Examen de premiere session de decembre 2018
X3IT040 (Statistical and symbolic language modeling)
All documents allowed. You can also use electronic devices to consult information. Not to communicate and receive assistance.
75 mn.

1 Search of an optimum dependency tree

The scores of the dependencies of the sentence "Jean n'y voit qu'un jeu" are the following (the governors are on the first column, the dependents are on the first line):

	$	*Jean*	n'	y	*voit*	qu'	*un*	*jeu*
$		0,1			0,8			0,1
Jean								0,2
n'		0,1				0,6		
y		0,1	0,2					
voit		0,5	0,3	0,3		0,4		0,7
qu'			0,5				0,2	
un								
jeu		0,2		0,7	0,2		0,8	

1. Draw the best dependency graph (with cycles).
2. Where is the cycle of the graph ?
3. Use the spanning tree algorithm to find the best non-projective dependency tree.

2 Active learning

We will use the following notations

- Σ is the alphabet. Here $\Sigma = \{a, b\}$
- The table is $\langle S, E, T \rangle$
- S is the set of states (some being red, others blue). $S \subset \Sigma^*$
- E is the set of experiments. $E \subset \Sigma^*$
- T is the table. $\forall u \in S$, $\forall v \in E$, $T[u][v]$ contains the result of the MQ for uv.

2.1 Closed by prefixes, by suffixes

One rule which must be followed in algorithm L^* is that set S should be closed by prefixes and set E should be closed by suffixes. We intend to understand why.

1. A set E is closed by suffixes if $\forall wu \in E$, we have $u \in E$. Are the following sets closed by prefixes?
   - $\{\lambda, a, aa, ba\}$
   - $\{\lambda, a, aa, bb\}$

2. Let us consider this table:

	λ	a	aaa
λ	1	1	0
a	1		
b			

Is this table consistent? closed? complete? What happens if you try to build the automaton? Justify why E being suffix-complete is important.

3. Build an example showing why the set S should be prefix complete. Develop the example as above.

2.2 About counter-examples

In the classical L^* algorithm, when a counter-example is produced in response to a strong equivalence query, this example is treated in such a way as to introduce more rows. More precisely, if w is a counterexample, than the algorithm adds a new row for each string ua where u is a prefix of w.

An alternative idea has also been tried: use the counterexample to create new columns, not new rows.

1. Define how a table is to be modified (as above) when counterexample w is received.
2. We suppose $\Sigma = \{a, b\}$ and the Oracle has returned the following queries:
   - (round 1) $(\lambda, 1)$, $(a, 1)$, $(b, 1)$ and an equivalence query was made (describe this query).
   - (round 2) Counterexample *aabab* was returned. Draw the table at this point.
   - What membership queries are you going to make at this point? We suppose the target used by the Oracle is represented below. Run the algorithm from this point on.
   - At this point is your table adequate to make another equivalence query? If not, what do you do? Run the algorithm until you converge.

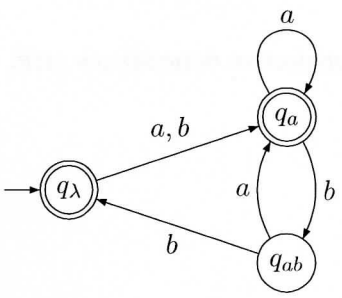

Figure 1: The target automaton

3 Statistical Language modelling

1. What is the *back-off* (repli) for a language model?
2. Write the formulation to compute the probability of a word sequence W with length L with an n-gram model or order N. (You can use an example, e.g. a bi- or tri-gram then generalize)
3. What is the main difference between a statistical and a neural language model?
4. Why do we need gated recurrent units? Explain which problem they solve.
5. What is a conditional language model? (in opposition to non-conditional language model). Illustrate your explanation with an example.