Named Entity Recognition

- What is NE?
- What isn't NE?
- Problems and solutions with NE task definitions
- Problems and solutions with NE task
- Some applications

Why do NE Recognition?

- Key part of Information Extraction system
- Robust handling of proper names essential for many applications
- Pre-processing for different classification levels
- Information filtering
- Information linking

NE Definition

- NE involves **identification** of *proper names* in texts, and **classification** into a set of predefined categories of interest.
- Three universally accepted categories: **person**, **location** and **organisation**
- Other common tasks: recognition of date/time expressions, measures (percent, money, weight etc), email addresses etc.
- Other domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

<PER>**Prof. Jerry Hobbs**</PER> taught CS544 during <DATE>**February 2010**</DATE>.
<PER>**Jerry Hobbs**</PER> killed his daughter in <LOC>**Ohio**</LOC>.

- Identify mentions in text and classify them into a predefined set of categories of interest:
  - Person Names: **Prof. Jerry Hobbs**, **Jerry Hobbs**
  - Organizations: **Hobbs corporation**, **FbK**
  - Locations: **Ohio**
  - Date and time expressions: **February 2010**
  - E-mail: **mkg@gmail.com**
  - Web address: **www.usc.edu**
  - Names of drugs: **paracetamol**
  - Names of ships: **Queen Marry**
  - Bibliographic references:

There are Eleven types of entities in Name as given below.

Person: Person entities are limited to humans. A person may be a single individual or a group.

Organization: Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.

Location: Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.

Facilities: Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.

Locomotives: A locomotive entity is a physical device primarily designed to move an object from one location to another, by (for example) carrying, pulling, or pushing the transported object. Vehicle entities may or may not have their own power source.

Artifacts: Artifact entities are objects or things, which are produced or shaped by human craft, such as tools, weapons/ammunition, art paintings, clothes, ornaments, medicines.

Entertainment: Entertainment entities denote activities, which are diverting and hold human attention or interest, giving pleasure, happiness, amusement especially performance of some kind such as dance, music, sports, events.

Cuisine's: This entity refers to various type of food, prepared in different manners such as Chinese food, South-Indian, North-Indian foods.

Organisms: Organism entities are living things and have the ability to act or function independently such as humans, viruses, bacteria etc. Here we have not taken into consideration plants, those have been classified separately as 1.10.

Plants: These entities are living things having photosynthetic, eukaryotic, multicellular organisms of the kingdom Plantae, containing chloroplasts, having cellulose cell walls, and lacking the power of locomotion.

Disease: This entity refers to the state of a disordered or incorrectly functioning organ, part, structure, or system of the body resulting from the effect of genetic or developmental errors, infection, poisons, nutritional deficiency or imbalance, toxicity, or unfavorable environmental factors; illness; sickness; ailment such as fever, cancer etc.

Problems in NE Task Definition

Person vs. Artefact: "**Sandwich** wants his bill." vs "Bring me a **sandwich**."

Organisation vs. Location : "**England** won the World Cup" vs. "The World Cup took place in **England**".

Company vs. Artefact: "shares in **MTV**" vs. "watching **MTV**"

Location vs. Organisation: "she met him at **Heathrow**" vs. "the **Heathrow** authorities"

Basic Problems in NE

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
  - John Smith (company vs. person)
  - May (person vs. month)
  - Washington (person vs. location)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. “may”

More complex problems in NER

- Issues of style, structure, domain, genre etc.
  - Punctuation, spelling, spacing, formatting, ….all have an impact

Dept. of Computing and Maths
Manchester Metropolitan University
Manchester
United Kingdom

List Lookup Approach

- System that recognises only entities stored in its lists (gazetteers).
- Advantages - Simple, fast, language independent, easy to retarget
- Disadvantages – collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Shallow Parsing Approach

- Internal evidence – names often have internal structure. These components can be either stored or guessed.

location:

CapWord + {City, Forest, Center}

e.g. Sherwood Forest

Cap Word + {Street, Boulevard, Avenue, Crescent, Road}

e.g. *Portobello Street*

Shallow Parsing Approach

- External evidence - names are often used in very predictive local contexts

Location:

"to the" COMPASS "of" CapWord

e.g. *to the south of* ***Loitokitok***

"based in" CapWord

e.g. *based in* ***Loitokitok***

CapWord "is a" (ADJ)? GeoWord

e.g. ***Loitokitok*** *is a friendly city*

Difficulties in Shallow Parsing Approach

- **Ambiguously capitalised words** (first word in sentence)

  [All American Bank] vs. All [State Police]

- **Semantic ambiguity**

  "John F. Kennedy" = airport (location)

  "Philip Morris" = organisation

- **Structural ambiguity**

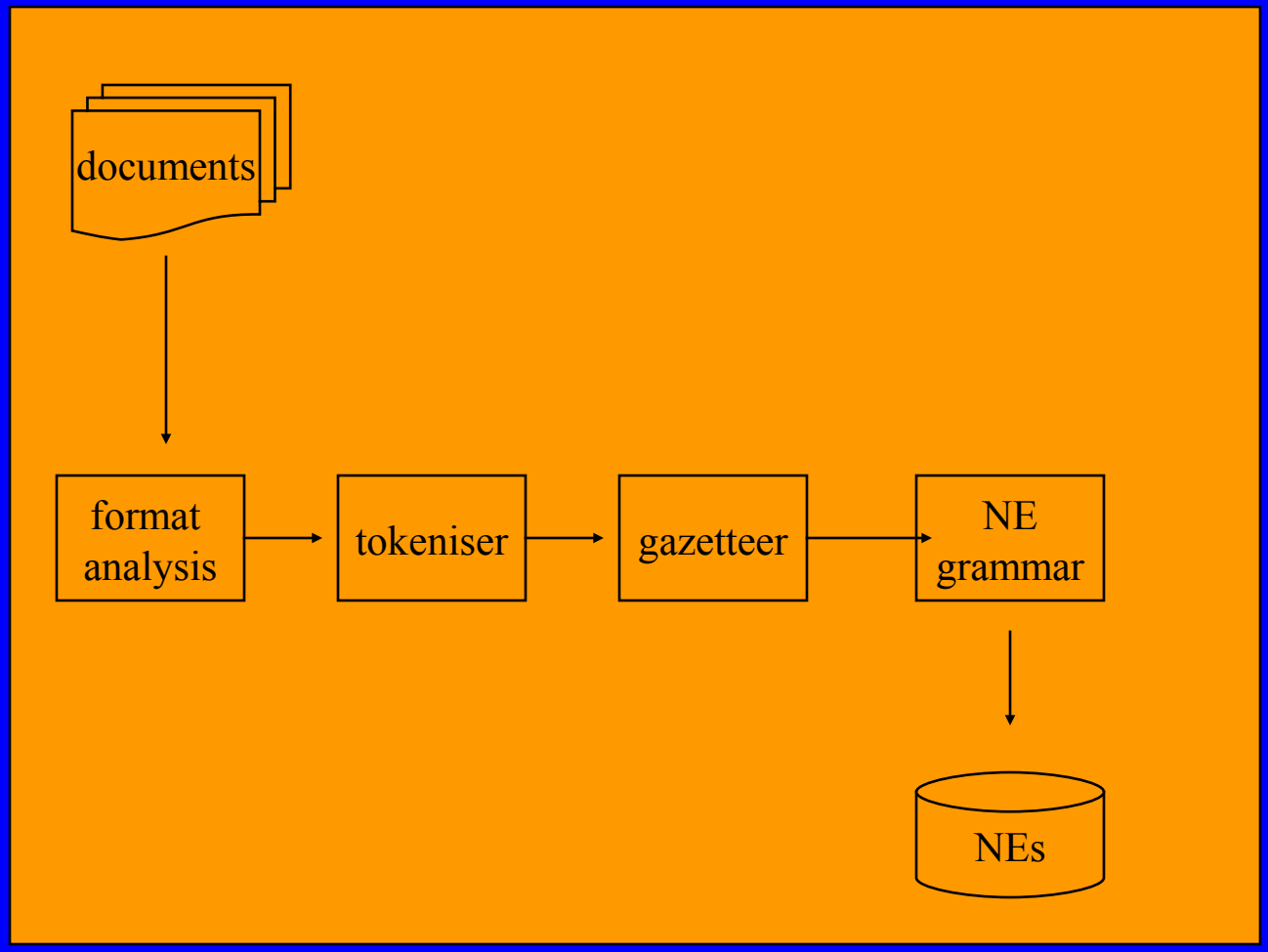
  [Cable and Wireless] vs. [Microsoft] and [Dell]

  [Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith].

Technology

- JAPE (Java Annotations Pattern Engine)

NE System Architecture



Modules

- **Tokeniser**
  - **segments text into tokens, e.g. words, numbers, punctuation**
- **Gazetteer lists**
  - **NEs, e.g. towns, names, countries, ...**
  - **key words, e.g. company designators, titles, ...**
- **Grammar**
  - **hand-coded rules for NE recognition**

JAPE

- Set of phases consisting of pattern /action rules
- Phases run sequentially and constitute a cascade of FSTs over annotations
- LHS - annotation pattern containing regular expression operators
- RHS - annotation manipulation statements
- Annotations matched on LHS referred to on RHS using labels attached to pattern elements

Tokeniser

- Set of rules producing annotations
- LHS is regular expression matched on input
- RHS describes annotations to be added to AnnotationSet

```
(UPPERCASE _LETTER)
   (LOWERCASE_LETTER)* >
    Token; orth = upperInitial; kind = word
```

Gazetteer

- Set of lists compiled into Finite State Machines
- Each list has attributes MajorType and MinorType (and optionally, Language)

city.lst: location: city

currency_prefix.lst: currency_unit: pre_amount

currency_unit.lst: currency_unit: post_amount

Named entity grammar

- hand-coded rules applied to annotations to identify NEs
- annotations from format analysis, tokeniser and gazetteer modules
- use of contextual information
- rule priority based on pattern length, rule status and rule ordering

Example of JAPE Grammar rule

```
Rule:     Location1
Priority: 25
( ( { Lookup.majorType == loc_key,
      Lookup.minorType == pre}
    { SpaceToken} )?
  { Lookup.majorType == location}
  ( {SpaceToken}
    { Lookup.majorType == loc_key,
      Lookup.minorType == post} ) ?
)
 : locName -->
       :locName.Location = { kind = “gazetteer”, rule =
         Location1
     }
```

MUSE

- **MUlti-Source Entity recognition**
- **Named entity recognition from a variety of text types, domains and genres.**
- **2 years from Feb 2000 – 2002**
- **Sponsors: GCHQ**

PASTA

- **Protein Active Site Template Acquisition**
- **Aim: Use of IE techniques to create a database of protein active site data to support protein structure analysis**
- **Partners: Dept. of Computer Science, Information Studies, Mol. Biology and Biotechnology, Univ. of Sheffield**
- **Sponsors: BBSRC-EPSRC Bioinformatics Initiative**

MUMIS

- **MUltiMedia Indexing and Searching environment**
- **Application of IE technology to multimedia, multilingual video indexing in football domain**
- **2 years: June 2000 - 2002**
- **CTIT (NL), University of Sheffield (UK), DFKI (D), Max Planck Institute (D), University of Nijmegen (NL), ESTeam (SWE), VDA (NL)**