NLP

Introduction to NLP

Relation extraction

Relation Extraction

- Links between entities
  - Works-for
  - Manufactures
  - Located-at

MUC

- Annual competition
  - DARPA, 1990s
- Events in news stories
  - Terrorist events
  - Joint ventures
  - Management changes
- Evaluation metrics
  - Precision
  - Recall
  - F-measure

MUC Example

<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO> Copyright (c) 1989 Jiji Press Ltd.; </SO>
<TXT>
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 NEW TAIWAN 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, BRIDGESTON SPORTS OFFICIALS SAID.
THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID.
BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUB PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.
WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</DOC>

Figure 2: A sample article from the MUC-5 English joint ventures task.

```
<TEMPLATE-0592-1> :=
    DOC NR: 0592
    DOC DATE: 241189
    DOCUMENT SOURCE: "Jiji Press Ltd."
    CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
<TIE_UP_RELATIONSHIP-0592-1> :=
    TIE-UP STATUS: EXISTING
    ENTITY: <ENTITY-0592-1>
            <ENTITY-0592-2>
            <ENTITY-0592-3>
    JOINT VENTURE CO: <ENTITY-0592-4>
    OWNERSHIP: <OWNERSHIP-0592-1>
    ACTIVITY: <ACTIVITY-0592-1>
<ENTITY-0592-1> :=
    NAME: BRIDGESTONE SPORTS CO
    ALIASES: "BRIDGESTONE SPORTS"
             "BRIDGESTON SPORTS"
    NATIONALITY: Japan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-2> :=
    NAME: UNION PRECISION CASTING CO
    ALIASES: "UNION PRECISION CASTING"
    LOCATION: Taiwan (COUNTRY)
    NATIONALITY: Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-3> :=
    NAME: TAGA CO
    NATIONALITY: Japan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-4> :=
    NAME: BRIDGESTONE SPORTS  TAIWAN  CO
    LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
    TYPE: COMPANY
    ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
<INDUSTRY-0592-1> :=
    INDUSTRY-TYPE: PRODUCTION
    PRODUCT/SERVICE: (39 "20,000 IRON AND 'METAL WOOD' [CLUBS]")
<ENTITY_RELATIONSHIP-0592-1> :=
    ENTITY1: <ENTITY-0592-1>
             <ENTITY-0592-2>
             <ENTITY-0592-3>
    ENTITY2: <ENTITY-0592-4>
    REL OF ENTITY2 TO ENTITY1: CHILD
    STATUS: CURRENT
<ACTIVITY-0592-1> :=
    INDUSTRY: <INDUSTRY-0592-1>
    ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
    START TIME: <TIME-0592-1>
<TIME-0592-1> :=
    DURING: 0190
<OWNERSHIP-0592-1> :=
    OWNED: <ENTITY-0592-4>
    TOTAL-CAPITALIZATION: 20000000 TWD
    OWNERSHIP-%: (<ENTITY-0592-3> 10)
             (<ENTITY-0592-2> 15)
             (<ENTITY-0592-1> 75)
```

Figure 3: A sample filled template from the MUC-5 English joint ventures task.

Example from
Grishman and Sundheim 1996

Other Examples

- Job announcements
  - Location, title, starting date, qualifications, salary
- Seminar announcements
  - Time, title, location, speaker
- Medical papers
  - Drug, disease, gene/protein, cell line, species, substance

Filling the Templates

- Some fields get filled by text from the document
  - E.g., the names of people
- Others can be pre-defined values
  - E.g., successful/unsuccessful merger
- Some fields allow for multiple values

Approaches

- View IE as a sequence labeling problem
  - Use HMM
- Use patterns
  - E.g., regular expressions
- Features
  - Capitalization (initial, allcaps), contains digits, spelling (e.g., suffixes), punctuation

Perl Regular Expressions

^	beginning of string; complement inside []
$	end of string
.	any character except newline
*	match 0 or more times
+	match 1 or more times
?	match 0 or 1 times
|	alternatives
( )	grouping and memory
[ ]	set of characters
{ }	repetition modifier
\	special symbol

Perl Regular Expressions

*a**	zero or more
a+	one or more
a?	zero or one
a{*m*}	exactly *m*
a{*m*,}	at least *m*
a{*m*,*n*}	at least *m* but at most *n*
repetition?	shortest match

Perl Regular Expressions

\t	tab
\n	newline
\r	carriage return (CR)
*	asterisk
\?	question mark
\.	period
\x*hh*	hexadecimal character
\w	Matches one alphanumeric (or '_') character
\W	matches the complement of \w
\s	space, tab, newline
\S	complement of \s
\d	same as [0-9]
\D	complement of \d
\b	"word" boundary
\B	complement of \b
[x-y]	inclusive range from x to y

Sample Patterns

- Price (e.g., $14,000.00)
  - `\$[0-9,]+(\.[0-9]{2})?`
- Date (e.g., 2015-02-01)
  - `^(19|20)\d\d[- /.](0[1-9]|1[012])[- /.](0[1-9]|[12][0-9]|3[01])$`
- Email
  - `^[_a-z0-9-]+(\.[_a-z0-9-]+)*@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})$`
- Person
- May include HTML code
- May include POS information
- May include Wordnet information

Sample Input for NER

```
( (S
    (NP-SBJ-1
      (NP (NNP Rudolph) (NNP Agnew) )
      (, ,)
      (UCP
        (ADJP
          (NP (CD 55) (NNS years) )
          (JJ old) )
        (CC and)
        (NP
          (NP (JJ former) (NN chairman) )
          (PP (IN of)
            (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) ))))
      (, ,) )
    (VP (VBD was)
      (VP (VBN named)
        (S
          (NP-SBJ (-NONE- *-1) )
          (NP-PRD
            (NP (DT a) (JJ nonexecutive) (NN director) )
            (PP (IN of)
              (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate) ))))))
    (. .) ))
```

Sample Output for NER (IOB format)

file_id	sent_id	word_id	iob_inner	pos	word
0002	**1**	**0**	**B-PER**	**NNP**	**Rudolph**
0002	**1**	**1**	**I-PER**	**NNP**	**Agnew**
0002	1	2	O	COMMA	COMMA
0002	1	3	B-NP	CD	55
0002	1	4	I-NP	NNS	years
0002	1	5	B-ADJP	JJ	old
0002	1	6	O	CC	and
0002	1	7	B-NP	JJ	former
0002	1	8	I-NP	NN	chairman
0002	1	9	B-PP	IN	of
0002	**1**	**10**	**B-ORG**	**NNP**	**Consolidated**
0002	**1**	**11**	**I-ORG**	**NNP**	**Gold**
0002	**1**	**12**	**I-ORG**	**NNP**	**Fields**
0002	**1**	**13**	**I-ORG**	**NNP**	**PLC**
0002	1	14	O	COMMA	COMMA
0002	1	15	B-VP	VBD	was
0002	1	16	I-VP	VBN	named
0002	1	17	B-NP	DT	a
0002	1	18	I-NP	JJ	nonexecutive
0002	1	19	I-NP	NN	director
0002	1	20	B-PP	IN	of
0002	1	21	B-NP	DT	this
0002	1	22	I-NP	JJ	British
0002	1	23	I-NP	JJ	industrial
0002	1	24	I-NP	NN	conglomerate
0002	1	25	O	.	.

Evaluating Template-based IE

- For each test document
  - Number of correct template extractions
  - Number of slot/value pairs extracted
  - Number of extracted slot/value pairs that are correct

Relation Extraction

- Person-person
  - ParentOf, MarriedTo, Manages
- Person-organization
  - WorksFor
- Organization-organization
  - IsPartOf
- Organization-location
  - IsHeadquarteredAt

ACE Evaluation

- 2002 newspaper data
- Entities:
  - Person, Organization, Facility, Location, Geopolitical Entity
- Relations:
  - Role, Part, Located, Near, Social

Relation Extraction

- Core NLP task
  - Used for building knowledge bases, question answering
- Input
  - **Mazda North American Operations** *is headquartered in* **Irvine, Calif.**, and oversees the sales, marketing, parts and customer service support of Mazda vehicles in the United States and Mexico through nearly 700 dealers.
- Output
  - IsHeadquarteredIn (Mazda North American Operations, Irvine)

Relation Extraction

- Using patterns
  - Regular expressions
  - Gazetteers
- Supervised learning
- Semi-supervised learning
  - Using seeds

Extracting IS-A Relations

- Hearst's patterns
  - X and other Y
  - X or other Y
  - Y such as X
  - Y, including X
  - Y, especially X
- Example
  - Evolutionary relationships between the platypus and other mammals

Supervised Relation Extraction

- Look for sentences that have two entities that we know are part of the target relation
- Look at the other words in the sentence, especially the ones between the two entities
- Use a classifier to determine whether the relation exists

Example

- English
  - **Beethoven** *was born* in December **1770** in Bonn
  - *Born* in Bonn in **1770**, **Beethoven** …
  - After his *birth* on December 16, **1770**, **Beethoven** grew up in a musical family
  - **Ludwig van Beethoven** (**1770**–1827)
  - While this evidence supports the case for 16 December **1770** as **Beethoven**'s *date of birth*

Example (non-English)

- German
  - **Ludwig van Beethoven** *wurde* am 17. Dezember **1770** in Bonn *getauft*
  - **Ludwig van Beethoven** *wurde* in Bonn, 15. Dezember **1770**, eine Familie ursprünglich aus Brabant in Belgien *geboren*
  - Der *Geburtstag* von **Ludwig van Beethoven** wurde im Winter **1770** in Bonn nicht genau dokumentiert
- Spanish
  - **Ludwig van Beethoven** *nació* en Bonn el 17 de diciembre de **1770**
  - *Nacido* en Bonn **1770**, **Beethoven** …
  - **Ludwig van Beethoven**, *nace* en diciembre de **1770**

Semi-supervised Relation Extraction

- Start with some seeds, e.g.,
  - **Beethoven** *was born* in December **1770** in Bonn
- Look for other sentences with the same words
- Look for expressions that appear nearby
- Look for other sentences with the same expressions

Evaluating Relation Extraction

- Precision P
  - correctly extracted relations/all extracted relations
- Recall R
  - correctly extracted relations/all existing relations
- F1 measure
  - F1 = 2PR/(P+R)
- If there is no annotated data
  - only measure precision

Conclusion

- Probabilistic NLP
- Part of Speech Tagging
- Hidden Markov Models
- Information Extraction

NLP