Chapter 1

Introduction

1.1 What we're going to learn

We're going to learn how to model the statistical-distributional properties of language using very simple tools that have become widely used in the computational world. Once we're on top of the basics, we'll go on to ask whether these tools represent an alternative theory of language, or less grandiosely, whether these tools can provide insight into how the theory of language should be developed.

1.2 What is "Statistical NLP"?

What does Statistical natural language processing—or NLP—refer to? It refers to a class of models that can be used in the domain of natural language processing.

The models. To say these models are "statistical" is to say several things. First, the models exhibit certain mathematical properties. Second, the models govern the *frequency* with which various constructions might occur. Finally, the models are not specified in their entirety in advance, but are "trained" on the basis of data. While these models have very explicit mathematical structure, it is not at all clear whether they have any *linguistic* structure. It is, then, aren *extremely* interesting question whether linguistic models of language structure and statistical models of language structure can inform each other...

Natural language processing. This term is unfortunately ambiguous.

What it refers to in the present context is *not* how human beings process language. Rather it refers to how we might process language with a computer. It is of course an *exceedingly* interesting question whether theories of human and computer processing can inform each other...

1.3 Why do this?

Why should we believe that modeling the statistical properties of language has some utility? There are several ways to go at this. On the one hand, one can argue that the statistical distribution of linguistic elements has implications for linguistic questions. On the other hand, one can also argue that this statistical distribution has computational applications.[1]

1.3.1 Linguistic motivation

In general, generative linguistics has not treated the frequency with which some construction might occur in a language.

In syntax, for example, Chomsky once argued that parasitic gap constructions provided critical evidence for the way syntax should be organized, e.g. sentences like "Which book did you read t without buying t?". Yet sentences of this type are actually quite rare. On the orthodox view, their rarity is irrelevant. Should syntactic theory also account for the frequency of sentences of different types?

In phonology, on typological grounds, one can argue that *closed* syllables are more "marked" than *open* syllables. A closed syllable is one where the rightmost element is a consonant; an open syllable is one where the rightmost element is a vowel. For example, in a word like *candy*—where the syllables are *can* and *dy*—the first syllable is closed and the second open. This markedness relation entails that all languages have open syllables; only a subset have closed ones. Yet, in a language like English, which has both kinds of syllables, closed syllables are much more common than open syllables. This is irrelevant to orthodox phonological theory. Should phonological theory account for the frequency of syllable types?

On the other hand, there is evidence that frequency of items is something that people are sensitive to. I'll give two examples: from grammaticality judgments and from acquisition.

[1] One can argue that it has other applications as well (Hammond 1999).

Coleman & Pierrehumbert (1997) and Frisch *et al.* (2000) conduct a series of experiments demonstrating that the frequency of phonological elements correlates with judgments of well-formedness. For example, subjects are presented with nonsense words like [blɪk], and asked either to judge them on a scale from 1 to 7, or simply to indicate whether they are well-formed or not. Both studies show that these judgments correlate with the probabilities of the phonological constituents that make up the nonsense words. Therefore the main task of generative grammar, "grammaticality" judgments, are a function of frequency.

Zamuner (2001) considers the very earliest stages of phonological acquisition and asks what factors govern the acquisition of coda consonants. What we might expect is that the least "marked" elements are acquired first. This is what the most natural innatist theory would maintain. She shows, instead, that what is acquired first is what is most common in the ambient language. Thus, in acquisition, the most frequent stuff is the stuff that seems to be acquired first.

Presumably the same arguments could be made in the syntactic domain.[2]

You can make a similar argument from learnability theory (Gold 1967). Consider the following set of hypothetical languages:

$$\begin{aligned} H_0 &= \{a, aa, aaa, \ldots\} \\ H_1 &= \{a\} \\ H_2 &= \{a, aa\} \\ H_3 &= \{a, aa, aaa\} \\ H_i &= \{a, aa, \ldots, a_i\} \end{aligned} \tag{1.1}$$

Assume that this is the set of possible grammars. Assume, moreover, that the learner can only learn from some finite presentation of positive evidence. That is, the learner will witness some number of grammatical sentences and at some point must venture a guess about what the grammar is.

Consider what such a learner would do if exposed to $\{a, aaaa\}$. If the learner were to routinely guess H_4 for this, what kind of (finite) evidence would ever compel her to guess H_0? Likewise, if the learner were to routinely guess H_0 for this, what kind of (finite) evidence would compel her to guess H_4?

It can be shown that this learnability paradox is resolvable with implicit

[2]Pull this from the "Satiation" paper in *LI*? Also, another paper in *Language* that Andrew C. knows the citation of. Until I do this, solicit ideas from the class.

negative evidence. Such evidence comes from the fact that the learner might *expect* to be exposed to forms of some specific sort by some specific time. In the absence of such forms, the learner concludes that they are generally ruled out, as opposed to simply missing from the learner's experience.[3]

1.3.2 Computational applications

Modeling the statistical distribution of linguistic elements has dramatic computational applications. There are two general domains: what I will refer to as *prediction* and *identification*. Let's consider a few examples.

Prediction. A statistical model of language allows us to use linguistic context to *predict* subsequent linguistic elements. One example of this is speech recognition.[4] The basic idea in speech recognition is that acoustic objects are mapped to linguistic elements, e.g. phonemes, words, sentences. Consider now the problem of identifying some acoustic event as representative of some linguistic element.

There are actually two kinds of evidence for what this element might be. One kind of evidence is the set of acoustic properties it instantiates. Another kind of evidence is what kind of linguistic element we are *expecting*. For example, imagine we are confronted with some acoustic event that could be identified as *hat* or *had*. If the preceeding context is "the cat in the ...", then it is fairly obvious which of these to choose. This is how statistical language modeling can aid speech recognition.[5]

Identification. The statistical properties of a language sample can be used to identify that sample. For example, statistical modeling is used in language identification.[6] One can develop statistical models of different languages and then use those models to identify languages based on a limited sample. Presumably the same technique can be used to identify topics and authorship.

---

[3] Given that experience is finite at any particular point in time, does this suggest that H_0 is unlearnable? Not necessarily. If the frequency of very long sequences is so low as to make it reasonable to suppose that the amount of time that has passed is insufficient to have expected to have heard really long forms, one might conclude that they simply hadn't been heard yet.

[4] See Jelinek (1997).

[5] We will look more closely at the mathematical underpinnings of these ideas in chapter 10.

[6] Cite Dunning paper?

Statistical models can also be used as part of other fairly traditional computational-linguistic tasks, e.g. tagging and parsing. For example, in the tagging domain, one can identify more readily the part of speech of some token by consideration of how frequently its possible tags occur in some statistical context. Likewise, similar factors can guide the parsing of different structural possibilities. We will treat these topics in chapters 10.2 and 9.

There are a huge number of other applications.

1.4 What follows

In the remainder of this text, we survey some of the basic techniques used in these domains: n-gram models, Hidden Markov Models, probabilistic context-free grammars, etc.

These are not incredibly complex notions, but they require some formal background to make them accessible. We therefore treat basic formal language theory and probability theory first.[7]

Finally, we consider the ontological status of these computational methods. Are they—or can they be—theories of language? Or are they simply convenient techniques with no import for linguistic theory?

The text is accompanied by a set of programs available on my website.[8]

[7] I will assume the reader is familiar with the basics of linguistic theory.

[8] `http://www.u.arizona.edu/~hammond`