

Information Extraction - Introduction

Pawan Goyal

CSE, IIT Kharagpur

Week 10, Lecture 3

Goal: “machine reading”

Goal

Acquire structured information knowledge from unstructured text



Information Extraction (IE) Systems

- Find and understand limited relevant parts of texts

Information Extraction (IE) Systems

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text

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:

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

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

- Organize information so that it is useful to people
- Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

Information Extraction (IE)

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

What sort of information?

Information Extraction (IE)

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

What sort of information?

IE Systems extract clear, factual information

Information Extraction (IE)

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

What sort of information?

IE Systems extract clear, factual information

- Roughly: *Who did what to whom when?* etc.

Information Extraction (IE)

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

What sort of information?

IE Systems extract clear, factual information

- Roughly: *Who did what to whom when?* etc.

E.g., *Gathering earnings, profits, headquarters etc. from company reports*

- The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.

Information Extraction (IE)

Definition

Information extraction is the task of finding structured information from unstructured or semi-structured text.

What sort of information?

IE Systems extract clear, factual information

- Roughly: *Who did what to whom when?* etc.

E.g., *Gathering earnings, profits, headquarters etc. from company reports*

- The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
- **headquarters**("BHP Biliton Limited", "Melbourne, Australia")

Information Extraction (IE)

Example

In 1998, Larry Page and Sergey Brin founded Google Inc.

Example

In 1998, Larry Page and Sergey Brin founded Google Inc.

We can extract the following information,

- FounderOf(Larry Page, Google Inc.),
- FounderOf(SergeyBrin, Google Inc.),
- FoundedIn(Google Inc., 1998)

Example

In 1998, Larry Page and Sergey Brin founded Google Inc.

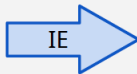
We can extract the following information,

- FounderOf(Larry Page, Google Inc.),
- FounderOf(SergeyBrin, Google Inc.),
- FoundedIn(Google Inc., 1998)

Such information can be used by search engines and database management systems to provide better services to end users.

Biomedical domain

- A large amount of scientific publications
- Need to look for discoveries related to particular genes, proteins or other biomedical entities
- Biomedical entities often have synonyms and ambiguous names
- **Critical task:** automatically identify mentions of biomedical entities in text and link them to their corresponding entries in existing knowledge bases.



Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...

textual abstract:
summary for human

structured knowledge extraction:
summary for machine

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip on flights to some cities also served by lower-cost carriers. **American Airlines**, a **unit of AMR**, immediately matched the move, **spokesman Tim Wagner** said. **United**, a **unit of UAL**, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

Relation types

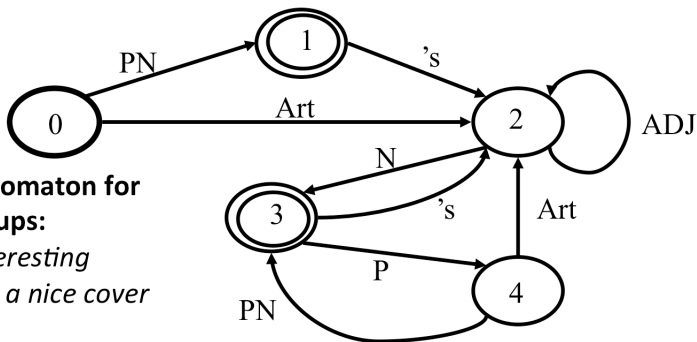
For generic news text ...

Relations		Examples	Types
Affiliations			
	Personal	<i>married to, mother of</i>	PER → PER
	Organizational	<i>spokesman for, president of</i>	PER → ORG
	Artifactual	<i>owns, invented, produces</i>	(PER ORG) → ART
Geospatial			
	Proximity	<i>near, on outskirts</i>	LOC → LOC
	Directional	<i>southeast of</i>	LOC → LOC
Part-Of			
	Organizational	<i>a unit of, parent of</i>	ORG → ORG
	Political	<i>annexed, acquired</i>	GPE → GPE

Relation extraction: 5 easy methods

- Hand-built patterns
- Bootstrapping methods
- Supervised methods
- Distant supervision
- Unsupervised methods

Hand-written Information Extraction: use regex



Finite Automaton for Noun groups:

*John's interesting
book with a nice cover*

Rule-based Extraction Examples

Determining which person holds what position in what organization

Rule-based Extraction Examples

Determining which person holds what position in what organization

[person], [position] of [org]

Vuk Draskovic, leader of the Serbian Renewal Movement

Rule-based Extraction Examples

Determining which person holds what position in what organization

[person], [position] 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

Rule-based Extraction Examples

Determining where an organization is located

Rule-based Extraction Examples

Determining where an organization is located

[org] in [loc]

NATO headquarters in Brussels

Rule-based Extraction Examples

Determining where an organization is located

[org] in [loc]

NATO headquarters in Brussels

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

KFOR Kosovo headquarters

Intuition from Hearst (1992)

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

- What is Gelidium?

Intuition from Hearst (1992)

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

- What is Gelidium?
- How do you know?

Intuition from Hearst (1992)

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

- What is Gelidium?
- How do you know?

Automatic Acquisition of Hyponyms

- Y such as $X((, X) * (, \text{ and/or } X)$
- such Y as X
- X or other Y
- X and other Y
- Y including X
- Y , especially X

Examples of Hearst patterns

Hearst pattern	Example occurrences
X and other Y	...temples, treasures, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

Patterns for learning meronyms

Berland and Charniak's patterns

- Selected initial patterns by finding all sentences in a corpus containing *basement* and *building*

Patterns for learning meronyms

Berland and Charniak's patterns

- Selected initial patterns by finding all sentences in a corpus containing *basement* and *building*

whole NN[-PL] 's POS part NN[-PL]
part NN[-PL] of PREP {the|a} DET mods [JJ|NN]* whole NN
part NN in PREP {the|a} DET mods [JJ|NN]* whole NN
parts NN-PL of PREP wholes NN-PL
parts NN-PL in PREP wholes NN-PL

... building's basement ...
... basement of a building ...
... basement in a building ...
... basements of buildings ...
... basements in buildings ...

Problems with hand-built patterns

- Requires hand-building patterns for each relation!
 - ▶ hard to write; hard to maintain
 - ▶ there are zillions of them
 - ▶ domain-dependent
- Don't want to do this for all possible relations!
- Plus, we'd like better accuracy
 - ▶ Hearst: 66% accuracy on hyponym extraction
 - ▶ Berland and Charniak: 55% accuracy on meronyms