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

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 $w_1 w_2 w_3$
- 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 [$_{\text{PERSON}}$ Pele]'s condition has improved, according to a [$_{\text{TIME}}$ Thursday evening] statement from a [$_{\text{LOCATION}}$ Sao Paulo] hospital.

NER, Time, and Event Extraction

- Brazilian football legend [$_{\text{PERSON}}$ Pele]'s condition has improved, according to a [$_{\text{TIME}}$ Thursday evening] statement from a [$_{\text{LOCATION}}$ Sao Paulo] hospital.
- There had been earlier concerns about Pele's health after [$_{\text{ORG}}$ Albert Einstein Hospital] issued a release that said his condition was "unstable.“
- [$_{\text{TIME}}$ Thursday night]'s release said [$_{\text{EVENT}}$ Pele was relocated] to the intensive care unit because a kidney dialysis machine he needed was in ICU.

Biomedical Example

- Gene labeling
- Sentence:
  - [$_{\text{GENE}}$ BRCA1] and [$_{\text{GENE}}$ BRCA2] are human genes that produce tumor suppressor proteins

NLP