CS510: Advanced Topics in Information Retrival **Fall 2017**

Assignment 3

Kevin Li(zhuol2) *Handed In: September 22, 2017*

1. N-Gram Language Model

   a. Each word depends only on the past $n-1$ words.

   b. This assumption is reasonable because this is exactly the approximation for languages. When you see "For ex", the next letter is "a" with higher probability because it depends on these words we found in the past few positions. But it is independent with what we have before that.

   c. It massively simplifies the problem of estimating the language model from data.

2. Unigram Language Model

   a. Because

$$\begin{aligned} & p(the\ sun\ rises\ in\ the\ east\ and\ sets\ in\ the\ west) \\ = & p(the)p(sun)p(rises)p(in)p(the) \\ & p(east)p(and)p(sets)p(in)p(the)p(west) \end{aligned}$$

   there are 7 unique parameters, as there are 8 in total but probabilities sum up as 1. Their estimated values can be

$$p(the) = 3/11,\ p(sun) = 1/11,\ p(rises) = 1/11,\ p(in) = 2/11$$

$$p(east) = 1/11,\ p(and) = 1/11,\ p(sets) = 1/11,\ p(west) = 1/11$$

   b. There will be 11 parameters, as there are 12 in total but probabilities sum up as 1. Their estimated values can be

$$p(the) = 3/11,\ p(sun) = 1/11,\ p(rises) = 1/11,\ p(in) = 2/11$$

$$p(east) = 1/11,\ p(and) = 1/11,\ p(sets) = 1/11,\ p(west) = 1/11$$

$$p(a) = 0,\ p(from) = 0,\ p(retrieval) = 0,\ p(BM25) = 0$$

3. Bigram Language Model

   a. Because

$$\begin{aligned} & p(the\ sun\ rises\ in\ the\ east\ and\ sets\ in\ the\ west) \\ = & p(the)p(sun|the)p(rises|sun)p(in|rises)p(the|in) \\ & p(east|the)p(and|east)p(sets|and)p(in|sets)p(the|in)p(west|the) \end{aligned}$$

there are $12 \times 12 + 12 = 156$ unique parameters, because we should consider all the pair combination as conditional probabilities in the vocabulary. According to our estimator we can have

$$\begin{aligned}
p(sun|the) &= 1/3 \\
p(the|in) &= 1 \\
p(from|the) &= 0 \\
p(BM25|retrieval) &\text{ can not be estimated from the document} \\
p(east|west) &= 0 \\
p(east|rises) &= 0 \\
p(sets|\#) &= 0 \\
p(the|\#) &= 1
\end{aligned}$$

b.
- The problem is many parameters will have 0 probabilities because they do not appear in the document. This is not for estimation as they do have chances to appear.
- It occurs in 2.b as well.
- For bigram model, the problem is more severe, because the bigram model has more parameters with same vocabulary, thus leading to more 0 probabilities.
- The solution is to backoff to shorter N-grams, eventually to unigram. (i.e. smoothing)

4. Smoothing

a. Dirichlet Prior Smoothing

$$p(w|d) = \frac{c(w,d) + \mu p(w|REF)}{|d| + \mu} = \frac{|d|}{|d| + \mu}\frac{c(w,d)}{|d|} + \frac{\mu}{|d| + \mu}p(w|REF)$$

We can then prove the two lemmas

- If $|d| \to \infty$, we have

  $$\frac{|d|}{|d| + \mu} \to 1$$

  and

  $$\frac{\mu}{|d| + \mu} \to 0$$

  So

  $$\lim_{|d| \to \infty} p(w|d) = \frac{c(w,d)}{|d|}$$

  The unigram language model smoothed with a Dirichlet prior becomes equivalent to one estimated using the maximum likelihood estimate.

- If $\mu \to \infty$, we have

  $$\frac{|d|}{|d| + \mu} \to 0$$

and

$$\frac{\mu}{|d|+\mu} \to 1$$

So

$$\lim_{\mu\to\infty} p(w|d) = p(w|REF)$$

The unigram language model smoothed with a Dirichlet prior becomes equivalent to the background language model used in the smoothing.

b. Jelinek-Mercer smoothing:

$$p(w|d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda p(w|REF)$$

Katz-Backoff smoothing:

$$p(w|d) = \frac{max(c(w,d)-\delta, 0) + \delta|d|_u p(w|REF)}{|d|}$$

The advantage that Jelinek-Mercer over Katz-Backoff is the performance when the word count is low. If $c(w,d) < \delta$, Katz-Backoff will ignore the ML estimation, which leads to inaccuracy.

5. Application of Smoothing

a. By Dirichlet Prior Smoothing and $\mu = 4$

$$\begin{aligned} p(w|d) &= \frac{c(w,d)+\mu p(w|REF)}{|d|+\mu} \\ &= \frac{|d|}{|d|+\mu}\frac{c(w,d)}{|d|} + \frac{\mu}{|d|+\mu}p(w|REF) \\ &= \frac{11}{11+4}\frac{c(w,d)}{|d|} + \frac{4}{4+11}p(w|REF) \\ &= \frac{11}{15}\frac{c(w,d)}{|d|} + \frac{4}{15}p(w|REF) \end{aligned}$$

Then we can have:

$$\begin{aligned} p(the) &= 11/15 \times 3/11 + 4/15 \times 0.17 = 0.2453 \\ p(sun) &= 11/15 \times 1/11 + 4/15 \times 0.05 = 0.08 \\ p(rises) &= 11/15 \times 1/11 + 4/15 \times 0.04 = 0.0773 \\ p(in) &= 11/15 \times 2/11 + 4/15 \times 0.16 = 0.176 \\ p(east) &= 11/15 \times 1/11 + 4/15 \times 0.02 = 0.072 \\ p(and) &= 11/15 \times 1/11 + 4/15 \times 0.16 = 0.1093 \\ p(sets) &= 11/15 \times 1/11 + 4/15 \times 0.04 = 0.0773 \\ p(west) &= 11/15 \times 1/11 + 4/15 \times 0.02 = 0.072 \end{aligned}$$

There are small changes with each probability. It depends on $p(w|REF)$ and our ML estimator for unigram model. If some word got higher probability in reference, we increase the result somehow (e.g. ”and”), otherwise decrease a bit.

b. For $\mu = 0.01$

$$p(w|d) = (11c(w, d) + 0.01p(w|REF))/11.01$$

Then

$$\begin{aligned}
p(the) &= (11 \times 3/11 + 0.01 \times 0.17)/11.01 = 0.273 \\
p(sun) &= (11 \times 1/11 + 0.01 \times 0.05)/11.01 = 0.091 \\
p(rises) &= (11 \times 1/11 + 0.01 \times 0.04)/11.01 = 0.091 \\
p(in) &= (11 \times 2/11 + 0.01 \times 0.16)/11.01 = 0.182 \\
p(east) &= (11 \times 1/11 + 0.01 \times 0.02)/11.01 = 0.091 \\
p(and) &= (11 \times 1/11 + 0.01 \times 0.16)/11.01 = 0.091 \\
p(sets) &= (11 \times 1/11 + 0.01 \times 0.04)/11.01 = 0.091 \\
p(west) &= (11 \times 1/11 + 0.01 \times 0.02)/11.01 = 0.091
\end{aligned}$$

For $\mu = 100$

$$p(w|d) = (11c(w, d) + 100p(w|REF))/111$$

Then

$$\begin{aligned}
p(the) &= (11 \times 3/11 + 100 \times 0.17)/111 = 0.180 \\
p(sun) &= (11 \times 1/11 + 100 \times 0.05)/111 = 0.054 \\
p(rises) &= (11 \times 1/11 + 100 \times 0.04)/111 = 0.045 \\
p(in) &= (11 \times 2/11 + 100 \times 0.16)/111 = 0.162 \\
p(east) &= (11 \times 1/11 + 100 \times 0.02)/111 = 0.027 \\
p(and) &= (11 \times 1/11 + 100 \times 0.16)/111 = 0.153 \\
p(sets) &= (11 \times 1/11 + 100 \times 0.04)/111 = 0.045 \\
p(west) &= (11 \times 1/11 + 100 \times 0.02)/111 = 0.027
\end{aligned}$$

The results match my intuition as in 4.a. When $\mu \to 0$, the result becomes equivalent to one estimated using the maximum likelihood estimate. When $\mu \to \infty$, the result becomes equivalent to the background language model used in the smoothing.

c. We have Jelinek-Mercer smoothing as

$$p(w|d) = (1 - \lambda)\frac{c(w, d)}{|d|} + \lambda p(w|REF)$$

So for $\lambda = 0.01$

$$\begin{aligned}
p(the) &= 0.99 \times 3/11 + 0.01 \times 0.17 = 0.272 \\
p(sun) &= 0.99 \times 1/11 + 0.01 \times 0.05 = 0.091 \\
p(rises) &= 0.99 \times 1/11 + 0.01 \times 0.04 = 0.090 \\
p(in) &= 0.99 \times 2/11 + 0.01 \times 0.16 = 0.182 \\
p(east) &= 0.99 \times 1/11 + 0.01 \times 0.02 = 0.090 \\
p(and) &= 0.99 \times 1/11 + 0.01 \times 0.16 = 0.092 \\
p(sets) &= 0.99 \times 1/11 + 0.01 \times 0.04 = 0.090 \\
p(west) &= 0.99 \times 1/11 + 0.01 \times 0.02 = 0.090
\end{aligned}$$

So for $\lambda = 0.5$

$$\begin{aligned}
p(the) &= 0.5 \times 3/11 + 0.5 \times 0.17 = 0.221\\
p(sun) &= 0.5 \times 1/11 + 0.5 \times 0.05 = 0.070\\
p(rises) &= 0.5 \times 1/11 + 0.5 \times 0.04 = 0.065\\
p(in) &= 0.5 \times 2/11 + 0.5 \times 0.16 = 0.170\\
p(east) &= 0.5 \times 1/11 + 0.5 \times 0.02 = 0.065\\
p(and) &= 0.5 \times 1/11 + 0.5 \times 0.16 = 0.125\\
p(sets) &= 0.5 \times 1/11 + 0.5 \times 0.04 = 0.065\\
p(west) &= 0.5 \times 1/11 + 0.5 \times 0.02 = 0.055
\end{aligned}$$

So for $\lambda = 0.9$

$$\begin{aligned}
p(the) &= 0.1 \times 3/11 + 0.9 \times 0.17 = 0.180\\
p(sun) &= 0.1 \times 1/11 + 0.9 \times 0.05 = 0.054\\
p(rises) &= 0.1 \times 1/11 + 0.9 \times 0.04 = 0.045\\
p(in) &= 0.1 \times 2/11 + 0.9 \times 0.16 = 0.162\\
p(east) &= 0.1 \times 1/11 + 0.9 \times 0.02 = 0.027\\
p(and) &= 0.1 \times 1/11 + 0.9 \times 0.16 = 0.153\\
p(sets) &= 0.1 \times 1/11 + 0.9 \times 0.04 = 0.045\\
p(west) &= 0.1 \times 1/11 + 0.9 \times 0.02 = 0.027
\end{aligned}$$

From the results we can find:

- For larger λ, the value of parameters depends more on the reference. While the smaller the λ, the value depends more on ML estimation.
- Comparing with 5.a and 5.b, we find the relativity among parameters are the same. A parameter is larger in this smoothing method is also larger in the other smoothing method. Different smoothing method doesn't change the relative relations among parameters, but only got differences on values.