April 28, 2003

Nigel Ward

Statistical Approaches to Natural Language Processing

CS 4390/5319

Overview of the Final

Like the tests so far, the final exam will have three types of question:

1. Knowledge Questions, testing your knowledge of the facts of speech and language, models, and basic concepts,
2. Skills Questions, testing your ability to apply standard analysis techniques, models, and algorithms,
3. Integrative Questions, testing your understanding of: the suitability of the various models and techniques for different tasks, of the engineering issues involved in building systems, etc.

You are encouraged to bring 1 sheet of *handwritten* notes to refer to during the exam. You may use both sides. Write your name on it, as it will be collected together with the exam.

Possible Integrative Questions

1. Why is NLP hard? List two or three aspects.

   Now try to illustrate these aspects using the following sentences, or, if not possible, explain what is unusually simple about these sentences.

   "You can't be in London for long without going to the zoo."

   "Vowels are sounds produced by vocal cord vibration (voicing) and a relatively open vocal tract."

2. Why is it so much easier to build an understander for a computer language (i.e. a compiler) than an understander for a natural language?

3. It is easy to build a system that does NLP better, in some respects, than humans do. Give two examples where this is the case.

4. Discuss ways in which the nature of understanding differs as a function of task. For example, understanding for machine translation, understanding for information extraction, understanding for information retrieval, and understanding for commands to a robot.

5. For each of these systems, state what characteristic(s) of the task allowed it to be a success:
   (a) Student (the algebra problem solver)
   (b) Meteo (the English-to-French weather forecast translator)
   (c) Eliza (the pattern-matching psychoanalyst)
6. For disambiguation, what are the advantages and disadvantages of using preferences (including statistical preferences), rather than using hard constraints?
7. For building machine translation systems, what are the advantages and disadvantages of each of the following models: direct, transfer, interlingua?
8. In what ways is a conversation like a waltz? (If you built a robot capable of waltzing with a human, what parts of its software could you adapt to use in a conversational program?)
9. Chomsky (1957) writes (page 16) "the notion 'grammatical in English' cannot be identified in any way with the notion "high order of statistical approximation to English"' and says that statistics alone could not account for the difference in grammaticality between *"I saw a fragile whale."* and *"I saw a fragile of."*

   Could modern statistical techniques handle his example? Explain how or explain why not.

   Is Chomsky's claim therefore invalid? Explain why or why not.
10. For each of the following tasks, which of the following techniques are most useful? If no answer is appropriate, explain why. (Note that some techniques may not be useful for anything.)

    (a) case grammar
    (b) HMM-based part-of-speech tagging
    (c) probabilistic CFG
    (d) top-down procedural parsing
    (e) unification
    (f) scripts
    (g) semantic grammar
    (h) Eliza-style pattern-matching
    (i) case grammar
    (j) other _____

    ___ serving as a language model to aid speech recognition for a dictation system
    ___ machine translation of weather forecasts
    ___ machine translation of newspaper articles about politics
    ___ understanding e-mail messages about printer problems
    ___ automatically checking the grammar of English sentences written by non-natives
    ___ understanding questions about the names, addresses, and phone numbers of people listed in a database
    ___ an interface to a system for giving advice about planning subway trips
    ___ extracting information from newspaper stories

11. Consider each of the following as an application for machine translation. Rank the difficulty of each from 1 (easy, possible today) to 4 (very very hard), Also, for each task, say briefly what makes it easy or hard. (adapted from JM Exercise 21.12)

    (a) letters between an American girl and her Chinese pen-pal ___

    (b) electronic mail ___

    (c) articles in chemistry journals ___

    (d) magazine advertisements ___

    (e) children's storybooks ___

    (f) history books ___

    (g) an English-speaker wanting to read articles on Japanese anime-lovers web sites ___

    (h) an English-speaker wanting to post an article to a Japanese animi-lovers web site ___

12. Discuss techniques and metrics for evaluating the performance of each of the following:

    CFG grammars
    parsers of various types
    language models
    machine translation systems
    information extraction systems
    natural language interfaces to databases

13. Recall that we said that a good grammar should have the following attributes:

    A produces/accepts all the sentences of a language

    B does not produce sentences which are ungrammatical (does not overgenerate)

    C is not too ambiguous

    D produces reasonable parse trees for all sentences of a language (so a semantic interpreter can subsequently extract the meaning)

    E can be parsed efficiently (without taking too much time or disk space)

    F is useful even for parsing incomplete and ungrammatical inputs

    For machine translation, which of the above attributes are
    most important? _____ least important? _____

    For a natural language interface to a database ...
    most important? _____ least important? _____

    For a language model for a speech understanding system ...
    most important? _____ least important? _____

    For information extraction from newspaper stories ...
    most important? _____ least important? _____

14. On a song I heard the phrase "I was hit by a Maine car", but when I looked a the song lyrics, it turned out the words were actually "mink car".

    (a) What factors likely contributed to this mis-hearing?

    (b) Would those factors also be likely to cause a speech recognition systems to get the same error?

    (c) How could you fix a speech recognition system so that this error were less likely to occur?

    (d) Could you fix a speech recognition system so it would *never* make this kind of error? How or why not?

15. Consider the following sentence about bridge:

    > *Declarer would have all the time in the world to force out the ace and king of the suit, and come to nine tricks via one spade, two hards, five diamonds, and the ace of clubs.*

    (a) underline all NPs

    (b) what is the grammatical role played by the word *out*?

    (c) what is the grammatical role played by the first word *to*?

    (d) what is the grammatical role played by the second *to*?

    (e) what words or phrases are conjoined by the first *and*?

    (f) what words or phrases are conjoined by the second *and*?

    (g) what words or phrases are conjoined by the third *and*?

    Now, suppose we wanted a parser capable of producing a representation which included the answers to the above questions. For each question, give a parsing techniques which would be suitable for reliably answering questions of that type, and explain how it would work in this specific sentence.

16. Jurafsky and Martin's exercise 21.6

17. Jurafsky and Martin's exercise 21.7

18. Explain how each of the following can affect user behavior, and how that in turn can affect dialog success rates.

    (a) personality of the voice

    (b) wording of the prompts

    (c) speed of the prompts

19. Give 3 performance criteria for spoken dialog systems, and discuss the extent to which the correlate.

20. Telephone linemen need to know the most recent local weather information before they start work. It would be nice to provide a system that will automatically provide weather information for any specified region of West Texas over the telephone. Draw a diagram showing the modules needed for such a system, and for each module specify whether it is a standard component that you could buy or download, something that you would need to customize, or something you would have to build from scratch.