CSC 594 Topics in AI – Applied Natural Language Processing

Fall 2009/2010

Overview of NLP tasks (text pre-processing)

Typical Text Pre-processing Step

- Given a raw text (in a corpus), we typically pre-process the text by applying the following tasks in order:
  1. **Part-Of-Speech (POS) tagging** – assign a POS to every word in a sentence in the text
  2. **Named Entity Recognition (NER)** – identify named entities (proper nouns and some common nouns which are relevant in the domain of the text)
  3. **Shallow Parsing** – identify the phrases (mostly verb phrases) which involve named entities
  4. **Information Extraction (IE)** – identify relations between phrases, and extract the relevant/significant "information" described in the text

Source: Andrew McCallum, UMass Amherst

1. Part-Of-Speech (POS) Tagging

- POS tagging is a process of assigning a POS or lexical class marker to each word in a sentence (and all sentences in a corpus).

Input: `the lead paint is unsafe`

Output: `the/Det lead/N paint/N is/V unsafe/Adj`

2. Named Entity Recognition (NER)

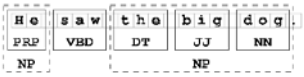
- NER is to process a text and identify named entities in a sentence
  - e.g. "U.N. official Ekeus heads for Baghdad."

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

4

3. Shallow Parsing

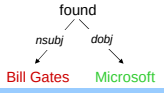
- Shallow (or Partial) parsing identifies the (base) syntactic phases in a sentence.



[$_{NP}$ He] [$_{V}$ saw] [$_{NP}$ the big dog]

- After NEs are identified, **dependency parsing** is often applied to extract the syntactic/dependency relations between the NEs.

[$_{PER}$ Bill Gates] founded [$_{ORG}$ Microsoft].



Dependency Relations
nsubj(Bill Gates, found)
dobj(found, Microsoft)

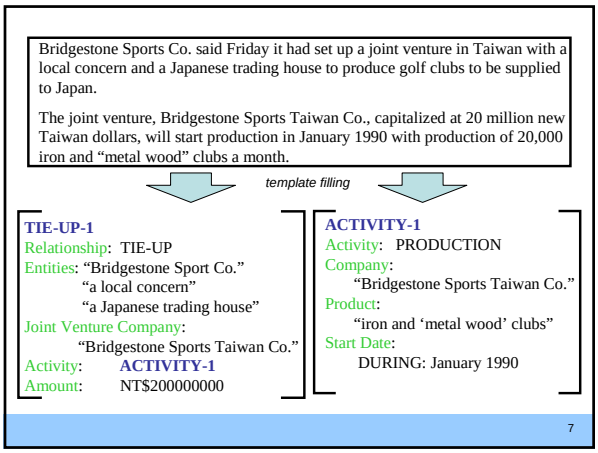
5

4. Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured text
- Transform unstructured information in a corpus of texts or web pages into a structured database (or templates)
- Applied to various types of text, e.g.
  - Newspaper articles
  - Scientific articles
  - Web pages
  - etc.

Source: J. Choi, CSE842, MSU

6


Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied 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.
template filling
TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000
ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990
7