UCT AI SOCIETY

INTRODUCTION TO LANGUAGE MODELLING: N-GRAM MODELS

Andrej Jovanović
contact.me.maddox@gmail.com
maddox-j.github.io

TODAY'S STRUCTURE

WHAT IS LANGUAGE MODELLING

N-Gram language models: theory and assumptions.

CODE-WITH-ME

Let's build.

FOOD FOR THOUGHT

What have we not considered?

WHAT IS A LANGUAGE MODEL?

IN THE MORNING, I DRINK SOME

____

FOR FUN, I ________

BUT! HOW DO WE TRANSFER THIS INTO SOME CONCRETE, MATHEMATICAL NOTION?

PROBABILITY ESTIMATION

$$P(\textit{the} \mid \textit{its water is so transparent that})$$

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

BREAKING IT DOWN WITH PROBABILITY THEORY

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

$$P(w_1, w_2, w_3, w_4, \ldots, w_n) = P(w_{1:n})$$

$$P(w_{1:n}) = P(w_1)P(w_{2:n}|w_1)$$

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_{3:n}|w_1, w_2)$$

$$\begin{aligned} P(w_{1:n}) &= P(w_1)P(w_2|w_1)P(w_3|w_{1:2})\ldots P(w_n|w_{1:n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_{1:k-1}) \end{aligned}$$

MARKOVIAN ASSUMPTIONS

GLOBAL MODEL

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2})\ldots P(w_n|w_{1:n-1})$$
$$= \prod_{k=1}^{n} P(w_k|w_{1:k-1})$$

FIX A HISTORY

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-N+1:n-1})$$

BI-GRAM MODEL

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$$

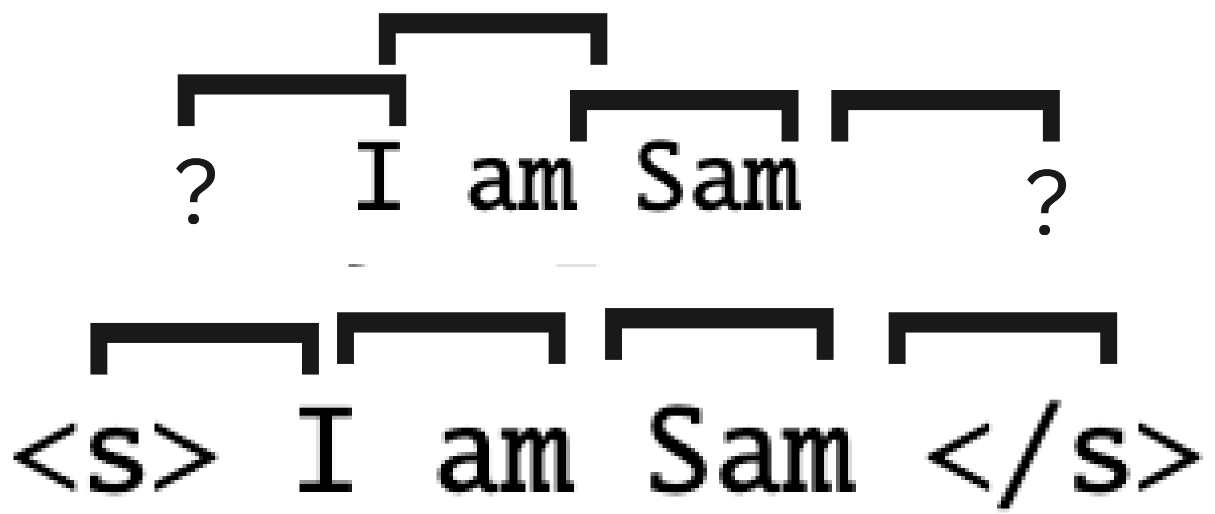
$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

MAXIMUM LIKELIHOOD ESTIMATION

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

PADDING!



WORKED EXAMPLE [1]

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 3.2 Bigram probabilities for eight words in the Berkeley Restaurant Project corpus of 9332 sentences. Zero probabilities are in gray.

Here are a few other useful probabilities:

$P(\texttt{i}|\texttt{<s>}) = 0.25$ $P(\texttt{english}|\texttt{want}) = 0.0011$
$P(\texttt{food}|\texttt{english}) = 0.5$ $P(\texttt{</s>}|\texttt{food}) = 0.68$

$$\begin{aligned} P(\texttt{<s> i want english food </s>}) &= P(\texttt{i}|\texttt{<s>})P(\texttt{want}|\texttt{i})P(\texttt{english}|\texttt{want}) \\ &\quad P(\texttt{food}|\texttt{english})P(\texttt{</s>}|\texttt{food}) \\ &= .25 \times .33 \times .0011 \times 0.5 \times 0.68 \\ &= .000031 \end{aligned}$$

LET'S DO SOME CODING!

OTHER THINGS TO CONSIDER?

- OOV WORDS?

$$P(w_{1:4}) = P(w_1| < s >)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3)P(< /s > |w_4)$$

$$P(w_{1:4}) = a * b * c * 0 * e = 0$$

- NUMERICAL STABILITY

$$logP(w_{1:2}) = logP(w_1| < s >) + logP(w_2|w_1) + logP(< /s > |w_2)$$

REFERENCES

[1] Speech and Language Processing (3rd ed. draft) – Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/