Advanced AI Techniques (WS05)

Excercise sheet 4

Deadline: 06.12.05

Excercise 1 (4 points)

a.) Part-of-speech-tagging *is the task of assigning (grammatical) word categories or "tags" to the individual words in a sentence. Manually assign tags to the words in the sentence* ***"I saw a man with a telescope"****. Use the following set of tags:*

- ***N*** *- a noun.*
- ***Vi*** *- an intransitive verb.*
- ***Vt*** *- a transitive verb.*
- ***D*** *- a determiner (like "the","a").*
- ***PR*** *- a preposition.*

b.) Parsing *a sentence means recovering its internal grammatical structure in a so-called parse tree. Find two different possible parse trees for the sentence* ***"I saw a man with a telescope"*** *to show the syntactic ambiguity in this sentence. Use the tags given above and the following non-terminals in the tree:*

- ***S*** *- the whole sentence.*
- ***NP*** *- a noun phrase.*
- ***VP*** *- a verb phrase.*
- ***PP*** *- a prepositional phrase, which is a preposition followed by a noun phrase. Prepositional phrases can occur both in noun phrases and verb phrases.*

Excercise 2 (4 points) Definite clause grammars *are an extension to context free grammars that include arguments in non-terminal symbols and unification.*

- *Give the most general unification (substitution and result of unification) for the following pairs of non-terminals, or say why they cannot be unified:*
  1. $PN(Number, Case)$ *and* $PN(singular, Case)$
  2. $NP(N)$ *and* $VP(singular)$

3. $VP(Any, accusative)$ *and* $VP(Number, accusative)$
4. $A(s(N))$ *and* $A(s(s(M)))$

- *As explained in the lecture, DCGs are strictly more expressive than CFGs and can, for example, represent the language* $\{a^n b^n c^n | \; n \in \mathbb{N}\}$*. Show how to derive the sentence* $aabbcc$ *using the grammar presented in the lecture.*

Excercise 3 (4 points)

Consider the following bigram statistics of a corpus over the vocabulary $\{a, b, c, d\}$*:*

	a	*b*	*c*	*d*	$N(w_x)$	$T(w_x)$	$Z(w_x)$
a	*20*	*20*	*20*	*0*	*60*	*3*	*1*
b	*0*	*0*	*60*	*0*			
c	*0*	*0*	*400*	*0*			
d	*...*	*...*	*...*	*...*			

The word in row i *is followed* n *times by the word in column* j *in the text, e.g.* b *is followed by* c *60 times.*

a.) Assume we want to use Witten-Bell smoothing to compute the bigram probability estimates. Compute the necessary statistics $N(w_x)$*,* $T(w_x)$*,* $Z(w_x)$ *for* $w_x \in \{a, b, c\}$*. From these, compute the bigram probabilities* $p^*(w_i|w_x)$ *for* $w_i \in \{a, b, c, d\}$ *and* $w_x \in \{a, b, c\}$ *according to the formulas given in the lecture.*

b.) A more advanced way of estimating probabilities for bigrams that have never been observed in the training data is to use backoff smoothing. In the case of bigrams, backoff smoothing means reverting back to (corrected) unigram probabilities if the bigram has never been observed. As an example, for $w_x = b$ *compute the backoff factor* $\alpha(w_x)$ *and then the final probability estimates for the unseen bigrams starting with* b*.*