Einführung in die Computerlinguistik
Part-of-Speech Tagging

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

2015-11-16

Overview

1. Motivation
2. Background
3. Probabilistic POS tagging

Outline

Part-of-speech tagging: Definition

- Part-of-speech tagging is the process of disambiguating the syntactic category of a word in context.
- Example: "book" is either a verb or a noun.
- In the context "the book" it can only be a noun.
- In the context "to book a flight" it can only be a verb.
- Part-of-speech tagging assigns to "book" the correct syntactic category in context.

Is part-of-speech tagging hard?

- The example of "book" in the phrase "the book" is easy.
- The rule "a word after 'the' cannot be a verb" takes care of it.
- Are all cases of part-of-speech tagging this easy? Example of an ambiguous context with two possible parts of speech?

Hard example

The	representative	put	chairs	on	the	table
AT article	NN noun	VBD verb-d	NNS noun-s	IN prep	AT article	NN noun
AT article	JJ adjective	NN noun	VBZ verb-z	IN prep	AT article	NN noun

In this case, finding the correct parts of speech for the sentence is more difficult. Exercise: Information available to pick correct tagging?

Questions

- Is this just a weird example or are part-of-speech ambiguities frequent?
- What's an example of a frequent English word that is not ambiguous with respect to syntactic category?
- Are part-of-speech ambiguities frequent in other languages?

Why part-of-speech tagging?

- Part-of-speech tagging is used as a preprocessing step.
- It is solvable: Very high accuracy rates can be achieved (sometimes 99%).
- It helps with many things you want to do with text, e.g., chunking, information extraction, question answering and parsing.

Part-of-speech tagging of tweets

ikr	smh	he	asked	fir	yo	last
!	G	O	V	P	D	A
name	so	he	can	add	u	on
N	P	O	V	V	O	P
fb	lololol					
∧	!					

Ta

is a preprocessing step for man NLP tasks.

Outline

Setup

- We will first look at the Brown corpus tag set.
- Early work on part-of-speech tagging was done on the Brown corpus.
- It's still an important corpus in NLP.

Creators of Brown corpus:
W. Nelson Francis & Henry Kučera (Brown University)

Brown corpus tags

Tag	Part Of Speech
AT	article
BEZ	the word “is”
IN	preposition
JJ	adjective
JJR	comparative adjective
MD	modal
NN	singular or mass noun
NNP	singular proper noun
NNS	plural noun
PERIOD	. : ? !
PN	personal pronoun

Tag	Part Of Speech
RB	adverb
RBR	comparative adverb
TO	the word “to”
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle, gerund
VBN	verb, past participle
VBP	verb, non-3rd person singular present
VBZ	verb, 3rd singular present
WDT	wh-determiner: “what”, “which”, . . .

Are these typical syntactic categories? Tag: “Peter arrived in London on Tuesday”

What information can we use for tagging?

- Let's look again at our example sentence:
  "The representative put chairs on the table."
- What information is available to disambiguate this sentence syntactically?

Hard example

The	representative	put	chairs	on	the	table
AT article	NN noun	VBD verb-d	NNS noun-s	IN prep	AT article	NN noun
AT article	JJ adjective	NN noun	VBZ verb-z	IN prep	AT article	NN noun

In this case, finding the correct parts of speech for the sentence is more difficult. Exercise: Information available to pick correct tagging?

Two main sources of information

1. The context of the ambiguous word:
   the words to the left and to the right
   - Example: for a JJ/NN ambiguity in the context "AT __ VBZ", NN is much more likely than JJ.
2. A word's bias for the different parts of speech
   - Example: "put" is much more likely to occur as a VBD than as an NN.

Information sources

- Information source 2: The frequency of the different parts of speech of the ambiguous word
- This source of information lets us do 90% correct tagging of English very easily: Just pick the most frequent tag for each word.
- For most words in English, the distribution of tags is very uneven: there is one very frequent tag and the others are rare.

Notation

w_i	the word at position i in the corpus
t_i	the tag of w_i
w^l	the l^{th} word in the lexicon
t^j	the j^{th} tag in the tag set
$C(w^l)$	the number of occurrences of w^l in the training set
$C(t^j)$	the number of occurrences of t^j in the training set
$C(t^j t^k)$	the number of occurrences of t^j followed by t^k
$C(w^l : t^j)$	the number of occurrences of w^l that are tagged as t^j

Notation: Example

the	representative	put	chairs	on	the	table
w_1	w_2	w_3	w_4	w_5	w_6	w_7
w^5	w^{81}	w^3	w^4	w^1	w^5	w^6
AT	NN	VBD	NNS	IN	AT	NN
article	noun	verb-d	noun-s	prep	article	noun
t_1	t_2	t_3	t_4	t_5	t_6	t_7
t^{16}	t^{12}	t^2	t^9	t^3	t^{16}	t^{12}

$$
\begin{array}{lcl|lcl}
C(w^5) & = & 2 & C(w^4) & = & 1 \\
C(t^{16}) & = & 2 & C(t^2) & = & 1 \\
C(t^{16}t^{12}) & = & 2 & C(t^{12}t^2) & = & 1 \\
C(t^{16}t^2) & = & 0 & C(w^5w^{81}) & = & 1 \\
C(w^5 : t^{16}) & = & 2 & C(w^5 : t^{12}) & = & 0
\end{array}
$$

Notation: Exercise

Confidence/NN in/IN the/AT pound/NN is/BEZ widely/RB expected/VBN to/TO take/VB another/AT sharp/JJ dive/NN if/IN trade/NN figures/NNS for/IN September/NNP ,/, due/JJ for/IN release/NN tomorrow/NN ,/, fail/VBP to/TO show/VB a/AT substantial/JJ improvement/NN from/IN July/NNP and/CC August/NNP 's/POS near-record/JJ deficits/NNS ./. Chancellor/NNP of/IN the/AT Exchequer/NNP Nigel/NNP Lawson/NNP 's/POS restated/VBN commitment/NN to/TO a/AT firm/JJ monetary/JJ policy/NN has/VBZ helped/VBN to/TO prevent/VB a/AT freefall/NN in/IN sterling/NN over/IN the/AT past/JJ week/NN ./. Give the values of the following: w_4,

t_5, $C(w_8)$, $C(t_9)$, $C(t_1 t_2)$, $C(w_3 : t_3)$

Supervised learning

- Labeled training set: each word is annotated (or marked or tagged) by a linguist, with correct part-of-speech
- Train a statistical model on the training set
  - Result: A set of parameters (= numbers) that were learned from the specific properties of the training set
- Apply statistical model to new text that we want to analyze for some task (information retrieval, machine translation etc)

Tagged training corpus/set: Example

Confidence/NN in/IN the/AT pound/NN is/BEZ widely/RB expected/VBN to/TO take/VB another/AT sharp/JJ dive/NN if/IN trade/NN figures/NNS for/IN September/NNP ,/, due/JJ for/IN release/NN tomorrow/NN ,/, fail/VBP to/TO show/VB a/AT substantial/JJ improvement/NN from/IN July/NNP and/CC August/NNP 's/POS near-record/JJ deficits/NNS ./. Chancellor/NNP of/IN the/AT Exchequer/NNP Nigel/NNP Lawson/NNP 's/POS restated/VBN commitment/NN to/TO a/AT firm/JJ monetary/JJ policy/NN has/VBZ helped/VBN to/TO prevent/VB a/AT freefall/NN in/IN sterling/NN over/IN the/AT past/JJ week/NN ./.

Outline

Contents of this section

- Parameter estimation: context parameters
- Parameter estimation: bias parameters
- Noisy channel model
- Greedy tagging
- Viterbi tagging
- Exam: estimation of context/bias parameters

Parameter estimation: Context

- The conditional probabilities $P(t^k|t^j)$ are the context parameters of the model.
- This will be our formalization of the first source of information in tagging: the context.
- Note that this is a very impoverished model of context.
  - Limited horizon, Markov assumption: we assume that our memory is limited to a single preceding tag.
  - Time invariance, stationary: we assume that these conditional probabilities don't change. (e.g., the same at the beginning and at the end of the sentence)

Parameter estimation: Context

- How can we estimate $P(t^k|t^j)$?
- For example: how can we estimate $P(\text{NN}|\text{JJ})$?
- First: maximum likelihood estimate
- Training text: long tagged sequence of words

Tagged training corpus/set: Example

Confidence/NN in/IN the/AT pound/NN is/BEZ widely/RB expected/VBN to/TO take/VB another/AT sharp/JJ dive/NN if/IN trade/NN figures/NNS for/IN September/NNP ,/, due/JJ for/IN release/NN tomorrow/NN ,/, fail/VBP to/TO show/VB a/AT substantial/JJ improvement/NN from/IN July/NNP and/CC August/NNP 's/POS near-record/JJ deficits/NNS ./. Chancellor/NNP of/IN the/AT Exchequer/NNP Nigel/NNP Lawson/NNP 's/POS restated/VBN commitment/NN to/TO a/AT firm/JJ monetary/JJ policy/NN has/VBZ helped/VBN to/TO prevent/VB a/AT freefall/NN in/IN sterling/NN over/IN the/AT past/JJ week/NN ./.

Parameter estimation: Context

- How can we estimate $P(t^k|t^j)$?
- For example: how can we estimate $P(\text{NN}|\text{JJ})$?
- $$\hat{P}_{ml}(t^k|t^j) = \frac{\hat{P}_{ml}(t^j t^k)}{\hat{P}_{ml}(t^j)} \approx \frac{\frac{C(t^j t^k)}{C(.)}}{\frac{C(t^j)}{C(.)}} = \frac{C(t^j t^k)}{C(t^j)}$$
- $$\hat{P}_{ml}(\text{NN}|\text{JJ}) = \frac{C(\text{JJ NN})}{C(\text{JJ})}$$

Parameter estimation: Context

$$\hat{P}_{ml}(t^k|t^j) = \frac{\hat{P}_{ml}(t^j t^k)}{\hat{P}_{ml}(t^j)} \approx \frac{\frac{C(t^j t^k)}{C(.)}}{\frac{C(t^j)}{C(.)}} = \frac{C(t^j t^k)}{C(t^j)}$$

$$\hat{P}_{laplace}(t^k|t^j) = \frac{C(t^j t^k) + 1}{C(t^j) + |T|}$$

Parameter estimation: Word bias

- What about the second source of information: frequency of different tags for a word?
- We need to estimate: $P(t_i|w_i)$
- Actually: $P(w_i|t_i)$
- Example: $P(\text{book}|\text{NN})$

Parameter estimation: Word bias

- How to estimate $P(\text{book}|\text{NN})$
- $$\hat{P}_{ml}(w^l|t^j) = \frac{\hat{P}_{ml}(w^l : t^j)}{\hat{P}_{ml}(t^j)} = \frac{\frac{C(w^l : t^j)}{C(.)}}{\frac{C(t^j)}{C(.)}} = \frac{C(w^l : t^j)}{C(t^j)}$$
- $$\hat{P}_{ml}(\text{book}|\text{NN}) = \frac{C(\text{book} : \text{NN})}{C(\text{NN})}$$

Parameter estimation: Word bias

$$\hat{P}_{ml}(w^l|t^j) = \frac{\hat{P}_{ml}(w^l : t^j)}{\hat{P}_{ml}(t^j)} = \frac{\frac{C(w^l:t^j)}{C(.)}}{\frac{C(t^j)}{C(.)}} = \frac{C(w^l : t^j)}{C(t^j)}$$

$$\hat{P}_{laplace}(w^l|t^j) = \frac{C(w^l : t^j) + 1}{C(t^j) + |V|}$$

Tagged training corpus/set: Example

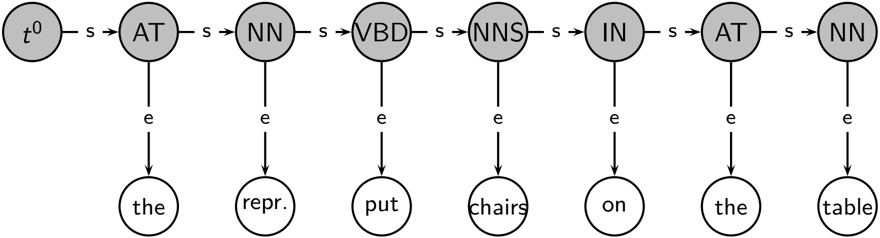
Confidence/NN in/IN the/AT pound/NN is/BEZ widely/RB expected/VBN to/TO take/VB another/AT sharp/JJ dive/NN if/IN trade/NN figures/NNS for/IN September/NNP ,/, due/JJ for/IN release/NN tomorrow/NN ,/, fail/VBP to/TO show/VB a/AT substantial/JJ improvement/NN from/IN July/NNP and/CC August/NNP 's/POS near-record/JJ deficits/NNS ./. Chancellor/NNP of/IN the/AT Exchequer/NNP Nigel/NNP Lawson/NNP 's/POS restated/VBN commitment/NN to/TO a/AT firm/JJ monetary/JJ policy/NN has/VBZ helped/VBN to/TO prevent/VB a/AT freefall/NN in/IN sterling/NN over/IN the/AT past/JJ week/NN ./. Estimate $P(\text{take}|\text{VB})$ and $P(\text{AT}|\text{IN})$

Parameter estimation: Word bias

- What about the second source of information: frequency of different tags for a word?
- We need to estimate: $P(t_i|w_i)$
- Actually: $P(w_i|t_i)$
- Example: $P(\text{book}|\text{NN})$

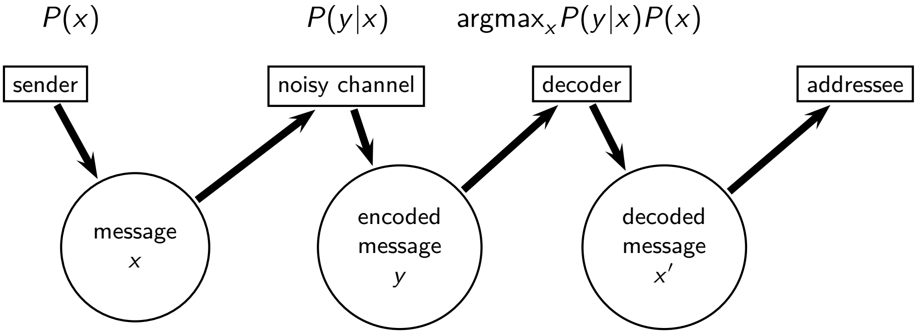
$P(w|t)$ versus $P(t|w)$

(s = sequence, e = emission)

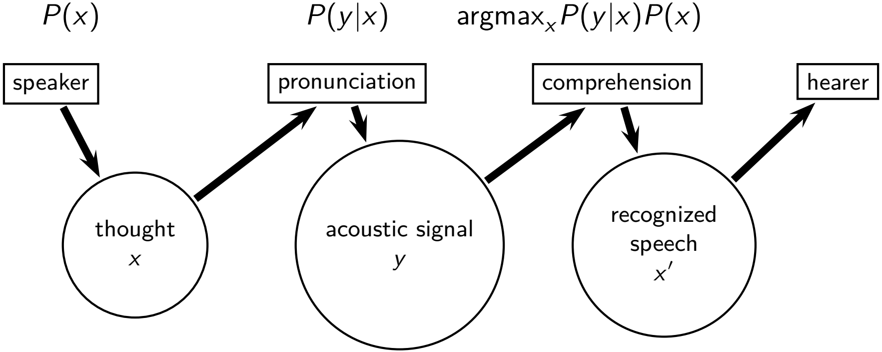


- The tags generate the words (not vice versa).
- Hence: The tags are given and the words are conditioned on the tags ...
- ... and the correct formalization is $P(w|t)$.

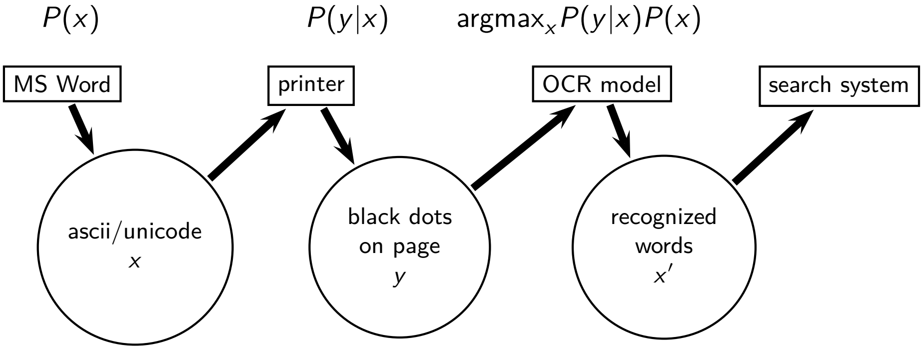
Noisy channel: Information theory / telecommunications



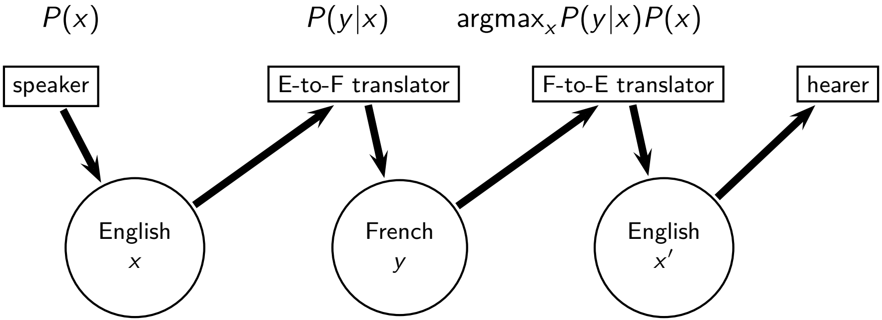
Noisy channel: Speech recognition



Noisy channel: Optical character recognition

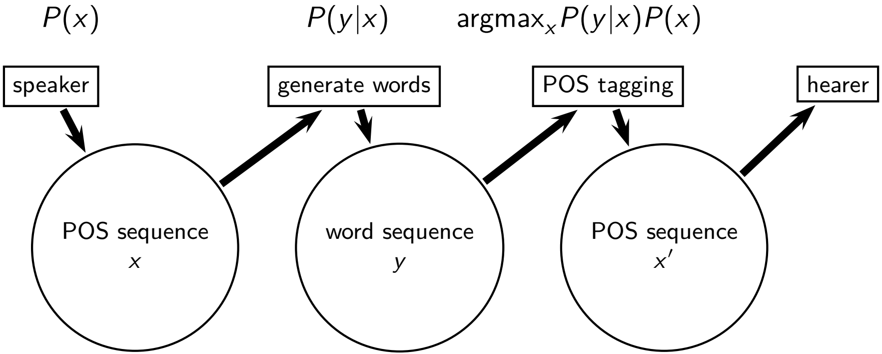


Noisy channel: French-to-English machine translation

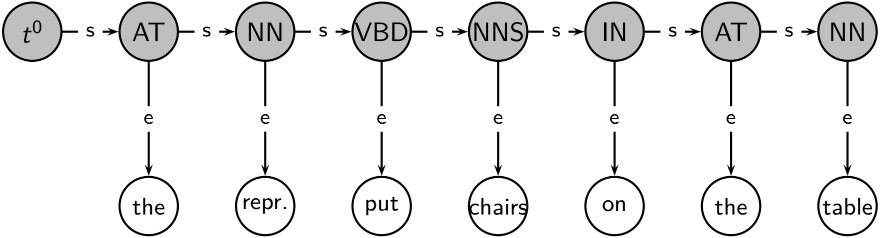


Noisy channel for part-of-speech tagging?

Noisy channel: Part-of-speech tagging



Noisy channel: Part-of-speech tagging



Exercise: How do we actually do the tagging?

- Context: $P(t_{i+1}|t_i)$
- Word bias: $P(w_i|t_i)$
- Given a sequence of words (a sentence), how do we compute the corresponding (disambiguated) part-of-speech sequence?
- Example:
  - Input: the representative put chairs on the table
  - Output: AT NN VBD NNS IN AT NN
- How can we do this?

"Greedy" tagging

- Suppose we've tagged a sentence up to position i.
- Then simply choose the tag t for the next word w_{i+1} that is most probable.
- At position i, choose tag that maximizes: $P(t_i|t_{i-1})P(w_i|t_i)$
- Let's do this for: "The representative put chairs on the table."
- $P(\text{VBP}|\text{NN})P(\text{put}|\text{VBP})$
- $t_3 = \text{VBP}$ maximizes $P(t_3|\text{NN})P(\text{put}|t_3)$

Problems with greedy tagging

- What can go wrong with greedy tagging?
- Example?
- A representative put costs 20% more today than a month ago.

Notation (2)

w_i	the word at position i in the corpus
t_i	the tag of w_i
$w_{i,i+m}$	the words occurring at positions i through $i+m$ (alternative notations: $w_i \cdots w_{i+m}$, $w_i, \ldots, w_{i+m}$, $w_{i(i+m)}$)
$t_{i,i+m}$	the tags $t_i \cdots t_{i+m}$ for $w_i \cdots w_{i+m}$
w^l	the l^{th} word in the lexicon
t^j	the j^{th} tag in the tag set
$C(w^l)$	the number of occurrences of w^l in the training set
$C(t^j)$	the number of occurrences of t^j in the training set
$C(t^j t^k)$	the number of occurrences of t^j followed by t^k
$C(w^l : t^j)$	the number of occurrences of w^l that are tagged as t^j
T	number of tags in tag set
W	number of words in the lexicon
n	sentence length

Part-of-speech tagging: Problem statement

- We define our goal thus: Given a sentence, find the most probable sequence of tags for this sentence.
- Formalization of this goal:

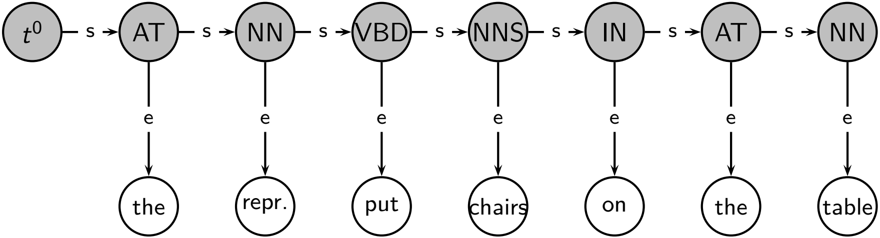
$$t_{1,n} = \arg\max_{t_{1,n}} P(t_{1,n}|w_{1,n})$$

Simplifying the argmax (1)

$$
\begin{aligned}
t_{1,n} &= \arg\max_{t_{1,n}} P(t_{1,n}|w_{1,n}) && (1)\\
&= \arg\max_{t_{1,n}} P(t_{0,n}|w_{1,n}) && (2)\\
&= \arg\max_{t_{1,n}} \frac{P(w_{1,n}|t_{0,n})P(t_{0,n})}{P(w_{1,n})} && (3)\\
&= \arg\max_{t_{1,n}} P(w_{1,n}|t_{0,n})P(t_{0,n}) && (4)\\
&= \arg\max_{t_{1,n}} [\prod_{i=1}^{n} P(w_i|t_{0,n})]P(t_{0,n}) && (5)
\end{aligned}
$$

$P(w|t)$ versus $P(t|w)$

(s = sequence, e = emission)



- The tags generate the words (not vice versa).
- Hence: The tags are given and the words are conditioned on the tags ...
- ... and the correct formalization is $P(w|t)$.

Simplifying the argmax (2)

$$
\begin{aligned}
&= \arg\max_{t_{1,n}} [\prod_{i=1}^{n} P(w_i|t_{0,n})] P(t_{0,n}) && (6)\\
&= \arg\max_{t_{1,n}} [\prod_{i=1}^{n} P(w_i|t_i)] P(t_{0,n}) && (7)\\
&= \arg\max_{t_{1,n}} [\prod_{i=1}^{n} P(w_i|t_i)][\prod_{i=1}^{n} P(t_i|t_{0,i-1})] && (8)\\
&= \arg\max_{t_{1,n}} [\prod_{i=1}^{n} P(w_i|t_i)][\prod_{i=1}^{n} P(t_i|t_{i-1})] && (9)\\
&= \arg\max_{t_{1,n}} \prod_{i=1}^{n} [P(w_i|t_i) P(t_i|t_{i-1})] && (10)
\end{aligned}
$$

Simplifying the argmax (3)

$$
\begin{aligned}
&= \arg\max_{t_{1,n}} \prod_{i=1}^{n} [P(w_i|t_i)P(t_i|t_{i-1})] && (11)\\
&= \arg\max_{t_{1,n}} \sum_{i=1}^{n} [\log P(w_i|t_i) + \log P(t_i|t_{i-1})] && (12)
\end{aligned}
$$

Do you recognize these parameters? What's the difficulty if you want to tag based on this?