Tagging

Motivation and Background

Outline

Next Word Prediction

The Probability of a Sequence

Training and Testing

Parts of Speech

English Parts of Speech

Tagging

Word Prediction

NY Times exercise shows that we seem to have the ability to predict future words in an utterance.

How?

- domain knowledge
- syntactic knowledge
- lexical knowledge

Claim

A useful part of the knowledge needed to allow Word Prediction (guessing the next word) can be captured using simple statistical techniques.

In particular, we'll rely on the notion of the *probability* of a sequence (e.g., sentence).

Why do this?

Why would anyone want to predict a word?

You don't really. But if you say you can predict the next word, it means you can rank the likelihood of sequences containing various alternative words.

- you can assess the likelihood/goodness of a sentence

Statistical Methods Abound in NLP

Resolving part of speech ambiguity

- *fly* can be a noun or a verb
- find the most likely word class, given the surrounding words

Resolving word sense ambiguity

- the noun *bank* can be the side of a river or a financial instituion
- again, find the most likely word class, given the surrounding words

Speech recognition

- find the most likely word sequence given a signal

Corpus Analysis

Statistical methods use large corpora (or databases) of natural language, which have been marked up (or "annotated") with phenemena of interest (e.g., parts of speech, word sense, parse trees).

- annotation must frequently be done by humans

These corpora are then used to find statistics.

- e.g., in 95% of cases, if *fly* follows an article (e.g., *the*, *a*), it is a noun

These statistics are then used to analyze new (unannotated) sentences.

- in the sentence *The fly flies*, the most likely part of speech tagging is *art n v*

Essentially we use probability theory to find the likelihood of one interpretation over another, and hence the maximally likely interpretation.

Handwriting Recognition Example

Assume a note is given to a bank teller, which the teller reads as *I have a gub.* (example courtesy of a Woody Allen film)

NLP to the rescue ...

- *gub* is not a word
- - *gun* and *gull* are words, but *gun* has a higher probability in the context of a bank

Real Word Spelling Errors

They are leaving in about fifteen *minuets* to go to her house.
The study was conducted mainly *be* John Black.
The design *an* construction of the system will take more than a year.
Hopefully, all *with* continue smoothly in my absence.
Can they *lave* him my messages?
I need to *notified* the bank of [this problem.]
He is trying to *fine* out.

Real Word Spelling Correction Ex.

Collect a list of commonly substituted words

- piece/peace, whether/weather, their/there ...

Whenever you encounter one of these words in a sentence, construct the alternative sentence as well

Assess the goodness of each and choose the one (word) with the more likely sentence

Example

- blah blah blah the whether
- blah blah blah the weather

Counting Words in Corpora

Probabilities are based on counting things, so ...

- what to count?
  - words (this chapter), word classes, word senses, speech acts ...
  - what is a word? (e.g., are *cat* and *cats* the same word?)
- where to find the things to count?
  - Corpora (online collections of text and speech)

Simple N-Grams

An N-gram model uses the previous N-1 words to predict the next one:

- $P(w_n \mid w_{n-1})$

Approximating Shakespeare

As we increase the value of N, the accuracy of the n-gram model increases.

Unigrams

- *Every enter now severally so, let*
- *Hill he late speaks; or! a more to leg less first you enter*

Bigrams

- *What means, sir. I confess she? then all sorts, he is trim, captain.*
- *Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry.*

Approximating Shakespeare (cont.)

Trigrams

- *Sweet prince, Falstaff shall die.*
- *This shall forbid it should be branded, if renown made it empty.*

Quadrigrams

- *What! I will go seek the traitor Gloucester.*
- *Will you not tell me who I am?*

Approximating Shakespeare (cont.)

There are 884,647 tokens, with 29,066 word form types, in about a one million word Shakespeare corpus.

Shakespeare produced 300,000 bigram types out of 844 million possible bigrams. Thus, 99.96 % of the possible bigrams were never seen (have zero entries in the table).

The quadrigrams are worse. What's coming out looks like Shakespeare because it is Shakespeare.

All those zeroes are causing problems.

Some Useful Empirical Observations

A small number of events occur with high frequency.

A large number of events occur with low frequency.

You can quickly collect statistics on the high frequency events.

You might have to wait an arbitrarily long time to get valid statistics on low frequency events.

Some of the zeroes in the table are really zeroes. Some are simply low frequency events you haven't seen yet.

Another Example

`% deroff -w /usr/man/man1/csh.1 | sort |`
`uniq -c | sort -r`

yields got 1626 unique types out of about 11,000 tokens, with the top seven in terms of frequency being:

- 709 the
- 304 is
- 275 to
- 250 of
- 192 and
- 188 command
- 184 csh

But 740 of the 1626 only occur once!

Example (continued)

Even if we add in more and more man pages...

- very few/no additional new high frequency types. But lots of additions to the counts for the existing high frequency types.
- more and more new single instance types

N-Gram Training Sensitivity

If we repeated the Shakespeare experiment but trained on a Wall Street Journal corpus, there would be little overlap in the output.

- implications for corpus design

Methodology: Training and Testing

Probabilities come from a <u>Training</u> portion of a corpus, which is used to design the model.

- overly narrow corpus: probabilities don't generalize
- overly general corpus: probabilities don't reflect task or domain

A separate <u>Test</u> portion of the corpus is used to *evaluate* the model, typically using standard *metrics*.

- held out test set
- cross validation
- evaluation differences should be statistically significant

Review

N-grams trained from some corpus

Evaluation issues

However, N-gram language models are an extremely impoverished attempt to capture what we know about likely sequences of words.

Clearly, syntactic, semantic, and discourse expectations play a role.

New Topic: Syntax

Up until now we've been dealing with simple-minded (though useful) notions of what sequences of words are likely.

Now we will turn to the study of how words

- are clustered into classes
- group with their neighbors to form phrases and sentences
- depend on other words

We'll start with syntactic word classes.

- Parts of Speech
- English Parts of Speech
- Tagging

Parts of Speech

Most tagsets implicitly encode fine-grained specializations of eight basic parts of speech (POS, word classes, morphological classes, lexical tags).

- noun, verb, pronoun, preposition, adjective, conjunction, article, adverb

These categories are based on *morphological* and *distributional* similarities and not, as you might think, semantics.

In some cases, this is straightforward (at least in a given language), in other cases it's not.

The Distribution of Tags

Tags follow all the usual frequency-based distributional behavior.

- most word types have only one part of speech
- of the rest, most have two, etc.
- the most frequently occurring word types tend to have multiple tags (and as we'll see later, they also tend to have more meanings)

Therefore, while it's easy to determine the correct tag for most wordtypes, it isn't necessarily so easy to tag most texts.

Notes on Tagsets

There are various tagsets for formally coding POS.

The choice of tagset pretty much depends on the nature of the application.

Since accurate tagging can be performed with relatively large tagsets it makes sense to use one of the larger standard sets. If it makes distinctions you don't need, you can merge the finer grained tags.

Why Tag POS?

Language Modeling

Pronunciation

Stemming

Parsing

Word Sense Disambiguation

Information Extraction

English Word Classes

Open (lexical) class types

- new words are coined/borrowed
- nouns, verbs, adjectives, adverbs

Closed (functional) class types

- relatively fixed membership
- small class, but frequently occur
- pronouns, prepositions, conjunctions, articles . . .

Nouns

Take possessives, occur in plural form, occur with determiners . . .

Proper nouns (*Diane*)

Common nouns

- count nouns (*professor, student, computer*)
- mass nouns (*snow, air*)

Vary in

- number (singular, plural)
- gender (masculine, feminine, neuter)
- etc.

Verbs

Refer to actions, activities, processes, states (*throw, walk, have*)

Tenses: present, past, future, . . .

Other inflection: number, person

Voice: active, passive

Standard morphological forms, as previously discussed.

- stem or non-3rd-person-sg (*eat*)
- -s form or 3rd-person-sg (*eats*)
- -ing participle or progressive (*eating*)
- past participle (*eaten*)

Irregular verbs

Auxiliary subclass of verbs are closed class, however.

Adjectives and Adverbs

Adjectives (semantically) describe properties or qualities.

- color (*blue*)
- age (*old*)
- value (*good*)

Adverbs also modify something, often verbs but also other adverbs and verb phrases.

- directional or locative (*downhill*)
- degree (*extremely*)
- manner (*slowly*)
- temporal (*yesterday*)

Pronouns

Roughly, a shorthand for referring to a noun phrase or entity.

- personal (*you, I, me*)
- possessive (*my, yours, mine, ours*)
- wh- (*what, who, whom, whoever, where, when*)

Vary in

- person
- gender
- number
- case (in English, nominative, accusative, possessive, 2nd possessive, reflexive)

Pronouns (continued)

English pronouns from CELEX on-line dictionary, with frequency counts from a 16 million word corpus.

it	199,920	how	13,137	yourself	2,437	no one	106
I	198,139	another	12,551	why	2,220	wherein	58
he	158,366	where	11,857	little	2,089	double	39
you	128,688	same	11,841	none	1,992	thine	30
his	99,820	something	11,754	nobody	1,684	summat	22
they	88,416	each	11,320	further	1,666	suchlike	18
this	84,927	both	10,930	everybody	1,474	fewest	15
that	82,603	last	10,816	ourselves	1,428	thyself	14
she	73,966	every	9,788	mine	1,426	whomever	11
her	69,004	himself	9,113	somebody	1,322	whosoever	10
we	64,846	nothing	9,026	former	1,177	whomsoever	8
all	61,767	when	8,336	past	984	wherefore	6
which	61,399	one	7,423	plenty	940	whereat	5
their	51,922	much	7,237	either	848	whatsoever	4
what	50,116	anything	6,937	yours	826	whereon	2
my	46,791	next	6,047	neither	618	whoso	2
him	45,024	themselves	5,990	fewer	536	aught	1
me	43,071	most	5,115	hers	482	howsoever	1
who	42,881	itself	5,032	ours	458	thrice	1
them	42,099	myself	4,819	whoever	391	wheresoever	1
no	33,458	everything	4,662	least	386	you-all	1
some	32,863	several	4,306	twice	382	additional	0
other	29,391	less	4,278	theirs	303	anybody	0
your	28,923	herself	4,016	wherever	289	each other	0
its	27,783	whose	4,005	oneself	239	once	0
our	23,029	someone	3,755	thou	229	one another	0
these	22,697	certain	3,345	'un	227	overmuch	0
any	22,666	anyone	3,318	ye	192	such and such	0
more	21,873	whom	3,229	thy	191	whate'er	0
many	17,343	enough	3,197	whereby	176	whenever	0
such	16,880	half	3,065	thee	166	whereof	0

Prepositions

Prepositions occur before noun phrases.

Semantically, they are relational.

- spatial (*on*)
- temporal (*after*)
- ...

of	540,085	through	14,964	worth	1,563	pace	12
in	331,235	after	13,670	toward	1,390	nigh	9
for	142,421	between	13,275	plus	750	re	4
to	125,691	under	9,525	till	686	mid	3
with	124,965	per	6,515	amongst	525	o'er	2
on	109,129	among	5,090	via	351	but	0
at	100,169	within	5,030	amid	222	ere	0
by	77,794	towards	4,700	underneath	164	less	0
from	74,843	above	3,056	versus	113	midst	0
about	38,428	near	2,026	amidst	67	o'	0
than	20,210	off	1,695	sans	20	thru	0
over	18,071	past	1,575	circa	14	vice	0

Conjunctions

Conjunctions join two things.

- coordinating (*and, or, but*): things are of equal stature
- subordinating (*that, if, because, although*): one element has an embedded status
- ...

and	514,946	yet	5,040	considering	174	forasmuch as	0
that	134,773	since	4,843	lest	131	however	0
but	96,889	where	3,952	albeit	104	immediately	0
or	76,563	nor	3,078	providing	96	in as far as	0
as	54,608	once	2,826	whereupon	85	in so far as	0
if	53,917	unless	2,205	seeing	63	inasmuch as	0
when	37,975	why	1,333	directly	26	insomuch as	0
because	23,626	now	1,290	ere	12	insomuch that	0
so	12,933	neither	1,120	notwithstanding	3	like	0
before	10,720	whenever	913	according as	0	neither nor	0
though	10,329	whereas	867	as if	0	now that	0
than	9,511	except	864	as long as	0	only	0
while	8,144	till	686	as though	0	provided that	0
after	7,042	provided	594	both and	0	providing that	0
whether	5,978	whilst	351	but that	0	seeing as	0
for	5,935	suppose	281	but then	0	seeing as how	0
although	5,424	cos	188	but then again	0	seeing that	0
until	5,072	supposing	185	either or	0	without	0

Articles

Articles (determiners) begin noun phrases, and thus help identify them.

- articles: *the* (definite), *a, an* (indefinite)
- demonstratives: *this, that*

A particularly small class, but also very frequent.

- *the* 1,071,676
- *a* 413,887
- *an* 59,359
- sometimes others

Other Parts of Speech

Prepositions vs. Particles

- *She ran up a hill/bill.*

Phrasal Verbs

- *The plane took off./Take it off.*

Interjections

- *Ouch!*

Penn Treebank Tagset Formalization

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	“to”	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential ‘there’	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*’s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one’s*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

How are nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and articles formalized?

How would you tag the following sentence from the Brown Corpus?

- The grand jury commented on a number of other topics.

Penn Treebank (continued)

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Tagged sentence from Brown Corpus

- The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

UCREL C5 Tagset Fragment

Tag	Description	Example
PNX	reflexive pronoun	*itself, ourselves*
POS	possessive *'s* or *'*	
PRF	the preposition *of*	
PRP	preposition (except *of*)	*for, above, to*
PUL	punctuation – left bracket	(or [
PUN	punctuation – general mark	. ! , : ; - ? ...
PUQ	punctuation – quotation mark	' ' "
PUR	punctuation – right bracket	) or]
TO0	infinitive marker *to*	
UNC	unclassified items (not English)	
VBB	base forms of *be* (except infinitive)	*am, are*
VBD	past form of *be*	*was, were*
VBG	-ing form of *be*	*being*
VBI	infinitive of *be*	
VBN	past participle of *be*	*been*
VBZ	-s form of *be*	*is, 's*
VDB	base form of *do*(except infinitive)	*does*
VDD	past form of *do*	*did*
VDG	-ing form of *do*	*doing*
VDI	infinitive of *do*	*to do*
VDN	past participle of *do*	*done*
VDZ	-s form of *do*	*does*
VHB	base form of *have* (except infinitive)	*have*
VHD	past tense form of *have*	*had, 'd*
VHG	-ing form of *have*	*having*
VHI	infinitive of *have*	
VHN	past participle of *have*	*had*
VHZ	-s form of *have*	*has, 's*
VM0	modal auxiliary verb	*can, could, will, 'll*
VVB	base form of lexical verb (except infin.)	*take, live*
VVD	past tense form of lexical verb	*took, lived*
VVG	-ing form of lexical verb	*taking, living*
VVI	infinitive of lexical verb	*take, live*
VVN	past participle form of lex. verb	*taken, lived*
VVZ	-s form of lexical verb	*takes, lives*
XX0	the negative *not* or *n't*	
ZZ0	alphabetical symbol	*A, B, c, d*

Part of Speech Tagging

Part of speech tagging is simply assigning the correct part of speech for each word in a corpus.

Input

- a set of tags (a tagset)
- a dictionary that tells you the possible tags for each word (including all morphological variants)
- a text to be tagged (string of words)

Output

- single best tag for each word
- e.g., Book/VB that/DT flight/NN

Why is Tagging Hard?

Example

- Book/VB that/DT flight/NN
- Does/VBZ that/DT flight/NN serve/VB dinner/NN

Tagging is a type of disambiguation

- *book* can be NN or VB
  - *Can a read a book on this flight?*
- *that* can be DT or complementizer
  - *My travel agent said that there would be a meal on this flight.*

How Hard is the Tagging Problem?

11.5% of English words in the Brown corpus are ambiguous

But 40% of tokens in the Brown corpus are ambiguous

Unambiguous (1 tag)	**35,340**	
Ambiguous (2–7 tags)	**4,100**	
2 tags	3,760	
3 tags	264	
4 tags	61	
5 tags	12	
6 tags	2	
7 tags	1	(“still”)

The Brown Corpus

The Brown Corpus of Standard American English was the first of the modern, computer readable general corpora. Compiled by W. N. Francis and H. Kucera, Brown University, Providence, RI.

Corpus consists of one million words of American English texts printed in 1961.

For a long time, Brown and LOB (British) corpora were the only easily available online, so many studies have been done on these corpora.

Studying the same data allows comparison of findings without having to take into consideration possible variation caused by the use of different data.

But . . . ?

There is also a tagged version of the Brown Corpus

`http://www.hit.uib.no/icame/brown/bcm.html`

The Brown Corpus (cont.)

The corpus consists of 500 texts, each consisting of just over 2000 words, compiled from 15 different text categories

1. press (reportage) (44 texts)
2. press (editorial) (27 texts)
3. press (reviews) (17 texts)
4. religion (17 texts)
5. skills and hobbies (36 texts)
6. popular lore (48 texts)
7. belles-lettres (75 texts)
8. miscellaneous: govt. and house organs (30 texts)
9. learned (80 texts)
10. fiction (general) (29 texts)
11. fiction (mystery) (24 texts)

12. fiction (science) (6 texts)
13. fiction (adventure) (29 texts)
14. fiction (romance) (29 texts)
15. humor (9 texts)

Evaluation

How do you know how well you've done?

- use a tagged test corpus

How do you know if you're doing well?

Doing ok?

Doing great?

Making stupid claims?

A Common Evaluation Metric

Percent Correct

- the percentage of all tags in the test set where the tagger and a human labeled "gold standard" agree
- typically 96-97% for POS tagging

Evaluation Issues

Lower Bound/Baseline

- your goodness metric has to take into account some baseline that any dumb approach could achieve
- in the case of POS tagging, this is the most frequent tag heuristic (unigrams)
- can get you 90% for POS!
- thus, a result of 99% for POS tagging is less impressive than for say speech act tagging, where the baseline is much lower

Upper Bound/Ceiling

- this is all dependent on how good the human taggers did their job
- if there's a 3% error in your training corpus, then reporting 99% accuracy will make you look silly

Another Evaluation Metric

Agreement via Kappa

- adjusts for a baseline and thus normalizes for task difficulty
- the ratio of the proportion of times that two classifiers agree (corrected for chance agreement) to the maximum proportion of times that the classifiers could agree (corrected for chance agreement)
- $$\frac{P(A) - P(E)}{1 - P(E)}$$
- P(A) = percent correct
- P(E) = expected agreement (by chance)

Error Analysis

Once you have some results, you can improve your tagging scheme by doing error analysis.

Contingency Tables/Confusion Matrices are a way of visualizing what went wrong, where

- usually a table of raw counts
- rows indicate the correct tag
- columns indicate the output of the tagger
- $cell(row_i, column_j)$ contains the number of times that an item with the correct class x was classified as class y

Example

```
- Tagger Output
- A B C
A 0 0 5
B 0 1 4
C 0 7 1
```

- A was mistagged as C 5 times (i.e., 5 times when the tagger proposed C, it should have been a A)
- B ...
- C ...
- error analysis suggests that you might want to
  - merge tags B and C,
  - rewrite the coding instructions