


Language Technology I

Relation Extraction Exercises

- 1. Please list the differences between the information extraction task and the full text understanding task.***

Information Extraction: The Nature of the Task



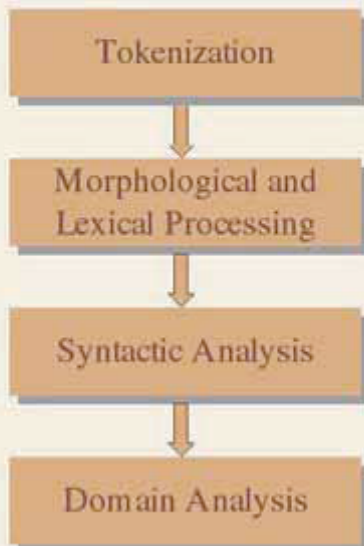
- * Delimited criteria of relevance/topics are specified in advance
- * Fixed and limited representational format
- * Clear criteria of success are at least possible
- * Corollary features:
 - Typically only parts of the text are relevant
 - Often only part of a relevant sentence is really relevant
 - Can be targeted at large corpora

Text Understanding: The Nature of the 'Task'

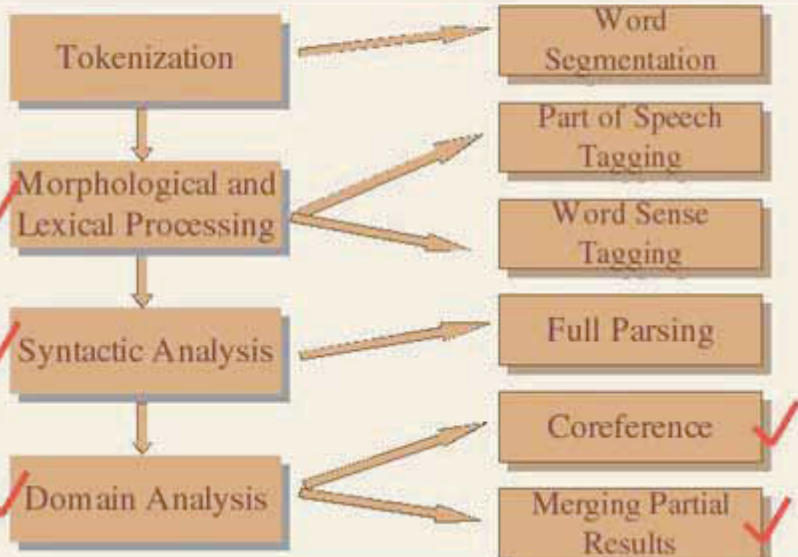


- * No predetermined specification of, or limit to the semantic and communicative aspects of interest
- * Representation of meaning must be rich and flexible enough to capture all the meaning (?) of the text.
- * No clearly defined criteria of success
- * Corollary features:
 - Every bit of the text is relevant
 - Can't (yet) be applied to large bodies of text

A Bare-Bones Extraction System



Flesh for the Bones



2. Please calculate Precision and Recall of the following information extraction task.

$$\text{Precision} = \frac{\text{correct answers extracted by the system}}{\text{all answers extracted by the system}}$$

$$\text{Recall} = \frac{\text{correct answers extracted by the system}}{\text{total possible answers}}$$

Text

1. 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.
2. Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.
3. "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."
4. Richard Stallman, founder of the Free Software Foundation, countered saying...

Possible Answers distributed in the different paragraphs

1. 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.
2. Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.
3. "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."
4. Richard Stallman, founder of the Free Software Foundation, countered saying...

In the following, we list all possible names and their types in each paragraph

1. Microsoft Corporation (organization name), CEO (title/position name), Bill Gates (person name)
2. Microsoft (organization name), Gates (person name)
3. Bill Veghte (person name), Microsoft (organization name), VP (title/position name)
4. Richard Stallman (person name), founder (title/position name), Free Software Foundation (organization name)

Answers exacted by the System

1. 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.
2. Today, claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.
3. "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."
4. Richard Stallman, founder of the Free Software Foundation, countered saying...

In the following, we list all the extracted names and their types in each paragraph

1. Microsoft Corporation (organization name), CEO (title/position name), Bill Gates (person name), Orwellian (person name)
2. Windows (person name)
3. Bill Veghte (person name), Microsoft VP (person name)
4. Richard Stallman (person name), founder (title/position name), Free Software Foundation (organization name)

Please calculate for each name type the precision and recall values of the named recognition task of the system.

Right Answer:

Name Types	Precision	Recall
Person Name	50%	75%
Organization Name	100%	50%
Title/Position Name	100%	66.7%

3. Information Extraction in the Management Succession Domain

Text for Information Extraction

1. Cash-strapped Figgie International Inc. eliminated its quarterly dividend and received a temporary cash infusion from a new lender.
2. The Willoughby, Ohio, industrial company also named a longtime outside director as a new vice chairman, although it stopped short of bringing in a total outsider to bolster management.
3. Meanwhile, Figgie signaled that its 1993 operating losses and year-end adjustments would be greater than previously expected when reported in two to three weeks.
4. The company's Class A share fell \$1.875, or 15%, to \$10.875 in Nasdaq Stock Market trading.
5. The conglomerate said it received a new \$40 million, one-year renewable loan from CIT Group, which is secured by receivables. That eliminated the need to ask its other banks, which have extended a \$150 million credit line, for more funds. "With CIT's support, we have addressed our near-term cash flow needs while we continue to put into place the elements of a more comprehensive deleveraging of the balance sheet," Chairman Harry E. Figgie Jr. said in a statement.
6. Nevertheless, Figgie sought to calm its syndicate of bankers at a meeting yesterday. The bankers expressed concerns about the company's financial plight and the need for new management, according to a person who attended the meeting. A company spokesman confirmed that those issues were discussed, but said they were "not representative of the whole meeting."
7. Wall Street was hoping for stronger outside management to help Figgie. Instead, the company named a director, 66-year-old Walter M. Vannoy, who has been on the board since 1981. Although the current vice chairman, Harry E. Figgie III, 40, will continue to hold that title, Mr. Vannoy will be second in command, the company said.
8. Mr. Vannoy formerly served as vice chairman of McDermott International Inc. and as president and chief operating officer of Babcock & Wilcox.
9. The company said suspension of the six-cents-a-share quarterly dividend, which would have been payable in March on both Class A and Class B common stock, will save it about \$4 million a year. "The board felt the dividend suspension was prudent until we effect a profitability turnaround."
10. The actions, coupled with the company's previously announced plan to sell its Rawlings Sporting Goods division and two other divisions, are the first phase of a turnaround plan to be disclosed within the next two weeks, Figgie said. Among other things, the company wants to lop \$200 million off its \$450 million of debt this year.

a. Named entity recognition

Please extract all names of the following types from the above text and specify the paragraph number, where the name occurs.

- i. Person names, e.g., **Walter M. Vannoy (7)**
- ii. Location names, e.g., **Ohio (2)**,
- iii. Company/organization names, e.g., **Figgie International Inc. (1)**
- iv. Position names, e.g., **director (2)**, **vice chairman (2)**

b. Relation recognition

Please extract the person-position and person-company relations from the above texts

- i. Person-Position Relationships: a relationship between a person and his position in a company

Person	Position
Harry E. Figgie Jr. (5)	Chairman (5)
Walter M. Vannoy (7)	Director (7)
Harry E. Figgie III (7)	vice chairman (7)
Mr. Vannoy (8)	vice chairman (8)
Mr. Vannoy (8)	president and chief operating officer (8)

- ii. Person-Company Relationships: the affiliation of a specific person

Person	Company
Harry E. Figgie Jr. (5)	CIT Group (5)
Walter M. Vannoy (7)	Figgie (7)
Harry E. Figgie III (7)	Figgie (7)
Mr. Vannoy (7)	Figgie (7)
Mr. Vannoy (8)	McDermott International Inc. (8)
Mr. Vannoy (8)	Babcock & Wilcox (8)

c. Filling templates

Please try to fill the following database records by extracting corresponding information from the above text. The database record corresponds to a management succession event, which contains four attributes

- i. Person In: person, who is named for the position
- ii. Person Out: person, who resigned from the position
- iii. Position: position, which needs a personnel change
- iv. Company: company or organization, where the personnel changes take place
- v. Time: when the personnel change takes place

Right Answer:

Person IN	Person OUT	Position	Company	Time
Walter M. Vannoy (7)		Director (7)	<u>Figgie (7)</u>	
	<u>Mr. Vannoy (8)</u>	vice chairman (8)	<u>McDermott International Inc. (8)</u>	
	<u>Mr. Vannoy (8)</u>	president and chief operating officer (8)	<u>Babcock & Wilcox (8)</u>	

- d. Identify linguistic patterns as relation extraction rules for Person_In, Person_Out,

Right Answer:

1. *Person_In*:
<Company> **named** <Person: Person_In> as <Position>
2. *Person_Out*:
<Person: Person_Out> **formerly served as** <Position> of <Company>

4. Please give a definition of the components of the semantic model of relation extraction and the major tasks of the relation extraction.

Right Answer:

Components:

- Entities - Individuals in the world *that are mentioned in a text*
 - Simple entities: singular objects
 - Collective entities: sets of objects of the same type *where the set is explicitly mentioned in the text*
- Relations – Properties that hold of tuples of entities.
- *Complex Relations* – Relations that hold among entities and relations
- Attributes – one place relations are attributes or individual Properties
- Temporal points and intervals
- Relations may be timeless or bound to time intervals
- Events – A particular kind of simple or complex relation among entities involving a change in relation state at the end of a time interval.

Tasks:

- Named-Entity Recognition
- NE Type Classification
- Coreference Resolution
- Relation/Event Extraction

5. Please describe the relationship between a system for learning relation extraction rules and a relation extraction system. In which system, the two systems are integrated into one.

Right Answer:

- **Relationship**

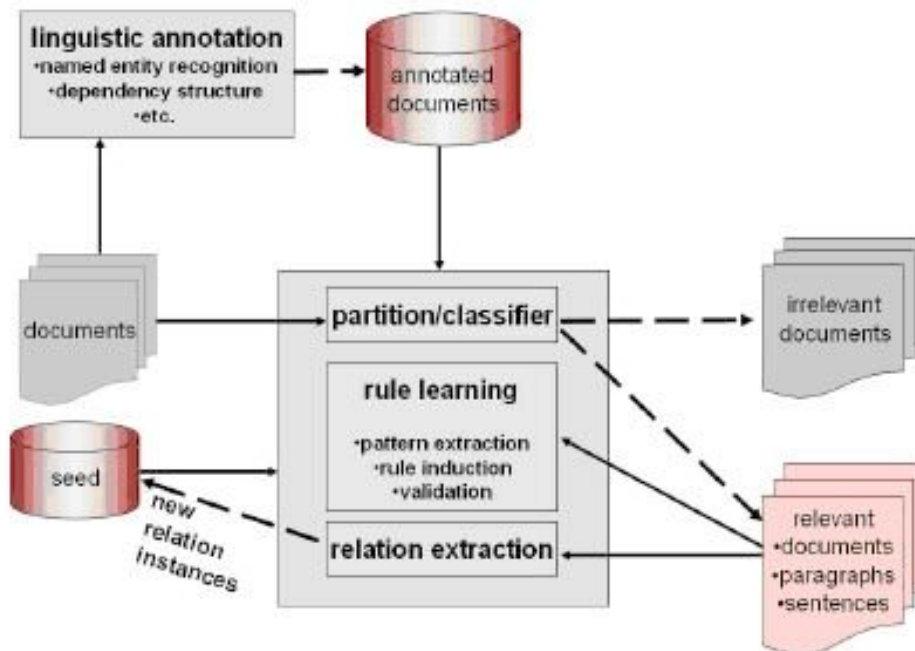
In general, the learning system tries to learn extraction rules. The learned rules will be used as rule resources for the relation extraction system.

- **In DIPRE, Snowball and DARE (dare.dfki.de), rule learning and relation instance extraction are combined into one system within a bootstrapping framework.**

References:

1. Feiyu Xu, Hans Uszkoreit, Hong Li
A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity
Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 6/2007
http://www.dfki.de/lt/publication_show.php?id=3013
2. Feiyu Xu
Bootstrapping Relation Extraction from Semantic Seeds
PHD-Thesis, Saarland University, 2007
<http://www.dfki.de/~feiyu/thesisfeiyuxu.pdf>
3. Feiyu Xu, Hans Uszkoreit, Hong Li
Task driven coreference resolution for relation extraction
Proceedings of the European Conference for Artificial Intelligence ECAI 2008, Patras, Greece, 8/2008
http://www.dfki.de/lt/publication_show.php?id=3016
4. **DARE System: dare.dfki.de**

6. Please describe the basic learning algorithm of the DARE system



Right Answer:

In DARE, the learning and extraction processes interact with each other and are integrated in a bootstrapping framework. The whole algorithm works as follows:

1. Input:

- A set of un-annotated free natural language texts
- A trusted set of relation instances, initially chosen ad hoc by the users, as seed.

2. Text/Passage retrieval: Apply seeds to the documents and divide them into relevant and irrelevant documents. A document is relevant if its text fragments contain a minimal number of the relation arguments of a seed and the distance among individual arguments does not exceed the defined width of the textual window.

3. Pattern extraction: Annotate the relevant text fragments with named entities and linguistic structures and extract linguistic patterns which contain seed relation arguments as their linguistic arguments.

4. Rule induction: Induce relation extraction rules from the set of patterns using compression and generalization methods.

5. Rule Ranking: Rank the rules based on their domain relevance and the trustworthiness of their origin

6. Relation extraction: Apply induced rules to the corpus, in order to extract more relation instances.

7. Ranking and validation: Rank and validate the new relation instances.

8. Stop if no new rules and relation instances can be found.

7. Please describe the rule components of the DARE rule presentation and derive extraction rules and their projections from the dependency tree structure in the example below.

The target relation: *Prize-Awarding Event*, containing four arguments about a person or an organization wins a particular prize in a specific area and in a certain year:

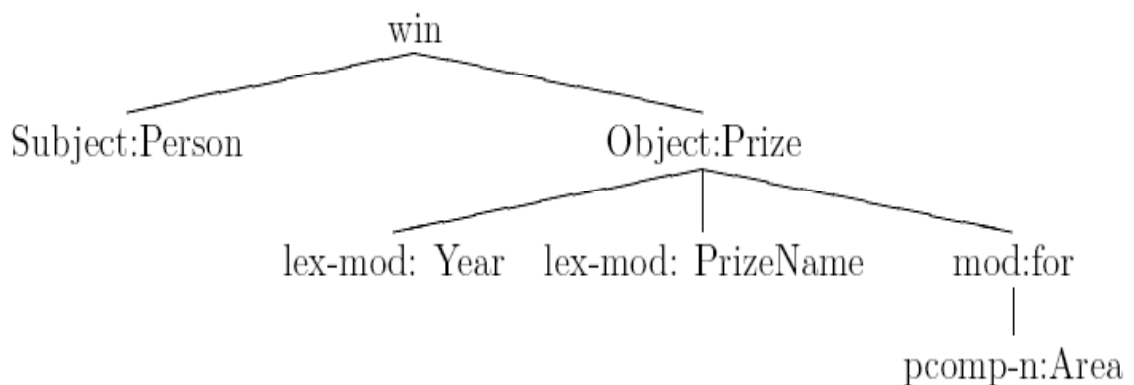
<recipient, award, area, year>

Seed: *<Mohamed ElBaradei, Nobel, Peace, 2005>*

An example sentence matched with the above seed:

Mohamed ElBaradei, won the 2005 Nobel Prize for Peace on Friday for his efforts to limit the spread of atomic weapons.

Dependency Parser Tree of the Example Sentence (simplified):



Right Answer:

A DARE rule has three components:

1. **rule name:** *ri* ;
2. **output:** a set *A* containing *n* arguments of the *n*-ary relation, labelled with their argument roles;
3. **rule body:** an AVM containing:
 - head: the linguistic annotation of the top node of the linguistic structure;
 - daughters: its value is a list of specific linguistic structures (e.g., subject, object, head, mod), derived from the linguistic analysis, e.g., dependency structures and the named entity information;
 - rule: its value is a DARE rule which extracts a subset of arguments of *A*.

Three rules can be derived from the above dependency tree:

(1) extracts the semantic argument *area* from a prepositional phrase, while (2) extracts three arguments *year*, *prize* and *area* from the complex noun phrase and calls the rule (1) for the argument *area*.

(1)

```
Rule name:: area_1
Rule body:: [ head [ pos preposition
                lex-form "for"
              ],
            daughters < [ pcomp-n [ head [1] Area ] ] > ]
Output:: <[1]Area>
```

(2)

```
Rule name:: year_prize_area_1
Rule body:: [ head [ pos noun
                lex-form "prize"
              ],
            daughters < [ lex-mod [ head [1] Year ],
                [ lex-mod [ head [2] Prize ],
                mod [ head [ pos preposition
                        lex-form "for"
                      ],
                    rule area_1:: <[3]Area > ] ] ] > ]
Output:: <[1]Year, [2]Prize, [3]Area>
```

(3) is the rule that extracts all four arguments from the verb phrase dominated by the verb “win” and calls (2) to handle the arguments embedded in the linguistic argument “object”.

(3)

Rule name:: recipient_prize_area_year_1

Rule body::

head	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>verb</td> </tr> <tr> <td style="padding-right: 10px;">mode</td> <td>active</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“win”</td> </tr> </table>	pos	verb	mode	active	lex-form	“win”								
pos	verb														
mode	active														
lex-form	“win”														
daughters	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">subject</td> <td style="padding-left: 10px;"> <table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td>[1 Person]</td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px;">object</td> <td style="padding-left: 10px;"> <table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td style="padding-left: 10px;"> <table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px;">rule</td> <td>year_prize_area_1:: <[4]Year, [2]Prize, [3]Area></td> </tr> </table> </td> </tr> </table>	subject	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td>[1 Person]</td> </tr> </table>	head	[1 Person]	object	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td style="padding-left: 10px;"> <table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px;">rule</td> <td>year_prize_area_1:: <[4]Year, [2]Prize, [3]Area></td> </tr> </table>	head	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table>	pos	noun	lex-form	“prize”	rule	year_prize_area_1:: <[4]Year, [2]Prize, [3]Area>
subject	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td>[1 Person]</td> </tr> </table>	head	[1 Person]												
head	[1 Person]														
object	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">head</td> <td style="padding-left: 10px;"> <table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px;">rule</td> <td>year_prize_area_1:: <[4]Year, [2]Prize, [3]Area></td> </tr> </table>	head	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table>	pos	noun	lex-form	“prize”	rule	year_prize_area_1:: <[4]Year, [2]Prize, [3]Area>						
head	<table border="0" style="border-collapse: collapse;"> <tr> <td style="padding-right: 10px;">pos</td> <td>noun</td> </tr> <tr> <td style="padding-right: 10px;">lex-form</td> <td>“prize”</td> </tr> </table>	pos	noun	lex-form	“prize”										
pos	noun														
lex-form	“prize”														
rule	year_prize_area_1:: <[4]Year, [2]Prize, [3]Area>														

Output:: <[1]Recipient, [2]Prize, [3]Area, [4]Year>