Natural Language Processing

N-Gram Models

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se
Introduction

“Once upon a time, there was a . . .”

- How can we guess the next word?
- Estimate $P(w|\text{“Once upon a time, there was a”})$ for any $w$
- This is a probabilistic language model
Language Models

- Language models are crucial in many NLP applications.
- Example from speech recognition:
  
  \[
  \begin{align*}
  & she \text{ drank two beers} \\
  & she \text{ drank too beers} \\
  & she \text{ drank too deers}
  \end{align*}
  \]

- Other NLP applications that make use of language models:
  1. Statistical machine translation
  2. Part-of-speech tagging
  3. Spell checking
  4. Optical character recognition
Let $w_1, \ldots, w_n$ be an arbitrary sequence of words.

We can compute $P(w_1, \ldots, w_n)$ using the chain rule:

\[
P(w_1, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1, \ldots, w_{n-1})
\]

\[
= \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1})
\]

But how do we find $P(w_i|w_1, \ldots, w_{i-1})$?
Estimation

- In theory, we can estimate $P(w_i|w_1, \ldots, w_{i-1})$ from data:
  \[
  \hat{P}(w_i|w_1, \ldots, w_{i-1}) = \frac{f(w_1, \ldots, w_i)}{f(w_1, \ldots, w_{i-1})}
  \]

- In practice, this becomes infeasible as $i$ grows larger.
- With a vocabulary of 100,000 words, there are:
  
  - $10^5$ possible unigrams ($i = 1$)
  - $10^{10}$ possible bigrams ($i = 2$)
  - $10^{15}$ trigrams ($i = 3$)
  - $10^{20}$ 4-grams ($i = 4$)
  - …
N-Gram Models

- We have to make independence assumptions
- In an $n$-gram model, we assume:
  \[ P(w_i|w_1, \ldots w_{i-1}) = P(w_i|w_{i-n+1}, \ldots, w_{i-1}) \]
- Words are dependent only on $n-1$ preceding words
  1. Unigram ($n=1$): $P(w_i|w_1, \ldots, w_{i-1}) = P(w_i)$
  2. Bigram ($n=2$): $P(w_i|w_1, \ldots, w_{i-1}) = P(w_i|w_{i-1})$
  3. Trigram ($n=3$): $P(w_i|w_1, \ldots, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})$
Text Generation

1-gram model

1. your something
2. you she to offices the possible his of of his said sight, was laughing had.
3. white was not full meet old be to made, you no I. described that power he the,
   man, And,
4. was Captain That she point labyrinth now must be far from. door had the from again what almost result fill, for coming as a with made
5. his then a country-town by you ’ ago Men?
Text Generation

2-gram model

1. Then here is the mud-bank what you, and instantly, and two officers waiting at once more valuable as I asked.

2. I may place is her husband and illegal constraint and outstanding, not recognised shape of finding that your heart, for communication between this man.

3. Mrs. Toller knows I mean that I have done very heartily at the 11: That is a lad, his neighbour.

4. Then there has offered to its centre one left this case, upon me to violin-land, though the corner and hurried across the very large staples.

5. Holmes ran up by old-fashioned shutters of treachery to attend to which I thought I have a foreigner, too late Ezekiah Hopkins, with this rather cumbrous.
Text Generation

3-gram model

1. However, when last seen, but now I will leave no survivor from a solution by the Underground and hurried me into a bedroom, which boomed out every quarter of a brickish red.

2. ’I beg that you have ever done yet, among the trees and wayside hedges were just being lighted as we stepped from her imprudence in allowing this brute to trace some geese which were new to me.

3. Holmes had sat up in my uncle’s life, and that a woman.

4. James and his hand and at the open, and has seen, but there are a thousand details which seem to have been hanged on far slighter evidence, I thought of!

5. Mr. Windibank draws my interest every quarter and pays it over to him.
Text Generation

4-gram model

1. Seeing that his passion was becoming ungovernable, I left him and returned towards Hatherley Farm.

2. You will excuse me, said my wife, and in order to see whether the objections are fatal, or if he had been to the side from which I could see that two of them were of the war he fought in Jackson’s army, and afterwards from your gesture, that Miss Rucastle was perfectly happy, and that I can.

3. I rang the bell and called for the weekly county paper, which contained a verbatim account of the matter, but you do not see the point.

4. It hadn’t pulled up before she shot out of the window?

5. Why does fate play such tricks with poor, helpless worms?
Text Generation

5-gram model

1. But what is it you wish?
2. He was too good and kind to leave me so.
3. You may remember the old Persian saying, ’There is danger for him who taketh the tiger cub, and danger also for whoso snatches a delusion from a woman.’
4. The paper was made in Bohemia, I said.
5. You will observe, said Holmes, are you sure about this whistle and metallic sound?
Evaluation

- Higher $n$-grams capture more linguistic structure
- But higher $n$-grams also require more training data
- What is the optimal trade-off?
- And how can we evaluate models more exactly?
Quiz

How do we compute $P(I\ love\ you)$ in a bigram model?

1. $P(I)P(love)P(you)$
2. $P(I)P(love|I)P(you|love)$
3. $P(I)P(love|I)P(you|I\ love)$