Closed Book, Open Notes (two handwritten pages), 75 minutes. *Please answer concisely.*

1. [10 points] True or False
   a. T F For a language model, lower entropy means higher perplexity.
   b. T F Sentiment analysis should start with exclusion of words on the stoplist.
   c. T F For information retrieval, stemming can be as good as lemmatization.
   d. T F For machine translation, stemming can be as good as lemmatization.
   e. T F Given a grammar, a sentence that has multiple possible parse trees is called “ambiguous”.
   f. T F Word embeddings are used in information retrieval primarily because it’s faster to compute similarity in 50-300 dimensions than in 500,000 dimensions.
   g. T F Multilingual word embeddings, as developed by Mikolov, require a large bitext (a corpus of matched sentences across two languages), and no monolingual corpora.
   h. T F Multilingual word embeddings, as developed by Mikolov, require large multilingual corpora but no bitext.
   i. T F Long short-term memory (LSTM) models can be used to classify inputs that are sequences.
   j. T F In speech recognition, the input to the Acoustic Model is a sequence of features, each representing aspects of the sound at one moment in time.

2. [2] Give examples illustrating how a single word of English may, depending on the context, have different parts of speech.

3. [4] In some ways, speech recognition is easier for dictation than it is for dialog systems, but in some ways the opposite is true. Explain.
4. [10] Briefly define or explain
   a. Smoothing

   b. Crawling

   c. Query expansion

   d. Polysemy

   e. Glottis

5. [6] In English the pronunciation of the past tense morpheme is context-dependent. In particular, it is pronounced /t/ after an unvoiced consonant (as in whiffed, kissed), /d/ after a voiced sound (as in loved, flinged, freed), and /ɛd/ in the context of /d/ or /t/ (as in padded, petted). Describe the situation using at least two of the following terms: phonological rules, diphthong, coarticulation, fricative.
6. [4] Dialog systems sometimes need to confirm the user input. To do this they need to use “confidence measure” which represents how confident the speech recognizer is that it has correctly understood the user’s input. Assuming that the speech recognizer can computing $P(w)P(o|w)$ for all words in the lexicon, write an equation for computing the confidence.

7. [1] What is the Euclidean distance between the points (-1, 5) and (2,1)?

8. [1] What is the cosine distance between the points (0,6) and (3,2)?

9. [2] If a certain dialog system has heard a user utterance in a state where $P(\text{yes}) = .4$, $P(\text{no}) = .3$ and $P(\text{yeah}) = .2$, and the acoustic model tells us that $P(o|\text{yes}) = .003$, $P(o|\text{no}) = .001$ and $P(o|\text{yeah}) = .005$, what word did the user most likely say? Show your work.
10. Imagine you plan to build a tool to help prospective UTEP CS graduate students find a research advisor whose expertise matches their interests. This tool could take many forms: chatbot, spoken dialog system, search engine, etc. Which would be best? Why?

11. Craigslist is a website where users can post services available or items for sale. It has a long list of prohibited content, including weapons, fireworks, body parts, medical devices, alcohol, tobacco, ivory, food stamps, birth certificates, military items, lottery tickets, and so on. Removing such prohibited content is a labor-intensive process. Describe how you might build a system to flag suspicious posts for possible removal. Your system should be robust to mangled keywords (“I have 200 Powerb**l tickets for sale at half price, not stolen, totally legit!!”), and circumlocutions (“high quality wafflemakers with any name and date of birth you want, indistinguishable from the official State of Texas ones (for entertainment purposes only, not to be used to apply for a drivers license etc.)”).

   a. Specify how you would formalize the task and measure system performance.
   b. Describe the structure of a system for this task, specifying the main technologies/components.
   c. Describe what training data you would use and how you would use it.