Word Classes
&
Part-of-Speech Tagging

Martin Jansche
recycling Prof. Hirschberg's slides

CS 4705

2004-10-04

What is a word class?

- Words that somehow ‘behave’ alike:
  - Appear in similar contexts,
  - Perform similar functions in sentences,
  - Undergo similar morphological changes.
- Why do we want to identify them?
  - Pronunciation (poor person’s homograph disambiguation)
  - Stemming (poor person’s morphological analysis)
  - Semantics (poor persons’ word sense disambiguation)
  - Richer language models (back off to part of speech)
  - Parsing (almost-parsing, supertagging)

In a nutshell

A small number of categories from traditional grammar:

- Nouns (*agreement*, *humanism*)
- Verbs (*agree*, *certify*)
- Adjectives (*insidious*, *oily*)
- Adverbs (*highly*, *seldom*)
- Prepositions (*from*)
- Determiners (*the*, *every*)
- Conjunctions (*and*, *if*, *but*)
- Particles (*up*, *off* in certain contexts)

Nouns

- Appear in similar contexts:
  - after determiners (*the N*, *every N*, *no N*)
  - modified by adjectives (*an oily N*)
  - modified by relative clauses (*the N [I saw yesterday]*)
- Perform similar functions in sentences:
  - noun phrases appear as arguments of predicates
  - free nominal modifiers (*Friday*, *no way*)
- Undergo similar morphological changes:
  - plural inflection (*agreement-s*)
  - derivation (*fish-y*, *noon-ish*, *item-ize*, *mis-adventure*, *e-card*)

Verbs

- Appear in similar contexts:
  - *to V*
  - modified by adverbs, prepositional phrases (*to V warmly*, *to V in the shower*)
- Perform similar functions in sentences:
  - predicate of a clause
  - arguments in certain constructions (*heard him V*)
- Undergo similar morphological changes:
  - inflection for subject agreement (*agree-s*)
  - past tense and past participle (*agree-(e)d*)
  - derivation (*dis-agree*, *agree-ment*, *communicat(e)-ion*)

Adjectives

- Appear in similar contexts:
  - modifying nouns (*every Adj student*)
  - modified by certain degree adverbs (*a very Adj person*)
  - after a form of *be* (*this room is Adj*)
- Perform similar functions in sentences:
  - with *be* as a predicate
  - arguments in certain constructions (*consider him V*)
- Undergo similar morphological changes:
  - inflection for degree (*oil(y)i-er*, *oil(y)i-est*)
  - derivation (*un-clean*, *oil(y)i-ness*, *the poor/meek/oily*)

Adverbs

- Appear in similar contexts:
  - modifying verbs (*to Adv go where no-one has gone before*)
  - modifying adjectives (*a(n) Adv qualified applicant*)
  - modifying adverbs (*a(n) Adv highly qualified applicant*)
  - modifying determiners (*hardly any, almost all*)
  - modifying prepositions (*he drove Adv into a brick wall*)
- Perform similar functions in sentences:
  - predicate modifiers
- Undergo similar morphological changes:
  - (not very systematic)

Prepositions

- Appear in similar contexts:
  - before noun phrases (*drove Prep the ocean*)
- Perform similar functions in sentences (when combined with a noun phrase):
  - as arguments of verbs (*accuse somebody of something*, *charge somebody with something*)
  - as optional modifiers indicating time, location, manner (*eat lunch before/after/during the meeting*)
- Undergo similar morphological changes:
  - none (except for category blending, like *nearest*)

Determiners

- Appear in similar contexts:
  - before nouns (plus nominal modifiers) (*each/every/a/the/this/that/no honest businessman*, *few/most/all/the/these honest businessmen*)
- Perform similar functions in sentences:
  - form noun phrases
- Undergo similar morphological changes:
  - none

Conjunctions

- Appear in similar contexts:
  - before sentences (*Smith was worried Conj the government was out to get him*)
- Perform similar functions in sentences (Conjunction + sentence):
  - arguments in certain constructions
- Undergo similar morphological changes:
  - none

Particles

Catch-all category. Contains hard to classify items, e.g. in multi-word verb forms (*give up*, *cave in*), or fixed constructions (*the more you buy the more you save*).

But things are not as simple

- Pronouns and proper names occur in (approximately) the same contexts as noun phrases, hence need to be tagged like noun phrases.

- Most nouns usually require determiners, but some cannot (easily) be used with determiners (*garlic*).

- All words can be nouns when quoted – no *if*s, *and*s or *but*s.

- Verbs contain a closed subclass of so-called auxiliary verbs, which have idiosyncratic negations (*aren't*, *cannot*), irregular paradigms, and can invert with their subjects.

- Adverbs appear to be a heterogeneous category, since they can modify verbs, adjectives, determiners etc.

Sentence with part-of-speech tags

From the Brown corpus (Francis & Kucera, 1964–1979):

The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place.

With tags (slightly simplified):

The/AT Fulton/NP County/NP Grand/NP Jury/NP said/VBD Friday/NP an/AT investigation/NN of/IN Atlanta/NP s/PPL recent/JJ primary/JJ election/NN produced/VBD no/AT evidence/NN that/CS any/DTI irregularities/NNS took/VBD place/NN ./.

What gets tagged?

- White space delimited strings?
  Need to decide how to tag *grand* in *grand jury*.

- More abstract tokens?
  Need to decide how to tag *'s* in *Atlanta 's*.

- Do not underestimate tokenization issues.

What tags get assigned?

- Tags should characterize the local syntactic function of a word in its context.
- The Brown Corpus has 80+ tags
- The Penn Treebank (PTB) has 40+ tags
- Differences in tag inventories:
  - granularity
  - treatment of special words (*to*, *no*, *there*, . . .)
  - presence or absence of information about internal structure of a word

Don’t mix categories

According to the tagset of the Penn Treebank, words fall into the following classes:

- nouns, verbs, etc.,
- those that are *to*,
- foreign words.

For example, *perestroika* is tagged as a foreign word, though it patterns with the proper nouns; *laissez-fair* is tagged as a foreign word, but it behaves like an adjective.

The problem is that part-of-speech and foreignness are orthogonal dimensions.

Make useful distinctions

The Brown Corpus tagset distinguishes the verbs *be*, *have*, *do* in addition to the closed class of auxiliaries and the open class of non-auxiliary verbs. These distinctions are not useful, since they can easily be recovered.

The Penn Treebank always tags *to* as TO, which doesn't reveal anything about whether it is a preposition, infinitival marker, etc.

Try to be consistent

The Penn Treebank contains residual variation that does not seem to be due to natural variation in the data.

The Wall Street Journal portion of the PTB is divided into 25 sections (00–24). We can trace the relative frequency of two complementary parts of speech of the same word across the 25 sections.

before

preposition (IN): 89.67%

adverb (RB): 10.18%

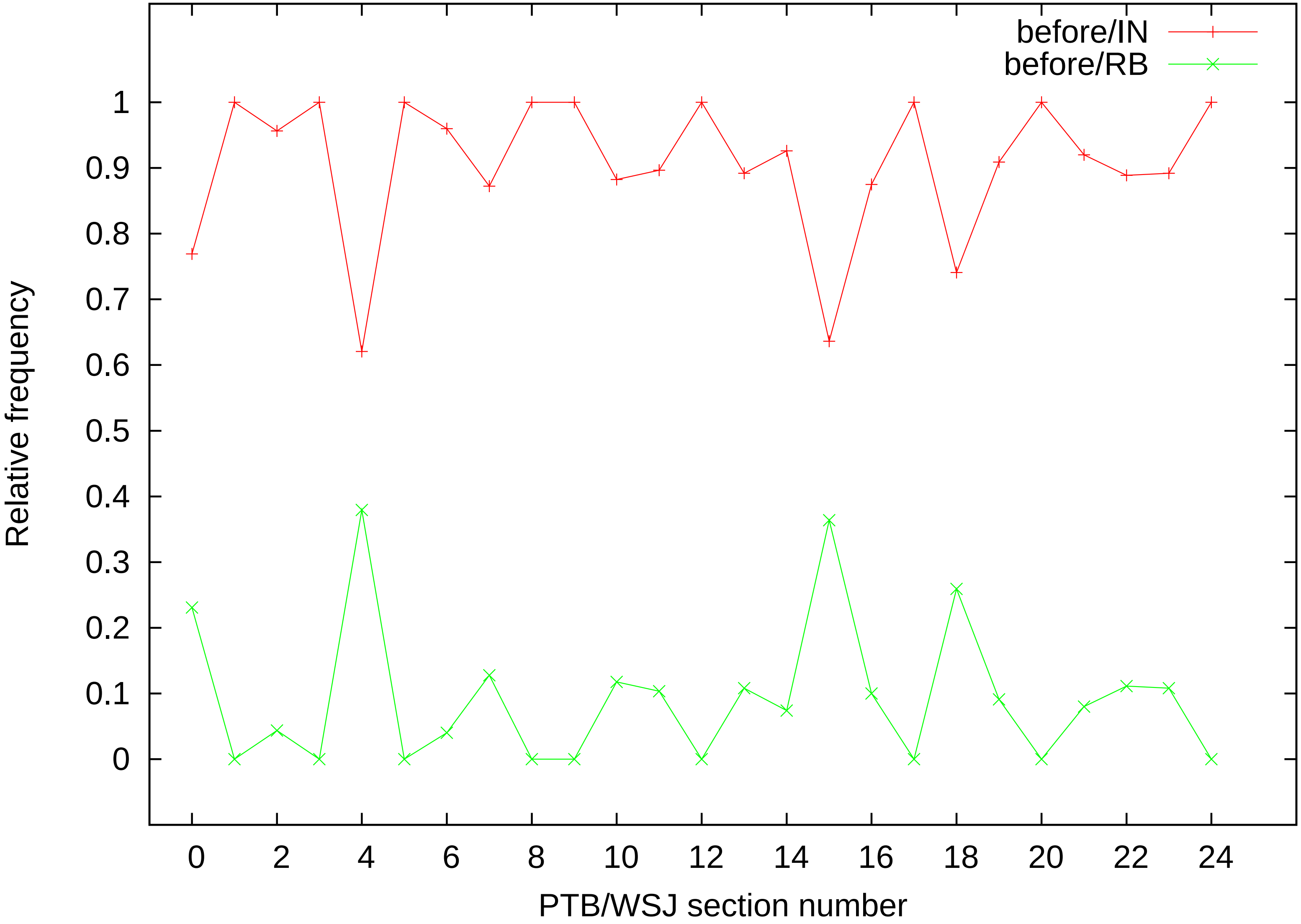
In a riveting day of hearings before the House Banking Committee, [...]

She had seen cheating before, but these notes were uncanny.

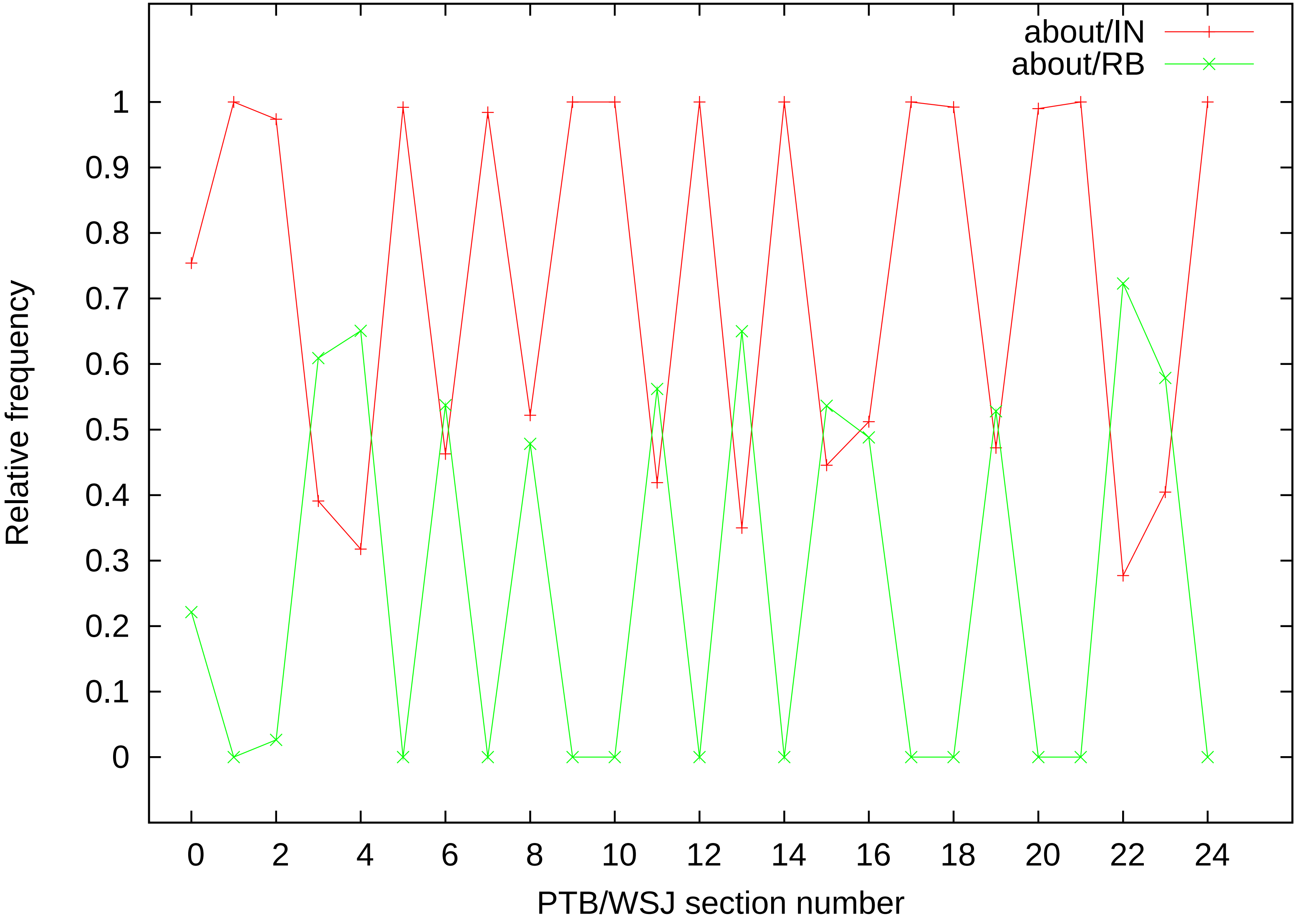
The department said orders for nondurable goods [...] fell 0.3% in September [...] after climbing 0.9% the month before.

[...] so it was a week before three of the four crates could be located in a bonded warehouse [...]

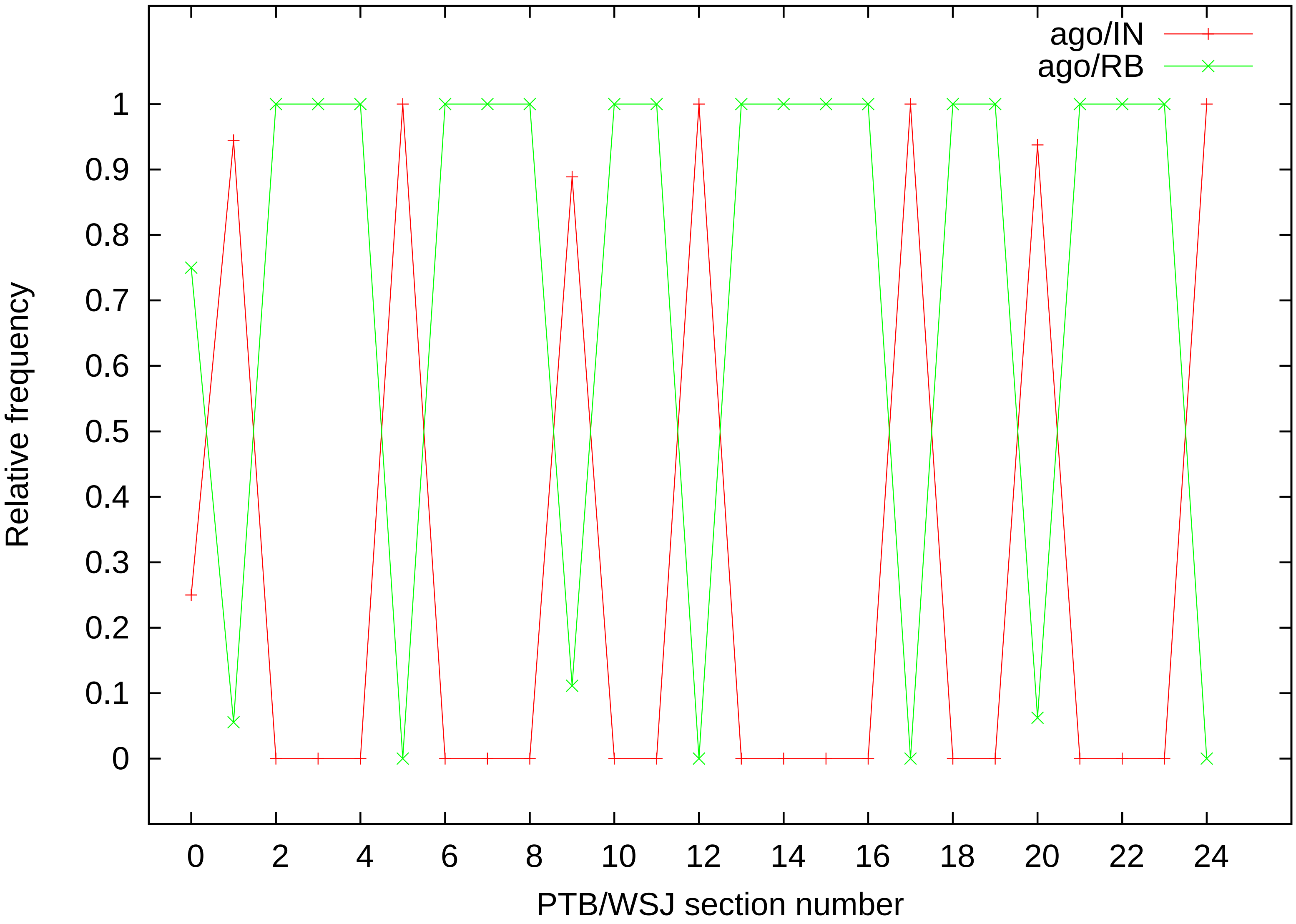
before


before/IN
before/RB
Relative frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
16
18
20
22
24
PTB/WSJ section number

about


about/IN
about/RB
Relative frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14
16
18
20
22
24
PTB/WSJ section number

ago


ago/IN
ago/RB
Relative frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14
16
18
20
22
24
PTB/WSJ section number

Automatic part-of-speech tagging

Given a sentence, find its (top n) "best" part-of-speech assignment(s).

She knows you like the back of her hand.

British Left Waffles on Falklands

Foot Heads Arms Body

Potential sources of disambiguation

- Many words have only one POS tag (*agreement*, *humanism*).

- Others have a single most likely tag that is far more probable than other tags (*dog*, *very*, *well*, *see*).

- Tags tend to co-occur regularly with certain tags (Det N, Det Adj), or avoid certain other tags (Det V, Det Prep).

Approaches to POS tagging

- Hand-crafted rules
- Largely automatically constructed systems:
  - Stochastic approaches
  - Other machine learning approaches

The Brill tagger

Tagline: transformation-based error-driven learning.
Really: greedy empirical risk minimization.

- Tag each word initially with most likely POS (determined by counting which tags the word appears with in the annotated training data).
- Examine a fixed set of possible transformations of tag assignments.
- Find the transformation that yields the biggest reduction of training error.
- Apply this transformation and repeat from step 2 until no further improvements are possible.

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS
Stage 0:	JJ	VBD	NNS	IN	NPS

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS
Stage 0:	JJ	VBD	NNS	IN	NPS

change VBD to NN when preceded by JJ

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS
Stage 0:	**JJ**	**VBD**	NNS	IN	NPS
change VBD to NN when preceded by JJ					
Stage 1:	JJ	**NN**	NNS	IN	NPS

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS
Stage 0:	JJ	VBD	NNS	IN	NPS

change VBD to NN when preceded by JJ

Stage 1:	JJ	NN	NNS	IN	NPS

change NNS to VBZ when preceded by NN

Example run of the Brill tagger

Input:	British	Left	Waffles	on	Falklands
Annotation:	JJ	NN	VBZ	IN	NPS
Stage 0:	JJ	VBD	NNS	IN	NPS
change VBD to NN when preceded by JJ					
Stage 1:	JJ	**NN**	**NNS**	IN	NPS
change NNS to VBZ when preceded by NN					
Stage 2:	JJ	NN	**VBZ**	IN	NPS

Learned [(JJ,VBD)→(JJ,NN), (NN,NNS)→(NN,VBZ)].

Implementing the Brill tagger

- Naive implementation can easily be done as a homework project.

- Two (solved) problems:
  - During learning, want to find the best transformation quickly.
  - When using the learned classifier, do we need to generate all intermediate tag assignments? (No!)

Stochastic n-gram taggers

We assume we know the conditional probability $\Pr(t \mid w)$ of a tag sequence t given a word sequence w. The best tag assignment for a word sequence w can be chosen as

$$\hat{t} = \arg\max_{t} \Pr(t \mid w)$$

By Bayes' theorem:

$$\Pr(t \mid w) = \frac{\Pr(w \mid t)\ \Pr(t)}{\Pr(w)}$$

Since w is given, $\Pr(w)$ is constant and can be ignored. Therefore:

$$\hat{t} = \arg\max_{t} \Pr(w \mid t)\ \Pr(t)$$

The source model

In general, if $t = (t_1, \ldots, t_n)$, then

$$
\begin{aligned}
\Pr(t) &= \Pr(t_1, \ldots, t_n) \\
&= \Pr(t_1)\ \Pr(t_2, \ldots, t_n \mid t_1) \\
&= \Pr(t_1)\ \Pr(t_2 \mid t_1)\ \Pr(t_3, \ldots, t_n \mid t_1, t_2) \\
&\ \ \vdots \\
&= \prod_{i=1}^{n} \Pr(t_i \mid t_1, \ldots, t_{i-1})
\end{aligned}
$$

Markov assumption: instead of conditioning on the full history $t_1, \ldots, t_{i-1}$ limit this to the k final terms $t_{i-k}, \ldots, t_{i-1}$.

A bigram source model

Approximate $\Pr(t)$ by a bigram (first order) model on tags.

$$\Pr(t) = \prod_{i=1}^{n} \Pr(t_i \mid t_{i-1})$$

Use standard techniques (lower-order interpolation, back-off) for estimating nonparametric bigram models from data. Not very challenging if tag inventory is small: for the PTB only around 2000 parameters.

The channel model

Let $t = (t_1, \ldots, t_n)$ as before and $w = (w_1, \ldots, w_n)$. Then:

$$\begin{aligned} \Pr(w \mid t) &= \Pr(w_1, \ldots, w_n \mid t_1, \ldots, t_n) \\ &= \prod_{i=1}^{n} \Pr(w_i \mid w_1, \ldots, w_{i-1}, t_1, \ldots, t_n) \end{aligned}$$

Assume that the probability of w_i depends only on t_i. This gives us a simple zeroth-order conditional model

$$\Pr(w \mid t) = \prod_{i=1}^{n} \Pr(w_i \mid t_i)$$

Nonparametric models of this form are easy to estimate from data if the set of words is assumed to be known.

Decoding with the overall model

$$\begin{aligned}
\hat{t} &= \arg\max_t \ \Pr(w \mid t) \ \Pr(t) \\
&= \arg\max_t \ \left(\prod_{i=1}^{n} \Pr(w_i \mid t_i)\right) \ \left(\prod_{i=1}^{n} \Pr(t_i \mid t_{i-1})\right) \\
&= \arg\max_t \ \log \left(\prod_{i=1}^{n} \Pr(w_i \mid t_i) \Pr(t_i \mid t_{i-1})\right) \\
&= \arg\max_t \ \sum_{i=1}^{n} \log \Pr(w_i \mid t_i) + \log \Pr(t_i \mid t_{i-1})
\end{aligned}$$

Use the Viterbi algorithm to find the maximizing t.

The Viterbi algorithm

A special case of a single-source algebraic path computation over weighted directed acyclic graphs (dawgs). Closely related problems: finding the shortest path in a dawg from a fixed source vertex; counting the number of paths through a da(w)g from a fixed source vertex; computing string edit distance.

Need the blackboard for this.

Extensions

- Don't have enough/any manually tagged data? Use partially supervised or unsupervised learning. Straightforward stochastic approach, but more algorithmically involved. Brill's approach can also be adapted.

- Have more than one annotation? Can't decide which annotation is correct for *child seat* or *Caribbean cooking*? No problem for stochastic approaches: treat the manually provided tag sequences as a probability distribution, proceed with partially supervised training.

- Need to reserve some probability mass for unseen words in the channel model. Fall back on internal morphology, suffixes, word length, or any other overt feature of the word.

Evaluation

- For any empirical AI task, we need to address how to evaluate our solutions. For POS tagging, compare predicted POS sequence against a known Gold Standard and count discrepancies (Hamming distance).
- Need human annotators to generate Gold Standard. Not all is gold, however – we saw evidence of discrepancies among human annotators in the PTB. These can be quantified with the kappa statistic.
- Need a performance baseline, typically provided by the simplest functional system one can think of. In this case, assigning the most likely tag to each word achieves above 90% accuracy.

Summary

- Part-of-speech tagging is a conceptually simple NLP problem. Many of the techniques discussed here carry over to slightly different tasks (named entity extraction, base-NP chunking) or can be refined for more complex tasks.
- Simple techniques come close to human performance.
- Many open issues remain, including unseen words and use of context beyond n-grams.
- Next Class: Guest Lecture by Owen Rambow
  - Read Chapter 9
  - Homework 1 due