NLP - Spring 2020 - Sample Midterm Exam

(1) You are given the following training data for the prepositional phrase (PP) attachment task.

v	n1	p	n2	Attachment
join	board	as	director	V
is	chairman	of	N.V.	N
using	crocidolite	in	filters	V
⋮	⋮	⋮	⋮	⋮

Where the attachment value of V indicates that p attaches to v and the attachment value of N indicates that p attaches to $n1$.

In order to resolve PP attachment ambiguity we can train a probability model: $P(A = N \mid v, n1, p, n2)$ which predicts the attachment A as N if $P > 0.5$ and V otherwise.

a. (6pts) To define $P(A = N \mid v, n1, p, n2)$ using n-gram probabilities, since we are unlikely to see the same four words $v, n1, p, n2$ in novel unseen data, in order for this probability model to be useful we need to take care of zero counts.

Provide a Jelinek-Mercer style *interpolation* smoothing model $\hat{P}(A = N \mid v, n1, p, n2)$ for this PP attachment probability model. Do **not** use recursive interpolation in your solution.

Assume that our training data is large enough to contain all the prepositions we might observe in unseen data. You cannot assume that all the verbs and nouns in the unseen data were seen in training. Your solution must have a 4-gram and at least two trigrams, at least two bigrams and at least one unigram.

In the interpolation model if you use any new variables then provide the constraints that the variables must obey such that $\hat{P}$ continues to be a valid probability.

Answer:

$$\begin{aligned}
\hat{P}(A = N \mid v, n1, p, n2) \quad = \quad & \lambda_1 P(A = N \mid v, n1, p, n2) \\
& + \lambda_2 P(A = N \mid v, p, n2) \\
& + \lambda_3 P(A = N \mid n1, p, n2) \\
& + \lambda_4 P(A = N \mid v, p) \\
& + \lambda_5 P(A = N \mid n1, p) \\
& + \lambda_6 P(A = N \mid p)
\end{aligned}$$

To be a well-formed interpolation model, $\sum_i \lambda_i = 1$. There are many other solutions such as for instance, the 3-gram model could have different choices for the conditioning context, e.g. $v, n1, p$ instead of $v, p, n2$ and similarly for the bigrams. The solution must have at least one four-gram, two trigrams, two bigrams and one unigram (the preposition).

b. (4pts) Assume you are given **pre-trained** word embeddings for each of the words in the prepositional phrase dataset. The embeddings we get from the word appearing as the center or target word are: t_v, t_{n1}, t_p, t_{n2}. The embeddings we get from the word appearing in the context are: c_v, c_{n1}, c_p, c_{n2}.

Given a definition of $\hat{P}(A = N \mid v, n1, p, n2)$ we can define a classifier that predicts noun attachment if $\hat{P}(A = N \mid v, n1, p, n2) \geq 0.5$ and verb attachment if $1 - \hat{P}(A = N \mid v, n1, p, n2) > 0.5$ otherwise.

Your task is to use **only** the dot products $c_{n1} \cdot t_p$ and $c_v \cdot t_p$ in order to provide $\hat{P}(A = N \mid v, n1, p, n2)$. The goal is to use the pre-trained word vectors to provide the probability for noun/verb attachment.

Ignore the word embeddings for $n2$ (c_{n2} and t_{n2}) for this question. Do **not** sum over the entire vocabulary to obtain a probability distribution.

Answer:

$$\hat{P}(A = N \mid v, n1, p, n2) = \frac{\exp(c_{n1} \cdot t_p)}{\exp(c_{n1} \cdot t_p) + \exp(c_v \cdot t_p)}$$

(2) (3pts) You are given a text of words: $w_1, \ldots, w_N$ and you proceed to estimate bigram probabilities with the maximum likelihood estimate using the bigram frequencies $c(w_i, w_{i-1})$ and unigram frequencies $c(w_i)$:

$$\hat{P}(w_i \mid w_{i-1}) = \frac{c(w_i, w_{i-1})}{c(w_{i-1})}$$

Write down the definition of a backoff smoothed distribution $P_{bo}(w_i \mid w_{i-1})$ where we backoff from the bigram probability $c(w_i, w_{i-1})$ to the unigram $P(w_i)$. using a new function $c^*(w_{i-1}, w_i)$ which uses the absolute discounting method and $\alpha(w_{i-1})$, where $\alpha(w_{i-1})$ is chosen to make sure that $P_{bo}(w_i \mid w_{i-1})$ is a proper probability:

$$\alpha(w_{i-1}) = 1 - \sum_{w_i} \frac{c^*(w_{i-1}, w_i)}{c(w_{i-1})}$$

Provide the definition of $P_{bo}(w_i \mid w_{i-1})$ and $c^*(w_{i-1}, w_i)$. Assume that $1 = \sum_{w_i} P(w_i)$.

Answer:

$$P_{bo}(w_i \mid w_{i-1}) = \begin{cases} \frac{c(w_{i-1}, w_i) - D}{c(w_{i-1})} & \text{if } c(w_{i-1}, w_i) > 0 \\ \alpha(w_{i-1}) P_{bo}(w_i) & \text{otherwise} \end{cases}$$

where D is set to some value less than one using held out set.

(3) (2pts) Consider embedding a sentence by summing the GloVe embeddings of each word in the sentence. Pick which statement is true from the list below about the embeddings of the sentences *we are winning* and *we are winning we are winning we are winning* (assume the words are in the vocabulary).

1. The embeddings of the two sentences are equal.
2. The embeddings of the two sentences are close when using Euclidean distance, but not cosine distance.
3. The embeddings of the two sentences are close when using cosine distance, but not Euclidean distance.
4. The embeddings of the two sentences are close when using cosine distance and Euclidean distance.
5. None of the above.

Answer: The true statement is: The embeddings of the two sentences are close when using cosine distance, but not Euclidean distance.

(4) We define a log-linear model that estimates a distribution $\Pr(s|w)$ where s is a sentiment label for a word, $s \in \{\text{positive}, \text{negative}\}$ and w is a word from the vocabulary V. The set V contains the words *amazing* and

horrible as well as additional words (so that $|V| > 2$). We want our log-linear model to specify the following probabilities (some probabilities are left undefined):

$$\begin{aligned}
\Pr(\text{positive} \mid \textit{amazing}) &= 0.9 \\
\Pr(\text{negative} \mid \textit{horrible}) &= 0.9 \\
\Pr(\text{positive} \mid w) &= 0.6 \text{ for any word } w \text{ other than } \textit{amazing}, \textit{horrible} \\
\Pr(\text{negative} \mid w) &= 0.4 \text{ for any word } w \text{ other than } \textit{amazing}, \textit{horrible}
\end{aligned}$$

The value for probabilities Pr(negative | *amazing*), Pr(positive | *horrible*) are left unspecified. It is assumed they are given values such that the following condition is satisfied for every $w \in V$:

$$\sum_s \Pr(s \mid w) = 1.0$$

You are given exactly four features: $f_1(w, s), \ldots, f_4(w, s)$:

$$\begin{aligned}
f_1(w, s) &= 1 \text{ if } w \text{ is } \textit{amazing} \text{ and } s \text{ is positive, 0 otherwise} \\
f_2(w, s) &= 1 \text{ if } w \text{ is } \textit{horrible} \text{ and } s \text{ is negative, 0 otherwise} \\
f_3(w, s) &= 1 \text{ if } w \notin \{\textit{amazing}, \textit{horrible}\} \text{ and } s \text{ is positive, 0 otherwise} \\
f_4(w, s) &= 1 \text{ if } w \notin \{\textit{amazing}, \textit{horrible}\} \text{ and } s \text{ is negative, 0 otherwise}
\end{aligned}$$

Hint: For some input (w', s') if for all k the value of $f_k(w', s') = 0$ then $\Pr(s'|w') = \frac{e^0}{Z}$ where Z is the normalization term.

a. (4pts) For your feature vector $\{f_1, f_2, f_3, f_4\}$ let the parameter vector be $\{v_1, v_2, v_3, v_4\}$. Write down the expressions for the following using the log linear model:

1. Pr(positive | *awesome*)
2. Pr(negative | *amazing*)

Answer:

1. $\Pr(\text{positive} \mid \textit{awesome}) = \frac{e^{v_3}}{e^{v_3} + e^{v_4}}$
2. $\Pr(\text{negative} \mid \textit{amazing}) = \frac{e^0}{e^{v_1} + e^0}$

b. (6pts) Define the values of the parameters v_1, v_2, v_3, v_4 for the log-linear model that can model the distribution provided above perfectly.

Answer:

$$\begin{aligned}
\Pr(\text{positive} \mid \textit{amazing}) &= \frac{e^{v_1}}{e^{v_1} + e^0} = 0.9 \\
\Pr(\text{negative} \mid \textit{horrible}) &= \frac{e^{v_2}}{e^0 + e^{v_2}} = 0.9 \\
\Pr(\text{positive} \mid w) &= \frac{e^{v_3}}{e^{v_3} + e^{v_4}} = 0.6 \text{ for any word } w \text{ other than } \textit{amazing}, \textit{horrible} \\
\Pr(\text{negative} \mid w) &= \frac{e^{v_4}}{e^{v_3} + e^{v_4}} 0.4 \text{ for any word } w \text{ other than } \textit{amazing}, \textit{horrible}
\end{aligned}$$

From the above we get the parameter values $v_1 = v_2 = \log(9)$ and $v_3 = \log(6)$ and $v_4 = \log(4)$. There are many other equivalent values as long as they give the right probability in the table above.