To what extent are Pinker's criticisms of the original Rumelhart and McLelland encoding scheme still valid for neural network approaches to inflectional morphology in general? Has the problem of how to encode English verb forms as layers of input and output perceptrons been effectively solved?

Stefan Sabev

22nd of November 2010

The problem to build a model which will fully explain the formation of past tense verbs in English is still unsolved. Although we have two major models - “Words and rules” by Steven Pinker and SPE by Noah Chomsky. Both, undoubtedly, masterpieces fail to capture all the different aspects of irregular verbs and to describe all the different patterns and exceptions. The Pinker model just proposes that rules are applied if no irregular form is stored, but this completely ignores the patterns which one undoubtedly will find. On the other side Chomsky's model does not provide anything for the stem-stem similarity. There have been different attempts to build one unified theory that will be enough to show all the aspects of irregular verbs, but till present day, no such theory is present. That is why people have started using alternate approaches to past tense formation as creating neural networks for the purpose.

Even though used only for several decades neural networks are now an important part of modern science, because they can be used in various fields of contemporary science. Rumelhart and McLelland proposed a entirely new approach to forming past tense forms by using a neural network. Their approach to the problem also known as connection or parallel distributed processing was viewed as revolutionary. It was even perceived as a paradigm shift in the field of linguistics. Pinker clearly states in his book that is undeniable that our brain computes language by networks of interconnected brain cells (neurons), but that R&M's neural network doesn't have the ability to organize things in modules based on their similarities and instead uses Hume's law of contiguity that if one one objects appear with another you associate them and that if a third objects is similar to the first it should share the former's associations. This network is called pattern associative memory or perceptron. The model of Rumelhart and McLelland actually uses features that one can find in Chomsky-Halle's theory. The input takes the verb stem and computes the output directly from it. They encoded words in a different way by representing them as a set of different vocal features – stop-high-stop, back-nasal-], nasal-stop-] etc. After summing up all the different combinations of features Rumelhart and McLelland had 460 inputs. Every one of those inputs activates one or several perceptrons. Perceptron can be perceived as an artificial neuron with incoming and out-coming connections If exceeded a certain threshold it will fire to its outputs.The bigger the sum of all the incoming connection' weights the bigger the chance for the unit to fire afterwards. Each one of them is activated (turned on) when the matching verb stem is entered. This basically means that one word can activate multiple perceptrons as long as it matches their encoding scheme. On the output side we have the same number of perceptrons – 460. This means that there exists a set of words that have the same structure and would eventually turn on the same input and output units even though they do not have the same pattern in forming past tense. The network is shown a pair of words (input-expected output) the weight of the incoming connections was either decreased or increased by a iota (a very small fraction) and consequently the same procedure was applied to the threshold. The way the neural networks is trained is that it is fed with a set of 420 words each of them shown 200 times. The network predicted most of the past tense verbs when tested with the familiar set of words - “look” - “looked”, “melt” and “melted”, “hit” and “hit” and it even guessed the correct form of “go” - “went”. When presented a new set of unknown 86 verbs it was correct about 75% of the past tenses of regular verbs and made predictable overgeneralisation errors as “digged” and “catched” for all of the new irregulars which is actually really close to the U-shaped curve which can be observed in small children. The neural network produced correct results for families of similar-sounding old irregular verbs like “cling”-”clung”, “sip”-”sept”, “slip”-”slept”, “bid”-”bid”, but it produced blends such as “gaved” and “stepted” which are common errors found in children. It was less likely to assign the suffix -ed to words that belong to bigger irregular families like “feel”, but more likely to verbs from smaller families like “blow”. Given the complexity of our brain the fact that a two-layer network with no hidden layers in-between can simulate children's behaviour in the process of gradually acquiring the correct forms of past tense verbs indicates encouraging progress has been made in the field of cognitive science.

Disadvantage of the network is that it can only be used to produce past tense forms and cannot be modified so that given the latter form it returns the infinitive form of a verb. The neural network is incapable of completing a task like this while a human instantly associates the verb with its past tense form and vice versa. When we say “went” we perfectly know that it actually means go in the past.

Second, this connectionism model computes every single detail of the past tense form. However many of the different details – as the choice between -t, -d and -id – can be found in fifteen distinct parts of our language system and most likely they are computed by a single monolithic phonology rule and just change according to morphology and syntax rather than having fifteen different networks for every one of those areas.

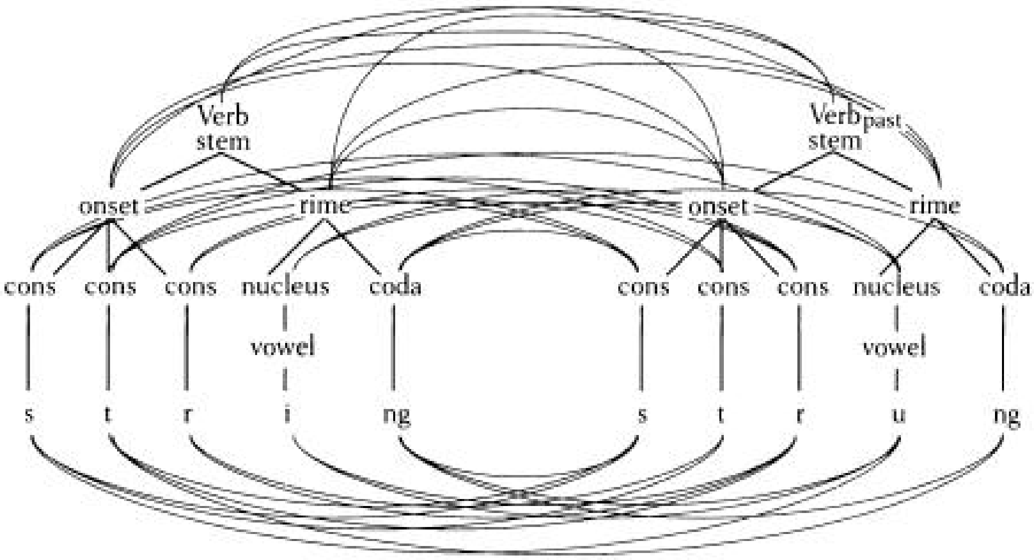
Another thing is that the connection cannot distinguish two words that have the same sound, but different meaning as mete-meted, meet-met and break-broke, brake-braked. This can be solved by adding additional modules each of which will be responsible for the meaning of words. However, words with similar meaning do not form past tense forms according to the same rules. Typical example of this would be hit-hit and strike-stroke. Both irregulars, but belong to different verb families.

Rumelhart and McLelland used a new way of encoding words into Wickelphones. A Wickelphone is a collection of three phonemes – for instance the verb draught consists of the following: [dr, dra, rau,aug,ugh,ght and ht]. Rumelhart and McLelland actually wanted to represent every word by its phonetical features according to the SPE model, so rather than a collection of phonemes it's a collection of Wickelfeatures. The set of all possible Wickelphones is about 64 000. This is a creative alternative but it brings a whole lot of problems connected with this encoding. It may seem really strange that at a first glance relatively easy task as registering a word in a database of units, but it is – even Wickelphones, which are a very creative and genuinely good approach, cannot solve the problem efficiently.

As in every language, English has lots of similarities between words – drink-shrink-think, glow-blow, ring-bring and many other examples spread across the language. Many of those tuples of words share Wickelphones, which strengthens the connection between them. As for many this means that they form past tense in a similar way (drink-drank, shrink-shrank), but for others this is wrong – drink-drank, but think-thought, ring-rang, but bring-brought. So when the network in the process of supervised learning receives several times those Wickelphones it increases the weight of the incoming connections and reduces the threshold of the unit, which matches the output. This procedure makes it impossible for the exceptions from the rule to be inflected correctly, because it is automatically redirected to the incorrect output. For instance it is very likely that the network will overgeneralise and produce “thinked” instead of “thought”. In other cases words are really alike, but share no Wickelphones at all – as given in the book examples “silt” and “slit”. Our brain perceives them as similar, but the encodings finds no connections between them. Another problem is that if a word has a Wickelphone that repeats itself several times there is no way to represent this and influence the results as the input can be either on or off. For unusual sounding words like pump or jump the network could not produce any past tense form, while for others it did completely bizarre for us things – turning “squat” into “squakt” or “mail” into “membled”. For us it seems very natural just to add the suffix -ed to those words, but since there was nothing like those words in the training set, the model just does not know how to react and this turns into weird for us past tense forms.

Cottrell and Plunkett (1994) suggest a different model. They encode all the words as cluster of Consonant-Vowel-Consonant (CVC), Consonant-Consonant-Vowel (CCV) or Vowel-Consonant-Consonant (VCC). There are the three valid combinations which are enough to correctly represent all the English words. This was more successful than the Rumelhart and McLelland model as it produced the appropriate forms for similar inputs which are not dependent on the output similarity. But even that model cannot effectively solve the problem with the encoding of the word.

At the end of chapter 4 Pinker suggests a minor change to the "Words and rules" model that will actually improve it. In that new modification words are linked to words, but words are linked to bits of words. However, these bits are not Wickelphones, but rather stems, onsets, rimes, vowels and consonants and features. It would look something like this:



This pictures is taken from the book "Words and rules" by Steven Pinker for illustrative purposes.

This way of encoding will mean that will with similar sounding (string, sling, stick, stink) will overlap some of the nodes. As a result "the irregular verbs show the kinds of associative effects found in connectionist pattern association." This modified version of the "Words and rules" model makes a strong prediction that regular and irregular inflection are psychologically and more importantly, neurologically distinguishable. However, this raises the question of how can they be distinguishable if people can generalise. The answer to that question is that irregular inflection depends on memorised words, while regular inflection can be applied to any word and it doesn't matter if the word is readily retrievable from the memory. It can do such, because it does not depend on memory access, but it is actually applying a symbol-processing rule or operation which applies to any verb. If this modified theory is correct it would be able to give an answer to the very old debate between associationism and rationalism – both theories are right, but they are right about different parts of the mind.

Even thought there have been major research interest in the area the problem of efficiently encoding words into some representation is still unsolved. Wickelphones so far are maybe one of the best way to manage to get around the problem, however they are a lots and they have disadvantages no one can deny. A word with multiple Wickelphones will make no difference as the unit responsible for it can be either on or off. The strengthening of the connections between certain verb families sharing common Wickelphones is another major problem, because it makes it virtually impossible for the network to output the correct form of the exception. The modified words-and-rules theory proposed by Pinker seems to be able to solve most of the present problems, however it is not proven completely correct yet.

References:

1. “Words and rules”, Steven Pinker, 1999
2. “Learning the past tense forms in a recurrent network: Acquiring the mapping from meaning to sounds”, Garrison Cottrell and Kim Plunkett, 1994