Statistical Natural Language Processing

Introduction

- **Long sentences most often give rise to ambiguity when conventional grammars are used to process the same**
- **Processing may yield large no of analyses**
- **It is here statistical notion help to avoid/ resolve ambiguity**

Corpus

- ❑ **Corpus : Collection of written text or spoken words of language**

- ➢ **<u>Types of Corpus</u>**

- ▪ **Textual Corpus : Content of a complete book, newspaper, magazine, web pages, journals , speeches etc..**
- ▪ **Corpus of spoken words**
- ▪ **Corpus for a specific domain : Tourism , law etc..**
- ▪ **Annoted Corpus : Rather than being a collection of raw text some corpus contain extra information regarding their content**
- ▪ **Parallel Corpus : A collection of texts which have been translated into one or several other languages**

  - Use in language translation activities

Concordance and Collocation

- **Concordance : An index or list of important words in a text (how often a word occurs (frequency))**
- **Collocation : Collection of words observed together**

  e.g.

  - **Rakhi gifts**
  - **Chrimas gifts**
  - **Chain smoker**
  - **Chain pulling**
  - **Exteremely beautiful**

Counting the elements in a corpus

- It yields valuable information regarding the probability of the occurrence of a word
- Probability can be use to predict a word that will follow
- Issues:
  - Should the punctuation marks be treated as a word
  - Case sensitization (IN an in) and (books and book) singular , plural considered distinct one
- Types : The no of distinct words in the corpus
- Tokens : Total no of words in corpus

Counting the elements in a corpus

Sentence

The former means the no of distinct words in the corpus while the latter stands for the total number of words in the corpus

Types : 14

Tokens : 24

Why Statistical/Probabilistic language models?

- Assign a probability to a sentence
- Machine Translation:
  - P(high winds tonight) > P(large winds tonight)
- Spell Correction
  - The office is about fifteen <u>minuets</u> from my house

  P(about fifteen <u>minutes</u> from) > P(about fifteen <u>minuets</u> from)
- Speech Recognition
  - P(I saw a van) >> P(eyes awe of an)
- Summarization, question-answering, etc., etc.!!

Probabilistic Language Modeling

- **Compute the probability of a sentence or sequence of words**

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- **Probability of an upcoming word**

$$P(w_5 | w_1,w_2,w_3,w_4)$$

- **A model that computes either of these:**

$$P(W) \quad \text{or} \quad P(w_n | w_1,w_2 \ldots w_{n-1})$$

is called a language model

How to compute P(W)

- **How to compute this joint probability:**

  P(its, water, is, so, transparent, that)

- **Intuition: let's rely on the Chain Rule of Probability**
  - **Definition of conditional probabilities**

    $P(B|A) = P(A,B)/P(A)$

    Rewriting: $P(A,B) = P(A)P(B|A)$
  - **More variables**

    $P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$
  - **The Chain Rule in General**

$$P(x_1,x_2,x_3,\ldots\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots\ldots\ldots\ldots\ldots P(x_n|x_1,\ldots,x_{n-1})$$

Compute joint probability of words

- The Chain Rule applied to compute joint probability of words in sentence

  $P(\text{"its water is so transparent"})$

  $= P(\text{its}) \times P(\text{water} \mid \text{its}) \times P(\text{is} \mid \text{its water})$

  $\times P(\text{so} \mid \text{its water is}) \times P(\text{transparent} \mid \text{its water is so})$

How to estimate these probabilities

- **Could we just count and divide?**

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- **Not, possible computationally, two many possible sentences**

❑ **Markov Assumption : The probability of a word depends on the probability of a limited history**

❑ **Generalization : The probability of a word depends on the probability of n previous words**

Markov Assumption

P (the | its water is so transparent that)

≈ P (the | that)

≈ P (the | transparent that)

- Generalize Formula

$$P(w_1,w_2,w_3,\ldots\ldots,w_n) = \prod_i P(w_i|w_{i-k},\ldots,w_{i-1})$$

In other words,

$$P(w_i \mid w_1, w_{2,}\ldots,w_{i-1}) \approx P(w_i \mid w_{i-k},\ldots,w_{i-1})$$

N- gram Models and its Applications

- **It is about predicting the n^{th} word from n-1 words**
- **What would be the next word in the following sentence**

  He is going to __________
- **Here predicting 5^{th} word from previous 4 words so it is 5-gram**

➢ **<u>Applications</u>**

- **In OCR**
- **Correcting a sentence**
- **Speech Recognition**
- **In translation**

Simplest case: Unigram model

$$P(w_1,w_2,w_3,\ldots\ldots,w_n) \approx \prod_i P(w_i)$$

$$P(w_i \mid w_1, w_2,\ldots,w_{i-1}) = P(w_i)$$

- Some automatically generated sentences from a unigram model

thrift did eighty said ….(random sequence of words)

e.g.

This is a sentence

Unigrams: This,

is,

a,

sentence

Bigram model

$$P(w_1,w_2,w_3,\ldots\ldots,w_n) = \prod_i P(w_i|w_{i-1})$$

$$P(w_i \mid w_1, w_{2,}\ldots,w_{i-1}) \approx P(w_i|w_{i-1})$$

- Some automatically generated sentences from a Bigram model

outside new car parking lot of the agreement……..

e.g.

This is a sentence

Bigrams: This is,

is a,

a sentence

Estimating Bigram probabilities

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Example 1: Estimating bigram probabilities on toy corpus

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $\quad P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $\quad P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $\quad P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $\quad P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

How to check one sentence is more probable than other?

<s> I am Sam </s>
<s> Sam I am </s>

- <s> I am Sam </s>

P (I | <s>) * P(am | I)* P(Sam | am) * P (</s> | Sam)

= 2/3 * 2/3 * 1/2 * 1/2

= 1/9

- <s> Sam I am </s>

P (Sam | <s>) * P(I | Sam)* P(am | I) * P (</s> | am)

= 1/3 * 1/2 * 2/3 * 1/2

= 1/18

- I am Sam is more probable than Sam I am