Chapter NLP:III

III. Words

- Word-level Phenomena
- Morphological Analysis
- Word Classes
- Named Entities

Named Entities

Entities

An entity represents an object from the real world. They are a basic semantic concept in natural language processing.

- Named entities are objects that can be denoted with a proper name.

  `Prof. Dr. Abdul Nachtigaller in Finsterberge at Nachtschule`

- Numeric entities are values, quantities, proportions, ranges, or similar.

  `in this year  2018-10-18  $ 100 000  60-68 44`

Named Entities

Named Entities

Named entities are the semantic equivalent of proper nouns: Everything that can be referred to by name is an entitiy.

The most common types of named entities are:

- **PER** (Person): people, characters, ...
  Turing is a giant of computer science.
- **LOC** (Location): regions, rivers, ...
  The Ilm is a small river.
- **ORG** (Organization): companies, sports teams, ...
  The IPCC warned about climate change.
- **GPE** (Geo-Political Entity): countries, states, ...
  Weimar lies in Thuringia.

[Deere and Co. ORG] said it reached a tentative agreement with the [machinists union ORG] at its [Horicon, Wis. LOC] plant, ending a month-old strike by workers at the facility.

Named Entities

Named Entities

A more complete set of entities is used by OntoNotes. [Weischedel et al]

Names, Named Entities	
PERSON	People, including fictional
NORP	Nationalities, parties, . . .
FACILITY	Buildings, highways, . . .
ORGANIZATION	Companies, institutions, . . .
GPE	Countries, cities, . . .
LOCATION	mountains, rivers, . . .
PRODUCT	Vehicles, foods, . . .
EVENT	Hurricanes, sports events , . . .
WORK OF ART	Titles of books, songs, . . .
LAW	Named documents, laws
LANGUAGE	Any named language

Values	
DATE	Dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	weights, distances, /dots
ORDINAL	“first”, “second”
CARDINAL	other numerals

Although there is a linguistic difference between *entities* and *values* they are often treated as equivalent in NLP.

Remarks:

Named entity tagsets vary by corpus and use case:

- Spacy uses the OntoNotes Tagset for its English models.
- 7 Entity types (NameType) are recognized by Universal Dependencies
  GEO (Geographical Name), GIV (Given Name), SUR (Surname), NAT (Nationality), COM (Company), PRO (product), OTH (other)
- 6 Entity types by WNUT Emerging Entity Recognition [Derczynski et al.]
  PERSON, LOCATION (GPE, facility), CORPORATION, PRODUCT (tangible goods, well-defined services), CREATIVE-WORK (song, movie, book), GROUP (music band, sports team, non-corporate organisations)
- 64 Entity types (incl. subtypes) by the BBN Pronoun Coreference and Entity Type Corpus
  BBN annotates entity types and subtypes from 3 groups of entities in XML:

<ENAMEX TYPE="ORGANIZATION:CORPORATION">*Deere and Co.*</ENAMEX> *said it reached a tentative agreement with the*<ENAMEX TYPE="PER_DESC"> *machinists* </ENAMEX> <ENAMEX TYPE="ORG_DESC:OTHER"> *union*</ENAMEX> *[. . .] ending a* <TIMEX TYPE="DATE:AGE"> *month-old* </TIMEX> *strike*.

Named Entities

Named Entity Recognition

Finding and labeling entities in a text is called Named Entity Recognition. Alternative: Named Entity Tagging, Named Entity Resolution.

NER is a span recognition problem. Entities often span multiple tokens, so a tagger needs to:

- Distinguish entities from non-entities.

  apple vs. [apple ORG]
- Find the boundaries of the entity.

  the [brandenburg LOC] gate *vs* [the brandenburg gate LOC]

  on [Washington's LOC] [Capitol Hill LOC]
- Disambiguate different entity types.

  [Washington PER] *vs* [Washington LOC] *vs* [Washington GPE] *vs* [Washington ORG]

Span recognition problems are typically solved by BIO tagging.

Named Entities

BIO Tagging

Idea: Model NER as a sequence labeling problem and tag word-by-word. Encode boundary and entity type in each tag.

BIO tagging:

1. Assign the first token in an entity a B for beginning and its tag.
   on Washington's/B-LOC Capitol/B-LOC Hill stands ...
2. Assign all non-first tokens in an entity a I for inside and its tag.
   on Washington's/B-LOC Capitol/B-LOC Hill/I-LOC stands ...
3. Assign all non-entity tokens an O for outside.
   on/O Washington's/B-LOC Capitol/B-LOC Hill/I-LOC stands/O ...

As span recognition problem: [Deere and Co. ORG] said it ...

As sequence labeling problem: Deere/B-ORG and/I-ORG Co./I-ORG said/O it/O ...

Now we can solve NER with any sequence labeler.

Remarks:

- Two popular variations of BIO are IO and BIOES.
- IO is a generalization of BIO and encodes less information. Each B-TAG is instead tagges as I-TAG. It might be easier to fit models with IO if resources are scarse.
- BIOES is an extention of BIO and encode more information. The last token in an entity is tagges as E-TAG for ending. If entities consist of only one token, it is tagged as S-TAG for single.