Lecture 6
Sequence Tagging

LTAT.01.001 – Natural Language Processing
Kairit Sirts (kairit.sirts@ut.ee)
24.03.2021
Plan for today

- The sequence tagging problem
  - POS tagging
  - Morphological tagging
  - (CONLL-U data format)
  - Named Entity Recognition

- Evaluating Sequence tagging tasks
- Old school methods for sequence tagging
Sequence tagging problem

- Many NLP problems can be viewed as sequence tagging/labelling.
- Each token in a sequence is assigned a tag/label.
- Labels of tokens are dependent on the labels of other tokens in the sequence, particularly their neighbors.

```
<table>
<thead>
<tr>
<th>Label</th>
<th>Label</th>
<th>Label</th>
<th>Label</th>
<th>Label</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>cat</td>
<td>sat</td>
<td>on</td>
<td>the</td>
<td>mat</td>
</tr>
</tbody>
</table>
```
Sequence tagging task examples

- POS (part-of-speech) tagging
- Named entity recognition
- Chinese word segmentation
- Shallow parsing (also called chunking)
- Semantic role labeling
POS tagging
POS tags

- Part-of-speech tags, syntactic categories, word classes
- POS tags give information about the word and its neighbors
- Useful for many other NLP tasks: information extraction, syntactic parsing, information retrieval, summarization

Janet will back the bill

Proper noun  Modal verb  Verb  Article  Noun
Main POS categories

● Nouns
  ● Common nouns – things (chair), events (lecture), abstractions (justice), verb-like terms (swimming) etc
  ● Proper nouns – proper names of people (John), countries (Estonia), organizations (University of Tartu) etc

● Verbs – words referring to actions and processes (to draw, to ponder)
● Adjectives – words describing properties or qualities (black, young)
● Adverbs – words modifying (mostly) verbs

Unfortunately, John walked home extremely slowly yesterday
Open and closed class words

Open class words
- Nouns
- Verbs
- Adjectives
- Adverbs

Closed class words
- **Prepositions:** on, under, over
- **Determiners:** a, an, the
- **Pronouns:** she, who, I, others
- **Conjunctions:** and, but, or, as
- **Auxiliary verbs:** can, may, are
- **Particles:** up, down, on, off
- **Numerals:** one, two, first
Universal POS tags

<table>
<thead>
<tr>
<th>Open class words</th>
<th>Closed class words</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADP</td>
<td>PUNCT</td>
</tr>
<tr>
<td>ADV</td>
<td>AUX</td>
<td>SYM</td>
</tr>
<tr>
<td>INTJ</td>
<td>CCONJ</td>
<td></td>
</tr>
<tr>
<td>NOUN</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>PROPN</td>
<td>NUM</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>PART</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCONJ</td>
<td></td>
</tr>
</tbody>
</table>

Source: https://universaldependencies.org/u/pos/index.html
POS tagging

- UD POS tags

```
DET NOUN VERB ADP DET NOUN
```

```
The cat sat on the mat
```

- WSJ POS tags

```
DT NN VBD IN DT NN
```

```
The cat sat on the mat
```
Morphological tagging
Morphology

- Morphology studies the internal structure of words

availability NOUN +nom +pl

availabilities

avail +able +ity +es

avail_abil_iti_es
Morphemes

- Morphemes are the smallest units of language that carry a semantic meaning.

**availabilities**

avail +able +ity +es

- Derivational suffix that transforms an adjective into a noun
- Nominative plural
- Inflectional suffix
- Derivational suffix that transforms a verb into an adjective
- Verbal root
Lexical and grammatical morphemes

- Lexical morphemes carry themselves a semantic meaning. Most of them can stand on their own.
  - Boy, table, yellow, run, waste etc

- Grammatical morphemes cannot stand on their own. Their role is to modify the meaning of a lexical morpheme or specify the relationships between lexical morphemes
  - -s, -ing, -able, at, in, on
The role of grammatical morphemes

- Overlap with syntax and semantics

“I put the book on the table” vs “Panin raamatu lauale” (in Estonian)

“Giraffe bit the zebra” vs “Kaelkirjak hammustas sebrat” or “Sebrat hammustas kaelkirjak”
Morphological analysis

- The task of finding all possible morphological tags for a word
- A morphological analysis consists of:
  - Lemma/stem
  - POS
  - Morphological attributes/features

**Question:**

*Should morphological analysis be done in context or can you analyse each word in isolation?*
Morphological analysis

- The task of finding all possible morphological tags for a word
- A morphological analysis consists of:
  - Lemma/stem
  - POS
  - Morphological attributes/features

**Question:**
Should morphological analysis be done in context or can you analyse each word in isolation?

**Answer:** Can be done in isolation

(something) has lasted

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>kestnud</td>
<td>/<em>V</em> nud,</td>
</tr>
<tr>
<td></td>
<td>/<em>S</em> pl n,</td>
</tr>
<tr>
<td></td>
<td>/<em>A</em> //</td>
</tr>
<tr>
<td></td>
<td>/<em>A</em> sg n,</td>
</tr>
<tr>
<td></td>
<td>/<em>A</em> pl n,</td>
</tr>
</tbody>
</table>

kestnud
Morphological tagging

- Can be treated as a sequence tagging task
- Conceptually very similar to POS tagging – instead of POS tags there are now morphological tags
- The UD (universal dependencies) datasets also contain morphological analyses for many languages
### Morphological tagset sizes in UD corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>Tagset Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>349</td>
</tr>
<tr>
<td>Chinese</td>
<td>31</td>
</tr>
<tr>
<td>Czech</td>
<td>2630</td>
</tr>
<tr>
<td>English</td>
<td>117</td>
</tr>
<tr>
<td>Estonian</td>
<td>662</td>
</tr>
<tr>
<td>Finnish</td>
<td>2052</td>
</tr>
<tr>
<td>French</td>
<td>228</td>
</tr>
<tr>
<td>German</td>
<td>684</td>
</tr>
<tr>
<td>Korean</td>
<td>11</td>
</tr>
<tr>
<td>Russian</td>
<td>693</td>
</tr>
</tbody>
</table>
# Universal morphological features

<table>
<thead>
<tr>
<th>Lexical features</th>
<th>Inflectional features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nominal</strong>*</td>
<td><strong>Verbal</strong>*</td>
</tr>
<tr>
<td>PronType</td>
<td>Gender</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
</tr>
<tr>
<td>Poss</td>
<td>NounClass</td>
</tr>
<tr>
<td>Reflex</td>
<td>Number</td>
</tr>
<tr>
<td>Foreign</td>
<td>Case</td>
</tr>
<tr>
<td>Abbr</td>
<td>Definite</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
</tr>
<tr>
<td></td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td>Clusivity</td>
</tr>
</tbody>
</table>
Universal Dependencies data

- [https://universaldependencies.org/](https://universaldependencies.org/)
- Manually annotated data in more than 70 languages
  - POS tags
  - Lemmas
  - Morphology
  - Syntax
- All datasets are in CONLL-U format
CONLL-U format

- A standard tabular format for certain type of annotated data
- Each word is in a separate line
- 10 tab-separated columns on each line:
  1. Word index
  2. The word itself
  3. Lemma
  4. Universal POS
  5. Language specific POS
  6. Morphological features
  7.-9. Information related to syntactic information
  10. Any other annotation
CONLL-U format: example

# text = They buy and sell books.
1  They  they  PRON  PRP  Case=Nom|Number=Plur
2  buy  buy  VERB  VBP  Number=Plur|Person=3|Tense=Pres
3  and  and  CONJ  CC  _
4  sell  sell  VERB  VBP  Number=Plur|Person=3|Tense=Pres
5  books  book  NOUN  NNS  Number=Plur
6  .  .  PUNCT  .  _

# text = I had no clue.
1  I  I  PRON  PRP  Case=Nom|Number=Sing|Person=1
2  had  have  VERB  VBD  Number=Sing|Person=1|Tense=Past
3  no  no  DET  DT  PronType=Neg
4  clue  clue  NOUN  NN  Number=Sing
5  .  .  PUNCT  .  _
The role of computational morphology in NLP

- Still largely an unexplored area. Why?
- Intuitively, modeling morphology should help to reduce the vocabulary sparsity problems
- Morphological agreement:
  - For instance, verbs and nouns must agree in number
  - In German: der Mann geht vs die Männer gehen (the man goes vs men go)
- Potentially could be useful for many downstream tasks:
  - Machine translation
  - Natural language generation
  - Language modeling (for speech recognition)
Named Entity Recognition
NER: Named entity recognition

- Find all *named entities* in the text and label their types

At the party **Thursday night** at **Chateau Marmont**, **Cate Blanchet** barely made it up in the elevator.
### Types of named entities

<table>
<thead>
<tr>
<th>Type</th>
<th>Tag</th>
<th>Sample Categories</th>
<th>Example sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>PER</td>
<td>people, characters</td>
<td>Turing is a giant of computer science.</td>
</tr>
<tr>
<td>Organization</td>
<td>ORG</td>
<td>companies, sports teams</td>
<td>The IPCC warned about the cyclone.</td>
</tr>
<tr>
<td>Location</td>
<td>LOC</td>
<td>regions, mountains, seas</td>
<td>The Mt. Sanitas loop is in Sunshine Canyon.</td>
</tr>
<tr>
<td>Geo-Political</td>
<td>GPE</td>
<td>countries, states, provinces</td>
<td>Palo Alto is raising the fees for parking.</td>
</tr>
<tr>
<td>Entity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facility</td>
<td>FAC</td>
<td>bridges, buildings, airports</td>
<td>Consider the Tappan Zee Bridge.</td>
</tr>
<tr>
<td>Vehicles</td>
<td>VEH</td>
<td>planes, trains, automobiles</td>
<td>It was a classic Ford Falcon.</td>
</tr>
</tbody>
</table>
Types of named entities correspond to proper names

Often also **temporal expressions** and **numeric expressions** are extracted

Types of named entities may depend on the domain
- Names of genes and proteins in biomedical texts
- Names of college courses
- Commercial products
- Works of art
Citing high fuel prices, [United Airlines ORG] said [Friday TIME] it has increased fares by [$6 MONEY] per round trip on flights to some cities also served by lowercost carriers. [American Airlines ORG], a unit of [AMR Corp. ORG], immediately matched the move, spokesman [Tim Wagner PER] said. [United ORG], a unit of [UAL Corp. ORG], said the increase took effect [Thursday TIME] and applies to most routes where it competes against discount carriers, such as [Chicago LOC] to [Dallas LOC] and [Denver LOC] to [San Francisco LOC].
Named entity recognition

1. Find text segments corresponding to entities
   - Segmentation ambiguity

2. Classify entities
   - Type ambiguity

<table>
<thead>
<tr>
<th>Name</th>
<th>Possible Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington</td>
<td>Person, Location, Political entity, Organization, Vehicle</td>
</tr>
<tr>
<td>Downing St.</td>
<td>Location, Organization</td>
</tr>
<tr>
<td>IRA</td>
<td>Person, Organization, Monetary Instrument</td>
</tr>
<tr>
<td>Louis Vuitton</td>
<td>Person, Organization, Commercial Product</td>
</tr>
</tbody>
</table>
Type ambiguity

● [Washington _PER_] was born into slavery on the farm of James Burroughs.
● [Washington _ORG_] went up 2 games to 1 in the four-game series.
● Blair arrived in [Washington _LOC_] for what may well be his last state visit.
● In June, [Washington _GPE_] passed a primary seatbelt law.
● The [Washington _VEH_] had proved to be a leaky ship, every passage I made …
NER as Sequence Tagging task

- The standard approach to NER is word-by-word sequence tagging

- The text is tagged with BIO or IO tagging
  - B – start of a named entity
  - I – inside of a named entity
  - O – outside of a named entity

- The tag is coupled with a type label:
  - B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC etc
NER labelling

[American Airlines \texttt{ORG}], a unit of [AMR Corp. \texttt{ORG}], immediately matched the move, spokesman [Tim Wagner \texttt{PER}] said.

- O – outside
- B-ORG – beginning of the Organization entity
- I-ORG – inside of the Organization entity
- B-PER – beginning of the Person entity
- I-PER – inside of the Person entity
- etc
NER labelling

[American Airlines\textsubscript{ORG}], a unit of [AMR Corp. \textsubscript{ORG}], immediately matched the move, spokesman [Tim Wagner\textsubscript{PER}] said.

<table>
<thead>
<tr>
<th>Word</th>
<th>BIO label</th>
<th>IO label</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B-ORG</td>
<td>I-ORG</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>I-ORG</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>AMR Corp.</td>
<td>B-ORG</td>
<td>I-ORG</td>
</tr>
<tr>
<td>World</td>
<td></td>
<td></td>
</tr>
<tr>
<td>matched</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>the</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>move</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>spokesman</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Tim</td>
<td>B-PER</td>
<td>I-PER</td>
</tr>
<tr>
<td>Wagner</td>
<td>I-PER</td>
<td>I-PER</td>
</tr>
<tr>
<td>said</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>.</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Another version is BIOES format

- B – begin of the entity
- I – inside of the entity
- O – outside of entities
- E – end of the entity
- S – single word entity
New NER annotations for Estonian

- Entity set: PER, ORG, LOC, GPE, TITLE, PROD, EVENT, DATE, TIME, PERC, MONEY
- Reannotated the “old” NER corpus (ca 240K words)
- Annotated ca 130K words of new data (news texts, opinions, social media)
- Hierarchical annotations up to three levels

John Skytte Institute of Political Studies
Gazetteers

- Dictionaries that contain lists of entities
- Place names, person names etc
- Geographical names: www.geonames.org
- Census data for person names
- Derived from text corpora
- Lists of months, weekdays etc
- Corporations, commercial products etc
- Proteins, genes etc
Recap

- Sequence tagging – assign each word a label from a finite set of labels
- POS and morphological tagging and NER are canonical sequence tagging task examples.

- What other NLP problems can be formalised as a sequence tagging task?
Evaluating a sequence tagger
Evaluating a POS/morphological tagger

- **Accuracy** = \( \frac{\text{Number of correctly predicted tags}}{\text{Number of all words}} \)

- **Accuracy of OOV words**

- **Precision and recall per POS/morphological label**
  - Aggregate with micro- or macro-averaging
Evaluating a NER system

- Use precision, recall and F-score
- Evaluation on the entity level, not the word level

<table>
<thead>
<tr>
<th>Word</th>
<th>True label</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B-ORG</td>
<td>O</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>AMR</td>
<td>B-ORG</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Corp.</td>
<td>I-ORG</td>
<td>I-ORG</td>
</tr>
</tbody>
</table>

True positives:
False positives:
False negatives:

Precision:
Recall:
Evaluating a NER system

- Use precision, recall and F-score
- Evaluation on the entity level, not the word level

<table>
<thead>
<tr>
<th>Word</th>
<th>True label</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B-ORG</td>
<td>O</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>AMR</td>
<td>B-ORG</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Corp.</td>
<td>I-ORG</td>
<td>I-ORG</td>
</tr>
</tbody>
</table>

True positives: 1
False positives: 0
False negatives: 1

Precision: 1
Recall: 0.5
Evaluating a NER system

- Consider **segmentation errors**
- True entity: ORG(1, 2), predicted entity ORG(1, 1)

<table>
<thead>
<tr>
<th>Word</th>
<th>True label</th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B-ORG</td>
<td>B-ORG</td>
<td></td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>AMR</td>
<td>B-ORG</td>
<td>B-ORG</td>
<td></td>
</tr>
<tr>
<td>Corp.</td>
<td>I-ORG</td>
<td>I-ORG</td>
<td></td>
</tr>
</tbody>
</table>

True positives:
False positives:
False negatives:
Precision:
Recall:
Evaluating a NER system

- Consider **segmentation errors**
- True entity: ORG(1, 2), predicted entity ORG(1, 1)

<table>
<thead>
<tr>
<th>Word</th>
<th>True label</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B-ORG</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Airlines</td>
<td>I-ORG</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>unit</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>AMR</td>
<td>B-ORG</td>
<td>B-ORG</td>
</tr>
<tr>
<td>Corp.</td>
<td>I-ORG</td>
<td>I-ORG</td>
</tr>
</tbody>
</table>

- True positives: 1
- False positives: 1
- False negatives: 1
- Precision: 0.5
- Recall: 0.5
Standard evaluation scripts

- Sometimes there exist more or less standard evaluation scripts for certain tasks.
- NER – conlleval script, originally in perl but nowadays python versions can be found
- conll2018_ud_eval.py – for POS and morphological tagging (but also token and sentence segmentation, lemmatization and dependency parsing), available from https://universaldependencies.org/conll18/evaluation.html
Old-school methods for sequence tagging
Sequence tagging approaches

1. Hidden Markov Models
2. Log-linear sequence tagger (logistic regression at each step)
3. Conditional random fields
4. Recurrent neural networks (new school)