29 Modeling language acquisition and representation: connectionist networks

Ping Li

Connectionism, Parallel Distributed Processing (PDP), or neural networks have had a profound impact on cognitive sciences in the last two decades. Language, as one of the central human cognitive components, has received in-depth treatments since the beginning of connectionist research. The acquisition of the English past tense (Rumelhart & McClelland, 1986), the recognition of speech (McClelland & Elman, 1986), and the processing of sentences (McClelland & Kawamoto, 1986) are among the earliest domains of connectionist research in the original PDP models. Connectionism has since been applied to the study of many domains involving language, including language acquisition, normal and impaired word reading, and language organization in the brain (e.g. Elman et al., 1996; Plaut et al., 1996; Miikkulainen, 1997; Small et al., 1995). Unfortunately, connectionist models or modeling have had very limited influences on Chinese psycholinguistics as a whole. To date, there are very few connectionist models that are designed specifically to account for the processing or representation of the Chinese language. This lack of interaction between connectionism and Chinese psycholinguistics is lamentable. On the other hand, however, this lack opens new avenues for research. In this chapter, I present research from our laboratories that explores the issue of linguistic representations and acquisition in connectionist networks, with particular reference to Chinese, in both monolingual and bilingual contexts.

Connectionism: an overview

Connectionist representation and learning

A close parallel has been drawn between the human mind and the digital computer in the classical conceptualization of human cognition. The essence of this computer-metaphor view is that cognitive operations are serial (one step at a time), discrete (symbol-based processing), and modular (domain-specific

* Preparation of this chapter was made possible by a grant from the National Science Foundation (BCS-0131829).

processes) (see Bates & Elman, 1993, for a review). In stark contrast to this view, connectionism advocates that cognitive processing is parallel, distributed, and interactive in nature. The connectionist proposal is reflected most clearly in its views on knowledge representation and acquisition. First, in terms of knowledge representation, connectionism argues for "distributed representation": a given concept is represented not by a single unit or node (as in classical cognitive models) but by multiple units or nodes in concert, the result of which is a pattern of activation of relevant micro-features that distribute across multiple units. Second, in terms of knowledge acquisition, connectionism argues for learning through the adaptation of weights, the strengths of connections that hold between multiple units. Because of distributed representation, one cannot directly identify crisp concepts or rules in a connectionist system, and therefore the acquisition of relevant concepts or rules entails the learning of appropriate activation patterns, that is, relevant distributed representations. This process can only be achieved by the accumulation and adjustment of weights between units that will lead to the appropriate activation patterns.

Connectionist theories assume a high degree of interactivity between various levels of information processing, in contrast to classical theories that assume highly modularized, often serial, and "informationally encapsulated" processes (see Fodor, 1983, for the latter). A typical connectionist network consists of three layers of units: the input layer, the hidden layer, and the output layer. The input layer receives information from input representations (e.g. orthographic representations of Chinese characters), the output layer provides output representations produced by the network (e.g. classifications of Chinese characters), and the hidden layer forms the network's internal representations as a result of the network's learning of the input–output relationships (e.g. the orthographic similarities between characters at different stages of learning). In a standard feed-forward network, information processing goes from the input layer to the hidden layer, and then from the hidden layer to the output layer. Learning occurs when the connection weights are adjusted so that the system can compute the input-to-output function successfully. Different networks use different algorithms to achieve learning; the most widely used learning algorithm in psychological and cognitive studies is "back propagation" (Rumelhart, Hinton & Williams, 1986), according to which each time the network learns the input-to-output mapping (in a forward cycle), the discrepancy (or error) between the actual output (produced by the network) and the desired output (provided by the modeler) is calculated, and is propagated back to the network so that the relevant connection weights can be adjusted relative to the amount of error. Continuous weight adjustments in this way lead the network to fine-tune its strength of connections in response to regularities in the input–output relationships. At the end of learning, the network derives an optimal set of weight configurations so that it can take on any pattern in the input and produce the correct output

pattern (for technical details, consult Anderson, 1995; Dayhoff, 1990; Hertz, Krogh & Palmer, 1991; for psychologically relevant details, read Bechtel & Abrahamsen, 1991; Ellis & Humphreys, 1999; Rumelhart, McClelland & the PDP Research Group, 1986).

Clearly, connectionist ideas are more biologically motivated than concepts in the classical view of cognition: notions such as multiple processing units, activation, and connection weights provide more neurally plausible constructs to conceptualize human information processing than do discrete symbols, rules, linguistic true structures, and the like. The human brain consists of a massive network of neurons working together, often in parallel. However, the resemblance between most current connectionist systems and the human biological system remains at a superficial level, and more work is needed to make connectionism more biologically grounded.

Connectionist language processing

It needs only a simple twist to link connectionism to language processing. Because of its properties in representation and learning, connectionism was quickly applied to solve many classical problems in language acquisition, speech perception, and lexical and sentence processing (vol. II of the PDP books, Rumelhart et al., 1986; for a more recent overview, see Ellis & Humphreys, 1999; for a summary in Chinese, see Li, 2002a). In the eyes of connectionism, rules in linguistic theories provide only a formal (and convenient) way of describing languages and linguistic behaviors, but their psychological reality is doubtful. This view is clearly opposed to generative linguistic theories that Universal Grammar (UG) is innate and psychologically real (Chomsky, 1988; see also an brief exposition of UG by Chien & Lust, this volume, and Yang, this volume). Connectionism argues that linguistic representations are "emergent properties," emergent, not built in, owing to the interaction of the learning system with the linguistic environment (MacWhinney, 2001a). Connectionist systems demonstrate capabilities in inducing syntactic and semantic structures from the input through detecting regularities in the form–meaning mapping process. As a shortcut to the understanding of emergentism, structured, "rule-like," representations in a connectionist network can emerge in much the same way as a hexagonal structure emerges from the honeycomb: every honeybee packs a small amount of honey into the honeycomb from a given angle, but no honeybee has a grand plan (or a genetically determined rule) for making the hexagonal shape (Bates, 1984). Individual honeybees are like individual units in connectionist networks, and when they work in concert according to given dynamics (e.g. maximizing the packing density of spheres), they create structures as if they have an innate rule.

Perhaps the most well-known connectionist model of language is Rumelhart and McClelland's (1986) model of the acquisition of the English past tense, a simple model that produced classical developmental effects such as U-shaped learning. In the empirical literature, it is observed that young children go through the following stages in the acquisition of grammatical morphology (see Bowerman, 1982): initially, they produce correct forms such as *feet* (plural), *broke* (past tense), and *untie* (prefix); at the second stage, they produce a significant amount of errors in each of these domains, for example, *foots*, *breaked*, and *untighten*; finally, they recover from these errors. This U-shaped developmental pattern, characteristic of early language learning, has been taken as strong evidence for the child's internalization of a linguistic rule – at the second stage, children overgeneralize the acquired rule and apply the plural *-s*, the past tense *-ed*, and the prefix *un-* to all words (nouns and verbs), irrespective of whether the word takes these affixes in the target language. In the context of past-tense acquisition, traditional rule-based accounts, such as the one advocated by Pinker (1991) and Pinker and Prince (1988), argued for a dual learning process, one mechanism involving the acquisition of a regular morphological rule (to account for overgeneralization errors), and the other the associative learning of exceptions (to account for recovery from errors). In contrast, Rumelhart and McClelland showed that one can account for the acquisition of both the regular and irregular past-tense forms with a single mechanism realized in connectionist learning with distributed representation and adaptive connection weights. Overgeneralizations in this view reflect the child's ability to extract statistical regularities in the linguistic input (e.g. phonological patterns in input–output mapping, for regular and irregular verbs alike) and the child's ability to use the extracted patterns productively.

The debate on the utility of the past-tense model has continued to this day (MacWhinney & Leinbach, 1991; Pinker, 1999; Plunkett & Marchman, 1991; Seidenberg, 1997), and it has centered on the core issue of whether language learning should be characterized as a symbolic, rule-based process or as a connectionist, statistical learning process. However, there are two major gaps in this debate, which we took as points of departure for our investigation (Li, 2003a). First, most studies have focused on the statistical regularities in the phonological properties of words that govern the use of the English past tense, while few studies paid attention to the meaning structure of the verbs with which past-tense forms are used. Empirical evidence suggests that verb semantics plays an important role in children's acquisition of the past tense (Brown, 1973). Second, most of this debate has revolved around the standard feed-forward connectionist models, while few studies paid attention to other clusters of models that are more cognitively and biologically plausible for language. Several limitations are known to exist with feed-forward networks that learn through back propagation, especially in the context of language acquisition (MacWhinney,

2001b). These considerations motivated our research in connectionist semantic acquisition (Li, 1993; Li & MacWhinney, 1996; Li & Shirai, 2000), and led us to look for models that bear more resemblance to language learning in the natural setting (Li, 2003a; Farkas & Li, 2001, 2002a, b).

Lexical representations in self-organizing connectionist networks

A self-organizing model of the lexicon

In response to the aforementioned gaps in current connectionist modeling, we turned to a class of models called self-organizing neural networks. Self-organizing networks belong to "unsupervised learning" models, in which learning occurs without an explicit teaching signal or "supervisor" as in "supervised learning" models with back-propagation. They provide psycholinguistically more plausible models – in the natural setting, language learning, especially organization and reorganization of the mental lexicon, is largely a self-organizing process that proceeds without explicit teaching (MacWhinney, 2001b).

Self-organization in these networks typically occurs in a two-dimensional map (a self-organizing map, or SOM; Kohonen, 1982, 2001), where each processing unit in the network is a location on the map that can uniquely represent one or several input patterns. At the beginning of learning, an input pattern randomly activates a set of units on the map that surrounds the best matching unit (the winner). Once these units become active in response to a given input, the weights of the winner and those of its neighboring units are adjusted such that they become more similar to the input and will therefore respond to the same or similar inputs more strongly the next time. This process continues until all the inputs can elicit specific response units in the map. As a result of this self-organizing process, SOM gradually develops concentrated areas of units on the map that capture input similarities, and the statistical structures implicit in the input are preserved on a two-dimensional space. In this way, SOM extracts a compressed but efficient representation for the complex input patterns.

Several appealing properties, in particular, the process by which the maps develop ordered representation of lexical categories, make SOMs suitable as models of the lexical system (Ritter & Kohonen, 1989; Miikkulainen, 1993, 1997; Li, Farkas & MacWhinney, 2004). To model the interactive nature of the lexicon, Miikkulainen (1993, 1997) connected several SOMs via Hebbian learning (Hebb, 1949), according to which the associative strength between two units is increased if the units are both active at the same time. Upon training of the network, for example, a lexical form (orthographic or phonological) representation of the word is presented to the network and, simultaneously, the semantic representation of the same word is also presented to the network.

Through self-organization, the network forms an activity on the lexical map in response to the form input, and an activity on the semantic map in response to the semantic input. Through Hebbian learning, the network establishes associations between the two maps: initially all units on one map are fully connected to all units on the other map, but as learning continues, the associations become focused, such that in the end only the maximally responding units (winners) are associated across maps. The combination of Hebbian learning with self-organization is important in that it can account for the process of how the learner establishes relationships between forms and meanings and between forms and forms (e.g. one can also connect phonology and orthography), on the basis of how often they co-occur and how strongly they are co-activated in the representation. In what follows, I summarize three models that rely on principles of self-organization and Hebbian learning that we have developed in modeling language processing and acquisition in Chinese.

Modeling character acquisition

There have been only two preliminary attempts to model the processing of Chinese characters using connectionist architecture.[1] First, Chen and Peng (1994) proposed a connectionist model of recognition and naming in Chinese, using a standard feed-forward network. Their model consisted of orthographic representations at the input layer, mapped to phonological representations at the output layer. Orthographic representations of Chinese characters focused on radical components and their structural relationships in the character. A major success of the model was its ability to show frequency effects of characters and to distinguish regular and irregular characters in naming. A second model, the interactive constituency model, was proposed by Perfetti, Liu, and Tan (2002) and Perfetti and Liu (this volume). Perfetti et al. were interested in building a more general model of reading in Chinese rather than modeling specific effects in naming. Their model included four levels of interactive constituency: radical, orthography, phonology, and semantics. The input units were 144 radicals that begin activation in orthography, which then activate phonology or semantics of the characters. One interesting pattern from their model was the oscillation effect: the onset of inhibition of orthographically similar primes coincides with the onset of facilitation of phonological priming. This effect matches with empirical observations on the time course of orthographic and phonological priming (Perfetti & Tan, 1998).

Although both models attest to the utility of connectionist networks, their methods in character representation and in network architecture were still crude.

[1] Although not a modeling enterprise, Taft's (1994) research attempted to account for character recognition on the basis of the classic interactive-activation model of word recognition (McClelland & Rumelhart, 1981), a precursor to the PDP models.

The first model was limited to specific naming effects and it is not clear how it can generalize to other domains of reading and acquisition. The second model used a very limited vocabulary (204 characters) and it is not clear how the model can scale up to a larger lexicon because of its localist representation (i.e. one unit per character). In addition, both models were not designed to encode the phonology of radicals, hence were unable to capture the role that radicals play in character naming (and the effects of regularity by frequency in naming). Finally, both models relied on the standard feed-forward architecture. In contrast to these two models, Xing, Shu, and Li (2004) presented a self-organizing connectionist model of character acquisition. Xing et al.'s model aimed at two goals. First, it wanted to test the usefulness of self-organizing neural networks in orthographic acquisition. Second and more important, it attempted to evaluate the degree to which connectionist models can inform us of the complex structural and processing properties of Chinese orthography. The most serious obstacle to this goal is the faithful representation of the complex orthographic similarities of Chinese characters. Perfetti et al.'s localist representations did not solve the problem, and Chen and Peng's distributed representations were also limited to the particular characters used in their training. Xing et al. analyzed a large-scale character database, the *UCS* Chinese character database (Standards Press, 1994), and examined the strokes, components, and structures for each of the 20,902 characters in the database. On the basis of this analysis, they incorporated the component features, shapes, stroke structures, radical positions, and stroke numbers, encoded in a 60-unit vector representation of characters. For example, component features included single, separate, crossing, and connecting; radical positions included top, bottom, left, right, middle, inner, etc.

To model the acquisition of characters, Xing et al. selected their input characters from the School Chinese Corpus (Shu et al., 2003) that consists of 2,570 characters from elementary school textbooks used in Beijing. The network was trained on three batches of roughly 300 characters each, which occurred in grades 1, 3, and 5 in the corpus. The training progressed by pairing the orthographic representations of these characters in one map with their phonological representations in the other (the PatPho representations of Chinese; see Li & MacWhinney, 2002). Once learning was completed, the network was tested on novel words for character naming, and the testing words varied in their frequency (high and low) and regularity (regular and irregular).

Simulations from this model revealed several interesting patterns. First, the model developed clearly structured representations for Chinese characters, indicating the validity of both the representational method and the self-organizing process. Second, the tests with novel characters in the model showed both frequency effects and regularity effects in character acquisition, and, more important, the interaction between the two: regularity effects were only marginal for

high-frequency characters, but were pronounced for low-frequency characters and novel characters. Xing et al. further conducted analyses on the naming errors produced by the network and found that the network's "awareness" of regularity increased with training grade: in grade 1, the network tended to read novel characters as totally irrelevant characters, but in grades 2 and 3, it became more likely to read the character in the pronunciation of its phonetic or as another character having a similar phonetic part. These developmental patterns match up well with empirical observations such as those reported by Shu, Anderson, and Wu (2000).

Modeling lexical category formation

While the Xing et al. study was concerned with orthographic representation and acquisition, Li (2002b) conducted another simulation on the development of semantic representations and lexical categories in Chinese. The model differs from the above in one crucial aspect in that a special recurrent network was used to acquire semantic and grammatical information of words as part of the DevLex model, a self-organizing model for the development of the lexicon (see Farkas & Li, 2001, 2002a, b; Li & Farkas, 2002; Li, Farkas & MacWhinney, 2004 for details). The DevLex model computes transitional probabilities between words in a large-scale text corpus, which serves as the basis for lexical semantic representations of these words.

An important issue in connectionist language processing is the emergence of lexical categories in these networks. Elman (1990) showed that a simple recurrent network derives meaningful representations of lexical categories (e.g. nouns and verbs, animates and inanimates) when the network learns to predict the next word in the processing of sentences presented word by word. In a similar fashion, Li, Farkas, & MacWhinney (2004) showed that the DevLex model develops categorical representations dynamically at different stages when the network is exposed to parental speech in the input. Li's (2002b) simulations on the *Corpus for Modern Chinese Research* (CMCR, Beijing Language Institute, 1995) were consistent with these findings in English. The CMCR corpus contains about 1.2 million word tokens, recorded from various contemporary written sources (e.g. newspapers). Three hundred most frequent words (which covers 39 percent of the entire corpus) from this corpus were extracted and submitted to our model. The resulting semantic map displayed clear grammatical and semantic categories: nouns and verbs were separated by the network, and so were prepositions, adverbs, pronouns, particles, numerals, and classifiers; within each grammatical class, semantically similar words were also grouped together as clusters (see figure 5 of Li, 2002b). These results argue clearly for the emergence of categorical representations of language rather than for a predetermined modular structure in the mental lexicon (assuming that the

modular structures exist in the adult speaker's linguistic representations; see Pulvermüller, 1999).

Modeling bilingual language processing

Bilingualism is in dire need of formal models (Li, 2003b). So far there are only a handful of models (connectionist or otherwise) that are implemented to account for bilingual language processing. Li and Farkas (2002) presented a connectionist model of bilingual lexical and sentence processing, the SOMBIP, which was a variant of the DevLex model for the bilingual context (see also summary in Hernandez, Li & MacWhinney, 2005). They applied the model to the Hong Kong Bilingual Child Language Corpus (Yip & Matthews, 2000; see also Yip, this volume) that contains transcripts of conversations between a child and the researchers, including his native English-speaking father and native Cantonese-speaking mother. The parental speech from this corpus (with about 185,279 word tokens) served as input to our network. The network was trained to learn 400 most frequent words (types) in the corpus (184 Chinese words and 216 English words), which cover about 56 percent of the total words in the corpus.

Results from this study indicate that SOMBIP was able to model a number of classical effects in bilingual language representation and processing, for example, phonological and semantic priming effects within and across languages. Most important, Li and Farkas showed that the network dynamically separated the Chinese lexicon from the English lexicon, in that it developed distinct lexical representations for the two languages after learning (cf. figure 2 of paper). Within each lexicon, the network further distinguished various grammatical and semantic categories in its representation (e.g. nouns vs. verbs, state verbs vs. activity verbs; cf. figure 3 of paper). Although on the surface these results are consistent with empirical arguments for language-specific (or distinct) representations of the bilingual mental lexicon, our model emphasizes the dynamic property of representation in learning – the representations reside in an integrated network, but are functionally distinct for each language, subject to change and development (see similar arguments made by Bialystok, 2001, and Grosjean, 1998). Clearly, such interactive patterns have emerged from the network's analyses of the statistical characteristics of the input in the bilingual's two languages, for example, similarities and differences between category members and between languages. The ability of the network to distinguish categories and languages developmentally without relying on preformed representational modules provides another example of the classical emergentist argument: patterns of linguistic behaviors that are different or otherwise dissociated need not arise from distinct mechanisms in the representational system (MacWhinney, 2001a; Rumelhart & McClelland, 1986), but can emerge naturally from the learning

of input characteristics within a single system. The SOMBIP or DevLex model provides the necessary, psycholinguistically plausible, learning mechanisms for such a system.

Concluding remarks

In this chapter I provided an overview of connectionism and the principles underlying connectionist language processing, and presented a sketch of a developmental self-organizing model of lexical processing applied to the Chinese language. A number of theoretical and architectural considerations of our model are discussed, and preliminary simulation results from the model are analyzed with respect to Chinese character acquisition, lexical category formation, and bilingual language representation. The modeling results provide significant insights into the mechanisms for language processing and acquisition in Chinese, especially with regard to linguistic representations, in both the monolingual and bilingual contexts.

The heatwave of connectionism has never hit the field of Chinese psycholinguistics, and this chapter provides only a starting point for fertile explorations in future studies. We may arrive at a better understanding of a whole range of issues with the connectionist research method, including lexical and morphological representation and acquisition, lexical and sentence processing in monolingual and bilingual contexts, and impaired language development and language production in Chinese.