Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard Karls Universität Tübingen

Germany

Part-of-speech (POS) Tagging

- Part-of-speech tagging refers to the assignment of (disambiguated) morpho-syntactic categories, in particular word class information, to individual tokens.
- Part-of-speech tagging requires a pre-defined tagset and a tagset assignment algorithm.
- Disambiguation of part-of-speech labels takes local context into account.

Criteria for the Construction of Tagsets

Geoffrey Leech proposed general guidelines for the design of tagsets:

- **Conciseness:** Brief labels are often more convenient to use than verbose, lengthy ones.
- **Perspicuity:** Labels which can easily be interpreted are more user-friendly than labels which cannot.
- **Analysability:** Labels which are decomposable into their logical parts are better (particularly for machine processing).

Tagset Design and Use

- Standardization
  - Cross-linguistic guidelines for tagsets and tagging corpora have been proposed by the Text Encoding Initiative (TEI)

    Link: `www.tei-c.org`
- Tagset size
  - Trade-off between linguistic adequacy and tagger reliability
  - The larger the tagset, the more training data are needed for statistical part-of-speech taggers

Tagsets for English (1)

Tagsets are often developed in conjunction with corpus collections.

- The Brown Corpus tagset
  - First used for the annotation of the Brown Corpus of American English
  - Later adapted for the annotation of the Penn Treebank of American English

Tagsets for English (2)

- CLAWS
  - First designed for the annotation of the Lancaster-Oslo-Bergen corpus (LOB corpus). LOB is the British English counterpart of the Brown Corpus of American English.
  - Later adapted for the annotation of the British National Corpus (BNC), the largest corpus of British English with approximately 100 million words of running text.

Part-of-speech Tagging – An Example

Example from BNC using C7 (adapted version of CLAWS) tagset:

Perdita&NN1-NP0; ,&PUN; covering&VVG; the&AT0; bottom&NN1; of&PRF; the&AT0; lorries&NN2; with&PRP; straw&NN1; to&TO0; protect&VVI; the&AT0; ponies&NN2; ’&POS; feet&NN2; ,&PUN; suddenly&AV0; heard&VVD-VVN; Alejandro&NN1-NP0; shout-ing&VVG; that&CJT; she&PNP; better&AV0; dig&VVB; out&AVP; a&AT0; pair&NN0; of&PRF; clean&AJ0; breeches&NN2; and&CJC; polish&VVB; her&DPS; boots&NN2; ,&PUN; as&CJS; she&PNP; ’d&VM0; be&VBI; playing&VVG; in&PRP; the&AT0; match&NN1; that&DT0; afternoon&NN1; .&PUN;

Part-of-speech Tagging – An Example

The codes used are:

AJ0:	general adjective	POS:	genitive marker
AT0:	article	PNP:	pronoun
	neutral for number		
AV0:	general adverb	PRF:	of
AVP:	prepositional adverb	PRP:	prepostition
CJC:	co-ord. conjunction	PUN:	punctuation
CJS:	subord. conjunction	TO0:	infinitive to
CJT:	that conjunction	VBI:	be
DPS:	possessive determiner	VM0:	modal auxiliary
DT0:	singular determiner	VVB:	base form of verb

Part-of-speech Tagging – An Example

The codes used are:

NN0:	common noun, neutral for number	VVD:	past tense form of verb
NN1:	singular common noun	VVG:	-ing form of verb
NN2:	plural common noun	VVI:	infinitive form of verb
NP0:	proper noun	VVN:	past participle form of verb

General Issues Visible in the Example

- Tags are attached to words by the use of TEI entity references delimited by ‘&’ and ‘;’.
- Some of the words (such as *heard*) have two tags assigned to them. These are assigned in cases where there is a strong chance that there is not sufficient contextual information for unique disambiguation.
- Approximation of a logical tagset (possible trade-off with mnemonic naming conventions).

Example (Penn Treebank tagset)

all	DT	I am busy *all* afternoon on Thursday
	PDT	if you move *all* the way to the fourth of August
	NN	*all* I have open is the morning
	RB	you said you were *all* full
along	RB	so moving right *along* let us see
	RP	we can just take them *along*
	IN	I was thinking more *along* the lines of December
begin-	NN	you are thinking *beginning* of next week
ning	VBG	I would only have time *beginning* at the 21st
	JJ	I am gone the whole *beginning* part of the week

Penn Treebank Tagset

CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subord. conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle

Penn Treebank Tagset (2)

NN	Noun, sg or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd per. sg. present
NNP	Proper noun, sg	VBZ	Verb, 3rd per. sg. present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Tagsets for other Languages

- German: Stuttgart/Tübingen Tagset (STTS)

  Link: `www.sfs.uni-tuebingen.de /Elwis/stts/stts.html`

- MULTEXT-East: Tagsets for Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene)

  Link: `http://nl.ijs.si/ME/`

The Stuttgart-Tübingen Tagset STTS

- The STTS is a set of 54 tags for annotating German text corpora with part-of-speech labels.
- The STTS guidelines (available on the website) explain the use of each tag by illustrative examples to aid human annotators in consistent corpus annotation by STTS tags.
- It was jointly developed by the Institut für maschinelle Sprachverarbeitung of the University of Stuttgart and the Seminar für Sprachwissenschaft of the University of Tübingen.

The Stuttgart-Tübingen Tagset STTS

1. Nomina (N)
2. Verben (V)
3. Artikel (ART)
4. Adjektive (ADJ)
5. Pronomina (P)
6. Kardinalzahlen (CARD)
7. Adverbien (ADV)
8. Konjunktionen (KO)
9. Adpositionen (AP)
10. Interjektionen (ITJ)
11. Partikeln (PTK)

Table 1: Tags for major word classes

STTS Tags

POS =	Description	Examples
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präpos.; Zirkumpos. links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*

STTS Tags (2)

POS =	Description	Examples
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit “] A big fish [” übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit “zu” und Infinitiv	*um [zu leben], anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit, wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*

STTS Tags (3)

POS =	Description	Examples
NN	normales Nomen	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch], irgendein [Glas]*

STTS Tags (4)

POS =	Description	Examples
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpron.	*mein [Opa], deine [Oma]*
	Relativpronomen	
PRELS	substituierend	*[der Hund,] der*
PRELAT	attribuierend	*[der Mann ,]* *dessen [Hund]*

STTS Tags (5)

POS =	Description	Examples
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe], wessen [Hut]*
PWAV	adverbiales Interrogativ– oder Relativpronomen	*warum, wo, wann, worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen*
PTKZU	"zu" vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*

STTS Tags (6)

POS =	Description	Examples
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten], zu [schnell]*
TRUNC	Kompositions–Erstglied	*An– [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit “zu”, voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*

STTS Tags (7)

POS =	Description	Examples
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*

STTS Tags (8)

POS =	Description	Examples
$,	Komma	**,**
$.	Satzbeendende Interpunktion	**. ? ! ; :**
$(	sonstige Satzzeichen; satzintern	**– [ ] ( )**

Automatic POS Tagging: Basic Issues

- Use a word list or lexicon and disambiguate or tag without lexicon or word list?
- If there is more than one possible tag for a word, how to select the correct one?
- The unkown word problem: What happens if the word is not in the word-tag list?
- How rich is the tagset?
  - word = full form (incl. morphological information), or
  - word = lemma (word class information without morphology)?

POS Tagging: Main Approaches

- Rule-based approach:

  Write local disambiguation rules.

- Stastistical approach:

  Compile statistics from a corpus to train a statistical model.

- Machine learning approach:

  Compile (weighted) patterns of features and values from a corpus to train a classifier.

Rule-Based Approach

- Leading ideas:
  - Usually only local context needed for disambiguation.
  - Formulate context-sensitive disambiguation rules.
- Example:

?	VBZ	$\rightarrow$	not NNS
NNS	?	$\rightarrow$	not VBZ

Problems with Rule-Based Approach

- Rules can only be used when necessary context is not ambiguous.
- There are too many ambiguous contexts.
- The rules are dependent on the tagset.
- Manual encoding is time-consuming.

Statistical Approach

- Collect table of tag frequencies from hand-annotated training corpus.
  - E.g.: freq(DT NN) = 10 171, freq(TO NN) = 5
- But the frequency for rare tags is low.
  - freq(NN POS) = 36, freq(POS) = 71
  - in comparison: freq(NN) = 24 211
- Solution: Compute conditional probability:
  - P(NN|DT) = (freq(DT NN))/(freq(DT)) = 0.420,
  - P(POS|NN) =(freq(NN POS))/(freq(NN)) = 0.0015

Hidden Markov Taggers

For a given sentence or word sequence, Hidden Markov Model taggers (HMM taggers) compute the most likely sequence of tags $t_{1,n}$ by finding the maximum value for the following formula:

P(word | tag) * P(tag | previous $n-1$ tags)

In symbols:

$$\underset{t_{1,n}}{\texttt{arg}\,\texttt{max}} P(t_{1,n} \mid w_{1,n}) = \prod_{i=1}^{n} P(w_i \mid t_{1,n}) * P(t_n \mid t_{1,n-1}) \\ * P(t_{n-1} \mid t_{n-1,n-2}) * \cdots * P(t_2 \mid t_1)$$

Simplifying HMM Taggers

Computation with the previous formula is very complex since it takes into account the entire left context for a given word.

Solution:

- Experience from rule-based approaches:
  - Local context is sufficient for disambiguation in most cases.
  - Selection of the current tag often depends on the preceding tag.
- Thus, limit the window of context tags that enter into the computation.

First-order HMM Taggers

Only bi-grams of tags (tag and its preceding tag) are considered:

P(word | tag) * P(tag | previous tag)

In symbols:

$$\underset{t_{1,n}}{\arg\max} P(t_{1,n} \mid w_{1,n}) = \prod_{i=1}^{n} P(w_i \mid t_i) * P(t_i \mid t_{i-1})$$

Second-order HMM Taggers

Only tri-grams of tags (tag and its preceding tag) are considered:

P(word | tag) * P(tag | two previous tags)

In symbols:

$$\underset{t_{1,n}}{\texttt{arg}\,\texttt{max}} \mathrm{P}(\mathrm{t}_{1,n} \mid \mathrm{w}_{1,n}) = \prod_{i=1}^{n} \mathrm{P}(\mathrm{w}_i \mid \mathrm{t}_i) * \mathrm{P}(\mathrm{t}_i \mid \mathrm{t}_{i-1}\,\mathrm{t}_{i-2})$$

Obtaining Probabilities

- Conditional probabilities for tag sequences and for word (given a tag) are computed from the frequency tables generated from the training corpus.

- The size of the training corpus needed for good results is proportional to the size of the tagset.

Sequences of Ambiguous Tags

sentence:	would	you	mind	working	on	Saturday
possible tags:	MD	PRP	VB	VBG	IN	NP
		PRP$	VBP	NN	RP	
			NN		RB	

- In order to find the most likely path, the probabilities in each path are multiplied
- P(MD PRP VB VBG IN NP) = P(PRP|MD) x P(VB|PRP) x P(VBG|VB) x P(IN|VBG) x P(NP|IN) = 0.032 x 0.161 x 0.087 x 0.027 x 0.335 = 0.00000405541
- P(MD PRP$ NN NN RB NP) = P(PRP$|MD) x P(NN|PRP$) x P(NN|NN) x P(RB|NN) x P(NP|RB) = 0.002 x 0.053 x 0.076 x 0.056 x 0.038 = 0.000000171432

Advantages of Statistical Approach

- Very robust, can process any input strings
- Training is automatic, very fast
- Can be retrained for different corpora/tagsets without much effort

Disadvantages of Statistical Approach

- Requires a great amount of (annotated) training data.
- The linguist cannot influence the performance of the trained model.
- Changes in the tagset $\rightarrow$ changes in the word list (+ changes in the morphology) + changes in the corpus
- Can only model local dependencies.

Freely Available POS Taggers

- TnT Computerlinguistik Saarbrücken, HMM tri-gram tagger,

  `http://code.google.com/p/hunpos/`
- Brill Tagger transformation-based error-driven,

  `http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/parsing/taggers/brill/0.html`
- TreeTagger by Helmut Schmid, University of Stuttgart

  `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`