Named Entity Recognition

Dan Smith

UEA
University of East Anglia

Overview

- Why is named entity identification important?
- What is NER?
- Shallow parsing
- NLTK

3

Motivation

- Named entities are important:
  - identify places, people, organisations, …
  - improve search
  - help question answering
- Robust handling of proper names essential for many applications

4

WHAT IS NAMED ENTITY IDENTIFICATION?

5

What is NER?

- 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, ...), email addresses etc.
- Other domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

6

NER issues

- Ideally need to do more than just matching text strings with pre-defined lists of names.
- Need to recognise when terms are being used as NEs in a given context
- This is difficult... so for IR it's usually sufficient to identify the NEs regardless of context

7

NER issues

- Variation of forms
  - John Smith, Mr John Smith, J Smith, …
- Ambiguity of NE types
  - "John Smith's gone to London…" (person)
  - "John Smith's Brewery in Tadcaster" (organisation)
- Punctuation, formatting, spacing can all be important in NER

8

Gazetteer-based approach

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

9

Naïve gazetteer lookup

KEEP UP ON (Vietnam) YOUR READING (UK) WITH AUDIO BOOKS (Louisiana, USA)

Audio books (Louisiana, USA) are highly popular (S.Carolina, USA) with library (Pennsylvania, USA) patrons in the town (Mass., USA)

of (Turkey) Springfield, (Virginia, USA) Greene (Maine, USA) County, MO. (Norway) "People are mobile (Alabama, USA)

and busier, and audio books (Louisiana, USA) fit into that lifestyle" says Gary (Indiana, USA)

Sanchez, (Dominican Republic) who oversees the library's (Pennsylvania, USA) $2 million (Kentucky, USA) budget...

10

NE issues (2)

- Some types of NE change rapidly so are difficult to list exhaustively
  - organisations
  - personal names
- Many NEs have multiple meanings
  - May: month, person, May Day, May Ball, …
  - Christian Dior: person or organisation
- Most NEs are multi-token sequences

11

SHALLOW PARSING

12

Shallow parsing: basic approach

1. Split document into sentences
2. PoS tagging
3. Chunking to identify noun phrases
4. Apply rules to identify NEs in noun phrases

- NLTK has modules for these tasks, including a pre-trained NE recogniser
  nltk.ne_chunk()
  that adds NE class labels

13

Shallow parsing rules

- Names often have internal structure. These components can be either stored or guessed.

location:

CapWord + {City, Forest, Center}

e.g. Thetford Forest

CapWord + {Street | Boulevard | Avenue | Crescent | Road}

e.g. *London Street*

14

Shallow parsing approach

- Based on external evidence
  - names are often used in very predictable local contexts

  location:

  "to the" COMPASS "of" CapWord

  to the south of ***Norwich***

  "based in" CapWord

  based in ***Norwich***

  CapWord "is a" (ADJ)? GeoWord

  Norwich *is a friendly city*

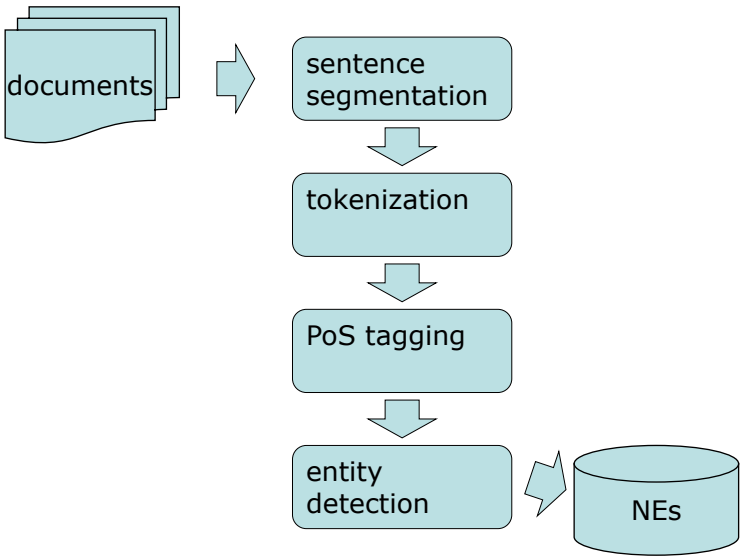
15

Issues with shallow parsing approach

- Ambiguously capitalised words
  - first word in sentence
    [All American Bank] vs. All [District Council]
- Semantic ambiguity
  - "John F. Kennedy" = airport (location)
  - "Philip Morris" = organisation
- Structural ambiguity
  - [Cable and Wireless] vs. [Microsoft] and [Apple]
  - [Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith]

16

NER system architecture



17

Modules

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

18

NER with NLTK

ORGANIZATION	Wiggin LLP, WHO
PERSON	Eddy Bonte, President Obama
LOCATION	River Yare, Mount Everest
DATE	June, 2008-06-29
TIME	two fifty a m, 1:30 p.m.
MONEY	175 million US Dollars, GBP 10.40
PERCENT	twenty pct, 18.75%
FACILITY	Nelson's Column, Stonehenge
GPE (geo-political entity)	South East Asia, Midlothian

19

Resources

- Bird S., Klein E., Loper E. (2009) *Natural Language Processing with Python*, O'Reilly (Chapter 7)
  http://www.nltk.org/book/