Learning Nonadjacent Dependencies with a Recurrent Neural Network

Igor Farkaš*

Department of Applied Informatics, Comenius University,
Mlynská dolina, 84248 Bratislava, Slovakia
farkas@fmph.uniba.sk

Abstract. Human learners are known to exploit statistical dependencies of language elements such as syllables or words during acquisition and processing. Recent research suggests that underlying computations relate not only to adjacent but also to nonadjacent elements such as subject/verb agreement or tense marking in English. The latter type of computations is more difficult and appears to work under certain conditions, as formulated by the variability hypothesis. We model this finding using a simple recurrent network and show that higher variability of the intervening syllables facilitates the generalization in the continuous stream of 3-syllable words. We also test the network performance in case of more realistic, two intervening syllables and show that only a more complex training algorithm can lead to satisfactory learning of nonadjacent dependencies.

1 Introduction

Statistical learning appears to be an important mechanism in language development and processing. Humans exploit distributional cues at various levels that help them discover structural dependencies in the language [1,2,3]. These processes are likely to occur unconsiously in the form of implicit learning [4]. In addition to adjacent dependencies, languages tend to comprise relationships between constituents that are conveyed in nonadjacent structure. For example in English, these nonadjacent dependencies exist between subject nouns and verbs in number agreement (e.g. *the boys living next door are naughty*), or between auxiliaries and inflectional morphemes (e.g. *is sleep-ing*). Any mechanism used broadly in language acquisition must therefore, in some way, be capable of learning nonadjacent regularities.

This problem was previously tackled using artificial languages (ALs) and the evidence for tracking nonadjacent probabilities, at least in the continuous streams of syllables, appears contrasting [5,6,7]. Earlier experiments with learning ALs failed to show generalization from statistical information unless additional perceptual cues (i.e. pauses between words or phonological features of phonemes) were available, suggesting that distributional information alone is

* Supported by the grant 1/0361/08 of the Slovak Grant Agency for Science.

M. Köppen et al. (Eds.): ICONIP 2008, Part II, LNCS 5507, pp. 292–299, 2009.

not sufficient to support the discovery of the underlying grammatical-like regularity embedded in a continuous speech stream. With this evidence in mind, Peña et al. [6] argued that generalization and speech segmentation are different processes maintained by separate mechanisms: statistical computations are used in segmentation, but these are distinct from algebraic rule-like computations that would account for generalizations of the distant structure. Peña et al. experimented with learning the continuous stream of 3-syllable words of the form A_iXB_i with $i = 1, 2, 3$ (and three Xs), where A_i exactly predicts B_i. The participants preferred words A_iXB_i, over "part words" (PWs), such as B_jA_iX or XA_iBj (i.e. the triples crossing word boundaries), which was taken as an evidence of successful word segmentation (because the subjects probably took advantage of nonadjacent dependencies between syllables that helped them automatically segment the continuous stream). Next, they were tested whether in addition to segmentation, they could also detect structural regularity in the stream. For that purpose, Peña et al. introduced "rule words" (RWs), such as A_iXB_i, where the intervening (embedded) X appeared in the stream but never in mid-position (i.e. $X \in \{A_j, B_j | j \neq i\}$). This makes RWs congruent with generalization: Unlike PWs, they have a novel surface form (but a familiar deep form). When the subjects' task was to decide between PWs and RWs, no preference for RWs was found, which was interpreted as no generalization (failure to discover the underlying regularity).

However, as promptly suggested by Gómez [8], this could have been due to low variability of X (henceforth, n_X), because she had found that sufficiently large variability ($n_X = 24$) resulted in successful generalization to novel surface structures (RWs). Onnis et al. [9,7] replicated this finding and the results of their experiments led them to fine-tune the variability hypothesis by postulating that generalization occurs at both extremes of variability – zero or large variability. The hypothesis states that when large variability disrupts adjacent dependencies, learners will seek alternative sources of predictability, such as nonadjacent dependencies. In the zero variability case, the reversal effect is observed: the common elements X share the same contextual frames (e.g. *don't-eat-it, he's-eat-ing).* Onnis et al. [7] showed that with sufficiently large n_X, tracking nonadjacent dependencies can result in simultaneous word segmentation and generalization of the embeddings (at the absence of any additional cues). The segmentation of the continuous stream is itself difficult because decreased transitional probabilities (due to high n_X) are known to lead to segmentation within word boundaries [1].

Here we model the effect of variability with a simple recurrent network (SRN; [10] using Peña et al.'s data. SRNs have been successfully applied for various sequential learning tasks, but, to our knowledge, not yet to this type of data with nonadjacent dependencies. In an earlier paper, Garzón [11] used an SRN in this specific task but he did not focus on the variability hypothesis. In experiment 1, we show that generalization accuracy improves with larger variability of embedding. In experiment 2, we simulate the same task in case of more realistic dependencies – embeddings consisting of two syllables rather than one.

2 Simulations

2.1 Experiment 1

Input data. We used streams composed of three different words generated by ALs of the form A_iXB_i, where $P(B_i|A_i) = 1$. In each AL, the three frames $A_i\_B_i$ were combined with embedding X (hence forming various words) whose variability n_X was systematically manipulated. To avoid any biases caused by specific frames, we ran multiple simulations for the same n_X using different frame triples. All three frames had the same probability of occurrence, and so had each X, i.e. $P(X|A_i) = 1/n_X$ and $P(B_i|X) = 0.33$. Hence, all variability conditions had the same transitional probabilities, except $P(X|A_i)$ which depends on n_X. Each syllable was represented as a consonant-vowel pair, taken from the pool of 8 consonants (*b,d,g,p,t,k,r,l*) and 5 vowels (*a,e,i,o,u*), respectively, amounting to 40 possible syllables in total (e.g. *ba, gi, ke*). The 3 frames were randomly chosen with the constraint that no consonant or vowel (except one vowel) was repeated within the same AL (e.g. *da_te, pi_gu, ro_ka*). Words had the embedding formed by syllables that did not occur in the frames (set of size 34). Following Peña et al., PWs had the form B_jA_iX or XB_iA_j. RWs contained embeddings X in A_iXB_i taken from the remaining two frames, i.e. $X \in \{A_j, B_j | j \neq i\}$. This setup allows the following prediction: If a learner computes adjacent statistical probabilities, he should prefer PWs over RWs, at least in the large variability condition (because PWs imply higher transitional probabilities than RWs). Conversely, if the learner computes nonadjacent dependencies he would rely on the most statistically reliable ones, namely $P(B_i|A_i)$, i.e., he would segment correctly at word boundaries, and hence prefer RWs.

Method. We trained an SRN within the next-syllable-prediction paradigm in the stream, given the current syllable at the input. In each simulation, the weights were randomly initialized within [-.1,.1]. Learning rate was set to 0.1 and momentum to 0.8. Each syllable was represented as the concatenation of two localist codes (a consonant and a vowel), to avoid any similarities within consonants or within vowels that might introduce bias into computations. Hence, the network had 13 input and 13 output units. We chose 20 hidden units and 20 context units. In each variability condition (given by n_X), we ran 10 simulations, each using different frames, implying different training and testing sets. For training we used 100 concatenated words (the same words were necessarily repeated within the given set, due to combinatorial limitations), randomly ordered and without pauses. Each simulation lasted 600 epochs. Each testing set contained 12 words. The next syllable (target) was considered to be predicted correctly if the location of both maxima on two output units (one for the consonant and one for the vowel) matched those of the target. To assess the performance of our SRNs, we had to come up with an appropriate procedure that would correspond to the experimental design in Peña et al. In their experiment 2, the subjects were asked to compare RWs and PWs, hearing one after the other, and decide which of the two stimuli sounded more like a word. To match this binary decision task, we compared the prediction errors for both RW

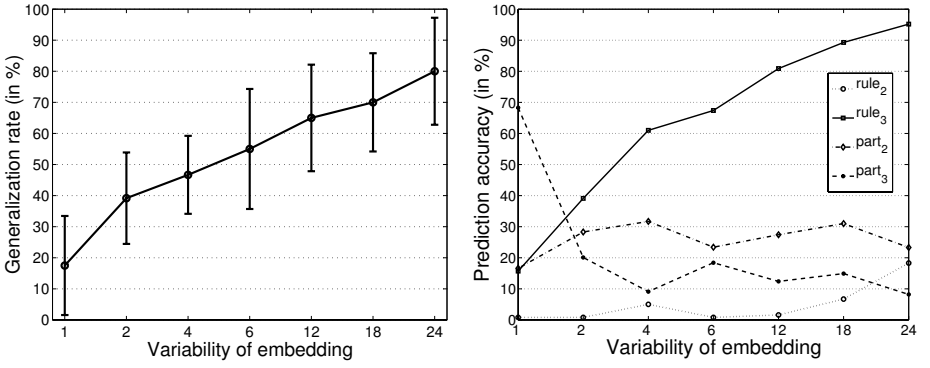

Fig. 1. (a) Average generalization rate in unsegmented artificial languages of type A_iXB_i. (b) Average errors for RWs and PWs, for predicting the X (index 2) and B_i (index 3) syllables. Standard deviations (not shown) were below 20%.

and PW test sets in each simulation as follows: For both test sets we recorded network prediction errors (squared Euclidean distance between the target and the output vectors) in each prediction step. The word prediction error was taken as the sum of prediction errors for the second (X) and the third syllables (B_i). These summed errors for 12 RWs and 12 PWs were then sorted ascendingly. The proportion of RW errors found in the first half of the sorted list was interpreted as the generalization accuracy.

Results. As shown in Figure 1a, the generalization accuracy grows with increasing variability of embedding. When $n_X \geq 12$, the network prefers RWs significantly more often than PWs. Qualitatively, this result is in agreement with experiment 2 in Onnis et al. (2004) although they reported a lower average rate for $n_X = 24$ (64% vs. 80% predicted by our networks). For $n_X = 3$ they reported 42% average generalization rate, which is a very good match with the networks (when considering the average for $n_X = 2$ and 4). We can gain more insight into the model behavior by looking at separate predictions of X and B_i (predictions of the first syllables A_i are not informative and were observed to remain at expected rate 0.33). These SRN predictions for RWs (we will refer to them as RWs of type1) and PWs (syllables X and B) are shown in Figure 1b. It can be seen that whereas predictions within PWs do not improve with higher n_X, predictions of B_i in RWs (denoted as rule$_3$) do significantly. This accounts for preference of RWs over PWs for higher n_X, expressed by lower summed errors in most cases. Similar ascending curve was observed also in case of predicting B_i in RWs which were constructed in a different way (as in Newport & Aslin, 2004) – using novel X syllables that did not appear during training at all (henceforth, RWs of type2). Whichever X is used in RWs, SRN is observed to predict the third RW syllable. Gradual increase of accuracy in predicting B_i in RWs, combined with the previous "input-buffering" step (remembering the first syllable) could be interpreted as the computational implementation of the gradual switching from tracking adjacent to remote dependencies, once the former become less reliable.

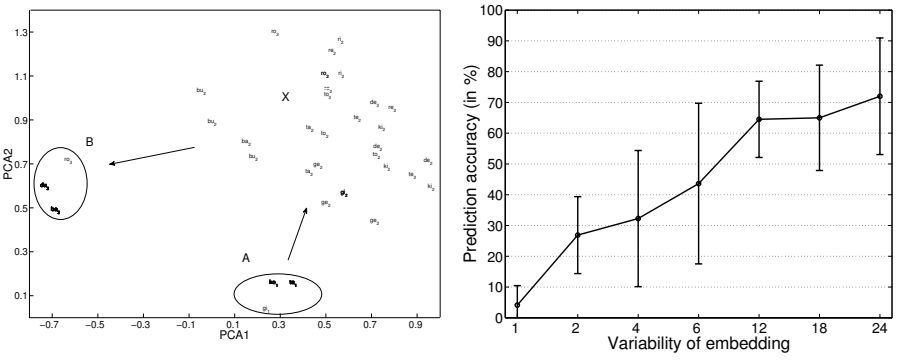

Fig. 2. (a) Layout of hidden unit activations of a trained SRN projected by the principal component analysis method. SRN was trained on language *ko_du*, *gi_ba*, *te_ro* with $n_X = 18$. Z_j denotes the input syllable Z presented at the j-th position within a word. (b) Average prediction accuracy for B_i syllables in rule-words, in unsegmented artificial languages of type A_iXYB_i.

This invariant behavior with respect to RWs of both types can be seen if we look at hidden unit activations of the SRN during testing. Figure 2a shows the two-dimensional (linear) projections of these activation vectors, in case of large variability of embedding ($n_X = 18$). Activations corresponding to A_i syllables are clearly separated, and so are the activations corresponding to B_i syllables. The largest cluster (X) comprises hidden unit activations for intervening inputs, covering syllables used during training (e.g. *ta,de,ki,re*), and also those used in RWs of both type1 (e.g. *ba,ro,gi,te*) and type2 (*ri,to,bu,ge*). Clearly, hidden unit activations document that the first and the last word syllables are distinctly represented in SRN.

However, in case of $n_X < 12$, such a distinction was observed to deteriorate. Although cluster B remained fairly separated, clusters A and X tended to merge, whereas the mutual distance between cluster A and X-B merged cluster became tended to be smaller, too. This may be the reason for lower prediction rates.

Our results do not match the zero part of the variability hypothesis, because preferences for RWs for $n_X = 1$ are very low in Figure 1a. However, according to Onnis et al. [9,7], high RW preference (and hence, generalization) in experiment 1 was only demonstrated in case of segmented artificial speech. If the zero-variability hypothesis turned out to also apply to a continuous stream, it would be a challenge to find a model that could account for that.

2.2 Experiment 2

Input data. To investigate whether an SRN can handle longer dependencies, we created ALs of the type A_iXYB_i, with three different frames per language, and varying embeddings XY within the frames. We considered a simplified design in that for given n_X both X and Y syllables were taken from the same set and

varied randomly (i.e. yielding n_X^2 combinations). In this experiment we focused on predictions of RWs which were constructed using existing 3 frames combined with 4 novel embeddings (i.e. 16 possible XY pairs). Out of all possible RWs, we randomly chose 12 of them for testing in each simulation.

Method. Data representation was the same as in experiment 1, as well as the network architecture. However, SRN trained for this task using standard error back-propagation algorithm failed to learn these more distant dependencies. Hence, we used an online version of the real-time recurrent learning (RTRL; [12] which is known to be a more powerful training algorithm for recurrent networks [13]. In this case, SRN was set to have 18 hidden units, was trained for 500 epochs and the learning rate was decreased to 0.05. Other network parameters were the same as in experiment 1.

Results. Figure 2b shows the prediction accuracy for B_i syllables within RWs averaged over 10 simulations. This ability is interpreted as generalization ability for novel words, although predictions of X and Y were very low, inversely related to n_X. Lower prediction accuracy and higher standard deviations compared to previous case (see the rule_3 curve in Figure 1b) suggest that this learning task faces greater difficulty.

Peña et al. [6] and Onnis et al. [7] also used two-syllable embeddings in their experiments, but they considered segmented rather than continuous speech. It may be that due to segmentation cue, tested subjects do not find the two-syllable embeddings more difficult in terms of learning generalization. However, our simulations suggest that using two-syllable variable embeddings in case of unsegmented stream does complicate learning. This network prediction could be tested in an experiment with human subjects using unsegmented speech.

3 General Discussion

Statistical learning of dependencies between elements in a sequence is an automatic process widely expoited by humans during processing of temporal structures. Earlier work showed that underlying computations are related to adjacent elements, but more recent work suggests that they also pertain to nonadjacent elements. The latter task appears to be more difficult, perhaps due to learner's bias towards adjacent transitional probabilities that could be perceptually easier to track. In addition, with nonadjacent elements the learner faces a combinatorial problem, since the number of possible nonadjacent probabilities that can be tracked grows exponentially with the length of the embedding. Therefore, it might be that remote computations can only be carried out under certain conditions.

In search for these conditions, earlier research claimed that learning nonadjacent dependencies is only possible given the availability of additional cues. Peña et al.'s conclusion was that pauses between words are necessary, Newport & Aslin [5] stated that phonological cues are required, which explained their finding why only nonadjacent segments could be learnt but not syllables. However, learning nonadjacent dependencies can occur even in a continuous stream of data without any additional cues, provided that the variability of embedding is sufficiently

large [7]. Our experiment confirms this computational capability using a sequential learning device, Elman's SRN, that only relies on the order of elements in a sequence. This also implies that the system is capable of focusing on nonadjacent regularities within the frames without having to apply higher algebraic rule-like computations as hypothesized by Peña et al. Actually, the support for ubiqitous associative learning mechanisms was also expressed in the follow-up work that convincingly questioned the line od reasoning used in Peña et al. [14].

In experiment 2 we observed qualitatively the same behavior of SRN in case of longer, two-syllable embeddings, but only if a more powerful RTRL training algorithm was substituted for standard error back-propagation. This more realistic case is very relevant, since natural languages contain remote dependencies, even with typically longer and varying span of embedding (such as A_iXYB_i and A_iXYZB_i). Therefore, suitable experiments and computational simulations should be the focus of subsequent research. In sequence learning literature, earlier work had shown that nonadjacent dependencies spanning identical embedded sequences (of 3 elements and more) are not learnt by human learners and provide an especially difficult learning problem even for large SRNs [15].

On the other hand, tracking remote dependencies requires features reminiscent of learning context-free languages (CFLs). Recurrent neural nets have been shown to have a potential to learn CFLs [16,17]. However, learning processes in these cases are studied on a higher, more abstract level, typically employing only a few symbols (such as the a^nb^n language). This shifts the processing up away from the syllable-based level that involves a considerably higher number of elements.

In summary, tracking remote dependencies is a crucial linguistic ability, whose underpinnings we are just starting to uncover. There are various questions that remain unanswered, one of them being whether adjacent and nonadjacent dependencies require separate learning processes, or the same general process can be employed under a wide range of conditions. Previous results [8,7] and our simulations suggest that the learning system may be capable of various statistical computations seeking the most reliable sources of information. This is consistent with hypotheses of the "reduction of uncertainty" [18] and the simplicity principle [19], stating that the learning system tends to choose the simplest hypothesis about the available data by seeking its invariant patterns. When transitional probabilities are high, adjacent elements are perceived as invariant. When large variability disrupts adjacent probabilities, learners will tune to alternative sources of invariance, potentially between remote elements.

References

1. Saffran, J., Newport, E., Aslin, R.: Word segmentation: The role of distributional cues. Journal of Memory and Language 35, 606–621 (1996)
2. Mintz, T., Newport, E., Bever, T.: The distributional structure of grammatical categories in speech to young children. Cognitive Science 26, 393–424 (2002)
3. Redington, M., Chater, N., Finch, S.: Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science 22, 425–470 (1998)

4. Cleeremans, A., Destrebecqz, A., Boyer, M.: Implicit learning: news from the front. Trends in Cognitive Sciences 2(10), 406–416 (1998)
5. Newport, E., Aslin, R.: Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology 48, 127–162 (2004)
6. Peña, M., Bonatti, L., Nespor, M., Mehler, J.: Signal-driven computations in speech processing. Science 298, 604–607 (2002)
7. Onnis, L., Monaghan, P., Christiansen, M., Chater, N.: Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society, pp. 1047–1052. Lawrence Erlbaum, Mahwah (2004)
8. Gómez, R.: Variability and detection of invariant structure. Psychological Science 13(5), 431–436 (2002)
9. Onnis, L., Christiansen, M., Chater, N., Gómez, R.: Reduction of uncertainty in human sequential learning: Evidence from artificial language learning. In: Proceedings of the 25th Annual Conference of the Cognitive Science Society, pp. 886–891. Lawrence Erlbaum, Mahwah (2003)
10. Elman, J.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
11. Garzón, F.: Non-adjacent transitional probabilities and the induction of grammatical reglarities. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 767–772. Lawrence Erlbaum Associates, Mahwah (2005)
12. Williams, R., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Computation 1, 270–280 (1989)
13. Doya, K.: Recurrent networks: Recurrent learning. In: The Handbook of Brain Theory and Neural Networks, pp. 796–800. MIT Press, Cambridge (1995)
14. Perruchet, P., Tyler, M., Galland, N., Peereman, R.: Learning non-adjacent dependencies: No need for algebraic-like computations. Journal of Experimental Psychology: General 133, 573–583 (2004)
15. Cleeremans, A., McClelland, J.: Learning the structure of event sequences. Journal of Experimental Psychology: General 120, 235–253 (1991)
16. Rodriguez, P., Wiles, J., Elman, J.: A recurrent neural network that learns to count. Connection Science 11(1), 5–40 (1999)
17. Bodén, J., Wiles, J.: Context-free and context-sensitive dynamics in recurrent neural networks. Connection Science 12(3/4), 197–210 (2000)
18. Gibson, E.: An Odyssey in Learning and Perception. MIT Press, Cambridge (1991)
19. Chater, N.: Reconciling simplicity and likelihood principles in perceptual organization. Psychological Review 103, 566–581 (1996)