Rule-Based POS

- An Introduction
- EngCG Tagger – ENGTWOL

Rule-based Tagging

- 2 stage early solution:
  - Dictionary is used to assign to each word a list of potential POS tags
  - Large list of hand-written rules is applied to disambiguate specific tags
- Modern solutions:
  - 2 stage architectures
  - Vastly larger dictionary and rule sets
- EngCG tagger is based on Constraint Grammar approach

EngCG Tagger (Voutilainen, 1995)

- The EngCG ENGTWOL Lexicon:
  - Based on two-level morphology
  - Using 56,000 entries for English word stems (Heikkilä, 1995)
  - Counting a word with multiple POS as separate entries
  - Not counting inflected and many derived forms
  - Each entry is annotated with a set of morphological and syntactic features

ENGTWOL (Voutilainen, 1995)

Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
entire	ADJ	ABSOLUTE ATTRIBUTIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	IMPERATIVE VFIN
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

Sample lexical entries from the ENGTWOL lexicon.

EngCG Tagger (Voutilainen, 1995)

SG – singular

-SG3 – other than third-person singular

ABSOLUTE – non-comparative and non-superlative adjective

NOMINATIVE – non-genitive

PCP2 – past participle

PRE – predeterminers

NOINDEFDETERMINER – words like furniture do not appear with the indefinite determiner a

SV – verb appears with a subject (nothing occurred)

SVO – verb appears with a subject and an object (I showed the film)

SVOO – with a subject and two complements: (She showed her the ball)

EngCG Tagger

- In the first stage of tagger,
  - each word is run through the two-level lexicon transducer and the entries for all possible POS are returned.
- Example: *Pavlov had shown that salivation...*

Pavlov	**PAVLOV N NOM SG PROPER**
had	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown	**SHOW PCP2 SVOO SVO SV**
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	**N NOM SG**

EngCG Tagger

- Second stage: Apply the constraints
- A set of about 3,744 constraints (EngCG-2 system) are then applied to the input sentences to rule out incorrect POS.
- To eliminate tags that are inconsistent with the context, use the constraints in a negative way

ADVERBIAL-THAT RULE
Given input: "that"
if
(+1 A/ADV/QUANT); /* *if next word is adj, adverb, or quantifier* */
(+2 SENT-LIM); /* *and following which is a sentence boundary,* */
(NOT -1 SVOC/A); /* *and the previous word is not a verb like* */
/* *'consider' which allows adj as object complements* */
then eliminate non-ADV tags
else eliminate ADV tags

Ex: *It is not that odd,*

ENGTWOL (Voutilainen, 1995)

- The system also includes probabilistic constraints, and also make use of other syntactic information
- Rule to express the constraint : complementizer '*that*'

COMPLEMENTIZER-THAT RULE
Given input: "that"
if
/* *if the previous word is verb* */
/* which expects a compliment *'believe', 'think' or 'show'* */
/* *and if 'that' is followed by the beginning of a noun phrase, finite verb*/*
then CS tag

Ex: *"I thought <u>that</u> you might like some milk"*

Thank You