INF5820

Natural Language Processing - NLP

H2009

Jan Tore Lønning
jtl@ifi.uio.no

Part of Speech Tagging

INF5830

Lecture 2

Aug 31 2009

Part of speech tagging

- Example: Oslo-Bergen-tagger

POS Tagging

- J&M: “The process of assigning a part-of-speech or lexical class marker to each word in a collection.”

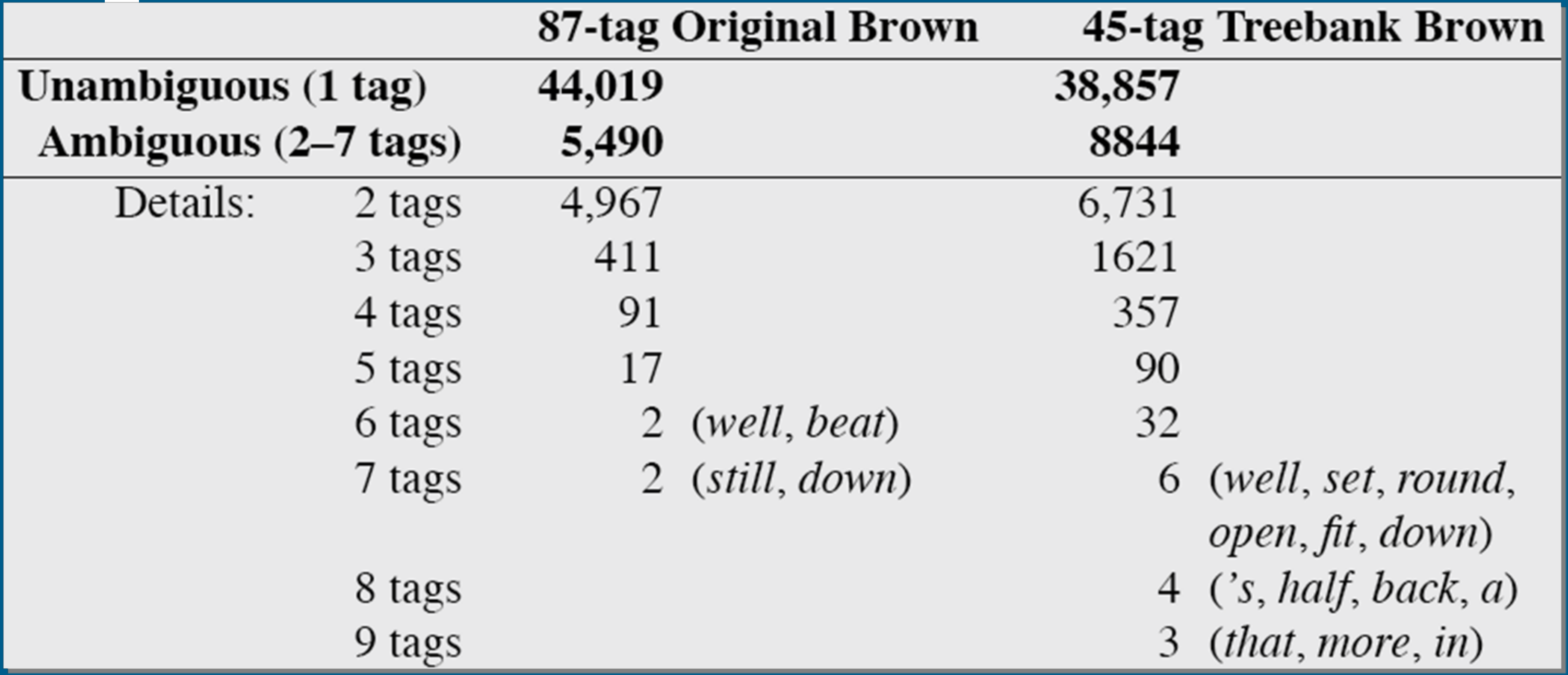
WORD	tag
the	**DET**
koala	**N**
put	**V**
the	**DET**
keys	**N**
on	**P**
the	**DET**
table	**N**

POS Tagging

- Words often have more than one POS: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

These examples from Dekang Lin

How Hard is POS Tagging? Measuring Ambiguity



		87-tag Original Brown		**45-tag Treebank Brown**	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*’s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

Methods for POS Tagging

1. Rule-based tagging
   - (ENGTWOL)
2. Stochastic
   1. Probabilistic sequence models
      - HMM (Hidden Markov Model) tagging
      - MEMMs (Maximum Entropy Markov Models)
3. Transformation-based tagger (Brill)
   1. Rule-based +
   2. Relearning

Different approaches

Deep	Grammars, parsing	
	CG: Syntactic categories	
Shallow, low-level	Rule-based tagging (CG)	HMM-tagging, MaxEnt-tagging
	Rule-based Hand-written	Stochastic Machine learning

CG-tagger

- Steps in the tagging process:
  1. Preprocessing
     1. Tokenization: from characters to tokens
     2. Sentence segmentation
  2. Morphological analysis, multi-tagging
     1. Assign all possible tags to all tokens
  3. Disambiguation
     1. Remove contextually impossible tags (using a set of hand-written rules)
     2. Keep 1+ tags for each token

2. Morphological analysis – multi-tagging

- Assign all possible tags to all tokens
- Alt.1 Fullform lexicon, containing
  - All words: *run, runs, running, ran, run, …*
  - With associated tags
- Alt. 2 Lexeme lexicon (*run*)+
  - Morphological analyzer:
    - *run, runs, ran, running …*
    - Tag
  - Efficiency
  - Finnish: 2000 forms of a noun, 12000 forms of a verb

Part of speech tagging

- Example: Oslo-Bergen-tagger

CG-tagger

- Steps in the tagging process:
  1. Preprocessing
     1. Tokenization: from characters to tokens
     2. Sentence segmentation
  2. Morphological analysis, multi-tagging
     1. Assign all possible tags to all tokens
  3. Disambiguation
     1. Remove contextually impossible tags (using a set of hand-written rules)
     2. Keep 1+ tags for each token

Example: Adverbial “that” rule

- Eliminates all readings of “that” except the one in
  - “It isn’t *that* odd”

```
Given input: “that”
    If
    (+1 A/ADV/QUANT)  ;if next word is adj/adv/quantifier
    (+2 SENT-LIM)       ;following which is E-O-S
    (NOT -1 SVOC/A)      ; and the previous word is not a
                            ; verb like “consider” which
                            ; allows adjective complements
                            ; in “I consider that odd”
    Then eliminate non-ADV tags
    Else eliminate ADV
```

Hand-written rules

```
#:1898
 REMOVE:1898 (verb perf-part) IF
              (-1 %til%)
              (-2 %og%)
              (NOT -3 %av%)
;
#  "De kjørte til låven og til huset (ikke perf-part)"))
```

- Eks
  - # 3044
  - #2391-92
  - #2421
  - #5088 – spesifik
- Regelformat: http://visl.sdu.dk/cg2_howto.html

Tagging vs parsing

- A tagger faces the same two tasks as a grammar-based parser
- Ambiguity:
  - Choose the correct tag sequence between several candidates
- Coverage:
  - Assigning tags to words not in the lexicon:
    - Proper names
    - New words
    - Compounds
    - typos

CG-syntax

- After POS-tagging/Morph. Disambiguation:

4. Map tags to sets of possible syntactic functions
5. Run disambiguator for synt. Function

- Uses similar types of rules and processing as morph. Analyzor
- See examples

CG-rule format for tagging

- Rules may refer to
  - Morph. Categories (tags)
  - Word forms
- Rules may be general:
  - Part of a tag (=class of tags), e.g. all verbs.
  - Sets of words
- Specific: single words
- Contexts:
  - Local, neighbors
  - Anywhere in the sentence
- Rule-format developed over time: CG, CG2, CG3

CG-processing

- Two layers of rules:
  - All normal rules are tried first
  - The heuristic rules
- Possible rule conflicts (within a layer)
  - Determined by rule-order (outside the formalism)
- Rules compiled into finite automata
  - Easily combined
  - Fast processing

Ambiguity

- A CG-tagger leaves ambiguities:

	VB	PRP$	NN
PRP	VBD	PRP	VB
I	saw	her	duck

- How to determine the possible parses?
  - PRP VB PRP$ NN
  - PRP VBD PRP$ NN
  - PRP VBD PRP VB
- In contrast to the impossible ones:
  - PRP VBD PRP VB
  - + 4more

Coverage: unknown words

- All possible tags?
  - No – too many
- Spell correction? (typos)
- Guess tags:
  - From morphology:
    - -ing: VBG, JJ, N
    - Norw.: -er: V_pres, N_pl
    - Starting capital: proper name
  - From frequency
    - Proper names
    - Nouns
- Norw., German, etc:
  - Compound analysis

Stochastic tagging:

HMM-TAGGING

And then

- Some statistics:
  - Product rule
  - Stochastic variable
- J & M,Chap. 5, slide 26-36
- Morkov-models slides