Assignment 5
Language & Computers | LINS313

Due: Monday Aug 15, by 11:59pm

The homework is due to Canvas by 11:59pm on its due date. Typeset is preferable, but if you do handwrite it and scan it, please ensure that the writing is clearly legible. If you don't have access to a scanner, there are many free scanner apps for smartphones which will give you a pretty good quality image.

A perfect solution is worth 50 points. **Make sure you show your work**; this will account for the majority of your grade. If any of these instructions or questions do not make sense to you, reach out to me right away.

Problem 1: Language Models(10 Points) You are given the following corpus, preprocessed to have punctuation removed and start and end tokens inserted.

START I am Sam END
START Sam I am END
START I am Sam END
START I do not like green eggs and Sam END
START I do so like green eggs and ham END

i. Construct a bigram Markov model for this corpus (where each word is conditioned on the previous word). Draw the state diagram as a FSA (finite state automaton), with transition arrows annotated for probability.[1]

   - What is the bigram model probability of the sentence *I do not like green eggs and Sam*? How about *I am Sam*?

[1] To see the example we used in class with accompanying explanation, check out Alexander Dejeu's tutorial at https://hackernoon.com/from-what-is-a-markov-model-to-here-is-how-markov-models-work-1ac5f4629b71.

- Generate one sentence using this model that isn't in the corpus. What is it? What is its probability? Can you generate more new sentences? If so, write some others.
- Bonus (1 point): Pick one of these sentences. What is the perplexity of the sentence? Perplexity $PP(W)$ of a sentence W is defined as

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_n)}}$$

ii. Now construct a trigram Markov model (where each word is conditioned on the previous *two* words) and draw the FSA for it. This one will more states!

- What is the trigram model probability of the sentence *I do not like green eggs and Sam?* How about *I am Sam*?
- Generate one sentence using this model that isn't in the original corpus. What is it? What is its probability? Can you generate more new sentences? If so, write some others.
- Bonus (1 pount): Pick one of these sentences. What is the perplexity of the sentence?

Problem 2: Markov Chain Reaction (10 Points) For this problem you will use a Markov chain generator to create a mashup of two texts of your choosing, by automatically constructing a language model based on two texts and then randomly generating new text. Experiment with the word-level n-gram generator at `https://ankiewicz.com/technology/markov-generator/#` and the character-level generator at `https://projects.haykranen.nl/markov/demo/`. Paste both texts into the same box (this is your corpus). Experiment with the settings, specifically the *order* of the model, to find a combination that produces output to your liking.

Tips: To get the best results, your texts should be as long as possible. So that one text does not dominate the language model, the two texts should also be of *comparable* length.

i. Report on the texts you chose to combine, the settings you landed on, and give a few examples (sentence-paragraph length that you like). Why did you

prefer these settings to the others you tried? How is the generated output different?

ii. What are the differences between character level language models and word-level language models? What can character level models do that word-level models cannot? What can word-level models do that character level models cannot?

Problem 3: Winograd Schema (10 Points)

Part 1 Here are a few Winograd schema challenge questions. For each one, a) identify the anaphor whose reference changes depending on the variant chosen, and b) name the correct antecedent of the anaphor for each variant of the sentence.

(1) Sam tried to paint a picture of shepherds with sheep, but they ended up looking more like [dogs/golfers].

(2) Mary tucked her daughter into bed, so that she could [sleep/work].

(3) We went to the lake, because a shark had been seen at the ocean beach, so it was a [dangerous/safer] place to swim.

(4) Jane knocked on Susan's door but she did not [answer/get an answer].

(5) The man lifted the boy onto his [bunk bed/shoulders].

Part 2 Winograd Schema Challenge questions consists of three parts.

i. A sentence or brief discourse that contains the following:

- Two noun phrases of the same semantic class (male, female, inanimate, or group of objects or people),
- An ambiguous pronoun that may refer to either of the above noun phrases, and
- A special word and alternate word, such that if the special word is replaced with the alternate word, the natural resolution of the pronoun changes.

ii. A question asking the identity of the ambiguous pronoun, and

iii. Two answer choices corresponding to the noun phrases in question.

A machine will be given the problem in a standardized form which includes the answer choices, thus making it a binary decision problem.

Write two novel challenge questions, and give the same answers as above: what is the anaphor with varying reference? To what does the anaphor refer in either variant of the sentence?

Problem 4: Word vectors (10 Points) Go to this website, which lets you play around with word vectors http://vectors.nlpl.eu/explore/embeddings/en/visual/. Click "Similar Words" and choose settings: English Gigawords, All of Them, High/Medium/Low.

i. Choose a word of your choice and find its 10 closest semantic associates. Report them and write a few sentences about what you found.

ii. The program automatically guesses the part of speech you want. But you can also control it like BREAK_NOUN vs. BREAK_VERB. Come up with a word for which you can pass a different part of speech tag (NOUN or VERB, etc.) that you think will cause the list of related words to be very different. Report and discuss the results.

iii. Try entering GREEN_ADJ. What are the 10 words most associated with it? Try other colors (BLUE_ADJ, RED_ADJ, etc.). Do you think the model has learned something about the space of colors? "Green" can also mean "inexperienced" or "environmentally conscious." Are those meanings here? Why or why not?

Problem 5: Translation (10 Points) Here is a parallel corpus consisting of two sentence pairs.

Mary unlocked the car door.	Can you unlock the door?
Maria schloss die Autotiir auf.	Kannst du die Tiir aufschliessen?

Word translations: German "du" means *you*, and "die" corresponds to *the*. German "kannst" translates to the English verb *can*. German "Auto" translates to *car*,

and "Tiir" means *door*. The German verb "aufschliessen", *to unlock*, has a separate prefix: In the past tense, the prefix "auf" (un-) separates from the rest of the verb to make "schloss... auf".

Part 1 Draw the alignment between each English/German sentence pair in the form of connecting lines. Some of the alignments may involve may be one-to-many or many-to-one.

Part 2 Based on the alignment you have annotated, what are the translation probabilities P(Tiir|door) and P(die|the)?

Part 3 Based on this two-sentence parallel corpus, what is the fertility of the English word unlock? In particular, what is $n(1|unlock)$, the probability that unlock will yield one word of output in the German translation? What is $n(2|unlock)$?