English past tense: An improved connectionist model and a dual route model

An overview by Eric Auer

15. November 2001

In this writeup, I am to compare a dual route model by *Pinker* to improved connectionist models, for example the one by *Plunkett and Marchman* (1993). This task is complicated by two problems: I am lacking some detailed information, and the P/M 1993 model is not exactly a very good and recent connectionist model. So there will be a special section on *Marcus'* criticism on the P/M 1993 model, and some other parts of the discussion will be sketchy (especially for the Pinker model). Still, I think this text will give you an overview of our new pair of competitors and their strenghts and weaknesses.

1 Introducing the new competitors

1.1 A multilayer network

The connectionist model described, or better, criticized, in this section is the one described by Plunkett and Marchman 1993: While the Rumelhart and McClelland model described in the previous writeup uses a very simple *Perceptron* style network to solve the riddles of English past tense by means of a pattern associator, the P/M model is doped with a *multilayer* network. But as we shall see, multilayer networks trained using the backpropagation rule provide no automatic solution for every problem that older/simpler connectionist models have.

While Perceptrons have been proved not to be able to do anything more than a linear separation of input space, multilayer networks can learn arbitrary functions to map their input to output patterns. This will involve the emergence of a private representation of the problem space in the so-called *hidden layers*, meaning all layers but the input and output ones. Thus, the private representation is not thaught – only the training set of input and output patterns is.

1.2 A dual route model

Pinker, on the other hand, suggests to combine the best of both worlds into a dual route model: A set of (symbolic) rules would handle the regular inflection in what seems to be a very straightforward way, while a (connectionist) pattern associator memory is to handle the irregulars. The idea is that the pattern associator makes a good combination of memorizing irregular forms while taking into account subregularities as well. A verb that is taken to

be irregular by the pattern associator will be handled as such, *blocking* the default route of regular inflection.

Pinker thus avoids the old separation of connectionist and rationalist models. In general, he makes a very good point in taking into account that the strength of symbolic models lies in regular inflection, while connectionist models are good in finding and processing more subtle redundancies in their task – such as the ones in irregular verb inflection. The problem, however, is that Pinker shows no clear way of training the model. I believe only the connectionist part is trained with the irregular verbs, and outside the training mode some kind of measure is deduced from the pattern associator output to decide which route to take. The measure would be somehow telling how strong/clear the pattern associator reaction on a certain verb is. This would, however, require all verbs to be tagged as regular or irregular by a teacher in order to make training work. The rule based part does not try to be psychologically plausible in any way, the rules are simply taken to be there.

2 The three hurdles revisited

In the last writeup, I was describing how well the competitors would handle *stem-past similarity* (the past form is usually similar to the stem), *change-change similarity* (the transformation of a stem into a past form is subregular) and *stem-stem similarity* (the groups of verbs undergoing the same kind of transformation have in general similar stems).

The rule based account had a nice explanation for the first two similarities: Only a few features need to be added or changed if stem and past are similar, and fewer rules are needed if the inflection is not different for every verb, but rather subregular. At least the model by Chomsky that I had mentioned did not, however, explain stem-stem similarity. The verbs were simply tagged to tell which rule would apply to which verb. Part of the tagging can in principle be done by done by rules, but the nature of the task is also very suited for handling by memory rather than by rules.

The pattern associator model on the other hand almost implies stem-stem similarity and it can also be argued that change-change similarity is natural for a pattern associator, but on the other hand, there is no reason whatsoever why the model should prefer stem-past similarity in what it does. In fact, a rule/regularity saying „the past tense is the stem read backwards" would be learnable just as easily as the normal English inflection, while being highly unnatural for human languages.

Whether the improved connectionist model can overcome the mentioned problem is not clear: It depends on the way the input and output are encoded and on the question of internal handling of the stem. Some other models (not the P/M 1993 one) did for example encourage using a (partial) copy of the stem in the process of output creation. Other models such as the ones based on the *Simple Recurrent Network* structure by *Elman* do actually encode the time structure of a problem as such, giving a nice account to all three kinds of similarity at least at first glance while strongly rejecting unnatural inflection such as the stem reversal mentioned above. The P/M 1993 model, however, did encode the time structure in a flat way: All of the stem is fed into the network at once, and all the output is collected at once, so the

stem-reversal problem is still there. As I am lacking some details of information about that model, I do give no further predictions on the performance of the P/M model here.

The dual route model still lacks some explanation on how the regular part and the regular/irregular distinction is learned, and as the parts are not described in great detail, it is unclear which of the problems of the purely rule-based and connectionist are still there in the combined model. But still I do think that the combined model will do at least as good as any of its parts: We can give an useful „implementation" of stem-stem and change-change similarity – the connectionist part – and of stem-past and change-change similarity – the rule based part. So the overall performance and plausibility will depend on how well both parts are integrated. The rule based part will not be very psychologically plausible from a connectionist point of view, but the rules can be shaped based on the assumption that the mind wants to keep things simple (and in some way logical or intuitive). The importance of plausible learning regimes is real, but on the other hand, the dual route model still shows a number of other interesting properties.

3 Problems with P/M 1993

Marcus 1995 dedicates an entire article on problems and trickery associated with the P/M 1993 model and the R/McC model discussed earlier. I will summarize some of his points here, along with some more information about the P/M 1993 model.

As Marcus points out, P/M were using some tricks in their graphs: While comparing their network to data describing how a child (Adam) handles the problem, they compare the vocabulary size of the network to the age of the child in a linear way and made other hard to explain decisions on how several graphs were plotted. Some of the differences showing up after fixing these issues: The network stops overregularizing completely after a while and does overregularize far less than the child (only one fifth, counted in types).

The P/M network needed high proportions of regular tokens in the input – about twice as high as in the „input" for the child – in order to be able to generalize the default inflection. While a child shows a period where no overregularizations are made (the idea is that they first do not use the past tense at all, then just memorize the forms, and only when they begin to find regularities, they start to overregularize and sometimes irregularize), the model did not show this effect. I do, however, account this to general problems with the way the problem is presented: Children are exposed to verb inflection long before they get the idea on when (and how) to use it, so one cannot tell whether they would produce proper inflection in the period where they simply do not use the past tense.

A more important problem is that P/M again used unrealistic manipulations in the training environment – alledgedly to induce something like the U-shaped learning curve in human inflection learning. They changed the speed of vocabulary growth (in terms of training epochs per new verb) suddenly, claiming to simulate the human vocabulary spurt in that way. Marcus argues that they simply deprived the network of enough time for learning at this point to force it into producing overregularization errors. He may very well be right with that. Also, children

usually start overregularizing a year after they start their vocabulary spurt (at the age of 16 months – overregularization starting at the age of about 29 months).

Another model of Plunkett and Marchman used a constant training set, taking into account type and token frequencies in 1991, but this is not discussed here. The training set thus contained several instances of each of the few irregular verbs (high token frequency and low type frequency) and many regular verbs (high type frequency but represented with a low token frequency for each in the training set). The other model started with training a certain training set for a while, then continuing by adding more verbs at regular intervals (with an 80 percent probability of using regular verbs).

As a last important point, the P/M model showed a different error pattern distribution than the child: The model overregularized no-change verbs more often, children do so for vowel-change verbs. This means a child is more likely to use *singed* for *sang*, thus failing to use the vowel-change irregularity, while the model was more likely to use *hitted* for *hit*. The numbers are 5.1 and 3.0 percent of vowel- and no-change overregularizations for Adam and 0.8 and 1.1 percent for the network. Also, while children produce more overregularization errors (like in *hit/hitted*), the network produces more irregularization errors (like in *flow/flew*) and hybrids (like in *sing/sanged*).