Minor Examination

Course Name: Natural Language Understanding **Code:** CSL 7640

Total Scores-40 **Time:** 2 hours

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the parts to any particular question should appear together.

Q1 (a). Explain how ambiguity in natural language is different/similar to the ambiguity in formal language. Give examples of lexical, and syntactic ambiguities (*scope, prepositional phrase attachment*), and state how these ambiguities could pose challenges in further processing. Explain how semantic information and context could help in resolving such ambiguities. **[3+ 5 + 3]**

Q1(b). Explain with examples and intuition how discourse level tasks introduce more complexities in comparison to semantics and syntactic level tasks. **[4]**

Q2(a). Consider following four tags for the Part-of-Speech (PoS) tagging problem.

N-Noun; A-Adjectives; V-Verb; O-Others

Historically, AI was used to understand and recommend information. Now, Generative AI can also help us create new content. Generative AI builds on existing technologies, like Large Language Models (LLMs) which are trained on large amounts of text and learn to predict the next word in a sentence. For example, "peanut butter and ___" is more likely to be followed by "jelly" than "shoelace"

Assign appropriate PoS tags to every token in the above sentences. **[5]**

Q2(b). Consider a first order HMM, and compute the transition and emission probabilities from the tagged sequence of Q2(a) (***treating this as a training set***).

Show how using Viterbi decoding, the sequence of tags will be assigned to the following sentence (***make appropriate assumption as and when required***):

AI was used to understand and recommend information.. **[8+3]**

Q3 (a). Considering proper nouns as the potential named entities, mark the named entities for the example paragraph of Q2 (a). Use the BIO (*where B, I and O denote the beginning, Intermediate and outside of NE*) notation for marking the boundaries of the NEs.

Q3(b). Explain with examples how contextual information influences Named Entity Recognition.

Q3 (c). Give an example where context could help in NE classification.

[4+3+2]

Best of Luck

Major Examination

Course Name: Natural Language Understanding
Code: CSL 7040

Full Marks-60 **Time:** 3 hours

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the components to any particular question should appear together.

Q1(a). *Attention in NMT is equivalent to Alignment in SMT.* Explain this with proper intuition.

Q1 (b). Consider the following pair of parallel sentences in English-Hindi

India is a diverse country→भारत एक विविधतापूर्ण देश है (transliterated form: *bhaarat ek vividhataapoorn desh hai*)

Construct the word alignment table, and subsequently the phrase table from the word alignment information.

Q1(c). Consider a Pivot based NMT system (with three languages). What kind of transfer learning strategy will be adapted when the source is a more distant language to pivot than the target; and the target is a more distant language to pivot than the source? (*explain with appropriate architectural diagrams and necessary steps*). **(3+8 +6)**

Q2 (a). With an example of product review sentence (two aspects, each having different sentiment classes, *viz.*positive and negative), describe how a bi-directional LSTM by feeding target specific encoded representations at every time step help improve the performance over a model that does not make target specific representations (*for aspect based sentiment analysis*)?

Q2 (b). Show an example scenario, where a single word exhibits two different sentiments.

(6+2)

(Q3). (a). For semantics, why is the count based method (e.g. tf-idf) not as robust as the prediction based method (e.g.. word2vec)?

(Q3). (b). Explain with necessary architectural diagrams and optimization criteria how a Skip-gram word embedding model be implemented using CBOW based word embedding model. **(3+7)**

Q4(a). What is the intuition behind treating every task of NLU as a next word prediction problem in LLM?

Q4 (b). *Temperature sampling and top-k sampling for LLM generation are equivalent-* Justify in favor or against this claim with proper intuitions and explanations.

Q4(c). How caching of attention scores done in LLM and how does it benefit? How parameter efficient training is carried out in LLM? **(3+4+6)**

Q5 (a). Determine the coreference chain from the following text:

The development of AI has created new opportunities to improve the lives of people around the world, from business to healthcare to education. It has also raised new questions about the best way to build fairness, interpretability, privacy, and safety into these systems.

Q5 (b). Create the instances (*positive and negative*) for training of a machine learning based anaphora/coreference resolution model for the example of Q5 (a) (*consider noun phrases as the possible markable*).

Q5(c). For every instance of (b) above, extract the features to satisfy the constraints (e.g. number agreement, gender agreement etc). **(3+5+4)**

All the best