Language Model using Counting
–AI Language Technology

Dr. Tiansi Dong, Prof. Dr. Christian Bauckhage
dongt@bit.uni-bonn.de, christian.bauckhage@iais.fraunhofer.de

Feb., 2018
Target of this Lecture

- Understand language model based on N-gram
Target of this Lecture

- Understand language model based on N-gram
- Understand the concept of Perplexity
Target of this Lecture

- Understand language model based on N-gram
- Understand the concept of Perplexity
- Understand Smoothing techniques
What can we do to understand this language (a little bit)?

- Counting!
  - What can we count?
  - The number of vocabulary, how often do two (three, ...) words appear in sequence?
A language from Alien

What can we do to understand this language (a little bit)?

- Counting!
What can we do to understand this language (a little bit)?

- Counting!
- What can we count?
A language from Alien

- What can we do to understand this language (a little bit)?

- Counting!

- What can we count?

- The number of vocabulary, how often do two (three, ... ) words appear in sequence?
sentence $s$ is a sequence of $N$ words: $s = w_1 w_2 \ldots w_N$
The N-gram Language Model

- sentence $s$ is a sequence of $N$ words: $s = w_1 w_2 \ldots w_N$
- 1-gram model. Counting the number of $w_i$ appearing in $s$
The N-gram Language Model

- sentence $s$ is a sequence of $N$ words: $s = w_1 w_2 \ldots w_N$
- 1-gram model. Counting the number of $w_i$ appearing in $s$
- 2-gram model. Counting the number of $w_i w_{i+1}$ appearing in $s$
The N-gram Language Model

- Sentence $s$ is a sequence of $N$ words: $s = w_1 w_2 \ldots w_N$
- 1-gram model. Counting the number of $w_i$ appearing in $s$
- 2-gram model. Counting the number of $w_i w_{i+1}$ appearing in $s$
- 3-gram model. Counting the number of $w_i w_{i+1} w_{i+2}$ appearing in $s$
What is Perplexity?

- Perplexity is the measurement of language model
Perplexity is the measurement of language model quality.
What is Perplexity?

- Perplexity is the measurement of language model
- What does Perplexity measure?
- The quality of language model
Compute Perplexity: Example

- What we have is a train corpus: *this is the cat that killed the rat that ate the malt that lay in the house that jack built*
What we have is a train corpus: *this is the cat that killed the rat that ate the malt that lay in the house that jack built*

We want to measure, in terms of perplexity, a test sentence: *this is the house that jack built*
Compute Perplexity: Example - trigram

- a train corpus: \( s_1 \ s_2 \ this \ is \ the \ cat \ that \ killed \ the \ rat \ that \ ate \ the \ malt \ that \ lay \ in \ the \ house \ that \ jack \ built \ end \)
Compute Perplexity: Example - trigram

- a train corpus: \( s_1 \) \( s_2 \) this is the cat that killed the rat that ate the malt that lay in the house that jack built end

- Unigrams: this is the cat that killed rat ate malt lay in house jack built end

\[
P(\text{snt}_\text{test}) = \frac{1}{N} \prod_{i=1}^{N} p(w_i|w_{i-2}w_{i-1})
\]
Compute Perplexity: Example - trigram

- a train corpus: $s_1 \ s_2 \ this\ is\ the\ cat\ that\ killed\ the\ rat\ that\ ate\ the\ malt\ that\ lay\ in\ the\ house\ that\ jack\ built$ end
- Unigrams: this is the cat that killed rat ate malt lay in house jack built end
- Totally $V = 15$ unigrams. Why?
Compute Perplexity: Example - trigram

- a train corpus: \( s_1 \ s_2 \text{ this is the cat that killed the rat that ate the malt that lay in the house that jack built end } \)
- Unigrams: this is the cat that killed rat ate malt lay in house jack built end
- Totally \( V = 15 \) unigrams. Why?
- the test sentence \( snt_{test} \text{ “this is the house that jack built”} \) has \( N = 7 \) words, 10 symbols: “s_1 \ s_2 \text{ this is the house that jack built end}”
Compute Perplexity: Example - trigram

- a train corpus: \( s_1 \ s_2 \ \text{this is the cat that killed the rat that ate the malt that lay in the house that jack built end} \)
- Unigrams: this is the cat that killed rat ate malt lay in house jack built end
- Totally \( V = 15 \) unigrams. Why?
- the test sentence \( snt_{test} \) “this is the house that jack built” has \( N = 7 \) words, 10 symbols: “\( s_1 \ s_2 \ \text{this is the house that jack built end} \)”
- The perplexity of \( snt_{test} \)

\[
P(snt_{test}) = p(snt_{test})^{-\frac{1}{N}} = \left( \prod_{i=1}^{N+1} p(w_i|w_{i-2}w_{i-1}) \right)^{-\frac{1}{N}} \tag{1}
\]
The perplexity of $snt_{test}$

$$\mathcal{P}(snt_{test}) = p(snt_{test})^{-\frac{1}{N}} = \left(\prod_{i=1}^{N+1} p(w_i|w_{i-2}w_{i-1})\right)^{-\frac{1}{N}}$$

$$= \left(\prod_{i=1}^{N+1} \frac{C_{train}(w_{i-2}w_{i-1}w_i) + 1}{C_{train}(w_{i-2}w_{i-1}) + V}\right)^{-\frac{1}{N}}$$

$$= \left(\frac{C_{train}(s1 \ s2 \ this) + 1}{C_{train}(s2 \ this) + 15} \cdots\right)^{-\frac{1}{7}}$$

$$= \left(\frac{1 + 1}{1 + 15} \cdot \frac{0 + 1}{1 + 15}\right)^{-\frac{1}{7}}$$

$$= 11.89$$