Homework 2 - HMMs

Manasa Bollavaram

March 18, 2017

1 Bigram HMMs

1.1 Pre-processing data:

Data is pre-processed before employing it to build the language model. Words of each tweet containing special characters, numbers, combination of both are replaced by special words such as words containing # are replaced by HASHTAG, words containing combination of alphabets and numbers are replaced by ALPHANUMERIC. Special words used to replace the existing words in the tweets are ALPHANUMERIC, WEBADRESS, USERNAME, HASHTAG, PUNCTUATION, NUMBERS, EMPTY.

1.2 OOV Handling:

After pre-processing the training corpus, the word vocabulary count turned out tu be 43002.63% of the vocabulary has words with the frequency of occurence 2 or 1. These words account to only 3.5% of total occurences (total occurences = 729778 and these words occurences =25613). 80% of these words are selected to UNK.

1.3 Smoothing:

Linear interpolation is used for computing transition probabilities and addk smoothing is used to calculate emission probabilities. Parameters used for linear interpolation are $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$ where λ_1 is the unigram probability coefficient and λ_2 is the bigram probability coefficient.From homework 1, a trend of decreasing perplexity is observed when λ_2 is higher and therefore a value of 0.7 is selected. $K = 0.001$ is used for addk smoothing. As seen in homework 1, perplexity usually decreases with smaller magnitude of k and hence 0.001 is selected for smoothing.

1.4 Model performance:

Performance measures for dev and test sets using above smoothing parameters and techniques are
Model performance on $devset$=92.09

Model performance on *testset*=90.34

1.5 Error analysis:

Examples:

1. **Pre-processed tweet:** rt USERNAME : i'm screaming emoticon WEBADRESS
Given tags:[@ L V , U]
Predicted tags:[@ L V **G** U]
Emoticon is incorrectly tagged. Emoticons are not pre-processed by scanning for their special endcoded texts and replacing with special words. Replacing with a string category would reduce the inaccuracy.

2.**Pre-processed tweet:** what the heck even is ALPHANUMERIC
Given tags:[O D N R V A]
Predicted tags:[O D N R V **$**]
Adjective is incorrectly tagged as numeral.

3.**Pre-processed tweet:** seventh individual charged by serious HASHTAG office in investigation into alstom and alleged HASHTAG WEBADRESS via USERNAME
Given tags:[A A V P A # N P N P &̂ A N U P @]
Predicted tags:[A N V P A **N** N P N P &̂ **A** # U P @]
Hashtag is incorrectly tagged as common noun and vice-versa.

4.**Pre-processed tweet:**rt USERNAME : PUNCTUATION USERNAME PUNCTUATION and more in today's top casting notices PUNCTUATION WEBADRESS WEBADRESS emoticon
Given tags:[@ , @ , , A A N , , & R P S A V N , U U ,]
Predicted tags:[@ , @ , , A A N , , & **A** P S A V N , U U ,]
Adverb is incorrectly tagged as pre or post-position.

Replacing special words with the categories such as HASHTAG, ALPHANUMERIC results in incorrect tagging, coming up with better categories to replace the words such that the tagger identifies the role of the word in the sentence would help in better tagging.

2 Trigram HMMs

2.1 Formal equation for joint probability model

$$\begin{aligned}
p(\mathbf{s}, \mathbf{y}) &= p(x_1, x_2, ...x_n, y_1, ...y_n, y_{n+1}) \\
&= q(y_{n+1} = \text{STOP}|y_{n-1}, y_n) q(y_1|y_{-1} = \text{START}, y_0 = \text{START}) e(x_1|y_1) \prod_{i=2}^{n} q(y_i|y_{i-2}, y_{i-1}) e(x_i|y_i)
\end{aligned}$$

where $y_{n+1} = \text{STOP}$, $y_0 = \text{START}$ and $y_{-1} = \text{START}$; $q(...)$ is the transition probability and $e(...)$ is the emission probability.

2.2 Viterbi Decoding Algorithm

Probability of a sequence $1, 2, .., k$ in the given sequence of $1, 2, ..n$ is

$$\begin{aligned}
p(x_1, ..x_i, y_1, ..y_i) &= q(y_i|y_{i-2}, y_{i-1}) e(x_i|y_i) p(x_1, ..x_{i-1}, y_1, ..y_{i-1}) \\
&= \frac{q(y_i, y_{i-1}|y_{i-2})}{q(y_{i-1}|y_{i-2})} e(x_i|y_i) q(y_{i-1}|y_{i-2}, y_{i-3}) e(x_{i-1}|y_{i-1}) p(x_1, ..x_{i-2}, y_1, ..y_{i-2})
\end{aligned} \tag{1}$$

Maximizing the probability for sequence $1, 2, ...i-2$ can be done as following:

$$\begin{aligned}
\pi(i, y_i, y_{i-1}) &= \max_{y_1, y_2..y_{i-2}} \frac{q(y_i, y_{i-1}|y_{i-2})}{q(y_{i-1}|y_{i-2})} e(x_i|y_i) p(x_1, ..x_{i-1}, y_1, ..y_{i-1}) \\
&= \max_{y_{i-2}} \frac{q(y_i, y_{i-1}|y_{i-2})}{q(y_{i-1}|y_{i-2})} e(x_i|y_i) \max_{y_1, y_2..y_{i-3}} \frac{q(y_{i-1}, y_{i-2}|y_{i-3})}{q(y_{i-2}|y_{i-3})} e(x_{i-2}|y_{i-2}) p(x_1, ..x_{i-2}, y_1, ..y_{i-2}) \\
&= \max_{y_{i-2}} \frac{q(y_i, y_{i-1}|y_{i-2})}{q(y_{i-1}|y_{i-2})} e(x_i|y_i) \pi(i-1, y_{i-1}, y_{i-2})
\end{aligned}$$

Recursive viterbi algorithm developed using above expression is:

Base case:

$$\begin{aligned}
\pi(1, y_1, y_0) &= q(y_1|y_{-1} = \text{START}, y_0 = \text{START}) e(x_1|y_1) \\
\text{backpointer}(1, y_1, y_0) &= -1 \\
\pi(2, y_2, y_1) &= q(y_2|y_0 = \text{START}, y_1) e(x_i|y_i) \\
\text{backpointer}(2, y_2, y_1) &= -1
\end{aligned}$$

For $i = 3...n$

$$\begin{aligned}
\pi(i, y_i, y_{i-1}) &= \max_{y_{i-2}} q(y_i|y_{i-2}, y_{i-1}) e(x_i|y_i) \pi(i-1, y_{i-1}, y_{i-2}) \\
\text{backpointer}(i, y, y_{i-1}) &= \arg\max_{y_{i-2}} q(y_i|y_{i-2}, y_{i-1}) e(x_i|y_i) \pi(i-1, y_{i-1}, y_{i-2})
\end{aligned}$$

Termination Step

$$p(\mathbf{s}, \mathbf{y}) = \max_{y_{n-1}, y_n} q(y_{n+1} = \text{STOP}|y_{n-1}, y_n)\pi(n, y_n, y_{n-1})$$
$$\text{backpointer}(n+1, y_n, y_{n-1}) = \arg\max_{y_{n-1}, y_n} q(y_{n+1} = \text{STOP}|y_{n-1}, y_n)\pi(n, y_n, y_{n-1})$$

where backpointers help to backtrack and retrieve the maximum probable sequence.

2.3 Smoothing and OOV Handling

Corpus is pre-processed using the same techniques as for bigram and UNKed the same way. Linear interpolation is employed for transition probabilities and addk smoothing for emission probabilities.$\lambda_1 = 0.2$,$\lambda_2 = 0.4$ and $\lambda_3 = 0.4$ are used for linear interpolation and $K = 0.001$ is used for addk smoothing.

2.4 Performance measure

Performance measure using the design decisions and smoothing parameters described above are provided below:

Model performance on *devset*=91.45
Model performance on *testset*=89.23

2.5 Error analysis

Examples:

1. **Pre-processed tweet:** rt USERNAME : i'm screaming emoticon WEBADRESS
Given tags:[@ L V , U]
Predicted tags:[@ L V **G** U]

2.**Pre-processed tweet:** what the heck even is ALPHANUMERIC
Given tags:[O D N R V A]
Predicted tags:[O D N R V **$**]

3.**Pre-processed tweet:** seventh individual charged by serious HASHTAG office in investigation into alstom and alleged HASHTAG WEBADRESS via USERNAME
Given tags:[A A V P A # N P N P &̂ A N U P @]
Predicted tags:[A N V P A **#** N P N P &̂ ˆ # U P @]

4.**Pre-processed tweet:**rt USERNAME : PUNCTUATION USERNAME PUNCTUATION and more in today's top casting notices(tag wrong) PUNCTUATION WEBADRESS WEBADRESS emoticon

Given tags:[@ , @ , , A A N , , & R P S A V N , U U ,]
Predicted tags:[@ , @ , , A A N , , & **A** P S A V N , U U ,]

Incorrect tagging is same as for the bigram except for the third tweet. Larger sentences trigram tends to give accurate tags compared to bigram.