Intro to Text Processing
Lecture 5

Houda Bouamor
Behrang Mohit
Today: Part of Speech Tagging

• Reminder: Homework 1
  – due Wednesday at 10:00 am
Motivation

• Richer processing of text ➔ abstraction over words

  – Example: Question answering
    • Mariem (NNP) went (V) to (P) the (Det) school (NN)
    • Where did Mariem go?
      – (asking about location ➔ noun)

• Moving up from words to level of phrases, clauses, sentences: syntax
Task: POS tagging

- Part of Speech: POS

- **Input:** a tokenized sentence
  - $w = w_1 \; w_2 \; w_3 \; \ldots \; w_n$

- **Output:** the POS tags for the sentence
  - $t = t_1 \; t_2 \; t_3 \; \ldots \; t_n$
Example:

Ahmed Hassan received a gift.

proper noun  proper noun  verb  determiner  noun  punctuation

descriptor  descriptor  name  past tense  indefinite  singular  end-of
sentence

• Can move deeper and richer
How are POS tags defined?

• Casually: based on basic language knowledge
  – Noun: a person, place, thing, ...
  – Verb: action word

• Linguists: distributional properties
  – Open vs. close
  – Some correlation with the casual version
Linguistic Tests

• If $w$ fits into “$w$ ideas spread quickly” then $w$ is an adjective.

• If we can take suffix –s to make a plural, then $w$ is a noun.
# General POS categories

<table>
<thead>
<tr>
<th>Open classes</th>
<th>Closed classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>prepositions</td>
</tr>
<tr>
<td>verbs</td>
<td>determiners</td>
</tr>
<tr>
<td>adjectives</td>
<td>numerals</td>
</tr>
<tr>
<td>adverbs</td>
<td>auxiliary verbs</td>
</tr>
</tbody>
</table>
Fine-grained classes

Open classes

- nouns
- verbs
- adjectives
- adverbs

proper
common
count
Fine-grained classes

Open classes

- nouns
- verbs
- adjectives
- adverbs

- directional
- degree
- manner
Tagsets

• Different ways to select the set of possible part of speech.
• For English:
  – Brown Corpus: 87 tags
  – Penn Tree Bank: 45 tags
Tag set in rich morphology

- Languages with rich morphology:
  - A word can be a complicated structure
    - Prefix + stem + suffix + suffix2
    - Arabic example: wṣyaktobun $\Rightarrow$ w + s + yaktob + un
      - And will write they $\Rightarrow$ and they will write
  - Should we have tags for every complex word construction?
    - It leads to thousands of tags

- Open problem and discussion on the right level of tag granularity
# English POS Tag Set (partial)

- **NNP** – Proper Noun (John)
- **NN** – Singular Noun (book)
- **NNS** – Plural Noun (books)
- **IN** – preposition (in, on ..)
- **MD** – Modal (can, may, ought)
- **JJ** – adjective (red)
- **JJR** – comparative adjective (larger)
- **VB** – verb base form (go)
- **VBD** – verb past tense (went)
- **VBG** – verb present tense (going)
- **VBN** – verb past participle (gone)
- **VBP** – verb present tense (go)
- **VBZ** – verb present tense 3rd person (goes)
- **DT** – Determiner (the, a)
- **WDT** – wh-determiner (what, which)
- **PRP** – personal pronoun (I, you)
- **PRP$** - possessive pronoun (her, his..)
- **RB** – adverb (fast)
- **RBR** – adverb comparative (faster)
- **JJS** - adjective superlative (fastest)
- **CD** – number (1, 2, one)
- **PUNC** – punctuation (.,?:;)
- **DEL** – Delimiter (<s>)
- ...
Sample POS tagged corpus

IN In DT an NNP Oct. CD 19 NN review IN of `` `` DT The NN Misanthrope '' '' IN at NNP Chicago POS 's NNP Goodman NNP Theatre -LRB- -LRB- `` `` VBN Revitalized NNS Classics

VBP Take DT the NN Stage IN in NNP Windy NNP City , , '' '' NN Leisure CC & NNS Arts -RRB- -RRB- , , DT the NN role IN of NNP Celimene , , VBN played IN by NNP Kim NNP Cattrall , , VBD was RB mistakenly VBN attributed TO to NNP Christina NNP Haag . .

NNP Ms. NNP Haag VBZ plays NNP Elianti . .

NNP Rolls-Royce NNP Motor NNPS Cars NNP Inc. VBD said PRP it VBZ expects PRP$ its NNP U.S. NNS sales TO to VB remain JJ steady IN at IN about CD 1,200 NNS cars IN in CD 1990 . .

DT The NN luxury NN auto NN maker JJ last NN year VBD sold CD 1,214 NNS cars IN in DT the NNP U.S.
Challenges in POS tagging

• Knowing a word, what is its tag?
  – How to handle unknown words?

• If the word has more than one possible tag, which one to pick?
  – Need to resolve **ambiguity**
The problem of ambiguity

• Time flies like an arrow.
  – flies ➔ noun (plural)? or verb (3rd person)?

• Book that flight.
  – Book ➔ Noun or verb?
Sample of ambiguities

| 7 down | 5 's | 5 forecast |
| 6 that | 5 run | 5 fit |
| 6 set | 5 repurchase | 5 first |
| 6 put | 5 read | 5 East |
| 6 open | 5 present | 5 counter |
| 6 hurt | 5 out | 5 cost |
| 6 cut | 5 many | 5 close |
| 6 bet | 5 less | 5 bid |
| 6 back | 5 left | 5 beat |
| 5 vs. | 5 Japanese | 5 a |
| 5 the | 5 in | |
| 5 spread | 5 hit | |
| 5 split | 5 half | |
| 5 say | 5 further | |

317 down RB
200 down RP
138 down IN
10 down JJ
1 down VBP
1 down RBR
1 down NN
POS tagging as a classification task

• Difficulty varies based on the tag set
• Given a human-labeled corpus, how can we train classifier?
Classifying POS tags

• Remember:
  – **Input:** a tokenized sentence
    • \( w = w_1 \ w_2 \ w_3 \ldots \ w_n \)
  – **Output:** the POS tags for the sentence
    • \( t = t_1 \ t_2 \ t_3 \ldots \ t_n \)

• Baseline Statistical approach:
  – \( \text{argmax} \ P(t \mid w) \) for known words
  – \( \text{argmax} \ P(t) \) for unknown words
The Noisy-channel Paradigm

I will fix the job situation

Noisy Channel

I bla fix the bla situation

What did he say?

Decoder
Find that $X^*$ that maximizes $P(X \mid Y)$

$X^*$
POS tagging with the noisy channel

• I want to write a sentence
  – I first choose a sequence of tags $t=t_1, t_2, \ldots, t_n$ (source)
  – The tags go through the channel, *noise* corrupts them into *words*.
  – $w=w_1, w_2, \ldots, w_n$ is *observed*.
  – What were the original tags I had in mind?
POS tagging with the noisy-channel

He can can the can

What were the tags he had in mind?
Statistical POS tagging

• We have \( W = w_1, w_2, w_3 \ldots w_n \)

• We need to find: \( \arg\max_T P(T|W) \)
  
  – Remember, last time we used Bayes rule to reverse the conditional probability.

\[
\arg\max_T P(T|W) = \arg\max_T P(W|T) P(T)
\]
POS tagging with the noisy-channel

Noisy Channel

P(T)

P(W|T)

What tags sequences could he have chosen?

How could the channel have transformed tags to words?

What were the tags he had in mind?
Statistical POS tagger (1)

1. Find \( \text{argmax}_T P(T|W) = \text{argmax}_T P(W|T) P(T) \)
   - \( P(T) = \text{probability of tag sequence } t_1, t_2, t_3 \ldots t_n \)
   - \( P(W|T) = \text{probability of words given the tags} \approx P(w_1|t_1) \times P(w_2|t_2) \times \ldots \times P(w_n|t_n) \)
   - Simplifying assumptions
     - Words are independent of each other.
     - A word’s identity depends only on its tag.

2. Training:
   - \( p(w_i|t_i) \) can be estimated from annotated training data
Statistical POS tagger (2)

• \( P(T) = \text{probability of tag sequence } T = t_1, t_2, \ldots, t_n \)

• Similar to the language model derivation
  – Chain rule and N-gram approximation
    \[
    P(t_1, t_2, t_3 \ldots t_n) = P(t_1 \mid <s>) \times P(t_2 \mid t_1) \times P(t_3 \mid t_1, t_2) \times P(t_4 \mid t_1, t_2, t_3) \times \ldots \times P(t_n \mid t_1, t_2, \ldots, t_{n-1})
    \]
    \[
    P(t_1, t_2, t_3 \ldots t_n) \approx P(t_1 \mid <s>) \times P(t_2 \mid t_1) \times P(t_3 \mid t_2) \times P(t_4 \mid t_3) \times \ldots \times P(t_n \mid t_{n-1})
    \]
  – Each tag depends on the previous tag.
Training a bigram tagger

• \( \text{argmax}_T P(T|W) = \text{argmax}_T P(W|T) P(T) \)
  
  \[ = P(w_1|t_1) P(w_2|t_2) \ldots (w_n|t_n) P(t_2|t_1) P(t_3|t_2) \ldots P(t_n | t_{n-1}) \]

Training:

\( P(w_i|t_j) \) gets estimated from an annotated corpus.
\( P(t_i|t_j) \) gets estimated from an annotated corpus.
Available tools

• Illinois POS tagger
  http://cogcomp.cs.illinois.edu/demo/pos/

• Stanford POS tagger

• TreeTagger
  http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/