Name: **(You may work in pairs)**

Statistical Lexical Category Disambiguation

Tutorial 5

The Parser Files

- We'll continue looking at the Viterbi tagger implementation in Prolog: `Tutorial5/viterbi2.pl`

1. Reduced relative clauses

Using the TNT tagger, consider the ambiguous verb in reduced relative clauses (file `relatives.txt`). Again, using both `susanne` and `wsj` models, identify which sentences each model predicts will be difficult. Which best matches the behavioral evidence?

- The man taken to the police station escaped.
- The man taken to the police station escaped.
- The man fought at the police station escaped.
- The man held at the police station escaped.
- The man who was fought at the police station escaped.
- The man who was held at the police station escaped.

2. Viterbi in Prolog

Recall the Viterbi algorithm for category disambiguation, which is much more efficient because it only computes the best path, and thus avoids keeping track of any paths that cannot lead to the best bath. The HMM has been modified slightly to distinguish transitive (`vt`) and intransitive verbs (`vi`), and allow a simple adverbial to appear after the verbs (but see the probabilities). The aux state has been removed.

- The man <u>fought</u> yesterday.
- The man <u>held</u> the man.

You can try it again, just to make sure it's running:

```
?- mps([the,man,fought,the,man],S).
```

a) Show the best POS sequence, and it's probability, for both sentences above:

3. Reduced relative clauses

Now recall the ambiguous verbs in (reduced) relative clauses (file `relatives.txt`), which we will heavily simplify here (removing the PP that could appear after the first verb, and reducing the lexicon):

i.	The man <u>taken</u> yesterday fought.	*Unambiguous past-part: vpp*
ii.	The man <u>fought</u> yesterday fought.	*Ambiguous: vpp, vi, vt*
iii.	The man <u>held</u> yesterday fought.	*Ambiguous: vpp & vt*

Modify the HMM to accept these sentences too (don't worry about object relative clauses):

a) You will need to add a state for past-participle (`vpp`).
b) What transitions do you need to add from current states in the HMM, to `vpp`?
c) What transitions do you need to add from `vpp` to current states?
d) Estimate the probabilities for these transitions.
e) Adjust the old and new transition probabilities so all transitions leaving a state sum to 1.0.
f) Add the new words you need, that can be emitted by each state, again, adjust the probability (these needn't sum to 1.0, since you only have a partial lexicon, and will generally be small).

Remember you don't need to worry about ruling out ungrammatical sequences, just bi-gram transitions. Also remember, some words are ambiguous, and can be emitted by more than one state.

Show the transitions and emissions that you added:

4. Accounting for the human preferences:

a) Recall the explanation for the difficulty in these sentences for why (iii) is easy for people, while (ii) is more difficult:

b) Show the POS sequences your HMM assigns to the three sentences:

c) If necessary adjust the probabilities in your HMM so that your tagger explains/predict the difficulty people have with (ii) only. Now show the sequences you get for all 3 sentences:

5. Unreduced relative clauses

Extend the HMM somewhat further, to allow the following unreduced relative clauses (and don't worry about other kinds of relative clauses).

- The man who was taken yesterday fought.
- The man who was fought yesterday fought.
- The man who was held yesterday fought.

a) You will need to add states for the relative pronoun "who", and for the passive auxiliary "was", and of course transitions to/from the new states. Show the transitions and emissions you added:

b) Show below the POS sequence your HMM assigns to all 8 example sentences in this tutorial (Exercises 2-5):