Lecture 3 IE Concepts

Fang Li
Dept. of Computer Science

Contents

- IE Definition, History and Concepts
- IE Technologies
- **IE Evaluations** with classroom exercise

What is IE? (**old** definition)

- Information Extraction (IE) aims to extract the **facts** from documents.
- IE extracts information from **actual texts** by computer at high speed, which are normally from publicly available electronic sources
- Map them into predefined, structured representations (e.g.,templates),

What is IE ? (Definition)

- Information Extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in *unstructured data sources,* such as natural language text, making the information more suitable for information processing tasks. (**new definition**)

IE History: MUC, ACE, TAC overview

- 1987~1998
- MUC: **M**essage **U**nderstanding **C**onference
- 1999 ~ 2008
- ACE: **A**utomatic **C**ontent **E**xtraction
- 2008 ~ today
- TAC: **T**ext **A**nalysis **C**onference (2008 ~ now)
- **In the form of a competition**

Participants submit their results and compare with human-made results.

MUC, ACE, TAC (research tasks)

- **MUC**: named entity recognition, coreference resolution, template element construction, element construction, scenario template production.
- **ACE**: detection & tracking of entities, recognition of semantic relations, recognition of events
- **TAC**: Entity Discovery and Linking, knowledge base population,…

IE example

Mr. **Murdoch** moved to Los Angeles from New York to focus on the filmed entertainment operations that were then under **Barry Diller**, **Fox** **chief executive;**

IE definition:

Management Succession

Identification:
Classification
structuring

Management Succession

Organization:	Fox
Post:	chief executive
Person In:	Murdoch
Person Out:	Barry Diller

Some Concepts of IE

- Named Entity: Individuals in the world *that are mentioned in the text with* ***a name****.*
- Relation: Properties that hold of two entities over **a time interval**.
- Event: A particular kind of relation among entities, implying **a change** in relation state at the end of the time interval. Different entities play different **roles** in the relation.

Some Concepts of IE (cont.)

- **Linguistic Mention**
  - A particular **linguistic phrase**
  - Denotes a particular *entity*, *relation*, or *event*
    - A noun phrase, name, or possessive pronoun
    - A verb, nominalization, compound nominal, or other linguistic construct relating other linguistic mentions
- **Linguistic Entity**
  - **Equivalence class** of mentions with same meaning
    - Co-referring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

From Douglas E. Appelt

Example

- **Linguistic Mention**:
  - 上海交通大学 (named entity)
  - 上交大,SJTU (Abbreviations)
  - 位于上海西南角著名的高等学府 (a phrase)
  - SJTU ...., 它是世界百强大学之一 (pronoun)
- **Linguistic Entity**: all of them

Example of linguistic mention and linguistic entity

Bridgestone Sports Co said Friday **it** has **set up a joint venture in Taiwan** with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, *Bridgestone Sports TaiWan Co.*, capitalized at 20 million Tainwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month. The monthly output will be later raised to 50,000 units, **Bridgestone Sports** spokesman Tom White, said.

IE Tasks

- Recognition of **entity**, **relation** or **event**.
- **Coreference resolution**

Real World

people, company and so on, such as:

Bridgestone Sports Co

These mentions may represent the same entity.

1. Bridgestone Sports Co
2. It
3. Bridgestone Sports
4. The company

IE task: How to identify it?

- **ComplexWords**: recognition of multiwords and proper named entities.
- **Basic Phrases**: Sentences are segmented into noun groups, verb groups, and particles.
- **Complex Phrases**: Complex noun groups and complex verb groups are identified.
- **Domain Events**: semantic structures are built that encode the information about entities and events contained in the pattern.
- **Merging Structures**: Semantic structures from different parts of the text are merged if they provide information about the same entity or event.

Complex Words

- For example: “set up”, ” Bridgestone Sports Co.”
- IBM is a company, DNA is not.
- XYZ's sales.
- Vaclav Havel, 53, former president of the Czech Republic.

Basic Phrases

Company Name:	Bridgestone Sports Co.
Verb Group:	said
Noun Group:	Friday
Noun Group:	it
Verb Group:	had set up
Noun Group:	a joint venture
Preposition:	in
Location:	Taiwan
Preposition:	with
Noun Group:	a local concern
Conjunction:	and
Noun Group:	a Japanese trading house
Verb Group:	to produce
Noun Group:	golf clubs
Verb Group:	to be shipped
Preposition:	to
Location:	Japan

Complex Phrases

- the attachment of appositives to their head noun group: “The joint venture, Bridgestone Sports Taiwan Co."
- the construction of measure phrases “20,000 iron and metal wood clubs a month"
- the attachment of “of" and “for" prepositional phrases to their head noun groups: “production of 20,000 iron and metal wood clubs a month"
- noun group conjunction: “a local concern and a Japanese trading house"

Domain Events

The domain event patterns:

① <Company/ies> <Set-up> <Joint-Venture> with <Company/ies>

② <Produce> <Product>

③ <Company> <Capitalized> at <Currency>

④ <Company> <Start> <Activity> in/on <Date>

1

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	–
Activity:	–
Amount:	–

2

Activity:	PRODUCTION
Company:	–
Product:	"golf clubs"
Start Date:	–

3

Relationship:	TIE-UP
Entities:	–
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

4

Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	–
Start Date:	DURING: January 1990

Merging Structures

1+3

Relationship:	TIE-UP
Entities:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	–
Amount:	NT$20000000

- assign each entity and object to the appropriate event template.
- Merge them if they are consistent.

2+4

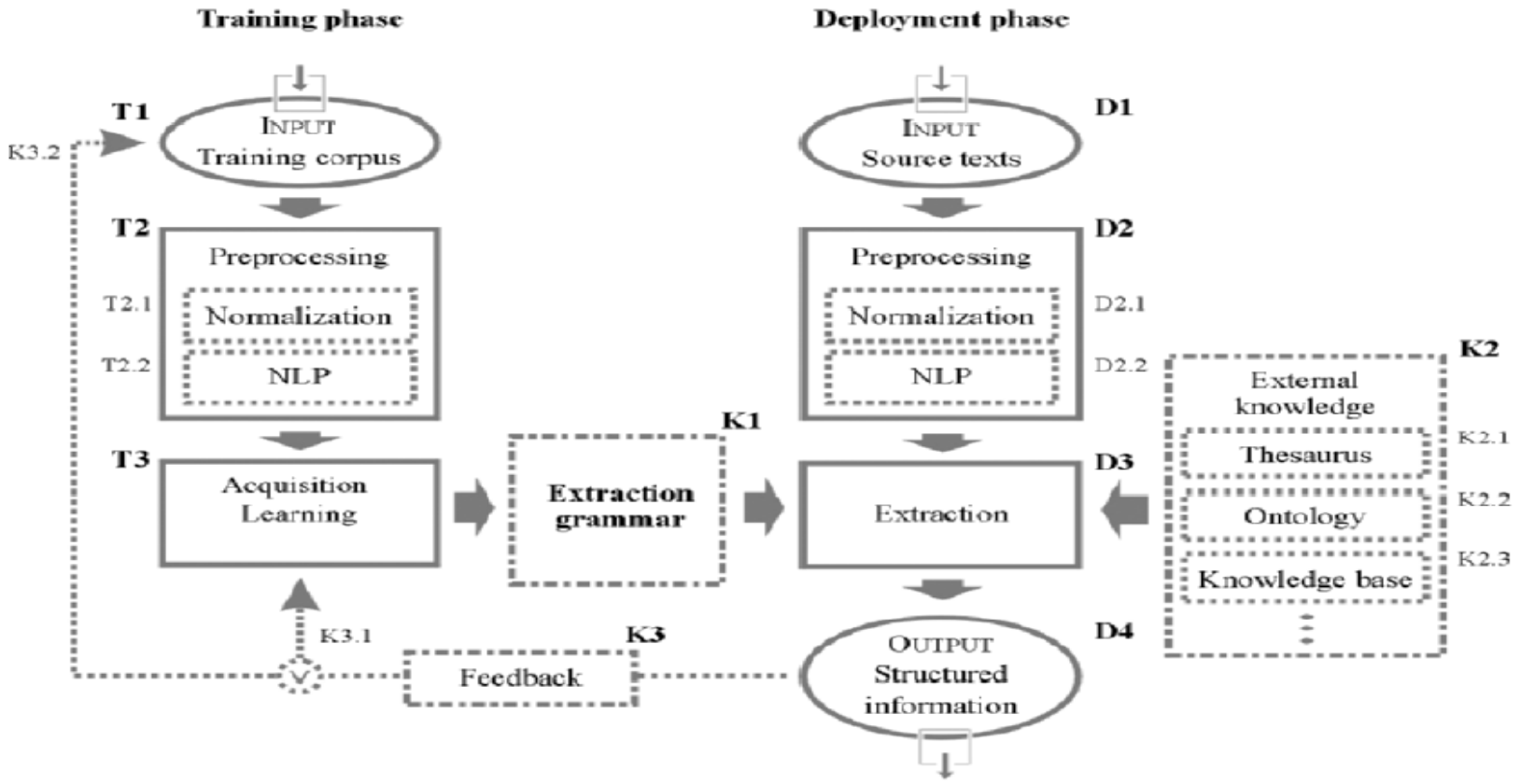
Activity:	PRODUCTION
Company:	"Bridgestone Sports Taiwan Co."
Product:	"iron and 'metal wood' clubs"
Start Date:	DURING: January 1990

Diversity of IE source

- Unstructured IE
- Semi-structured IE

- Single Doc.
- Multiple Doc.

The Common Extraction Process



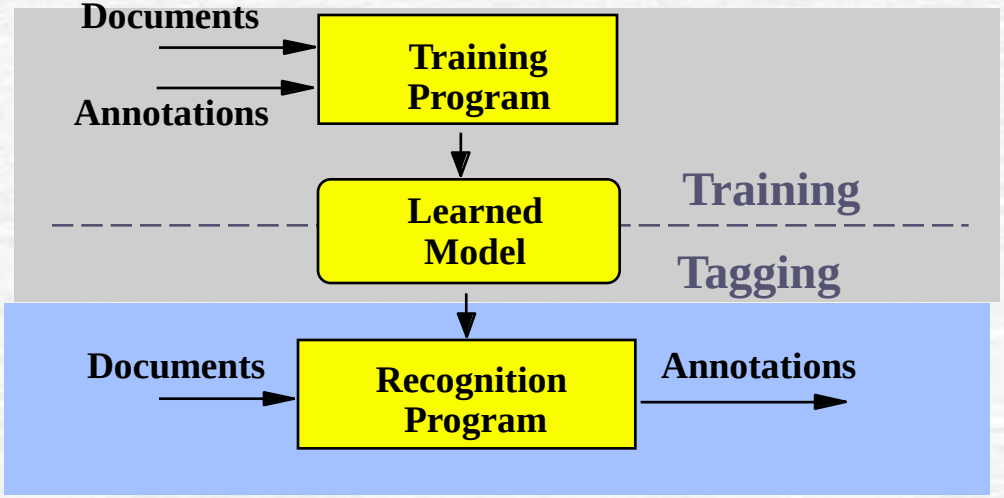
Legend	
T	component of training phase
D	component of deployment phase
K	knowledge component

A typical information extraction system.

Two Basic Approaches to IE

- **Knowledge Engineering Approach:**
  Grammars are constructed by hand
  Domain patterns are discovered by human

- **Automatic Learning Approach:**



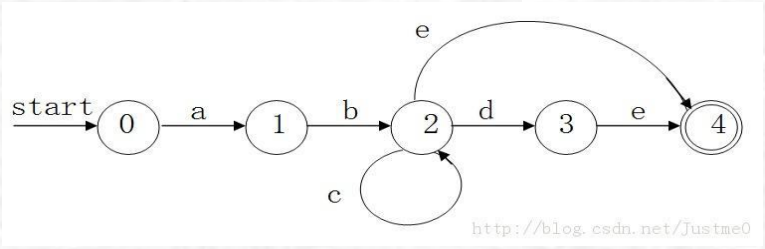
Learn to recognize information from examples
(text "annotated" with correct answers)

IE Approaches

- **Knowledge engineering**
- **Automatically learning**
  - Statistical learning
  - Machine learning
  - Deep learning
- **Hybrid approach**

Knowledge Engineering

- Adopts human **linguistic knowledge** to build grammatical and semantic rules for the components in IE systems.
- *Finite-state automata*:.
- *Cascaded automata*



Knowledge Engineering (advantages and disadvantages)

- The best performing systems.
- Human ingenuity in establishing and tuning patterns is still in the lead.

- Very **laborious** development process
- Domain adaptation might require **reconfiguration**
- Needs experts who have both, **linguistics and domain** expertise.

Machine learning

- **Inductive learning: learn a function from examples (simplest form)**

f is the **target function,** An **example** is a pair (*x*, *f(x)*)

Task: find a **hypothesis** *h* **such that** ***h*** **≈** ***f***
given a training set of examples

  - Ignores prior knowledge
  - Assumes examples are given

Machine learning (cont)

- Supervised learning

Given $D = \{\mathbf{X}_i, \mathbf{Y}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $D^{new} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$

- Unsupervised learning

Given $D = \{\mathbf{X}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $D^{new} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$

- Semi-supervised learning

a small amount of labeled data with a large amount of unlabeled data.

Machine Learning Method (advantages and disadvantages)

- Rules are automatically derived from the training data.
- System can be customized to a specific domain without interfering from any developers.
- **Training data** may be difficulty supplied or expensive to obtain.
- Changes to specifications may require **re-annotation** of large quantities of training data

Statistically Learning

- Depends on corpus analysis and statistics, which is an empirical approach.
- often use some machine learning models:

- HMM (Hidden Markov Model)
- SVM (Support Vector Machines)
- MEM (Maximum Entropy modeling)

Statistically Learning (advantages and disadvantages)

- Analyze and discover fairly **fine distinction of language phenomena**
- Build a statistical model of actual language
- Resolve some **practical problems** of actual language texts
- Relies on statistical corpus including domain and distribution of language phenomena, to a great extent.

统计学习vs.机器学习

	学习函数方法	解释性	注重点
统计学习	假设→验证	强	模型的可解释。
机器学习	不假设，交叉验证	弱	模型的可预测性

hybrid methods

- Combines the above approaches for giving play to their strong points.

What works best?

- Use rule-based approach when
  - Resources (e.g., lexicons,lists) are available
  - Rule writers are available
  - Training data scarce or expensive to obtain
  - Extraction specs likely to change
  - Highest possible performance is critical
- Use trainable approach when
  - Resources unavailable
  - No skilled rule writers are available
  - Training data is cheap and plentiful
  - Good performance is adequate for the task

Evaluation for IE

- **Intrinsic Evaluation**, i.e., the performance of the extraction task is measured.
- **Extrinsic Evaluation**, i.e., measuring the performance of another task in which information extraction is an integral part.

Evaluation for IE （cont.）

A golden standard is used to evaluate the result of systems

		B	
		Yes	No
A	Yes	20	5
	No	10	15

- Golden standard: **human made**
- inter-annotator agreement: e.g. more than 80% (Cohen's kapp coefficient)

Cohen's kappa coefficient: $(p_0-p_e)/(1-p_e)$

p_0实际标注精度，p_e随机精度

Evaluation for IE

$$precision = \frac{\#correct\_answers}{\#answers\_produced}$$

$$recall = \frac{\#correct\_answers}{\#total\_number\_of\_correct\_answers}$$

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

$$\beta = 1 \Rightarrow F1 = \frac{2}{1/P + 1/R}$$

β Is a parameter representing relative importance of P and R. e.g. β =1 means R and P equal weighting, β =0 means only P.

How to Evaluate

	Expert says yes	Expert says no	
System says yes	a	b	$a + b = k$
System says no	c	d	$c + d = n - k$
	$a + c = r$	$b + d = n - r$	$a + b + c + d = n$

where

n = number of classified objects

k = number of objects classified into the class C_i by the system

r = number of objects classified into the class C_i by the expert.

Precision = a / (a+ b)

Recall = a / (a + c)

Accuracy = (a+d)/ n

b is the wrong answers (false positive)

c is missing answers (false negatives)

Evaluation for IE (cont.)

- High precision means that the **extracted information does not contain any or only very few errors.**
- High recall refers to the situation where **all or almost all information to be extracted is actually extracted**.
- Accuracy is computed as the proportion of correct assignments to a class in all assignments.

Accuracy or Precision ?

	Correct	Not correct
System selected	0	0
System not selected	10	990

If there are 1000 examples, 10 are correct, 990 are not correct. System finds nothing. What are the accuracy?

Accuracy = 99% -- no meaning

Precision is important.

How to Evaluate **Multiple Classes**

- Often multiple classes are assigned, in order to evaluate the whole system, **macro averaging** and **micro-averaging** are used.

	Expert says yes	Expert says no
System says yes	10	10
System says no	10	970

Class 1

	Expert says yes	Expert says no
System says yes	90	10
System says no	10	890

Class 2

	Expert says yes	Expert says no
System says yes	100	20
System says no	20	1860

All classification decisions

Macro-averaged precision: $(0.5 + 0.9)/2 = 0.7$ Averaged over classes

Micro-averaged precision: $100/120 = 0.83$ Over all binary classification decision

How to Evaluate **Multiple Classes** (cont.)

- **Macro-Averaging**: gives equal weight to every category (*category-pivoted measure*).
- **Micro-Averaging**: gives equal weight to every document (it is called a *document-pivoted measure*)

Summarization

- What is Information Extraction? √
- What are the **general methods** for IE? √
- What are the **evaluation metrics** for IE ? √

References

- Textbook chapter 1, 2, 8
- Douglas E.Appelt,” Introduction to Information Extraction” (Tutorial for IJCAI-99)

Chinese Language Processing Platform:

- http://ictclas.nlpir.org/nlpir
- http://www.ltp-cloud.com
- http://nlp.qq.com

IE sources

- http://www.ontotext.com/kim
- http://callisto.mitre.org
- http://timeml.org/site/tango/tool.html
- http://complingone.georgetown.edu/~linguist/compling.html
- http://gate.ac.uk/
- http://nltk.org