Natural Language Processing
SoSe 2014

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Exercise 1: Language Modelling & Part-of-Speech Tagging

Dr. Mariana Neves

May 7th, 2014

Tasks

- Language Modelling

- Part-of-speech tagging

Genia corpus

- Genia corpus (Treebank)

  - Around 2,000 abstracts in XML format

  - Biomedical domain

  - Sentence split, tokenized, POS tagged

- Available at:

  - http://www.nactem.ac.uk/genia/genia-corpus/treebank

```xml
1  <?xml version="1.0" encoding="utf-8"?>
2  <!DOCTYPE Annotation>
3  <?xml-stylesheet type="text/css" href="GENIA_tree_bracket.css"?>
4  <Annotation>
5  <PubmedArticleSet>
6  <PubmedArticle>
7  <MedlineCitation>
8  <PMID>1309833</PMID>
9  <Article>
10 <ArticleTitle><sentence id="S1"><cons cat="NP"><cons cat="NP"><cons cat="NP"><tok cat="NN">Cortisol</tok> <tok cat="
11 <Abstract>
12 <AbstractText><sentence id="S2"><cons cat="S"><cons cat="NP" id="i33" role="SBJ"><tok cat="JJ">Primary</tok> <tok ca
13 <sentence id="S3"><cons cat="S"><cons cat="NP" id="i34" role="SBJ"><tok cat="PRPP">Its</tok> <tok cat="NN">occurrence
14 <sentence id="S4"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok ca
15 <sentence id="S5"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">The</tok> <tok cat="JJ">first</tok> <tok cat=
16 <sentence id="S6"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="NN">Hydrochlorothiazide</tok> <tok cat="NN">therap
17 <sentence id="S7"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">The</tok> <tok cat="JJ">second</tok> <tok cat
18 <sentence id="S8"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">All</tok> <tok cat="CD">four</tok> <tok cat="
19 <sentence id="S9"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="JJ">Low</tok> <tok cat="NN">dose</
20 <sentence id="S10"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="DT">All</tok> <tok cat="NNS">patients</tok></con
21 <sentence id="S11"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="DT">The</tok> <tok cat="JJ">diurn
22 <sentence id="S12"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="EX">There</tok></cons> <cons cat="VP"><tok cat=
23 <sentence id="S13"><cons cat="S"><cons cat="NP" id="i36" role="SBJ"><tok cat="VBG">Circulating</tok> <tok cat="JJ">ad
24 <sentence id="S14"><cons cat="S"><cons cat="NP" id="i37" role="SBJ"><tok cat="NN">Glucocorticoid</tok> <tok cat="NNS"
25 <sentence id="S15"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok c
26 <sentence id="S16"><cons cat="S"><cons cat="NP" id="i38" role="SBJ"><cons cat="NP"><tok cat="DT">A</tok> <tok cat="J
27 <sentence id="S17"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><tok cat="DT">the</tok> <tok c
28 <sentence id="S18"><cons cat="S"><cons cat="PP"><tok cat="IN">As</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">a
29 <sentence id="S19"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">t
30 <sentence id="S20"><cons cat="S"><cons cat="NP" role="SBJ"><tok cat="JJ">Partial</tok> <tok cat="NN">cortisol</tok> <
31 <sentence id="S21"><cons cat="S"><cons cat="PP"><tok cat="IN">In</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">t
32 <sentence id="S22"><cons cat="S"><cons cat="NP" role="SBJ"><cons cat="NP"><tok cat="NN">Therapy</tok></cons> <cons c
33 </Abstract>
34 </Article>
35 </MedlineCitation>
36 </PubmedArticle>
37 </PubmedArticleSet>
38 </Annotation>
39
```

```
<sentence id="S9">
<cons cat="S">
    <cons cat="NP" role="SBJ">
        <cons cat="NP">
            <tok cat="JJ">Low</tok>
            <tok cat="NN">dose</tok>
            <tok cat="NN">dexamethasone</tok>
            <tok cat="NN">therapy</tok>
        </cons>
        <cons cat="PRN">
            <tok cat="LRB">(</tok>
            <cons cat="NP">
                <cons cat="QP">
                    <tok cat="CD">1-1.5</tok>
                </cons>
                <tok cat="NN">mg/day</tok>
            </cons>
            <tok cat="RRB">)</tok>
        </cons>
    </cons>
<cons cat="VP">
    <tok cat="VBD">was</tok> ...
</cons>
</sentence>
```

word

part-of-speech tag
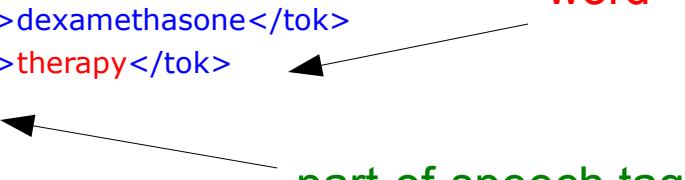
Task 1: Language modelling (Session 2)

- Calculate the perplexity of the corpus

- Split in training (90%) and test (10%) sets

  - Randomly and per document

  - Build a language model using the training set (Maximum Likelihood, bigram)

  - Use Laplace (Add-one) smoothing for the zero probabilities

  - Estimate probabilities on the test set

Task 2: Part-of-speech tagging (Session 4)

- Calculate precision for the corpus

- Split in training (90%) and test (10%) sets

  – Randomly and per document

  – Learn the model using the training set (Hidden Markov
    Model, bigram)

  – Predict POS tags for the test set

Task 2: Part-of-speech tagging

Gold-standard
(test set)

Predictions
(test set)

FN TP FP
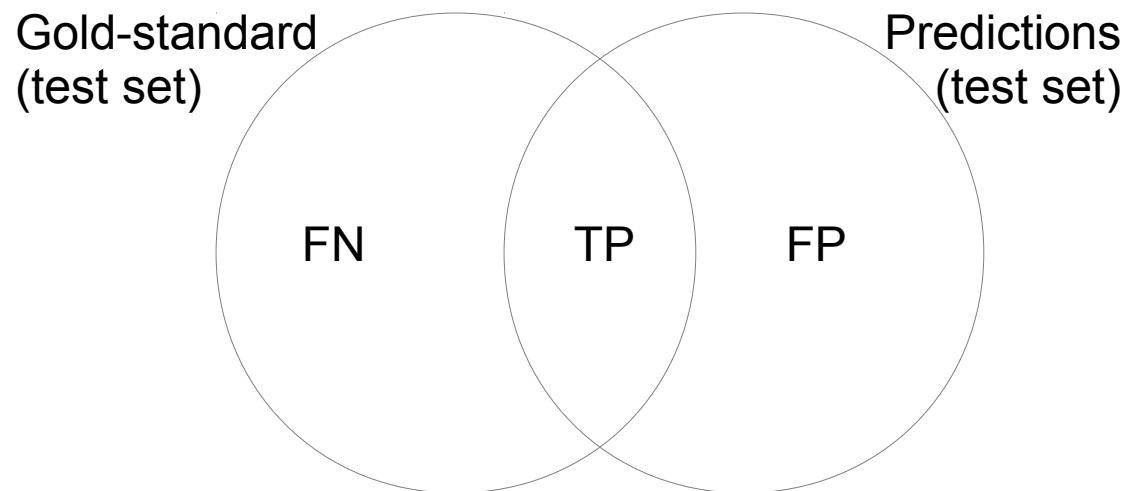
$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F-measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Task 2: Part-of-speech tagging

Secretariat[NNP] is[VBZ] expected[VBN] to[TO] race[VB] tomorrow[NR] .

Secretariat[NNP] is[VBZ] expected[VBN] to[TO] race[NN] tomorrow[NR] .

5 TP, 1 FP [NN], 1 FN [VB]: Precision = 5/6

Exercise 1

- Deadline: June 11th
  - Mail with results (perplexity/precision) and source code
- 20% final grade
  - 30% for both tasks