This paper presents a localist multimodal neural network that uses Hebbian learning to acquire one-word child language from child directed speech (CDS) comprising multi-word utterances and queries in addition to one-word utterances. The model implements cross-situational learning between linguistic words used in child directed speech, the accompanying perceptual entities, conceptual relations and inferred communicative intentions. In 90 cases out of 117, the network successfully generates one-word utterances that may be viewed as being semantically equivalent to the CDS input used to train the network. The model also successfully emulates the one-word speech of a child in 12 out of 28 cases, despite its localist nature, thereby suggesting that Hebbian learning, as used in most models of cognitive development, is capable of cross-situational learning, a key component of multimodal temporal cognitive acquisition tasks, of which child language acquisition is one.

1. Introduction

Child language acquisition is a highly complex cognitive process that, despite holding the fascination of many researchers over the years, still remains unresolved. Recently neural network models have been used as a computational means for understanding the child language acquisition process. Examples include the autoassociation model by Plunkett, Sinha, Moller and Strandsby [1], models based on Miikkulainen’s [2] Hebbian-linked self-organising maps architecture [3][4], and, more recently, the counterpropagation network model by Nyamapfene and Ahmad [5].

In this class of models of child language acquisition a neural network model of child language acquisition is trained by submitting to it, on a cycle by cycle basis, single word forms and their associated extralinguistic information. For instance, the Plunkett et al model [1] is trained by submitting to it, on a cycle by cycle basis, an image representation and its associated word label. Similarly, in the Abidi and Ahmad neural multi-net model [3] and the Nyamapfene and Ahmad counterpropagation network model [5], training proceeds by submitting
to the network, on a cycle by cycle basis, the phonetic representation of a single word taken from a child’s speech corpus alongside its corresponding perceptual entity, conceptual relation and perceived communicative intention. Consequently, the training and simulation of current neural network models is akin to a child acquiring its first language through exposure to the linguistic output of another child undergoing the same process, and then emulating that child.

However, in real life the child is most likely to acquire her speech from her caregivers who in a family setting may be the child’s parents and older siblings. In general, caregivers communicate with infants using a specially articulated version of normal speech known as “child directed speech” (CDS) or “motherese” [6][7]. Unlike the training methodology adopted by current neural network models, the CDS that infants hear consists of multi-word statements and queries in addition to single word utterances. Firstly, infants have to segment the individual words in the CDS, and this they can do by the time they are seven and a half months old [8]. Then they have to associate the individual words in the CDS stream their corresponding referents in the linguistic environment. It is this second aspect of child language acquisition which is the focus of this paper.

It has been suggested that infants may learn word-meaning associations by pairing individual spoken words with several co-occurring possible referents over a period of time and then statistically deciding on the most appropriate word-referent pair. This form of learning is termed cross-situational learning [9]. In this paper a localist multimodal neural network uses Hebbian learning [10] to acquire one-word child language from child directed speech (CDS) through cross-situational learning.

The rest of this paper is organised as follows: In the next section a discussion of the role of child directed speech to the acquisition of child language is given. This is followed by a review of two computational models that, when given multi-word input, use cross-situational learning to map individual spoken words to their respective meanings. Following this, the localist neural network model for early child language acquisition from CDS is introduced, and the data used in simulating and analysing this model is presented. A discussion of the simulation results of the model is then presented. The paper then concludes by summing up the contributions of the model to early child language research, and reveals ongoing work to further improve the model.

2. Role of Child-Directed Speech in Child Language Acquisition

Language acquisition normally takes place in the context of a rich interaction between the child and its parents [11]. Early conversations are restricted to familiar settings and to objects that are present thereby greatly
simplifying the child’s problem of learning the words for things [6]. For instance, mothers’ speech to one- and two- year olds consists of simple, grammatically correct, short sentences that deal with the child’s interests: actions, objects, people and events that are present in the ‘here’ and now' [11][14].

CDS possesses features that may help the child to segment speech into words, phrases and sentences [11]. For instance, single-word utterances are quite frequent, and words are articulated clearly and slowly with distinct pauses between sentences. In addition, mothers tend to repeat isolated phrases and words following the complete utterance. Consequently, CDS can be viewed as a highly specialized language with the necessary affective qualities to engage the child in language, and one which allows the child to remain focused on the provider of the input thereby maximizing language learning [7].

3. Related Computational Work

Siskind [12] developed a mathematical model based on cross-situational learning whose input is a series of multi-word utterances, each paired with a set of possible meanings for the utterance as a whole. Each utterance is viewed as an unordered collection of word symbols, and the model maps each word symbol in an utterance to a set of conceptual expressions that represent the meanings of different senses of that word. A set of inference rules embodying several constraints on the word-learning problem, including the constraint that the meanings of individual words in an utterance contribute non-overlapping portions to the meaning of the whole utterance, are then used over several trials to model the child’s progression to mapping individual words to their meanings. Throughout the series of trials, as the model learns the meanings of some of the words, it uses that knowledge to further constrain the possible meanings of other words in an utterance, resulting in faster learning.

When presented with an artificial corpus, Siskind’s model is able to learn a homonymous lexicon despite noisy multi-word input in the presence of referential uncertainty. However, the model’s algorithm is very complex and unpractical for empirical mother-child corpora since it can not infer, through generalisation, or otherwise, the meanings of new multi-word utterances.

Yu and Ballard [13] have recently used a machine translation model to learn word-object association probabilities in a natural corpus comprising interactions between a mother and a pre-verbal infant. In their model, they assume that word-object pairs are latent variables underlying the spoken words and extra linguistic information that constitute the corpus. By formalising the task of child language
acquisition as an expectation maximisation problem, they develop a learning algorithm that associates words and their referent objects in a manner that maximises the likelihood of the audio-visual observations in the corpus. Using word-object co-occurrence statistics, they assign initial values for word-object association probabilities, and on the E-step of the Expectation Maximisation algorithm, they compute counts for all word-object pairs. These values are then used on the M-step of the Expectation Maximisation algorithm to refine the word-object association probabilities. The E and M steps are repeated until the association probabilities converge. By weighting the word-object associations using prosodic information and joint attention information, the model demonstrated that recall improves in the presence of social cues. However, the model fails to take into account the speaker’s communicative intention as well as the conceptual relations between the referent objects and events as suggested by the literature on child language acquisition [14][15].

4. Experimental Method

Child language acquisition may be viewed as a form of ‘social convergence’ in which the child with a certain socio-cognitive capacity attempts to make sense of the contextualised language of the adults in her environment [16]. In this regard, the child learns to determine the entities the caregiver is speaking about, as well as the conceptual relations between the objects in the environment as well as the caregiver’s communicative intentions and associates these with the linguistic words uttered by the caregiver. This task is simulated by a neural network that comprises five sets of nodes, namely spoken word nodes, actor nodes, object nodes, conceptual relation nodes and communicative intention nodes which are connected through learning. Each node encodes a single entity.

The word nodes encode the caregiver’s utterances as multiword sequences. Each node encodes a word, and in our model, the nodes in a sequence are associated in accordance with temporal Hebbian links.

For each CDS utterance, the associated actor and object nodes are simultaneously activated, along with the corresponding conceptual relation and communicative intention nodes and Hebbian learning used to update the weights between activated modal nodes. Hence the extent to which individual words in each CDS utterance co-occur with agents, objects, conceptual relations and inferred communicative intentions is established in an automatic, self-organising manner in accordance with the Hebbian weight update equation [17][18]:

\[
\Delta w_{ij} = \eta \left( d_i - w_{ij} \right)
\]  

(1)
where $\Delta w_{ij}$ denotes the change in weight from unit $i$ to unit $j$, $a_i$ and $a_j$ denote the activation levels of units $i$ and $j$ respectively, and $\varepsilon$ denotes the learning rate. The term $\{a_i - w_{ij}\}$ ensures that weights do not grow without bound, thereby minimising the possibility of weight saturation. Eq.1 captures the conditional probability that a sending end node was active given that the receiving node was active \cite{17}\cite{18}, i.e.

$$w_{ij} = P(a_i|a_j)$$

Consequently, whenever a given receiving node $j$ is active, if a sending unit $i$ also tends to be active, the interconnecting weight will tend to be high. In contrast, whenever a given receiving unit is active, if a sending node tends not to be active, the interconnecting weight between the two will tend to be low. In this way, the Hebbian learning yields weights that reflect conditional probabilities of activities, and in turn yield interconnecting weights that represent correlations in the environment.

5. Data

The simulation described in this paper uses the child language acquisition data in the Bloom 1973 corpus \cite{15}. This corpus is found in the Child Language Data Exchange System (CHILDES) corpora \cite{19}.

5.1. Child-Directed Speech (CDS) Training Data

The data set for training the network model and assessing its ability to generate one-word child language utterances is taken from the earliest sample in the Bloom 1973 corpus, i.e. the sample taken at age 1 year 4 months and 21 days. This dataset comprises 195 utterances directed at Alison by her mother, the perceptual entities underpinning each utterance (as identified from the utterances and their accompanying annotations), the conceptual relationships between the entities, and the mother’s communicative intentions as inferred from the discourse. The perceptual entities associated with each utterance were categorised into Actors and Objects based on the roles they play in the conceptual relationship between them. In this paper the term extralinguistic information is used to collectively refer to the communicative intention, conceptual relation, actors, and objects associated with a CDS utterance.

During network training CDS utterances and their associated extralinguistic information are simultaneously applied to the network on a cycle-by-cycle basis. A learning rate small enough to guarantee the convergence of the network
weights to stable values is selected and the CDS utterances are applied until there is no change in the value of the weights. In the work reported in this paper a learning rate $\varepsilon = 0.01$ is used to train the network over 500 cycles.

To assess the trained network’s ability to make one-word utterances, 117 distinct extralinguistic terms identified from the CDS training dataset are used. When an extralinguistic term is applied to the network, each word node computes individual activations for conceptual relations, communicative intentions, actors and objects. The resultant word activation is the product of these four individual modal activations, and the word with the highest activation product is deemed the winner.

5.2. One-Word Stage Child Language Test Data

According to Bloom [15], children at the one-word stage use single word utterances to talk about the conceptual relations between perceptual entities. 28 utterances made by the child Alison in the Bloom 1973 corpus [15] were identified, and for each utterance, the extralinguistic information encoding the perceived conceptual relation is applied to the trained model. In each case, the word node with the highest activation is deemed to represent the model’s linguistic response to the input.

6. Results and Discussion

This section presents and discusses the results of the network in simulating one-word child language on the basis of extra-linguistic information derived from motherese utterances as well as the network’s ability to emulate the one-word utterances taken from Alison’s utterances in the corpus.

6.1. One-Word Utterances from CDS Extra-Linguistic Data

Table 1 lists examples of the one-word responses made by the trained network when prompted by extralinguistic information from the CDS training dataset.

Of the 33 single word CDS utterances in the training data, the model successfully recalled 32 of the utterances when prompted with the corresponding extralinguistic information. For instance, for the single word CDS input “up”, the model correctly responds with the output “up”.

In the failed response, the model responded with “fell” instead of “down”. The concept of “down” had been trained using the situation when the doll fell down from the chair. In this situation, the CDS sequence is as follows:

1. Hey, look!
When presented with the extralinguistic information for the word “down” in line 4, the model responds with the word “fell”. This is possibly due to the fact that both “fell” and “down” occur in the same situation, and other different situations would be needed to enable the model to correctly distinguish between them. The fact that the model successfully learns from single word utterances would suggest their importance as learning aids in child language acquisition, where it has been noted that single word utterances are quite dominant in CDS [20].

The remaining 84 extra-linguistic inputs with multi-word CDS utterances, the network manages to generate 57 one-word utterances that may be regarded as equivalent, in a semantic sense, to the associated multi-word CDS expressions. For the remaining 27 extra-linguistic inputs the network generates single-word utterances that are difficult to classify as having the same meaning as the associated CDS multi-word utterances.

An example of when the network gives a linguistic response that can be regarded as appropriate is when the network simulates Alison’s response to extra-linguistic response associated with CDS utterances made as Alison and her mother covered a jar with its lid. In this case, the extra-linguistic information presented to the network is as follows: communicative intention – comment, conceptual relation – lid covers jar, actors – Alison and mom, objects – lid and jar. The network model responds to this extra-linguistic information with the single word ‘cover’ - i.e. the word ‘cover’ gives the highest output activation to this extra-linguistic information. Although Alison never uses the word “cover” in
her one-word utterances in the corpus, the word “cover” seems to be an appropriate one-word equivalent for the corresponding multi-word CDS expression: ‘OK, let’s cover it up and put it away.’

Fig. 1 shows all the words in the model that responded with non-zero activation to the extralinguistic information comprising: communicative intention – *comment*, conceptual relation – *lid covers jar*, actors – *Alison and mom*, objects – *lid and jar*.

An example of when the network gives a one-word utterance that may be regarded as inappropriate is when extralinguistic information associated with CDS utterance “there are no more cookies” is presented to the network. The extra-linguistic information for this CDS utterance is: communicative intention – *comment*, conceptual relation – *object disappearance*, actors – *cookies*, objects – *none*. The network model responds to this extralinguistic information with the single word ‘are’ - i.e. the word ‘are’ gives the highest word output activation to this extra-linguistic information.

Fig. 2 shows all the words in the model that responded with non-zero activation to the extralinguistic information comprising: communicative intention
From Alison’s one-word utterances, the correct response should have been ‘gone’, which is the second highest activation after ‘are’. Also, the network responds with ‘gone’ when presented with extra-linguistic information for similar circumstances such as the disappearance of bubbles, and the disappearance of the microphone. This may be because the word ‘gone’ appears explicitly in the CDS expressions for the disappearance of bubbles and the disappearance of the microphone, which is not the case with the disappearance of the cookies where the expression uttered by Alison’s mother is ‘there are no more cookies’. This situation could possibly be remedied by incorporating inhibitory links between word nodes to ensure that inappropriate words are inhibited. Nevertheless, these results do suggest that the Hebbian weight update algorithm does try to associate applied extralinguistic information with the most appropriate one-word utterance, even in situations when such associations do not appear explicitly in the motherese used to train the model.

Figure 2. Activation plot for all the words that responded with non-zero activation to the extralinguistic information comprising: communicative intention – comment, conceptual relation – object disappearance, actors – cookies, objects – none.
6.2. *Emulating Alison’s One-Word Child Language*

The output of the network, in response the extra-linguistic terms taken from Alison’s utterances, falls into four categories when compared to Alison’s utterances, as shown in Table 2:

<table>
<thead>
<tr>
<th>Category</th>
<th>Exact Matches</th>
<th>Equivalent Matches</th>
<th>Unrelated Matches</th>
<th>No Output generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Utterances</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

There are six exact matches between the network’s output and Alison’s one-word utterances. In addition, the network output may also be considered to be equivalent to Alison’s one-word utterances in another six of the data items. For example, the network uses the word ‘eat’ instead of ‘cookie’ to request a cookie, and the word ‘fell’ instead of ‘down’ to refer to the event ‘pig falls.’ As we have discussed in section 6.1, this may be due to the word used by Alison being unavailable in the CDS used to train the network. However, the network is able to associate an appropriate word in the CDS utterance to the corresponding extralinguistic information.

In another eight of the data items, the network generated outputs which bear no discernible relationship to Alison’s utterance in terms of semantic meaning. For instance, the network generated the word ‘little’ to name a chair and the word ‘are’ to refer to the disappearance of cookies. As discussed in section 6.1, this may be due to the word used by Alison being unavailable in the CDS used to train the network. In this instance, however, an inappropriate word in the CDS utterance is statistically matched by the Hebbian training algorithm to the corresponding extralinguistic information.

In the remaining eight utterances the network was unable to output a one-word utterance in response to the inputted extralinguistic information. In all these cases, the particular combination of perceptual entities, conceptual relation and inferred communicative intention in Alison’s utterances had not been used at all in the CDS used to train the network. In these instances, the network failed to generalise to some appropriate linguistic output. This, in my opinion, is due to the localist scheme used in constructing the model whereby a single entity is represented by a single node. A distributed representation where data items are encoded as vectors of the features making them up would enable the network to generalise to data items used during training.
7. Conclusion and Future Work

As noted in [21], most cognitive processes, including language acquisition, "generally involve an interplay between a number of sources of information...each aspect of the information in the situation can act on other aspects, simultaneously influencing other aspects and being influenced by them.” Incorporating multimodal processing in neural network models of cognitive processing, as has been done in this paper, may therefore help to make them more biologically plausible. In addition, the model presented in this paper is arguably one of the first Hebbian implementations of cross-situational learning. With cross-situational learning increasingly being viewed as an important cognitive learning mechanism, this model therefore presents some justification for the predominance of Hebbian learning in models of cognitive development [18].

As has been pointed out, however, the model presented in this paper exhibits some shortcomings as a consequence of its localist nature. An equivalent model based on distributed representation is currently being investigated. In addition, the progression of child language acquisition from one-word utterances to two-word utterances is also being investigated by incorporating multimodal temporal processing into the model as suggested in [22].

References

4. P. Li, I. Farkas and B. MacWhinney, Neural Networks 17, 1345 (2004).