Recitation notes on Kneser Ney

Kartik Goyal

September 5, 2016

Abstract

1 Notation

• Given a sequence/sentence \(w_1 w_2 w_3 \ldots w_n\), \(w_i^j\) refers to the substring \(w_i \ldots w_j\), \(\forall \ i < j\). Also \(c(w_i^j)\) refers to the count of the substring \(w_i^j\) in the corpus.

• Given vocabulary \(V\), its size is denoted by \(|V|\) and \(\sum_w\) is a shorthand for \(\sum_{w \in V}\).

• \(N_1 + (\bullet w_i^j) = |\{w_{i-1} : c(w_{i-1}^j) > 0\}|\)

• \(N_1 + (w_i^j \bullet) = |\{w_{j+1} : c(w_{i}^{j+1}) > 0\}|\)

• \(N_1 + (\bullet w_i^j \bullet) = |\{(w_{i-1}, w_{j+1}) : c(w_i^{j+1}) > 0\}| = \sum_{w_{j+1}} N_1 + (\bullet w_i^{j+1})\)

• \((a)_+ = \max(a, 0)\)

2 Basic equation

For N-gram KN models,

\[
 p_{KN}(w_i|w_{i-n+1}^{i-1}) = \frac{(c'(w_{i-n+1}^i) - D)_+}{\sum_{w_i} c'(w_{i-n+1}^i)} + \alpha(w_{i-n+1}^i)p_{KN}(w_i|w_{i-n+1}^{i-2}) \tag{1}
\]

Hence, for bigram KN models,

\[
 p_{KN}(w_i|w_{i-1}) = \frac{(c'(w_{i-1}^i) - D)_+}{\sum_{w_i} c'(w_{i-1}^i)} + \alpha(w_{i-1})p_{KN}(w_i) \tag{2}
\]

D is to be treated as a hyperparameter which is typically < 1. For the highest order model, \(c'(w) = c(w)\). KN expression for lower order models is elaborated...
in the latter sections. For lower order modes, the \( c(w_{i-n+1}^{i-1}) \), is replaced by a value dependent upon the fertility of the relevant ngram. The lowest level of recursion is the unigram and its expression is:

\[
p_{KN}(w_i) = \frac{N_{1+}(w_i)}{N_{1+}(\bullet \bullet)}
\]  

(3)

3 Deriving \( \alpha \)

If \( c(w_{i-n+1}^{i-1}) > 0 \), then using the fact the \( \sum_{w_i} p_{KN}(w_i|\text{context}) = 1 \), we sum the LHS and RHS of eqn 1 over the whole vocabulary:

\[
\sum_{w_i} p_{KN}(w_i|w_{i-n+1}^{i-1}) = \sum_{w_i} \frac{(c(w_{i-n+1}^{i-1}) - D)+}{c(w_{i-n+1}^{i-1})} + \alpha(w_{i-n+1}^{i-1}) \sum_{w_i} p_{KN}(w_i|w_{i-n+2}^{i-1})
\]

which is equal to

\[
1 = \sum_{w_i; c(w_{i-n+1}^{i-1}) > D} \frac{(c(w_{i-n+1}^{i-1}) - D)+}{c(w_{i-n+1}^{i-1})} \sum_{w_i; c(w_{i-n+1}^{i-1}) > D} \frac{D}{c(w_{i-n+1}^{i-1})} + \alpha(w_{i-n+1}^{i-1})
\]

Since, we are working with \( 0 < D < 1 \) and the counts \( c \) are integers, we can write the above expression as:

\[
1 = \sum_{w_i} \frac{(c(w_{i-n+1}^{i-1})}{c(w_{i-n+1}^{i-1})} - \frac{D}{c(w_{i-n+1}^{i-1})} N_{1+}(w_{i-n+1}^{i-1}\bullet) + \alpha(w_{i-n+1}^{i-1})
\]

which leads us to:

\[
1 = 1 - \frac{D}{c(w_{i-n+1}^{i-1})} N_{1+}(w_{i-n+1}^{i-1}\bullet) + \alpha(w_{i-n+1}^{i-1})
\]

giving the final expression:

\[
\alpha(w_{i-n+1}^{i-1}) = \frac{D}{c(w_{i-n+1}^{i-1})} N_{1+}(w_{i-n+1}^{i-1}\bullet)
\]

4 Edge Cases

- If \( c(w_{i-n+1}^{i-1}) = 0 \), then the first expression in the RHS of eqn 1 is undefined. In this case when the context is not at all present in the corpus, keep on backing of completely to the lower order KN models till you come across a context with non-zero counts.

- If a new type \( w_i \) is seen, then you have two options, either return a zero probability or back off to a uniform model that returns smoothed \( \frac{1}{|V|} \). Generally, the first option is often implemented.
• For the lower order KN models we define $c'(w_{i-n+1}^i)$ in equation 1 differently i.e. for this case, $c'(w_{i-n+1}^i) = N_{1+}(w_{i-n+1}^i)$, hence the expression for lower order KN models is:

$$p_{KN}(w_i | w_{i-n+2}^i) = \frac{(N_{1+}(w_{i-n+2}^i) - D)_{+}}{\sum_{w} N_{1+}(w_{i-n+2}^i)} + \alpha(w_{i-n+2}^i) p_{KN}(w_i | w_{i-n+3}^i)$$

This gives us exactly the same expression as the one in the lecture slides.

References