In this lecture we examine some of the philosophical themes and leading ideas that motivate statistical approaches to linguistics and natural language and begin exploring what can be learned by looking at statistics of texts. We will use some terminology that will be introduced later.

1 The study of natural Language

The study of language is concerned with two basic questions:

- What kinds of things do people say?
- What do these things say/ ask/ request about the world?

The first point covers aspects of the *structure of language*. The second pertains to semantics, pragmatics and discourse – how to connect utterances to the world.

Most of *corpus linguistics* is about the first point. But, patterns of use can also imply deep understanding, and therefore corpus based techniques may also be used to address the second question. In some sense, if one wants to make significant progress in NLP, one should hope that the fact that something can be said, statistically, about the first point, would facilitate progress on the second. Even strongly, these statistical regularities could be the only reason that natural language is such a powerful communication channel.

Wittgenstein: The meaning of the word is defined by the circumstances of its use.

However, most of the statistical *discoveries* are done in the context of the first question, so we will discuss it in the rest of this lecture.

Traditional Linguistics view of language is that "People produce grammatical sentences". As a consequence of this basic view the theory developed cares about whether the sentence is structurally well formed and less about whether this is the kind of thing people say, or whether this is semantically strange.

This document (as well as the lecture...) exemplifies that this distinction is not sufficient. People can use ill-formed language and be perfectly clear and vice versa. In many cases you can hear non native speakers of a language use sentences that are *grammatically correct* but are not what people typically say. (E.g. "Open the Radio".)

Chomsky's famous example:

(1) Colorless green ideas sleep furiously

vs.

(2) Furiously sleep ideas green colorless

was meant to make a point against statistics as an approach to language. Both these sentences have never occurred. Hence, in any statistical model, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet, (1), though nonsensical is grammatical, while (2) is not.

But, what does it really show?

- Is grammatical/non grammatical the significant distinction?
- Can statistical approaches "discover" that one sentence is grammatical and the other is not? (Important point: representation).

For an additional set of examples: think of sentences for which the question of whether they are grammatical or not is open.

2 Some problems with non-statistical approaches

[Abn96] provides a very good survey of problems in natural language that non-statistical approaches to linguistics will find hard to handle. We mentioned only a few of them here.

2.1 Non Categorical Phenomena in Language

In many cases, a binary decision on the meaning or the role of a word is impossible.

Language Change: Words change meaning and their part of speech in the sentence. For example, the word **while** used to be a noun that describes a time period, as in to take a while.. Now it is used more often as a complementizer that introduces clauses, as in while you were out....

It is clear that these kinds of changes are not categorical.

Blending POS The word **near** used to be a preposition, as in he lives near the station.. Now it is used also as an adjective, as in we will review that decision in the near future..

2.2 The ambiguity of Language

Ambiguity exists in almost any natural language decision, and at almost any level. The examples we gave last time are all a result of ambiguities that, in principle, can be resolved in several ways, only one of which makes sense to humans. The following example pertains more to ambiguities that reflect on the close coupling of syntactics and semantics.

Most traditional approaches attempt first to determine the *structure* of the sentence and then use it to determine other things, like "who did what to whom". Semantic analysis is done, if at all, only after the syntactic analysis. Can they be decoupled. Consider the sentence:

Our company is training workers .

[Our company NP] [[is aux] [[training V] [workers NP] VP] VP]

Here training workers is understood correctly. is training is the Verb group.

[Our company NP] [[is V] [[[training V] [workers NP] VP] NP] VP]

Here is is the main verb and training workers is used like a gerund as in our problem is training workers.

[Our company NP] [[is V] [[training $AdjP$] [workers N] NP] VP]

Here is is also the main verb and training modifies workers as in training wheels.

This is an example of a sentence with (at least) three different syntactic analysis (parses). Examples of these sort exist in almost any non-trivial, or long enough sentence. Prepositional phrases can always be used as examples here since that typically have several possible attachments, only one of which makes sense. E.g.:

I wore the shirt with the short sleeves.

Long sentences may have hundreds different syntactically legitimate parses. The sentence

List the sales of the products produced in 1973 with the products produced in 1972.

is reported to have 455 different parses by one parsing system.

In addition to these *legitimate* ambiguities, that are also many problems with the fact that language, when used, does not produce well-formed sentences. Speech data, discourse data, e-mail, etc., all make use of ill-formed sentences. In many cases, it is not even clear whether a good parse exists; perhaps all one can do is extract some key phrases and use them to make sense of the sentence.

Thanks for all you help

Has a legitimate parse; its meaning is "thanks for all those that you help". Most likely, you will read here your rather than you, since the parse is just more likely.

2.3 Robustness; Scaling up

A Natural Language Processing system is required to be good at making disambiguation decisions (word sense, category, syntactical structure, semantic scope,...). Even if one could write down a good set of constrains and preference rules as a basis for a system that make natural language inferences of these sort, we still need to address

- Scaling up beyond small and domain specific applications.
- Practicality: time consuming to build if we want reasonable coverage
- Brittleness (e.g., in the face of using metaphors)

This, before we even touched upon the problem of world knowledge that is required if we want to perform any significant inferences in natural language. In many respects, this is just an instance of the general Knowledge Representation problem in AI, and it is hard to imagine that it is possible to get around this problem without learning. See a good discussion of this point in the introduction to [Cha93].

3 Preliminary notes about the Statistics of language

With the above as motivation, let us consider what can be learned about language by looking at texts and extracting statistics from it. See a pointer to *Unix for Poets, K. Church* for a nice "do it yourself" introduction. As a running example we will use the text of *Tom Sawyer* by Mark Twain. There are a few questions we can ask, for which the answer is easily derived by taking statistics.

As we are taking statistics, keep in mind the question – are we getting any closer to understanding Tom Sawyer?

What are the most common words?

Common words in Tom Sawyer

Word	Frequency	Use
The	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive)
in	906	preposition
that	877	complementizer, demonstrative pronoun
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	(proper) noun
with	642	preposition

The table presents the most common words in Tom Sawyer. The list is dominated by function words (determiners, prepositions, complementizers) with the addition of the word Tom, the only indication to the content. How representative is that?

How many words are in the text?

- There are **71,370** word tokens.
  It that enough to collect statistics on ?
- The text takes 0.5 MB (500k characters)
  a very small corpus relative to those being used today in corpus based NLP.

- There are only **8018** different words.
  There is a ratio of 1:9 between word types and tokens. This is a fairly small number. It can be attributed to the fact that this is a children's book. The same size text of news would have about **11,000** different words (yielding a ration of 1:6 or so).

Two important issues to address here are:

- Does the ratio, #(types)/#(tokens) depend on the *size of the corpus*?
- Here, on average, words occur 9 times. But, what is the distribution? It turns out that word types have very uneven distribution.

What is the distribution of words?

The table presents the number of times the ith most common word in the text occurs:

Frequency of Frequencies of Word Types

# of occurrences	# of words
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
>100	102

If we look more carefully on the data we can see that:

- The most common 12 words (over 700 occurrences) account for 1% of the text.
- The most common 100 words account for 50.9% of the text.
- Almost half (3993/8018=49.8%) of the words occur once!
- Over 90% of the word types occur less than 10 times . (only 540+99+102=741 occur >10 times)

This data shows that language used in text does have some non-uniform and perhaps interesting statistics. On the other hand it can be used to reflect on approaches that use mostly "bag of words" approaches.

3.1 Zipf's Laws

Are these observations linguistically significant? Probably not. Perhaps these can be used as an indication of Authorship or writing style. The significance of these observations is mostly in that they indicate that statistical NLP is hard. It is hard to predict much about the behavior of words if we do not observe them in the text.

Will the phenomenon of "long tail" go away when we use a larger corpus? No. This is exhibited by Zipf's Laws - the first results is corpus linguistics.

Zipf (1929) made some extreme claims about unifying principles he discovered and their implication to the understanding of the human nature, in a series of works about what he called "The principle of the least effort". From our point of view, the most important result is the

First Zipf's Law (1929):

Let

- f be the frequency of a word type in a large corpus (# of occurrences)
- r be the position of the word in a list of words ranked according to frequency.

Then, f is proportional to $1/r$. Equivalently, there exists some constant K, s.t

$$f \cdot r = K$$

.

That means that the 50th most common word should occur three times more often in the text than the 150th most common word. The following table presents an empirical evaluation of Zipf's law on the text of Tow sawyer.

Empirical Evaluation of Zipf's Law

Word	Freq.	Rank	f x r	Word	Freq.	Rank	f x r
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5335	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	COULD	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

The empirical evaluation shows that the law pretty much holds. A little less than expected for the first 3 words, a little more around 100. But, still good as a rough description of the frequency distribution in human language. What happens in other languages? Notice that Zipf's law were formulated for German, and since then have be studied in many other languages.

Based on his findings Zipf developed a theory that claimed as follows (very informally):

Both speaker and listener are trying to minimize their effort.
The *speaker:* by using a small vocabulary of common words and
The *hearer*, by having a large vocabulary of rare words, to reduce ambiguity.

Regardless of this, however, the practical consequence of the law is important:

For most words, our data about their use will be exceedingly sparse.

Mandelbrot (the Fractals guy, 1954) studied Zipf's laws extensively, and derived more general relationships between f and r. One other issue to learn from this – not everything is distributed according to the *normal distribution*. In this case, we get some kind of *hyperbolic distribution*.

Zipf phrased some other, less known, laws. The second Zipf's law had to do with *word meaning*. If m is the number of meanings a word has then m is proportional to the square root of f.

For example, according to this law, words of rank 10,000 average 2.1 meanings, words of rank 5,000 average 3 meanings, words of rank 2,000 average 4.6 meanings and so on. (m behaves like $\sqrt{f} \approx 1/\sqrt{r}$

The third Zipf's law had to do with word clumps. The idea was to measure, for each content word, the number of words between consecutive occurrences of them in the text. If F is the frequency

of intervals lengths, and the interval length is I, then F is proportional to the inverse of I. That is, content words tend to occur near each other. Other Zipf's laws have to do with morphology (inverse relation between the frequency of words and their length, etc.

Are Zipf's laws surprising?

In order to try to understand that, assume you generate words according to the following model:

Uniformly choose one of 27 characters (26 letters + blank).

Then,

$$Prob[\text{word of n characters is generated}] \approx= 1/27(26/27)^n$$

(we do not distinguish between cases with repetitions). That means that there are

- 26 times more words of length $n+1$ then words of length n, and
- words of length n are more frequent than words of length n+1.

It can be shown that these two combine to guarantee the regularity of Zipf's laws and of Manelbrot's refinements. This may indicate that Zipf's laws may not be so valuable as a characterization of language. But, the basic insight is still important. There is some interesting statistics of language, but

Frequency based approaches are hard since most of the words are rare.

Should we move beyond words?

Given the above observations with respect to single words, we may think of trying to learn about the language by acquiring statistics from longer sequences of words.

A *Collocation* is a phrase which is formed by a combination of "parts" and that has an existence beyond the sum of its parts. Examples may include:

- Compounds. e.g., disk drive or
- Phrasal verbs e.g., make up or
- Phrases e.g., bacon and eggs.

Any phrase that people repeat because they have heard others using it, is a candidate for a collection. (E.g., "el-em-en-o-p" is an example that young kids may remember as a collocation.)

As such, collocations are important in Translation and in Information extraction, as well as many other Natural Language tasks. Probably these phrases should be in the dictionary, since their

meaning, in most cases, is not composed from the meaning of the words they are formed of. Notice the collocations can be long and discontinuous (e.g., put [something] on).

To the extent that most of language use is *people reusing phrases* and constructions they have heard before, collocation are very important

What about statistics of collocations?

The definition we've given is *not constructive*; it is hard to identify them. We will start just by collecting statistics of sequences of 2 words (also called bigrams) as taken from the New York Times. This is shown in the table below.

Commonest bigrams in the New York Times

Frequency of Collocation	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

We can see that the common syntactic constructions are those involving individually extremely common words. In addition, most of them follow the form [preposition, determiner]. In this way, it is hard to say that we have gained something by looking at these pairs. If we want to gain something from looking at pairs of words, we need to gather statistics more carefully. For example, we could

- Take into account the frequency of each of the words.

- Remove POS sequences that are not interesting (e.g., [preposition, determiner])
- Keep only POS sequences of interest (e.g., [adj-noun], [noun-noun]

But, doing it this way, we need to be able to tag the text for part-of-speech. Only in order to gather reasonable statistics.

The next table shows that this indeed yields much better results.

Frequent bigrams after filtering

Frequency	Word 1	Word 2	POS pattern
11487	New	York	AN
7261	United	States	AN
5412	Los	Angeles	NN
3301	last	year	AN
3191	Saudi	Arabia	NN
2699	last	week	AN
2514	vice	president	AN
2378	Persian	Gulf	AN
2161	San	Francisco	NN
2106	President	Bush	NN
2001	Middle	East	AN
1942	Sadam	Hussein	NN
1867	Soviet	Union	AN
1850	White	House	AN
1633	United	Nations	AN
1337	York	City	NN
1328	oil	prices	NN
1210	next	year	AN
1074	chief	executive	AN
1073	real	estate	AN

To summarize, we exhibits some of the problems in traditional approaches to language processing, but also some of the difficulties in applying *simple minded* statistical analysis. Simple minded search does not work, and there was a need to add information, beyond what is available in the text, even to collect statistics on the text. (Do we have a chicken and egg problem?)

4 Handouts:

- Steve Abney's paper on Statistics and Linguistics.

References

[Abn96] S. P. Abney. Statistical methods and linguistics. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, 1996.

[Cha93] E. Charniak. *Statistical Language Learning*. MIT Press, 1993.