A Sentence Generation Network that Learns Surface and Abstract Syntactic Structures

Martin Takac*, Lubica Benuskova, and Alistair Knott

Dept. of Computer Science, University of Otago
PO Box 56, Dunedin 9054, New Zealand
takac@ii.fmph.uniba.sk

Abstract. In this paper we present a connectionist model of sentence generation based on the novel idea that sentence meanings are represented in the brain as sequences of sensorimotor signals which are replayed during sentence generation. Our model can learn surface patterns in language as well as abstract word-ordering conventions. The former is achieved by a recurrent network module; the latter by a feed-forward network that learns to inhibit overt pronunciation of predicted words in certain phases of sensorimotor sequence rehearsal. Another novel element of the model is adaptive switching of control based on uncertainty (entropy) of predicted word distributions. Experiments with the model show that it can learn the syntax, morphology and semantics of a target language and generalize well to unseen meanings/sentences.

Keywords: sentence generation, language acquisition, neural network

1 Introduction

Sentence generation can be viewed as the problem of encoding a "message" as a sequence of words. A *simple recurrent network* (SRN) has proven to be particularly successful in learning sequential dependences [4]. When trained for predicting the next word in the sequence, it can implicitly form syntactic categories and learn probability distributions conditioned by grammatical rules of the target language [5]. However, pure SRN models have difficulty generalizing to patterns that were rarely or never seen during training, even though they conform to abstract grammatical rules [2, 7]. As a workaround, models were suggested that separate rules from their content (words) and learn sequences of more abstract elements, e.g. semantic roles [2], abstract word classes [8] or multi-word phrasal units [3].

In this paper, we present a model of sentence generation that combines learning surface patterns, such as idioms or fixed expressions, with learning of abstract rules. The key novel idea of the model is its representation of sentence meaning as a *sequence* of semantic representations, rather than as a static assembly of active units (Sect. 2).

* Also at Dept. of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia.

The presented model is also interesting from an architectural point of view. It consists of several modules: a content-blind control network for learning abstract rules, a context-independent vocabulary, and a SRN for learning surface patterns. These modules need to be mutually coordinated and employed in different phases of sentence generation. Control passing among the modules is driven by another neural network that is trained to use the *entropy* of the different modules (i.e. their degree of confidence in their predictions) to select which module is in control of processing.

In the rest of the paper we introduce the architecture in more detail (Sect. 3), describe an experiment exploring the model's ability to acquire different word-ordering conventions and surface patterns (Sect. 4) and present the results of this experiment (Sect. 5).

2 Meanings Represented as Sensorimotor Sequences

Declarative sentences typically describe *episodes* – events or states. We focus on concrete episodes that can be described by transitive sentences (e.g. *John kisses Mary*). The semantic structure of an episode can be modelled as a collection of thematic roles (e.g. AGENT, PATIENT, ACTION) with associated fillers. A connectionist model must employ a scheme for binding semantic objects to particular roles. The scheme we use is motivated by the embodied view on cognition, namely that high-level semantic representations of concrete episodes are delivered by sensorimotor (SM) routines. In our model, the experience of a transitive episode involves a canonical sequence of SM operations – a *deictic routine* [1] (Table 1, for an extensive body of evidence see [6]). Each operation takes place in an initial context, generates a reafferent signal and establishes a new context. We also assume that experienced episodes can be stored in working memory as prepared SM sequences that can be internally replayed. In our model, in order to express an episode verbally, a speaker needs to internally replay the episode's stored SM sequence, in a mode where the replayed signals generate linguistic side-effects. In this account, the syntactic structure of a sentence is in part a reflection of the structure of the underlying SM routine.

Table 1. The time course of signals occurring during the replay of a deictic routine 'an agent grasps a cup' from working memory.

Sustained signals	Transient signals			
	Initial context	Operation	Reafferent signal	New context
$plan_{attend\_agent,attend\_cup,grasp}$	C_1	*attend_agent*	*agent_rep*	C_2
$plan_{attend\_agent,attend\_cup,grasp}$	C_2	*attend_cup*	*cup_rep*	C_3
$plan_{attend\_agent,attend\_cup,grasp}$	C_3	*grasp*	*agent_rep*	C_4
$plan_{attend\_agent,attend\_cup,grasp}$	C_4		*cup_rep*	

3 Architecture

The complete model of language production consists of several functional modules that work together: an **episode rehearsal network**, which replays a working memory episode representation to generate a sequence of SM signals; a **word production network**, which maps individual SM signals onto word forms; a **control network**, which determines the points during episode rehearsal when these word forms should be pronounced; and a **word sequencing network** which learns surface regularities in word sequences, and several other components (Fig. 1). We will now explain each module in turn.

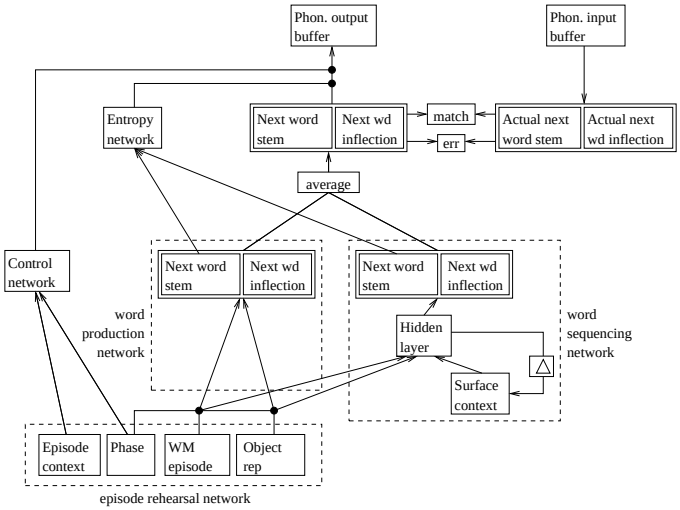

Fig. 1. The complete model of language production. Besides gating overt pronunciation, the control network and the entropy network coordinate mode switching between episode rehearsal and surface word sequencing. Localist linear neurons in the 'Next word stem' and 'Next word inflection' blocks are combined using the softmax function and represent probability distributions.

3.1 The Episode Rehearsal Network

The **episode rehearsal network** (Fig. 2) consists of four parts. The **working memory (WM) episode** area stores a prepared SM sequence (a plan). The plan is tonically active during the whole rehearsal of a particular episode, but generates transient activity in two areas ('context' and 'current object') when it is replayed. As well as supporting the rehearsal of SM sequences, WM episode representations also provide the semantics of inflected verbs in our model. They encode a planned motor action, but also planned actions of attention to the agent

and patient, which we assume can surface as agreement inflections on the verb stem. The semantics of nouns come from the 'current object' area. We model the syntactic differences between nouns and verbs using the different temporal dynamics of current object and WM episodes in the episode rehearsal network.

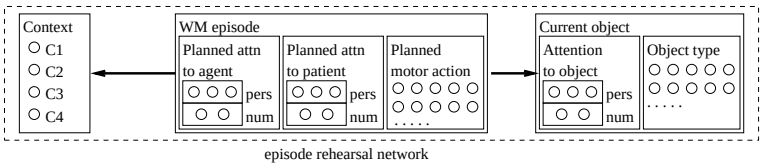

Fig. 2. The episode rehearsal network. The 'Phase' part (Fig. 1) is not shown. Each block with circles uses 1-hot localist coding.

The **current object** area holds a transient representation of the currently attended object. During the course of episode rehearsal, this area alternately holds representations of the agent and the patient. Besides a person and number information (in the same format as in the WM episode), it codes the type of the currently attended object.

The **context** and **phase** (Fig. 1) areas hold a representation of the current stage during episode rehearsal. This representation helps to drive the episode rehearsal process. In our simulation there are four possible contexts (see Table 1), each of which has two phases (see Sect. 3.2). The thick arrows in the diagram reflect the fact that the sequence of transient representations in the current object and context areas are generated by a WM episode representation.

3.2 The Word Production Network

The episode rehearsal network provides input to the **word production network** (WPN), which learns a context-independent lexicon in the form of a mapping from single concepts to single words. It also provides input to the word-sequencing network that will be described in the next section.

The WPN consists of one layer of linear perceptrons taking input from all the units in the WM episode and the current object parts of the episode rehearsal system. The input connections are gated by inhibitory links from a cyclic pattern generator (depicted as 'Phase' in Fig. 1) so that at any time input comes either wholly from the current object or wholly from the WM episode. During episode rehearsal, the pattern generator cycles through two phases in each context, providing first an opportunity to read out the current object representation (phase a), and then the tonically active WM episode (phase b) – i.e. to read out first a noun and then a verb.

The WPN is trained on the utterances of mature speakers, paired with episode representations stored in the episode rehearsal network. We are simulating an infant who experiences episodes in the world and who also hears

mature speakers talking. The mature utterances the system hears are stored in a **phonological input buffer** as sequences of target words in exactly the same format as words generated by the WPN.

3.3 The Control Network

For a given semantic input from the episode rehearsal network, a word/inflection output is computed via two independent paths – the WPN and the word-sequencing network (WSN). We envisage their joint averaged output (JAO) to be a premotor articulatory plan that may or may not be overtly pronounced, depending on the decision of other module – the **control network** that gates the connection from the joint output to the phonological output buffer.

The episode rehearsal network gated by the phase generator delivers a structured sequence of semantic signals to the WPN and WSN. For a transitive episode, the sequence is as shown below.

Context/phase	C1a	C1b	C2a	C2b	C3a	C3b	C4a
SM signal	AGENT	WM EP.	PATIENT	WM EP.	AGENT	WM EP.	PATIENT

Note that this sequence contains multiple occurrences of each concept. Different languages have different word-ordering conventions (e.g. English has the SVO – Subject Verb Object word order, while Māori has VSO); the model has to learn on which occasion the JAO should be pronounced/witheld. This is the task of the control network – a feed-forward network with one hidden layer, which is trained on the match between the predicted word (JAO) and the current word in the phonological input buffer. The desired output is binary (1 for 'pronounce' in case of match, 0 for 'withold' in the case of a mismatch).

Note that the control network is content blind in the sense it takes no input from actual semantic concepts, just from the 'context' and 'phase' parts of the episode rehearsal system. Hence it has a potential to learn abstract syntactic rules in terms of contexts/phases when the overt pronunciation should be suppressed. For example, for SVO language, pronunciation should be suppressed in all context/phases but $C1a$, $C1b$, $C2a$; for VSO in all but $C1b$, $C3a$, $C4a$.

3.4 Learning Surface Patterns

The model described so far can learn a lexicon and a set of abstract word-ordering conventions for a given language. However, languages also contain surface patterns such as idioms or fixed expressions – sequences of words that occur together with particularly high frequency and that contribute their meaning collectively rather than individually, e.g. *Winnie the Pooh*. Other surface patterns take the form of statistical tendencies, where in some context a particular word is more likely to occur than other words. Idioms violate the one-to-one correspondence between concepts and words; hence we need to extend the model with a device that can generate more than one word for a particular semantic concept present at the input. In our model, this is the **word-sequencing network** – a variant of

a SRN. Input and output-wise, it mimics the WPN (Fig. 1). Both networks are trained by the 'actual' next word replayed from the phonological input buffer. However, the WSN has a hidden layer with recurrent connections, which enables it to learn commonly occurring sequential patterns in its training data.

3.5 The Entropy Network

Since the WSN is able to produce more than one word for any given semantic input, the model needs to decide when to pass control back to the episode rehearsal network, i.e. when to deliver the next semantic signal.

Imagine the WSN is to produce an idiomatic expression – say *Winnie the Pooh*. This expression describes a single semantic signal. Given this signal, the WSN should begin by predicting the word *Winnie* with high confidence, and then, after copying back the surface context, the word *the* and then (after another copy operation) the word *Pooh*, both with high confidence. But after this point, the network can no longer be so confident. Like a regular SRN, it can at best predict the category of the following word (e.g. predicting a higher likelihood for verbs), but not a particular content word. This indicates that the episode rehearsal network should deliver the next semantic signal.

As a (inverse) measure of confidence, we use the *entropy* in the word stem part of the WSN output. If the predicted word has high entropy (many competing alternatives), it should not be overtly pronounced and the control should be passed back to the episode rehearsal network. An exact threshold for the entropy can be task dependent and can change in time, so we use an adaptive feed-forward network (called the **entropy network**) that learns the right value. It takes as its input the entropies of the WSN and WPN outputs and is trained on the same Boolean 'match' signal as the control network. Besides control passing, the output of the entropy network has a gating function similar to that of the control network, i.e. to suppress the overt pronunciation of words predicted with low confidence.

3.6 Sentence Generation in the Trained System

As already mentioned, sentence generation in our conception involves replaying a particular episode from working memory, generating a sequence of semantic signals in the episode rehearsal system, from which a sequence of words is produced in the phonological output buffer. The trained model alternates between two modes of iteration. In one mode, the episode rehearsal system is in control. This system iterates through the sequence of SM signals until it reaches a context at which the control network allows a word to be overtly pronounced. In the other mode, the WSN is in control. At each iteration, the WPN and WSN jointly predict a probability distribution for the next word given the currently active SM signal. If they can confidently predict the next word, the word is pronounced, the WSN updates its surface context layer and the model carries on in this mode until the networks can no longer confidently predict the next word.

3.7 Training the System

During training, the model alternates between the same two modes as during generation. In the first mode, episode rehearsal advances (and the control network is trained) until a context/phase is reached in which the control network gives the 'pronounce' signal. Then the network switches into the word sequencing mode. As long as the WPN and WSN predict the next word with sufficient confidence and it matches the actual word in the phonological input buffer, they keep predicting (based on a changing surface context), being trained, and advancing the phonological input buffer. If the prediction does not match or has a low confidence, the actual word stays in the phonological input buffer, the surface context is not copied and the model switches back to the episode rehearsal mode. Details of the training algorithm are given in [9].

4 Experiment

The model we have just described[1] was trained on an artificial language with an English vocabulary (105 words), morphology featuring number (Sg, Pl) inflections of nouns, number and person (1st, 2nd, 3rd) inflections on verbs, subject-verb agreement, irregular plurals (leaves, fish, teeth, women, etc.) and personal pronouns, and the SVO word order. The language consisted of 127088 transitive sentences, out of which roughly 80 % were regular transitive sentences such as *Mice bite-3pl dog-sg*, the rest contained continuous **idioms** such as *Mia-sg lick-3sg* **ice cream**-*sg* (13 %) and idioms interleaved with a noun phrase, such as *Daddy-sg* **kiss**-*3sg me* **good bye** (6.4 %).

The model was trained on a sample of 4000 randomly selected sentences paired with their meanings (sequences of semantic signals) for 25 epochs. After each training epoch, the weights were frozen and the model was tested for sentence generation on a set of 4000 previously unseen meanings. All results were averaged over 10 runs with different initial random weights of connections and different training/test samples of the target language.

To test the ability of the model to acquire all possible word-ordering conventions, we created another five target languages with the same vocabulary, morphology and similar idioms, but with different basic word-ordering (SOV, VSO, VOS, OVS, OSV) and ran 10 runs for each target language in the same way as for the SVO language.

5 Results

The control network was able to learn correct word-ordering rules with 100 % success for all the six word-orders. We also recorded the network's overall *generation accuracy*, measured as the proportion of correctly generated sentences.

[1] We conducted a preliminary study of a pure SRN (enhanced with spatially represented semantic roles) on an analogous language production task [10]. Although successful in generating certain types of unseen sentences, this network has a problem in principle – it cannot generalize across semantic roles, as argued in [2].

We considered an utterance to be correctly generated for a given meaning, if all thematic roles were expressed with semantically appropriate words, the sentence was syntactically correct (i.e. it complied with the transcription rules) and all the words had correct morphology (inflections). Averaged over all six word-orders, the models achieved 96.6 % (SD=2.7 %) accuracy on training sets and 94.1 % (SD=4.3 %) accuracy on test sets. Given that they were trained on 3 % of target sentences, the model achieved good generalisation ability.

6 Conclusion

The main goal of this paper was to introduce a novel connectionist architecture for sentence generation which is able to learn both abstract grammatical rules and surface patterns in a language. The experiments reported here show that our network can generate regular sentences but also sentences containing a variety of idiomatic surface structures. The main technical innovation which permits this is our use of sequences to represent sentence meanings (episodes). From the perspective of embodied cognition, this is helpful in connecting semantic representations to the sensorimotor system. From the perspective of syntax, it is helpful in supporting a rich model of patterns in language.

Acknowledgments. This research was supported by VEGA 1/0439/11 grant and a BuildIT postdoctoral fellowship grant for Martin Takac.

References

1. Ballard, D., Hayhoe, M., Pook, P., Rao, R.: Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 20(4), 723–767 (1997)
2. Chang, F.: Symbolically speaking: a connectionist model of sentence production. Cognitive Science 26, 609–651 (2002)
3. Dominey, P., Hoen, M., Inui, T.: A neurolinguistic model of grammatical construction processing. Journal of Cognitive Neuroscience 18(12), 2088–2107 (2006)
4. Elman, J.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
5. Elman, J.: Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7, 195–225 (1991)
6. Knott, A.: Sensorimotor cognition and natural language syntax. MIT Press, Cambridge, MA (in press)
7. Marcus, G.F.: Rethinking Eliminative Connectionism. Cognitive Psychology 37(3), 243–282 (1998)
8. Pulvermüller, F, Knoblauch, A.: Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain. Neural networks 22(2), 161–172 (2009)
9. Takac, M., Benuskova, L., Knott, A.: Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. Technical Report OUCS-2011-01, University of Otago, New Zealand (2011)
10. Takac, M., Knott, A., Benuskova, L.: Generation of idioms in a simple recurrent network architecture. Technical Report OUCS-2010-02, University of Otago, New Zealand (2010)