NLP
Introduction to NLP

*Information Extraction*
Information Extraction

• Usually from unstructured or semi-structured data
• Examples
  – News stories
  – Scientific papers
  – Resumes
• Entities
  – Who did what, when, where, why
• Build knowledge base
Named Entities

• Types:
  – People
  – Locations
  – Organizations
    • Teams
    • Newspapers
    • Companies
  – Geo-political entities

• Ambiguity:
  – London can be a person, city, country (by metonymy) etc.

• Useful for interfaces to databases, question answering, etc.
Times and Events

• **Times**
  – Absolute expressions
  – Relative expressions (e.g., “last night”)

• **Events**
  – E.g., a plane went past the end of the runway
Sequence labeling

• Many NLP problems can be cast as sequence labeling problems
  – POS – part of speech tagging
  – NER – named entity recognition
  – SRL – semantic role labeling

• Input
  – Sequence $w_1w_2w_3$

• Output
  – Labeled words

• Classification methods
  – Can use the categories of the previous tokens as features in classifying the next one
  – Direction matters
Named Entity Recognition (NER)

• Segmentation
  – Which words belong to a named entity?
  – Brazilian football legend Pele's condition has improved, according to a Thursday evening statement from a Sao Paulo hospital.

• Classification
  – What type of named entity is it?
  – Use gazetteers, spelling, adjacent words, etc.
  – Brazilian football legend [PERSON Pele]'s condition has improved, according to a [TIME Thursday evening] statement from a [LOCATION Sao Paulo] hospital.
NER, Time, and Event extraction

- Brazilian football legend [PERSON Pele]'s condition has improved, according to a [TIME Thursday evening] statement from a [LOCATION Sao Paulo] hospital.

- There had been earlier concerns about Pele's health after [ORG Albert Einstein Hospital] issued a release that said his condition was "unstable."

- [TIME Thursday night]'s release said [EVENT Pele was relocated] to the intensive care unit because a kidney dialysis machine he needed was in ICU.
Biomedical example

• Gene labeling
• Sentence:
  – [GENE BRCA1] and [GENE BRCA2] are human genes that produce tumor suppressor proteins
MUC

• Annual competition
  – DARPA, 1990s

• Events in news stories
  – Terrorist events
  – Joint ventures
  – Management changes

• Evaluation metrics
  – Precision
  – Recall
  – F-measure
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 KAOSIUNG, 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.

Figure 2: A sample article 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
l 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
\xhh 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)[0-9]{2})-[0-9]{2}-[012][0-9]$ 

- 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
NLP