Mind-II

Understanding the acquisition of verb morphology through connectionist simulations

Dustin Mitchell

March 26, 2001

1 Issue and importance

An overarching theme of human uniqueness research involves the type of uniqueness. Some researchers believe that humans are endowed with specific mechanisms which are completely absent in the child. In the area of language, Noam Chomsky is the most prominent proponent of this view. He holds that humans possess a unique "language acquisition device" [Chomsky, 1986, p. 3] which allows humans to learn language, but which is absent in other life forms. The opposing (majority) view is that humans possess no special mechanisms, but that humans have more mental power — a better-developed brain — than other animals. The species' mental abilities are above some threshold value at which sophisticated language becomes possible, while other species' abilities are below. This view gains great support from research showing that general mechanisms such as connectionist networks can, given sufficient power, learn human language.

Language, however, is very complicated. Thus, in exploring the acquisition of language in children, researchers must limit their focus considerably. Researchers do not try to teach a computer an entire language. Instead, they focus on specific parts of the process. Elman [Elman, 1993] researched the capacity for connectionist systems to decode the meaning of sentences with sophisticated structure. Some research has been done about simulations learning to link arbitrary words with specific meanings (see [Regier, 1996]), but the field of semantics requires reference to means of perception that computers do not share with animals, so this research has made little progress. Research into the phonetic structure of words, or morphology, has shown great promise.

English, classified as an uninflected language, does inflect its verbs to indicate tense, number, voice, etc. This inflection is highly irregular, and children learning it display an interesting phenomenon called U-shaped learning: in early learning they are very accurate in their inflections. As the learn, however, they make more mistakes on the same words. These mistakes are overgeneralizations — applications of morphological rules to exceptional verbs. As the child continues to learn, these errors are corrected. The research in this paper attempts to show that a general learning device — without any specific language mechanism — can show this behavior.

2 Theories and hypotheses

Noam Chomsky did not address past-tense learning specifically, but his more general theories apply without modification to the issue. He is a proponent of the concept of a universal grammar, which "may be regarded as a characterization of the genetically determined language facility." [Chomsky, 1986, p. 3] He hypothesizes the existence of a universal grammar (UG) because of what he calls Plato's problem: the language to which a young child is exposed in its lifetime is not sufficient to explain the child's abilities. As an example, he cites

> the structure-dependence of rules, the fact that without instruction or direct evidence, children unerringly use computationally complex structure-dependent rules rather than computationally simple rules that involve only the predicate "leftmost" in a linear sequence of words.[Chomsky, 1986, pp. 7-8]

Here he is speaking of syntax, but his comments could just as well apply to word morphology: at a surprisingly young age, children learn to apply certain phonetic rules to the generation of the past tense of a verb. For instance, a child learns quickly that many past-tense verbs are formed by suffixing a vowel and a dental to the present tense: "want" → "wanted", "talk" → "talked", etc. To ensure that the children have not just memorized these forms, researchers test them with novel words. Given "wob", a child will often generate "wobbed". Chomsky posits that the information with which a child is presented is not enough to give the child knowledge of these rules. This leads him to the conclusion that humans have an innate predisposition toward a set of rules. This set of rules is called the universal grammar.

Chomsky uses introspection and examples to show there is not enough evidence to establish rules. He offers no other evidence for the validity of his conclusions. However, other researchers have devised their own theories which have the support of empirical experiments.

Two such researchers in the area of past-tense learning in connectionist systems are Rumelhart and McClelland [Rumelhart and McClelland, 1986, henceforth RM]. Where Chomsky believes that rules are, at least to some extent, explicitly represented in the mind, RM argue that both the phonetic rules and the exceptions can be encoded implicitly within the general structure of the mind. Further, they argue that this encoding can develop given the information available to a child.

To test their theory, RM constructed a connectionist system capable of taking a representation of an English present-tense verb as input and presenting the past tense of that verb as output. Their system was not designed with any predispositions to language — it was just a set of linked network nodes. Their theory predicts that if this system is sufficiently rich then it has the capacity to implicitly encode the rules of English tense-changes, including exceptions. Additionally, the model should display the same U-shaped learning pattern observed in children, showing it is relevant to child development. To be sure the system is applying rules and not memorizing forms, the researchers must observe its performance on novel verbs. If it applies regular patterns to a verbs it has never seen, then it must have some representation of a rule.

Unfortunately, RM's work has several technical shortcomings, which will be described in the next section. In an effort to overcome these shortcomings and reinforce RM's arguments, Plunkett and Marchman [Plunkett and Marchman, 1991, Plunkett and Marchman, 1993] repeated RM's experiment, incorporating advances in technology and connectionist theory. Their intention was to show even more conclusively that a connectionist model could display the same U-shaped learning pattern observed in children.

Marcus [Marcus, 1995] argues with the validity of these researchers' conclusions, but presents no theories of his own. His arguments are similar in kind to the arguments used by John Searle [Searle, 1980] against the Turing test: it is difficult to judge what is going on inside of a system by merely observing its behavior. He also points out an alternate explanation for some of the results claimed by Plunkett and Marchman.

3 Methods and evidence

Chomsky's views about language acquisition are based principally on introspective thought experiments. In the broadest sense, he is wrestling with the mechanisms by which humans bridge the gap between perception and understanding. In [Chomsky, 1986] he addresses syntactic issues, arguing that the sophisticated and subtle structures that all speakers of English understand cannot be learned by known techniques: "there is little hope in accounting for our knowledge in terms of such ideas as analogy, induction, association, reliable procedures, good reasons, and justification in any generally useful sense, or in terms of 'generalized learning mechanisms' (if such exist)." [Chomsky, 1986, p. 12] Before this quotation, he offers convincing arguments that the first six techniques will not suffice individually, but he never addresses "generalized learning mechanisms" directly. This is precisely the mechanism the other researchers claim is responsible for human language acquisition.

At about the time Chomsky was writing [Chomsky, 1986], Rumelhart and McClelland were experimenting with a connectionist model for human acquisition of verb morphology rules. In designing the model, they were faced with several difficult decisions. First, how can a word of variable length be presented to a network in a way that will allow it to draw the proper conclusions? The English map from spelling to pronunciation is far too complicated to expect the network to handle that as well as the past-tense forms. A simple phonetic array (i.e. several independent bits of input for each phoneme in the word) seems effective, but is not compatible with variations in word length. The researchers' final decision was to use

Wickelfeatures, which encode phonetic changes; that is, one particular Wickelfeature might encode a change from a voiced vowel to a dental consonant. The researchers selected 460 such Wickelfeatures which appear commonly in English words.

The researchers designed their model as a pattern association map; that is, two arrays of 460 nodes with each node in one array connected directly to every node in the other array. RM attached a predesigned layer to convert from a phonological representation of a word to a Wickelfeature representation to the input nodes of this map, and a predesigned *binding network* to the outputs to convert back to a phonological representation. For each present-tense verb presented to the model, the system displays a word on its output. This output is compared to the correct past-tense, and the network is adjusted with a simple algorithm to weaken incorrect connections and strengthen correct connections between nodes. This process is iterated many times, and gradually the system learns to produce the proper forms.

Given this basic model, there are still many variables under the control of the experimenters. The word choice is a very important one. RM selected English words from an established dictionary of words in common usage. They used word-frequency data from this dictionary to classify their selected words into low-, medium- and high-frequency verbs. They began the simulation by presenting the model with only the high-frequency verbs. After several iterations through these verbs, the medium-frequency verbs were introduced. Finally, the low-frequency verbs were introduced to the model. The researchers made decisions about the time between introduction of each new set of verbs, in hopes of emulating the exposure a child receives to verbs and their past tenses.

Rumelhart and McClellands' results appear quite promising. They display U-shaped learning, especially in the irregular verbs: during the first (high-frequency) phase, accuracy on irregulars and regulars increase at the same rate. At the onset of the second phase, however, accuracy on the irregulars drops sharply, while accuracy on the regulars continues to grow. The accuracy on irregulars soon begins to increase again, but does not reach the accuracy of the regulars within the time of the trial.

The researchers also examined the types of errors made by the model. In particular, they addressed overregularization errors: situations where the model produces an incorrect past tense in accordance with a rule, such as “go” $\rightarrow$ “goed”. Their research showed that regularization errors account for almost all of the errors occurring at the beginning of the second phase.

Measurements were also take of the type of sub-regularities in the irregular verbs, but these findings do not bear significantly on the topic of this paper. Response of the trained network to novel verbs, however, is very important. As described in the last section, if the network can correctly process verbs it has never seen before, then it must somehow represent the rules. Indeed, the RM model has a 91% accuracy rate on the low-frequency set, which is used as a novel-verb set.

Plunkett and Marchman [Plunkett and Marchman, 1991, Plunkett and Marchman, 1993] (henceforth referred to as PM) point out several serious problems with the RM model. First, the model never achieves the level of accuracy found in a child. This argues that the model is not adequately representing child development, but does not affect the credibility of RM's conclusions, which were based only on the differences in performance on different sets of words. Part of the reason for this inadequacy may be the single-layer nature of the RM model. At the time of RM's research, available computer power could not efficiently simulate a multi-layer network.

PM also point out that the Wickelfeature representation of verbs, in particular the binding network on the output end, perform a great deal of the work in this situation. “Wickelfeature representations presuppose a theory of the phonological regularities present in the English past tense system.” [Plunkett and Marchman, 1991, p. 46]

Most significantly, however, PM object to the input set. First, after the first 10 epochs, the vocabulary size is increased suddenly from 10 to 420 words. The proportion of irregular verbs in the first set is 80%, but after the vocabulary expansion the proportion drops to 20%. The sudden drop in accuracy on irregular verbs can be attributed directly to this change; during the first 10 trials irregulars are the majority of the words seen, and the network will learn them well. After the change, however, irregulars are in the minority and thus will not be learned as quickly as regulars.

PM also criticize the RM model for its extreme granularity of frequency. Within the high, medium and low frequency classes, all verbs have the same frequency. Clearly a child is exposed to verbs with a continuum of frequencies, and these frequencies change over the course of the child's development.

PM rectified many of these problems in their own model, which is described in detail in [Plunkett and Marchman, 1993]. The PM model has three layers: an input layer, a layer of hidden units, and an output layer. The researchers apply a more advanced learning algorithm called *back propagation* to allow this multi-layered network to learn. To counter the problems with word representation and length, PM constructed a dictionary of uniform-length strings of phonemes which do not correspond to English verbs. These "verbs" ere then matched with past-tense forms, with the pairs divided into four classes: arbitrary mapping, identity mapping, vowel change, and regular mappings. Arbitrarily mapped verbs have no relation to their past tenses; identity mapped verbs have identical past and present tense forms; vowel change verbs differ from their past tense by substitution of an internal vowel; regular verbs form their past tense by suffixation.

Verb types were chosen to mimic the vocabulary of a child who has mastered the past tense. Two out of 500 verbs are arbitrary (.4%), 91.6% are regular, 4% are vowel-change, and 4% are identity. To begin the simulation, however, a different set were selected: 50% regular, 20% identity, 20% vowel change, and 10% arbitrary. This choice was made based on research on childrens' language use.

To address the frequency granularity observed in the RM model, PM use a technique called *epoch expansion*. Under this scheme, new verbs are introduced, one at a time, at regular time intervals. In particular, PM introduced a new verb every 5 epochs for the first 100 verbs and every epoch for the remaining 400 epochs. They use a weighted selection system to continuously vary the frequency from the initial to the final configuration.

The researchers analyzed the errors produced by the model, as well as its performance on a set of novel verbs. Like RM, they classified errors by type: suffixation (where an incorrect ending is placed on the verb), identity (the verb is not changed), blend (vowel change and suffixation), or unknown. As expected, unknown errors become less common as the training continues, but interestingly the network still produces unclassifiable responses to regular verbs at the end of the trial. For each type of verb stem, particular types of errors were prevalent, which the researchers take to mean that the model is very sensitive to the verb phonology.

The researchers next separated the individual verbs which undergo U-shaped development; that is, they are mastered by the network, then incorrectly mapped, then mastered again. They found that 15% of identity verbs and 30% of vowel change verbs showed this characteristic.

As discussed in the previous section, network performance on novel verb stems is an important factor in understanding the mechanisms developed inside of the model. The PM model displays a strong tendency to regularize verbs with no distinguishing phonological markings, and also a weak tendency to perform the correct vowel-change on a vowel-change verb. A more interesting look at novel-verb performance is in terms of vocabulary size. A direct correlation was found between final vocabulary size and rate of regularization of indeterminate verbs.

Marcus [Marcus, 1995] responded harshly to the evidence presented by PM. He points out several places in the latter paper where the graphs and statistics are grossly misleading. More significantly, however, he points out that the onset of PM's U-shaped learning coincides precisely with the acceleration of the training regime from one verb every 5 epochs to a verb every epoch. Although PM claim that the discontinuity is motivated by a discontinuity observed in children, Marcus points out that these events are by no means simultaneous in children.

Although both the RM model and the PM model have significant problems, especially in representing U-shaped learning, they do display remarkable performance on novel verbs.

4 Conclusions and extensions

Research using connectionist networks has, as yet, been unable to model past tense morphology acquisition in children with complete accuracy. This should not be expected of the research, however. The acquisition of verb morphology is only a small component in the complicated process of language acquisition, and its interactions with the other components are very significant. For instance, verbs have many other forms which must be also generated by the child and reconciled with the past tense form. Children must also listen carefully to discover the link between the present and past tense of a single verb, while the model has both presented to it without any other stimulus. Children are also exposed to some direct teaching, including negative feedback ("no, Adam, 'went', not 'goed"').

The significant achievement of the connectionist research is to show that a general learning device can extract and implicitly encode morphological rules from a surprisingly low amount of input data. This conclusion comes from the models' performance on novel verbs, which was never brought into question. When presented with novel verbs, both models responded in a rule-governed manner, demonstrating that they had some representation of the rules. The U-shaped learning behavior reported by PM and RM, though not in correspondence with U-shaped learning in children, lends support to this point: when the vocabulary to which the model is exposed reaches a certain size and structure, the model changes over from rote memorization to rule construction.

This latter conclusion is only one possible interpretation, however. Here, again we are faced with the issue of determining the mechanism generating a behavior only by observing the behavior. It is possible that the models did not adjust the weights strongly enough in the initial trials, and that when more vocabulary was imposed the earlier learning became insignificant, effectively forcing the network to start its learning over. Indeed, this would explain the peculiar shape of both models' U-shape: initial learning followed by an almost instantaneous drop in accuracy, which is soon followed by a gradual increase. This instantaneous drop may represent the network effectively resetting itself to begin learning again.

Given modern increases in computing power, it would be interesting to see another model created which addresses the problems inherent in both the RM and PM models. Perhaps researchers could sample the vocabularies actually presented to children at many points in development and build a vocabulary schedule from this data. This data could then be presented to a model which is capable of handling English verbs in their full phonological forms. The introduction of some noise into the data would also be very appropriate: occasional "mistakes" where the wrong past tense form is presented for a verb; occasional misspellings of verbs; and occasional malformation of past-tenses. The construction of this vocabulary and model would be very time-consuming, and may still not represent U-shaped learning in children as well as we would like.

Expert systems are decision-making systems which are constructed with explicit knowledge of a problem. It would be interesting to see research on the use of expert systems to simulate past-tense construction. There is no learning process involved, but given certain limits on the size of the system it would certainly make errors, and these errors could be analyzed in the same way they were for the connectionist systems.

Research in the area of verb morphology acquisition is an exploration of the techniques that should be applied to acquisition of other aspects of language, and to language itself. Connectionist general learning systems have not yet been shown incapable of replicating the astonishing process of language acquisition in children, but there is still a great deal of research to be done.

References

[Chomsky, 1986] Chomsky, N. (1986). *Knowledge of Language.* Praeger, Westport, CT.

[Elman, 1993] Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

[Marcus, 1995] Marcus, G. F. (1995). The acquisition of the english past tense in children and multilayered connectionist networks. *Cognition*, 56:271–279.

[Plunkett and Marchman, 1993] Plunkett, K. and Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, 48:21–69.

[Plunkett and Marchman, 1991] Plunkett, K. and Marchman, V. A. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38:43–102.

[Regier, 1996] Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* MIT Press, Cambridge, MA.

[Rumelhart and McClelland, 1986] Rumelhart, D. and McClelland, J. (1986). On learning the past tense of English verbs. In J.L. McClelland, D.E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of congition. Vol. 2: Psychological and biological models.* (pp. 216-271). MIT Press, Cambridge, MA.

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. In M. A. Boden (Ed.), *The Philosophy of Artificial Intelligence.* pp. 67-88. Oxford University Press, New York.

5 Abstracts