In particular, we will rely on the notion of the probability of a sequence (e.g., the next word) can be captured using simple statistical techniques. A useful part of the knowledge needed to allow word prediction (guessing words in an instance) shows that we seem to have the ability to predict future

**Outline**

- Tagging
  - English Parts of Speech
  - Parts of Speech
  - Training and Testing
  - The Probability of a Sequence
  - Next Word Prediction

**Word Prediction**

- Lexical Knowledge
- Syntactic Knowledge
- Domain Knowledge

NY Times exercise shows that we seem to have the ability to predict future
Handwriting Recognition Example

Speech recognition

- find the most likely word sequence given a signal
- resolve word sense ambiguity

Resolving word sense ambiguity
- find the most likely word class given the surrounding words
- if it can be a noun or a verb

Find part of speech ambiguity
- you can assess the likelihood/goodness of a sentence
- can rank the likelihood of sentences containing various alternative words.

Why would anyone want to predict a word?

Corpus analysis

Statistical Methods Abound in NLP

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

In the sentence: "The fly flew the most likely part of speech tagging is...

These corpora are then used to find statistics...
An N-gram model uses the previous N-1 words to predict the next one.

Example sentence: 'He is going to the game.'

Collect a list of commonly substituted words.

Real Word Spelling Correction

Corpora (online collections of text and speech)
- what to find the things to count
- what is a word? (e.g. are cat and cars the same word?)
- words (this chapter); word classes; word sense; speech acts...

Problems are based on counting things, so...

Counting Words in Corpora
frequency events you haven’t seen yet. Some of the events in the table are rarely seen. Some are simply low frequency events. You might have to wait an abnormally long time to get valid statistics on low frequency events. You can quickly collect statistics on the high frequency events. A large number of events occur with high frequency. A small number of events occur with high frequency.

**Approximate Shakespeare (cont.)**

All those zeroes are causing problems. Could it be Shakespeare? The quadrants are worse. What’s coming out looks like Shakespeare be-case it is Shakespeare. Enters in the table. This, 99.9% of the possible bigrams were never seen (have zero Shakespeare produced 300,000 bigram types out of 884 million possible bigrams. There are 884,000 tokens. Shakespeare contain.

The empty Henry: Why dare stand forth thy company, forsooth? he is this puddle hill, without meaning, as I confess she leave all sorts, he is thin, Captain. Bigrams: What’s life? speech, or a more to leg det you order, for they. Empty cannot even severely. led.

As we increase the value of N, the accuracy of the n-gram model increases.
evaluation differences should be statistically significant

- Cross validation
- Hold out test set

Using standard metrics, a separate test portion of the corpus is used to evaluate the model. Specifically:

- Overly general corpora: Probabilities don’t reflect task or domain.
- Overly narrow corpora: Probabilities don’t generalize design the model.

Probabilities come from a training portion of a corpus, which is used to

---

**Methodology: Training and Testing**

---

Example (continued):
The distribution of tags

First, it is necessary so easy to tag most texts.
Therefore, while it is easy to determine the correct tags for most words, it is not so easy to tag things that are more meaningful.

The most frequent occurring words tend to have multiple tags (and of the rest, most have two, etc.)

Most word tags have only one part of speech.

Tags follow all the usual frequency-based distributional behavior.

Parts of Speech

- English parts of speech
- Parts of speech

We'll start with syntactic word classes.

Dependency on other words

Group with their neighbors to form phrases and sentences

Now we will turn to the study of how words

Evaluation issues

Next, grammatical and discourse expectations play a role.

Syntactic, semantic, and discourse expectations play a role.

To capture what we know about likely sequences of words.

However, N-gram language models are an extremely impoverished attempt.

N-grams trained from some corpus
etc.

- Gender (masculine, feminine, neuter)
- Number (singular, plural)
- Plural (mass nouns, count nouns)
- Common nouns
- Proper nouns

Nouns

- Nouns, pronouns, prepositions, conjunctions, articles
- Vary in small class, but frequently occur
- Relatively few membership

Closed (functional) class types

New words are coined/borrowed

Open (lexical) class types

English Word Classes

Information extraction

Word sense disambiguation

Parsing

Semantics

Parsing

Language modeling

Why the POS?

Notes on tags
English pronouns from CELEX on-line dictionary, with frequency counts from

Pronouns (continued)

2.5 million word counts.

Very in

adj.

with (which, who, whom, whichever, whomever)

defined (what, how, whoever, whoever, whom)

numeral

gender

person

Any in

adj.

item (which, who, whom, whichever, whomever, whom)

defined (what, how, whoever, whoever, whom)

numeral

gender

person

Any in

adj.

item (which, who, whom, whichever, whomever, whom)

defined (what, how, whoever, whoever, whom)

numeral

gender

person

Pronouns

Irregular verbs

Past participle (geden)

Past participle of passive (engaged)

Past participle of present-ge (gone)

Past of non-present-f (eat)

Standard morphological forms, as previously discussed.

Voice: active, passive

Other inflection: number, person

Tense: present, past, future...

Refer to actions, activities, processes, states (shows, wants, has, etc.)

Adjectives and Adverbs

Adjectives (semantically) describe properties of qualities.

Verb
Conjunctions

- **coordinating (and, or, but)**: things of equal stature
- **subordinating (that, for, because, although)**: one element has an ended...

Conjunctions join two things.

Parents occur before noun phrases.

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

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

How are nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and articles formulated?
Part of Speech Tagging

A text to be tagged (string of words)

4 words: in a corpus

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

Part of Speech Tagging

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z

A,B,c,d

Z
The Brown Corpus

There is also a tagged version of the Brown Corpus

The Brown Corpus of one million words of American English text printed in

The Brown Corpus of Standard American English was the first of its kind.

11.2% Action (myself) (49 words)
10.0 Action (yourself) (49 words)
9.8 Repeat (68 words)
8.8 Miscellaneous: ever and there or once or twice (70 words)
7.0 Novel: letters (70 words)
5.9 Poetical verse (68 words)
5.7 Shills and Bobbies (52 words)
4.0 Action (himself) (41 words)
3.0 Poetical: nouns (38 words)
2.1: Pronouns (1,467 words)
1.1: Unambiguous (1 word)

Unambiguous (2,769 words)
ambiguous (1 word)

But 60% of tokens in the Brown Corpus are ambiguous

11.5% of English words in the Brown Corpus are ambiguous

Why is the Tagging Problem?

- My brain again said that here would be a real on this flash.
  - that can be a DT or complementizer
  - Can a read a book on this flash?
  - book can be an or VP
  - Can be a type of disambiguation

Example

does/VP the/DT higher/NN serve/VB dimension/NN
book/NN the/DT higher/NN

help://www.net/ide/javascript/index.js
will make you look silly.

If there is a 3% error in your training corpus, then reporting 99% accuracy
this is all dependent on how good the human taggers did their job.

Upper Bound/Ceilings

speech act tagging, where the baseline is much lower.

This is a result of 99% for POS tagging is less impressive than for say

\% can get you 90% for POS?

in the case of POS tagging, this is the most frequent tag heuristic (unti-

dumb approach could achieve

your goodness metric has to take into account some baseline that any

Lower Bound/Baseline

<table>
<thead>
<tr>
<th>Evaluation Issues</th>
</tr>
</thead>
</table>

Making stupid claims?

Doing more?

How do you know if you're doing well?

Use a tagged test corpus

How do you know how well you're doing?

<table>
<thead>
<tr>
<th>A Common Evaluation Metric</th>
</tr>
</thead>
</table>

- [12] Action (science) (6 texts)
- [14] Action (romance) (29 texts)
- [15] Humor (9 texts)
Example

- For the coding instructions
- "a" and "b and c"
- Error analysis suggests that you might want to
- ...c...
- ...b...
- It should have been a (A)
- A was misclassified as C 5 times (i.e., 5 times when the target proposal C, ...0 7 1
...0 1 4
...0 0 5
...A 5
- Target Output
- Target Output

Error Analysis

Next you have some results, you can improve your tagging scheme by doing

Another Evaluation Metric