Maria Frey
LING492B
18 March 2021

Lab 3 Questions

PART ONE:

Question 1:
In the bigram model some examples of unnatural words are:

1. # P T EY T IY # IPA: [pteɪti]
   a. This is an awkward word because of the P T sequence; this is generated because # P is possible in English and P T is also possible, (for example in any past tense of verbs ending in ‘p’ or in words like ‘uptown’ and ‘abrupt’) and so the probability of this sequence is not low.
2. # K M AA N IH NG T EH R AH D # IPA: [kmɑnɪŋtɛɹʌd]
   a. This is unnatural because of the K M sequence which is possible word-medially in words like ‘blackmail’ and ‘embankment’ so the K M sequence is not unlikely for a bigram.
3. # D AH P R D AA R UW D # IPA: [dʌpɹdɑɹud]
   a. This is unnatural because of the P R D combination; P R is possible in many words such as ‘print’ and ‘compress’, and R D is possible in the past tense of verbs ending in ‘r’ as well as words like ‘aboard’ and ‘yesterday’. Therefore, a sequence like P R D is likely for a bigram model.
4. # P L Z # IPA: [plz]
   a. This is unnatural because of the lack of vowels, though it is a possible sequence of consonants. This is generated because # P is very common in English words, P L is also very common, L Z is common, and so is Z #. Therefore this is perfectly fine for the bigram.
5. # G N AY EH N EY T IH P AE P AY Z D EY V AH N T S T EH R AH N T AH Z #
   IPA: [gnaɪɛneɪtɪpæpaɪzdeɪvʌntstɛɹʌntʌz]
   a. This is a funny example because it’s unnaturally long and it’s a combination of a bunch unnatural sequences like G N and T S T and unlikely sequences like AY EH (there are only 6 of these in the training data and 2 of them are the name Rafael/Raphael!). The length of this is so long because the probabilities of all of these sequences is so low and the unnatural sequences are possible because those sequences exist in words of English.

In the trigram model some examples of unnatural words are:

1. # # F L UW P ER V IH ZH W AA CH D AW T P OW K T IH NG #
   IPA: [flupɝvɪdʒwɑtʃdaʊtpoʊktɪŋ]
   a. This is unnatural because it is too long. There aren’t enough words that long like this in English, so it just seems like gibberish. If it were broken up into more words it would be more natural, but it has too many codas that seem like word boundaries. I could believe [flupɝvɪdʒ] as its own word or even [flupɝvɪdʒwɑtʃ] but any longer and it doesn’t seem right. This was probably generated because the nature of the model is

    to keep adding phonemes until their probabilities equal 1, so it was just randomly generated like this.

2. # # B AA K IH NG K IH NG # IPA: [bɑkɪŋkɪŋ]
    a. This is unnatural to me solely because of [kɪŋkɪŋ]. It just sounds unnatural or like it needs to be two words [bɑkɪŋ] and [kɪŋ]. It makes sense that the model generated this because [kɪŋ] is very common in English so according to the model the sequence K IH NG K IH NG shouldn’t be problematic.
3. # # S T ER AH L IH NG T AY T AH L EY S # IPA: [stɝʌlɪŋtaɪtʌleɪs]
    a. This one is also unnatural because of the way it sounds, so it’s nothing that the model did wrong, it just sounds off to the human ear.
4. # # EH M AH L EY SH AH N EY T AH L EH N R IY S AH N T IH V EY SH AH N # IPA: [ɛmʌleɪʃʌneɪtʌlɛnɹisʌntɪveɪʃʌn]
    a. This is awkward because of its length. Just like example 1 this seems like it should be multiple words, especially because of the replication of English morphology. [ɛmʌleɪʃʌn] should be its own word because the use of the morpheme ‘-tion’ is almost always word-final. Since the model is just predicting words and doesn’t know morphology, this word is a perfectly fine representation of the training data.
5. # # G L AE M B UW M ER SH AE N T EH R D Z # IPA: [glæmbumɝʃæntɛɹdz]
    a. This doesn’t sound like a natural word of English because of the lack of unstressed vowels like /ə/ (I’ve been considering AH to be more like [ə] than [ʌ]). Since unstressed syllables in English usually become /ə/, most long words have at least one of them. The model was not specifically tuned to produce words according to this rule, so it makes sense that even though it is less likely, long words can be produced without AH.

Question 2:
The trigram was definitely able to generate more English-like words than the bigram. The trigram even generated that were morphologically meaningful! Because the trigram uses more information when generating the words, it makes sense that the words it makes are better overall. The awkward words generated by the bigram are much less likely to be generated by the trigram because the context is bigger. For example, the trigram wouldn’t generate a word like P L Z or a sequence like P R D because they never happen in the training data. It was hard to find words that the trigram generated that would be nearly impossible or completely unnatural in English, most of words that I thought were unnatural just didn’t sound right.

PART TWO:
Collaborator: Barbora Hlachova

I wasn’t able to get the code running properly for part two, so I’m not able to answer these questions