Closed Book Examination

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Engineering

Date: Monday 18th January 2010
Time: 14.00 – 16.00

Please answer any THREE Questions from the FIVE questions provided

The use of electronic calculators is NOT permitted
1. a) Assuming a unigram language model, compare the probabilities under that model of the following two word sequences: "I want to eat British food." and "I to food eat British want." Is this relation desirable? Explain why or why not. (5 marks)

b) Under a bigram language model, provide the probability of a sequence which includes one previously unseen word pair. Explain the consequences of this step, stating to what extent it improves behaviour. If it leads to non improvement, explain how to modify the model to improve behaviour. (6 marks)

c) What value is computed by the Forward algorithm? How does the algorithm compute this value efficiently? (4 marks)

d) Hidden Markov Models are often used for speech recognition. What do the emission probabilities represent in speech recognition? Why is a Hidden Markov Model necessary for speech, rather than a basic Markov model? (5 marks)

2. a) Critically discuss the difference between transfer based machine translation and interlingua based machine translation. Discuss advantages and disadvantages of each design. (6 marks)

b) i) Provide the definition of EBMT. (2 marks)

ii) Describe the difference between rule based vs example based MT (EBMT) systems. (4 marks)

c) The distinction between rationalistic and empiricist approaches has dominated the field of NLP. Critically discuss the implications of this distinction for the field. (8 marks)
3. a) What is the scope of automatic term recognition? Provide the definition of “term”. Give at least four applications of automatic term recognition in NLP and discuss what problems may arise when tackling term recognition. (5 marks)

b) “One morning I shot an elephant in my pajamas. How he got into my pajamas I don’t know” Groucho Marx, Animal Crackers, 1930

The humorous reading in the sentence above is due to what type of ambiguity? (2 marks)

Draw two parse trees for this ambiguous sentence; the humorous reading where the elephant is in the pajamas and the reading where a human being did the shooting in his pajamas. (6 marks)

c) Text mining techniques are used in many applications. Describe and define the main steps in text mining and provide examples of typical application areas where the technology is being used. (4 marks)

d) Prof. Maddox announced the formation of La Jolla Genomics together with his two sons. La Jolla Genomics is headquartered in the Maddox family’s hometown of La Jolla, Ca.

Identify the named entities an information extraction system would extract and assign them a category (type). (3 marks)
4. a) In a real application, the parsing process must be robust. Explain what this means. Give at least two examples of phenomena that make this necessary. (4 marks)

b) Explain the use of an agenda in chart parsing and illustrate for at least two parsing strategies. (4 marks)

c) Here is a small grammar for an artificial language:

\[
\begin{align*}
V & \rightarrow \text{burn} \\
N & \rightarrow \text{chillies} \\
\text{Adj1} & \rightarrow \text{hot} \\
\text{Adj2} & \rightarrow \text{red} \\
\text{Det} & \rightarrow \text{these} \\
\text{Conj} & \rightarrow \text{and} \\
S & \rightarrow \text{NP} \ \text{VP} \\
\text{NP} & \rightarrow \text{Det} \ \text{Adjp} \ N \\
\text{Adjp} & \rightarrow \epsilon \\
\text{Adjp} & \rightarrow \text{Adj1} \ \text{Adjp} \ \text{Adj2} \\
\text{VP} & \rightarrow \text{V} \\
\text{VP} & \rightarrow \text{VP} \ \text{Conj} \ V
\end{align*}
\]

The grammar covers strings like:

- These chillies burn
- These hot red chillies burn and burn

but also:

- These chillies burn and burn
- These hot red chillies burn

Add a list-valued feature to the grammar to restrict it so that the number of occurrences of hot, red and and is always the same. The following will help you get started:

\[
\begin{align*}
S & \rightarrow \text{NP} \ \text{VP} \\
\langle \text{NP count} \rangle & = \langle \text{VP count} \rangle \\
\text{Adjp} & \rightarrow \epsilon \\
\langle \star \ \text{count} \rangle & = \text{empty}
\end{align*}
\]

Why is it significant that this can be done? (12 marks)
5. a) What syntactic knowledge can be used in resolving the ambiguity in word senses? Give an example and explain how syntactic knowledge can help and also its limitations. (3 marks)

b) F-measure is an evaluation metric for Information Retrieval and Information Extraction Systems. Define balanced F-measure, using true positives (TP), false positives (FP), and false negatives (FN). (4 marks)

c) In the Vector Space Model (VSM), a document is represented as a vector, and each dimension corresponds to a term.

i) If a term occurs in the document, its value in the vector is non-zero, and can be calculated using term weighting. Describe the following commonly used weighting methods: TF, IDF, and TF-IDF. (4 marks)

ii) Define the cosine similarity measure, which is often used to calculate similarity between vectors of documents. (4 marks)

iii) The VSM has many applications in the field of natural language processing. Describe how this model can be used to automatically find semantically similar terms, given a set of documents and the assumption that semantically similar words tend to occur in similar contexts. (5 marks)

END OF EXAMINATION