Chapter 3: Basics of Language Modeling
Section 3.1.
Language Modeling in Automatic Speech Recognition (ASR)

All graphs in this section are from the book by Schuckat-Talamazzini unless indicated otherwise.
### Complexity of Speech Recognition Applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarmstoschalter</td>
<td>1</td>
</tr>
<tr>
<td>Menü-Steuerung (J/N)</td>
<td>2</td>
</tr>
<tr>
<td>Zahlen/Ziffern</td>
<td>10 + x</td>
</tr>
<tr>
<td>Gerätebedienung</td>
<td>20 – 200</td>
</tr>
<tr>
<td>Auskunftsdialo</td>
<td>500 – 2000</td>
</tr>
<tr>
<td>Alltagssprache</td>
<td>8000 – 20000</td>
</tr>
<tr>
<td>Diktiermaschine</td>
<td>20 000 – 50 000</td>
</tr>
<tr>
<td>Deutsch ohne Fremdwörter</td>
<td>ca. 300 000</td>
</tr>
</tbody>
</table>
What makes Speech Recognition difficult

Guten Morgen, Herr Hauptkommissar Thanner. Gibt es irgendetwas Neues im Fall “Verbmbil”?

Morgen, Thanner. Irgendwas Neues im Fall “Verbmbil”?

morgen thanner irgendet was neues im fall verbmbil

moanlanaignwasneuesifallverbmbiihl

der Text in “Schönschrift”

spontan gesprochene Sprache

Großschreibung? Satzzeichen?

kontinuierliche Sprache

Aussprachevarianten

artikulatorische Verschleifung

Störungen & Verzerrungen des akustischen Kanals

Überlagerung d. Fremdstimmen „Cocktailparty-Effekt”
Milestones in ASR

From: Schukat-Talamazzini: Spracherkennung

- DARPA SUR
  - digitale Filterung
  - HMM zur Spracherkennung Triphone
  - Lineare Vorhersage
  - Sprecherverifikation
  - schnelle DFT
  - Baum-Welch-Algorithmus
  - LPC-Metrik
  - DTW-Einzelworterkenner
  - Dynamische Zeitverzerrung (DTW)

- DARPA Speech & Natural Language
  - HARPY
  - IBM Tangora
  - stochastische $N$-Gramme
  - mel-Cepstrum
  - LBG-Algorithmus
  - Sprachsynthese vom Text
  - $n$ beste Wortketten
  - SPHINX
  - TDNN
  - Phonotopische Karte
  - Mikrofonfelder
  - backpropagation-Algorithmus

- Signalprozessoren

Introduction into ASR

Source: Steve Young: the HTK book
Basic Components of an ASR-System

Speech Signal → Feature Extraction → Search

\[ \hat{W} = \text{argmax } [P(A|W) \cdot P(W)] \]

Recognized word sequence \( \hat{W} \)

Acoustic Model: \( P(A|W) \)

Language Model: \( P(W) \)
The Markov Assumption

Cutting history to $M-1$ words:

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

Special cases:

- **Trigram**
  $$P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx P(w_i \mid w_{i-2}, w_{i-1})$$

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

- **Unigram**
  (uniform distribution)
  $$P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx \frac{1}{W}$$
Section 3.2.
A Measure for LM Quality: Preplexity
Motivation

• Speech recognition can be computationally expensive
• For research and development: need a simple evaluation metric for fast turn-around times in experiments
Test your Language Model

What’s in your hometown newspaper???
What’s in your hometown newspaper today
Test your Language Model

It’s raining cats and ???
Test your Language Model

It’s raining cats and dogs
Test your Language Model

President Bill ??
Test your Language Model

President Bill Gates
Definition of Perplexity

Let $w_1, w_2, \ldots, w_N$ be an independent test corpus not used during training.

Perplexity

$$PP = P(w_1, w_2, \ldots, w_N)^{-1/N}$$

Normalized probability of test corpus
Example: Zerogram Language Model

"Zerogram" (uniform distribution)

\[ P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx \frac{1}{W} \]

\[
PP = P(w_1, w_2, \ldots, w_N)^{-1/N} \\
= \left( \prod_{i=1}^{N} \frac{1}{W} \right)^{-1/N} \\
= W
\]

Interpretation: perplexity is the “average” de-facto size of vocabulary
Alternate Formulation

Assumption:
Use an M-gram language model
History $h_i = w_i - M + 1 \ldots w_i - 1$

Idea:
collapse all identical histories
Alternate Formulation

Example:

M=2

Corpus:

“to be or not to be”

h2=“to” w2=“be”

h6=“to” w6=“be”

a calculate P(“be”|”to”) only once and scale it by 2
Alternate Formulation

\[ PP = P(w_1, w_2, \ldots, w_N)^{-1/N} \]

\[ = \left( \prod_{i=1}^{N} P(\text{w}_{i} \mid h_i) \right)^{-1/N} \]

\[ = \exp \left( \log \left( \prod_{i=1}^{N} P(\text{w}_{i} \mid h_i) \right)^{-1/N} \right) \]

Use Bayes decomposition

Try to get rid of product
Alternate Formulation

\[
N(w,h): \text{absolute frequency of sequence } h,w \text{ on test corpus}
\]

\[
f(w,h) \text{ relative frequency of sequence } h,w \text{ on test corpus}
\]

\[
= \exp \left( -\frac{1}{N} \log \left( \prod_{i=1}^{N} P(w_i \mid h_i) \right) \right)
\]

\[
= \exp \left( -\frac{1}{N} \sum_{i=1}^{N} \log(P(w_i \mid h_i)) \right)
\]

\[
= \exp \left( -\sum_{w,h} f(w,h) \log(P(w \mid h)) \right)
\]
Alternate Formulation: Final Result

\[ PP = P(w_1 \ldots w_N)^{-1/N} = \exp \left( - \sum_{w,h} f(w, h) \log(P(w | h)) \right) \]
Alternate Formulation: Zerogram Example

"Zerogram" (uniform distribution) $P(w_i \mid w_1, w_2, \ldots w_{i-1}) \approx \frac{1}{W}$

$$PP = \exp\left(-\sum_{w,h} f(w,h) \log(P(w \mid h))\right)$$

$$= \exp\left(-\log\left(\frac{1}{W}\right) \sum_{w,h} f(w,h)\right)$$

Identical result

$$= \exp\left(-\sum_{w,h} f(w,h) \log\left(\frac{1}{W}\right)\right)$$

$$= \exp\left(\log(W) \times 1\right) = W$$
Perplexity and Error Rate

Perplexity is correlated to word error rate

Power law relation
Quality Measure: Mean Rank

• Definition:
  • Let w follow h in the test text
  • Sort all words after a given history according to $p(w_i|h)$
  • Determine position of correct word w
  • Average over all events in the test text
Alternative: Average Rank

We are to know the problems. We use have to do the resolve.

Alternative: Average Rank

issue: necessary
data: time
above: people
issues: operators
tools: old

next: mailroom
be: marketplace
issues: provision
be: reception
next: shop

next: important
next: meetings
be: months
issues: years
next: to weak
issues: days
next:
<table>
<thead>
<tr>
<th>Measure</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>0.955</td>
</tr>
<tr>
<td>Mean Rank</td>
<td>0.957</td>
</tr>
</tbody>
</table>
Quality Measure: worst Score

• Which fraction of the test text have a score worse than $S_c$
• Correlation: 0.919
Summary

• Speech recognition
  • Basic task
  • Language model as a component

• Perplexity:
  • Measures quality of language model