

# The Use of Corpora in Teaching and Research

## Part 1: Theoretical Preliminaries

Richard Sproat

<http://www.linguistics.uiuc.edu/rws/>  
rws@uiuc.edu

# Overview

- Introduction to Corpora and Corpus Based Methods
- Concordancing
- An example: Part of speech tagging

# Corpora and Accessing Corpora

- The Pendulum.
- Electronic Corpora and what they are useful for.
- Concordancing.

# The “Pendulum”

- “Corpus linguistics” had a long and venerable history.
  - ★ Philologists in the 19th century collected and indexed (‘concordanced’) corpora of early texts
  - ★ Lexicographers pored over corpora to find usages of words
  - ★ Linguists of the 1950’s like JR Firth and Zellig Harris pursued corpus-based analyses.

Firth proposed his famous maxim: “you shall know a word by the company it keeps” .

## The “Pendulum”

- Then the advent of generative linguistics in the 1960’s put a big dent in all that.
  - ★ Chomsky presented arguments against statistical approaches to linguistics. Some of these arguments were bogus, others well taken. A lot of this was tied up with the simultaneous war with the behaviorists (e.g. B.F. Skinner).
  - ★ And new techniques — in particular introspection about grammaticality — were proposed.

## The “Pendulum”

- But starting in the late 1980's the “pendulum” began to swing back, for several reasons:
  - ★ Purely “symbolic” approaches to NLP are too fragile.
  - ★ More and more corpora have become available.
  - ★ Computers have become faster and therefore more able to make use of the corpora that we have.

A lot of the early work started at the industrial research labs (AT&T Bell Labs, IBM), who usually had better machines than anyone else.

The result: if you go to an Association for Computational Linguistics meeting today, the complexion of the program will be totally different from what it was in the 1980's

# The “Pendulum”

- But the “pendulum” metaphor implies a dichotomy, where really none exists.

# The Growth of Electronic Corpora

- In the mid 1980's there weren't very many publicly available corpora
- One of the main corpora was the Brown Corpus (E.g. [http://clwww.essex.ac.uk/w3c/corpus\\_ling/content/corpora/list/private/brown/brown.html](http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html))
  - ★ Collected at Brown University based on texts available in 1961
  - ★ Was a “balanced” corpus: texts from a wide variety of genres (news, novels, sports, medicine . . . )
  - ★ Contained 1 million words divided into 500 texts of 2,000 words each.
  - ★ This makes the distribution of words in the Brown corpus a bit odd.
- Much used by psycholinguists interested in using frequency balanced stimuli
- Later on a part-of-speech tagged version of the Brown corpus was produced. Many early statistical taggers were trained on this.

# Some Corpora Available from the Linguistic Data Consortium

<http://www.ldc.upenn.edu>

Types of text available (inter alia):

- Newswire/newspaper
- Broadcast news
- Spontaneous speech transcriptions (e.g. Switchboard)
- Parallel multilingual texts
- Annotated text: Treebanks

## Newsire Data

Language	Approx # Words
Arabic	400 million
Chinese	1,450 million (characters)
Portuguese (European)	195 million
French	100 million
German	90 million
English	2.5 billion
Japanese	30 million
Korean	143K articles
Spanish (various)	170 million

# Newsire Data

- These are probably undercounts.
- A year of the *Associated Press* newswire is approximately 40 million words.
- Lots of corpora by other agencies of course. E.g. the British National Corpus (100M words) (<http://www.natcorp.ox.ac.uk/>)
- And of course this pales in comparison with what's available in (very) raw form on the web.
  - ★ Which company do you think is hiring natural language people right now?

## Parallel Multilingual Text

Language Pair	Size
French/English (Canadian Hansards)	2.9 million parallel sentences
English/Chinese (Hong Kong Parallel Text)	59M Eng. words/98M Chinese chars
English/French/Spanish (UN Archives)	58M words for English

## Treebank Data

Language	# Words
English	1 million
Arabic	700K
Chinese	500K

- Chinese treebank includes (obviously) word segmentations
- There are other treebanking projects elsewhere

## What Can be Done with Corpora: A Preview

The basic answer is *language modeling*: given a sequence of words that I've already seen, what's the most likely word to follow?

With appropriate abstractions on “sequence” and “word” this covers a lot of ground

- Part-of-speech tagging: predict the next word on the basis of the previous tags (I know this seems odd, but . . . )
- Language modeling for speech recognition:  
It's easy to recognize speech  
It's easy to wreck a nice beach

## What Can be Done with Corpora: A Preview

- Sense disambiguation: use statistics of cooccurring words to figure out which sense of a word is intended (cf. Firth's maxim):

Beside the three violins and two violas there were a cello and a **bass**.  
The waters were teeming with blackmouth **bass**.

- Spelling correction: I don't know weather you noticed this error
- Handwriting recognition: I have a gub
- Text normalization: you are trying to predict the next actual word given abbreviated input:

57 ST E/1st & 2nd Ave Huge **drmn** 1 **BR** 750+ **sf**, lots of sun & **clsts**. Sundeck & **Indry facils**. **Askg** \$187K, **maint** \$868, **utils incld**. Call **Bkr** Peter 914-428-9054.

## What Can be Done with Corpora: A Preview

- Parsing: here the “sequence” would be a richer structure such as a set of already predicted tree nodes
- Machine translation: here you want to predict the most likely word (sequence) in language X given a word sequence in language Y.

In all cases one does better if one has a model of the *domain* that one is trying to model. Hence one typically trains one’s models on text that approximates the kind of text that one is going to be dealing with.

NB: A “balanced” corpus is often not as useful, precisely because you are not tuned for a particular domain, but rather (e.g.) “general English”. (But what exactly **is** general English . . . )

## Noisy Channel Model: A Quick Glance

- The basic idea: a lot of problems in NLP can be construed as the problem of reconstructing an underlying “truth” given possibly noisy observations.
- This is very much like the problem that Claude Shannon (the “father of Information Theory”) set out to solve for communication over a phone line.
  - ★ Input  $I$  is clean speech
  - ★ The channel (the phone line) corrupts  $I$  and produces  $O$  — what you hear at the other end

## Noisy Channel Model: A Quick Glance

- Can we reconstruct  $I$  from  $O$ ?

Answer: you can if you have an estimate of the probability of the possible  $I$ 's and an estimate of the probability of generating  $O$  given  $I$ :

$$\operatorname{argmax} P(I)P(O|I)$$

First term  $P(I)$  is the *language model* and the second term  $P(O|I)$  is the *channel model*.

The probabilities  $P(I)$  and  $P(O|I)$  are estimated from corpora: usually one never has enough data to estimate all the probabilities directly, so one has to apply one or another *smoothing* or *backoff* technique to provide estimates for parameters one has not seen or not seen frequently enough.

# Simple Concordancing

- A concordancer is a program that allows for rapid access to particular terms or term sequences in a corpus.
- In general a linear search through a corpus is *not* viable.
- So you need an indexing scheme

## Inverted Index

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

## Typically you'd do some preprocessing. . .

<start/>

Alice was beginning to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped into the book her sister was reading , but it had no pictures or conversations in it , ' and what is the use of a book , ' thought Alice ' without pictures or conversation ? '

<pbreak/>

So she was considering in her own mind ( as well as she could , for the hot day made her feel very sleepy and stupid ) , whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her .

<end/>

## Index Positions

<start/> 1

Alice 2 was 3 beginning 4 to 5 get 6 very 7 tired 8 of 9 sitting 10 by  
 11 her 12 sister 13 on 14 the 15 bank 16 , 17 and 18 of 19 having 20  
 nothing 21 to 22 do 23 : 24 once 25 or 26 twice 27 she 28 had 29  
 peeped 30 into 31 the 32 book 33 her 34 sister 35 was 36 reading 37 ,  
 38 but 39 it 40 had 41 no 42 pictures 43 or 44 conversations 45 in 46  
 it 47 , 48 ‘ 49 and 50 what 51 is 52 the 53 use 54 of 55 a 56 book 57  
 , 58 ’ 59 thought 60 Alice 61 ‘ 62 without 63 pictures 64 or 65  
 conversation 66 ? 67 ’ 68

<pbreak/> 69

So 70 she 71 was 72 considering 73 in 74 her 75 own 76 mind 77 ( 78 as  
 79 well 80 as 81 she 82 could 83 , 84 for 85 the 86 hot 87 day 88 made  
 89 her 90 feel 91 very 92 sleepy 93 and 94 stupid 95 ) 96 , 97 whether  
 98 the 99 pleasure 100 of 101 making 102 a 103 daisy-chain 104 would

105 be 106 worth 107 the 108 trouble 109 of 110 getting 111 up 112 and  
113 picking 114 the 115 daisies 116 , 117 when 118 suddenly 119 a 120  
White 121 Rabbit 122 with 123 pink 124 eyes 125 ran 126 close 127 by  
128 her 129 . 130

<end/> 131

# Sort

' 59  
' 68  
( 78  
) 96  
, 117  
, 17  
, 38  
, 48  
, 58  
, 84  
, 97  
. 130  
: 24  
<end/> 131  
<pbreak/> 69  
<start/> 1  
? 67

Alice 2  
Alice 61  
Rabbit 122  
So 70  
White 121  
' 49  
' 62  
a 103  
a 120  
a 56  
and 113  
and 18  
and 50  
and 94  
as 79  
as 81  
bank 16  
be 106  
beginning 4  
book 33

book 57

.

.

.

Sproat

Theoretical Preliminaries

Concordancing

## Construct an Index to the Index

' 1 2  
( 3 3  
) 4 4  
, 5 11  
. 12 12  
: 13 13  
<end/> 14 14  
<pbreak/> 15 15  
<start/> 16 16  
? 17 17  
Alice 18 19  
Rabbit 20 20  
So 21 21  
White 22 22  
' 23 24  
a 25 27  
and 28 31

as 32 33  
bank 34 34  
be 35 35  
beginning 36 36  
book 37 38  
but 39 39  
by 40 41  
close 42 42  
considering 43 43  
conversation 44 44  
conversations 45 45  
could 46 46  
daisies 47 47  
daisy-chain 48 48  
day 49 49  
do 50 50  
eyes 51 51  
.  
.  
.

## Using the Indices

- To find locations of a given word, look up the word in the index to the index, and find its start and end points.
- Then seek to the start point in the main index and read the corpus positions referenced there until you get to the end point in the main index.
- For each corpus position, output a *window* of text around the corpus position
- If you want to look up a word sequence (e.g. *Alice was beginning*) then look up the least frequent word first (you would also have computed a *histogram*) and then look for the other words in windows around the least frequent word

## An Example: Part of Speech Tagging

- Part of speech (POS) tagging is simply the problem of placing words into equivalence classes.
- Notion of part of speech tags can be attributed to Dionysius Thrax, 1st Century BC Greek grammarian who classified Greek words into eight classes: noun, verb, pronoun, preposition, adverb, conjunction, participle and article.

## An Example: Part of Speech Tagging

- Tagging is arguably easiest in languages with rich (inflectional) morphology (e.g. Spanish) for two reasons:
  - ★ It's more obvious what the basic set of tags should be since words fall into morphologically distinct classes.
  - ★ The morphology gives important cues to what the part of speech is: *cantaremos* is highly likely to be a verb given the ending *-ar-emos*.
  - ★ It's arguably hardest in languages with minimal (inflectional) morphology:
    - \* there are fewer cues in English than there are in Spanish
    - \* for some languages like Chinese, cues are almost completely absent and linguists can't even agree on whether (e.g.) Chinese distinguishes verbs from adjectives.

# The Penn Treebank Tagset

- 46 tags, collapsed from the Brown Corpus tagset
- Some details:
  - ★ *to/TO* not disambiguated
  - ★ Verbs and auxiliaries (*have, be*) not distinguished (though these were in the Brown tagset).
- Some links:
  - ★ <http://www.computing.dcu.ie/~acahill/tagset.html>
  - ★ <http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html>
  - ★ <http://www.scs.leeds.ac.uk/amalgam/tagsets/upenn.html>

# The Penn Treebank Tagset

- Link for the original Brown corpus tags:
- <http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html>
- Motivations for the Penn tagset modifications
  - ★ “the Penn Treebank tagset is based on that of the Brown Corpus. However the stochastic orientation of the Penn Treebank and the resulting concern with sparse data led us to modify the Brown tagset by paring it down considerably” (Marcus, Santorini and Marcinkiewicz, 1993).
  - ★ eliminated distinctions that were lexically recoverable: thus no separate tags for *be*, *do*, *have*.
  - ★ as well as distinctions that were syntactically recoverable (e.g. the distinction between subject and object pronouns)

## Problematic Cases

Even with a well-designed tagset, there are cases that even experts find it difficult to agree on.

- adjective or participle?

a seen event, a rarely seen event, an unseen event,

- a child seat, \*a very child seat, \*this seat is child

but: that's a very MIT paper, she's soooooo California

- preposition or particle?

he threw out the garbage

he threw the garbage out

he threw the garbage out the door

\*he threw the garbage the door out

## Problem of Tagging

Want to be able to tag such things as:

*Can they can cans?*

*May may leave*

*He does not shoot does*

*You might use all your might*

*I am arriving at 3 am*

## How Hard is Tagging?

- Many words are unambiguous. From the Brown corpus:

# tags	# types with that many tags
1	35,340
2	3,760
3	264
4	61
5	12
6	2
7	1 “still”

- Baseline for English (Penn tagset) something like 91%.

# Various Approaches to Automatic Tagging

- Handwritten rules
- Source-channel (“HMM”)
- Maximum Entropy (MaxEnt)
- Transformation-based learning (the “Brill tagger”)

## Approaches to Automatic Tagging: Source-Channel Model

- Basic problem: uncover the underlying **signal** of POS tags as modified by the **noisy channel** that produces observable words from tags.
- For a bigram tagger this would give you the formula for the  $i$ th tag:

$$t_i = \operatorname{argmax}_j P(t_j | t_{i-1}) P(w_i | t_j)$$

- For the whole sentence then we want to maximize:

$$\prod_j P(t_j | t_{j-1}) P(w_j | t_j)$$

## Approaches to Automatic Tagging: Source-Channel Model

- Note that this can also be derived via Bayes' formula for a tag sequence  $T$  and word sequence  $W$ . Thus we want to maximize:

$$P(T|W)$$

which is given by

$$P(T|W) = \frac{P(T)P(W|T)}{P(W)}$$

But since we **know** the word sequence we can eliminate that and just maximize

$$P(T)P(W|T)$$

## Approaches to Automatic Tagging: Source-Channel Model

- Now if we assume a bigram model for  $P(T)$  and if we assume that the choice of  $w_j$  given  $t_j$  does not depend upon either  $w_{j-1}$  or  $t_{j-1}$  we can rewrite this as

$$\prod_j P(t_j|t_{j-1})P(w_j|t_j)$$

- Both the language model and the source model are estimated from tagged data.
  - ★ The ngram tag language model is estimated as we described previously for word ngrams
  - ★ The channel model  $P(W|T)$  is also estimated from the corpus, though typically it would be augmented with information from dictionaries.
- Accuracy rates of 96% are typical for this algorithm

## Approaches to Automatic Tagging: Source-Channel Model

- What do you do if you don't have tagged data?

You can assume an initial distribution of tags over the corpus (given a dictionary and perhaps some linguistically based guesses) and then use an algorithm such as **expectation maximization (EM)** — see Appendix D of Jurafsky and Martin, and later.

## Problems with Taggers (in English)

- Prenominal NN or NNP (proper name) or JJ (adjective):  
*Brown Corpus*
- RP (particle) or RB (adverb) or IN (preposition):  
*run up a pipe*  
*run up a bill*
- VBD (past verb) versus VBN (past participle) versus JJ:  
*The vase was broken yesterday: it was fine the day before.*  
*The vase was broken yesterday: but now it's fixed.*

## Some References

- Jurafsky, D. and Martin, J. 2000. *Speech and Language Processing*. Prentice-Hall.
- Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Baayen, H. 2000. *Word Frequency Distributions*. Kluwer.