This is a closed book and closed notes quiz. No electronic aids or cheat sheets are allowed.

There are 2 pages, 3 parts of questions, and 20 total points in this quiz.

The questions are printed on the back of this paper!

Please carefully read the instructions and questions before you answer them.

Please pay special attention on your handwriting; if the answers are not recognizable by the instructor, the grading might be inaccurate (NO argument about this after the grading is done).

Try to keep your answers as concise as possible; grading is not by keyword matching.

| Total | /20 |
1 True/False Questions (3pts×2)

For the statement you believe it is False, please give your brief explanation of it (you do not need to explain anything when you believe it is True). Note the credit can only be granted if your explanation is correct.

1. IBM translation model I assumes an one-to-one mapping between the words in the source language and the target language.
   **False, and Explain:** IBM translation model I can flexibly model all sorts of alignments, including one-to-one, one-to-many, and many-to-one. It only assumes a uniform alignment.

2. WordNet is organized as a fully connected graph, where synsets are the nodes on the graph.
   **False, and Explain:** WordNet is organized as a directed acyclic graph, and therefore it is not fully connected.

2 Multi-choice Questions (4pts×2)

1. Which of the following can be modeled as a sequence labeling problem: (b) (c)
   (a) Machine translation;
   (b) Part-of-speech tagging;
   (c) Named entity recognition;
   (d) Word semantic similarity comparison.

2. Bayes rule is the key ingredient in which of the following algorithm(s): (b) (c)
   (a) Language models;
   (b) Generative machine translation models;
   (c) Hidden Markov models;
   (d) Maximum Entropy Markov models.

3 Short Questions (6 pts)

1. How to use a Hidden Markov Model to perform tokenization in English?

   1. Prepare a corpus of sentences with BIO annotation of word boundaries. For example, Beginning of a word, Inside of a word, Outside of a word.
   2. Using maximum likelihood estimation to estimate the transition and emission probability in an HMM.
   3. Perform Viterbi algorithm with the trained HMM on new sentences for tokenization.