NLP Homework 1

Alvaro Faundez

October 2021

Part I

(10 points) Do exercise 3.4 from Chapter 3 in the textbook `https://web.stanford.edu/~jurafsky/slp3/3.pdf`

We are given the following corpus, modified from the one in the chapter:

- `<s> I am Sam </s>`
- `<s> Sam I am </s>`
- `<s> I am Sam </s>`
- `<s> I do not like green eggs and Sam </s>`

Using a bigram language model with add-one smoothing, what is $P(Sam|am)$? Include `<s>` and `</s>` in your counts just like any other token.

$$\begin{aligned} count^*(\texttt{am}, \texttt{sam}) &= 2 \\ count(\texttt{am}) &= 3 \\ |V| &= 10 \\ P(\texttt{am} \mid \texttt{sam}) &= \frac{count^*(\texttt{am}, \texttt{sam}) + 1}{count(\texttt{am}) + |V|} = \frac{2+1}{3+11} = 0.21428571428571427 \end{aligned} \tag{1}$$

Part II

1.3 Questions

1. (5 points) How many word types (unique words) are there in the training corpus? Please include the end-of-sentence padding symbol `</s>` and the unknown token `</unk>`. Do not include the start of sentence padding symbol `<s>`.

There are 41738 word types (unique words) in the training corpus.

2. (5 points) How many word tokens are there in the training corpus? Do not include the start of sentence padding symbol <s>.

There are 2468210 word tokens in the training corpus.

3. (10 points) What percentage of word tokens and word types in the test corpus did not occur in training (before you mapped the unknown words to </unk> in training and test data)? Please include the padding symbol </s> in your calculations. Do not include the start of sentence padding symbol <s>.

- Word tokens only in test corpus: 46
- Word tokens in test corpus 2769
- **Percentage of word tokens only in test corpus**: 1.6612495485734922
- Word types only in test corpus: 45
- Word types in test corpus: 1248
- **Percentage of word types only in test corpus**: 3.6057692307692304

4. (15 points) Now replace singletons in the training data with </unk> symbol and map words (in the test corpus) not observed in training to </unk>. What percentage of bigrams (bigram types and bigram tokens) in the test corpus did not occur in training (treat </unk> as a regular token that has been observed). Please include the padding symbol </s> in your calculations. Do not include the start of sentence padding symbol <s>.

- Bigrams types only in test corpus: 595
- Bigrams types in test corpus: 2300
- **Percentage of bigrams types only in test corpus**: 25.869565217391305%
- Bigrams tokens only in test corpus: 597
- Bigrams tokens in test corpus: 2669
- **Percentage of bigrams tokens only in test**: 22.367928062944923%

5. (15 points) Compute the log probability of the following sentence under the three models (ignore capitalization and pad each sentence as described above). Please list all of the parameters required to compute the probabilities and show the complete calculation. Which of the parameters have zero values under each model? Use log base 2 in your calculations. Map words not observed in the training corpus to the</unk> token.

Unigram Model

$$
\begin{aligned}
S &= \texttt{I look forward to hearing your reply .} \\
\log_2(P(\texttt{i})) &= \log_2(\frac{count(\texttt{i})}{count()}) = \log_2(\frac{7339}{2468210}) \\
&= -8.39366593438855 \\
\log_2(P(\texttt{look})) &= \log_2(\frac{count(\texttt{look})}{count()}) = \log_2(\frac{613}{2468210}) \\
&= -11.97529045258011 \\
\log_2(P(\texttt{forward})) &= \log_2(\frac{count(\texttt{forward})}{count()}) = \log_2(\frac{474}{2468210}) \\
&= -12.346290467372631 \\
\log_2(P(\texttt{to})) &= \log_2(\frac{count(\texttt{to})}{count()}) = \log_2(\frac{53048}{2468210}) \\
&= -5.540022976617652 \\
\log_2(P(\texttt{hearing})) &= \log_2(\frac{count(\texttt{hearing})}{count()}) = \log_2(\frac{209}{2468210}) \\
&= -13.527674584190008 \\
\log_2(P(\texttt{your})) &= \log_2(\frac{count(\texttt{your})}{count()}) = \log_2(\frac{1217}{2468210}) \\
&= -10.985920263557162 \\
\log_2(P(\texttt{reply})) &= \log_2(\frac{count(\texttt{reply})}{count()}) = \log_2(\frac{13}{2468210}) \\
&= -17.5345939981298 \\
\log_2(P(\texttt{.})) &= \log_2(\frac{count(\texttt{.})}{count()}) = \log_2(\frac{87894}{2468210}) \\
&= -4.811556652191113 \\
\log_2(P(\texttt{</s>})) &= \log_2(\frac{count(\texttt{</s>})}{count()}) = \log_2(\frac{100000}{2468210}) \\
&= -4.625393241834078 \\
\log_2(P(S)) &= \log_2(P(\texttt{I look forward to hearing your reply .})) \\
&= \log_2(P(\texttt{i}, \texttt{look}, \texttt{forward}, \texttt{to}, \texttt{hearing}, \texttt{your}, \texttt{reply}, \texttt{.}, \texttt{</s>})) \\
&= \log_2(P(\texttt{i})) + \log_2(P(\texttt{look})) + \log_2(P(\texttt{forward})) + \\
&\quad \log_2(P(\texttt{to})) + \log_2(P(\texttt{hearing})) + \log_2(P(\texttt{your})) + \\
&\quad \log_2(P(\texttt{reply})) + \log_2(P(\texttt{.})) + \log_2(P(\texttt{</s>})) \\
&= -8.39366593438855 + -11.97529045258011 + \\
&\quad -12.346290467372631 + -5.540022976617652 + \\
&\quad -13.527674584190008 + -10.985920263557162 + \\
&\quad -17.5345939981298 + -4.811556652191113 + \\
&\quad -4.625393241834078 \\
&= -89.74040857086109
\end{aligned}
$$

(2)

Bigram Model

$$
\begin{aligned}
S &= \texttt{I look forward to hearing your reply .} \\
\log_2(P(\texttt{<s>} \mid \texttt{i})) &= \log_2(\frac{count(\texttt{<s>}, \texttt{i})}{count(< s >)}) = \log_2(\frac{2006}{100000}) \\
&= -5.639534583824631 \\
\log_2(P(\texttt{i} \mid \texttt{look})) &= \log_2(\frac{count(\texttt{i}, \texttt{look})}{count(i)}) = \log_2(\frac{15}{7339}) \\
&= -8.93447718627382 \\
\log_2(P(\texttt{look} \mid \texttt{forward})) &= \log_2(\frac{count(\texttt{look}, \texttt{forward})}{count(look)}) = \log_2(\frac{34}{613}) \\
&= -4.172280422440442 \\
\log_2(P(\texttt{forward} \mid \texttt{to})) &= \log_2(\frac{count(\texttt{forward}, \texttt{to})}{count(forward)}) = \log_2(\frac{100}{474}) \\
&= -2.2448870591235344 \\
\log_2(P(\texttt{to} \mid \texttt{hearing})) &= \log_2(\frac{count(\texttt{to}, \texttt{hearing})}{count(to)}) = \log_2(\frac{6}{53048}) \\
&= -13.110048238932082 \\
\log_2(P(\texttt{hearing} \mid \texttt{your})) &= \log_2(\frac{count(\texttt{hearing}, \texttt{your})}{count(hearing)}) = \log_2(\frac{0}{0}) \\
&= -inf \\
\log_2(P(\texttt{your} \mid \texttt{reply})) &= \log_2(\frac{count(\texttt{your}, \texttt{reply})}{count(your)}) = \log_2(\frac{0}{0}) \\
&= -inf \\
\log_2(P(\texttt{reply} \mid \texttt{.})) &= \log_2(\frac{count(\texttt{reply}, \texttt{.})}{count(reply)}) = \log_2(\frac{0}{0}) \\
&= -inf \\
\log_2(P(\texttt{.} \mid \texttt{</s>})) &= \log_2(\frac{count(\texttt{.}, \texttt{</s>})}{count(.)}) = \log_2(\frac{82888}{87894}) \\
&= -0.08460143194821208 \\
\log_2(P(S)) &= \log_2(P(\texttt{I look forward to hearing your reply .})) \\
&= \log_2(P(\texttt{<s>}, \texttt{i}, \texttt{look}, \texttt{forward}, \texttt{to}, \texttt{hearing}, \texttt{your}, \texttt{reply}, \texttt{.}, \texttt{</s>})) \\
&= \log_2(P(\texttt{<s>} \mid \texttt{i})) + \log_2(P(\texttt{i} \mid \texttt{look})) + \log_2(P(\texttt{look} \mid \texttt{forward})) + \\
&\quad \log_2(P(\texttt{forward} \mid \texttt{to})) + \log_2(P(\texttt{to} \mid \texttt{hearing})) + \log_2(P(\texttt{hearing} \mid \texttt{your})) + \\
&\quad \log_2(P(\texttt{your} \mid \texttt{reply})) + \log_2(P(\texttt{reply} \mid \texttt{.})) + \log_2(P(\texttt{.} \mid \texttt{</s>})) \\
&= -5.639534583824631 + -8.93447718627382 + -4.172280422440442 + \\
&\quad -2.2448870591235344 + -13.110048238932082 + -inf + \\
&\quad -inf + -inf + -0.08460143194821208 \\
&= -inf
\end{aligned}
\tag{3}
$$

Bigram Model with Smoothing

$$
\begin{aligned}
S &= \texttt{I look forward to hearing your reply .} \\
\log_2(P(\texttt{<s>} \mid \texttt{i})) &= \log_2(\frac{count^*(\texttt{<s>}, \texttt{i}) + 1}{count(< s >) + |V|}) = \log_2(\frac{2006 + 1}{100000 + 41739}) \\
&= -6.142052348726812 \\
\log_2(P(\texttt{i} \mid \texttt{look})) &= \log_2(\frac{count^*(\texttt{i}, \texttt{look}) + 1}{count(i) + |V|}) = \log_2(\frac{15 + 1}{7339 + 41739}) \\
&= -11.582788837823436 \\
\log_2(P(\texttt{look} \mid \texttt{forward})) &= \log_2(\frac{count^*(\texttt{look}, \texttt{forward}) + 1}{count(look) + |V|}) = \log_2(\frac{34 + 1}{613 + 41739}) \\
&= -10.240859462550434 \\
\log_2(P(\texttt{forward} \mid \texttt{to})) &= \log_2(\frac{count^*(\texttt{forward}, \texttt{to}) + 1}{count(forward) + |V|}) = \log_2(\frac{100 + 1}{474 + 41739}) \\
&= -8.707188259410588 \\
\log_2(P(\texttt{to} \mid \texttt{hearing})) &= \log_2(\frac{count^*(\texttt{to}, \texttt{hearing}) + 1}{count(to) + |V|}) = \log_2(\frac{6 + 1}{53048 + 41739}) \\
&= -13.725046665121754 \\
\log_2(P(\texttt{hearing} \mid \texttt{your})) &= \log_2(\frac{count^*(\texttt{hearing}, \texttt{your}) + 1}{count(hearing) + |V|}) = \log_2(\frac{0 + 1}{209 + 41739}) \\
&= -15.35631440692812 \\
\log_2(P(\texttt{your} \mid \texttt{reply})) &= \log_2(\frac{count^*(\texttt{your}, \texttt{reply}) + 1}{count(your) + |V|}) = \log_2(\frac{0 + 1}{1217 + 41739}) \\
&= -15.390572037471506 \\
\log_2(P(\texttt{reply} \mid \texttt{.})) &= \log_2(\frac{count^*(\texttt{reply}, \texttt{.}) + 1}{count(reply) + |V|}) = \log_2(\frac{0 + 1}{13 + 41739}) \\
&= -15.349557686620518 \\
\log_2(P(\texttt{.} \mid \texttt{</s>})) &= \log_2(\frac{count^*(\texttt{.}, \texttt{</s>}) + 1}{count(.) + |V|}) = \log_2(\frac{82888 + 1}{87894 + 41739}) \\
&= -0.6451804614204727 \\
\log_2(P(S)) &= \log_2(P(\texttt{I look forward to hearing your reply .})) \\
&= \log_2(P(\texttt{<s>}, \texttt{i}, \texttt{look}, \texttt{forward}, \texttt{to}, \texttt{hearing}, \texttt{your}, \texttt{reply}, \texttt{.}, \texttt{</s>})) \\
&= \log_2(P(\texttt{<s>} \mid \texttt{i})) + \log_2(P(\texttt{i} \mid \texttt{look})) + \log_2(P(\texttt{look} \mid \texttt{forward})) + \\
&\quad \log_2(P(\texttt{forward} \mid \texttt{to})) + \log_2(P(\texttt{to} \mid \texttt{hearing})) + \log_2(P(\texttt{hearing} \mid \texttt{your})) + \\
&\quad \log_2(P(\texttt{your} \mid \texttt{reply})) + \log_2(P(\texttt{reply} \mid \texttt{.})) + \log_2(P(\texttt{.} \mid \texttt{</s>})) \\
&= -6.142052348726812 + -11.582788837823436 + -10.240859462550434 + \\
&\quad -8.707188259410588 + -13.725046665121754 + -15.35631440692812 + \\
&\quad -15.390572037471506 + -15.349557686620518 + -0.6451804614204727 \\
&= -97.13956016607362
\end{aligned}
\tag{4}
$$

Bigram Model with 0.5 discounting and Katz backoff

$$
\begin{aligned}
S &= \texttt{I look forward to hearing your reply .} \\
\log_2(P(\texttt{<s>} \mid \texttt{i})) &= \log_2(\frac{count^*(\texttt{<s>}, \texttt{i})}{count(\texttt{<s>})}) = \log_2(\frac{2005.5}{100000}) = -5.639894223622303 \\
\log_2(P(\texttt{i} \mid \texttt{look})) &= \log_2(\frac{count^*(\texttt{i}, \texttt{look})}{count(\texttt{i})}) = \log_2(\frac{14.5}{7339}) = -8.983386786754767 \\
\log_2(P(\texttt{look} \mid \texttt{forward})) &= \log_2(\frac{count^*(\texttt{look}, \texttt{forward})}{count(\texttt{look})}) = \log_2(\frac{33.5}{613}) = -4.193654073233009 \\
\log_2(P(\texttt{forward} \mid \texttt{to})) &= \log_2(\frac{count^*(\texttt{forward}, \texttt{to})}{count(\texttt{forward})}) = \log_2(\frac{99.5}{474}) = -2.2521186283546104 \\
\log_2(P(\texttt{to} \mid \texttt{hearing})) &= \log_2(\frac{count^*(\texttt{to}, \texttt{hearing})}{count(\texttt{to})}) = \log_2(\frac{5.5}{53048}) = -13.23557912101594 \\
\alpha_{\texttt{hearing}} &= 1 - \frac{\Sigma_w count^*(\texttt{hearing}, \texttt{your})}{count(\texttt{hearing})} = 1 - \frac{170.0}{209} = 0.1866028708133971 \\
\log_2(P(\texttt{hearing} \mid \texttt{your})) &= \log_2(\alpha_{\texttt{hearing}} \times \frac{P_{ML}(\texttt{your})}{\Sigma_{w \in B_{\texttt{hearing}}} P(\texttt{w})}) \\
&= \log_2(0.1866028708133971 \times \frac{0.00047387090619536564}{0.6730450391519687}) \\
&= -12.893950161003882 \\
\alpha_{\texttt{your}} &= 1 - \frac{\Sigma_w count^*(\texttt{your}, \texttt{reply})}{count(\texttt{your})} = 1 - \frac{874.5}{1217} = 0.2814297452752671 \\
\log_2(P(\texttt{your} \mid \texttt{reply})) &= \log_2(\alpha_{\texttt{your}} \times \frac{P_{ML}(\texttt{reply})}{\Sigma_{w \in B_{\texttt{your}}} P(\texttt{w})}) \\
&= \log_2(0.2814297452752671 \times \frac{5.061891356236445e-06}{0.8819099684217718}) \\
&= -19.239748589030043 \\
\alpha_{\texttt{reply}} &= 1 - \frac{\Sigma_w count^*(\texttt{reply}, \texttt{.})}{count(\texttt{reply})} = 1 - \frac{10.0}{13} = 0.23076923076923073 \\
\log_2(P(\texttt{reply} \mid \texttt{.})) &= \log_2(\alpha_{\texttt{reply}} \times \frac{P_{ML}(\texttt{.})}{\Sigma_{w \in B_{\texttt{reply}}} P(\texttt{w})}) \\
&= \log_2(0.23076923076923073 \times \frac{0.03422383683577278}{0.9201778670749203}) \\
&= -6.864316558703887 \\
\log_2(P(\texttt{.} \mid \texttt{</s>})) &= \log_2(\frac{count^*(\texttt{.}, \texttt{</s>})}{count(\texttt{.})}) = \log_2(\frac{82887.5}{87894}) = -0.08461013465181358 \\
\log_2(P(S)) &= \log_2(P(\texttt{I look forward to hearing your reply .})) \\
&= \log_2(P(\texttt{<s>}, \texttt{i}, \texttt{look}, \texttt{forward}, \texttt{to}, \texttt{hearing}, \texttt{your}, \texttt{reply}, \texttt{.}, \texttt{</s>})) \\
&= \log_2(P(\texttt{<s>} \mid \texttt{i})) + \log_2(P(\texttt{i} \mid \texttt{look})) + \log_2(P(\texttt{look} \mid \texttt{forward})) + \\
&\quad \log_2(P(\texttt{forward} \mid \texttt{to})) + \log_2(P(\texttt{to} \mid \texttt{hearing})) + \log_2(P(\texttt{hearing} \mid \texttt{your})) + \\
&\quad \log_2(P(\texttt{your} \mid \texttt{reply})) + \log_2(P(\texttt{reply} \mid \texttt{.})) + \log_2(P(\texttt{.} \mid \texttt{</s>})) \\
&= -5.639894223622303 + -8.983386786754767 + -4.193654073233009 + \\
&\quad -2.2521186283546104 + -13.23557912101594 + -12.893950161003882 + \\
&\quad -19.239748589030043 + -6.864316558703887 + -0.08461013465181358 \\
&= 9 \\
&= -73.38725827637026
\end{aligned}
\tag{5}
$$

6. (20 points) Compute the perplexity of the sentence above under each of the models.

$$S = \texttt{I look forward to hearing your reply .} \tag{6}$$

Unigram Model

$$\begin{aligned} M_S &= 9 \\ \log_2(P_{ML}(S)) &= -89.74040857086109 \\ l &= \frac{\log_2(P_{ML}(S))}{M_S} \\ &= \frac{-89.74040857086109}{9} \\ &= -9.971156507873454 \\ Perplexity(S) &= 2^{-l} = 1003.7306831109403 \end{aligned} \tag{7}$$

Bigram Model

$$\begin{aligned} M_S &= 9 \\ \log_2(P_{ML}(S)) &= -inf \\ l &= \frac{\log_2(P_{ML}(S))}{M_S} \\ &= \frac{-inf}{9} \\ &= -inf \\ Perplexity(S) &= 2^{-l} = inf \end{aligned} \tag{8}$$

Bigram Model with Smoothing

$$\begin{aligned} M_S &= 9 \\ \log_2(P_{ML}(S)) &= -97.13956016607362 \\ l &= \frac{\log_2(P_{ML}(S))}{M_S} \\ &= \frac{-97.13956016607362}{9} \\ &= -10.793284462897068 \\ Perplexity(S) &= 2^{-l} = 1774.607755085189 \end{aligned} \tag{9}$$

Bigram Model with 0.5 discounting and Katz backoff

$$\begin{aligned} M_S &= 9 \\ \log_2(P_{ML}(S)) &= -73.38725827637026 \\ l &= \frac{\log_2(P_{ML}(S))}{M_S} \\ &= \frac{-73.38725827637026}{9} \\ &= -8.154139808485585 \\ Perplexity(S) &= 2^{-l} = 284.86603528933665 \end{aligned} \tag{10}$$

7. (20 points) Compute the perplexity of the entire test corpus under each of the models. Discuss the differences in the results you obtained.

$$C = \{S : S \in Corpus_{test}\} \tag{11}$$

Unigram Model

$$\begin{aligned} M_C &= \Sigma_{S \in C} M_S = 2769 \\ \log_2(P_{ML}(C)) &= \Sigma_{S \in C} \log_2(P_{ML}(S)) = -27965.787053030697 \\ l_C &= \frac{\log_2(P_{ML}(C))}{M_C} \\ &= \frac{-27965.787053030697}{2}769 \\ &= -10.099598068989057 \\ Perplexity(C) &= 2^{-l_C} = 1097.190308743997 \end{aligned} \tag{12}$$

Bigram Model

$$\begin{aligned} M_C &= \Sigma_{S \in C} M_S = 2769 \\ \log_2(P_{ML}(C)) &= \Sigma_{S \in C} \log_2(P_{ML}(S)) = -inf \\ l_C &= \frac{\log_2(P_{ML}(C))}{M_C} \\ &= \frac{-inf}{2}769 \\ &= -inf \\ Perplexity(C) &= 2^{-l_C} = inf \end{aligned} \tag{13}$$

Bigram Model with Smoothing

$$\begin{aligned} M_C &= \Sigma_{S \in C} M_S = 2769 \\ \log_2(P_{ML}(C)) &= \Sigma_{S \in C} \log_2(P_{ML}(S)) = -31072.441669718453 \\ l_C &= \frac{\log_2(P_{ML}(C))}{M_C} \\ &= \frac{-31072.441669718453}{2} 769 \\ &= -11.221539064542599 \\ Perplexity(C) &= 2^{-l_C} = 2387.9204561337933 \end{aligned} \tag{14}$$

Bigram Model with 0.5 discounting and Katz backoff

$$\begin{aligned} M_C &= \Sigma_{S \in C} M_S = 2769 \\ \log_2(P_{ML}(C)) &= \Sigma_{S \in C} \log_2(P_{ML}(S)) = -23158.85155443307 \\ l_C &= \frac{\log_2(P_{ML}(C))}{M_C} \\ &= \frac{-23158.85155443307}{2} 769 \\ &= -8.363615584844013 \\ Perplexity(C) &= 2^{-l_C} = 329.381469853553 \end{aligned} \tag{15}$$

As expected the bigram model with Katz back-off has the lowest perplexity, 329.38, followed by the unigram model (perplexity 1097.19) and the bigram model with smoothing (2387.920). I wasn't expecting the huge difference between the unigram and bigram with smoothing, but as it was explained in class, smoothing does not perfome well.