vak #: 716631

Exam
Speech and Language Processing 1
5 November 2008

Introduction

This exam Speech and Language Processing 1 consists of 16 multiple choice questions (80 points) and two open questions (20 points). You may use the book "Speech and Language Processing", the slides and your notes.
The open questions may be answered in Dutch as well as in English.
The numbered grammar referred to in one of the multiple choice questions can be found in the final section of this document.
When the time is up or when you are finished you should hand in the answer form for the multiple choice questions as well as the answer of the open question (on a separate sheet of paper).

Tip:

> First fill in your answers on this question form; check the answers when you have completed all the questions; then fill in your answers on the answer form.

Success

Multiple choice questions

1. The language ABC is the regular language over the alphabet $\{a, b, c\}$ that contains all and only those strings that have *abc* as a substring. The language $s-ABC$ is the regular language over the alphabet $\{a, b, c\}$ that contains all and only those strings that have a single occurrence of *abc* as a substring.

   The table below shows the transition table of a non-deterministic finite state automaton M over the alphabet $\{a, b, c\}$; the initial state of M is 1; the only final (accepting) state is 4.

	Input		
State	a	b	c
1	1,2	1	1
2	$\emptyset$	3	$\emptyset$
3	$\emptyset$	$\emptyset$	4
4	4	4	4

   We say that a regular language L *is accepted by* M if for all elements x in L and only for elements in L it holds that: M accepts x.

   The regular language accepted by M is:

   (a) The language ABC.

   (b) A proper superset of ABC.

   (c) A proper subset of ABC, a superset of $s - ABC$.

   (d) The language $s - ABC$.

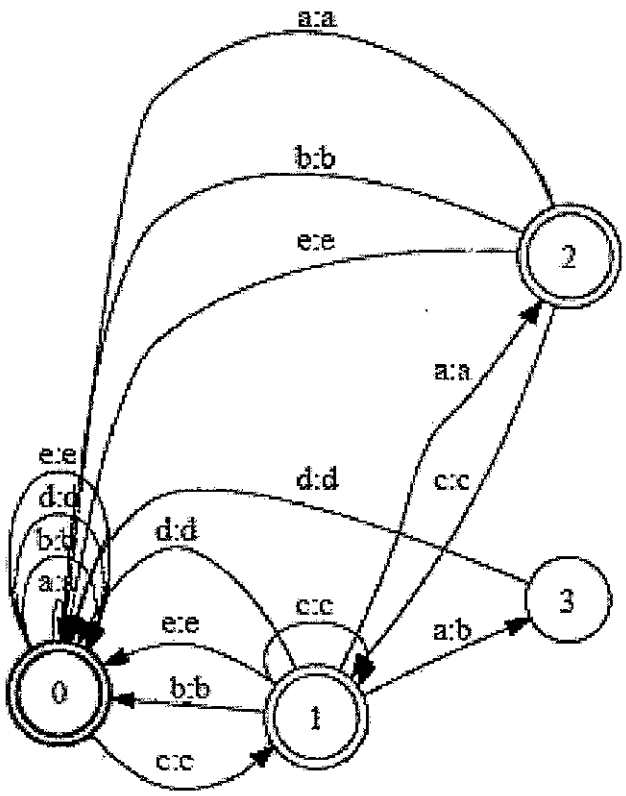
2. Which kind of word formation process has taken place to form the word *bombings*?

   (a) Inflection

   (b) Derivation

   (c) Both inflection and derivation

   (d) It depends whether the stem *bomb* is a noun or a verb



3. Consider the finite state transducer shown above.

   Which of the following rewrite rules properly specifies the context in which a is rewritten as b in this transducer?

   (a) $a \rightarrow b / \_\ d$

   (b) $a \rightarrow b / c\ \_\ d$

   (c) $a \rightarrow b / cc\ \_\ d$

   (d) $a \rightarrow b / \left\{ \begin{array}{c} a \\ b \\ d \\ e \end{array} \right\} c\ \_\ d \left\{ \begin{array}{c} a \\ b \\ d \\ e \end{array} \right\} \#$

4. In British (not American) spelling, the past tense of verbs such as *transfer* gets an extra *r* before the suffix. Of course, the *r* is not doubled if the verb already ends in two *r*'s:

Intermediate	*Surface*
transfer+ed	transferred
refer+ed	referred
deter+ed	deterred
err+ed	erred

Below you see the state-transition tables for two transducers designed to handle the r-doubling (or lack of it) in the examples above. We use PC-Kimmo notation where 0 is the empty symbol, + is the morpheme boundary symbol, and @ is the 'other' symbol. V and C stand for vowels and consonants, respectively. Final states are indicated with a colon (:) and non-final states with a dot after the state number.

Which of these transducers can correctly generate and recognize all of the four examples given above? (You don't need to worry about other words.)

```
RULE "One" 7 9
     V r 0 e d + C @ #
     V r r e d 0 C @ #
1:   2 1 0 2 1 0 1 1 1
2.   2 2 0 3 2 1 2 1 1
3.   1 4 0 3 1 0 1 1 0
4.   0 0 0 0 0 5 0 0 0
5.   0 0 6 0 0 0 0 0 0
6.   0 0 0 7 0 0 0 0 0
7.   0 0 0 0 1 0 0 0 0
```

```
RULE "Two" 8 7
     V r 0 e + @ #
     V r r e 0 @ #
1:   2 1 0 2 1 1 1
2.   1 2 0 3 1 2 1
3.   1 4 0 3 1 1 1
4.   0 0 0 0 5 0 0
5.   0 0 6 0 0 0 0
6.   0 0 0 7 0 0 0
7.   0 0 0 0 0 8 0
8.   0 0 0 0 0 0 1
```

(a) Only transducer "One"

(b) Only transducer "Two"

(c) Both transducers

(d) Neither of the transducers

5. Which of the following statements is true?

   (a) Morphological parsing takes spelling into account; stemming does not

   (b) Morphological parsing outputs morphological features; stemming does not

   (c) Morphological parsing can handle words that have been formed through derivation; stemming cannot

   (d) Stemming can deal with irregular word forms; morphological parsing cannot

6. Each of the two headlines below has two possible meanings: one is the intended one and one is a humorous (but unintended) meaning.

   - Headline 1: Fear of Bird Flu Grips Turkey
   - Headline 2: Sisters Reunited After 18 Years In Queue

   For which of the two headlines is the humorous meaning caused by syntactic ambiguity?

   (a) Headline 1

   (b) Headline 2

   (c) Both Headline 1 and Headline 2

   (d) Neither of the two headlines

7. Consider the following sentences. The sentences marked with * are not proper English sentences.

   the angry player is young
   the young coach is angry
   the former coach is angry at the player

   * the coach is former
   * the angry at the player coach is young
   * the coach is young at the player

Now consider the following context-free grammars. Which of these grammars generates the first three sentences given above, but none of the sentences marked with * ?

Grammar (i)

Rules:

S → NP VP	VP → Vbe AP	
NP → Det NP	AP → AP PP	
NP → AP N	AP → Adj	
NP → N	PP → P NP	

Lexicon:

N → player | coach
Adj → young | angry | former
P → at | on
Vbe → is
Det → the

Grammar (ii)

Rules:

S → NP VP	NP → N	
VP → Vbe AP	AP1 → Adj1	
VP → Vbe AP3	AP2 → Adj2	
NP → Det NP	AP3 → Adj3	
NP → AP1 N	AP → AP1 PP	
NP → AP2 N	PP → P NP	

Lexicon:

N → player | coach
Adj1 → angry
Adj2 → former
Adj3 → young
P → at | on
Vbe → is
Det → the

Grammar (iii)

Rules:

S → NP VP (AP1)
VP → Vbe
NP → N
NP → Det (AP2) N
AP1 → (Adj1 | Adj3) (PP)
AP2 → Adj1 | Adj2 | Adj3
PP → P NP

Lexicon:

N → player | coach
Adj1 → angry
Adj2 → former
Adj3 → young
P → at | on
Vbe → is
Det → the

(a) Grammar (i)

(b) Grammar (ii)

(c) Grammar (iii)

(d) None of these grammars

8. In French, the past participle of the verb *être* (to be) can have different forms depending on the subject's gender. For instance, we have:

   - *Il est assis* (He is seated)
   - *Elle est assise* (She is seated)

   (*est* is the third person singular form of the verb *être*)

   Now consider the sentence *Il est assise.* What is wrong with this sentence?

   (a) It is syntactically ambiguous

   (b) It violates agreement rules

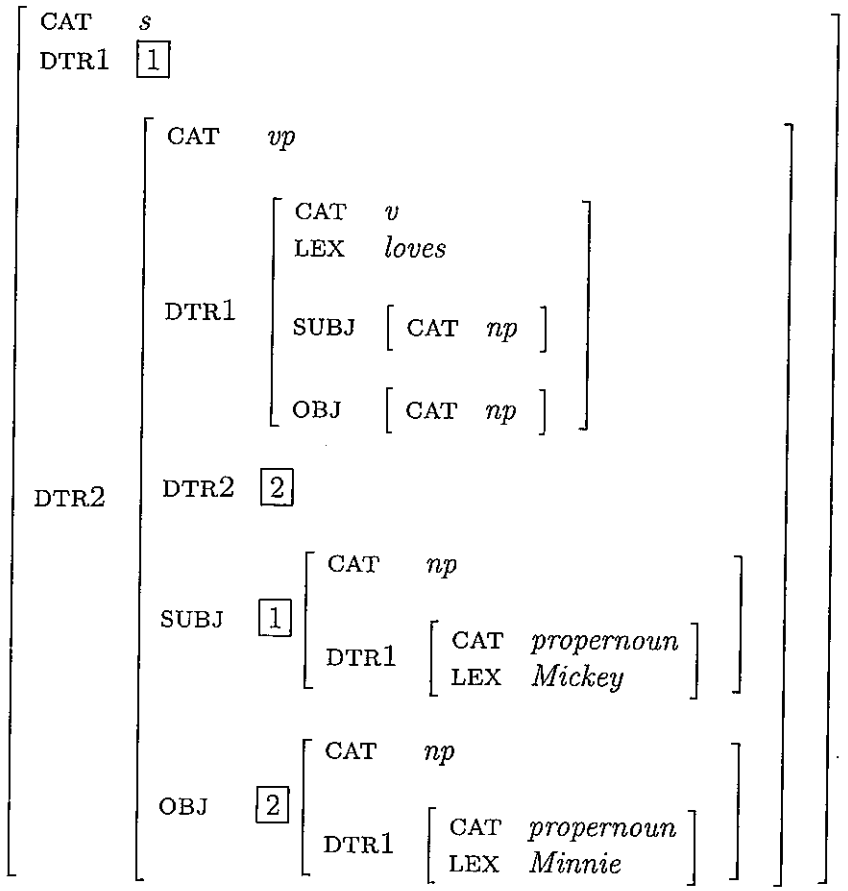
   (c) It violates morphological rules

   (d) It uses the wrong subcategorization frame for the verb

9. Which of the following feature structures is associated with the sentence "Mickey loves Minnie" by the grammar given in the Appendix?

(i)
$$\begin{bmatrix} \text{CAT} & s \\ \text{DTR1} & \boxed{1}\begin{bmatrix} \text{CAT} & np \\ \text{LEX} & \textit{Mickey} \end{bmatrix} \\ \text{DTR2} & \begin{bmatrix} \text{CAT} & vp \\ \text{LEX} & \textit{loves} \\ \text{SUBJ} & \boxed{1} \\ \text{OBJ} & \begin{bmatrix} \text{CAT} & np \\ \text{LEX} & \textit{Minnie} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(ii)


CAT s
DTR1 1
CAT vp
DTR1
CAT v
LEX loves
SUBJ CAT np
OBJ CAT np
DTR2
DTR2 2
SUBJ 1
CAT np
DTR1
CAT propernoun
LEX Mickey
OBJ 2
CAT np
DTR1
CAT propernoun
LEX Minnie

(iii)
$$\begin{bmatrix} \text{CAT} & s \\ \text{DTR1} & \boxed{1} \\ \\ \text{DTR2} & \begin{bmatrix} \text{CAT} & vp \\ \\ \text{DTR1} & \begin{bmatrix} \text{CAT} & v \\ \text{LEX} & loves \\ \text{SUBJ} & \boxed{1} \\ \\ \text{OBJ} & \boxed{2} \end{bmatrix} \\ \\ \text{DTR2} & \boxed{2} \begin{bmatrix} \text{CAT} & np \\ \\ \text{DTR1} & \begin{bmatrix} \text{CAT} & propernoun \\ \text{LEX} & Minnie \end{bmatrix} \end{bmatrix} \\ \\ \text{SUBJ} & \boxed{1} \begin{bmatrix} \text{CAT} & np \\ \\ \text{DTR1} & \begin{bmatrix} \text{CAT} & propernoun \\ \text{LEX} & Mickey \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(a) Feature structure (i)

(b) Feature structure (ii)

(c) Feature structure (iii)

(d) None of these feature structures

10. Consider the following grammar.

Rules:	Lexicon:
1. S → AP Noun	Noun → house | book | ...
2. S → Noun	Adj → nice | red | angry | ...
3. AP → AP Adj	
4. AP → Adj	

Assume you are using a top-down depth-first parser that applies the grammar rules in the order given above. What will happen when trying to parse the input *nice angry house*?

(a) The parser will fail

(b) The parser will loop forever

(c) The parser will succeed and output one parse

(d) The parser will succeed and output more than one parse

11. What is a correct Part of Speech tagging for the sentence *that oil can burn* ? (PoS tags are as follows: Det - determiner; Art - article; N - noun; V - verb; Pron - pronoun)

(a) *that/Art oil/N can/N burn/V*

(b) *that/Det oil/N can/V burn/V*

(c) *that/Pron oil/V can/V burn/V*

(d) *that/Pron oil/N can/V burn/V*

12. What is the best Part of Speech of the word *that* in the sentence *Did he go that far?*

(a) Conjunctive

(b) Adjective

(c) Adverb

(d) Pronoun

13. Given the "document" *abcabcabcabc* over the three letter alphabet a, b, c, the value for the maximum likelihood estimate of the conditional probability $P_{mle}(c|ab)$ is:

(a) 1.0

(b) 0.0

(c) 0.5

(d) 0.75

14. What is the advantage of the Maximum Entropy Markov Model (MEMM) over the Hidden Markov Model (HMM) for a statistical Part-of-Speech tagger

    (a) MEMM allows a more efficient Viterby algorithm than HMM
    (b) MEMM is more flexible for incorporating statistical models of features of words and contexts than HMM
    (c) MEMM allows smoothing of observation probabilities where HMM does not allow that.
    (d) MEMM can be applied for more types of languages than HMM

15. The word *"te"* is a misspelling. We use the noisy channel model to compute the most likely correct word. The table below shows the possible correct words (column *cw*), their prior probabilities and the conditional probabilities $P("te"|cw)$, the probability that the correct word in column *cw* is misspelled as *"te"*.

cw	$P(cw)$	$P(te\|cw)$
ate	0.0001	0.00001
ten	0.0015	0.001
tea	0.0005	0.001
the	0.005	0.00002

    What is the most likely correct word?

    (a) ate
    (b) ten
    (c) tea
    (d) the

16. What is the *minimum edit distance* between the words *"seventy"* and *"seventeen"* when we use the Levenshtein distance in which *insertion* and *deletion* have a cost 1 and *substitution* has cost 2 ?

    (a) 3
    (b) 4
    (c) 5
    (d) 6

Open questions

Question A

Canonical transcription is the official transcription of a word, given the language and, sometimes, the dialect. Explain with "real life" examples, why this canonical transcription is causing trouble for automatic speech recognition.

Question B

The basic idea of human language communication (spoken and written) is as follows: You speak with your mouth and the other is listening with her ears, or the other is writing a text with his hands and you are reading with your eyes. So, communication is between mouth and ears or body and eyes. Explain why in essence, this is not correct and how real human-to-human communication works.

Appendix: Grammar for question 9

Rules

$$
\begin{array}{lcl}
s \rightarrow np \;\; vp & & \\
\quad \langle s \text{ DTR1} \rangle & = & \langle np \rangle \\
\quad \langle s \text{ DTR2} \rangle & = & \langle vp \rangle \\
\quad \langle vp \text{ SUBJ} \rangle & = & \langle np \rangle
\end{array}
$$

$$
\begin{array}{lcl}
vp \rightarrow v \;\; np & & \\
\quad \langle vp \text{ DTR1} \rangle & = & \langle v \rangle \\
\quad \langle vp \text{ DTR2} \rangle & = & \langle np \rangle \\
\quad \langle vp \text{ OBJ} \rangle & = & \langle np \rangle
\end{array}
$$

$$
\begin{array}{lcl}
np \rightarrow propernoun & & \\
\quad \langle np \text{ DTR1} \rangle & = & \langle propernoun \rangle
\end{array}
$$

Lexicon

$$
\begin{array}{ll}
Mickey & \langle \text{LEX} \rangle = Mickey \\
& \langle \text{CAT} \rangle = propernoun
\end{array}
$$

$$
\begin{array}{ll}
Minnie & \langle \text{LEX} \rangle = Minnie \\
& \langle \text{CAT} \rangle = propernoun
\end{array}
$$

$$
\begin{array}{ll}
loves & \langle \text{LEX} \rangle = loves \\
& \langle \text{CAT} \rangle = v \\
& \langle \text{SUBJ CAT} \rangle = np \\
& \langle \text{OBJ CAT} \rangle = np
\end{array}
$$