English past tense: A rule based account and a perceptron connectionist model

An overview by Eric Auer

9. November 2001

In this short overview, I summarize the comparison of two completely different attempts of modelling the generation of past tense verb forms which can be found in more detail in a book by *Steven Pinker*, called „Words and Rules", chapter 4. The chapters before that one give an introduction, squeeze the regular verbs into very few rules, and show many regularities even in the set of irregular verbs. The discussion is focused on the English language. In a continuation of this work, I shall compare a more sophisticated connectionist model to a version of a dual route model sketched by Pinker (see the end of this text).

1 Background: From Humes and Hobbes to our „case"

There is no doubt that both regular and irregular verb forms have been existing for ages and neither of them is going to disappear in near future. People still feel patterns among irregulars, so assuming the irregulars would be leftovers of long-dead rules only stored in some kind of list is not really plausible. But what *is* a plausible model of the mind anyway? This question has been on the minds of many researchers for centuries, and the answers can often be put in one of two big groups: Rationalist and empiricist points of view.

In 1651, *Hobbes* describes human reasoning as special form of *computing*, or symbol processing, as modern *rationalists* would probably say. *Leibniz*, influenced by Hobbes and Wilkins (who had created an artificial language which derives words from concepts by some very regular algorithm), even expresses the hope of every controversial discussion being solvable by following exact logical rules in his famous saying *„Calculemus"* (let us calculate). Other important members of the rationalist movement include *Descartes*, and, among the linguists, *Humboldt* and last but not least *Chomsky*. *Chomsky*, along with *Halle*, developed very sophisticated and compressed sets of rules (see the theory of *generative phonology*) to capture, among other things, both the regular and irregular inflection of the English past tense.

A second movement, *empiricism*, has *Hume* as an early member: He assumed in 1748 *resemblance and contiguity in time and place* (and at an earlier stage cause and effect) to be the basic principles of thought processing. *Locke* also points to contiguity in time in 1689, when he claims this to be the mecha-

nism allowing us to learn the arbitrary pairing of words and concepts. Later, *Pavlov, Watson and Skinner* used very similar ideas in their theory of behaviorism. The members of this movement important for this text are *Rumelhart and McClelland*, targetting linguistic problems with „connectionism" – implemented in *a network serving as a pattern associator* in our case. This associator has no explicit rules at all.

2 The three hurdles both models need to take

If there were only regular verbs, the problem of past tense formation from (present tense) lexicon entries would be very easy: Only a few simple rules or patterns for appending the „–ed" and adjusting to the correct pronounciation need to be captured for that. Note that we also assume the past to be derived from the stem.

However, there are irregular verbs. Maybe not frequent in type, but very frequent as tokens: The set few irregular verbs is very similar to the set the most frequently used words: For example the very irregular verbs *be, do, have and go* are even irregular across a big number of languages, while at the same time being the most frequently used words there, describing four really basic concepts.

Those words are at the same time the most important of very few verbs where the mapping of different forms is a quite arbitrary mapping of strings, best learned by memorizing most of the instances as separate entries in some list. The real challenge are all the „normal" irregular verbs, showing many small subregularities that should have a plausible explanation.

Those subregularities show three main properties:

stem-past similarity Even for irregulars, the past tense is usually very similar to the stem: The only exceptions are *be/was* and *go/went*, while in all other cases only small changes apply, like for example in *ring/rang*.

change-change similarity Apart from the obious regular case, where the rule is roughly „add –ed" and not much more, the changes involved for irregular past tense formation can also be found to be taken from a small set of common regularities such as „convert i to a" (in *ring/rang, sing/sang, ...*) in most cases.

stem-stem similarity The example above leads to a third important similarity: Verbs sharing a certain kind of treating (like i/a conversion) are often also similar in sound (the above example can be described as part of a „consonant/i/ng group").

Having set up the parcours, both combattants will have to show how well they can handle the hurdles (provide a plausible model for the subregularities) in the next sections.

3 The rule-based model

Assuming a tendency to use few and simple rules, stem-past similarity and change-change similarity can be easily explained. But this is only what shows up at a first glance: Chomsky *did* capture the regular inflection and most ir-

regular forms in only a few rules, but his rules are a result of long research by adult linguists. One of the most radical assumptions is that the stored stems are in a form similar to the written form, or the spoken form of ancient English before the so-called *„Great Vowel Shift"* (a big change in vowel pronounciation centuries ago). This allows the rules to be very simple and elegant, but it is very improbable that a child aquiring knowledge of past tense forms is capable to deduce the preshifted forms and/or the rules to convert them from and to their pronounciation today.

A second big problem of this rule-based model is the handling of stem-stem similarity: Chomsky suggested to *tag* all entries in the mental verb lexicon with markers, stating which rules are responsible for creating the past tense of every single verb. The reason is that the classes of verbs being handled in the same way are fuzzy and blurred, a simple rule of „change i to a for verbs consisting of consonant/i/ng" does not capture them appropriately. The generally clever move to consider sound features rather than verbose sets of phonemes contributed to the elegant set of rules menioned above, but it did not help with the stem-stem similarity problem either.

And last but not least, some of the rules – as well as the tag set – are very artificial and rarely used, so the question arises why they should not be replaced by raw memorizing of forms. The patterns describing membership to a stem-stem similar group of irregular verbs call out for some „probabilistic" handling, and this is what leads us to the next model, the pattern associator of Rumelhart and McClelland (using associative memory as the „probabilistic" device).

4 The connectionist pattern associator

Rumelhart and McClelland used a radically different approach to the problem of English past tense in 1986 by using no rules at all, but a pattern associator memory: Their model took an input vector consisting of 460 binary information bits about the sound of a stem and generated 460 bits of binary output on how the past tense would sound like according to an amazingly simple rule: Every input bit (node) was connected to every output node, and by taking known pairs of correct stem/past forms as input and desired output, the network of connections was adjusted to capture the relations between input and output.

The model could thus be trained to treat similarities in stem sounds in similar ways (as the encoding of input and output vectors is a phonological one): Handling of stem-stem similarity can be explained better than with the rule based account.

But the big difference in structure also has some problems: The network of connections works as a pattern associator memory, none of the connections „has an idea of what it does to the data". To say it in another way, there are simply *no rules* in the sense of a Chomsky et al. model at all. As there are no rules, there is also no notion of a *simple rule*: Any arbitrary mapping of input sound patterns to output sound patterns can be learned, stem-past similarity (and even change-change similarity) looks like a complete coincidence to the network. The model also seems to put needless effort in handling phonology effects, because there are effects that apply to English in general and not only to the past tense – thus,

their handling should be done in a separate module, and not reproduced in every „inflection engine" (of which the past tense network is only one) from scratch.

Pinker complains about the mapping being only readable in one way, but I see no reason why the contiguity in the stem/past pairs should not be useable to learn how to recognize the correct stem when presented a past tense form as well as it is useable to learn how to generate a past tense form from a stem by the same mechanism of training a network to act as pattern associator / associate memory.

There is another issue Pinker is quite pessimistic about: Sometimes there are verbs with identical sounds but different inflection, for example in the three-fold case of *ring* in the three meanings „ring a bell", „wrestle" and „put a ring on a birds leg". In those cases, the network must not only use the phonological form of the stem as input, but also some semantic information. Relying on semantics alone would be the wrong way, because there are no semantic family resemblances among irregular verbs like the ones we do have in the (phonology/spelling based) stem-stem similarity introduced above. I do, however, see no problem in using the sound of the stem as the main input while still adding semantic clues (or even the output of other units such as a „this action/verb was just created from a noun/thing lexicon entry or concept" flag that would block irregular inflection) to the available input information.

The last but very tricky to solve weakness of simple pattern associator networks (perceptrons) is the proper encoding of the temporal structure of the input: Simply encoding words as unordered bags of sounds or sound features is obviously wrong, and having an array of those with one column per position has the fatal flaw that we neither know how many columns we should prepare nor does the input vector for „string" resemble to the one for „ring" in any way (we cannot handle shifts).

So a third attempt was to use so-called *Wickelphones* (after the psychologist *Wickelgren*): Wickelphones are triplets of phonemes in sequence, and the set of Wickelphones contained in a string can describe that string because the Wickelphones overlap and only chain together in the right order. But alas, Wickelphones cannot handle repetitions of substrings (having a substring several times does not change the set of Wickelphones – an unordered list of them or a set augmented with a marking of frequency for the members would solve this at first glance, but then we will fall to ambiguities in the ordering!) and sometimes they even fail to capture certain stem-stem similarities.

5 Sketch of a new „dual route" model

As we could see, both models showed some very appealing strenghts, but also gaping holes in plausibility of certain aspects. Pinker argues that as no single model solves the problem given, a combination of both may do better. Along with *Prince*, he researched this sketched model a bit further and gives a short overview at the end of the text I am summarizing here.

Rationalists have improved the rules for stem-stem similarity, and empiricists/connectionists created fancier network structures and data encodings (a perceptron can only find linear separable patterns; multilayer networks

overcome this problem. Also, creative solutions for the encoding of time and structure were found). Some researchers on the rationalist account have suggested to have two kinds of rules, the classical ones that generalize freely, and some that capture similarity patterns – those rules are called lexical redundancy rules, by, among others, *Aronoff, Bresnan, Jackendoff, Lieber and Spencer*. While some tuning attempts for connectionist networks look like just fiddling around, the invention of multilayer networks (among other alternatives) was one of the most important events to overcome the pessimism induced by some research by *Minsky* showing how weak standard perceptron models are [1] .

So while both sides improve their models, there is no strict reason to keep them separated apart from the centuries-old debate of empiricists and rationalist. Thus, Pinker suggests the verb inflection to be handled by a small set of clear rules (for phonology and regular verbs) on one hand but also some pattern associator memory (which will be good in capturing the subregular patterns that were problematic for the strictly rule based model). The pattern associator would detect irregular verbs passing by and block the rules for regular inflection in that case. For the encoding, Pinker suggests a structure similar to the trees rationalist („Cartesian") linguists use to describe the sound of verbs and syllables (which are composed of an onset of consonants and a rime consisting of the nucleus vowel and coda consonant in this model). This would be an encoding that preserves stem-stem similarity while being less sensitive to shifts, but there are clearly problems left both in implementation and handling by a network and in the encoding of non-trivial multisyllabic words.

[1] You got me: I am not sure if it was Minsky and I forgot who has invented backpropagation learning for multilayer networks. But I know that the alternatives include Hebb networks and maps categorizing input by moving nodes in a way preserving neighbourhood relations...