Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard-Karls-Universität Tübingen

Germany

Morphology: The Naive Solution

The simplest, but for most cases naive solution:

- Compile a full-form lexicon which lists all possible word forms together with their morphological analyses.
- If a given word has only one morphological analysis, the full-form lexicon stores exactly one reading.
- If a given word has more than one morphological analysis, the full-form lexicon stores all possible readings separately.

Morphological Analysis: Lemmatization

- Lemmatization refers to the process of relating individual word forms to their citation form (lemma) by means of morphological analysis.
- Lemmatization provides a means to distinguish between the total number of word tokens and distinct lemmata that occur in a corpus.
- Lemmatization is indispensible for highly inflectional languages which have a large number of distinct word forms for a given lemma.

Examples from English (1)

Input: *spies*

Analysis:

spies	spy+Noun+Pl
spies	spy+Verb+Pres+3sg

Input: *travelling*

Analysis:

travelling	travel+Verb+Prog
travelling	travelling+Adj
travelling	travelling+Noun+Sg

Examples from English (2)

Input: *foxes*

Analysis:

foxes	fox+Noun+Pl
foxes	fox+Verb+Pres+3s

Input: *moved*

Analysis:

moved	move+Verb+PastBoth+123SP
moved	moved+Adj

Examples from German (1)

Input: *Staubecken*

Analysis:

1. Stau+Noun+Common+Masc+Sg#
   Becken+Noun+Common+Neut+Sg+NomAccDat
2. Stau+Noun+Common+Masc+Sg#
   Becken+Noun+Common+Neut+Pl+NomAccDatGen
3. Staub+Noun+Common+Masc+Sg#
   Ecke+Noun+Common+Fem+Pl+NomAccDatGen

Examples from German (2)

```
<form>hat</form> <ENGLISH>has</ENGLISH>
<lemma wkl=VER typ=AUX pers=3 num=SIN modtemp=PRÄ>haben</lemma>
<lemma wkl=VER pers=3 num=SIN modtemp=PRÄ konj=NON>haben</lemma>

<form>man</form> <ENGLISH>one</ENGLISH>
<lemma wkl=PRO typ=IND kas=NOM num=SIN gen=ALG stellung=STV>man</lemma>

<form>mir</form> <ENGLISH>me</ENGLISH>
<lemma wkl=PRO typ=REF kas=DAT num=SIN gen=ALG pers=1>sich</lemma>
<lemma wkl=PRO typ=PER kas=DAT num=SIN gen=ALG pers=1>ich</lemma>

<form>gesagt</form> <ENGLISH>told</ENGLISH>
<lemma wkl=VER form=PA2 konj=SFT>sagen</lemma>
<lemma wkl=PA2 gebrauch=PRD komp=GRU>gesagt</lemma>

<form>,</form>
<lemma wkl=SZK>,</lemma>

<form>ja</form> <ENGLISH>right</ENGLISH>
<lemma wkl=ADV typ=MOD>ja</lemma>
```

Stemmers

- Stemmers are the simplest type of morphological analyzer.
- One of the main advantages of stemmers is that they do not require a lexicon.
- The function of a stemmer is to remove the most common morphological and inflectional endings from words.
- Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

Finite-State Morphology

- Basic Idea: Encode morphological analysis and generation as composition of finite-state transducers.
- Resources needed:
  - Morpho-syntactic lexicon that specifies which combinations of free and bound morphemes are grammatical.
  - Context-sensitive replacement rules for spelling alternations.

2-level Rules: Restriction Operators

Two-level morphology employs a set of particular restriction operators:

=>	the correspondence only occurs in the environment
<=	the correspondence always occurs in the environment
<=>	the correspondence always and only occurs in the environment
/<=	the correspondence never occurs in the environment

2-level Rules: Restriction Operators

Two-level morphology employs a set of particular restriction operators:

- => the correspondence only occurs in the environment
- <= the correspondence always occurs in the environment
- <=> the correspondence always and only occurs in the environment
- /<= the correspondence never occurs in the environment

Idea: Rules with restriction operators function as constraints on the mapping between lexical and surface form of morphs.

Toy Rules for English (1)

i:y-spelling

```
die+ing tie+ing
dy00ing ty00ing
```

Rule: i:y <= _ e:? +:0 i

Elision

```
agree+ed dye+ed hoe+ed hoe+ing
agre00ed dy00ed ho00ed hoe0ing
```

Rule: e:0 <= C { V, y } _ +:? e:e

with V = { a e i o u } and

C = { b c d f g h j k l m n p q r s t v w x y z sh ch }

Toy Rules for English (2)

Epenthesis (simplified!; c.f. Trost, p. 41, (2.32))

```
fox+s kiss+s church+s spy+s
foxes kisses churches spies
```

Rule: +:e <=> { C_{sib}, y:i, o:o } _ s

with C_{sib} = { s x z sh ch }

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; shouting&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).

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: `www.racai.ro/∼tufis/`

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.

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.
- Only local phenomena can be described.

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) = (P(DET NN))/(P(NN)) = 0.420,
  - P(POS|NN) =(P(NN POS))/(P(POS)) = 0.507

Obtaining Probabilities

- Conditional probabilities for tag sequences and for word (given a tag) are computed from the frequency tables generated from training corpus.
- The size of the training corpus needed for good results is proportional to the size of the tagset.

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 $\longrightarrow$ 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,

  `www.coli.uni-sb.de/~thorsten/tnt/`
- Brill Tagger transformation-based error-driven,

  `www.cs.jhu.edu/~brill/`