Noisy Channel Model and N Gram

27 November 2024 18:53

"observed data is a noisy version of the intended (original) message"

language model

Symbol Sequence → Noisy Encoding → o/p Sequence

$P(Y)$ → task independent

$P(X|Y)$ → depends on task

$P(X)$

Distribution of words that user intended to type

what user is likely to type given what they meant (noise)

Resulting distrib we actually see.

$$\text{Argmax}_Y \; P(X|Y)\,P(Y)$$

But how do you calculate the prob. of a string of words?

P("its water is so transparent") =
P(transparent) × P(transparent|so) × P(transparent|is so)
× P(transparent|water is so)
× P(transparent|its water is so)

Too much. → Simplify w/ Markov

$$P(\text{the}\,|\,\text{its water is so transparent that}) \approx P(\text{the}\,|\,\text{transparent that})$$

n-gram language models

look back (n-1) words.

Q. Why this start and end symbols are handy ?
If I do not have this, we will have $P(w_1)$ and $P(W_1/W_{1-1})$ which can have any value like 0 and we have to address that by smoothing. With Start and End, we have $P(w_1/\text{START})$ and $P(\text{END}/W_n)$ which are unlikely to have zero value

Bigram MLE (look back 1 word):

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

Eg1:

P(<s> I want chinese food </s>)
= P(I|<s>) x P(want|I) x P(chinese|want) x P(food|chinese) x P(</s>|food)

Bigram = pair of consecutive words.

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

no. of times "to spend" showed up in 9222 sentences.

Too big. Normalize by counts.

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

5/2533

	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

P(chinese|want)

"to food" has never occurred.

This is great and all, but what about long term dependencies?

- Syntactic dependencies
  - "The *man* next to the large oak tree near the grocery store on the corner is tall."
  - "The *men* next to the large oak tree near the grocery store on the corner are tall."

- Syntactic dependencies
  - "The ***man*** next to the large oak tree near the grocery store on the corner is tall."
  - "The ***men*** next to the large oak tree near the grocery store on the corner are tall."
- Semantic dependencies
  - "The ***bird*** next to the large oak tree near the grocery store on the corner flies rapidly."
  - "The ***man*** next to the large oak tree near the grocery store on the corner talks rapidly."