ANLP Tutorial Exercise Set 1 (for tutorial groups in week 2) WITH SOLUTIONS

v1.1
School of Informatics, University of Edinburgh
Sharon Goldwater

This week's tutorial sheet is primarily aimed at students who have less background in probability theory, to make sure that you have gone through the reading material and are learning to solve problems on your own. If this material is entirely new to you, don't expect that you will necessarily have absorbed all of it by now (this will take some time), but you should definitely try to work through these exercises and come with specific questions if you have trouble.

Exercise 1.
Suppose we have a group of first- and second-year students, each of whom comes from either the UK, Germany, or China. We choose a student uniformly at random from this group. Let F = "the student is in first year", U = "the student is from the UK", C = the student is from China", and G = "the student is from Germany". For this group of students, $P(U) = 0.6$, $P(G) = 0.3$, $P(C) = 0.1$, $P(F\,|\,U) = 0.7$, $P(F\,|\,G) = 0.5$, $P(F\,|\,C) = 0.6$.

a) What is the probability that we've chosen a first-year student?

b) What is the probability that the student is from the UK if it is a first-year student?

Solution 1.

a) Use the rule of total probability:

$$\begin{aligned} P(F) &= P(U)\,P(F\,|\,U) + P(G)\,P(F\,|\,G) + P(C)\,P(F\,|\,C) \\ &= (.6)(.7) + (.3)(.5) + (.1)(.6) \\ &= .63 \end{aligned}$$

b) Now use Bayes' Rule:

$$\begin{aligned} P(U\,|\,F) &= \frac{P(F\,|\,U)\,P(U)}{P(F)} \\ &= \frac{(0.7)(0.6)}{.63} \\ &= 2/3 \end{aligned}$$

Exercise 2.
(Exercise 4.6 from the Basic Probability Theory tutorial)
Suppose I want to choose a two-word name for a rock band via a random process. The first word will be chosen with uniform probability from the set {`yellow`, `purple`, `sordid`, `twisted`}. The second word will be either `quake`, with probability 1/3, or `revolution`, with probability 2/3.

a) What's the sample space for this experiment? How many outcomes are in it?

b) If I choose the first and second words independently, what is the probability that my rock band will be called `yellow revolution`?

c) Instead, I decide to condition the second word on the first. If the first word is a color, then I'll pick `quake` as the second word with probability 1/6. Assuming I still want the overall probability of a name with `quake` to be 1/3, what is the probability of `quake` if the first word is *not* a color? (Hint: use a joint probability table to help you answer this question. You may not need to fill in the entire table.)

Solution 2.

a) The sample space is all ordered two-word pairs using the words provided, and contains 8 possible pairs.

b) $P(\texttt{yellow revolution}) = P(\texttt{yellow})\, P(\texttt{revolution}) = (1/4)(2/3) = 1/6$.

c) Here's one way to solve the problem. Let C be the event that the first word is a color. Since there are two color words out of four equally likely words, $P(C) = 1/2$. Also, let Q be "second word is `quake`" and R be "second word is `revolution`". The initial JPT contains all the marginals: we were given $P(R)$ and $P(Q)$, we just computed $P(C)$, and $P(\neg C) = 1 - P(C)$. Also, we add $P(C,Q) = P(C)P(Q|C) = (1/2)(1/6) = 1/12$:

	Q	R	
C	1/12		1/2
$\neg C$			1/2
	1/3	2/3	

Now we can fill in $P(Q,\neg C) = P(Q) - P(Q,C) = 1/4$. And finally, compute $P(Q|\neg C) = \frac{P(Q,\neg C)}{P(\neg C)} = \frac{1/4}{1/2} = 1/2$.

Exercise 3.

Suppose I decide to probabilistically generate a "word", which can contain only characters in the set $\{\texttt{a},\texttt{b},\texttt{c},\texttt{d},\texttt{e}\}$. To generate the word, I start by generating the character `a`. Then, with probability q, I generate a single `b` and stop, otherwise I choose one of the other four characters (`a`, `c`, `d`, `e`) uniformly at random and keep going. I continue this process, always either generating a single `b` with probability q and stopping, or choosing one of the other four characters uniformly at random and continuing.

a) What are the possible words I might generate using this process? Draw an FSA that accepts all and only these words.

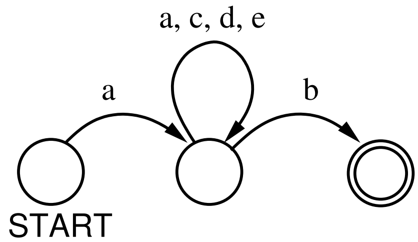
b) If L is a random variable representing the length of a word, give an equation for $P(L = n)$.

c) If $q = 0.3$, what is the probability that my generated word has 4 characters? What is the probability of generating the word `aaab`?

d) If $q = 0.3$, what is the probability of generating the word `aaab` given that I generate a 4-character word?

Solution 3.

a) The words are strings of characters consisting of an `a` followed by any sequence of `a`'s, `c`'s, `d`'s, and `e`'s followed by exactly one `b`. The FSA is:



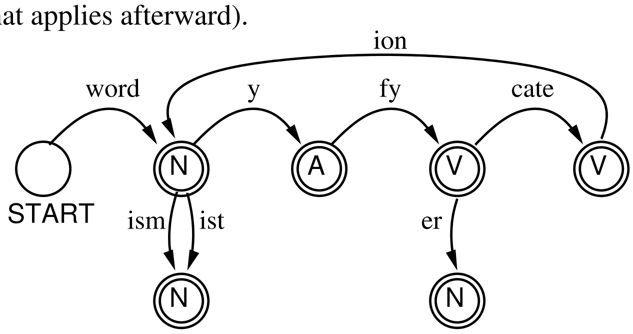
b) The equation is $P(L = n) = (1-q)^{n-2}(q)$. (Note the first character is "free", thus the $n-2$; also the fact that there are multiple different continuation characters is irrelevant, only the stopping prob matters.)

c) The probability of a 4-character word is $(0.7)^2(0.3) \approx 0.15$. However, there are many different possible 4-character words, so $P(\texttt{aaab})$ is much lower. After the first character, $P(a) = 0.7/4$, so $P(\texttt{aaab}) = (0.7/4)^2(0.3) \approx .0092$.

d) Note that $P(\texttt{aaab}, L = 4) = P(\texttt{aaab})$, so

$$\begin{aligned} P(\texttt{aaab} \,|\, L = 4) &= \frac{P(\texttt{aaab})}{P(L = 4)} \\ &= \frac{(0.7/4)^2(0.3)}{(0.7)^2(0.3)} \\ &= 1/16 \end{aligned}$$

Exercise 4.
In lecture 3, there is an example of an FSA for English derivational morphology that looks like this and generates words like *wordy*, *wordification*, etc. (assuming that spelling changes are fixed up by another FSA that applies afterward).



a) The example only includes a single stem, *word*, on the first transition arc. List three other stems that could go there. What kinds of words *can't* go there?

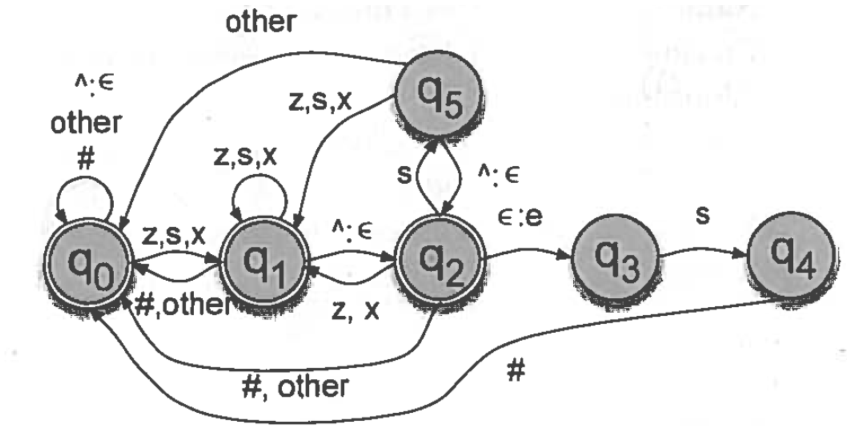
b) Consider the transitions labeled *er*, *ism*, *ist*. All of these end up in states labelled *N* (noun). What would happen if we removed the bottom two noun states and made these transitions end up in the same state where the *word* transition ends? Give some examples of words that are generated. Do these seem like possible words of English to you? (You might have different judgments than other people!)

Solution 4.
(material from Lec 3)

a) A few other stems that could go there: *Cameron*, *smell*, *fuzz*. Any word that isn't a noun cannot go there. I haven't been able to think of any noun stems that can't, but maybe someone in your tutorial can! As for nouns in general, it may depend on your judgment as to whether nouns with existing endings can go there or not (see also next question): for example my own judgement about *wordism+y* is a valid word is a bit fuzzy.

b) We'd get words like *wordismification*, *wordifierify*, *wordisty*, etc. I think these are legal, but pretty weird. Your mileage may vary.

Exercise 5.
The transducer from J&M Fig 3.17 is reproduced below. ('other' = none of {z,s,x,^,#,ε}).



a) What sequence of states would we go through to create the correct plural form for `axle^s#`? For `lass^s#`?

b) Can you think of any words that cause the transducer to go from state q_2 to q_5 and then continue on to an accepting (end) state? If not, can you at least say what properties would such a word need to have?

Solution 5.
(material from Lec 3)

a) `axle^s#` (showing also the consumed symbol to reach each state):
Start in q_0, then move to: q_0 (a), q_1 (x), q_0(l), q_0 (e), q_0 (^), q_1 (s), q_0 (#)
`lass^s#`:
Start in q_0, then move to: q_0 (l), q_0 (a), q_1(s), q_1 (s), q_2 (^), q_3 (ε), q_4 (s), q_0 (#)

b) The word would need to have a morpheme boundary preceded by one or more z's, x's, or s's, and also followed by an s and some additional characters. So the 's' after the morpheme boundary in this word couldn't be a plural 's' but the start of some other morpheme.

This made me think about compound words. The only one I've thought of so far is a bit contrived: `fox^sit#` (a verb meaning to watch someone's pet fox while they're out, similar to *babysit* or *dogsit*).

A student also pointed out that prefixed words will work too, and are less weird. For example *dissimilar*, *dissatisfy*.