# Language Technology
# http://cs.lth.se/edan20/
## Chapter 7: Part-of-Speech Tagging Using Rules

Pierre Nugues

Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

September 17, 2020

# The Parts of Speech

The parts of speech (POS) are classes that correspond to the lexical – or word – categories
Plato made a distinction between the verb and the noun.
After him, the word categories further evolved and grew in number until Dionysus Thrax formulated and fixed them.
Aelius Donatus popularized the list of the eight parts of speech: noun, pronoun, verb, participle, conjunction, adverb, preposition, and interjection.
Grammarians have adopted these POS for most European languages although they are somewhat arbitrary

# Part-of-speech Annotation

Sentence:

*That round table might collapse*

Annotation:

| Words | Parts of speech | POS tags |
|---|---|---|
| **that** | Determiner | DT |
| **round** | Adjective | JJ |
| **table** | Noun | NN |
| **might** | Modal verb | MD |
| **collapse** | Verb | VB |

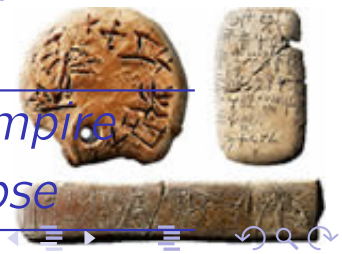
The automatic annotation uses predefined POS tagsets such as the Penn Treebank tagset for English

# Word Ambiguity

| | English | French | German |
|---|---|---|---|
| Part of speech | *can* modal | *le* article | *der* article |
| | *can* noun | *le* pronoun | *der* pronoun |
| Semantic | *great* big | *grand* big | *groß* |
| | *great* notable | *grand* notable | *groß* |

# POS Tagging

| Words | Possible tags | Example of use |
|---|---|---|
| **that** | Subordinating conjunction | *That he can swim is good* |
| | Determiner | *That white table* |
| | Adverb | *It is not that easy* |
| | Pronoun | *That is the table* |
| | Relative pronoun | *The table that collapsed* |
| **round** | Verb | *Round up the usual suspects* |
| | Preposition | *Turn round the corner* |
| | Noun | *A big round* |
| | Adjective | *A round box* |
| | Adverb | *He went round* |
| **table** | Noun | *That white table* |
| | Verb | *I table that* |
| **might** | Noun | *The might of the wind* |
| | Modal verb | *She might come* |
| **collapse** | Noun | *The collapse of the empire* |
| | Verb | *The empire can collapse* |



# Part-of-Speech Ambiguity in Swedish

The word *som* in the *Norstedts svenska ordbok*, 1999, has three entries:

1. *Om jag vore lika vacker som du, skulle jag vara lycklig.* (konjunktion)
2. *Bilen som jag köpte i fjol.* (pronomen)
3. *Som jag har saknat dig.* (adverb)

The part-of-speech difference can be significant:

Swedish. Compare the pronunciation of *vaken*, adjective, as in *Han är aldrig vaken innan klockan sju* and *vaken*, noun, as in *Vi fiskade i vaken i sjön*

English. Compare *object* in *I object to violence*, verb, or *I could see an object*, noun.

# Simple Grammatical Constraints are not Satisfying

Although, it makes no sense,

*I see a bird*

can be tagged as:

*I*/**noun** *see*/**noun** *a*/**noun** *bird*/**noun**

Because sequences of four nouns are possible in English as in:

*city school committee meeting.*

The disambiguation methods are based on

- Handcrafted rules
- Automatically learned rules
- Statistical methods

Currently disambiguation accuracy is greater than 95% for many languages

# POS Annotation with Rules

The phrase *The can rusted* has two readings
Let's suppose that *can*/modal is more frequent than *can*/noun in our corpus

First step: Assign the most likely POS
*The*/art *can*/modal *rusted*/verb

Second step: Apply rules
Change the tag from modal to noun if one of the two previous words is an article
*The*/art *can*/noun *rusted*/verb

This is the idea of Brill's tagger.

# Rule Templates

| Rules | Explanation |
|---|---|
| `alter(A, B, prevtag(C))` | Change A to B if preceding tag is C |
| `alter(A, B, nexttag(C))` | Change A to B if the following tag is C |
| `alter(A, B, prev2tag(C))` | Change A to B if tag two before is C |
| `alter(A, B, next2tag(C))` | Change A to B if tag two after is C |
| `alter(A, B, prev1or2tag(C))` | Change A to B if one of the two preceding tags is C |
| `alter(A, B, next1or2tag(C))` | Change A to B if one of the two following tags is C |
| `alter(A, B, surroundingtag(C, D))` | Change A to B if surrounding tags are C and D |
| `alter(A, B, nextbigram(C, D))` | Change A to B if next bigram tag is C D |
| `alter(A, B, prevbigram(C, D))` | Change A to B if previous bigram tag is C D |

# Learning Rules Automatically

Compare the hand-annotation of the reference corpus with the automatic one

| Automatic tagging | Hand annotation: gold standard |
|---|---|
| *The*/art *can*/modal *rusted*/verb | *The*/art *can*/noun *rusted*/verb |

For each error instantiate the templates
Rules correcting the error

```
alter(modal, noun, prevtag(art)).
alter(modal, noun, prev1or2tag(art)).
alter(modal, noun, nexttag(verb))
alter(modal, noun, surroundingtag(art, verb))
```

Rules introduce good and bad transformations
Select the rule that has the greatest error reduction and apply it

# Part-of-Speech Ambiguity in Swedish

The Swedish word *den* can be a determiner or a pronoun.
It corresponds to two entries in the *Nordstedts svenska ordbok* (1999, page 187):

- **den** artikel . . . som här antas vara känd . . . : **den** *nya bilen*
- **den** pron. personen eller företeelsen som är omtalad i sammanhanget . . . : *Var har du köpt kameran? Jag har fått* **den** *i present.*

Frequency information:

```
egrep -i "den  dt" talbanken.txt | wc -l
     820
egrep -i "den  pn" talbanken.txt | wc -l
     256
```

# Ambiguity Resolution in Swedish: The Baseline

Let us suppose that *den* is the only word to tag in the corpus and that it has two possible parts of speech: dt and pn.
Using the most frequent part of speech produces the annotations:

| Den | nya | läroplanen | innebär | också | ... |
|---|---|---|---|---|---|
| dt | jj | nn | vb_fin | ab | |

| Jag | har | fått | den | i | present |
|---|---|---|---|---|---|
| pn | vb_fin | vb | dt | pp | nn |

If the POS tagger is restricted to *den*, out of 820 + 256 = 1076 POS assignments,

$$\frac{820}{1076} = 76\%$$

are correct.

# Ambiguity Resolution in Swedish: The Rule Templates

Let us use two rules templates `alter(A, B, prev(C))` and `alter(A, B, next(C))` and instantiate them with the error on *Jag har fått den i present*.

| Jag | har | fått | den | i | present |
|---|---|---|---|---|---|
| pn | vb_fin | **vb** | (dt → pn) | **pp** | nn |

It yields:

1. Change `dt` to `pn` if previous POS tag is `vb`: `alter(dt, pn, prev(vb))`
2. Change `dt` to `pn` if next POS tag is `pp`: `alter(dt, pn, next(pp))`

Both rules produce a correct annotation on the training example.

# Ambiguity Resolution in Swedish: Selecting the Rules

Let us apply the two rules to all the occurrences of *den* in the corpus and ignore all the other words:

- The first rule corrects 15 wrong annotations of *den* and introduces 59 mistakes: $15 - 59 = -44$
- The second rule corrects 20 wrong annotations and introduces 5 mistakes: $20 - 5 = \mathbf{+15}$

The training step of Brill's tagger selects the most efficient rule, here `alter(dt, pn, next(pp))`.
Of course, this step is applied to all the ambiguous words and not only *den*.
We iterate the procedure until the error rate is below a certain threshold.

# Brill's Learning Algorithm

| St. | Operation | Input | Output |
|---|---|---|---|
| **1.** | Annotate each word of the corpus with its most likely part of speech | *Corpus* | *AnnotatedCorpus(1)* |
| **2.** | Compare pairwise the part of speech of each word of the *AnnotationReference* and *AnnotatedCorpus(i)* | *AnnotationReference* *AnnotatedCorpus(i)* | List of errors |
| **3.** | For each error, instantiate the rule templates to correct the error | List of errors | List of tentative rules |
| **4.** | For each instantiated rule, compute on *AnnotatedCorpus(i)* the number of good transformations minus the number of bad transformations the rule yields | *AnnotatedCorpus(i)* Tentative rules | Scored tentative rules |

# Brill's Learning Algorithm

| St. | Operation | Input | Output |
|---|---|---|---|
| **5.** | Select the rule that has the greatest error reduction and append it to the ordered list of transformations | Tentative rules | *Rule(i)* |
| **6.** | Apply *Rule(i)* to *Annotated-Corpus(i)* | *AnnotatedCorpus(i)* *Rule(i)* | *AnnotatedCorpus(i+1)* |
| **7.** | If number of errors is under predefined threshold, end the algorithm else go to step 2. | – | List of rules |

# First Brill's Rules

| # | Change From | Change To | Condition |
|---|---|---|---|
| 1 | NN | VB | Previous tag is TO |
| 2 | VBP | VB | One of the previous three tags is MD |
| 3 | NN | VB | One of the previous two tags is MD |
| 4 | VB | NN | One of the previous two tags is DT |
| 5 | VBD | VBN | One of the previous three tags is VBZ |

In the table, rules consider parts of speech only. This is the normal case and they are called unlexicalized.
Rules can also consider word values and they are called lexicalized.

| Change From | Change To | Condition |
|---|---|---|
| IN | RB | The word two positions to the right is *as* |

# Standard POS Tagsets: The Penn Treebank

| | | | | | |
|---|---|---|---|---|---|
| 1. | CC | Coordinating conjunction | 25. | TO | *to* |
| 2. | CD | Cardinal number | 26. | UH | Interjection |
| 3. | DT | Determiner | 27. | VB | Verb, base form |
| 4. | EX | Existential *there* | 28. | VBD | Verb, past tense |
| 5. | FW | Foreign word | 29. | VBG | Verb, gerund/present participle |
| 6. | IN | Preposition/sub. conjunction | 30. | VBN | Verb, past participle |
| 7. | JJ | Adjective | 31. | VBP | Verb, non-third pers. sing. pres. |
| 8 | JJR | Adjective, comparative | 32. | VBZ | Verb, third-pers. sing. present |
| 9. | JJS | Adjective, superlative | 33. | WDT | *wh*-determiner |
| 10. | LS | List item marker | 34. | WP | *wh*-pronoun |
| 11. | MD | Modal | 35. | WP$ | Possessive *wh*-pronoun |
| 12. | NN | Noun, singular or mass | 36. | WRB | *wh*-adverb |
| 13. | NNS | Noun, plural | 37. | # | Pound sign |
| 14. | NNP | Proper noun, singular | 38. | $ | Dollar sign |
| 15. | NNPS | Proper noun, plural | 39. | . | Sentence final punctuation |
| 16. | PDT | Predeterminer | 40. | , | Comma |
| 17. | POS | Possessive ending | 41. | : | Colon, semicolon |
| 18. | PRP | Personal pronoun | 42. | ( | Left bracket character |
| 19. | PRP$ | Possessive pronoun | 43. | ) | Right bracket character |
| 20. | RB | Adverb | 44. | " | Straight double quote |
| 21. | RBR | Adverb, comparative | 45. | ‘ | Left open single quote |
| 22. | RBS | Adverb, superlative | 46. | “ | Left open double quote |
| 23. | RP | Particle | 47. | ’ | Right close single quote |

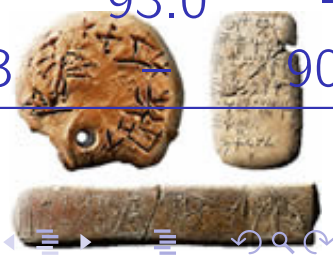
# An Example of Tagged Text from the Penn Treebank

Battle-tested/JJ Japanese/JJ industrial/JJ managers/NNS here/RB always/RB buck/VBP up/RP nervous/JJ newcomers/NNS with/IN the/DT tale/ NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
“/“ From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ”/” says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN the/DT Mitsui/NNP group/NN ’s/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.

# Measuring Quality: The Confusion Matrix

From Franz (1996, p. 124)

| ↓Correct | Tagger → | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DT | IN | JJ | NN | RB | RP | VB | VBD | VBG | VB |
| DT | 99.4 | 0.3 | – | – | 0.3 | – | – | – | – | – |
| IN | 0.4 | 97.5 | – | – | 1.5 | 0.5 | – | – | – | – |
| JJ | – | 0.1 | 93.9 | 1.8 | 0.9 | – | 0.1 | 0.1 | 0.4 | 1.5 |
| NN | – | – | 2.2 | 95.5 | – | – | 0.2 | – | 0.4 | – |
| RB | 0.2 | 2.4 | 2.2 | 0.6 | 93.2 | 1.2 | – | – | – | – |
| RP | – | 24.7 | – | 1.1 | 12.6 | 61.5 | – | – | – | – |
| VB | – | – | 0.3 | 1.4 | – | – | 96.0 | – | – | 0.2 |
| VBD | – | – | 0.3 | – | – | – | – | 94.6 | – | 4.8 |
| VBG | – | – | 2.5 | 4.4 | – | – | – | – | 93.0 | – |
| VBN | – | – | 4.6 | – | – | – | – | 4.3 | – | 90. |



# Recognizing Parts of Speech

Parts of speech denomination is comparable in Western European languages and roughly corresponds
They follow Donatus' teaching
(`http://htl2.linguist.jussieu.fr:8080/CGL/text.jsp?id=T28`)
If you are not sure, look up in a dictionary
Two common mistakes in the labs:

- Confusion between noun and the Swedish word *namn*.
  - A common noun, or more simply a noun, corresponds to *substantiv*
  - Proper noun, or name, (or proper name) corresponds to *namn* or *egennamn*.
- Possessive pronouns like *my*, *your*, *his*, *her*, ... are not real pronouns. They should be called possessive adjectives or determiners.

# Multext and Google's Universal POS tagset

| Part of speech | Multext | Universal POS, new version |
|---|---|---|
| **Noun** | N | NOUN, PROPN |
| **Verb** | V | VERB, AUX |
| **Adjective** | A | ADJ |
| **Pronoun** | P | PRON |
| **Determiner** | D | DET |
| **Adverb** | R | ADV |
| **Adposition (Preposition)** | S | ADP |
| **Conjunction** | C | CCONJ, SCONJ |
| **Numeral** | M | NUM |
| **Interjection** | I | INTJ |
| **Residual** | X | X |
| **Particle** | - | PART |
| **Ponctuation mark** | - | PUNCT |
| **Symbol** | - | SYM |

# Attributes for Nouns (Multext)

| Position | Attribute | Value | Code |
|---|---|---|---|
| 1 | Type | Common | c |
| | | Proper | p |
| 2 | Gender | Masculine | m |
| | | Feminine | f |
| | | Neuter | n |
| 3 | Number | Singular | s |
| | | Plural | p |
| 4 | Case | Nominative | n |
| | | Genitive | g |
| | | Dative | d |
| | | Accusative | a |