LING439/539 - Statistical NLP

Chapter 1. Introduction

Thursday, August 25 2016

Foundation of Statistical NLP

Foundation of Statistical Natural Language Processing by Chistopher D. **Manning** and Hinrich **Schütze**.

Statistical NLP and Frequency Distributions

”Statistical considerations are essential to an understanding of the operation and development of languages” (Lyons 1968, Introduction to Theoretical Linguistics).

Overview

- Linguistic Science aims to characterize and explain the multitude of linguistic phenomena we observe in conversation, writing, and other media
  - Cognitive aspects: how do humans acquire, use, and understand language ?
  - Grounding aspects: how do symbols (and language) reference the world ?
  - Language structures: what forms and structures do languages tend to take ?
- The last aspect (linguistic structures) is what we're concerned with here
  - Part of speech tagging, grammars, and other linguistic structures

Rules

Linguistic structures have historically been studied and characterized in terms of systems of rules:

- This has been the case for many hundreds of years
- This rule-based characterization became much more formal in the last decades as linguists specified detailed grammars
    - Task: determine whether sentences were well-formed

Well-formed:	*The puppy chased after a ball.*
Ill-formed:	*Chased puppy after the ball a.*

Rules are brittle

Rule-based formalisms can get very far, but they're still very rigid, and don't allow "wiggle-room" that describes how language is used in practice

- Semantic perspective: People produce utterances that have meaning for listeners which may be on the border of grammatical rules
- Linguistic creativity: Language is constantly evolving

"All grammars leak" (Sapir, 1921)

But there are clearly regularities:

- NP → (DET) NN *captures some of what's observed*
- NP → (DET) (ADJ*) NN *captures much of what's observed*
- NP → (DET) (ADJ*) NN (PP) *captures more*

Data-driven methods

- Re-frame the problem:
  - Originally: *What sentences are grammatical and ungrammatical ?*
  - Now: *What are the common patterns that occur when language is used ?*
- The major tool that we use to characterize patterns is **counting**, or more formally, **statistics**.
- Statistics is generally founded in probability theory, which is where we'll begin acquiring our theoretical background

Data-driven

We generally use *large* corpora to study language in use.

Rationalist vs. Empiricist Approaches

- An underlying theme in the study of language and cognition is the dichotomy between rationalism and empiricism
- Rationalism (1960s-1980s): A significant part of the knowledge in the human brain is not derived by sensory input, but is fixed in advance (presumably from genetics)
    - Chomsky, innate language facility
    - AI: Create intelligent systems by hand-coding then with knowledge and reasoning mechanisms from humans

Poverty of the stimulus by Chomsky

- It's difficult to understand how children can learn something as complex as language from simply observing the world and receiving both variable and noisy input
- Language facilities are innate: aspects of the brain are hard-wired at birth to be sensitive to language learning

Rationalist
The brain is hard wired to learn language

Tabula-Rasa
The brain is a complete ”blank slate”, and learns everything required from data

Middle Ground: Empiricism

Rationalist

The brain is hard wired to learn language

Empiricism

- Some very simple underlying cognitive abilities are present
- Some underlying sensitivity or preference to certain ways of organizing knowledge, and generalizing from sensory inputs
- The mind does not begin with detailed principles (morphological structure, case marking, etc.)

Tabula-Rasa

The brain is a complete "blank slate", and learns everything required from data

Empiricist Approach to NLP

- Specify a generic enough language model
- Induce the specific values of parameters by applying statistical analyses and machine learning methods on a large amount of linguistic data

In NLP, we generally need lots of data

Terminology

- Corpus: a large collection of texts
- Corpora: several such collections

Common Corpora for NLP

Brown Corpus

- Tagged corpus containing about 1M words
- Assembled at Brown University in the 60s and 70s
- Balanced:
  - Representative sample of American English (at the time)
  - Press, fiction, scientific/legal texts, etc

Penn Treebank

- Large corpus of syntactically annotated (parsed) sentences
- Largely from Wall Street Journal (WSJ) articles

WordNet:

- Large electronic dictionary of English
- Hierarchical: Taxonomic information
- Synsets: groups of words with identical (or nearly identical) meanings
- Some other kinds of relations (e.g. part-whole) included

Claude Shannon

- "The Father of Information Theory" - fundamental work on theory of communication, entropy, and signal processing.
- The approach we tend to use in Statistical NLP draws from the work of Shannon
  - Assign probabilities to linguistic events
  - Focus on sentences that are “usual” and “unusual”

Can language be modeled probabilistically ?

What is a ”word”?

- Simple definition: anything separated between whitespace
- But this has many issues (from *Tom Sawyer*):

424	Tom
84	Tom’s
80	Tom,
47	Tom.
29	”Tom,
18	Tom?”
16	Tom.”
13	Tom!”

- We normally perform **tokenization** to normalize the text and split it into linguistically meaningful units
  - After tokenization, we have *Tom* 762 times.

Word Frequencies

Brown Corpus:

- Over 1,000,000 word tokens
- 49,680 word types

Type vs Token:

- Type: a given word, independent of capitalization
  - e.g. “cat”
- Token: a specific instance of a word
  - e.g. “Cat”, “cat”, “CAT”, etc.

Frequency from *Tom Sawyer*

	Token		Type
4936	,	4936	,
3793	.	3793	.
3335	the	3711	the
3309	”	3309	”
2955	and	3095	and
1761	a	1830	a
1712	to	1715	to
1438	of	1446	of
1163	was	1316	it
1130	it	1253	he
	# of tokens = 7937		# of types = 7417

It occurs 186 times and *it* 1130 times.

Question:

What proportion of word types appear exactly once?

Text file: `http://tinyurl.com/gnjtovt`

Word Frequencies by Tokens

Word Frequencies by Tokens (≥ 100)

Word Frequency Distributions

Follow a power law distribution (long tail).

Brown Corpus:

- Over 1,000,000 work tokens
- 50,000 word types
- 22,000 word types occur only once (44.1%)

Corpus linguistics:

- A few very common words (often syntactic glue - functional words)
- A medium number of medium frequency words
- A large number of low frequency words

Brown corpus

freq	# of type	freq	# of type
1	21919	11	491
2	7191	12	431
3	3915	13	390
4	2461	14	318
5	1810	15	308
6	1274	16	316
7	1096	17	254
8	817	18	219
9	686	19	199
10	543	20	200

Zipf's Law

The frequency of a word is proportional to it's rank:

$$f \propto \frac{1}{r}$$

or, in other words:

There is a constant k such that $f \cdot r = k$.

Grounding this: the 50th most common word should occur with three times the frequency of the 150th most common word.

from *Tom Sawyer*:

```
...
  50th  237 would
...
 150th   77 last
```

The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correpond to the ranks and frequencies of the words in the Brown corpus. The line is the relationship between rand and frequency predicted by Zipf for $k = 100,000$, that is $f \times r = 100,000$.

Frequencies of different word types

Extremely frequent words

- Commonly determiners, prepositions, conjunctions, pronouns, some adverbs
- Punctuation
- Function words: convey the syntactic information in a sentence

Medium frequency words

- Common nouns, verbs, adjectives, and adverbs
- Content words: convey semantic information in a sentence

Low frequency words

- Also content words (nouns, verbs, adjectives, rare adverbs)
- New words, names, foreign words, numbers, etc.

Unknown words and Zipf's law

- Because of the highly-skewed distribution, many possible words don't appear in a corpus, and are therefore "unknown"
- Common words not found in the Brown corpus:
  - combustible, parabola, preprocess, deodorizer, ...
- Names of people and places, especially foreign names
- Domain-specific vocabulary: Science, medicine, etc.
- Neologisms (newly coined words)

We'll talk more about unknown words, and their consequences later.