A Connectionist Model of Categorization and Grounded Word Learning

Georg Dorffner, Michael Hentze, Georg Thurner
University of Vienna,
and Austrian Research Institute for Artificial Intelligence[1]

This paper reports about ongoing research on a connectionist model of the learning of single words and their meaning, grounded in perception. Similar to (Plunkett, Sinha, Moller, & Strandsby, 1993) it consists of a minimum of two sensory-based components with an adaptable link between them. Both components perform categorization of sensory stimuli plus current internal states using a version of "soft" competitive learning employing mechanisms known from adaptive resonance theory (Grossberg, 1987). Through repeated learning categorization gradually leads to distinct attractors (compressed activation states). The temporal co-occurence of such categories in both system components can elicit the building of strongly weighted connections between them via a specialized component designed to approximate the arbitrariness and discreteness of the function of words (their properties as symbols in the semiotic sense), to be robust against noise, and to reflect important psycholinguistic phenomena.

Introduction

One of the fundamental questions in language acquisition is: How come words to mean something, and how is a consistent mapping between meaning and sound achieved? This paper introduces a connectionist model trying to address this question. After a brief look at previous models for simple word acquisition we introduce a novel model which distinguishes itself from the other ones in many important respects. The specific aspects of word learning we want to account for, and which previously have either been neglected or not included in a satisfactory way, are the following:

- Words in principle are of a specific nature which is called being "arbitrary" in semiotics, in the sense that their form does not by itself reflect any aspect of their meaning (compare (Plunkett et al., 1993), who speak about "the emergence of symbols" in the context of word learning). Through this nature, the learning of words runs counter the otherwise more "natural" way of learning based on similarities among stimuli.

- The mapping between words and their meaning is learnt under extremely noisy conditions. In other words, this mapping does not simply consist in finding out which word stimulus (e.g. sound) is paired with which meaning (e.g. visual category), but must be a learning algorithm robust against extreme distortions of

[1]The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science, Research, and the Arts.

these pairings (for instance, against the fact that, say, in 60 % of the cases a visual category is *not* paired with the actually correct word).

- Word learning is strongly connected to the organization of conceptual knowledge and thus reflects many of the phenomena known about human categorization (concept formation), such as prototype and basic level phenomena (Rosch, 1978), the influence of language on the learning of taxonomic categories (Nelson, 1988), etc.

To model such aspects, several simplifying assumptions must be made, by which many otherwise important aspects of language are left out:

- Words and their recognition does not involve any sophisticated phonology or morphology but are simple physical entitites, being input to the system as stimulus pattern without sequential structure.

- Aspects of meaning that go beyond simple words and their references (such as the relationship between syntax and semantics) are neglected.

- The basic way to acquire meaning is to categorize sensory stimuli (e.g. visual patterns) and to map them onto a category of stimuli recognized as words. In this context, only content words, and among them only (or mainly) nouns are considered for the time being. Also, categories which are not entirely induced by sensory patterns (more "abstract" concepts, so-to-speak) are not considered for the moment. However, it will be hinted upon how both important aspects can be accounted for in extensions of our basic model.

Nevertheless, in spite of its apparent simplicity, the models discussed here permit the inclusion of many non-trivial aspects of word acquisition, such as the ones listed above. It should be noted here that the purpose of modeling (in this work, connectionist modeling) cannot (and should not) be to focus on all aspects of language at once, but to pinpoint a selected set of phenomena in need for an explanation or theory. Only if models (epecially the ones which can be implemented on a computer) are to be turned into systems that actually perform language understanding (such as has been the goal of Natural Language Understanding as part of Artificial Intelligence) they must necessarily encompass a broad variety of language aspects. Connectionist models, in principle, can be seen as the basis of novel systems performing language understanding. For the purpose of this work, however, this perspective is not focused upon. Instead, the purpose of modeling to provide a basis (a "vocabulary", so to speak) for forming theories (explanations) is at the center of research.

Previous models

Several connectionist models for aspects of word learning and language acquisition have been proposed in the literature. (Cottrell, Bartell, & Haupt, 1990), for instance, propose a model consisting of two separate multilayer perceptrons, each auto-associating a visual (image) and a locally encoded name pattern with each other, respectively, so as to form internal distributed representations of those patterns in the hidden layers. These hidden

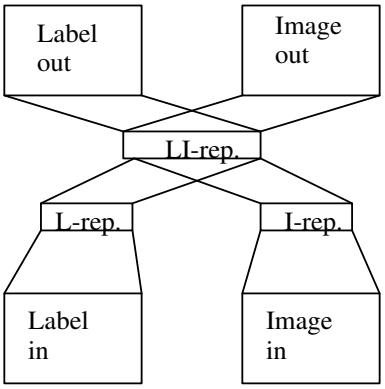

Figure 1: An outline of the model by Chauvin and Plunkett et al. consisting of a multilayer perceptrons with two separate and one common hidden layer

layers are then viewed as input and output to a third multilayer perceptron, learning to associate them with each other. After learning, a word pattern can be input and the corresponding image activated, and vice versa. Although no extensive psycholinguistic analysis is made, it is hinted that this kind of model can be seen as a first step toward a connectionist theory of meaning ("grounding meaning to perception").

(Chauvin, 1989) and, in an extension, (Plunkett et al., 1993) also use two separate input in an autoassociative mode, but unite the two components into one multilayer perceptron, trained with backpropagation (Rumelhart, Hinton, & Williams, 1986). They do this by introducing another common hidden layer, via which the autoassociations of both inputs onto themselves are learned (fig. 1). By this they suggest that learning of both patterns not only achieves a cross-association between them (permitting, similarly to (Cottrell et al., 1990), the activation of one pattern given the other), but also influence each other in an intricate, distributed way. 32 classes of binary visual patterns, with 5 random distortions each, were used as input on the visual side (preprocessed by projecting them onto a simple "retina"). Like in (Cottrell et al., 1990) orthogonal patterns (local representations) are used as label input. Their model exhibits many interesting effects, such as a vocabulary spurt (the rapid increase of learnt association between labels and visual patterns after a relativly slow progress), the temporal precedence of comprehension over production, and a prototype effect in the sense that several instances of a pattern category can lead to the generalization onto their prototype, although it had never been presented during training.

Another interesting model is that by (Schyns, 1991), who uses selforganizing feature maps to perform explicit categorization of sensory patterns. He suggests a separation of categorization and subsequent naming, reflected in separate components to achieve the two mechanisms.

Our model

The model we want to propose here is based on several previous similar models (Dorffner, 1992b, 1992a), extended by several aspects to eliminate weaknesses of earlier versions.

It is in the same spirit as most of the models discussed in the previous section, in that it consists of two separate modules learning on visual and ("acoustic") word patterns, respectively, which are then mapped onto each other. Two main components can be distinguished: Categorization and a mapping between categories.

A categorization model The model we propose rests upon the assumption that a major aspect of word learning is concept formation based on stimuli (and internal activation patterns), and that the basic aspect of concept formation is a process of categorization. Like the model by (Schyns, 1991), but unlike the models by (Chauvin, 1989; Plunkett et al., 1993), we assume a component that models categorization explicitly in a variation of competitive learning (compare (Grossberg, 1982; Rumelhart & Zipser, 1985)).

The main process of categorization in terms of connectionist activation patterns is the orthogonalization of responses. In other words, starting from intially disrtibuted responses to input stimuli, categorization means to achive patterns that are approximately orthogonal to each other when viewed as vectors. Forming a category, and thus the foundation of what we call a concept in the model, is thus a process of abstracting from similarities in the stimulus, making responses as dissimilar as possible.

As opposed to other competitive learning models, categorization here is seen as a process that is developed gradually during learning. In other words, instead of applying a strict winner-take-all (WTA) process (which sets the most active unit to maximum activation, and all others to zero), competition is achieved through inhibitive connections and a "rich-get-richer" effect. Through learning the ability of the winner to suppress the other units' activations is enhanced. This is a mechanism very similar to the ones implemented in the ART 3 model (Carpenter & Grossberg, 1990). Compared to ART3, in our model the mechanisms are greatly simplified, mainly due to the reason that we do not pay much attention to biological plausibility. The update rule used is known as *shunting equation* (Grossberg, 1982) or *interactive activation* (McClelland & Rumelhart, 1981). Clear categories are developed only after repeated presentation of stimuli, having a major consequence for word learning. It is assumed that the mapping of words to their meaning is learnt only after the category that constitues the meaning is distinguished sufficiently from other categories. A major side-effect of this approach is the easy accounting for protoype effects during word learning (Rosch, 1978). The degree to which a category is distinguishable from others (reflected through the introduction of a "goodness-of-fit" value defined below) is built into the learning rule adapting weights between the categories and their respective names (the words they are to be associated with). By doing this, patterns closer to the class prototype – which itself evolves during learning based on presentations of several instances of the category – will be mapped to words much more quickly than patterns farther away from the prototype (this is a different kind of prototype effect than in (Plunkett et al., 1993), reported by, among others, (Rosch, 1978)).

Goodness-of-fit is defined as a combination of two aspects:

1. a value expressing how close the pattern is to the class prototype, similar to the comparison in adaptive resonance theory (ART, (Grossberg, 1987)). This value is

simply the net input to the winner unit, which is computed as the dot product between the input pattern and the winner's weight vector (which evolves into representing a prototype of the category).

2. a value expressing the extent by which the winner can suppress the other units. This value is basically the activation of the winner minus the average activation of all the other units.

These two parts are weighted, giving more emphasis to the first aspect at the beginning of the learning process, while the second aspect becomes stronger after learning has proceeded for a while. This goodness-of-fit value is used in two respects:

1. At ever update of the categorization layer it is used to decide whether the input pattern is close enough to the prototype, very similar to the comparison in ART. Also very similar to ART, if a pre-set threshold is not reached, then the current winner is reset and the update repeated, allowing another unit to become the winner. If that unit has never been the winner before (thus does not represent a category yet) this can be considered as the recuritment of a new unit to represent a new class of patterns. This mechanism, as in ART, permits rather stable but yet flexible learning.

2. When adapting the weights between the category layer and the component mapping the category to a word (see below), the goodness-of-fit is used as a multiplicative factor. This is basically done to prevent the mapping of words to spurious patterns, expressing the underlying assumption that the mapping between a word and its sensory category (meaning) can only proceed if categorization in that particular case has evolved into sufficiently distinct responses.

Learning for categorization is done the following way. Every unit can store the information that it has served as the winner in categorization before (using a separate parameter). With this parameter, the unit can distinguish whether it is a newly recruited unit (new category) or represents a previously established category. In the former case, the weights are set identical to the input pattern. In the latter case a instar learning rule (Grossberg, 1982) is applied to the weights between the input and the category layer, pulling the weight vector closer toward the input vector. In both cases, the inhibitive intra-layer weights of the category layer are adapted as well, using a Hebbian rule increasing inhibition on each weight leading from an active unit to a more weakly activated unit.

Before each pattern presentation and after each learning step, both input patterns and corresponding weight vectors are normalized to unit length.

Mapping words to meanings As already hinted above, a category formed based on sensory stimuli is viewed as the potential internal substrate for the meaning associated to a word. In addition – contrary to most previous models – the identification of words is also seen as a categorization process, where words are physical entities causing patterns

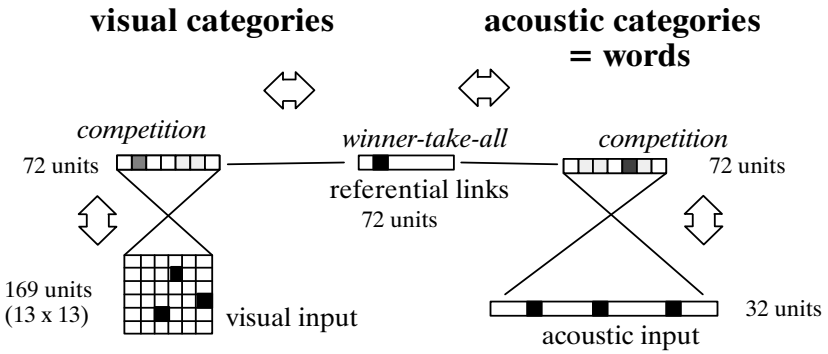

Figure 2: The model proposed in this paper, consisting of two inputs, two categorization layers and a layer linking two category representations

on a sensory input, which must be categorized. This more closely corresponds to the real-world situation than simply assuming orthogonal representations of words to be given (as is done in (Plunkett et al., 1993)). It can lead to a model that can account for effects rooted in the physical nature of words, and can avoid artefacts like assymetries caused by the fact that the categorization process of words would otherwise be side-stepped.

The entire model therefore consists of two categorization modules and an additional layer (or array of layers), called a "(referential) link layer", to connect two categorization layers, one from each side (see fig. 2). For the sake of describing the model, we call one side (the one that categorizes words) the "acoustic" side (assuming words are presented acoustically) and the other one the "visual" side.

The purpose of the link layer is to develop links between word categories and visual categories that obey the properties of word meanings, most importantly the words' arbitrariness. In other words, similarities among words as physical entities must not generally be allowed to influence the corresponding visual categories (see discussion below). To achieve this the following mechanisms were introduced:

- Each time the link layer is updated to form a link between category layers, a winner-take-all process is initiated, meaning that only one unit in the link layer is kept active (at maximum level), and all others are made inactive.

- A learning rule applied to connections between the link layer and the category layers is introduced that results in strong weights only between the winners in the link layer and the corresponding category layers. In other words, without explicit knowledge about the categories and their winners, the link layer learns to activate exactly one category unit. Thus, in spite of the permission of co-active units in the category layer, the weights are strengthened such that after learning only the winner (the unit representing the corresponding category) will be activated, erasing the influences of similarities to other category responses.

Weights in the other direction (from the category to the link layer) are adapted by simple outstar rules (Grossberg, 1982).

Data and Results

Similar to (Plunkett et al., 1993), for the simulations the following data was used:

- 5 distortions each of 32 pattern prototypes, such as to permit 10 % overlap between members of different classes, at the visual input. The patterns are binary and contain 10 active bits on an overall number of 169 units.
- 32 different binary patterns (4 active bits of 32 units each), also permitting about 10 % overlap between patterns, for the acoustic input.

Note that the acoustic patterns indeed show similarities with each other, the influence of which must be diminished during learning. Each class of visual patterns was arbitrarily assigned one of the 32 labels as the "correct" one. Each learning step consisted of randomly picking one visual pattern and the corresponding acoustic word pattern. Both patterns are presented, categorization is initiated, as well as the learning algorithm for the connections between category and link layers. Although learning of categories and learning the mapping between them need not be separated, for the sake of simplicity, two phases were distinguished for the reported simulations:

- A categorization phase, where each side is independently trained to form catgeories with sufficient goodness-of-fit. In all cases, the 32 categories were perfectly learned.
- A mapping phase, where the weights between inputs and categorization layers were kept fixed, and only the learning algorithm for adapting weights between category and link layers was applied.

Fig. 3 shows learning curves of one typical training sweep. The basic characteristics of these curves turned out to be robust over several runs with different initial weight initializations. Two performance measures are distinguished (Plunkett et al., 1993):

- Comprehension: For this, an acoustic pattern is input, the corresponding category layer, the link layer, the visual category layer and the visual input layer are updated (in that order), applying winner-take-all to the link layer. A response is counted as being correct, if the input pattern (among the 5 times 32 patterns) which is closest to the resulting pattern in the visual input (according to its Euclidean distance) is in the "correct" category (the one that was associated with the word input).
- Production: For this, a visual pattern (one out of the 5 times 32 patterns) is input, the corresponding category layer, the link layer, the acoustic category layer, and the acoustic input are updated (in that order), applying winner-take-all to the link layer. A response is counted as correct, if the input pattern (among the 32 possible acoustic patterns) which is closest to the resulting pattern in the acoustic input layer is the "correct" word.

Comprehension and production scores in fig.3 are percentages correct over all (5 times 32) visual and (32) acoustic patterns, respectively. The curves depcited show the

Figure 3: Learning curves for percentage correct naming (lower curve) and comprehension (upper curve) for netwoeks trained with 20 % (left curve) and 60 % noise (right curve). The comprehension and naming curves are identical for the left network.

performance, when 20 % and 60 % noise, respectively, is introduced, meaning that in 20 (60) % of the cases, the word chosen to be input together with a given visual pattern was picked randomly from the possible 32 word patterns. Only in the remaining 80 (40) % of the cycles, the "correct" word was chosen consistently. This reflects the real-world situation that only in a minority of the cases when a child is faced with an object to be named the appropriate word is uttered. In real life, of course, the words uttered in this context are not entirely random, but include words on a different level on taxonomic hierarchy (such as 'poodle' or 'animal' instead of 'dog') or semantic reference (such as 'brown' or 'furry' instead of 'dog'). But if a model can handle such highly noisy conditions, it has good prerequisites to show robust learning behavior in truly real-life situations. Several observations can be made:

- The model shows qualitatively realistic learning characteristics, including aspects like a kind of "vocabulary spurt" (Plunkett et al., 1993; McShane, 1979), where the number of correctly comprehended and produced words rises rather quickly. However, this spurt is not preceded by a period of slow learning.
- The model shows extreme robustness against noise (as defined above). Only above a level of 60 % noise, the performance would degrade considerably.
- The model also shows the typical temporal precedence of comprehension over production, as discussed in (Plunkett et al., 1993) (see also (Clark, 1983), as cited there), but only with considerable noise in learning.

The qualitative differences to models like (Plunkett et al., 1993) warrant a more elaborate analysis, comparing the two different approaches. This will be content of future research.

Possible extensions

It was mentioned above that several simplifying assumptions are made in the model to focus on some of the important aspects of word learning. The categorization and word mapping model, however, was conceived with many extensions beyond those simplifications in mind. Examples are the following.

- In an elaborate version, the categorization module consists not only of a single layer but of a pool of interconnected categorization layers. This permits the learning of multiple categories corresponding to the same or similar stimuli.
- Through their interconnections, previously learned categories can be input to subsequently learned ones. In addition, other components such as motor output or additional sensory stimuli can be input to a particular categorization layer. As a result, categories need not correspond to sensory stimuli alone, but can go beyond sensory classes to include more "abstract" categories, as well.
- Through mulitple connections via pools of link layers, synonyms (multiple words for one meaning) and homonyms (multiple meanings for one word) can be learnt.

Discussion and conclusion

This paper has introduced a connectionist model for many aspects of simple first word learning. It distinguishes itself in many respects from previous connectionist approaches to modeling similar phenomena. The main differences, among others, are

- an explicit account for categorization permitting the modeling of aspects like prototype or basic level effects (in the sense that prototype patterns and basic level categories can learned to be named more quickly than others),
- a mechanism permitting to model a word's arbitrariness in referring to categories (as "true" symbols in the semiotic sense),
- the robustness against noise.

It was also briefly discussed that the model, on first sight at least, is not able to fully reproduce some phenomena that can be modeled by alternative approaches. This might be rooted in the fact that more complex aspects of word learning need to be included (e.g. articulation or more realistiv visual and acoustic perception) – something which, in principle, should be possible to build into the model in more elaborate versions. It might also point to the fact that this model and alternatives resting more on the effects of backpropagation learning in multilayer perceptrons, will have to complement each other in the future.

In summary, this model was designed in order to enhance discussions around connectionist models in linguistics, and to point connectionists and linguists alike to strengths and weaknesses of several modeling approaches within connectionism.

References

Carpenter, G. & Grossberg, S. (1990). Art 3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural networks*, *3*, 129–152.

Chauvin, Y. (1989). Toward a connectionist model of symbolic emergence. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, pp. 580–487. Hillsdale: Lawrence Erlbaum.

Clark, E. (1983). Meanings and concepts. In Mussen, P. (Ed.), *Carmichael's Manual of Child Psychology* (4th edition)., Vol. 3. New York: Wiley.

Cottrell, G., Bartell, B., & Haupt, C. (1990). Grounding meaning in perception. In *Proceedings of the German Workshop for AI (GWAI)*. Heidelberg: Springer.

Dorffner, G. (1992a). A step toward sub-symbolic language models without linguistic representations. In Reilly, R. & Sharkey, N. (Eds.), *Connec tionist Approaches to Language Processing*. Hove: Lawrence Erlbaum.

Dorffner, G. (1992b). Taxonomies and part-whole hierarchies in the acquisition of word meaning - a connectionist model. In *Proceedings of the Annual Conference of the Cognitive Science Society*. Hillsdale: Lawrence Erlbaum.

Grossberg, S. (1982). *Studies of mind and brain*. Boston: Reidel Press.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, *11*(1), 23–64.

McClelland, J. & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: part 1. an account of basic findings. *Psychological Review*, *88*, 375–407.

McShane, J. (1979). The development of naming. *Linguistics*, *17*, 879–905.

Nelson, K. (1988). Where do taxonomic categories come from?. *Human Development*, *31*, 3–10.

Plunkett, K., Sinha, C., Moller, M., & Strandsby, O. (1993). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3 & 4).

Rosch, E. (1978). Principles of categorization. In Rosch, E. & Lloyd, B. (Eds.), *Cognition and Categorization*. Hillsdale: Erlbaum.

Rumelhart, D. & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*(1), 75–112.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. & McClelland, J. (Eds.), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Vol. 1. Cambridge: MIT Press.

Schyns, P. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, *15*(4).