CS460/626: Natural Language Processing/Speech, NLP and the Web
(Lecture 2–POS tagging)

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay
5th Jan, 2012
Two pictures
Part of Speech Tagging
What it is

- POS Tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags.
- The set of tags is called the Tag-set.
Definition

- Tagging is the assignment of a single part-of-speech tag to each word (and punctuation marker) in a corpus.

- "“_“ The_DT guys_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP . _.

- "“_“ The_DT guys_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP . _.

- "“_“ The_DT guys_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP . _.
Importance of POS tagging
Categories of words

- Words can be divided into classes that behave similarly.
- Traditionally eight parts of speech in English: noun, verb, pronoun, preposition, adverb, conjunction, adjective and article.
- More recently larger sets have been used: e.g. Penn Treebank (45 tags), Susanne (353 tags).
Why POS

- POS tell us a lot about a word (and the words near it).
  - E.g, adjectives often followed by nouns
  - personal pronouns often followed by verbs
  - possessive pronouns by nouns

- Pronunciations depends on POS, e.g. *object* (first syllable NN, second syllable VM), *content*, *discount*

- First step in many NLP applications
Open and closed classes

- Closed classes have a fixed membership of words: determiners, pronouns, prepositions
- Usually function word: frequently occurring, grammatically important, often short (e.g. *of, it, the, in*)
- Open classes: nouns, verbs, adjectives and adverbs
Open Class (1/2)

- **Nouns:**
  - Proper nouns (*Scotland, BBC*),
  - common nouns
    - count nouns (*goat, glass*)
    - mass nouns (*snow, pacifism*)

- **Verbs:**
  - actions and processes (*run, hope*)
  - also auxiliary verbs (*is, are, am, will, can*)
Open Class (2/2)

- Adjectives:
  - properties and qualities (age, colour, value)

- Adverbs:
  - modify verbs, or verb phrases, or other adverbs- Unfortunately John walked home extremely slowly yesterday
  - Sentential adverb: unfortunately
  - Manner adverb: extremely, slowly
  - Time adverb: yesterday
Closed class

- Prepositions: on, under, over, to, with, by
- Determiners: the, a, an, some
- Pronouns: she, you, I, who
- Conjunctions: and, but, or, as, when, if
- Auxiliary verbs: can, may, are
Where does POS tagging fit in:

- Discourse and Coreference
- Semantics Extraction
- Parsing
- Chunking
- POS tagging
- Morphology

Increased Complexity Of Processing
POS and Parsing
POS critical for Parsing: Stanford Parser output

Your query
My dog also likes eating sausage.

Tagging
My/PRP$ dog/NN also/RB Likes/VBZ eating/VBG sausage/NN ./.

Parse
(ROOT (S (NP (PRP$ My) (NN dog)) (ADVP (RB also)) (VP (VBZ likes) (S (VP (VBG eating) (NP (NN sausage)))))) ( . .)))

Typed dependencies
poss(dog-2, My-1) nsubj(likes-4, dog-2) advmod(likes-4, also-3) root(ROOT-0, likes-4) xcomp(likes-4, eating-5) dobj(eating-5, sausage-6)
POS and SMT
Experimental Framework

- Setup
  - Used Moses for Decoding
  - GIZA++ for word alignment
  - Evaluation was done using BLEU and Subjective evaluation in terms of Adequacy and Fluency

- Corpus Statistics

<table>
<thead>
<tr>
<th>Domain</th>
<th># Sentences</th>
<th># Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Tourism</td>
<td>13738</td>
</tr>
<tr>
<td>Tuning</td>
<td>Tourism</td>
<td>600</td>
</tr>
<tr>
<td>Testing</td>
<td>Tourism</td>
<td>400</td>
</tr>
</tbody>
</table>
# Adequacy and Fluency Scale

<table>
<thead>
<tr>
<th>Level</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Flawless Hindi with no grammatical errors.</td>
</tr>
<tr>
<td>4</td>
<td>Good Hindi with few minor morphology errors</td>
</tr>
<tr>
<td>3</td>
<td>No-Native Hindi with few minor morphology errors</td>
</tr>
<tr>
<td>2</td>
<td>Disfluent Hindi with correct phrases but ungrammatical overall</td>
</tr>
<tr>
<td>1</td>
<td>Incomprehensible</td>
</tr>
</tbody>
</table>

**Fluency Scale**

<table>
<thead>
<tr>
<th>Level</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>All meaning is conveyed</td>
</tr>
<tr>
<td>4</td>
<td>Most of the meaning is conveyed</td>
</tr>
<tr>
<td>3</td>
<td>Much of the meaning is conveyed</td>
</tr>
<tr>
<td>2</td>
<td>Less meaning is conveyed</td>
</tr>
<tr>
<td>1</td>
<td>None of the meaning is conveyed</td>
</tr>
</tbody>
</table>

**Adequacy Scale**
## Results (1/3)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.36</td>
<td>2.32</td>
<td>2.26</td>
</tr>
<tr>
<td>Baseline + Suffix</td>
<td>11.23</td>
<td>2.44</td>
<td>2.28</td>
</tr>
<tr>
<td>Baseline + Semantic features</td>
<td>9.00</td>
<td>1.6</td>
<td>1.42</td>
</tr>
<tr>
<td>Baseline + Morphology Analyzer</td>
<td>11.34</td>
<td>2.40</td>
<td>2.30</td>
</tr>
<tr>
<td>Baseline + POS</td>
<td>11.36</td>
<td>2.39</td>
<td>2.29</td>
</tr>
<tr>
<td>Baseline + WSD</td>
<td>11.36</td>
<td>2.35</td>
<td>2.25</td>
</tr>
<tr>
<td>Baseline+ Clause Wise</td>
<td>11.98 (11.95 is baseline)</td>
<td>2.31 (2.10 is baseline)</td>
<td>2.13 (2.03 is baseline)</td>
</tr>
</tbody>
</table>

Experimental Results
### Results (2/3)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline+ POS+ Morphology</td>
<td>11.43</td>
<td>2.54</td>
<td>2.42</td>
</tr>
<tr>
<td>Baseline+ Semantic+ Morphology</td>
<td>10.94</td>
<td>2.28</td>
<td>2.15</td>
</tr>
<tr>
<td>Baseline+ Semantic + POS</td>
<td>8.25</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Baseline+ WSD+POS</td>
<td>11.38</td>
<td>2.39</td>
<td>2.27</td>
</tr>
<tr>
<td>Baseline+ WSD + Morphology</td>
<td>11.35</td>
<td>2.35</td>
<td>2.25</td>
</tr>
<tr>
<td>Baseline+ WSD+ Semantic</td>
<td>9.02</td>
<td>1.67</td>
<td>1.39</td>
</tr>
<tr>
<td>Baseline+ POS+ Morphology</td>
<td>11.43</td>
<td>2.54</td>
<td>2.42</td>
</tr>
</tbody>
</table>
## Results (3/3)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>13.35</td>
<td>3.02</td>
<td>2.66</td>
</tr>
<tr>
<td>Baseline+ Suffix</td>
<td>13.50</td>
<td>3.04</td>
<td>2.82</td>
</tr>
<tr>
<td>Baseline+ Morph</td>
<td>13.58</td>
<td>3.04</td>
<td>2.82</td>
</tr>
<tr>
<td>Baseline+ POS</td>
<td>13.44</td>
<td>3.00</td>
<td>2.82</td>
</tr>
<tr>
<td>Baseline+ Semantic</td>
<td>12.65</td>
<td>2.72</td>
<td>2.52</td>
</tr>
<tr>
<td>Baseline+ POS+ Morphology</td>
<td>13.27</td>
<td>3.12</td>
<td>2.96</td>
</tr>
<tr>
<td>Baseline+ POS+ Morphology 2</td>
<td>9.02</td>
<td>3.1</td>
<td>2.92</td>
</tr>
<tr>
<td>Baseline+ Morph+ Semantic</td>
<td>12.51</td>
<td>2.8</td>
<td>2.30</td>
</tr>
<tr>
<td><strong>Reordering Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MWE and POS
Multi-Word Expressions (MWE)

- Necessary Condition
  - Word sequence separated by space/delimiter

- Sufficient Conditions
  - Non-compositionality of meaning
  - Fixity of expression
    - In lexical items
    - In structure and order
Examples – Necessary condition

- Non-MWE example:
  - Marathi: सरकार हक्काबक्का झाले
  - Roman: sarakAra HakkAbakkA J ZAle
  - Meaning: government was surprised

- MWE example:
  - Hindi: गरीब नवाज़
  - Roman: garIba navAjZa
  - Meaning: who nourishes poor
Examples - Sufficient conditions (Non-compositionality of meaning)

- Konkani: पोटांत चाबता
  - Roman: poTAMta cAbatA
  - Meaning: to feel jealous

- Telugu: చివర డోలడి దుడుము
  - Roman: ceVttu kiMXa pLI daru
  - Meaning: an idle person

- Bangla: মাটির মানুষ
  - Roman: mAtira mAnuSa
  - Meaning: a simple person/son of the soil
Examples – Sufficient conditions (Fixity of expression)

In lexical items

- **Hindi**
  - usane muj e gAll dl
  - *usane muj e gall pradAna kl

- **Bangla**
  - jabajjI bana karadaMda
  - *jl banabhara karadaMda
  - *jabajjI bana jela

- **English (1)**
  - life imprisonment
  - *lifelong imprisonment

- **English (2)**
  - Many thanks
  - *Plenty thanks
Examples – Sufficient conditions (In structure and order)

- English example
  - kicked the bucket (died)
  - the bucket was kicked
    (not passivizable in the sense of dying)

- Hindi example
  - उम्र कैद
  - umra kEda (life imprisonment)
  - umra bhara kEda
## MW task (NLP + ML)

<table>
<thead>
<tr>
<th></th>
<th>String + Morph</th>
<th>POS</th>
<th>POS+WN</th>
<th>POS + List</th>
<th>Chunk-ing</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rules</strong></td>
<td>Onomatopoeic Reduplication</td>
<td>Non-</td>
<td>Non-redup</td>
<td>Conjunct verb</td>
<td>Non-contiguous Complex Predicate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(tik tik, chham chham)</td>
<td>Onomatopoeic Reduplication</td>
<td>(raat din, dhan doulat)</td>
<td>(verbalizer list), Compound verb (vector verb list)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-redup (Syn, Anto, Hypo)</td>
<td></td>
<td>(salaha dena, has uthama)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Statistical</strong></td>
<td>Collections or fixed expressions</td>
<td>Conjunct verb</td>
<td>Non-contiguous Complex Predicate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(many thanks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Idioms will be list morph + look up*
Challenge of POS tagging

Example from Indian Language
Tagging of *jo, vaha, kaun* and their inflected forms in Hindi and their equivalents in multiple languages
DEM and PRON labels

- **Jo_DEM** ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai

- **Jo_PRON** kal aayaa thaa, vaha cricket acchhaa khel letaa hai
Disambiguation rule-1

- If
  - Jo is followed by noun
- Then
  - DEM
- Else
  - ...

False Negative

- When there is arbitrary amount of text between the *jo* and the noun

- *jo_??? bhaagtaa huaa, haftaa huaa, rotaa huaa, chennai academy a koching lenevaalaa* ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai
False Positive

- $Jo_{DEM}$ (wrong!) *duniyadarii samajhkar chaltaa hai*, ...

- $Jo_{DEM/PRON?}$ *manushya manushyoM ke biich ristoM naatoM ko samajhkar chaltaa hai*, ... (ambiguous)
False Positive for Bengali

- *Je_DEM* (wrong!) *bhaalobaasaa paay, sei bhaalobaasaa dite paare*  
  (one who gets love can give love)
- *Je_DEM* (right!) *bhaalobaasa tumi kalpanaa korchho, taa e jagat e sambhab nay*  
  (the love that you imagine exits, is impossible in this world)
Will fail

- In the similar situation for
  - Jis, jin, vaha, us, un

- All these forms add to corpus count
Disambiguation rule-2

- If
  - Jo is oblique (attached with ne, ko, se etc. attached)
- Then
  - It is PRON
- Else
  - <other tests>
Will fail (false positive)

- In case of languages that demand agreement between *jo*-form and the noun it qualifies

- *E.g.* Sanskrit

- *Yasya_PRON (wrong!) baalakasya aananam drshtyaa...* *(jis ladake kaa muha dekhkar)*

- *Yasya_PRON (wrong!) kamaniyasya baalakasya aananam drshtyaa...*
Will also fail for

- Rules that depend on the whether the noun following *jo/vaha/kaun or its form* is oblique or not
- Because the case marker can be far from the noun
- `<vaha or its form> ladakii jise piliya kii bimaarii ho gayiii thii ko ...`
- **Needs discussions across languages**
DEM vs. PRON cannot be disambiguated

IN GENERAL

At the level of the POS tagger

i.e.

Cannot assume parsing

Cannot assume semantics