Administrivia

- Homework 1 graded
POS Tagging

• Task:
  • assign the right part-of-speech tag, e.g. noun, verb, conjunction, to a word in context
  • in NLP: assign one of the 45(48) Penn tags

• POS taggers
  • need to be fast in order to process large corpora
    • Linear wrt. size of the corpora
  • POS taggers assign the correct tag without parsing the sentence
    • the walk : noun I took ...
    • I walk : verb 2 miles every day

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
<td>and, but, or</td>
<td>SYM</td>
<td>symbol</td>
<td>*, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>one, two, three</td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a, the</td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>existential “there”</td>
<td>there</td>
<td>VB</td>
<td>verb, base form</td>
<td>eat</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>mea culpa</td>
<td>VBD</td>
<td>verb, past participle</td>
<td>ate</td>
</tr>
<tr>
<td>IN</td>
<td>preposition or subordinating conjunction</td>
<td>of, in, by</td>
<td>(past tense)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>yellow</td>
<td>VBG</td>
<td>verb, gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td>bigger</td>
<td>VBN</td>
<td>verb, past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJIS</td>
<td>adj., superlative</td>
<td>widest</td>
<td>VBP</td>
<td>verb, non-3g pres</td>
<td>eat</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td>1, 2, One</td>
<td>VBZ</td>
<td>verb, 3g pres</td>
<td>eat</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>can, should</td>
<td>WDT</td>
<td>wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing, or mass</td>
<td>llama, snow</td>
<td>WP</td>
<td>wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td>llamas</td>
<td>WPS</td>
<td>possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>IBM</td>
<td>WRB</td>
<td>wh-adverb</td>
<td>here, where</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Caroinas</td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>all</td>
<td>$</td>
<td>pound sign</td>
<td>$</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>”</td>
<td>right quote</td>
<td>‘or ‘</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, you, he</td>
<td>”</td>
<td>left quote</td>
<td>“or “</td>
</tr>
<tr>
<td>PRPS</td>
<td>possessive pronoun</td>
<td>your, one’s</td>
<td>)</td>
<td>right parenthesis</td>
<td>[ . ( , { &lt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>quickly, never</td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>faster</td>
<td>.</td>
<td>sentence-final punct</td>
<td>. ?</td>
</tr>
<tr>
<td>RBRS</td>
<td>adverb, superlative</td>
<td>fastest</td>
<td>:</td>
<td>mid-sentence punct</td>
<td>: : . --</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>up, off</td>
<td>:</td>
<td></td>
<td>: : : --</td>
</tr>
</tbody>
</table>
How Hard is Tagging?

• Easy task to do well on:
  • naïve algorithm
    • assign tag by (unigram) frequency
  • 90% accuracy (Charniak et al., 1993)
  • POS tagging is not 100% accurate (*errors compound down the line...*)

• Brown Corpus (Francis & Kucera, 1982):
  – 1 million words
  – 39K distinct words
  – 35K words with only 1 tag
  – 4K with multiple tags (DeRose, 1988)

That’s 89.7% from just considering single tag words, even without getting any multiple tag words right
Penn TreeBank Tagset

Part-of-Speech Tagging Guidelines for the Penn Treebank Project

- [http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports](http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports)

Example:

- The duchess was entertaining last night.

  • From the Penn Treebank itself
  
  (VP (TO to)
   (VP (VB put)
    (NP (DT the) (NN genie))
    (ADV (PRT (RB back)))
    (PP (IN in)
     (NP (DT the) (NN bottle))))
  
  • Treebank (cited by textbook):

  (5.4) Mrs./NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG

  (5.5) All/DT we/PRP gonna/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN

  (5.6) Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

<table>
<thead>
<tr>
<th>Table 2: The Penn Treebank POS tagset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CC Coordinating conjunction 25. TO to</td>
</tr>
<tr>
<td>2. CD Cardinal number 26. OE Interjection</td>
</tr>
<tr>
<td>3. DT Determiner 27. VB Verb, base form</td>
</tr>
<tr>
<td>4. EX Existential there 28. VBD Verb, past tense</td>
</tr>
<tr>
<td>5. FW Foreign word 29. VBP Verb, gerund/present participle</td>
</tr>
<tr>
<td>6. IN Preposition/subordin. 30. VBN Verb, past participle</td>
</tr>
<tr>
<td>7. JJ Article 31. VBZ Verb, non-3rd ps. sing. present</td>
</tr>
<tr>
<td>8. JJR Adjective, comparative 32. VBN Verb, 3rd ps. sing. present</td>
</tr>
<tr>
<td>9. JJS Adjective, superlative 33. VBD wh-determiner</td>
</tr>
<tr>
<td>10. LS List item marker 34. WP wh-pronoun</td>
</tr>
<tr>
<td>11. MD Modal 35. WP Wh-pronoun</td>
</tr>
<tr>
<td>12. NN Noun, singular or mass 36. WDB Wh-adverb</td>
</tr>
<tr>
<td>13. NNS Noun, plural 37. # Proud sign</td>
</tr>
<tr>
<td>14. NNP Proper noun, singular 38. $ Dollar sign</td>
</tr>
<tr>
<td>15. NPS Proper noun, plural 39. . Enclase-final punctuation</td>
</tr>
<tr>
<td>16. PDT Predeterminer 40. , Comma</td>
</tr>
<tr>
<td>17. POS Possessive ending 41. ; Colon, comma</td>
</tr>
<tr>
<td>18. RBR Adverb 42. ) Right bracket character</td>
</tr>
<tr>
<td>19. RP Possessive pronoun 43. ] Right close single quote</td>
</tr>
<tr>
<td>20. RB Adverb 44. &quot; Straight double quote</td>
</tr>
<tr>
<td>21. RBR Adverb, comparative 45. ' Left open single quote</td>
</tr>
<tr>
<td>22. RBR Adverb, superlative 46. ' Left open double quote</td>
</tr>
<tr>
<td>23. RP Particle 47. ` Right open single quote</td>
</tr>
<tr>
<td>24. SW Symbol 48. &quot; Right open double quote</td>
</tr>
</tbody>
</table>

| mathematical or scientific |
The Part-of-Speech tagger has automatically labeled the input in the following way.

- NNP/ Helicopters
- NN/ will
- NNP/ patrol
- VBD/ the
- JJ/ temporary
- JJ/ no-fly
- JJ/ zone
- NN/ around
- NNP/ New
- NNP/ Jersey
- NNP/ Pos's
- NNP/ MetLife
- NNP/ Stadium
- NNP/ Sunday
- ,
- IN/ with
- NNP/ F-16s
- VB/ based
- IN/ in
- NNP/ Atlantic
- NNP/ City
- JJ/ ready
- TO/ to
- VBD/ be
- VBD/ be
- VBD/ scrambled
- IN/ if
- DT/ an
- JJ/ unauthorized
- NN/ aircraft
- VBD/ does
- VBD/ enter
- VBD/ off
- the
- VBD/ restricted
- NN/ airspace
- ./

- IN/ Down
- IN/ below
- ,
- JJ/ bomb-sniffing
- NNS/ dogs
- MD/ will
- VBP/ patrol
- VBD/ the
- NNP/ trains
- CC/ and
- NNP/ buses
- VB/ that
- VBD/ will
- VBP/ take
- VBD/ approximately
- VPP/ 30,000
- IN/ of
- DT/ the
- JJ/ 80,000-
- NNS/ spectators
- VBD/ to
- VBD/ to
- NNP/ Sunday
- NNP/ Pos's
- NNP/ Super
- NNP/ Bowl
- IN/ between
- DT/ the
- NNP/ Denver
- NNP/ Broncos
- CC/ and
- NNP/ Seattle
- NNP/ Seahawks
- ./

- DT/ The
- NNP/ Transportation
- NNP/ Security
- NNP/ Administration
- VBD/ said
- VBD/ that
- VBD/ VBD/ has
- VBD/ added
- IN/ about
- IN/ two
- NNP/ dozen
- NNP/ dogs
- VBD/ to
- VBD/ monitor
- NNP/ passengers
- VBD/ coming
- VBD/ in
- IN/ and
- VBD/ out
- IN/ of
- VBD/ the
- NNP/ airport
- VBD/ around
- VBD/ at
- VBD/ the
- NNP/ Super
- NNP/ Bowl
- .
POS Tagging and the Buffalo sentence

• Let's digress a bit and look at a worst case scenario. Consider the Buffalo sentence below ...
POS Tagger

• Buffalo buffalo = *buffalo from Buffalo*

• Since *buffalo* is a transitive verb, we can form:
  • NNP/Buffalo NNS/buffalo VBP/buffalo NNP/Buffalo NNS/buffalo
POS Tagger

- Object relative clause construction:
  - Buffalo buffalo (that) Buffalo buffalo buffalo
  - NNP/Buffalo NNS/buffalo NNP/Buffalo NNS/buffalo VBP/buffalo
POS Tagger

• Substitute the relative clause into the sentence:
  • NNP/Buffalo NNS/buffalo NNP/Buffalo NNS/buffalo VBP/buffalo VBP/buffalo NNP/Buffalo NNS/buffalo

Syntactic Analysis
POS Tagger

- The UIUC tagger:
POS Tagger

• Stanford Parser:

```
Your query
Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo .
Tagging
Buffalo/NNP buffalo/NNP Buffalo/NNP buffalo/JJ buffalo/NN buffalo/NN Buffalo/NNP buffalo/VBZ ./.
Parse
(ROOT
  (S
    (NP (NNP Buffalo) (NNP buffalo) (NNP Buffalo) (JJ buffalo) (NN buffalo) (NN buffalo) (NNP Buffalo))
    (VP (VBZ buffalo))
    (. .))))
```

NNP = proper noun,
JJ = adjective,
VBZ = verb 3rd person singular present
POS Tagger

• Berkeley Parser:

```
Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo .
```

`VBP = verb non-3^{rd} person singular present`
Tagging Methods

• 3 Basic Approaches
  • Manual rules
  • Machine Learning of rules
  • Statistical models (Hidden Markov Models)

not just Penn POS tags: more information available...
Rule-Based POS Tagging

• **ENGTWOL**
  • English morphological analyzer based on two-level morphology (*see JM Chapter 3*)
  • 56K word stems
  • processing
    • apply morphological engine
    • get all possible tags for each word
    • apply rules (1,100) to eliminate candidate tags
Rule-Based POS Tagging

see section 5.4

• ENGTWOL and ENGCG
  • (dead link)
  • http://www2.lingsoft.fi/cgi-bin/engcg

Test cases:
1. the **walk**: noun I took ...
2. I **walk**: verb 2 miles every day
3. I want to win
4. I went to the store
5. The duchess was entertaining last night.
Rule-Based POS Tagging

**ENGCG**

Type one or more English sentences (max. 100 words). For best results, use proper capitalization and punctuation.

```
the walk I took
(See the description of morphological tags, syntactic tags and other notations.)
"<the>"
  "the" <Def> DET CENTRAL ART SG/PL @DN>
"<walk>"
  "walk" N NOM SG  @SUBJ @OBJ
"<i>"
  "i" <*> <NonMod> PRON PERS NOM SG1 SUBJ @SUBJ
"<took>"
  "take" <as/SVOC/A> <for/SVOC/A> <SVO> <SV00> <SV> V PAST VFIN @FMAINV
```
Rule-Based POS Tagging

I walk 2 miles every day

(See the description of morphological tags, syntactic tags and other notations.)

"<i>"
  "i" <*> <NonMod> PRON_PERS NOM SG1 SUBJ @SUBJ
"<walk>"
  "walk" <SV> <SVO> V PRES -SG3 VFIN @+FMAINV
"<2>"
  "2" NUM CARD @QN>
"<miles>"
  "mile" N NOM PL @O-ADV
"<every>"
  "every" <Quant> DET CENTRAL SG/PL @QN>
"<day>"
  "day" N NOM SG @ADV
Rule-Based POS Tagging

I want to win

(See the description of morphological tags, syntactic tags and other notations.)

"<*i>"
  "i" <*> <NonMod> PRON PERS NOM SG1 SUBJ @SUBJ
"<want>"
  "want" <SVOC/A> <SVO> <SV> <P/for> V PRES -SG3 VFIN @+FMAINV
"<to>"
  "to" INFMARK> @INFMARK>
"<win>"
  "win" <SV> <SVO> V INF @-FMAINV
Rule-Based POS Tagging

I went to the store
(See the description of morphological tags, syntactic tags and other notations.)

"<*i>"
  "i"  <*>  <NonMod> PRON PERS NOM SG1 SUBJ @SUBJ
"<went>"
  "go"  <SVC/A>  <SV>  <P/for>  <P/with>  V  PAST  VFIN  @+FMAINV
"<to>"
  "to"  PREP  @ADV
"<the>"
  "the"  <Def>  DET  CENTRAL  ART  SG/PL  @DN
"<store>"
  "store"  N  NOM  SG  @<P
Rule-Based POS Tagging

The duchess was entertaining last night
(See the description of morphological tags, syntactic tags and other notations.)

"<*the>
  "the" <*> <Def> DET CENTRAL ART SG/PL @DN
"<duchess>
  "duchess" <Title> N NOM SG @SUBJ
"<was>
  "be" <SV> <SVC/N> <SVC/A> V PAST SG1,3 VFIN @+FAUXV
"<entertaining>
  "entertain" <SVO> <SV> PCP1 @-PMAINV
"<last>
  "last" <Genord> DET POST SG/PL @QN
"<night>
  "night" N NOM SG @ADVL
Rule-Based POS Tagging

- example in the textbook is:
  - *Pavlov had shown that salivation* ...
  - ... elided material is crucial

"<that>""that" <**CLB> CS @CS
"<that>""that" DET CENTRAL DEM SG @DN

"<**CLB> clause boundary (who)"
Rule-Based POS Tagging

- Examples of tags:
  - PCP2 past participle
  - **intransitive:**
    SV subject verb
  - **ditransitive:**
    SVOO subject verb object object

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Additional POS features</th>
</tr>
</thead>
<tbody>
<tr>
<td>smaller</td>
<td>ADJ</td>
<td>COMPARATIVE</td>
</tr>
<tr>
<td>entire</td>
<td>ADJ</td>
<td>ABSOLUTE ATTRIBUTIVE</td>
</tr>
<tr>
<td>fast</td>
<td>ADV</td>
<td>SUPERLATIVE</td>
</tr>
<tr>
<td>that</td>
<td>DET</td>
<td>CENTRAL DEMONSTRATIVE SG</td>
</tr>
<tr>
<td>all</td>
<td>DT</td>
<td>PREDETERMINER SG, PL, QUANTIFIER</td>
</tr>
<tr>
<td>dog's</td>
<td>N</td>
<td>GENITIVE SG</td>
</tr>
<tr>
<td>furniture</td>
<td>N</td>
<td>NOMINATIVE SG, NO IND, DETERNER</td>
</tr>
<tr>
<td>one</td>
<td>NUM</td>
<td>SG</td>
</tr>
<tr>
<td>she</td>
<td>PRON</td>
<td>PERSONAL, FEMININE, NOMINATIVE, SG</td>
</tr>
<tr>
<td>show</td>
<td>V</td>
<td>IMPERATIVE, VFIN</td>
</tr>
<tr>
<td>show</td>
<td>V</td>
<td>PRESENT, SG, VFIN</td>
</tr>
<tr>
<td>show</td>
<td>N</td>
<td>NOMINATIVE, SG</td>
</tr>
<tr>
<td>shown</td>
<td>PCP2</td>
<td>SVOO, SVO, SV</td>
</tr>
<tr>
<td>occurred</td>
<td>PCP2</td>
<td>SV</td>
</tr>
<tr>
<td>occurred</td>
<td>V</td>
<td>PAST, VFIN, SV</td>
</tr>
</tbody>
</table>

Old textbook figure 8.8
Rule-Based POS Tagging

- **example**
  - it isn’t that:adv odd

- **rule (from pg. 138)**
  - given input “that”
  - if
    - (+1 A/ADV/QUANT)
    - (+2 SENT-LIM)
    - (NOT -1 SVOC/A)
  - then eliminate non-ADV tags
  - else eliminate ADV tag

    cf. I consider that odd

    <SVOC/A> complex transitive with adjective complement (consider)
Rule-Based POS Tagging

- **examples**
  - it isn’t **that**:adv odd
  - I consider **that** odd

```
It isn’t that odd
(See the description of morphological tags, syntactic tags and other notations.)
"<it>"
  "it" <+> <NonMod> PRON NOM SG3 SUBJ @SUBJ
"<is>"
  "be" <+> <SVC/N> <SVC/A> V PRES SG3 VFIN @+PMAINV
"<_n’t>"
  "not" NEG-PART @NEG
"<that>"
  "that" ADV AD-A @AD-A>
"<odd>"  
  "odd" A ABS @PCOMPL-S
```

```
I consider that odd
(See the description of morphological tags, syntactic tags and other notations.)
"<i>"
  "i" <+> <NonMod> PRON PERS NOM SG1 SUBJ @SUBJ
"<consider>"
  "consider" <+> <Vcog> <SVOC/N> <SVOC/A> <as/SVOC/A> <SVO> V PRES -SG3 VFIN @+PMAINV
"<that>"
  "that" PRON DEM SG @OBJ
"<odd>"  
  "odd" A ABS @PCOMPL-O
```
Rule-Based POS Tagging

- Superceded by ENGCG-2 (4000 rules)
  
  Website seems not to work anymore...

```
English Machinese Phrase Tagger 4.6

<table>
<thead>
<tr>
<th>Text Baseform</th>
<th>Phrase syntax and part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>it</td>
<td>nominal head, pro-nominal</td>
</tr>
<tr>
<td>is</td>
<td>main verb, indicative present</td>
</tr>
<tr>
<td>n't</td>
<td>adverbal head, adverb</td>
</tr>
<tr>
<td>that</td>
<td>premodifier, adverb</td>
</tr>
<tr>
<td>odd</td>
<td>nominal head, adjective, sentence boundary</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Baseform</th>
<th>Phrase syntax and part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>nominal head, pro-nominal</td>
</tr>
<tr>
<td>consider</td>
<td>main verb, indicative present</td>
</tr>
<tr>
<td>that</td>
<td>premodifier, adverb</td>
</tr>
<tr>
<td>odd</td>
<td>nominal head, adjective, sentence boundary</td>
</tr>
</tbody>
</table>
```
Tagging Methods

• 3 Basic Approaches
  • Manual rules
  • Machine Learning of rules
  • Statistical models (Hidden Markov Models)
Transformation-Based POS Tagging (TBT)

section 5.6

• basic idea (Brill, 1995)
  • Tag Transformation Rules:
    • change a tag to another tag by inspection of local context
    • e.g. the tag before or after
  • initially
    • use the naïve algorithm to assign tags
  • train a system to find these rules
    • with a finite search space of possible rules
    • error-driven procedure
      • repeat until errors are eliminated as far as possible
  • assume
    • training corpus is already tagged
      • needed because of error-driven training procedure
TBT: *Space of Possible Rules*

- **Fixed window around current tag:**

  ![Diagram]

  - Prolog-based $\mu$-TBL notation (Lager, 1999):
    - $t_0 > t_0' \leftarrow t@[+/-N]$
    - "change current tag $t_0$ to new tag $t_0'$ if word at position $+/-N$ has tag $t$"
TBT: Rules Learned

- **Examples of rules learned**
  (Manning & Schütze, 1999) (µ-TBL-style format):
  - **NN > VB < T0@[−1]**
    - ... to walk ...
  - **VBP > VB < MD[−1,−2,−3]**
    - ... could have put ...
  - **JJR > RBR < JJ@[1]**
    - ... more valuable player ...
  - **VBP > VB < n’t@[−1,−2]**
    - ... did n’t cut ...
    - *(n’t is a separate word in the corpus)*

**Key:**

NN = noun, sg. or mass  
VB = verb, base form  
VBP = verb, pres. (¬3rd person)  
JJR = adjective, comparative  
RBR = adverb, comparative
The μ-TBL System

- Implements Transformation-Based Learning
  - Can be used for POS tagging as well as other applications
- Implemented in Prolog
  - code and data
- [http://www.ling.gu.se/~lager/mutbl.html](http://www.ling.gu.se/~lager/mutbl.html)
  - (dead link)
- Full system for Windows (based on Sicstus Prolog)
  - Includes tagged *Wall Street Journal* corpora
The \( \mu \)-TBL System

- Tagged Corpus (for training and evaluation)
- Format:
  - \( \text{wd}(P,W) \)
    - \( P \) = index of \( W \) in corpus, \( W \) = word
  - \( \text{tag}(P,T) \)
    - \( T \) = tag of word at index \( P \)
  - \( \text{tag}(T_1,T_2,P) \)
    - \( T_1 \) = tag of word at index \( P \), \( T_2 \) = correct tag
- (For efficient access: Prolog first argument indexing)
The µ-TBL System

• Example of tagged WSJ corpus:

  • `wd(63,'Longer'). tag(63,'JJR'). tag('JJR','JJR',63).`
  • `wd(64,maturities). tag(64,'NNS'). tag('NNS','NNS',64).`
  • `wd(65,are). tag(65,'VBP'). tag('VBP','VBP',65).`
  • `wd(66,thought). tag(66,'VBN'). tag('VBN','VBN',66).`
  • `wd(67,to). tag(67,'TO'). tag('TO','TO',67).`
  • `wd(68,indicate). tag(68,'VBP'). tag('VBP','VB',68).`
  • `wd(69,declining). tag(69,'VBG'). tag('VBG','VBG',69).`
  • `wd(70,interest). tag(70,'NN'). tag('NN','NN',70).`
  • `wd(71,rates). tag(71,'NNS'). tag('NNS','NNS',71).`
  • `wd(72,because). tag(72,'IN'). tag('IN','IN',72).`
  • `wd(73,they). tag(73,'PP'). tag('PP','PP',73).`
  • `wd(74,permit). tag(74,'VB'). tag('VB','VB',74).`
  • `wd(75,portfolio). tag(75,'NN'). tag('NN','NN',75).`
  • `wd(76,managers). tag(76,'NN'). tag('NN','NN',76).`
  • `wd(77,to). tag(77,'TO'). tag('TO','TO',77).`
  • `wd(78,retain). tag(78,'VB'). tag('VB','VB',78).`
  • `wd(79,relatively). tag(79,'RB'). tag('RB','RB',79).`
  • `wd(80,higher). tag(80,'JJR'). tag('JJR','JJR',80).`
  • `wd(81,rates). tag(81,'NNS'). tag('NNS','NNS',81).`
  • `wd(82,for). tag(82,'IN'). tag('IN','IN',82).`
  • `wd(83,a). tag(83,'DT'). tag('DT','DT',83).`
  • `wd(84,longer). tag(84,'RB'). tag('RB','JJR',84).`
The μ-TBL System
The µ-TBL System
The µ-TBL System

- **Recall**
  - percentage of words that are tagged correctly with respect to the reference (gold-standard)

- **Precision**
  - percentage of words that are tagged correctly with respect to what the tagger emits

- **F-score**
  - combined weighted average of precision and recall
  - Equally weighted:
    - $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$
The µ-TBL System

Concordance

31 occurrences tagged as VBN that should be VBD:

80966: Mr. Schlessberg thought dinner would be a better
81007: of 1993, the company reported revenue of $1.77 billion
81518: The selectiveness reflected an investors' move to arranged before dinner between the HUD
81701: the private meeting Mr. Wilson discussed. General
81716: chairman, the two men discussed pleasantries and the
81911: Then the dollar finished moderately stronger, barely reacting
82498: it, too.
82768: Raytheon, which used a 1930 British design to
83197: conceived the project but later sold most of its interest to
84111: chairman, Frank Lorenzo disclosed last month that the company expressed interest in buying a
84154: and Southwest Airlines have all

Index

- 31: VBD
- 31: VBD
- 31: VBD
- 28: VBD
- 22: VBD
- 20: VBD
- 19: VBD
- 15: VBD
- 15: VBD
- 14: VBD
- 12: VBD
- 11: VBD
- 10: VBD
- 9: VBD
- 8: VBD
- 8: VBD
- 7: VBD
- 6: VBD
The µ-TBL System

tag transformation rules are
  • human readable
  • more powerful than simple bigrams
  • take less “effort” to train
    • minimum size training corpus 250,000 words
Online Brill Tagger

- [http://cst.dk/online/pos_tagger/uk/](http://cst.dk/online/pos_tagger/uk/)

**CST's Part-Of-Speech tagger (Brill, with adaptations)**

The POS-tagger marks each word in a text with information about word class and morphological features, for example

"This page is about the Brill-tagger"

→

This/DT page/NN is/VBZ about/IN the/DT Brill-tagger/NNP

This page is about the Brill-tagger

[Process my text]
Statistical POS Tagging

• Textbook Section 5.5
  • describes HMM POS Tagging

• Personally, I’ve often used the MXPOST tagger ...
  • Java code (portable) and freely available
  • Maximum entropy tagging
  • Reference:
    • http://www.inf.ed.ac.uk/resources/nlp/local_doc/mxpost_doc.pdf
Jabberwocky

• Lewis Carroll (first stanza), suggested by Kristen Howe (538 student in 2012):

  Has nonsensical words.
  But not random, still obeys English syntax and rhyming ...

• 'Twas brillig, and the slithy toves Did gyre and gimble in the wabe; All mimsy were the borogoves, And the mome raths outgrabe.

  Brillig: Following the poem, the character of Humpty Dumpty comments: "'Brillig' means four o'clock in the afternoon,
  Slithy: Humpty Dumpty says: "'Slithy' means 'lithe and slimy'.
  Tove: Humpty Dumpty says "'Toves' are something like badgers,
  Gyre: "To 'gyre' is to go round and round like a gyroscope."
  Gimble: "To make holes as does a gimlet."
  Wabe: The characters in the poem suggest it means "The grass plot around a sundial"
  Mimsy: "'Mimsy' is 'flimsy and miserable' ."
  Borogove: Following the poem Humpty Dumpty says, "'borogove' is a thin shabby-looking bird
  Mome rath: Humpty Dumpty says following the poem: "A 'rath' is a sort of green pig; but 'mome' I'm not certain about. I think it's short for 'from home',
  Outgrabe: Humpty says "'outgrbing' is something between bellowing and whistling,
Jabberwocky

• A possible translation:
  • It was 4 o'clock, and the lithe and slimy badgers
  • Did go round and round and make holes in the grass around a sundial
  • All flimsy and miserable were the thin, shabby-looking birds
  • And the green pigs from home bellow/whistle
Jabberwocky

- Lewis Carroll (first stanza):
  - 'Twas brillig, and the slithy toves
    NN  CC  DT  NN  NNS
  - Did gyre and gimble in the wabe;
    VBD  NN  CC  NN  IN  DT  NN
  - All mimsy were the borogoves,
    DT  NN
    VBD  DT  NNS
  - And the mome raths outgrabe.
  - CC  DT  NN  NNS  NN

U of Illinois POS Tagger
http://cogcomp.cs.illinois.edu/demo/pos/

78% (18/23)
Jabberwocky

- Lewis Carroll (first stanza):
  - ’T was brillig, and the slithy toves NN
    VBD VBG CC DT JJ NNS
  - Did gyre and gimble in the wabe; VBD
    VBG CC NN IN DT NN
  - All mimsy were the borogoves, DT NN
    VBD DT NNS
  - And the mome raths outgrabe.
  - CC DT NN NNS VBP

CST Brill Tagger
Jabberwocky

- Lingsoft:
Jabberwocky

- Lingsoft:
Jabberwocky

- Lingsoft:
**Jabberwocky**

- **Lingsoft:**

```plaintext
"<*and>"
  "and"  <*> CC @CC
"<the>"
  "the"  <Def> DET CENTRAL ART SG/PL @DN>
"<mome>"
  "mome"  <*> N NOM SG @NN>
"<raths>"
  "raths"  <*> <NoBaseformNormalisation> N NOM SG/PL @NN> @<P
"<outgrabe>"
  "outgrabe"  <*> N NOM SG @PCOMPL-S @APP @<P
"<$..>"
```
Jabberwocky

Should do better than pure POS taggers?

- Sentence Parsing (Stanford parser):

```
'Twas brillig, and the slithy toves

Tagging
'Twas/RB brillig/JJ ,/ and/CC the/DT slithy/JJ toves/NNS

Parse
(ROOT
  (UCP
    (ADVP (RB 'Twas))
    (ADJP (JJ brillig))
    (, ,)
    (CC and)
    (NP (DT the) (JJ slithy) (NNS toves))))

UCP - Unlike Coordinated Phrase.
```
Jabberwocky

• Sentence Parsing (Stanford parser):

*Did gyre and gimble in the wabe;*

Tagging

*Did/NNP gyre/JJ and/CC gimble/JJ in/IN the/DT wabe/NN ;/:*

Parse

(Root
(S
(NP (NNP Did))
(ADJP (JJ gyre)
  (CC and)
  (JJ gimble))
(PP (IN in)
  (NP (DT the) (NN wabe)))
(;})))
Jabberwocky

• Sentence Parsing (Stanford parser):

```
All mimsy were the borogoves,
```

**Tagging**

```
All/DT mimsy/RB were/VBD the/DT borogoves/NNS ,/, 
```

**Parse**

```
(ROOT
 (S
   (NP (DT All))
   (ADVP (RB mimsy))
   (VP (VBD were)
     (NP (DT the) (NNS borogoves)))
   (, ,)))
)```
Jabberwocky

• Sentence Parsing (Stanford parser):

   And the mome raths outgrabe.

   Tagging
   And/CC the/DT mome/NN raths/VBZ outgrabe/JJ ./.

   Parse

   (ROOT
     (S (CC And)
       (NP (DT the) (NN mome))
       (VP (VBZ raths)
         (ADJP (JJ outgrabe))
       ))}

VBZ: 3rd person singular, ends in -s