Cours WebScience 2014

Information Extraction and Named Entity Recognition:

Getting simple structured information out of text

Elena Cabrio and Serena Villata

(equipe Wimmics)

CREDITS (for the slides): C. Manning and D. Jurafsky (Stanford University), A. Lavelli and B. Magnini (FBK), Andrew McCallum (University of Massachussets Amherst)

Information Extraction

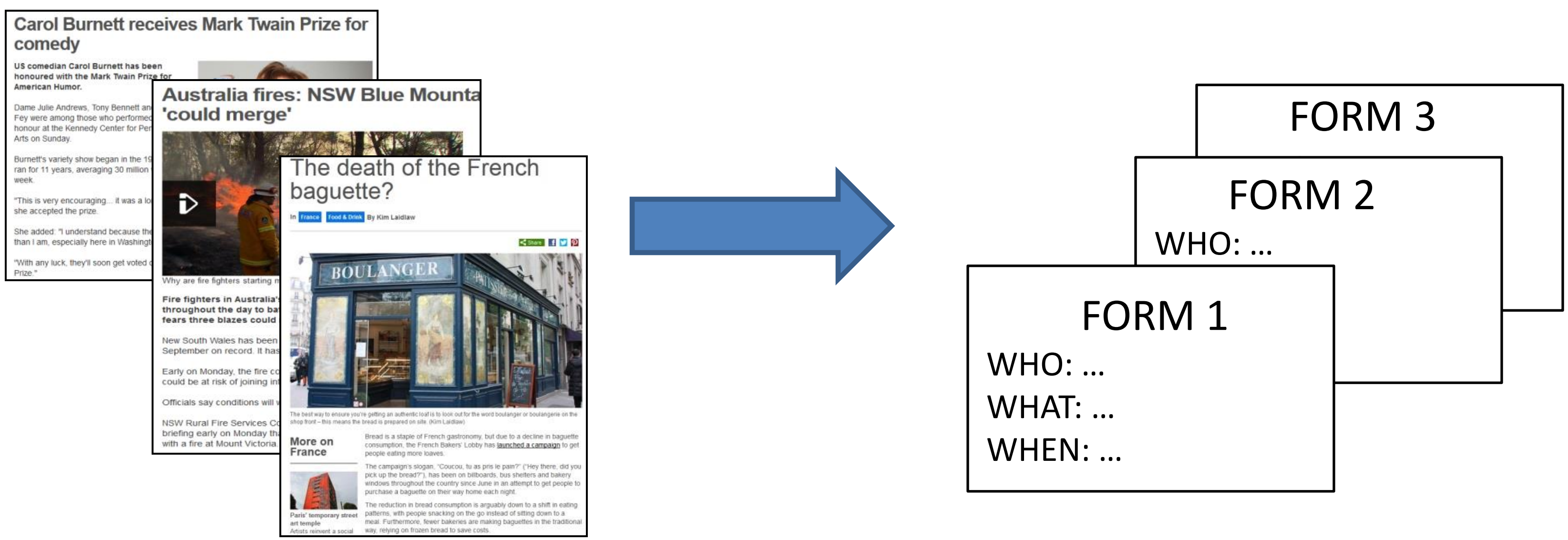
Information extraction (IE) systems:

- Find and understand **limited relevant parts of texts**
- **Gather information** from many pieces of text
- Produce a **structured representation** of relevant information:
  - ***relations*** (in the database sense),
  - a ***knowledge base***
- Goals:
  1. Organize information so that it is useful to people
  2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction

IE systems extract **clear, factual information**

Roughly: *Who did what to whom when?*



Mapping of texts into fixed format output (templates) representing the key information

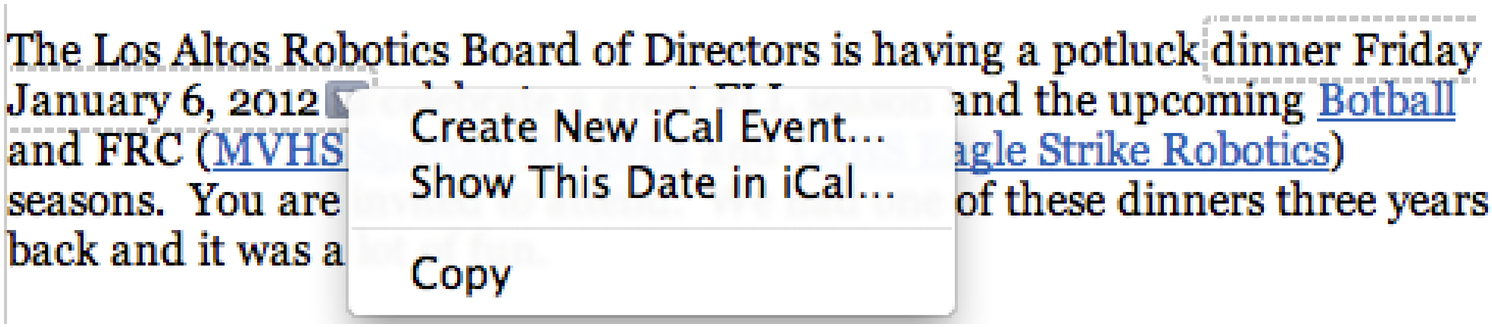
An example
(remember it for the Lab!!)

Vente Villa 4 pièces Nice (06000)
Réf. 12390: Sur les Hauteurs de Nice. Superbe villa moderne (190m2), 2 chambres et 1 suite parentale, 3 salles de bain. Très grand salon/salle à manger, cuisine américaine équipée. Prestations de haut standing. Vue panoramique sur la mer. Cette villa a été construite en 2005. 1 270 000 euros. Si vous êtes intéressés, contactez vite Mimi LASOURIS
06.43.43.43. 43

REAL ESTATE TEMPLATE
Reference: 12390
Prize: 1 270 000
Surface: 190 m2
Year Built: 2005
Rooms: 4
Owner: Mimi LASOURIS
Telephone: 06.43.43.43. 43

Low level Information Extraction

Is now available in applications like Apple or Google mail, and web indexing



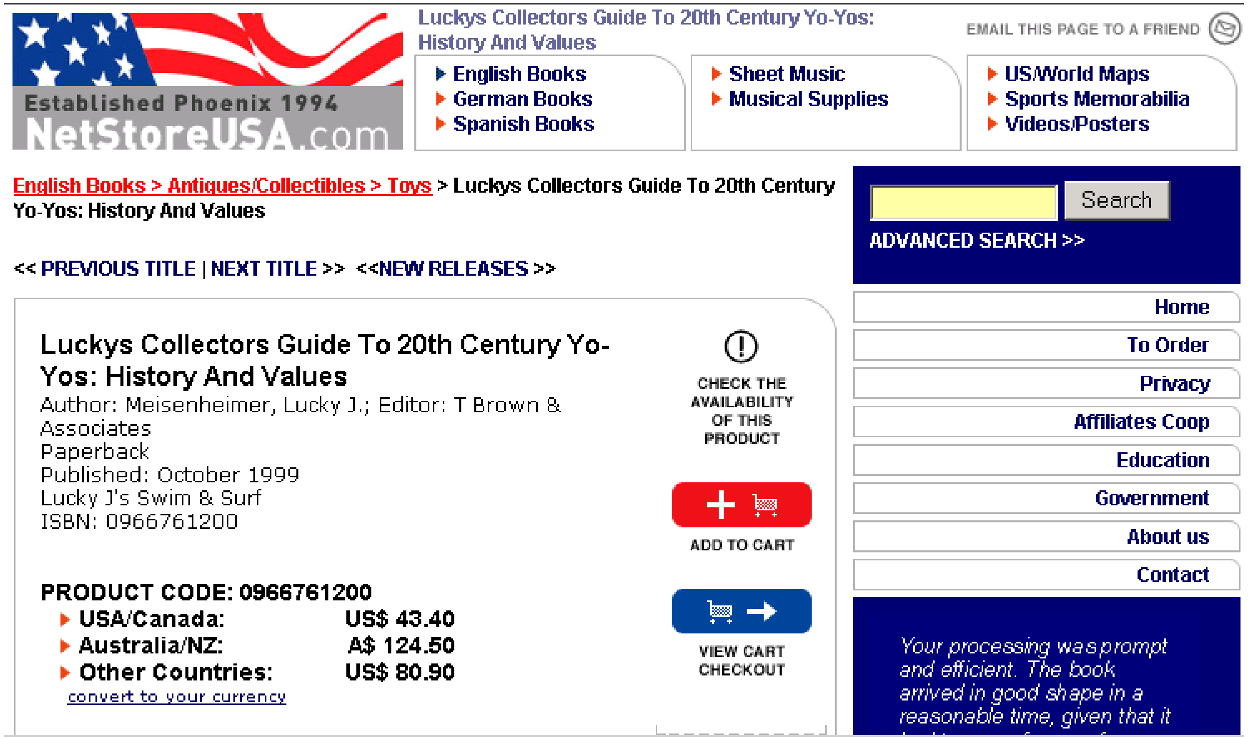
Often seems to be based on regular expressions and name lists

Why is IE hard on the Web?

A book, Not a toy

Title

Need this price



IE vs Information Retrieval

Information Retrieval

- User Query ⇨ Relevant Texts
- *Approach*: keyword matching
- *Query generality*: full

Information Extraction

- Linguistic analysis targeted to relevant information
- *User Query ⇨ Relevant Information*
- *Approach:* linguistic analysis
- *Query generality:* limited to target information

Named Entity Recognition

A very important sub-task: **identify** and **categorize**

- **Entities** (persons, organizations, locations)
- **Times** (dates, times and durations)
- **Quantities** (monetary values, measures, percentages and cardinal numbers)

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

A very important sub-task: **identify** and **categorize**

- **Entities** (persons, organizations, locations)
- **Times** (dates, times and durations)
- **Quantities** (monetary values, measures, percentages and cardinal numbers)

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person **Date** **Location** **Organization**

Named Entity Recognition

Crucial for Information Extraction, Question Answering and Information Retrieval

- Up to 10% of a newswire text may consist of proper names , dates, times, etc.

Relational information is built on top of Named Entities

Many web pages **tag** various **entities**, with links to bio or topic pages, etc.

- Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …

Apple/Google/Microsoft/… smart recognizers for document content

NER: Evaluation (remember it for the lab!)

The 2-by-2 contingency table:

	Correct	Not correct
Selected	TP	FP
Not selected	FN	TN

Precision: % of selected items that are correct

Recall: % of correct items that are selected

F-measure: weighted harmonic mean

NER task

Task: Predict entities in a text

Foreign	ORG	
Ministry	ORG	
spokesman	O	
Shen	PER	Standard evaluation is per entity, *not* per token
Guofang	PER	
told	O	
Reuters	ORG	
:	:	

Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

3 standard approaches to NER (and IE) (remember it for the Lab!)

1. **Hand-written regular expressions**
   - Perhaps stacked
2. **Using classifiers**
   - Generative: Naïve Bayes
   - Discriminative: Maxent models
3. **Sequence models**
   - HMMs
   - CMMs/MEMMs
   - CRFs

Hand written patterns for NER

If extracting from automatically generated web pages, **simple regex patterns** usually work.

- Amazon page
- <div class="buying"><h1 class="parseasinTitle">(.*?)</h1>

For certain restricted, common types of entities in unstructured text, **simple regex patterns** also usually work.

- Finding (US) phone numbers
- (?:\(?[0-9]{3}\)?[-.])?[0-9]{3}[-.]?[0-9]{4}

Natural Language Processing-based Hand-written Information Extraction

For unstructured human-written text, some NLP may help

- **Part-of-speech (POS) tagging**
  - Mark each word as a noun, verb, preposition, etc.
- **Syntactic parsing**
  - Identify phrases: NP, VP, PP
- **Semantic word categories** (e.g. from WordNet)
  - KILL: kill, murder, assassinate, strangle, suffocate

- Cascaded regular expressions to match relations
  - Higher-level regular expressions can use categories matched by lower-level expressions

Rules-based extracted examples

Determining which person holds what office in what organization

[person] , [office] *of* [org]

Vuk Draskovic, leader of the Serbian Renewal Movement

[org] (*named*, *appointed*, etc.) [person] Prep [office]

NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

[org] *in* [loc]

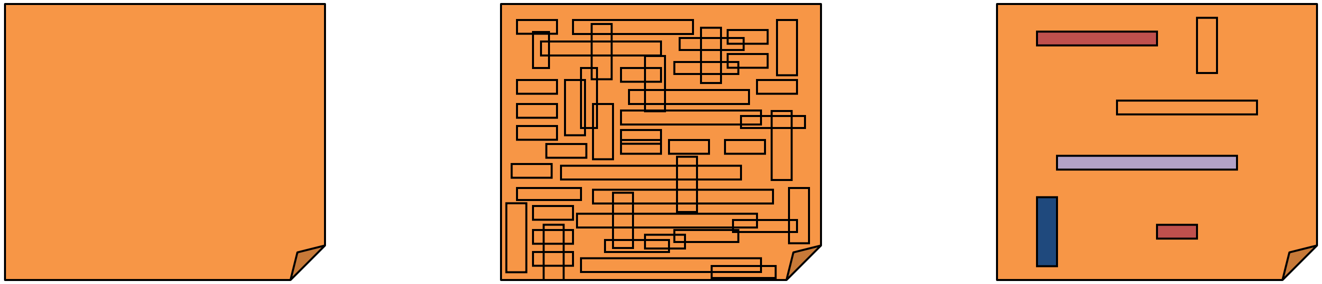
NATO headquarters in Brussels

[org] [loc] (*division*, *branch*, *headquarters*, etc.)

KFOR Kosovo headquarters

Naïve use of text classification for IE

- Use conventional classification algorithms to classify substrings of document as "*to be extracted*" or not.



- In some simple but compelling domains, this naive technique is remarkably effective.

The ML sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

The ML sequence model approach to NER

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Features for sequence labeling

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

The full task of Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Now Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

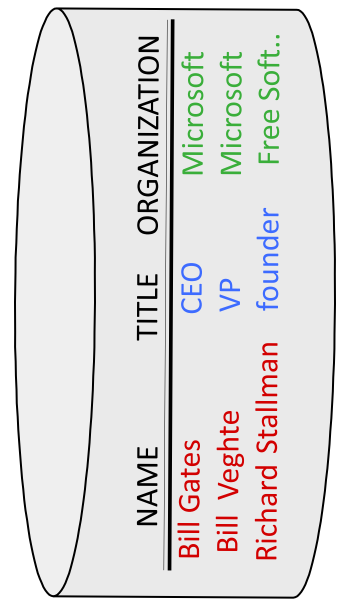
Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Gates
Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation



Arity of relations

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

Association task: Relation Extraction

Checking if groupings of entities are **instances of a relation**

1. **Manually engineered rules**

   Rules defined over words/entites:

   ```
   <company> located in <location>
   ```

   Rules defined over parsed text:

   ```
   ((Obj <company>) (Verb located) (*) (Subj <location>))
   ```

2. **Machine Learning-based**

   Supervised: Learn relation classifier from examples

   Partially-supervised: bootstrap rules/patterns from "seed" examples

Example

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire , is finding itself hard pressed to cope with the crisis…

Information Extraction System

Date	***Disease Name***	***Location***
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.
May 1995	**Ebola**	**Zaire**

Why Relation Extraction?

- Create new **structured knowledge bases**, useful for any app
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
  - Support question answering

    The granddaughter of which actor starred in the movie "E.T."?

    (acted-in ?x "E.T.")(is-a ?y actor)(granddaughter-of ?x ?y)!

How to build relation extractor

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
   - Bootstrapping (using seeds)
   - Distant supervision
   - Unsupervised learning from the web

Hand written patterns

"Agar is a substance prepared from a mixture of **red algae, such as Gelidium,** for laboratory or industrial use"

What does *Gelidium* mean?
How do you know?

Patterns for extracting IS-A relation (hyponyms)

```
"Y such as X ((, X)* (, and|or) X)"!
"such Y as X"!
"X or other Y"!
"X and other Y"!
"Y including X"!
"Y, especially X"!
```

(Hearst, 1992): Automatic Acquisition of Hyponyms

Extracting richer relations using rules

Intuition: relations often hold between specific entities

```
located-in (ORGANIZATION, LOCATION)
founded (PERSON, ORGANIZATION)
cures (DRUG, DISEASE)
```

Start with Named Entity tags to help extract relation!

Hand-built patterns for relations

Human patterns tend to be high-precision
- Can be tailored to specific domains

Human patterns are often low-recall
- A lot of work to think of all possible patterns!
- Don’t want to have to do this for every relation!
- We’d like better accuracy

Supervised Machine Learning

- Choose a set of **relations** we'd like to extract
- Choose a set of relevant **named entities**
- Find and **label data**
- Choose a representative corpus
- Label the Named Entities in the corpus
- Hand-label the relations between these entities
- Break into training, development, and test
- **Train a classifier** on the training set

Supervised Relation Extraction

- Can get high accuracies with enough hand-labeled training data, if test similar enough to training

- Labeling a large training set is expensive
- Supervised models are briattle, don't generalize well to different genres

Semi-supervised and unsupervised

Bootstrapping: use the seeds to directly learn to populate a relation

Gather a set of **seed pairs** that have relation R

Iterate:

1. Find sentences with these pairs
2. Look at the context between or around the pair and generalize the context to create patterns
3. Use the patterns for grep for more pairs

Bootstrapping

`<Mark Twain, Elmira>` **Seed tuple**

- Grep (google) for the environments of the seed tuple

  "Mark Twain is buried in Elmira, NY."
  `X is buried in Y`
  "The grave of Mark Twain is in Elmira"
  `The grave of X is in Y`
  "Elmira is Mark Twain's final resting place"
  `Y is X's final resting place.`

- **Use those patterns to grep for new tuples**
- Iterate

Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

Errors cascade (error in entity tag → error in relation extraction)

These are very rough, actually optimistic, numbers

- Hold for well-established tasks, but lower for many specific/novel IE tasks

TP

Extraction d'information structurée a partir des annonces immobiliers sur le Web

1. Choisir un site web d'annonces immobiliers
    1. CraigList (cotedazur.fr.craigslist.fr)
    2. PAP (http://www.pap.fr)

2. Extraire au moins 15 textes des annonces, et les sauvegarder dans des fichiers textuels. 5 textes serons utilisés comme set de développement, et les autres serons utilisés comme test set. Vous pouvez les extraire en choisissant une des stratégies suivantes:

1. Web Crawler (extraire les textes automatiquement des pages web, par exemple en utilisant la librairie java http://jsoup.org/)

2. En copiant les textes des sources HTML de la page

3. Analyser les annonces du set de développement pour identifier au moins 10 caractéristiques que vous jugez relevants pour décrire les biens immobiliers (par exemple: le type de bien, le prix, la surface, le nombre de pièces, etc.)

4. Générer un fichier CSV (how? http://www.computerhope.com/issues/ch001356.htm), ou Excell, pour stocker les informations relevants de chaque annonce du test set, par rapport aux caractéristiques choisies. L'extraction doit être automatique, en utilisant un des méthodes expliqués dans le cours:

1. Hand written patterns (patrons, regex écrites a la main)

2. Classifiers (en utilisant des algorithmes d'apprentissage automatique)

Des outils de TAL (Stanford parser, Tree Tagger) peuvent être utilisés pour tokeniser, lemmatiser, ou détecter les Part-of-Speech du texte.

5. Résultat attendu :

Reference annonce	Type de bien	Prix	Surface
1234	Appartement	230000 euros	$60m^2$

6. Évaluation:

1. Création du goldstandard : annoter les annonces de test manuellement, en utilisant les balises xml, comme il suit:

Superbe <TYPEDUBIEN> appartement </TYPEDUBIEN> standing <SURFACE>73m²</SURFACE>

2. Calculer la précision, le rappel et la F-measure du tableau obtenu grâce au votre extracteur d'information, par rapport aux annotations correctes que vous avez annoté manuellement (le goldstandard).

7. Écrire un rapport de au moins deux pages, qui décrit avec précision toutes les étapes que vous avez parcouru, les stratégies que vous avez choisi dans chaque étape, et les résultats obtenus.

8. **TP rendu** : rapport, fichier CSV, fichier goldstandard. Date limite: **lundi 2 juin**.

Envoyer par mail aux adresses: **elena.cabrio@inria.fr**; **serena.villata@inria.fr**