In Proceedings of the Sixth International Conference on Cognitive Modeling, 366-367.
Mahwah, NJ: Lawrence Earlbaum.

Bilingual Language Learning in Connectionist Networks

Jing Liang (jliang@andrew.cmu.edu)
Department of Modern Languages, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Introduction

Connectionist modeling opens up a venue to research on language learning from the statistical point of view. Network input construction offers a way to examine what kind of knowledge could be statistically learned and internal representations from network hidden units help to clarify how the knowledge thereafter could be represented in human minds (Elman, 1991; St. John & McClelland, 1990). Results, thus far, demonstrate the ability of general cognitive learning in various aspects of human language, morphology, syntax, etc., without recourse to innate linguistic rules. This study aims to extend the work of modeling monolingual language learning into the area of bilingual language learning in hopes of drawing further evidence to support the argument of general cognitive capacity in language learning by requiring the network to learn different syntactic structures of two languages using just one learning algorithm and with no innate linguistic structures built in.

Grammar and Corpora

Grammar in the current simulation is a stochastic context-free grammar (SCFG hereafter), including simple transitive and intransitive sentences, sentences with subject- and object-extracted relative clauses, and passive sentences. 10 nouns and 14 verbs in English used in Rhode & Plaut (1999) were borrowed because of good matching in light of meanings between Chinese and English. Actual verbs, however, drop to 7 when translated into Chinese since there is no morphosyntactic change on verbs in terms of tense and plurality. A program was written to generate 11,600 sentences in total with 5,788 sentences in English and 5,812 in Chinese. Additionally, some semantic and pragmatic constraints were incorporated into sentences to make this pseudo-language grammar less artificial. For example, only human beings can walk dogs; only animals can bite human beings, not vice versa; the subject in transitive sentences cannot be the same as the object.

Network Architecture

The network used in the simulation is a simple recurrent network (SRN hereafter), adapted from the one in Elman (1991). The SRN is characterized of the ability of human memory by copying internal representations in the hidden layer at one time step to the context layer and feeding back the input at the next time step together with the memory in the context to the hidden layer for more effective weight updating.

The input layer employs a localist form with each unit representing one word either in English or Chinese. In English, the auxiliary verb is bound with the main verb as one unit. There are no explicit grammatical rules encoded in the network input. Two groups of output units perform the tasks of word prediction and sentence comprehension in terms of agent identification. The SRN is trained to maintain information only from the network's current activation in order to produce appropriate output for the next step. Therefore, the word prediction task forces the network to better retain previous word information such that the network could extract agent information more accurately, particularly in the case of Chinese because it is more difficult to identify the agent in Chinese due to the fact that pre-posed Chinese relative clauses can cause the position of the agent keep changing as the sentence length varies. One small hidden layer is added between the input layer and the main hidden layer as transducers (Elman, 1991) in order to cope with the dramatic changes of sentence structures in bilingual language learning.

Network Training and Testing

The network was trained 400 epochs with the mixed corpus of Chinese and English to simulate the native bilingual learning situation. Different combinations of learning parameters were tried to find the optimal one. A learning rate 0.01, a batch size of 10 words per weight update, initial random weights between $\pm$ 1.0, and a momentum of 0 were used for non-incremental training with back-propagation. The overall performance was measured using mean squared errors.

100 testing sentences, 50 in Chinese and 50 in English, were excluded from the training corpus and later used to examine the generalization of the network performance. For the word prediction task, the test corpus was constructed such that it tested not on the actual next word, but the probabilistic distributions of the next word. Rather than activate the exact next word, the network should activate a likelihood of many words at the same time. Since the grammar is SCFG, theoretically the correct distributions for the next word can be computed based on the corpus. Optimally, the theoretically generated distributions could be matched with the actual output distributions.

Results and Discussion

The overall performance of the network was examined by computing mean squared errors weighted by the percentage of different sentence types in the training corpus. The error rate 0.021 for word prediction and 0.048 for agent identification show that the network successfully learned Chinese and English simultaneously. The network activated the correct output units for agent identification more than what it should not have 99% of the time for English and 97% of the time for Chinese. The relatively low accuracy

for Chinese is attributed to the fact that agent identification in Chinese bears the characteristic of unpredictability to some extent because potentially the agent could occur at any position in contrast with agent identification in English where the agent occurs only at the first word position unless in passive sentences with regard to this grammar.

The better evaluation of the network performance entails the scrutiny on the activations of individual sentences. The activation of the word prediction task for one Chinese sentence is shown in Figure 1, in which the units listed for every word indicate the probable units that could be activated after this word. It demonstrates that the network learned co-occurrence of words, especially the collocation between nouns and verbs. In this grammar, all nouns can be "bitten" by at least one other noun; therefore after the verb *yao (bite)* and the constructive particle *de*, all nouns were activated almost evenly. However, after the agent *goumen (dogs)* occurred, the network activated several frequently co-occurrent verbs, e.g. *fei (bark)*, *yao (bite), zhui (chase)*, much more than others. The activation level for the agent identification task also reached 0.954 with the target at 1.

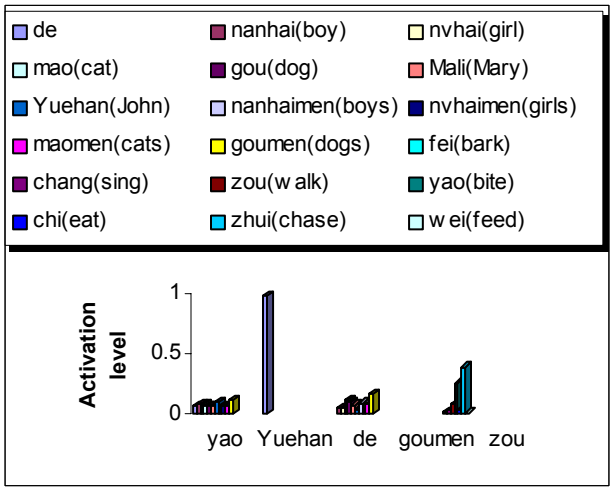

Figure 1: Activation levels for the word prediction task.

As the network successfully learned probabilistic distributions of language corpora, it also constructed internal representations of language structures in the hidden units. After the network was fully trained, weight matrices were extracted and entered into principal component analysis (PCA thereafter). PCA is useful for decreasing the dimensionalities in a multivariate space and characterizing the variation of each dimension in the hidden units. Sentence (1) and (2) were analyzed along the dimension of component 1 and 2 as indicated in Figure 2. Except for two nouns *gou (dog)* vs. *goumen (dogs)*, the principal component loading values for all other nouns are the same for two sentences. This indicates that the network learned to differentiate the structure of singularity vs. plurality in the state space.

(1) Yao Yuehan de gou (dog) zou. (Singular)
(2) Yao Yuehan de goumen (dogs) zou. (Plural)

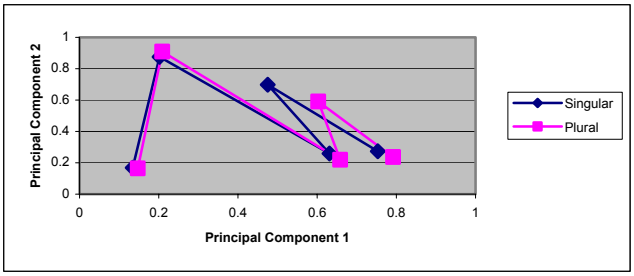

Figure 2: PCA for sentence (1) and (2).

Hierarchical clustering analysis was also used to evaluate whether the network constructed systematic structures for nouns and verbs. Verb similarity analysis was drawn from Chinese as an illustration, where similar verbs were connected earlier along the axis. As shown in Figure 3, the network captured the differences between transitive and intransitive verbs, even including slight ones. For example, zou/liu (*walk)* can be used as a transitive verb occasionally and chi (*eat)* in the corpus is mainly used as an intransitive verb. In conclusion, this simulation successfully testifies that the connectionist network has the ability to learn two different language structures implicitly as human beings do in natural language acquisition using statistical information available in language corpora and forming systematic structures of language knowledge in the end.

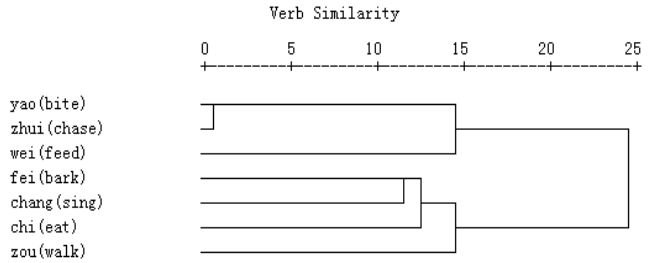

Figure 3: Hierarchical clustering analysis for Chinese verbs.

Acknowledgments

I want to thank Dr. David Plaut and Dr. Brian MacWhinney for some suggestions on network configuration and implementation. Thanks also go to Dr. Douglas Rohde for some tips related to LENS usage.

References

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7,* 195-225.

Rhode, D. L. T. & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72,* 67-109.

St. John, M., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217-257.