Review of assessment and delivery of COMP34412

Assessment

Before I look in detail at the exam I would also note that the assessment of this course consists of two pieces of coursework, counting 20%, and an exam, counting 80%. I view the coursework at least as much being part of the teaching materials as an exercise in assessing how well the students have assimilated the material in the lectures. As such, it tends to have a high average, which is combined with the exam average to produce the overall mark for the course, because I want them to do it and learn from it. The coursework is do-able, and if you do it you get full marks for it, so inevitably it has a high average. In the analysis below of the marks I will look at the three components (coursework, exam, overall) because my intention when setting the coursework and the exam is to arrive at an overall mark with about the right average. More to the point, given that the coursework covers topics that I regard as important, there are always questions on the exam which are closely related to the coursework. Several people did neither of the coursework exercises, thus losing 20% of the marks for the course immediately but also depriving themselves of the preparation for the questions relating to those items. Furthermore, some 16 out of the 38 people who did do coursework 2 did not attempt the final part. Given that question 4 on the exam was directly related to the material that that part of the coursework was supposed to consolidate, these students were in a less than ideal position to deal with this question.

I did run two revision sessions. The first was poorly attended, the second was extremely poorly attended, and the students who did come to the second session had very few things they wanted me to expand on. They did, however, ask about the nature and structure of the exam. The revision sessions would have been a good time to come and ask about the material related to the coursework that people had difficulty with.

I have done some analysis of the marks, in particular looking at how the coursework and the exam interact. I cannot, of course, link individual students' coursework marks to their exam marks, since I do not know which exam scripts belong to individual students. I have therefore simply attached the 39 coursework marks randomly to the 39 marks of the students who got more than 1/20 on Q4 of the exam (I do not regard it as a coincidence that there were 39 students who submitted at least one of the pieces of coursework and 39 students who got more than 1/20 on Q4).

On that basis, the overall average for the 39 who did at least one piece of the coursework is 66.4%, and the overall average across all students, including the ones who did no coursework and hence automatically deprived themselves of 20 marks, was 59.7%. I do not regard these as being way out of line of the target of 65% for a third year course.

The average mark on Q4 f**or people who got more than 1/20 for this question** was 51.2, compared to 55.7 and 59.4 for Q2 and Q3. If it was felt that the lower average for this question was due to people running out of time then it wouldn't be unreasonable to scale it by (55.7+59.4)/2*51.2 = 1.12. I would use the average for people who got more than 1/20 because, as noted, a substantial number of people did not do the related coursework, and I do not see any reason why their failure to do the coursework, and their consequential difficulties with this question, should benefit everyone else.

There are detailed comments on various questions below.

Q1(a) I was expecting autocorrelation as the answer to the first bit. If they told me that you could use FFTs then I gave the mark, though in that case I was happier if they said something about **how** you would use an FFT. Some people just said "Use PSOLA" with no description, which didn't get them anything. People who said things about MFCCs also got very little credit unless they retrieved the situation with a later part of their answer. For changing the pitch, I needed to know that you used the information you'd gathered about pitch in order to work what size chunks to insert/delete. Just inserting/deleting "frames" got you nothing, because doing that does change the pitch.

Q1(b) I was made aware (see above) that I had given a rather vague specification of precision and recall in the notes, so I marked this one pretty liberally -- if you did it right you got full marks, if you did something that could be interpreted as arising from what I said you got full marks. In order to not get full marks you

either had to do something that could not be so interpreted, or you had to clearly misuse the definitions that you had given. Some people gave me micro-P, -R and -F scores, which is fine. This was a place where the issues that were pointed out with the notes were relevant, so I treated it by allowing any coherent answer that came from either inside or outside my notes.

Q1(c) Since I couched this in terms of transition and emission probabilities, the fact that in the notes I gave a strange (=wrong) version of how you arrive at emission probabilities was irrelevant. Most people who did it right did so without writing screeds of stuff, so I'm also not convinced it was too long for the marks. Some people interpreted the normalisation step as one of the four, which made their lives easier. I needed some evidence that they knew that you have to explore all the parallel branches, and that you can't just choose the local best and follow that, for full marks. A nice diagram also helped. A pretty large proportion of the cohort got it basically right -- lots of 5/5s for it. I am reasonably happy about it.

Q1(d) A lot of people got the second part partly right, and a smaller number got it properly right. There were quite a few answers that were based on the assumption that if a word didn't have spin (or charm, or strangeness) then it didn't have a spin (or ...) affix, but I think it's clear from the wording of the question and from the examples I did in the lectures that there should be one root, "bcd", and a range of affixes.

Nothing emerged during the marking of Q1 beyond the things I had been made aware of. My description of precision and recall clearly had led to some problems, my description of Viterbi did not, and I don't feel that I drove them to write too much for this one.

Q2: my mark scheme listed a subset of the components. I gave pretty well full marks for anyone who just gave the ones I listed, including the fact that emission probabilities link feature vectors of MFCCs to states. That last bit could be expressed in loads of ways, but I did want something about acoustic preprocessing for full marks.

A few people got confused by the fact that the HTK typically compiles the grammar into a file called wdnet, which they thought had something to do with the lexical resource WordNet. I didn't actually take marks off for this, but I did read the rest of what they said more carefully to see if there were other confusions.

I did only give full marks for 2(c) if they said something about using the grammar during training -- either explicitly that you don't or that it can be useful for generating prompts to be recorded. There are some startlingly incomplete answers to Q2. These are not generally at the end of the paper, so they don't seem to be being caused by the paper being too long. I am really quite surprised by this one, because nearly everyone did the coursework, and I don't see how you could do the coursework and then say some of the things they say, particularly about the role of the grammar; for sure anyone who did the coursework used the grammar to generate prompts during training. I see very little evidence that anyone was confused by my use of the term "concatenative unit selection", and lots of people who clearly connected that to the discussion in the notes and the coursework.

Q3: I didn't see much evidence that calling MALT transition-based has caused a lot of problems (maybe student 94994182, but they don't do anything with Edmond's algorithm in the practical part, and they don't given any description of graph-based approaches, so I think they probably just didn't revise that part of the course). A lot of people just described the data structures and operations for MALT. Given that the coursework involved writing a set of rules to act as an oracle, and I talked at some length about how you can learn rules, I did require some mention of how you make the decisions for full marks.

I'm not convinced that the amount of working in the MALT practical example is excessive for the marks available (and for some people they're the cheapest 5 marks on the paper). Several people gave meticulous, scrupulous descriptions of Edmond's algorithm but then failed to apply it -- just drew the graph with some loops in and didn't do anything with them. I can't really see a reason for that. Some people drew graphs with only one instance of "love", which wasn't great. The rules I gave actually produce two loops: if people found one and then eliminated it properly I gave full marks.

(4)(a) Far too many people described finding the common ancestors of **words** rather than synsets. So many that I went back to the notes, but I think that the slides about the 2-dimensional structure of WordNet are

pretty clear, and I said things like "WordNet relations are between synsets: do I have to do disambiguation before I can use it?" so I'm not very sympathetic on this one.

(4)(b) A lot of people quoted my algorithm verbatim, but didn't explain it and didn't apply it to the example. I gave half marks for this, but to get more you had to either say something or apply it (or both). It's a slightly tricky algorithm, because I use continuations to handle the fact that the matching algorithm may need to backtrack, so if they didn't apply it to the example I did at least want something that showed that they understood it rather than just having memorised it. The question did say

> "this algorithm should allow 'I watched a black bird' to entail "I saw a bird', assuming that the dependency trees for these two sentences are as in Figure 3",

which I think is a pretty clear invitation to apply whatever algorithm you describe for this question to these trees.