,
  company(Hottest Burgers)))
Stage 6: Reference Resolution

- resolves anaphoric expressions against prior discourse

Frank Smith retired yesterday. He was executive vice president of ... 

- uses ontology = semantic classification hierarchy to test for compatible antecedents

Sony announced that Hideo Tanaka was joining the firm as vice president for information technology.
George Garrick, 40 years old, has served as president of Sony, Inc. for 13 years.

The company announced his resignation effective October.

[Date: June, 23, 2000]
Discourse-Level Analysis:

- Often information/facts from several sentences must be combined:
  
  Mr. Tokugawa resigned last week as vice president of Hitachi.
  
  ...  
  
  He will be succeeded by Ms. Nomura.

- apply inference rule
  
  - had-job(X, Job) & succeed(Y, X)  
    
    ➔ have-job(Y, Job)

- to infer
  
  - have-job(Ms. Nomura, vice president of Hitachi)
Example: Management Succession

- George Garrick, 40 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>European Information Services, Inc.</td>
<td>London</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
<tr>
<td>CEO</td>
<td>Nielsen Marketing Research</td>
<td>USA</td>
<td>George Garrick</td>
<td>In</td>
</tr>
</tbody>
</table>
MUC/Tipster Template Format

<TEMPLATE-20040629.0071> :=
  <SUCCESSION_EVENT-20040629.0071-1>
  <SUCCESSION_EVENT-20040629.0071-2>

...

<SUCCESSION_EVENT-20040629.0071-2> :=
  ORG_NAME: "European Information Services"
  PER_NAME: "George Garrick"
  POST: "president"
  NEW_STATUS: OUT
  ORG_LOCALE: London
Problems

- Performance
- Customization
- Where does the knowledge come from?
Customization

- Need to be adapted for every new domain of interest
- Because of modular system structure, customization entails collection of knowledge

- Factors in evaluation of system:
  - Performance
  - Flexibility/Customizability
Problems: Customization

To customize a system for a new extraction task, we have to develop

- (domain-specific) lexicons
- word classes for the domain
- new patterns for new types of events
- inference rules
- etc.

This can be a big job requiring skilled labor

- expense of customization limits uses of extraction
Phase I

- Manual customization tools
  - PET: PULS Extraction Toolkit
Lexical Analysis → Name Recognition → Partial Syntax → Scenario Patterns

Pattern Base → Scenario Patterns

Reference Resolution → Discourse Analyzer

Template Format → Output Generation

Lexicon → Semantic Concept Hierarchy

Inference Rules
Customization Environment

Lexicon Editor (2a)

Concept Editor (2b)

Template Editor (2c)

Pattern Editor (2d)

Lexicon

Semantic Concept Ontology

Template Ontology

Pattern Base

Pattern Editor (2d)

“Gold Standard” Corpus

Evaluation & scoring (2e)

Document Browser
Semantic patterns

;;; For <company> appoints <person> <position>

(defpattern appoint
  "np-sem(C-company)? rn? sa? vg(C-appoint) np-sem(C-person) ',,'?
  to-be? np(C-position) to-succeed?:
  company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes,
  position-at=8.attributes |
  np-sem(C-company)? rn? sa? vg(C-appoint) np-sem(C-person) ',,'?
  to-succeed as-position:
  company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes")
(whenpattern appoint (when-appoint 's))

(defun when-appoint (phrase-type)
  (let ((person-at (binding 'person-at))
    (company-entity (entity-bound 'company-at))
    (person-entity (essential-entity-bound 'person-at 'C-person))
    (position-entity (entity-bound 'position-at))
    (predecessor-entity (entity-bound 'predecessor-at))
    new-event)
    (not-an-antecedent position-entity) ...
Customization environment

- **User:**
  - Example of text that contains events
  - Correspondence to output structure

- **System:**
  - Automatically builds pattern
    - Trigger
    - Action
  - **Syntactic and semantic generalization**
Automatic knowledge discovery

- Machine learning techniques for acquiring
  - Patterns
  - Names and concept classes

- (Advanced topics)
What we will need, to build system

- **Fundamental analysis**
  - Sentence, word splitting/segmentation
  - Dictionary lookup / morphological analysis
  - Part of speech tagging (POS disambiguation)
  - Name recognition and classification
  - Syntactic analysis: shallow (or full) parsing
  - Reference resolution / discourse analysis

- **Can be rule-based or statistical**

- **Advanced methods (corpus-based, data-driven):**
  - Machine learning for automated customization
  - Discovering patterns
  - Discovering names and concept classes
Applications
Filling non-explicit slots

Slots not explicitly mentioned in text: e.g.

- Industry sectors
- Event types
- Relevance of event (to scenario)
- Confidence
- ...

Learn to extract values for slots based on features of the text
Example: industry sector

Many industry classifications (Reuters, ...)

- Multi-level hierarchical classification system
- ~800 lower-level sectors
  - 500 with 100+ documents (in last 3 years)
  - 300 with 600+ documents
- 40 top-level sectors

Features

- Keywords found in text
  - bigrams, ...
- Entity-names: company names
- Special NPs: company descriptors
Example: industry sector

Features

- **Entity-names: company names**
  - Classification by rote

- **Special NPs: company descriptors**
  - Laptop manufacturer → electronics
  - Bank → finance: banking
  - Brewery → alcoholic beverages

- **Keywords (and bigrams)**
  - Chair, bed, table → furniture
  - “Credit card” → finance: credit
  - “SIM card” → electronics
  - Oil → energy: petroleum | cosmetics | food
    - Body oil → cosmetics
    - Olive oil → food
Approach

Multi-class, multi-label problem

Map to: Array of binary classifiers—one per sector
- Rote
- Statistical: Naïve bayes, SVM, ...
- Learn thresholds for positive vs. negative class
- Each example classified by 300 classifiers
  - Compute TP, TN, FP, FN, F-measure, accuracy,...
## Sector classification

<table>
<thead>
<tr>
<th>Classifier</th>
<th>M-average</th>
<th></th>
<th></th>
<th></th>
<th>(\mu)-average</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rec</td>
<td>Pre</td>
<td>F1</td>
<td>Rec</td>
<td>Pre</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td><strong>Statistical classifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB+IG</td>
<td>31.3±0.9</td>
<td>21.9±0.6</td>
<td>19.7±0.6</td>
<td>31.5±0.5</td>
<td>22.4±0.6</td>
<td>26.2±0.5</td>
<td></td>
</tr>
<tr>
<td>NB+BNS</td>
<td>34.2±1.1</td>
<td>16.6±0.6</td>
<td>15.8±0.5</td>
<td>33.1±0.7</td>
<td>13.4±0.4</td>
<td>19.0±0.5</td>
<td></td>
</tr>
<tr>
<td>SVM+IG</td>
<td>31.9±1.3</td>
<td>59.2±1.1</td>
<td>37.1±1.2</td>
<td>30.5±0.4</td>
<td>72.7±0.6</td>
<td>42.9±0.4</td>
<td></td>
</tr>
<tr>
<td>SVM+BNS</td>
<td>32.7±0.9</td>
<td>55.2±1.0</td>
<td>36.2±0.7</td>
<td>30.1±0.5</td>
<td>70.8±0.6</td>
<td>42.2±0.5</td>
<td></td>
</tr>
<tr>
<td><strong>Rote classifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>36.8±0.8</td>
<td>65.2±1.0</td>
<td>44.5±0.7</td>
<td>45.9±0.5</td>
<td>60.5±0.4</td>
<td>52.2±0.5</td>
<td></td>
</tr>
<tr>
<td>descriptor</td>
<td>8.8±0.3</td>
<td>38.4±1.2</td>
<td>11.6±0.3</td>
<td>16.4±0.2</td>
<td>29.0±0.3</td>
<td>20.9±0.4</td>
<td></td>
</tr>
<tr>
<td>name+desc</td>
<td>39.4±0.8</td>
<td>63.3±0.7</td>
<td>46.2±0.7</td>
<td>48.5±0.5</td>
<td>57.8±0.5</td>
<td>52.8±0.4</td>
<td></td>
</tr>
<tr>
<td>name→desc</td>
<td>11.9±0.2</td>
<td>48.0±0.9</td>
<td>16.0±0.3</td>
<td>20.6±0.4</td>
<td>39.0±0.4</td>
<td>27.0±0.4</td>
<td></td>
</tr>
<tr>
<td>name+name→desc</td>
<td>39.2±0.8</td>
<td>60.0±0.8</td>
<td>44.8±0.6</td>
<td>48.5±0.5</td>
<td>54.5±0.4</td>
<td>51.3±0.4</td>
<td></td>
</tr>
<tr>
<td><strong>Combined classifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>name→SVM+IG</td>
<td>46.2±1.0</td>
<td>73.7±0.8</td>
<td>55.1±0.8</td>
<td>52.5±0.5</td>
<td>75.9±0.4</td>
<td>62.0±0.4</td>
<td></td>
</tr>
<tr>
<td>SVM+IG→name</td>
<td>47.0±1.2</td>
<td>67.7±0.9</td>
<td>53.7±1.1</td>
<td>49.9±0.3</td>
<td>73.9±0.3</td>
<td>59.6±0.3</td>
<td></td>
</tr>
<tr>
<td>name ∪ SVM+IG</td>
<td>52.2±1.1</td>
<td>66.3±0.8</td>
<td>56.9±0.9</td>
<td>57.7±0.4</td>
<td>71.1±0.3</td>
<td>63.7±0.4</td>
<td></td>
</tr>
<tr>
<td>name+desc→SVM+IG</td>
<td>48.4±1.1</td>
<td>69.2±0.7</td>
<td>55.5±0.9</td>
<td>56.2±0.5</td>
<td>70.0±0.3</td>
<td>62.4±0.4</td>
<td></td>
</tr>
<tr>
<td>SVM+IG→name+desc</td>
<td>46.7±1.0</td>
<td>70.2±0.8</td>
<td>54.6±0.8</td>
<td>53.8±0.5</td>
<td>71.2±0.4</td>
<td>61.3±0.4</td>
<td></td>
</tr>
<tr>
<td>name+desc ∪ SVM+IG</td>
<td>53.7±1.0</td>
<td>64.5±0.8</td>
<td>57.2±0.8</td>
<td>59.7±0.4</td>
<td>68.1±0.3</td>
<td>63.6±0.3</td>
<td></td>
</tr>
</tbody>
</table>

- Classification helps IE. IE helps classification
Micro vs macro average

![Graph showing test set sampled from original distribution with permuted 1 and permuted 2.

- X-axis: Sector label
- Y-axis: Number of instances
- Legend: Test set, permuted 1, permuted 2]
Micro vs macro average

![Graphs showing micro vs macro average](image-url)
Tracking trends about entities

Relate entities via IE and different media

Study how companies interact in different sources:
- how companies appear in news
- their presence on social media, and
- daily fluctuation in their stock prices

Social media measured: hits on Twitter, Wikipedia, ...

Complex complementary relationships, time-series analysis
Tracking trends about entities

- Collect news from global sources
- Apply IE to identify key entities
- Collect complementary data about stock prices
- Collect social media mentions, access logs, ...

Look for

- correlations between mentions of company or product in the news and visibility in social media
- patterns related to companies, products, or industry sectors

Goals: market data prediction, ...
Daily trends (and smoothed)

alibaba

alstom

malaysian airlines

GM
Cross-correlations (at different lags)

- **Number** = lag (days)
  - +: wiki follows news
  - -: news follows wiki
- **Color** = sign of $\rho$
  - blue = positive
  - red = negative

- Largest values on the diagonal