DALHOUSIE UNIVERSITY
Inspiring Minds

Winter 2017 (Jan9-Apr10)
Faculty of Computer Science

CSCI 4152/6509 — Natural Language Processing

Assignment 4

Due: *by midnight, Monday, Apr 10, 2017*
Worth: 120 marks
Instructor: Vlado Keselj, CS bldg 432, 902.494.2893, vlado@dnlp.ca
TA: Magdalena Jankowska, jankowsk@cs.dal.ca

Assignment Instructions:

All answers must be submitted throught the SVN by the duedate.

The Lab questions must be submitted in the appropriate directories as specified in the labs (and questions).

The answer files for the other questions, should be submitted in the SVN directory *csuserid*/`a4`, similarly to the previous assignment.

All files must be plain-text files, unless specified differently by the question.

In questions 2 and 3 you need to draw a graph and some parse trees, and in those cases you can submit .pdf, .jpg, or similar files. You can draw those graphs in some programs and produce a PDF file, or you can draw them by hand and submit a photo of your drawing. If you submit such photo, it must be legible.

1) (24 marks, files in *csuserid*/`lab6`) Complete the Lab 6 as instructed. In particular, you will need to properly:

a) (4 marks) Submit the file '`parse.prolog`' as instructed.

b) (4 marks) Submit the file '`dcg.pl`' as instructed.

c) (4 marks) Submit the file '`dcg-ptree.prolog`' as instructed.

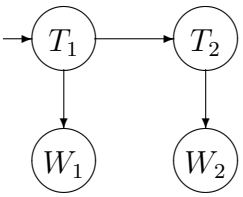
d) (4 marks) Submit the file '`dcg-agr.prolog`' as instructed.

e) (4 marks) Submit the file '`dcg-pcfg.prolog`' as instructed.

f) (4 marks) Submit the file '`dcg-agr2.prolog`' as instructed.

2) (70 marks, file: `a4q2.txt`, `a4q2.pdf`, `a4q2.jpg`, or similar)
A number of ambiguities in POS tagging comes from the combination of words "flies like"

since 'flies' can be a noun or verb, and 'like' can be a verb or preposition. One way to determine which combination of tags is more likely is to consider the following HMM model:



where we will consider W_1 = 'flies' and W_2 = 'like'. The values of the variables T_1 and T_2 come from the set of POS tags, and we will consider only the set $\{N, V, P\}$, which stand for nouns, verbs, and prepositions, respectively. We use a corpus to learn the model parameters, and based on this corpus we find that nouns occur approximately 3000 times, verbs 500 times, and prepositions 2000 times. We use these counts to make the estimate for $P(T_1)$. The estimates for $P(W_i|T_i)$ and $P(T_{i+1}|T_i)$ are based on the following counts:

T_i	W_i	count
N	flies	400
V	flies	200
P	flies	0
N	like	10
V	like	100
P	like	200

and

T_i	T_{i+1}	count
N	N	400
N	P	400
N	V	400
P	N	500
P	P	20
P	V	100

T_i	T_{i+1}	count
V	N	300
V	P	400
V	V	200

(a) (20 marks) Calculate the conditional probability tables for the model using the given data. Do not use any smoothing.

(b) (10 marks) Draw the corresponding factor graph.

If you are submitting your answer in a textual file, you can just describe what are the nodes of the factor graph and how they are connected, instead of drawing it; or you can draw it using ASCII characters, and assuming a fixed-width font.

(c) (40 marks) Using the message-passing algorithm discussed in class, find the most likely tags for T_1 and T_2 for given values of W_1 = 'flies' and W_2 = 'like'. The intermediate results, such as messages, must be included.

3) (26 marks, file: `a4q3.txt`, `a4q3.pdf`, `a4q3.jpg` or similar)
Let us consider the sentence:

they are hunting dogs

and the following grammar:

```
 S  →  NP VP     VBP  →  are       NNS  →  dogs
NP  →  they      NP   →  VBG NNS    VP  →  VBP VBG
VP  →  VBP NP    VBG  →  hunting    VP  →  VP NNS
```

where `S` is the start symbol, the non-terminal symbols are { `S`, `NP`, `VP`, `VBP`, `VBG`, `NNS` } and the terminals are { `they`, `are`, `hunting`, `dogs` }.

a) (5 marks) Briefly explain why the grammar is in CNF (Chomsky Normal Form).

b) (15 marks) Using the CYK algorithm, find all parse trees for the sentence "they are hunting dogs". Your answer must include the chart and the final trees.

c) (6 marks Indicate heads and dependencies in one of the trees. Show that tree in the bracketed notation.