Named Entity Recognition

27 November 2024 19:13

Locate + classify atomic elements into predefined categories.

- Entity name
- Same thing in all possible worlds. (Rigid designators)

Problems:

- Variation of NE
- Ambiguity of types — person v loc^n, person v organiz^n
- Punctuation

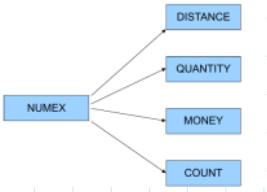
Annotating:

```
<ENAMEX TYPE= "PERSON" SUBTYPE_1= "INDIVIDUAL">
abc </ENAMEX>
```

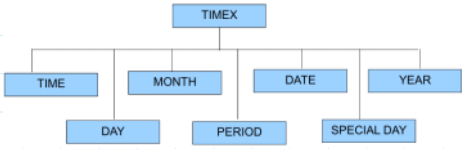
NE types

ENAMEX

NUMEX



TIMEX



A. Dictionary Based NER

- Use gazetteers — contains NE from all domains.
- Adv:
  - simple approach
  - high precision
- Disadv:
  - exhaustive dict is hard.
  - diff spellings.

B. Rule based NER

- Regular expression to get: ph no., email, capitalized names
  - capitalized names: first word of sentence? nested named entity? noun in German?
- Adv:
  - rich, expressive rules
  - good results.
- Disadv:
  - huge experience + grammatical knowledge
  - experts are expensive.
  - highly domain specific

C. Building Decision Trees

- Select attr @ each node
  - ↓
  - most useful for classifying samples.

Select word to check node

↓

most useful for classifying samples.

- Top-down greedy search thro' space of possible decision trees.
- Pick best attr and never look back.

IOB tagging

B = word in beginning chunk
O = outside any chunk
I = Inside chunk.

} named entity detection.