



JUNE 2010 SESSION EXAMINATIONS

LIN3022 Natural Language Processing

Monday 31st May 2010

9:15 ~ 11:15

Section A. Multiple choice

Answer all questions in this section. There is one, and only one, correct answer in each case. Each question carries four (4) marks.

Question 1

If you have a bigram language model, the *local history assumption* implies that:

- The probability of a word in a sequence is computed based on the following word.
- The probability of a word in a sequence is computed based on the previous word.
- The probability of a word in a sequence is independent of the previous or following words.
- The probability of a word in a sequence depends on whether it occurs at the beginning of the sequence or not.

Question 2

Suppose you wanted to build a Finite State Automaton (FSA) to accept a “sheep language”, which has strings like these:

- *ba*
- *baa*
- *baaa*
- ...

In other words, the string consists of an initial /b/ followed by any number of /a/. You have the option of either building a deterministic FSA or a non-deterministic FSA. Which of the following statements is true?

- Only a non-deterministic FSA can handle this language.
- No FSA will handle this language; only a context-free grammar can handle it.
- Both deterministic and non-deterministic FSAs can handle this language.
- No FSA will handle this language; a finite state transducer is needed.

Question 3

The *structural independence assumption* in a Probabilistic Context Free Grammar implies that:

- a) The probability of application of a rule is independent of the syntactic context in which that rule is applied.
- b) Given several rules with the same left hand side (i.e. rules of the form $A \rightarrow \beta$), the probability of any of these rules is independent of the probability of all the others.
- c) The order in which rules are applied to generate a parse tree does not matter to the overall probability of the tree.
- d) The probability of a parse tree for a sentence is independent of the probability of the sentence.

Question 4

In a Part of Speech Tagger based on a Hidden Markov Model, the *symbol emission probabilities* of the model express:

- a) The probability that a tag follows another tag.
- b) The probability that a tag occurs at the start of a sentence.
- c) The probability that a word has a particular tag.
- d) The probability that a word occurs after another word.

Question 5

Suppose you wanted to analyse the following two-sentence discourse, made up of sentences S1 and S2, using Rhetorical Structure Theory.

- (S1) You should check out the new mobile phone showroom.
- (S2) They have some really cheap prices.

This can be viewed as an instance of the MOTIVATION relation. Which of the following best describes the components of the relation?

- a) S2 is the satellite, whereas S1 is the nucleus.
- b) S1 is the satellite, whereas S2 is the nucleus.
- c) Both S1 and S2 are nuclei.
- d) Both S1 and S2 are satellites.

Question 6

If we construct a statistical language model and apply *smoothing*, we usually end up with:

- a) Probabilities that don't sum to a total of 1.
- b) Higher probabilities for more frequent items.
- c) Lower probabilities for observed items
- d) Higher probabilities for observed items

Question 7

Consider the following discourse, made up of two utterances U1 and U2:

(U1) Mary whispered quietly to Ted.

(U2) She had just seen a spider.

In this discourse, what kind of Centering transition do we have between the two utterances?

- a) A RETAIN transition.
- b) A SMOOTH SHIFT.
- c) A CONTINUE transition.
- d) A ROUGH SHIFT.

Question 8

Which of the following is the correct definition of a Probabilistic Context Free Grammar (PCFG)?

- a) A PCFG consists of (i) a set of non-terminal symbols, (ii) a set of terminal symbols, (iii) a set of productions, (iv) a designated start symbol.
- b) A PCFG consists of (i) a set of non-terminal symbols, (ii) a set of terminal symbols, (iii) a set of productions, (iv) a designated start symbol, (v) a probability value assigned to the designated start symbol.
- c) A PCFG consists of (i) a set of non-terminal symbols, (ii) a function assigning a probability value to each non-terminal.
- d) A PCFG consists of (i) a set of non-terminal symbols, (ii) a set of terminal symbols, (iii) a set of productions, (iv) a designated start symbol, (v) a function assigning a probability value to each production.

Question 9

Consider the following input knowledge base (KB) for a Natural Language Generation system, which represents three objects (e1, e2, e3), each of which has three properties (type, colour and size).

	type	colour	size
e1	chair	brown	large
e2	chair	red	small
e3	table	brown	large

Now, your system needs to select the content for a referring expression for e1. You use the Incremental Algorithm, with the preference order *type* >> *colour* >> *size*. What will your referring expression for e1 contain?

- a) Type only (“the chair”).
- b) Type and size (“the large chair”).
- c) Type and colour (“the brown chair”).
- d) Type, colour and size (“the large brown chair”).

Question 10

Suppose you have built a bigram (2-gram) and a 4-gram language model from the same corpus data. You now compare their *reliability* and *discrimination*. Which of the following would you expect?

- a) The bigram model is more reliable, but the 4-gram model has greater discrimination.
- b) The bigram model is superior on both reliability and discrimination.
- c) The 4-gram model is superior on both reliability and discrimination.
- d) The bigram model has greater discrimination but the 4-gram model is more reliable.

Section B.

Answer any 5 questions. Your answers should be clear and concise. Each question carries twelve (12) marks.

1. Explain the difference between top-down and bottom-up parsing.
2. Many NLP problems can be defined in terms of search. Give two examples of NLP tasks which can be defined in this way, and give a brief description of each.
3. What does a Document Planner do in a Natural Language Generation system? Give a brief explanation of one approach to Document Planning.
4. Explain the main features of the Transformation-Based Error-Driven Learning framework for Part of Speech Tagging.
5. Explain how Finite State Transducers can be used to carry out morphological parsing and generation.
6. Give an example of how Rhetorical Structure Theory can be used to analyse a discourse, and explain the relevance of RST for Document Planning for Natural Language Generation.
7. Describe the main principles of Centering Theory.
8. Describe the CKY algorithm for bottom-up parsing.
9. What are the main characteristics and limitations of Markov Models for Natural Language Processing?
10. Using examples, define the task of anaphora resolution and give a brief description of how Centering Theory can be used to solve this task.