WWN: Language Acquisition and Generalization using Association

By

Kajal Miyan

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Computer Science

2011

Abstract

WWN: Language Acquisition and Generalization using Association

By

Kajal Miyan

Based on some recent advances in understanding and modeling cortical processing for space [26] and time [55], we propose a developmental, general-purpose model for language acquisition using multiple motor areas. The thesis presents two main ideas: a) early language acquisition is a grounded and incremental process, i.e., the network learns as it performs in the real world b) language is a complex perceptual, cognitive and motor skill that can be acquired through associative learning and skill transfer principles described in [57]. The network architecture is informed by the existing neuroanatomic studies and the associative learning literature in psychology. Through the ventral pathway, the "what" motor learns, abstracts and feeds back (as recurrent top-down context) information that is related to the meaning of the text. Via the dorsal pathway, the "where/how" motor learns, abstracts and feeds back (as top-down context) information that relates to the spatial information of text, e.g., where is the text on a page. This is a major departure from the traditional symbolic and connectionist approaches to natural language processing (NLP) — the nature of the motor areas, i.e., actions or abstract meanings, play the role of "state hubs" in language acquisition and understanding. The "hubs" correspond to multiple concepts that form the state of the current context. As any human communicable concept can be either verbally stated (what) or demonstrated through actions (how), this model seems to be the first general purpose developmental model for general language acquisition, although the size of our experiments is still limited. Furthermore, unlike traditional NLP approaches, syntax is a special

case of actions. The major novelty in our language acquisition is the ability to generalize, going beyond a probability framework, by simulating the primary, secondary and higher order associations observed in animal learning through the generalization of internal distributed representations. A basic architecture that enables such a generalization is the overall distributed representation: not only a retina image but also an array of muscles is considered high-dimensional images. An emergent internal distributed representation is critical for going beyond experience to enable three types of generalization: member-to-class, subclass-to-superclass, member-to-member, and relation-specification. In our cortex inspired model, syntax and semantics are not treated differently, but as emergent behaviors that arise from grounded real-time experience.

Dedication

To my parents, Vijay Lakshmi and Brijendra.

And my sis, Prachi.

Acknowledgements

I would like to thank my advisor, Professor Juyang (John) Weng, for his insightful guidance throughout this thesis. I thank my labmates Matthew Luciw for his helpful discussions, and Mojtaba Solgi for his generosity whenever I needed help. I would like to sincerely thank my friends, John Felix Mahendran Issac and Dominic Jacob, for the valuable support they extended in a number of ways.

Lastly, I offer my regards to my family and friends who supported me in every respect during the completion of this project.

.

Have the courage to follow your heart and intuition. They somehow already know what you truly want to become. Everything else is secondary.

Steve Jobs

TABLE OF CONTENTS

LIST OF TABLES **ix**

LIST OF FIGURES **x**

1 Introduction **1**

2 Objective and Importance **5**
2.1 Challenges in NLP: Language Processing vs. Language Acquisition . 5
2.2 WWN: An Incremental Autonomous Language Learner 7

3 Literature Survey on Language Learning **9**
3.1 Language Schools of Thought 9
3.2 Computational Linguistics Models 10
3.3 Symbolic AI 11
3.4 Language Hierarchy 11
3.5 Neural nets 12
3.6 Language Grounded in Real World 13
3.7 Autonomous Mental development (AMD) model 13

4 Psychological grounding of the Model **15**
4.1 Early Language Acquisition 15
4.1.1 Association 16
4.1.2 Generalization 17
4.2 Written language 18

5 Network Architecture **19**
5.1 Neural Pathways 20
5.2 Area 20
5.3 Architecture 22
5.3.1 Creating new sentences through Generalization 26
5.4 Derivation of Formulations 27

6 Comparison with Finite Automata **29**

7 Machine Reasoning and Logic vs Human Reasoning **34**
7.1 Inductive Reasoning 36

8 Concept and Theme **37**
8.1 Language Association to Assist Generalization 37
8.1.1 Parroting 39
8.1.2 Member to Class Generalization: Association Aided by Associative Reasoning 39

8.1.3 Subclass to Superclass Generalization 41
8.1.4 Member to Member Generalization and Classification From Similarity 42
8.2 Attention Allows Generalization 45
8.2.1 Grounding leads to generalization 45
8.2.2 Specific Relations: Attention Makes Prediction Possible 47
8.3 Word Representation as Sensory Vector 50

9 Experiments 52
9.1 Data 52
9.2 Experiments and Results 54
9.2.1 Parroting 54
9.2.2 Canonical Word Representation in Sensory Vector 55
9.2.3 Member to Class Generalization 56
9.2.4 Member to Member Generalization 59
9.2.5 Subclass to Superclass Generalization 60
9.2.6 Classification from Similarity 62
9.3 Visualization of Layers X, Y and Z 66

10 Novelty and Contributions 71

11 Discussion 73
11.1 Language Processing Based On Grammar 73
11.2 Neural networks for language processing 74
11.3 Language Acquisition 75
11.4 Future work 76

12 Conclusion 78

BIBLIOGRAPHY 80

LIST OF TABLES

9.1 Sequence of words learnt with the passage of time 55

9.2 Sequence of words learnt with the passage of time, with bold words . 55

9.3 Sequence of words learnt with the passage of time, with italicized words 55

9.4 Training Sentence 1. Member to Class Generalization 59

9.5 Training Sentence 2. Member to Class Generalization 59

9.6 Training Sentence 1. Member to Member Generalization 60

9.7 Training Sentence 2. Member to Member Generalization 60

9.8 Training Sentence 3. Member to Member Generalization 60

9.9 Testing Sentence. Member to Member Generalization 61

9.10 Training Sentence 1. Subclass to Superclass Generalization 61

9.11 Training Sentence 2. Subclass to Superclass Generalization 61

9.12 Testing Sentence. Subclass to Superclass Generalization 62

9.13 Training Sentence 1. Classification from Similarity 62

9.14 Training Sentence 2. Classification from Similarity 63

9.15 Training Sentence 3. Classification from Similarity 63

9.16 Training Sentence 4. Classification from Similarity 63

9.17 Training Sentence 5. Classification from Similarity 64

9.18 Training Sentence 6. Classification from Similarity 64

9.19 Testing Sentence 1. Classification from Similarity 64

9.20 Testing Sentence 2. Classification from Similarity 64

LIST OF FIGURES

5.1 Architecture of the WWN network 21

5.2 WWN network demonstrating generalization 27

6.1 Finite Automata learning sentences 30

6.2 Comparing the network to Finite Automata 31

8.1 External network notation of WWN for classical learning 38

8.2 WWN network model for classical conditioning 38

8.3 Member to class generalization 39

8.4 External network notation of WWN for member to class generalization 40

8.5 WWN network demonstrating member to class generalization 40

8.6 External network notation for subclass to superclass generalization . . 41

8.7 Subclass to superclass generalization 42

8.8 WWN network demonstrating subclass to superclass generalization . 43

8.9 External network notation of WWN for member similarity 43

8.10 Member to member generalization 44

8.11 Member to member generalization in WWN network 44

8.12 Concept of grounding through the example of Peekaboo Unicorn . . . 46

8.13 Imitating classical conditioning 49

8.14 External network notation of WWN for imitating classical conditioning 49

9.1 Error rate of recognizing sequences 56

9.2 Error rate for all output states 57

9.3 Time Result for experiments 6 58

9.4 Results of experiments 2 to 5 65

9.5 Layers z_1 and z_2 66

9.6 Learning Sentence 1 67

9.7 Visualization for Pre-response vector 68

9.8 Learning Sentence 2 69

9.9 Learning Sentence 3 70

Chapter 1

Introduction

Humans often think in languages. It is known that languages assist in the mental processes of perception, cognition, behavior, thoughts and intelligence in humans. Although some animals are known to be able to communicate in certain ways yet none of them are known to have a rich, detailed, specific and complex symbolic system expression as humans do. Language, thus, indicates that humans have certain special cognitive abilities that help them to not only express themselves in a more unambiguous fashion but also develop into a powerful social community. Much of our adult intellect is conveyed, stored and enhanced through natural languages. Natural language not only consists of sounds, symbols, syntax and semantics peculiar to human communication, but more importantly, inside the brain it corresponds to brain-organized traces of sensorimotor experience grounded in the physical world, this could include seeing an object or hearing it or hearing about it or reading about it or reading descriptions about it etc. [4]. This perspective is supported by modern studies of language acquisition in developmental psychology. Hence, motor activities play both a preparation and an operational role in language acquisition [25]. It has also been shown that infants appear to use visual and auditory associations inherent in social contexts to learn native-language phonetic categories [56]. However, after 50 years of extensive research in the field of Natural Language Processing (NLP) in

Artificial Intelligence (AI), few efforts have been spent on simulating how an agent *acquires* a native language from real-world interactive experience, in fact much work in the traditional NLP community considers a language as a static set of symbols with handcrafted atomic (unbreakable) meanings, syntactic and semantic rules.

In contrast, early acquisition of language in children, seldom includes any explicit training of language rules. Children learn to pay attention to the desirable contexts and carry out an unrehearsed conversations with the co-participation of other motors [43]. Despite the lack of explicit syntactic rules, children acquire language skills interactively in ways similar to the acquisition of skills for other sensing modalities; language comes naturally to humans yet such skills are not genetically coded. In his 'Essay Concerning Human Understanding' John Locke introduces the logic of empiricism, he argues that language and ideas are not completely innate but develop from sensorimotor experience and the experience of reasoning [35].

Language acquisition is different from language processing. While the latter focusses on how language can be processed, which might include division of the sentences into grammatical chunks or tagging words as nouns, pronouns or adjectives etc., language acquisition is the grounded way of language processing. It allows the system to develop a deeper understanding of the subject as it more closely bound with the surroundings of the learner. Language acquisition takes inspiration from the way humans learn language through mere interactions in the beginning and later refining through teaching and experience enabling the learner to be able to pick up any language, native or non-native and to be able to effectively and easily communicate with others who are familiar with the same language. Language processing has been traditionally used many a times in order to solve several linguistic problems and though it been effective yet there is also a great scope for improvement as we will see in later sections this thesis explores and simulates language acquisition as an alternative to the traditional methods of learning and understanding language.

In the work reported in this thesis, we use a simplified cortex-like model to simulate the process of language acquisition via incremental, interactive, sensorimotor interactions. The network is able to display the linguistic abilities of a 5-6 year old child. Traditionally probabilistic or stochastic methods are used to determine the word to word relationship that helps create various phases and sentences. In our study we take a connectionists approach to not only learn to create relationships between the words occurring together but we mainly focus on creating new sentences from the knowledge that the system gathers on its own from an external source, which could be the environment or a supervisor, thus the term *interactive learning.* The network also learns as it performs or lives thus learning *incrementally.* But it will not be very efficient if one had to spoon-feed a child every information possible, there are time when based on the environment the child forms his/her own belief system. At the age of 5-10 this is a direct result of another major human brain capabilities, namely generalization, thinking which in turn are a result of or are facilitated by *secondary association.* Traditional linguistics does not use animal learning or primary and secondary association postulates in order to bind words/phrases/concepts together. We try to explain all the above brain's linguistic capabilities with the help of secondary associations that can be converted into primary associations through practice.

Our network focuses on learning via reading hence the network uses visual "where" and "what" pathways. The "what" pathway focuses on the word or the meaning of the word being read while the "where" pathway helps the system find the location of the word on the page, i.e., whether it is on the top of the page or the bottom or is it highlighted in some way. For tractability at this stage of system development, we use insulated words as distributed inputs (patterns) and use distributed motor outputs (also patterns). The model is not formally taught any grammatical rules governing a language but learns implicit rules on-the-fly though sensorimotor examples. It is also able to create new sentences using its learned reasoning and generalization capabilities. Theoretically the knowledge database of the system is infinite and there is no restriction on the amount of data the system can accumulate on its own.

Chapter 2

Objective and Importance

2.1 Challenges in NLP: Language Processing vs. Language Acquisition

The biggest challenge for the NLP community today is to connect concepts in complex sentences. This requires a deep understanding of not just the words, through surface grammatical structure, but also the semantic knowledge of the real world experiences. Though the types of grammars and the languages, described in the previous chapter, help us to study languages in a more systematic manner yet these formal languages are very structured and have very strict set of rules restricting improvisation, which is antecedent to human communication. Humans do not communicate within the bounds of rules nor have to have the knowledge of grammar in order to express effectively. It must be noted that natural languages are not formal languages. Though natural languages have grammatical rules to help formalization of languages yet humans do not parse every word or phrase to see which grammatical rule it fits to derive the context. Also the primary understanding of languages in humans does not develop through any formal schooling but is picked up through interaction and teaching of the parents and guardians. Humans do not follow a certain school of language, in fact language acquisition does not focus on syntax at all, humans pick up languages without knowing the grammar or rules behind it. Some children might pick up or acquire a second language in their childhood that has an entirely different grammatical structure than their first learnt language never realizing the actual grammatical difference or confusing between the rules for the two. An explicitly stated grammar is clearly not required for learning language.

Computer linguistics follows two approaches to solving the problems of natural language learning: *symbolic* and *stochastic*. Symbolic approach includes N-gram methods, inductive learning or finite state machine. The common factor in all the above approaches is the fact that all of them treat words only as symbols. The output is hand-designed and pre-determined. The stochastic methods are a relatively new approach in NLP. They are statistical antecedents of the symbolic methods and include Hidden Markov Models and Maximum Entropy modeling. None of these methods are *neuromorphic* but they do have a history of being successfully used for solving many problems. These methods are capable of representing infinite number of sequences and combinations of words but they have to be painstakingly designed based on a grammar or "syntax". Syntax is grammar, and grammar, as argued above, is restricting as it follows rules and unless the language abides rules, it is not considered correct.

The natural language processors involved in tagging or meaning analysis etc. usually try to break down the problem into smaller bits and pieces therefore relying more on text/speech segmentation or parse trees to try to understand the meaning or the correctness of the sentences. They thus try to divide the problem into more formal sub-parts and hope that those do not contain any ambiguities. Natural languages and their grammar, on the other hand, are not perfect they have anomalies and irregularities. Due to which though they are effective in theory yet they fail to do as well in practice. For e.g. anaphora like, *We gave the monkeys the bananas* *because they were hungry*, it is impossible for a computational linguistic system to identify clearly what does *they* here stand for. Another examples being similes, like, *as brave as a lion* or other figures of speech. On the contrary humans can easily identify that *they* stands for *monkeys*, as bananas cannot be hungry. They are able to understand syntactically wrong, semantically ambiguous, imperfect, grammatically incorrect sentences. Humans do not have to be grammar or language gurus to understand what is being said, in fact with ample experience they can even understand the meaning of unknown similes or expressions by merely following the context. Thus, as far as NLP is concerned, syntax can only take you so far but to be able to communicate like humans and to understand the real hidden meaning of what is being said. The system should have three basic capabilities "semantics", "grounding" and "experience". Our model makes use of the above to create a new approach for language processing that takes inspiration from humans to "acquire" language.

2.2 WWN: An Incremental Autonomous Language Learner

Unlike NLP, where a human programmer entrusted with the task of designing the system handcrafts each state and the outcome of a transition, in language acquisition, the system learns to device these transitions on its own by learning them autonomously from its surrounding environment that might or might not have a human teacher. Our method uses the latter to develop an autonomous language learner, it is unique in the sense that it is the first where-what network for language acquisition that takes visual word input in order to produce the correct action, which might include various language processing tasks, like, part-of-speech tagging, text segmentation recognizing syntactic ambiguity etc. The network need not learn everything before it starts performing, but instead should learn dynamically so that it can be corrected early if it learns some wrong information. It is all the more important as the network is not taught everything explicitly but instead draws associations and conclusions from what it has learnt so it becomes imperative that if we come across any wrong information learnt by the system, we correct it, just as small children are corrected by their teachers/parents if they say/do something wrong. Our network is an *incremental* learner that *learns as it goes* focusing mainly on language *understanding*. Hence unlike other systems that have to be trained or programmed before they can do anything, our network not only learns what it is taught but also learns as it is taught. The network can incrementally pick-up new tricks as it lives on and so it *grows stronger as it lives longer*.

The other major novelty of the system is the use of animal learning concepts like *classical conditioning* to develop links between words and corresponding concepts and properties aiding the system in the process of *reasoning*. It should be noted that early language learning is a skill and is acquired through sensorimotor interactions with the environment and hence skill transfer principles should work for transferring language skills just like any other motor related ones. This study tries to apply the above concepts to language learning. It uses classical conditioning to form links and associations between the various concepts of the world, formed through *generalization* as described later, without formally or explicitly creating the concept and distinction of “class”, “object”, “subclass” or “superclass” amongst the words. The study also presents generalization as an important technique very peculiar of human brain, helping in classification and more logical arrangement of knowledge.

Chapter 3

Literature Survey on Language Learning

Natural language understanding is an AI complete problem, it is not only hard to solve but if solved we will be also solving the central AI problem of creating a machine that passes the Turing test. Linguists have been trying to codify human language for ages. Several different fields of study, ranging from psychology to biology to neuroscience have studied the evolution and development of languages. There are two major schools of thoughts when it comes to the theory of learning the first language.

3.1 Language Schools of Thought

Nativists, led by Noam Chomsky believe that language is a human instinct, children learn languages without conscious effort, before they are aware of reward or punishment, or even before they can be formally taught. Chomsky argued that language was innate and the underlying principles of language were universal and inborn to all humans. He called it the *Universal Grammar* (UG) [8]. According to Chomsky, UG contains an "initial state" of the human language faculty, prior to any linguistic experience. Smolensky's Optimality Theory (OT) belonged to the same school of thought. OT filters out the structures that don't follow the universal rules from the input such that only conforming grammatical structures remain in the language. Chomsky-Schutzenberger, defined four kinds of grammars that result in formation of four different kinds of languages: general rewrite, context-sensitive grammar, context-free grammar and finite state. Together these grammars contain rules that can define any language known.

Non-nativists or "emergencionists" supported by the likes of Piaget, Mac Whinney, Bates and Snow, however opposed the idea of prior knowledge, or preference of a certain precursory or antecedent affinity.

3.2 Computational Linguistics Models

Today with electronic media becoming more and more popular, humans see a growing need of interaction with the technology in a more simpler and humane way this gave rise to a new field *Computational linguistics*, the latest study of language as seen in conjunction with computational use of it. Several computational linguistics models have been developed. The three main types of models are:

1. Symbolic AI: Handcrafted representation.

2. Neural nets: Emergent representation but weak with prior models.

3. Autonomous Mental development (AMD) model: Emergent representation but can reason with the help of Brain-mind model.

3.3 Symbolic AI

The main focus of computational linguists till of late was on Natural Language Processing (NLP) that for a long time used the grammars described later to solve problems like speech segmentation, text segmentation, parts of speech tagging, parsing and information retrieval etc. Soar [30], ACT-R [41], CYC [33] etc. are well know NLP models map symbols to symbols through handcrafted pathways. Finite Automata (FA) that we later compare to our model, is also one such method, along with its probability variants Markov Decision Processes (MDP), Partially Observable MDP (POMDP) and Bayesian nets. But these models are handcrafted and cannot evolve on their own. They are very restrictive and though all linguistic concepts can be modeled into states they have to be painstakingly designed by human programmers.

3.4 Language Hierarchy

Chomsky-Schutzenberger, in 1956, defined a containment hierarchy of classes of four kinds of formal grammars. It is called "the hierarchy of languages" because each successive type is a subset of the other. Type 1 *general rewrite*, that results in the formation of *unrestricted language* of the form $\alpha \rightarrow \beta$ that has no size or rule restriction and is the largest class that can be recognized by Turing machines. Type 2 or *context-sensitive grammar*, forms *context-sensitive language* that are infinite with rules like $\alpha A \beta \rightarrow \alpha \gamma \beta$, where α and β can be empty, A is non-terminal while γ could be either terminal or non-terminal, these can be recognized by linear bound automata. Type 3 or *context-free grammar*, generates *context-free language* with single non-terminal on the left and a string of terminal or non-terminals on the right of the form $A \rightarrow \gamma$. This is a very rich category of languages. This category includes all the programming languages. This kind of language can be recognized by non-deterministic pushdown automata. The final type of grammar is Type 4 or *finite state* that generates *regular language*. These languages follow the rule like, $A \rightarrow a$ and $A \rightarrow aB$ with a single nonterminal on the left-hand side and a single terminal, possibly followed (or preceded, but not both) by a single nonterminal on the right-hand side. These can be decided by finite state automata and obtained by regular expressions.

3.5 Neural nets

Neural networks or connectionist approach, developed in 1980s, attempts to model mental and psychological behaviors using networks with numeric, distributed internal representations. These networks have also been used to model distributed language representation. Unlike symbolic approaches, representations in such networks are emergent. These models have two main motivations. First, there is a need for parallel processing of knowledge from multiple sources in a systematic way without specifying or knowing which input component represents what meaning. Second, since the model is not symbolic the representation itself has a potential to tolerate noisy inputs, irregularities and "fuzziness" of real natural language.

Hinton, 1981, published some seminal work on distributed semantic representations [23]. Rumelhart and McClelland, 1986, [47] used distributed representation and semantic microfeatures to address the problem of case role assignment. Other early related studies that use networks include Hanson and Kegl [5] syntactic parsing, Allen [1] on question and answering, Sharkey [48] on prepositional attachment, Lange and Dyer [32] on inference, Smolensky [49] on variable binding. More recently, recurrent neural networks like Elman network [13] and Jordan network [9] use temporal states in models with context units. ARTMAP [42] was based on the concept of similarity measure for symbolic objects and can assign class labels to the objects.

These methods are weak mainly because neural networks are weak in generalization and reasoning.

3.6 Language Grounded in Real World

Along with all this, in recent times a lot of work has been done in the direction of binding *Physical grounding* with language. Studies and models like that of Zwaan et al. [58] and Roy and Mavridis [37] contributed to the understanding of grounded acquisition of language skill. Weng et al. [55] recently developed a cortex-like temporal processing model for incremental learning of text-motor behaviors for natural language.

This work is unique as it models a process of language acquisition using both the dorsal (where/how) and ventral (what) pathways so that words of the language not only have their meaning in terms of "what", but also in terms of "where/how". This is the first general-purpose model that is capable of dealing with multiple motor areas, including visual and auditory, for language processing. It shows how behaviors within a motor area and between different motor areas are integrated in contrast to the architecture with behavior-based robots [3] where a separate behavior arbitration module is used to determine the priority of inconsistent behaviors from different behavior modules, the behavior integration in our model is tightly integrated into the network itself.

3.7 Autonomous Mental development (AMD) model

Models like Multi Layer In-place Learning (MILN) [51], Where-what network 1(WWN-1) [27], WWN-2 [28], WWN-3 [36] and Brain-mind Model [53] belong to this category. These are also emergent models but can reason well as shown later in the thesis. [54] demonstrates the power of complex text processing using the framework of a general-purpose developmental spatiotemporal agent called Temporal Context Machines (TCM), demonstrating its power of forming online, active, abstract, temporal contexts.

Chapter 4

Psychological grounding of the Model

Language is a complex means of human expression having varied components including lexical-semantics, phonemes, grammar and prosody, just to name a few. Acquiring language includes acquiring and developing the above mentioned skills along with many others, but it all starts with the child starting to associate words through imitation and generalizing and forming informal concept categories as he/she gains more experience. In order to model language learning in humans we take inspiration from the study of psycholinguistics. We focus on modeling the above two phases of early language acquisition.

4.1 Early Language Acquisition

Piaget's early work on cognition emphasized the role of active experience in development of increasingly sophisticated mental structures for early language acquisition [15]. Humans learn natural language in the same fundamental manner as every other acquired skill, through active repetition. In several ancient cultures overt repetition was used to impart knowledge of scriptures to children, including India were children from the age of 5 would start reciting the Vedas aloud in order to memorize them and is termed *audiolingualism.* This could be accorded for the preservation of the more than 8000 languages in the world that do not have scripts. Even more strikingly, this strategy of language acquisition is not species specific; animals too learn all their life skills in a similar manner including being trained to respond to human words.

4.1.1 Association

The basic representation form of the early language is speech, as infants are introduced to language through listening and producing speech. According to [20], a language can be characterized as a continuous sequence of sounds forming structures to which our ears after a certain time get accustomed to and develop a certain amount of probability as to what word should/would follow a certain group of word, thus forming a structure that is not explicitly taught to a person but is slowly acquired as it listens to more examples or is taught and corrected by the teacher. This is called “association” of phrases and words. For e.g., after certain real life experience the sentence *Baby eats food* makes sense but *Newspaper eats house* does not because *Newspaper* is never associated with *house* through *eats.* This can further be modified to include the learner learning to associate/connect a word/phrase with another word/phrase just like an animal makes a connection between a neutral stimulus and a second rewarding/punishing stimulus based simply on the fact that they occurred together. This is called “associative learning” which is also a concept of classical conditioning as will be described in length later.

4.1.2 Generalization

As the children come across more words they become more familiar with the objects they start grouping similar objects and form certain notions about them. *Cognitive generalization* is the ability to apply and test concepts and classification criteria across a range of contexts and environments. For e.g. categorizing a *brown colored, Golden Retriever* and a *white colored, one-eyed St. Bernard* in *dog* category. Nathan Stemmer in [50] introduces this very powerful capability that the children apply while learning languages as a particular class of stimulus generalization in which the generalization process occurs through the semantic characteristics of the stimuli also known as "semantic generalization". Gomez [19] found that infants can generalize when they are presented with different samples generated by the same formal system thus being able to discern the structure if given sufficient evidence to support it. Generalizing concepts follows a U-shaped learning curve in children. Starting from the age of 2, taking a dip at the age of 5 when they start overgeneralizing and making errors. But then around the age of 10 they start learning the concept well enough to be able to use it correctly.

But stimulus generalization is not sufficient, as it must be combined with correct discrimination [6]. Discrimination involves the organism's ability to detect differences among stimuli and respond correctly to a specific stimulus. It should be noted that children also create their own concepts about the objects, e.g., they know *birds can fly* and if told that *penguin is a bird* they will think that *penguins can fly* unless and until corrected. This is another reason why the model must support dynamic learning. Both generalization and discrimination together result in the complete knowledge of an object.

4.2 Written language

Of all the languages in the world there are very few that have a written form yet most of the literature survives as written text. A lot of important knowledge has been documented and stored as books, encyclopedias and scriptures. The biggest knowledge bank, the whole world wide web is in the written form floating around in the internet. The first forms of written languages were pictorial but as language slowly evolved they became more and more conforming to the phonological word representations. The words are written as they are pronounced. Thus making it easier for children reading the text to be able to correlate the text with the phonological sounds. Phonological sounds or speech, however, are the basic representation of language because that is how infants are first introduced to languages. Hence it helps children to understand the written word better if they can read it aloud to convert the written codes into phonological representations. This leads to reading and though in the beginning children associate written language with the phonetics and phonetics to semantics, they slowly start associating written words to semantics directly.

Chapter 5

Network Architecture

As discussed earlier though imitation of speech and listening form a great part of early language acquisition in children [16], yet audition is only half the story, a great part of human cognitive capacity comes from association of auditory and visual sensory modalities, this is possible due to the lexical-semantic area. The lexical–semantic area is strategically located on the boundary between auditory and visual association cortex, receiving inputs from both Wernickes area and primary auditory cortex, as well as extrastriate visual cortex as noted by Hickok and Poeppel [24] and Price [44]. Due to this, the area also responds to visual-linguistic stimulus [4] so as to be able to relate words, heard or read, to an appropriate concept. Thus the system realizes that hearing "cat" or hearing the meow of a "cat" or reading the word "cat" or reading about a "cat" all provoke the same response as they are referring to the same object, which is "cat". We assume the input to our network to be such visual-linguistic stimulus and hence as a major novelty of this work, we introduce where-what pathways in the network simulating the brains dorsal and ventral pathways, found by Mishkim et. al. [40] through their lesion studies.

5.1 Neural Pathways

Our model is a simplified model of the lexical semantic area, and brings together important portions of the story without delving into minute details:
a) The dorsal pathway processes the "where/how" information required by its end motor — the arms. As an arm reaches a jingling toy, the location of the toy guides the action of the arm, but not directly the type of the toy. For visual-linguistics, the location of the text on a page is useful to understand the purpose. For example, a text at the top of a page might mean the title. For our network, we call the where motor as "placeholder motor" to identify the spatial characteristics of a page. To draw an analogy with a webpage, some of the text could be tagged, description tag, heading tag, emphasized, italicized or written in bigger font than normal.
b) The ventral pathway processes the "what" information required by its end motor — e.g., the vocal tract, which helps articulating a sentence. This is like a child listening or reading a sentence and then repeating it or trying to remember the meaning, at each time frame, with/without necessarily remembering the exact sentence and other details such as the prosody.

5.2 Area

The model takes inspiration from the previous neuromorphic networks like MILN [51], WWN-1 [27], WWN-2 [28] and WWN-3 [53]. The network has 3 main areas + 1 (computational layer), generic emergent area Y that is formed by sensory inputs from area X and motor inputs from area Z as illustrated in the Fig. 5.1. These areas are connected by bottom-up and top-down connections. Top-down connections are important as they are used as the supervisory signals to the network. We do not use error-back propagation as it is not biologically feasible, but top-down connections

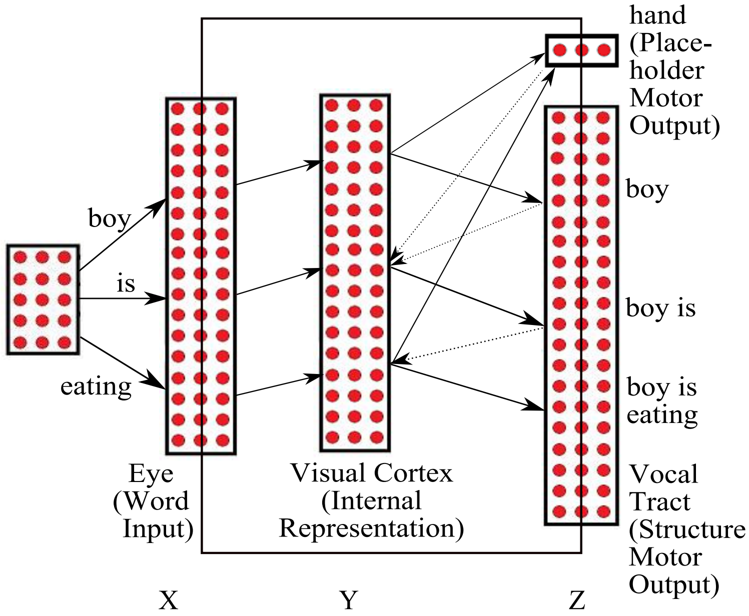

Figure 5.1: Architecture of the WWN network, consisting of 3 areas (X, Y, Z), showing the dorsal and ventral pathways. The system boundary represents the "skull-closed" architecture. The 1st layer is the Pre-processor Layer (Purely Computational, Not Biological). The later areas form a part of a simplified version of the lexical semantic area. The dorsal pathway, progressing from sensory area (X) to placeholder motor in area (Z), processes the where/how information to identify the spatial characteristics of a page. The ventral pathway, progressing from sensory area (X) to structure motor in area (Z), processes the what information required by its end motor the vocal tract which helps articulating a sentence (overt or covert action). The network is taught the sentence, *boy is eating*, each word at the 1st layer is an input to the network at a different time frame. Each input word provokes reaction from the neurons from different layers arrows represent the synapse transferred at a single instance of time. The dotted lines represent the top-down connections while the bottom-up connections are shown with solid arrows. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

have been known to exist between later to earlier cortical areas, e.g., V1 to LGN, but few networks have been able to successfully make use of them [18]. Positive feedbacks generally result in unstable systems (uncontrollable oscillations) hence most networks use top-down connections only during testing while turn them off in the training phase. Lateral inhibitions and lateral excitation can be used to solve this issue. [52] describe Lobe Component Analysis (LCA) as a model for a cortical feature level using lateral inhibitions to enable neurons to successfully detect different features.

5.3 Architecture

In the proposed network, we have 4 area layers Fig. 5.1:

Pre-processor layer (purely computational, not biological): Helps translate each input word into a binary encoding so that the representation takes up less memory space. The number of neurons in the layer is n if the number of unique words taught to the network is 2^n.

Sensory-input layer (X): Could be considered to be the retina that receives visual-linguistic stimulus in the form of words read, though for simplicity we do not model the visual input explicitly in this paper. Instead we simply provide the network with the canonical representation of the word in the form of the neuron/neurons excited by the word as it is received in its binary form from the previous pre-processor layer. Each word in the sentence is taken as input at a given instance of time in the order in which it appears in the sentence, this is similar to reading one word of a sentence at a time, from left to right in English or vice-versa in Persian. The network can be trained more than once on the same sentences as practice and review, which is like returning back to the beginning of the sentence if unable to understand. Number of neurons in this layer should be able to accommodate all the unique *"n" words* + ".", where "." indicates the end of each sentence/sequence, out of which we can create

$n \times n$ sentences.

We do not take a "bag of words" approach hence the word sequence plays an important role here. Not all word sequences formed by permutation and combinations of all words available will make complete sense in reality hence we discard those that don't and memorize or analyze the ones that do. Hence, e.g., if we are given two words *eats* and *cat* we can only create one sensible sentence out of the two, which is *cat eats* and hence we learn it while discarding *eats cat*.

Every word has a feature or a property. In our model this property is the placement of the word (whether the word appears in the title of the document or as a normal word etc.) or its font decoration (bold, italic etc.). The network tries to make out the importance of the word based on its feature that is received along with the input in this layer.

Visual Layer (Y): Neither the sensory input nor the human supervisor has a direct access to this layer. The representations in this layer are formed purely through the interactions between the neuronal synapses and signals coming in from the connections from other neurons. This makes our model "skull-closed", i.e., neither the teacher not the environment can directly modify the brain or the encoding of the system but can only manifest itself through sensory or motor inputs. Just as a teacher does not surgically wire-up his/her pupil's brain to teach it a concept but instead teaches it through experience (providing input through various senses).

The layer takes bottom-up input from the earlier sensory-input layer along with a supervised top-down input that could either be taught by a teacher through supervision or learnt from the past experience, to develop an internal representation of the knowledge so acquired, that involves learning a word sequence that might include the same words as the input or the word's meaning. It is important to notice that the network does not take a bag of words approach and only learns meaningful sentences/phrases. Every sentence ends with a ".", after which the network starts learning a new sequence.

Biologically, as depicted by the model of Hickok and Poeppel [24] and Price [44] and explained earlier in the section, language understanding and production requires two pathways the STG and the MTG. The STG connected to Broca's area, that creates the phonological loop, is used for the early acquisition of language in children. It only deals with phonemic speech output. Ferguson and Farwell [16] thought of this pathway to be able to provide an anatomical substrate for the imitation of speech. The MTG on the contrary, is important for carrying lexical-semantic information during the spontaneous production of established speech. For its simplicity we have not modeled all the brain areas supporting the above two pathways exactly as they are represented at the cortical level.

Motor cortex (Z): Consists of the motor neurons of the network that drive muscles. The placeholder motor could be the hand, reaching out to point the occurrence of a word. Similarly the structure motor could be the vocal-track helping in articulating thoughts in the form of speech (overt) or "self-talk" (intentional or covert).

Again, for its simplicity we have not modeled all the brain areas that along with the cortical connections, as described in [29], help in mapping word to articulation or mapping word to other language properties, e.g., semantics, grammar.

Skull-closed Cortical development: Before "birth" the network is not specialized in performing any particular task, it can only do so when it is trained after birth. During training, the lower layer (X), receiving the input from the external world, and higher layer (Z), receiving supervision signals from a human teacher, help the development of the cortical layer (Y). Assuming, the input from X is $\{x_1, x_2, ..., x_n\} \in x$, representing n unique words that create the sentences taught and bottom up weights, v_x map each input word to Y. Similarly, if the output in (Z) is $\{z_1, z_2, ..., z_m\} \in z$, for m unique sequence of words taught. We attach top-down weights, v_z, to map each output sequence to Y. Thus, the pre-action potential is,

$$y = \frac{x}{\|x\|} \cdot \frac{v_x}{\|v_x\|} + \frac{z}{\|z\|} \cdot \frac{v_z}{\|v_z\|}$$

which measures the degree of match between bottom-up and top-down inputs. The weight of the winning neuron is updated by a dually optimal Hebbian-like learning mechanism,

$$v_j = (1 - \rho(n_j))v_j + (\rho(n_j))y_j p$$

where j is the winning neuron and $v = (v_x, v_z)$ with yp being the product of pre-synaptic and post-synaptic activity of the firing neuron. $\rho(n_j)$ is the learning function that depends on n, age of the neurons, when a neuron wins its age is incremented by 1. Lateral inhibitions in the cortex allow only few top-k neurons to win. We can choose the number of winners or k, based on the amount of amount of generalization we want the network to learn. Hence,

$$j = \text{top-k-max}_{i=1}^{m}(y_i).$$

Layer Z is updated similarly but it has no top-down input.

The network exists in time, if we represent time-stamps as t-1 , t, t+1, t+2, ... , t+n. We must note that time is important but not critical for the function of the network. We expect time to become flexible after training. At t-1, the network gets bottom-up input as a word that is part of the sentence, it also receives the context, i.e., sequences of words that came before the current word in the sentence, these both inputs create a new state in t, to create top-down input for t+1; if t+1 is not the end of the sentence. Thus, if $V_x(t)$ and $V_z(t)$ are the weight vectors of Layer X and Z at time t respectively, and f is the area function, then,

$$y(t) = f(x(t-1), z(t-1), V_x(t-1), V_z(t-1))$$
$$z(t+1) = f(y(t), V_z(t))$$

5.3.1 Creating new sentences through Generalization

Words representing similar concepts tend to excite the same neurons thus creating a similar internal representation for words or phrases with similar meanings. This could be deemed similar to the concept of “partition” in set-theory, now if W is the set of all words and $\{p_1, p_2, ...p_n\} \in P$ be one of its partitions, where each member has similar internal representation, then to create a new sentence, let P have a “sequential” association with other partition $\{z_i\}$ ∈Z through R. Let us represent this relationship between the members of P and Z as $R(P, Z)$. Now taking up the case for each member, if $R(p_1, z_i)$ exists, then since $\{p_1, p_2, ...p_n\}$ is partition hence, $R(p_2, z_i), ..., R(p_n, z_i)$ also exist.

Further since, all members of P have similar representation,

$$y_{p1}(t) = f(p_1(t-1), z(t-1), V_{p1}(t-1), V_z(t-1))$$
$$\Rightarrow y_{p2}(t) = f(p_2(t-1), z(t-1), V_{p2}(t-1), V_y(t-1))$$

Hence, $y_{p1} = y_{p2} =$ top-k-max$_{i=1}^{m}(y_i)$.

More relationships can be defined later on but for now the network only deals with Equivalence classes and partitions, all members of the same partition have sibling relationships.

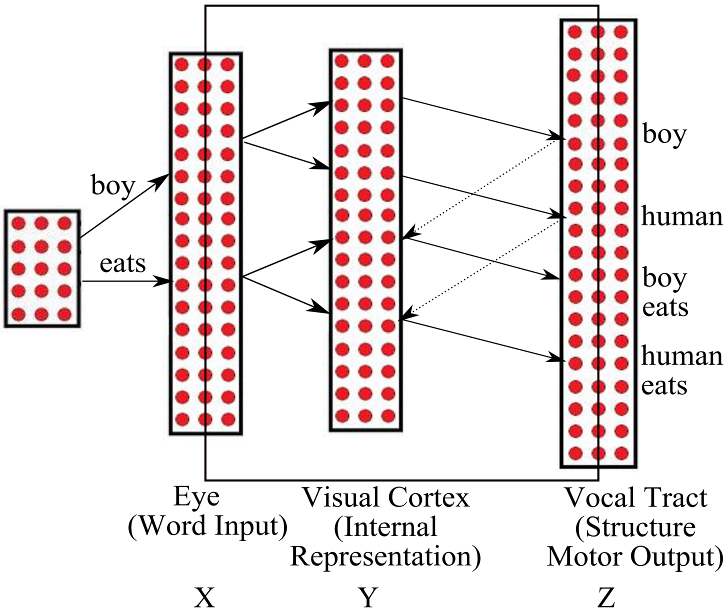

Figure 5.2: Demonstrating generalization: The network is taught the concept *boy is a human* resulting in the neurons representing *human* and *boy* co-firing, thus associating *human* with all the concepts that are associated with *boy* like *eats*. The dotted lines represent the top-down connections while the bottom-up connections are shown with solid arrows.

5.4 Derivation of Formulations

Many NLP methods are batch in the sense that all the training data are available as a batch for training. However, development is an incremental process — the agent must respond even while being trained.

In general, if $x_1, x_2, ..., x_n$ are the words in a sentence that act as a sequential input to the network, then the joint probability density of this sequence will be

$$\Pr(x_1, x_2, ..., x_n) = \Pr(x_1) \prod_{i=2}^{n} \Pr(x_i | x_1, x_2, ..., x_{i-1}) \tag{5.1}$$

However, estimation of this joint probability is expensive, and it does not lead to generalization required abstraction. Hence we introduce the concept of equivalent classes. Two sentences belong to the same equivalent class if they have the same meaning.

Now we can write (1) as,

$$\Pr(x_1, x_2, ..., x_n) = \Pr(x_1) \prod_{i=2}^{n} \Pr(x_i | \phi(x_1, x_2, ..., x_{i-1}))$$

where $\phi(x_1, x_2, ..., x_{i-1})$ is the equivalence class for $x_1, x_2, ..., x_{i-1}$. Traditionally, the above has been used for NLP. However, according to our above discussion, the purpose of cognition is to generate desired action, z_n, at each time. Thus, our formulation of a developmental agent is to focus on $\Pr(z_n|\phi(x_1, x_2, ..., x_{i-1}))$ instead of the sensory distribution $\Pr(x_i|\phi(x_1, x_2, ..., x_{i-1}))$. This is critical for "skull-closed" development because the teacher does not manipulate internal "brain" representation directly. Symbolic representation, on the other hand, corresponds to a "skull-open" approach as it is handcrafted. Furthermore, z_n is also general and flexible as it can correspond to any property of the input context. For example, the action can be directly related to the sensory class (e.g., state the name of input) or to other property of the sensory input (e.g., its location for correct arm reaching).

Lastly, the agent learns z_n recursively as the context that it needs to attend at the n-th time frame from any point in the past. The intractable problem of estimating very long temporal joint distribution above is converted into a single frame problem:

$$\text{top-k-max}_{z_n \in Z} \Pr(z_n | x_1, x_2, ..., x_{i-1}) \quad \approx \quad \text{top-k-max}_{z_n \in Z} \Pr(z_n | z_{n-1}, x_n)$$

where $z_{n-1} = \phi(x_1, x_2, ..., x_{n-1})$ and top-k means top-k actions to top matched probabilities.

Chapter 6

Comparison with Finite Automata

The two major schools symbolic AI and neural networks are divided ever since the re-kindling of neural networks in the 1980s. Weng [53] has established that a neural network can emulate any Finite Automata (FA) or its probabilistic variants such as Hidden Markov models (HMM), Markov Decision Process (MDP), Partially Observable MDP (POMDP) and Bayesian nets (also called semantic nets and belief networks). FA consists a finite set of states (Q) and transitions between the states due to a finite and non-empty set of input symbols (Σ). A new state is the result of the transition input at the current state, but there can be more than one transition paths that could be pursued from a current state, which could lead to different states, this could lead to an indeterminism, this can be resolved if we deploy a human who would choose between the transition paths and lead the logic to a particular state. Now the FA is deterministic whose mathematical model could be written as a quintuple, (Σ, Q, q_0, δ, A), where $q_0 \in Q$ is the initial state, $\delta\colon Q \times \Sigma \mapsto Q$ is a transition function and $A \in Q$ is a set of accepting states.

Let us try to design an FA that learns a phrase/sentence, ... *young cat looks.* Every transition, at a time instance, leads the current state to an intermediate/final output state. Some states have equivalent states that can be reached from the current state through a different transition, e.g. *young cat* can be called a *kitten* or

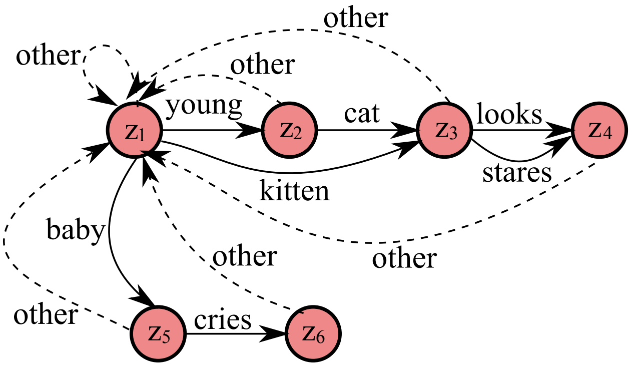

Figure 6.1: Hand-crafted FA learning sentences. It should be noticed that the states are pre-programmed and there is no brain that takes intellectual decisions about tasks. z_1 represents the start state, every sentence starts from z_1. Here the network has 6 states and 13 transitions between them. The dotted arrows show the error conditions, the machine returns to start state if the input to a certain state is not recognized.

looks can be replaced by a similar word *stares*. To define our states, $z_1 = q_0$, it can transits to another state z_2 if it receives an input, σ. Hence, $z_1 \xrightarrow{\text{young}} z_2$, similarly, $z_2 \xrightarrow{\text{cat}} z_3$. But instead the FA could reach z_3 from z_1 by following a different transition path, $z_1 \xrightarrow{\text{kitten}} z_3$. Fig. 6.1 describes the transition between the states in FA.

Our network can imitate all the actions of an FA. X and Z in our network correspond to the Q and Σ of an FA, the human teacher has the required representation of Z which means the human teacher knows the language and uses certain methods to teach it to the network or the learner. By canonical conversion from a symbolic set $\Sigma = \{\sigma_i \mid i = 1, 2, ..., n\}$ to an n-dimensional vector space X, i.e., σ_i corresponds to $x_i \in X$ where x_i is the only i-th component of X vector matrix to be 1 while all others are 0. We say they are equivalent, denoted by $x_1 \equiv \sigma_i$, in the sense of canonical conversion, similarly, the old state, $q_{\text{old}} \equiv z_j$, winner neuron in Z at time $t-1$, at the next time instance the input weight vectors to the layer Y from layers X and Z is (v_x, v_z). When the resulting neuron in Y fires, it stimulates a corresponding neuron in Z linked to this neuron to fire. Thus leading to the new state, q_{new}, which is the winner neuron in Z at time t. WWN hence learns the required behavior

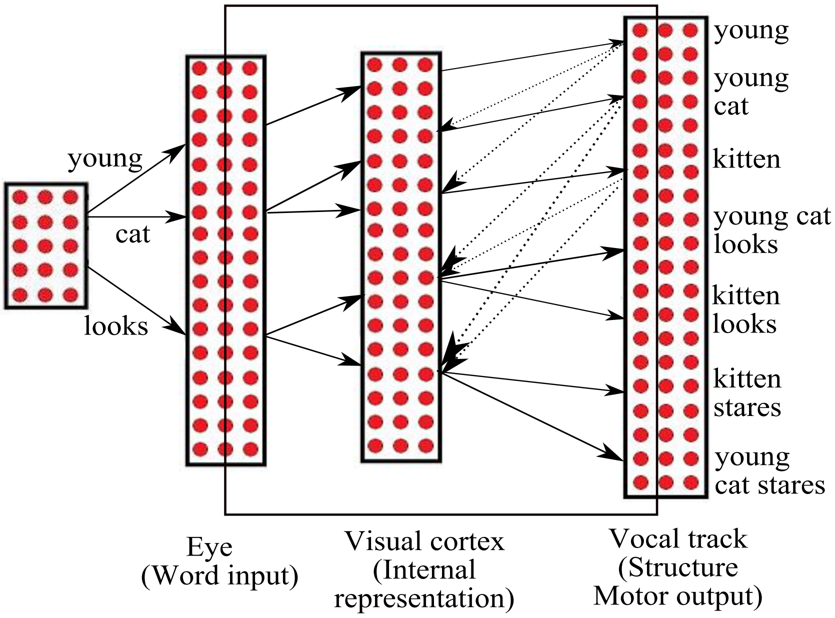

Figure 6.2: “Skull-closed” WW-Network that can learn the very same sentences as the FA but through autonomous learning. The network can use its previously learned states and the prowess of generalization to equate *young cat* with *kitten*. In the above figure the network learns few of the states shown in Fig. 6.1.

function f such that $z(t) = f(z(t-1), x(t))$, where $x(t) \in X, z(t-1) \in Z, z(t) \in Z$. In mathematical notation, f is a mapping, $f : X \times Z \mapsto Z$, just like the FA mapping of $\delta : Q \times \Sigma \mapsto Q$. We can now conclude that if an FA has c transitions then our network needs c neurons in Y to correctly map the relation $q_{\text{old}} \xrightarrow{\sigma_i} q_{\text{new}}$. Furthermore, equivalent transitions can be taken care of by our network through the *generalization* and *thinking* theories described earlier.

FA and several of its probabilistic variants like MDP, POMDP, HMM etc. as described in previous chapters have their states hand-crafted by humans. It is important to note, FA does not have a brain to learn of take any intelligent autonomous decisions, it can be called a “brainless” system. The output in FA is deterministic and static, i.e. a specific output symbol is attached to a transition or to a state. FA and similar systems cannot remember the exact input and transition route taken to reach a particular state. Our network on the other hand is dynamic with an emergent internal representation, which is formed through the network’s exposure to the sensory and motor fields through its peripheral layers X and Y respectively. As shown above WWN can learn all the states of an FA but can also reason to create new unseen states. Therefore, WWN solves the weakness of neural networks like types of agents in conducting goal-directed reasoning as pointed out by Minsky [39]. Furthermore, where FA is simply in the mind of the teacher it can learn more if designed further, WWN is a non-task specific incremental learner. WWN has both short-term memory and the capability to learn is able to do so thanks to the internal representation at layer Y. The network for a brief amount of time stores short history of $q_{\text{old}} \xrightarrow{\sigma_i} q_{\text{new}}$.

The compendious Oxford English Dictionary lists about 500,000 words in English language and a further half-million technical and scientific terms remain uncatalogued. Thus number of words is limited and the number of combinations that can be created with the words can be at most 1,000,000 × 1,000,000. If we try to design all the possible permutations and combinations in an FA we will soon realize that not only is it a labor-intensive process but it also takes up a lot of memory. Now consider the human brain. The number of neurons in an adult brain is 10^{11}, which is 10 times lesser than the total number of permutations that can be formed with all the words possible. Hence it will not be very feasible to use finite state machine to learn a language.

Keeping this in mind, let us question the logic of storing all the combinations of words in our system. The fact is our goal is not to find all word combinations possible but to know all sentences possible; the definition of a sentence states that a sentence is a group of words that “makes complete sense”. There will be far less word permutations that will result in the formation of such grammatically coherent sequences. Also, sentences are of finite length. Hence, it only makes sense to store the sequential phases. Our network does that in addition to having other means of creating new sentences on its own which decreases the storage demand on the system to a great extent.

Chapter 7

Machine Reasoning and Logic vs Human Reasoning

Many traditional artificial intelligence groups have used logic in order to reach feasible conclusions from known facts. The main strength of logic is, it is precise. It can not be broken if the facts are true and unambiguous. In linguistics logic has been used for similar reasons. But as language draws its main inspiration from the day-to-day life common sense becomes a defining factor. John McCarthy in his paper *Programs with Common sense* [38], creates a system that demonstrates common sense by logically deducing expected actions from a group of sentences. According to him *a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows.* But it is not very simple to model the real world in a computational system, for example to deduce the simple fact that “if you are in your car and you drive from home to the airport, then you reach the airport” using the following logical statement: *canachult(at(I,car), go(home,airport,driving), at(I,airport))*, computationally one needs to define a lot of sequences and objects with a substantial amount of logical scaffolding. McCarthy himself calculates the number of premises needed to make a computational system to get up from its desk and reach the airport is 17. These premises have to not only be pulled out of the memory but should be placed in correct order and interpreted correctly to proceed from the inspiration to the goal.

Humans on the other hand, seem to be able to do this a lot more easily. In fact humans do it more frequently as it is not possible to store the amount of information they are exposed to during their waking life in a single brain. Had it not been for common sense it would have been impossible for humans to be able to get anything done. So what is common sense in humans?

Lakoff and Johnson's *prototype theory of human categorization* [31], identifies some properties to be more central than others in objects. For e.g., though all birds have a beak, two eyes, two legs, yet of all these visual features feathers and wings seem to be most significant as they describe the object better and therefore are given more weightage than the other elements associated with the object. Similarly the ability of flight is more defining of a bird than its ability to walk. Humans are able to pay more attention to such central characteristics and group all objects into categories. These categories are not strictly scientific with hard boundaries, instead human categorization is more of a rough and fuzzy differentiation, not necessarily based on logic and reason. This is precisely where a logical framework fails. Moreover, human knowledge does not consist of absolute truths and absolute fallacies, and hence applying pure rules of mathematics cannot represent common sense. The wealth of human knowledge and the ability of the human brain to be able to categorize increases their ability to reason.

The distinction between the logical and the human reasoning is, the former is deductive, while the latter could be called as a combination of inductive reasoning along with scholasticism, rationalism, empiricism and associations formed due to experience.

7.1 Inductive Reasoning

Inductive reasoning is one of the reasoning processes. In contrast to the deductive reasoning, inductive reasoning does not support the logic that if the given premises are used to arrive at a particular conclusion then if premises are true then the conclusion if derived systematically and correctly will be true as well. Inductive reasoning does not rule out the possibility that even though the premises might be true yet the conclusions might not be so and that is because in inductive reasoning there is no logical movement from premise to conclusion. For e.g., given sparrow is a bird and birds can fly does not guarantee that kiwi too can fly, on the contrary there are several birds that cannot fly despite having wings and feathers. Similarly, a bird with a broken wing can not fly. Induction allows the system to doubt the correctness of the conclusion.

Chapter 8

Concept and Theme

8.1 Language Association to Assist Generalization

The network uses the concept of "primary and secondary association" or classical conditioning [12] to learn new concepts about objects. The occurrence of conditioned stimuli, CS_2 followed by CS_1, and CS_1 followed by a conditioned response (CR) trains a subject to correlate the occurrence of CS_2 to an otherwise unrelated CR.

$$CS_2 \xrightarrow{p} CS_1 \xrightarrow{p} CR \Longrightarrow CS_2 \xrightarrow{s} CR$$

($\xrightarrow{p}$, $\xrightarrow{s}$ means "primary" and "secondary" associations respectively, $\Longrightarrow$ means "results in"). Here the relationship between CS_2 and CS_1 and CS_1 and CR are primary relationships as they take place one after the other and in some case might be a result of the previous stimulus, but the relation between CS_2 and CR is secondary. It slowly develops as the agent experiences the same temporal routine over and over again. Thus transforming "primary" associations into "secondary". The above notation comes from psychology whereas the WWN-text network representation for the above is shown in Fig. 8.1.

Explaining the Pavlovian experiment in the above notation:

$$Tone \xrightarrow{p} Food \xrightarrow{p} Salivation \Longrightarrow Tone \xrightarrow{s} Salivation$$

Our network takes inspiration from the above theory. In our model, firing of neurons is equivalent to a stimulus or an event hence sequential firing of neurons could

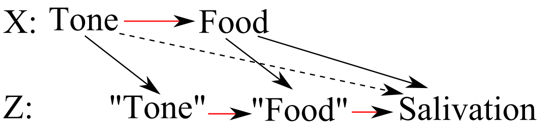

Figure 8.1: External network notation of WWN for classical conditioning. Only the external layers are shown in the diagram. 2 areas of WWN are seen, X as input and Z as output area. The red arrows show the progression in time. Words within “” are concepts while the once without quotes are actions. The black arrows show the learning loop in the network comprising of the primary associations only. The dotted arrows are the learned associations.

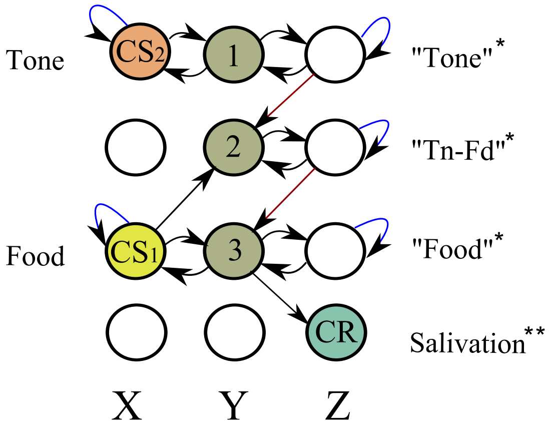

Figure 8.2: WWN network model for classical conditioning, as a special case of general process - autonomous thinking. WWN network imitating the Pavlov experiment. The dog hears a Tone each time the food is presented, after certain time the dog starts salivating when it hears the tone even though no food is presented. The 3 areas of WWN are seen, the blue arrows represent internal reverberating signals that make a sensation last in a neuron a little while longer than a single time step so that it associates the earlier stimulus of Tone with the stimulus presented after it Food and after sufficient training is able to associate Tone with Salivation (conditional response). Words within “” are concepts while the once without quotes are actions. * represents concepts being learned while ** represents concepts that have already been associated to each other.

result in the emergence of a new patterns and concepts.

8.1.1 Parroting

Early language learning as described in chapter 4.1.1 is the primary way of language acquisition. The learner is taught the language through speech or reading material, the learner repeats the taught premise, through overt or covert behavior. The output is supervised and both the input and the output come together to form primary associations between the words that occur together more often in sentences.

8.1.2 Member to Class Generalization: Association Aided by Associative Reasoning

The network is taught 3 concepts: object, class and feature. The object belongs a certain class and has a certain feature. The network is expected to relate this feature to the class and generalize the concept.

Co-firing of two neuron representing an "object" and its respective "class" can help form primary links between words. The network takes a word as input in Layer X. At the same time, the word is introduced to its parent class if any, parent class is actually a generalized concept of the word. E.g. the parent class of *Sparrow*

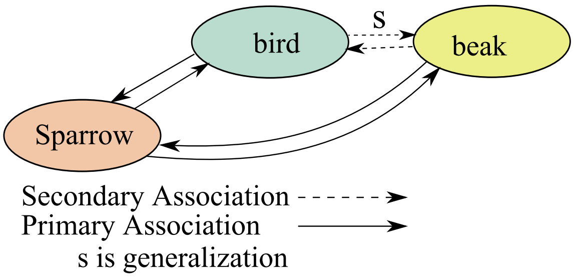

Figure 8.3: Member to class generalization: A concept "bird" is defined along with the concept of "Sparrow", if *Sparrow is a bird* and *Sparrow has a beak*, the network is then expected to figure out the generic knowledge that *birds have beaks* through association.

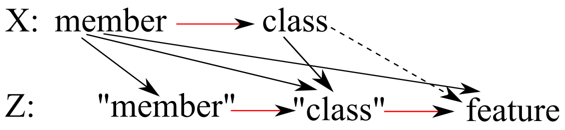

Figure 8.4: External network notation of WWN for member to class generalization. 2 areas of WWN are seen, X as input and Z as output area. The red arrows show the progression in time. Words within “” are concepts while the once without quotes are actions. The black arrows show the learning loop in the network comprising of only the primary associations. The dotted arrows are the learned associations.

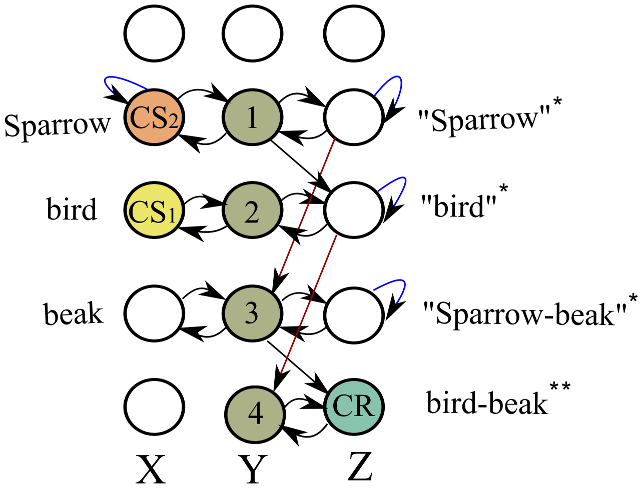

Figure 8.5: WWN network demonstrating member to class generalization: The three areas are shown X, Y and Z. The blue arrows show continuous reverberatory signals within the neuron such that it is able to hold a state for a little while longer than a single time step. The brown arrows from Z to Y represent the top-down connections. Reverberatory signals also run continuously between and within the areas, resulting in primary associations to be transformed into secondary associations as in classical conditioning. In this case, the network is presented with the fact that “Sparrow has a beak” and that “Sparrow” is associated with “bird” as “Sparrow is a bird”. The network later is able to generalize from a specific example to a whole class to learn “Birds have beaks”. Note that the network is never taught the concept of “class”, “subclass” or “object”.

could be *bird*. The parent class and the current word belonging to it co-fire. In the next time step the parent class word is taken as top-down input by layer Y, which is combined with the new bottom-up words (which might or might not have a parent classes or equivalent words) resulting in a new sentence sequence in the verbal motor. Thus both the word and the class combine together to form both context and structural rules. This has been explained in detail earlier in chapter 5 in the section 5.3.1. Though in current network we will have to explicitly teach the network that the parent-child relation is bidirectional, this could be considered as a topic for future research.

8.1.3 Subclass to Superclass Generalization

The network is taught 3 kinds of conceptual objects: object, subclass and superclass. The object belongs a certain class with is a subclass belonging to a different superclass. Every member of a set is the child of the set and every child has the same properties as the parent set. E.g., every *girl* is a *human* and can do everything that a *human* can or have all the properties of a *human*. Also, all neurons representing a “child/subclass” concept and its corresponding “parent/superclass” class co-fire.

So tracing the progression of time, we can see that,

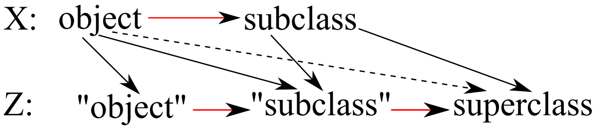

Figure 8.6: External network notation of WWN for subclass to superclass generalization. 2 areas of WWN are seen, X as input and Z as output area. The red arrows show the progression in time. Words within “” are concepts while the once without quotes are actions. The black arrows show the learning loop in the network comprising of only the primary associations. The dotted arrows are the learned associations.

The process is similar to classical conditioning [12] where simultaneous co-firing of

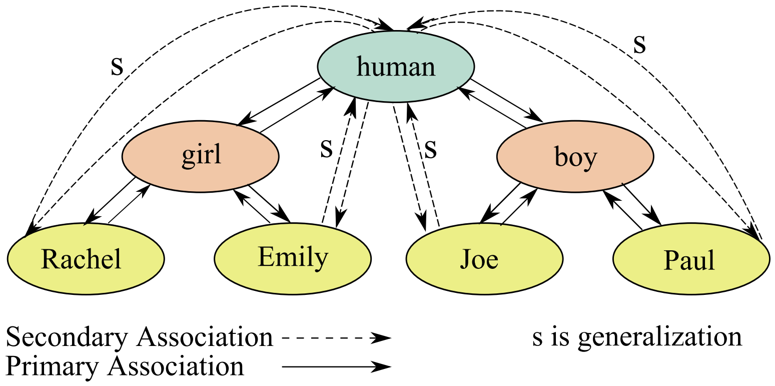

Figure 8.7: Subclass to superclass generalization: A concept "human" is defined along with the concept of "girl" and "boy", the left branch of the tree is explained thus, *girl is a human* and *Rachel and Emily are two girls*, the network is then expected to figure out that *Rachel and Emily are both humans* through association.

neurons is similar to simultaneous occurrence of stimuli.

8.1.4 Member to Member Generalization and Classification From Similarity

Network is taught to identify members of the same "partition", as defined earlier, and apply the property of one member to the other, while not confusing the members of separate partitions to be similar. For this the network is again introduced to a class and its two subclasses having different properties and features. But since the two classes are different the objects belonging the two also defer in features, i.e., though *Kiwi* and *Sparrow* are both birds yet the network understands that they have different flight capabilities. Thus we *classify based on similarity.*

The process involved in both the processes is association and not logical reasoning.

It should also be noted that in the current network we have to explicitly teach the network if the relation between the concepts is bidirectional or unidirectional,

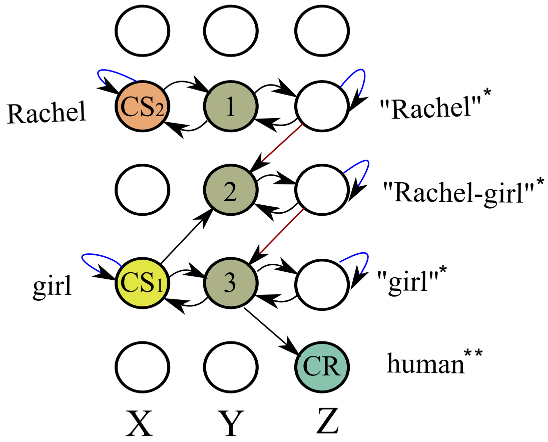

Figure 8.8: WWN network demonstrating subclass to superclass generalization: The three areas are shown X, Y and Z. The blue arrows show continuous reverberatory signals within the neuron such that it is able to hold a state for a little while longer than a single time step. The brown arrows from Z to Y represent the top-down connections. Reverberatory signals also run continuously between and within the areas, resulting in primary associations to be transformed into secondary associations as in classical conditioning. In this case, the first presented fact is that "Rachel" and "girl" are related and the "girl" and "human" are related is the second fact. The network is then able to connect the concepts to create a relation between "Rachel" and "human". Note that the network is never taught the concept of "class", "subclass" or "object".

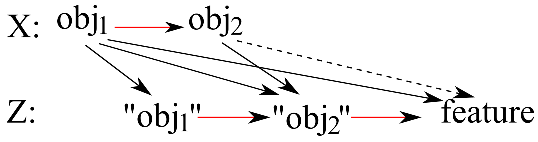

Figure 8.9: External network notation of WWN for member to member classification through similarity. 2 areas of WWN are seen, X as input and Z as output area. The red arrows show the progression in time. Words within "" are concepts while the once without quotes are actions. The black arrows show the learning loop in the network comprising of only the primary associations. The dotted arrows are the learned associations.

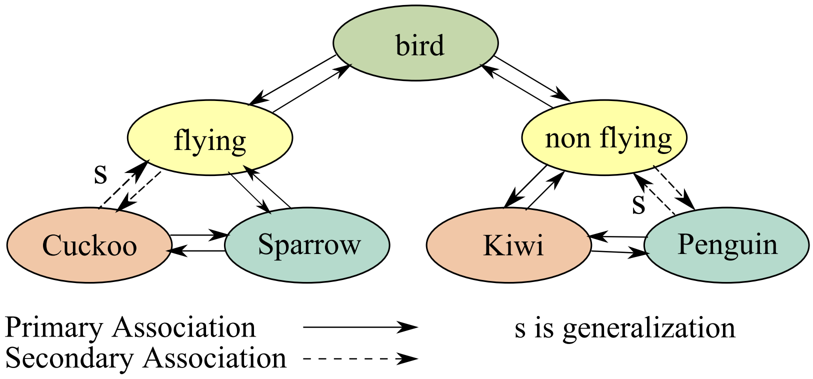

Figure 8.10: Member to member generalization: The concept of "bird" is taught, along with the fact that a bird could be a "flight bird" or a "non-flight bird", explaining the left branch of the tree. Learned sentences are *Cuckoo is a flying bird* and *Cuckoo has the same properties as a Sparrow*, the network tries to associate the concept of "Sparrow" with that of a "bird" figuring out that *Sparrow is a flying bird too.*

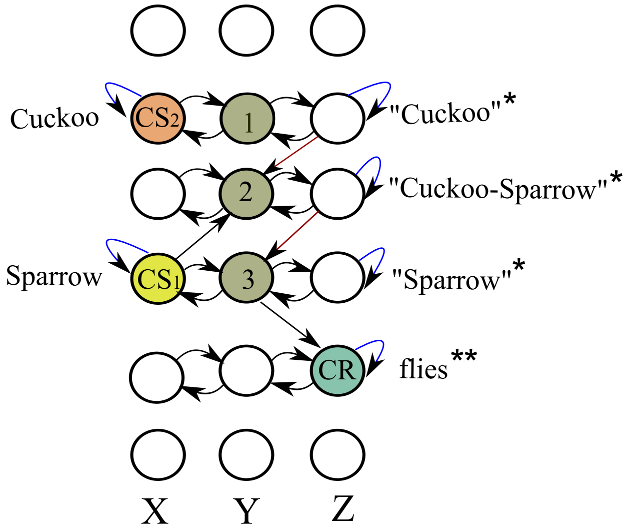

Figure 8.11: Member to member generalization: The three areas are shown X, Y and Z. The blue arrows show continuous reverberatory signals within the neuron such that it is able to hold a state for a little while longer than a single time step. The brown arrows from Z to Y represent the top-down connections. Reverberatory signals also run continuously between and within the areas, resulting in primary associations to be transformed into secondary associations as in classical conditioning. In this case, the network has already learned that "Cuckoo flies", in the time stamps shown in the figure the model learns that "Sparrow" has similar properties as the "Cuckoo" and hence later associates "Sparrow" to the property of flying.

i.e., we have to explicitly tell the network that the *bird* is the parent of *Cuckoo* and *Cuckoo* is the child node of *bird.* As the network gains more experience and learns more about parent-child relationships, it is able to relate the two objects in a bidirectional manner in its own. This could be used to create cause and effect and vice-versa relations which can be very useful in question-answer like conversations. This will result in the formation of various different kinds of relationships between objects. Not only that, this could later be later used in the formation of relationships between relationships. So, if the network is initially taught that Professor Snape is the teacher of Harry Potter then if the network has enough training it would realize that the pupil of the teacher is his student and if Professor Snape teaches Harry then Harry Potter must be Professor Snape's student or pupil.

8.2 Attention Allows Generalization

8.2.1 Grounding leads to generalization

According to Harnad [22], a symbol is grounded if the robot can pick out which category of sensorimotor projections it refers to. This might include attaching various pre-existing notions to the object or creating new ones based on the systems experience with it, these two major methods of acquiring grounded categories are known as, symbolic theft and sensorimotor toil, respectively. Symbolic theft as the nomenclature describes is the knowledge gained by the system from another source or a teacher. The metaphor "theft" here should not be taken literally as the system does not "rob" the teacher off his knowledge. Sensorimotor turmoil on the other hand refers to the system developing its own knowledge of an object through trial and error, learning in its own capacity. Of course to say the obvious, it is much easier and faster process to acquire categories through symbolic theft.

Thus grounding is an important precursor to a system developing an understanding of physical and metaphysical attributes of an object. But grounding is only helpful and effective if the symbols in the teachings are already grounded. Harnad explains it very well in his writings as the concept of a “Peekaboo Unicorn”, to explain it is Unicorn, or a white horse with a single horn, but it has the peculiar property that it vanishes without a trace whenever senses or measuring instruments are trained on it. Thus none of our senses can ever perceive it but a child can still be made to understand what it is if the child knows the concepts of horse, horn and vanishing. Children use the sensorimotor toil to gain first hand experiences about various things around them until they are enough to understand the language of parents to be able to gain from the experience of their elders. But parents can help their children to understand their surroundings better by encouraging them to touch a few things and play with them while discouraging them from playing with harmful things.

When the system comes across more objects that confer to having similar properties and behavior then grounding leads to generalization, which is a sensorimotor capacity that allows us to sort the world around into relatively orderly taxonomic kinds marked out by our differential responses to it [21].

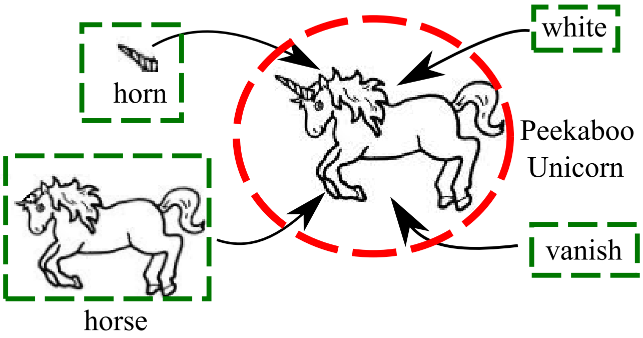

Figure 8.12: The concept of how Peekaboo Unicorn looks is developed through already grounded concept of horns, horse, white color and invisibility. If one understands all the latter four concepts they would be able to understand what a Peekaboo Unicorn is.

Our model focuses on this process in order to help the system learn generalization. Though very crude this could be used to create grounding and associating features to an object. To simplify matters, we introduce visual elements to the system through linguistic statements. In the beginning, we use the phrase “has a” to associate the corresponding object with the visual properties. E.g., *horse has a tail* or *horse has four legs.* Hence if *unicorn has a tail* and *unicorn has four legs* then the child would infer that the unicorn looks very much like a horse or in fact an animal.

$$object1\{visual - element1\} \xrightarrow{given} class - A$$

$$object1\{visual - element2\} \xrightarrow{given} class - A$$

$$object2\{visual - element1\}\&object2\{visual - element2\} \xrightarrow{deduced} class - A$$

8.2.2 Specific Relations: Attention Makes Prediction Possible

Human beings are able to communicate even in the most noisy places, furthermore they do not have to listen carefully to every word being said but have to merely pay attention to a few important words and concepts and how they are linked to each other. Though at times it is important to remember the exact words of the speaker, e.g. in highly crucial diplomatic meetings where quoting wrong ideas could be dangerous for the international relationships of the countries. But normally all the conversations that people indulge in, in day to day business are not as important hence only attending to important words and a little help from earlier experience can help the system to arrive at the correct context. This is, though, not always true but is deemed to be true in most of the circumstances. Computational linguistics system use similar stochastic measures to try to associate words that are more or less likely to appear together in order to decipher language in noisy backgrounds.

If the words familiar to the network appear in a sentence then the systems starts linking the words appearing in the same sentence. E.g. if 2 sentences are taught to the system, namely, *peach is a delicious fruit* and *peach is a sweet tasting fruit* then when the system comes across different sentences about *peach* linking it the same word *fruit* twice or more times, then it starts realizing that *peach* might be in some ways related to *fruit*. The more the two words occur together the more strongly bound are their concepts. That is, the neurons representing linked concepts, like *peach* and *fruit*, normally would fire one after the other but if the network is aware of the concept of *fruit* then after training it starts associating the concepts that do not occur in immediate vicinity but in the same sentence or while explaining the similar context. It is not necessary that the words should have a parent-child or object-property relationship, the relationship could be anything or nothing at all, but until the two words appear together they can be linked to each other.

Attention as described in [36] helps the network to be able to focus its intent on certain specific text to form a certain context. Top-down signals to the network can modify the course the sentence takes. As described earlier, our model uses association to predict in exact order in which words are meant to occur. One kind of phrase leading to a certain words while the other leading to something else. E.g. the sequence, *the man read the* leads to *newspaper*, while another sequence, *the man ate the* is more likely to be followed by *hamburger*. Thus the system tries to find out the most likely of the words to be followed by the phrase given to the network. But paying attention to the words *ate* and *read* can help the system to quickly identify the more likely to the two choices. Thus attention helps prediction to derive correct

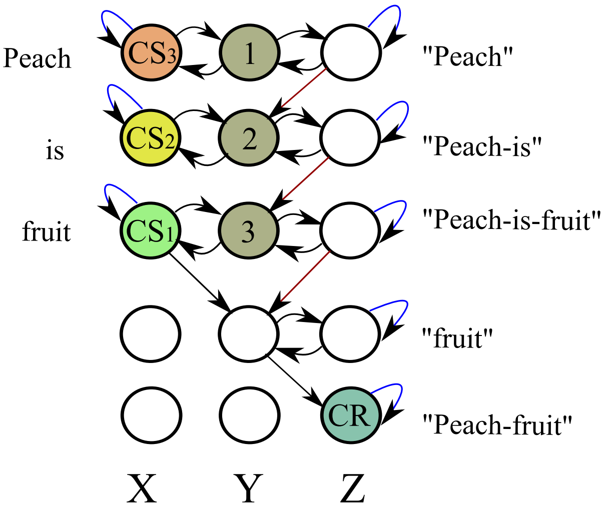

Figure 8.13: Imitating classical conditioning with the help of reverberating signals between neurons to form link between words that occur together. Learned sentence includes *Peach is a fruit*. When “Peach” and “fruit” concepts occur together in sentences several times the neurons representing the two concepts get wired to fire together. Hence as soon as the network gets “fruit” as input it is able to predict the occurrence of “fruit” in near future. Thus, in a way predicting the forthcoming word.

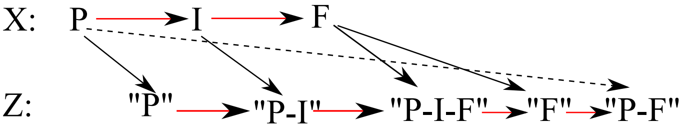

Figure 8.14: External network notation of WWN for imitating classical conditioning with the help of reverberating signals between neurons to form link between words that occur together. WWN notation showing the learning of the concepts shown in Fig. 8.13. P stands for “Peach”, I is for “is” and F for “fruit”. P-I-F in turn is a sequence “Peach is fruit” which is a more specific relation than “Peach-fruit” (P-F) relation.

conclusions.

8.3 Word Representation as Sensory Vector

In visual perception, a word is a sub-pattern in a complex background, as illustrated in our model. If a system that input computer words one at a time without background, we can map each word $\sigma \in \Sigma$ to a binary vector $x \in X$ of a fixed dimension d. Suppose that each element of x is either 0 or 1 and x is not a zero vector. All such binary vectors of dimension d can represent $2^d - 1$ words.

Given a fixed and sufficiently large d, it is beneficial to choose those binary vectors that give lower normalized inner product $r(x_i, x_j) = x_i \cdot x_j / (|x_i||x_j|)$ for all $x_i \neq x_j$. In the canonical mapping, we have $r(x_i, x_j) = 0$, for all $x_i \neq x_j$. This is too wasteful, as n words require a vector space of n dimension. This is also not necessary.

Define a mapping $m_i : \Sigma \to B_i \in B$, where B consists of binary vectors. We can define mapping m_1 as the canonical mapping. Define mapping m_i as a mapping whose range B_i is such that B_i contain all the binary vectors in B that have exactly i nonzero components. This canonical mapping is m_1.

Define the inter-set distance $d(A, B)$ between two sets A and B to be

$$d(A, B) = \min_{a \in A, b \in B} \left(1 - \frac{a \cdot b}{|a||b|} \right).$$

We have $d(B_1, B_2) = 1 - \frac{1}{\sqrt{2}}$. In general $d(B_i, B_j) = 1 - \frac{\min\{i,j\}}{\sqrt{i}\sqrt{j}}$. The larger the inter-set distance the better, as the network can distinguish vectors from different set using normalized inner products.

Define the within-set distance $d(S)$ to be

$$d(S) = \min_{a \in S, b \in S, a \neq b} \left(1 - \frac{a \cdot b}{|a||b|} \right).$$

We have $d(B_i) = 1 - \frac{i-1}{i} = \frac{1}{i}$. Likewise, the larger the within-set distance, the better.

From the above analytical results, we should choose a large d allowed by the computational resource. Then, we choose m_1, m_2, m_3, ... in such an order till all the words are mapped.

When we map Σ to $B' \subset B$, it is desirable also to pay attention to the distances between vectors in B'. Consider three words, “read”, “reader”, and “readership”. As these three words are similar, it is desirable for their binary vectors in B' to keep such similarity in the distance space of the normalized inner product. For example, one can assign $(1, 0, 0, ...)$, $(1, 1, 0, ...)$, and $(0, 1, 0, ...)$ to these three words, respectively. It is true that it is impractical for one to keep all pair-wise distances intact in the new space B, but a good mapping tends to give better performance with a limited amount of learning.

Chapter 9

Experiments

9.1 Data

The data consists of 6 to 7 word sentences in English, the words are randomly distributed. The problem space has a total of unique 1631×4 states, where 1631 is the number of "what" states while 4 is the number of "where" type. Thus the input vector is two dimensional, first representing the input word and the second representing the feature. This is different from the real input image, but is meant to simplify internal visualization and internal maturity.

Few sentences are related to each other as they talk about the same object or relate two objects, e.g., through "is-a" relationship. *Bird eats worm* and *Baby eats apple*, here both the sentences are talking about "eating". The network is trained on the same sentences a multiple number of times. An "is-a" relationship is defined in some sentences, to define an object-class relationship. For e.g., *Sparrow is a bird* and *Girl is a human*, relate the objects "Sparrow" and "Girl" to more general classes "bird" and "human" respectively. Needless to say objects in the same classes can be called equivalent and are supposed to have similar properties, i.e., if "girl" and "boy" both belong to the class "human" then both could share properties like "have hands", "can eat" etc. The network is taught 20 classes having at least 1 property each. Four type motors are “title”, “bold”, “italic” and the default type “word”. Every word in a sentence should have certain type feature that helps the network to identify the word more precisely.

The data set has been pre-synthesized, as we needed to ensure that certain relationships existed between the words and the concepts presented in a sentence or across various sentences, which could be used to demonstrate the capability of the system to create new sentences as in experiments 2, 3, 4 and 5 (to be described later). We wanted to demonstrate the capabilities of the system through the real world-like examples just as a child and develop grounding through simulated experience of language communication. For experiments 1, 6 and 7 that are only concerned with learning simple associations between words through *parroting* we have used random sentences that had been picked up from the certain online sources that catered to child learning.

The input to the system consists of words in the order in which they appear in the sentence, separated by ’.’s that indicate the ending of the sentence. Each time the network encounters a ’.’ it realizes the sentence has come to an end and it re-initializes the outmost layer ‘Z’. This allows the network to only learn the sequence of words appearing in the same sentence but not the sequence of the sentences themselves. The output consists of the word sequences/structural patterns or parts of sentences, e.g. in a 4 lettered sentence, $s = ABCD$, where A, B, C and D are 4 words that are all essential to the sentence meaning, without meaning-irrelevant words (e.g., stop words), here the word sequences learnt will be $s_1 = A$, $s_2 = AB$, $s_3 = ABC$ and $s_4 = ABCD$, $s_4 = s$ thus learning the *sequential* associations. Furthermore, while learning a new 4 lettered sentence $s_{new} = ABCZ$ does not have to re-learn s_1, s_2, s_3 but can directly learn s_{new}. As for the type motor, there are four of them, represented by <b> for words in bold font, <i> for italic font, <sub> for words in the title of the page and none for a simple word with no formatting or feature attached.

Total number of input sentences = 308.

Total distinct words in the sentences in the training set = 892.

Total number of states learnt = 1631.

Number of feature types = 4.

Number of classes = 20.

Hence, for experiment 1 the network dimensions are: X = [892 × 4], Y = [1631 × 4] and Z = [1631 × 4].

9.2 Experiments and Results

The following experiments show the major capabilities of the network. The configuration of the network is as follows. The network has 4 layers. The top-down weight = 0.7 and bottom-up weight = 0.3.

9.2.1 Parroting

The sentences had 1631 different word sequences/structural patterns or parts of sentences. In 2 epochs the network is able to imitate structural learning or the Audio/Verbal motor to perfection. The network is also able to learn the where/how type motor with no error in 2 epochs. The network with "where/how" pathway was found to be more efficient in recognizing the sequences than the network without it.
Input: Words (in the sequence in which they would appear in the sentence).
Output: Sequence of words learned.

Table 9.1: Sequence of words learnt with the passage of time

Time Frame	t_1	t_2
Input	EATING	HABITS
Output	EATING	EATING-HABITS

Table 9.2: Sequence of words learnt with the passage of time, with bold words

Time Frame	t_1	t_2	t_3	t_4	t_5
Input	Baby	likes	to	**eat**	apple
Output	Baby	Baby-likes	Baby-likes-to	Baby-likes-to -**eat**	Baby-likes-to -**eat**-apple

Table 9.3: Sequence of words learnt with the passage of time, with italicized words

Time Frame	t_1	t_2	t_3	t_4	t_5
Input	Bird	likes	to	*eat*	worms
Output	Bird	Bird-likes	Bird-likes-to	Bird-likes-to -*eat*	Bird-likes-to -*eat*-worms

In the given examples though the word "eat" is used in different contexts both the times yet if tested on the word alone the network will be able to identify the two based on the feature associated with the word, here the fonts italic (*eat*) and bold (**eat**). Words in SMALL CAPS represent title of the web-page. Results plotted in Fig. 1.

It should be noted that conceptually the input and output are synchronous but the programming model has a delay due to the computational requirements of the layers.

9.2.2 Canonical Word Representation in Sensory Vector

The network was tested on the same set of sentences, with different B_i, such that for the input vector to area X had i bit as 1 while the rest are 0. The network matrices change as we go on increasing the number of bits to see the performance. The

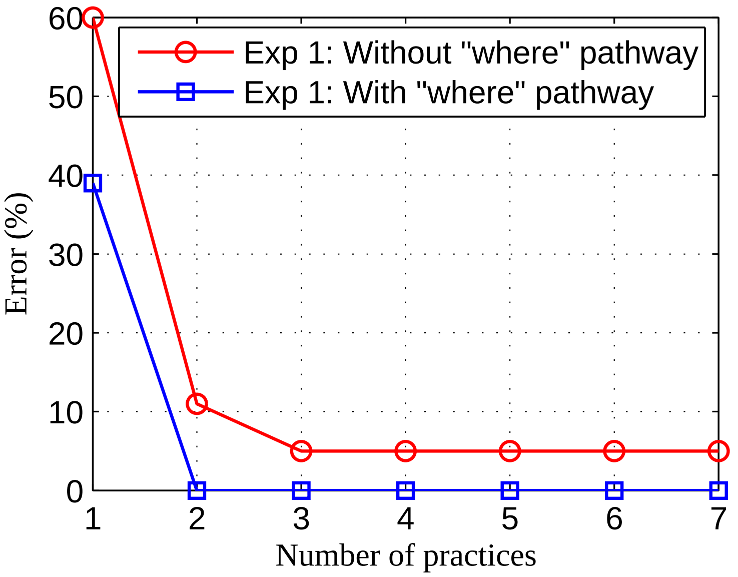

Figure 9.1: The graph plots the error rate of recognizing sequences, with and without "where" pathway in the network. The network with "where" pathway is able to reach 100% recognition rate within 2 epochs of training in all the experiments as it is able to identify words based on the additional feature.

network was tested on the task of parroting only, converged to 0 error rate within 2 epochs. We also plot the time and memory usage for all the various canonical and non-canonical representation, i.e. B_1, B_2, B_3 and B_4 in order to compare the performance. B_4 proves to be the most superior as compared to the rest as it is using the smaller matrix to represent the input vector.

9.2.3 Member to Class Generalization

The network is able to learn the properties of the objects and apply them to the classes perfectly and reaches 100% performance within 2 epochs. It should be noted that the concept of "class" or "member object" is not programmed into the network instead the network is taught to associate the same sequences with all co-firing neurons. Thus, if there is a sentence with n partitions in it, with each partition containing m members then it can learn a total of n^m sentences.

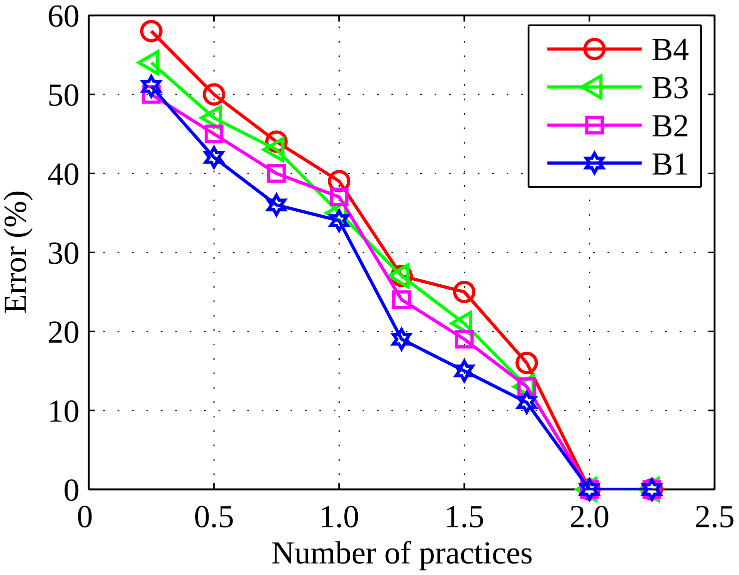

Figure 9.2: Error rate for all output states is plotted against the epochs. Total number of states is 1631. B1 is the canonical while B2, B3, B4 are non-canonical representations of the words with 2,3 and 4 bits 'on' in the input vector. We test only for the parroting task as explained in experiment 1. The epochs 0.2, 0.4, 0.6 and 0.8 mean that the network is trained on 20%, 40%, 60% and 80% of the data while being tested on the complete data used for the whole experiment. Similarly epoch 1.2 and so on mean that the network has been trained twice on 20% of the data but only once on the rest of the data, the testing set always consists of the complete data used for the experiment.
Network dimensions for B_1: X = [892 × 1]; for B_2: X = [44 ×4]; for B_3: X = [20 × 4]; for B_4: X = [15 × 4].

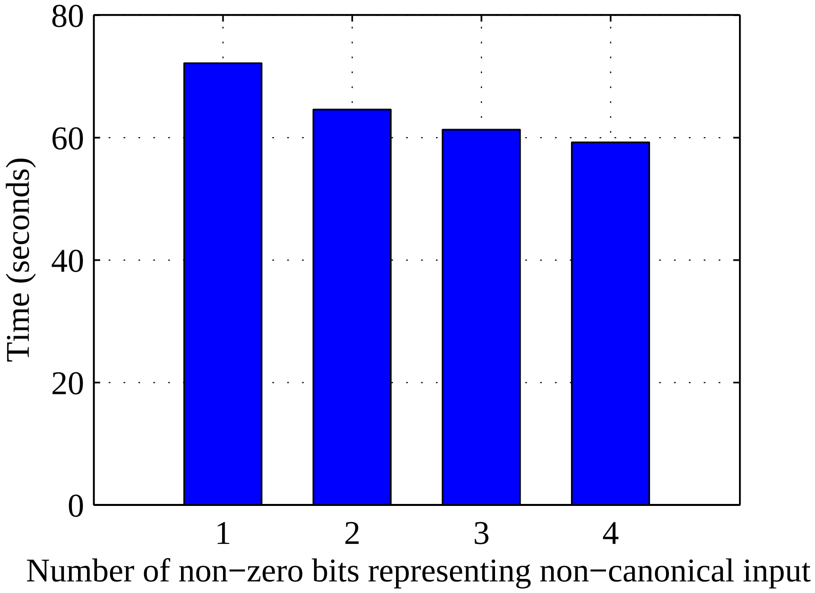

Figure 9.3: Time Result for Experiments 6: Time taken for the output error rate for non-canonical input representations to reach zero error is plotted against the corresponding respond density of input vector X. Bi has a response density i, i=1, 2, 3, 4. The canonical representation has a response density 1. B1 is the canonical while B2, B3, B4 are non-canonical representations of the words with 2,3 and 4 bits ‘on’ in the input vector

Input: Words (in the sequence in which they would appear in the sentence) and corresponding classes if any.

Output: Sequence of words learned and new sentences created through generalization of classes.

Example (Diagrammatically represented in Fig. 8.4):

Table 9.4: Training Sentence 1. Member to Class Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Apple	is	a	fruit
Output	Apple	Apple-is	Apple-is-a	Apple-is-a-fruit

Table 9.5: Training Sentence 2. Member to Class Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Apple	can	be	eaten
Output	Apple Fruit	Apple-can Fruit-can	Apple-can-be Fruit-can-be	Apple-can-be-eaten Fruit-can-be-eaten

9.2.4 Member to Member Generalization

The network was able to successfully reach a 100% detection.

Total number of input sentences = 308. Total distinct words in the sentences in the training set = 892. Total number of states learnt = 1631.

Input: Words (in the sequence in which they would appear in the sentence) and corresponding classes.

Output: Sequence of words learned and new sentences created through member to member generalization.

Example (Diagrammatically represented in Fig. 8.9):

Table 9.6: Training Sentence 1. Member to Member Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Cuckoo	is	a	bird
Output	Cuckoo	Cuckoo-is	Cuckoo-is-a	Cuckoo-is-a-bird

Table 9.7: Training Sentence 2. Member to Member Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Sparrow	is	a	bird
Output	Sparrow	Sparrow-is	Sparrow-is-a	Sparrow-is-a-bird

Table 9.8: Training Sentence 3. Member to Member Generalization

Time Frame	t_1	t_2
Input	Sparrow	fly
Output	Sparrow Bird	Sparrow-fly Bird-fly

9.2.5 Subclass to Superclass Generalization

The network was found to be able to successfully associate the given objects to the parent class of their subclass, without confusing the members of one subclass with another within 2 epochs. Each “member to class” relationship is explicitly taught

Table 9.9: Testing Sentence. Member to Member Generalization

Time Frame	t_1	t_2
Input	Cuckoo	fly
Output	Cuckoo Bird	Cuckoo-fly Bird-fly

to the network. Hence the network is taught both *Apple is a type of fruit* and *Fruit can be of type apple*. With ample experience the network would be able to create "member to class" and corresponding "class to member" relationships on its own.

Input: Words (in the sequence in which they would appear in the sentence) and corresponding classes.

Output: Sequence of words learned and new sentences created through subclass to superclass generalization.

Example (Diagrammatically represented in Fig. 8.10):

Table 9.10: Training Sentence 1. Subclass to Superclass Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Girl	is	a	human
Output	Girl	Girl-is	Girl-is-a	Girl-is-a-human

Table 9.11: Training Sentence 2. Subclass to Superclass Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Rachel	is	a	girl
Output	Rachel	Rachel-is	Rachel-is-a	Rachel-is-a-girl

Table 9.12: Testing Sentence. Subclass to Superclass Generalization

Time Frame	t_1	t_2	t_3	t_4
Input	Rachel	is	a	girl
Output	Rachel	Rachel-is	Rachel-is-a	Rachel-is-a-girl Rachel-is-a-human

9.2.6 Classification from Similarity

Experiment 5 tested if the network could identify members of the same class based on their features, without confusing the members of another class to be similar. The training sentences for the experiment consist of "is a" statements that allow the network to partition the objects into their perspective classes and the property training statements, that consist of 3 words, the object, property and whether the object has the given property, this is stated by "yes" or "no". The network is then given the *unseen* object and the property as inputs and tested by allowing it to predict the output of the 3^{rd} and the last time frame. The network predicts if the unseen member of the class has a certain property, by answering "yes" or "no". The network was able to map the correct property to the correct object and hence classify based on similarity of features. 100% performance was reached in 3 epochs.

Input: Words (in the sequence in which they would appear in the sentence) and corresponding classes.

Output: Sequence of words learned and new sentences created through member to member generalization.

Example:

Table 9.13: Training Sentence 1. Classification from Similarity

Time Frame	t_1	t_2	t_3	t_4
Input	Kiwi	is	a	weakwingedbird
Output	Kiwi	Kiwi-is	Kiwi-is-a	Kiwi-is-a-weakwingedbird

Table 9.14: Training Sentence 2. Classification from Similarity

Time Frame	t_1	t_2	t_3	t_4
Input	Sparrow	is	a	strongwingedbird
Output	Sparrow	Sparrow-is	Sparrow-is-a	Sparrow-is-a-strongwingedbird

Table 9.15: Training Sentence 3. Classification from Similarity

Time Frame	t_1	t_2	t_3	t_4
Input	Ostrich	is	a	weakwingedbird
Output	Ostrich	Ostrich-is	Ostrich-is-a	Ostrich-is-a-weakwingedbird

Table 9.16: Training Sentence 4. Classification from Similarity

Time Frame	t_1	t_2	t_3	t_4
Input	Cuckoo	is	a	strongwingedbird
Output	Cuckoo	Cuckoo-is	Cuckoo-is-a	Cuckoo-is-a-strongwingedbird

In the testing phase the input at time frame 3 is left blank as we let the network predict the output based on the inputs at the previous time frames.

Table 9.17: Training Sentence 5. Classification from Similarity

Time Frame	t_1	t_2	t_3
Input	Sparrow	flies	yes
Output	Sparrow	Sparrow-flies	yes

Table 9.18: Training Sentence 6. Classification from Similarity

Time Frame	t_1	t_2	t_3
Input	Kiwi	flies	no
Output	Kiwi	Kiwi-flies	no

Table 9.19: Testing Sentence 1. Classification from Similarity

Time Frame	t_1	t_2	t_3
Input	Cuckoo	flies	-
Output	Cuckoo strongwingedbird	Cuckoo-flies strongwingedbird-flies	yes yes

Table 9.20: Testing Sentence 2. Classification from Similarity

Time Frame	t_1	t_2	t_3
Input	Ostrich	flies	-
Output	Ostrich weakwingedbird	Ostrich-flies weakwingedbird-flies	no no

Graph in Fig. 9.4 plots the results of the experiments.

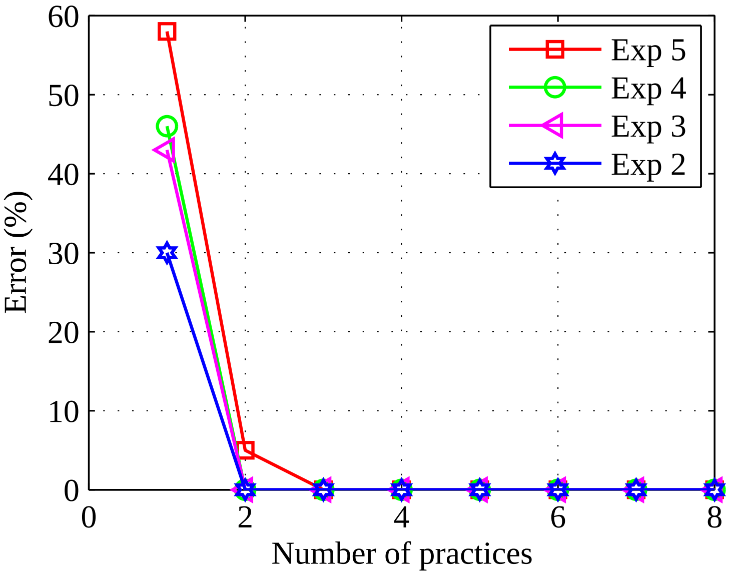

Figure 9.4: Results of Experiments 2 to 5: Total number of states is 1631. The network is able to reach 100% recognition rate within 2 epochs of training in all the experiments

9.3 Visualization of Layers X, Y and Z

The figure below shows visualization for areas x, y, z_1 and z_2 for 3 training sentences for experiment 1, *parroting.* The network is trained on 3 "type" motors, namely, bold, italic or simple words; if a word is not bold or italic it is a considered to be a simple word by default. The network is tested on the training sample itself. The input as discussed earlier is the sequence of words as they appear in a sentence. Layer x in the figure consists of words and is not shown as a matrix representation for the ease of understanding. Layer y, z_1 and z_2 on the other hand are the color-coded visual representations of the corresponding matrices. The network follows a winner-takes-all policy, the winner neuron is colored grey, while the non-firing neurons are represented by white squares. Layer z_1 represents the type motor while layer z_2 represents the structure motor.

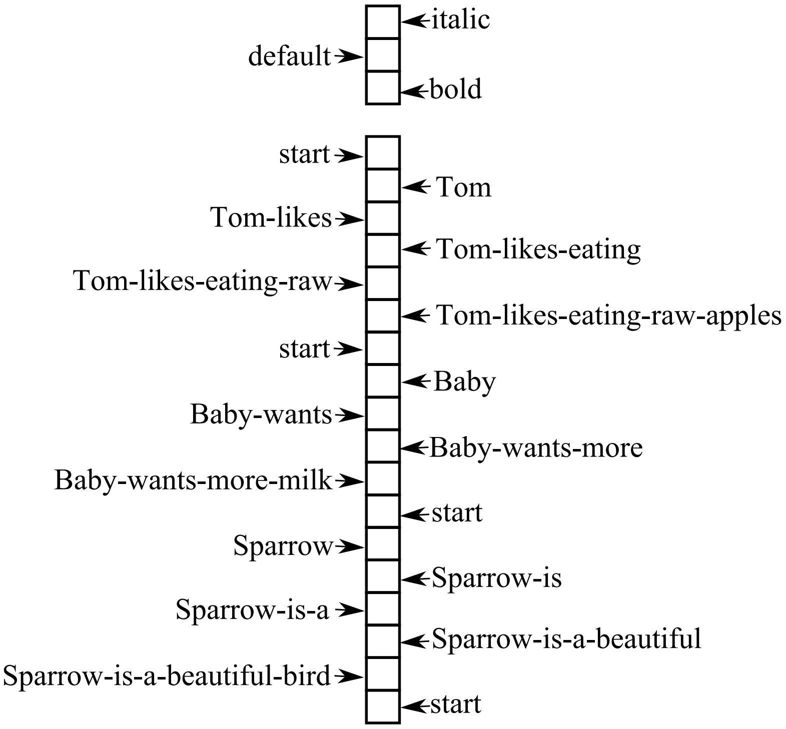

Figure 9.5: Visual representation of layers z_1 and z_2, each square represents a motor concept that is marked by an arrow next to it

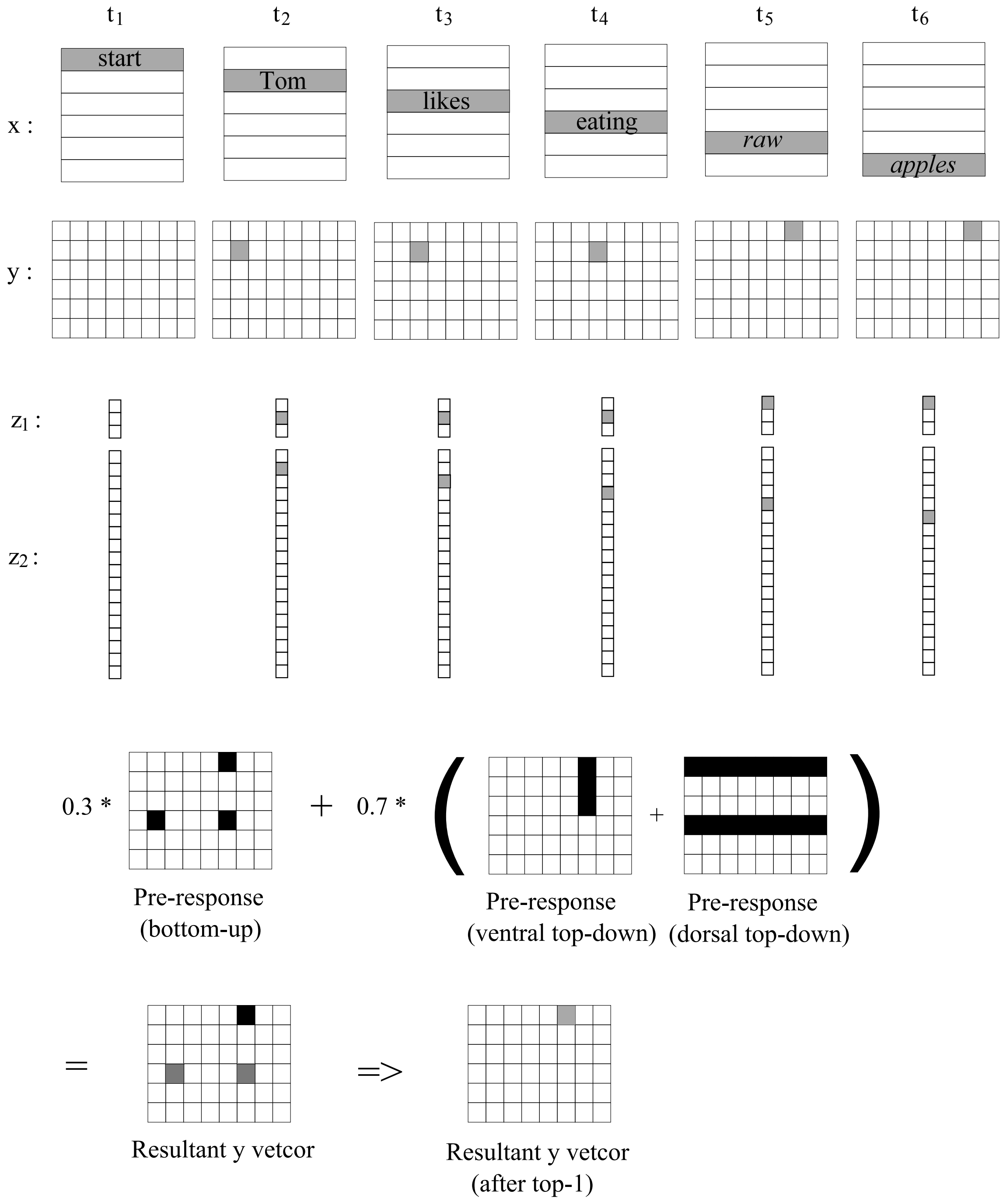

Figure 9.6: Learning Sentence 1: "Tom likes eating *raw apples*". Layer z and the outputs can be interpreted from the key in Fig. 9.5

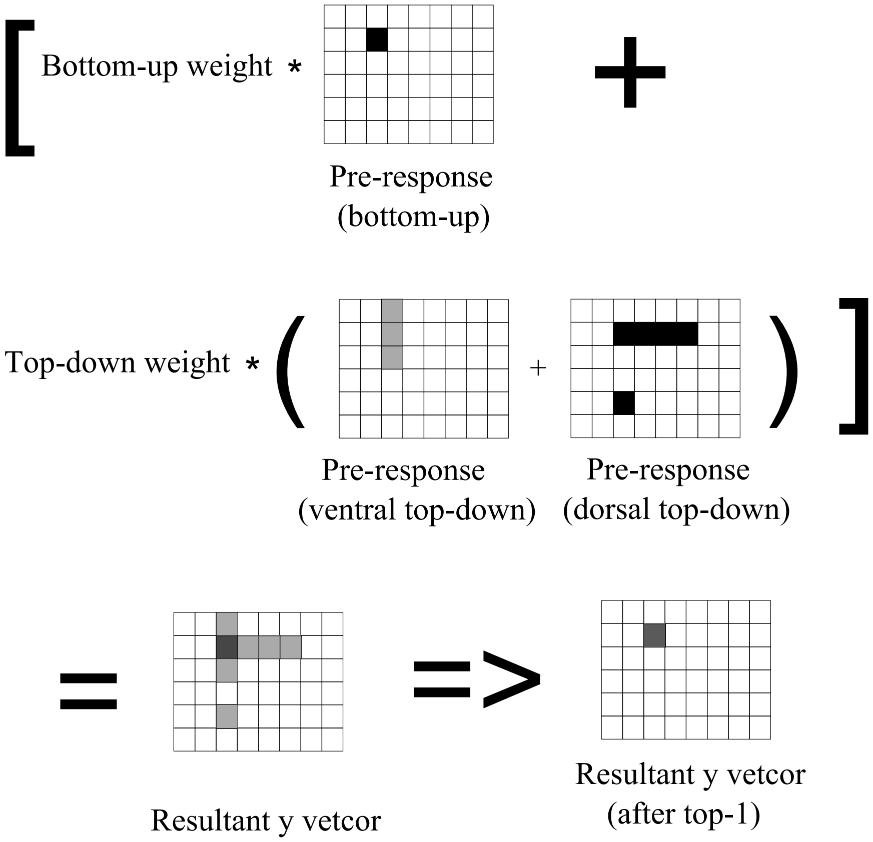

Figure 9.7: Visualization for Pre-response vector while learning the word “eating” in the sentence “Tom likes eating *raw apples*”. The top-down and bottom-up inputs are shown to result in the evolution of the the Y layer of which only 1 neuron having the highest response value is chosen to be the winner.

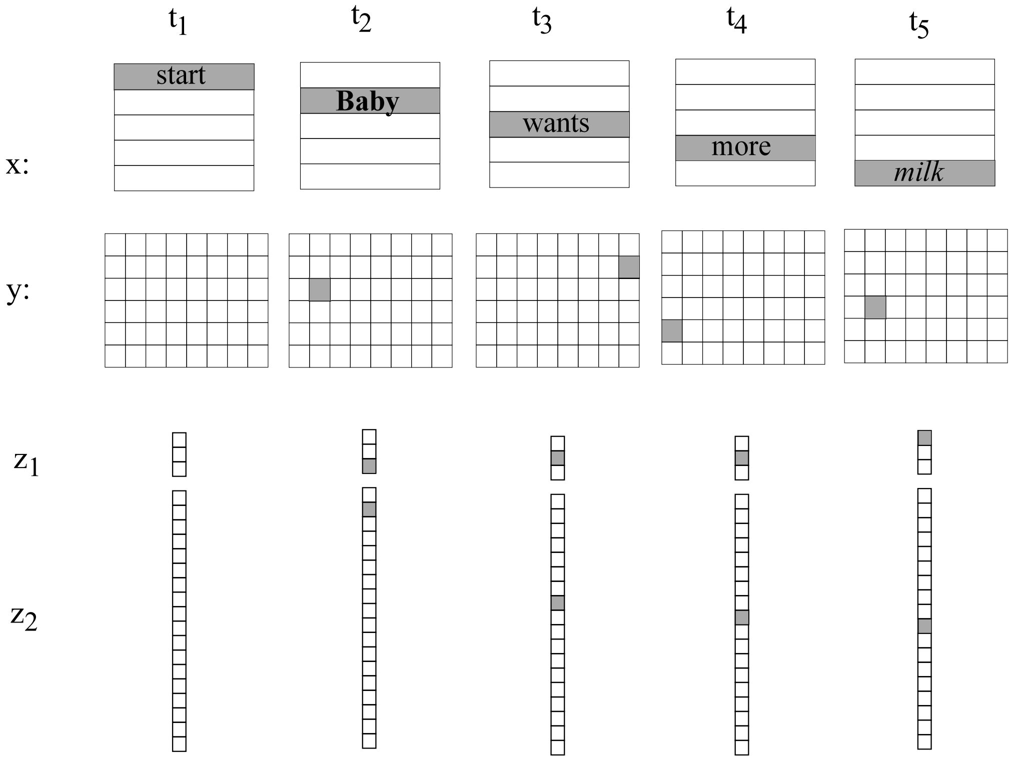

Figure 9.8: Learning Sentence 2: "**Baby** wants more *milk*". Layer z and the outputs can be interpreted from the key in Fig. 9.5

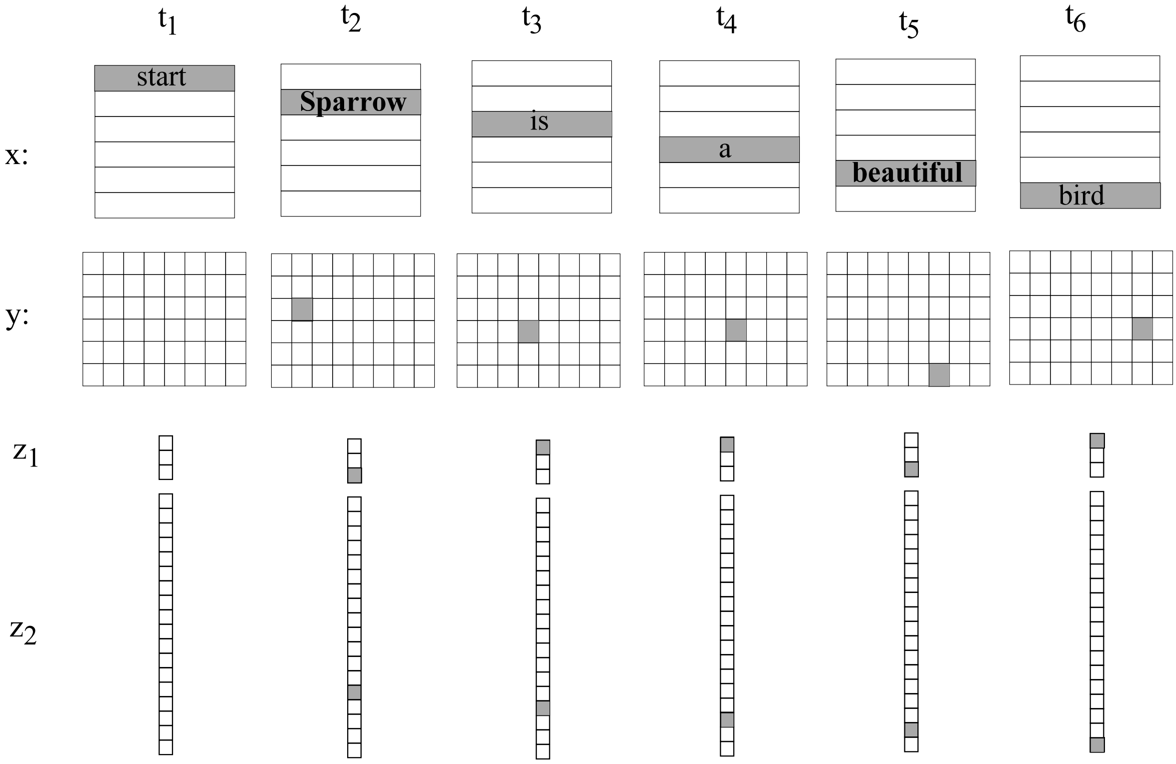

Figure 9.9: Learning Sentence 3: “**Sparrow** is a **beautiful** bird”. Layer z and the outputs can be interpreted from the key in Fig. 9.5

Chapter 10

Novelty and Contributions

The main novelties of the model are enumerated below:

1. Multi Layer In-place Learning (MILN) has been used for LA problem. This is the first computational model for language acquisition. Using multiple motor areas, the network demonstrates early language acquisition with the help of a neuromorphic, developmental, emergent general-purpose model.

2. The network simulates "where" and "what" concepts for language acquisition (LA). Thus creating a grounded model of language. One must remember that language does not only provoke vocal response, instead language could also invoke a physical response like a hand gesture. For e.g., if someone asks one to pass the salt on the table, the response of the person is not "yes, I will pass you the salt" instead if the person does intend to pass the salt he/she might merely reach out and pass the salt.

   The network's "what" motor is assisted by "where" motor as shown in the graph in Fig. 9.1.

3. Association based reasoning or generalization helps the network to create new semantics. The network uses classical conditioning methods for language skill transfer. Language is learned through the same principles as other physical

activities. It does not use hand-crafted language structure but allows primary and secondary associations, as seen in animal learning.

The network does not use logical reasoning instead uses generalization to create new sentences and reasoning, thereby broadening its own knowledge base. These new sentences might or might not be logically correct depending on the prior knowledge of the network. Thus the network is able to achieve better relation specificity. The following generalization methods help us model relationships between concepts:

- Member to class generalization
- Subclass to superclass generalization
- Member to member generalization

4. WWN does not treat syntax and semantics separately. The network is a general model of internal representation that integrates syntax and semantics, or concepts in general.

Chapter 11

Discussion

The model can be used not only for written language learning but for learning through audition as well as structural learning will remain the same. Furthermore, the model does not take a bag of words approach but is able to identify phrases that make complete sense versus those who don't. The network learns sequential word association, it is able to generalize correctly and hence create its own sentences. The network can also choose between multiple generalizations on the basis of what partition objects have the properties closest to the object in question.

11.1 Language Processing Based On Grammar

Language is not merely a mesh of words weaved together instead it is governed by rules that tie up the vocabulary called the grammar or syntax. Grammar aids language processing. It makes the language more structurally sound allowing computational linguists to take a statistical approach to solving problems of extracting information from speech and written text.

The paper mainly tackles *language understanding* and not exactly *language processing*. Language processing results only provide superficial solutions to linguistic problems. As explained earlier they are The machine might pretend to know what is being said but will have no knowledge of the actual meaning or context of the conversation to be able to be of real help as it has no understanding of actual language. Due to the above issue “semantics” becomes very important. In fact semantics is one of the main keys to the in-depth understanding of the language along with the other components like association and grounding. Through semantics we can actually understand the meaning while with the help of association and grounding we can find the unambiguous context to be able to react perfectly to the situation.

11.2 Neural networks for language processing

Many neural networks have tried to solve the language processing problem yet their complex nature seems to be overwhelming to most psycholinguists. Many have accused ANNs to be “black-boxes”. Further ANNs are also criticized in literature for being “cognitively implausible,” and failing to “capture generalizations”. Velde and Kamps in [17] have tried to model similar features; they call them productivity, dynamics (learning while training) and grounding. They have a strong model that divide a sentence into its grammatical constituents along with an ”agent” who causes something to happen and a ”theme” that is usually what is affected by the agent or its action. The network tries to deduce things by the agent and theme interactions along with the grammatical phrases they appear in. The aim of the paper is very similar to ours but its means are quite different. The main difference between Velde and Kamps’ feed forward network and our model is that they consistently use grammatical tags (nouns, pronouns etc.) to understand a sentence, which is very much like concepts known to the network. Our model on the other hand knows only two concepts; location and type, but type can represent any concept (e.g., noun, verb, noun phrase, verb phrase, etc). However, our model works for earlier language acquisition where the child has not learned any explicit formal grammar.

N-gram models are very commonly used for statistical modeling. They focus on short-length predictions of sequences, be it phonemes, alphabets or words though powerful they lack versatility as they cannot be used to model sentences of more than a certain length. Chelba and Jelinek's structured language model (SLM) aims to resolve this shortcoming [7]. The model uses a parser to create syntactic word-parse k-prefix of the word string to predict the next word and its POS tag while a constructor builds a binary branching structure of the sentence. SLM is thus able to capture long dependencies. SLM was first used for speech processing. Both the above methods use syntactic modeling. The neural network described in [14] uses SLM in a batch fashion to create a language predictor. This is very similar to the prediction. Further it uses error back-propagation method to create a recurrent network.

11.3 Language Acquisition

Language acquisition is a grounded approach to language processing. Unlike NLP, where a human programmer entrusted with the task of designing the system hand-crafts each state and the outcome of a transition, in language acquisition, the system learns to device these transitions on its own by learning them autonomously from its surrounding environment that might or might not have a human teacher. Our method uses the latter to develop an autonomous language learner, it is unique in the sense that it is the first where-what network for language acquisition that takes visual word input in order to produce the correct action, which might include various language processing tasks, like, part-of-speech tagging, text segmentation recognizing syntactic ambiguity etc. The network need not learn everything before it starts performing, but instead should learn dynamically so that it can be corrected early if it learns some wrong information. It is all the more important as the network is not taught everything explicitly but instead draws associations and conclusions from what it has learnt so it becomes imperative that if we come across any wrong information learnt by the system, we correct it, just as small children are corrected by their teachers/parents if they say/do something wrong. Our network is an *incremental* learner that *learns as it goes* focusing mainly on language *understanding*. Hence unlike other systems that have to be trained or programmed before they can do anything, our network not only learns what it is taught but also learns as it is taught. The network can incrementally pick-up new tricks as it lives on and so it *grows stronger as it lives longer.*

11.4 Future work

A few speculations could be made about the context and attention, context helps attention and attention helps balancing generalization and discrimination. The ventral motor pathway could be used to develop context such that if the network is taught *dog is an animal* and *sparrow is bird*, then "bird", "animal" and "not bird", "not animal" could be higher level concepts. The multiple levels of generalization could help the network to focus on the correct class. Generalization could be further used to fine tune our results through multiple muxel priming, as noted by [55]. We can introduce a Pre-TM layer, between the V2 and TM layers, which is a part of the motor hierarchy. The Pre-TM helps generalization when more than one motor neuron primes on at one time, both the object as well as its class fire together each time any of the object primes in Pre-TM, so that the network develops a concept of the class. Hence the class and object do not compete at the Pre-TM layer and are counted as the same, but they are perceived as individuals in the motor layer. Thus we are able to strike a balance between generalization and discrimination.

But there are a few questions that still need to be answered. How fine or coarse should the generalization be so that the network is able to create correct classification model? Though for now the teacher decides the value of k i.e., how many neurons should win in Layer Y, but ideally after ample experience the network should be able to decide it on its own.

Chapter 12

Conclusion

Although there are symbolic systems that model language acquisition [34], our system appears the first recurrent connectionist model for language acquisition without using any handcrafted internal representation. For e.g., traditional NLP systems require the human programmer to handcraft a static vocabulary and hand designate a word to each Hidden Markov Model (HMM). Further how such HMMs link with others are also handcrafted. In contrast, our network fully automatically develops all such Wirings and strengths through weight adaptation. In this sense, this seems the first truly "autonomous" developer for language acquisition in the sense that internal self-organization is fully autonomous after the "birth".

Comparing the open-style connectionist language networks (e.g., those used by Rogers and McClelland [46]) and many open-style symbolic network (e.g., [34]), the most obvious characteristic of our architecture is that the network is highly recurrent between the internal layer and the motor layers. While some modelers turned off recurrence during learning of their recurrent networks [10, 45], the major reason for us to succeed in dealing with such a high degree recurrence during learning was because of the series of cortex-like mechanisms of LCA [26]. The network is still at a very nascent stage. It is needless to say that it is far from reaching its potential in terms of richness, complexity and scale yet it does try to open new avenues by modeling a cortex like robust and efficient network and giving acceptable results. Language acquisition is not a trivial task but there are a lot of psychological motifs behind it, by studying language acquisition more we might finally understand concept about intelligence and thinking.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Allen R.B., Sequential connectionist networks for answering simple questions about a microworld. Proceedings of the 10th Annual Conference of the Cognitive Science Society, Montreal (1988).

[2] Allyssa McCabe, Language Games to be Played with Your Child. New York: Insight Books. (1992).

[3] Arkin R.C., Behavior-Based Robotics. MIT Press, Cambridge, Massachusetts (1998).

[4] Barsalou LW, Simmons WK, Barbey AK, Wilson CD. Grounding conceptual knowledge in modality-specific systems. Trends Cogn Sci. 7:8491 (2003).

[5] Benello J., Machie A.W., Anderson J.A., Syntactic category disambiguation with neural networks. Computer Speech and Language, p.203-217 (1989).

[6] Blanton, Richard L., Sensory Discrimination, Generalization and Language Training of Autistic Children, (1984).

[7] Chelba C., Jelinek F., Structured language modeling. Computer Speech and Language, 14:4, 283332. (2000).

[8] Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, Cambridge (1965).

[9] Cruse H., Neural Networks as Cybernetic Systems. Brains, Minds, Vol. 2, bmm615 (2006).

[10] Deco G., Rolls E.T., A Neurodynamical cortical model of visual attention and invariant object recognition. Vision Research, Vol. 40, p.2845-2859 (2004).

[11] DeJong, G.F., Skimming newspaper stories by computer, Technical Report 104, Yale University Department of Computer Science. (1977).

[12] Domjan M., The Principles of Learning and Behavior, 4th ed. Belmont, CA: Brooks/Cole, (1998).

[13] Elman, J.L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., Rethinking innateness: A connectionist perspective on development. MIT Press, Cambridge MA (1996).

[14] Emami A., Jelinek F., A Neural Syntactic Language Mode, Machine Learning, Vol.60, p. 195-227. Sept. (2005).

[15] Evans R. and Jones D., Metacognitive Approaches to developing Oracy, Developing Speaking and Listening with young children, (2009).

[16] Ferguson CA, Farwell CB. Words and sounds in early language acquisition. Language. 51:419439 (1975).

[17] Frank van der Velde and Marc de Kamps, A neural architecture for grounded cognition: Representation Structure Dynamics and Learning Proceedings of IJCNN2008 (WCCI2008), Hong-Kong (2008).

[18] Grossberg S., Raizada R., Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. Vision Research, 40:14131432, (2000).

[19] Gomez, R. L., Variability and detection of invariant structure. Psychological Science, 13(5), p.431436 (2002).

[20] Harris J., Early Language Development, Implications for Clinical and Educational Practice. Routledge, London (1990).

[21] Harnad S., The induction and representation of categories (1987).

[22] Harnad S., Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54, (1991).

[23] Hinton G.E., Implementing semantic netowrks in parallel hardware. In G.E. Hintons and J.A. Anderson (Eds) Parallel Models of Associative Memory, Hillsdale, NJ: Lawrence Erlbaum (1981).

[24] Hickok G., Poeppel D., Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition. 92:67–99, (2004).

[25] Iverson J.M., Developing language in a developing body: the relationship between motor development and language development. In Journal of child language. Vol. 37, No. 2, p.229-261 (2010).

[26] Ji Z., Weng J., Prokhorov D., Where-What Network 1: "Where" and "What" Assist Each Other Through Top-down Connections. Proc. IEEE International Conference on Development and Learning, Monterey, CA, p.61-63, Aug. 9-12 (2008).

[27] Ji Z., Weng J., WWN-2: A Biologically Inspired Neural Network for Concurrent Visual Attention and Recognition. Proc. IEEE International Joint Conference on Neural Networks, Barcelona, Spain, July 18-23, p.+1-8, (2010).

[28] Ji Z., Weng J., Prokhorov D., Where-What Network 1: "Where" and "What" Assist Each Other Through Top-down Connections, Proc. IEEE International Conference on Development and Learning, Monterey, CA, Aug. 9-12, p. 61-66 (2008).

[29] Joshua Knobe. Intentional action and side effects in ordinary language. Analysis, 63:190193, 2003.

[30] Laird J.E., Newell A., Rosenbloom P.S., Soar: An architecture for general intelligence. Artificial Intelligence, 33:164, (1987).

[31] Lakoff G., Johnson M., Metaphors We Live by. University of Chicago Press (1980).

[32] Lange T.E., Dyer M.G., High level inferencing in a connectionist network. Connection Science, p.181-217 (1989).

[33] Lenat D.B., CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):3338, (1995).

[34] Levinson S.E., Liu Q., Dodsworth C., Lin R., Zhu W., Kleffner M., The Role of Sensorimotor Function, Associative Memory and Reinforcement Learning in Automatic Acquisition of Spoken Language by an Autonomous Robot. Proc. of Workshop on Development and Learning, East Lansing, Michigan, p.95-100, April 5-7 (1999).

[35] Locke J., Essay Concerning Human Understanding (1689).

[36] Luciw M., Weng J., Where-What Network 3: Developmental Top-Down Attention for Multiple Foregrounds and Complex Backgrounds. International Joint Conference on Neural Networks, July 18-23, Barcelona, Spain, pp. +1-8, (2010).

[37] Mavridis N., Roy D., Grounded Situation Models for Robots: Where words and percepts meet, IEEE IROS (2006).

[38] McCarthy J., Programs with Common Sense. Proceedings of the Teddington Conference on the Mechanization of Thought Processes (1958).

[39] Minsky M., Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI Magazine, 12(2):3451, (1991).

[40] Mishkin M., Unterleider L.G., Macko K.A., Object Vision and Space Vision: Two Cortical Pathways. Trends in Neuroscicence, Vol. 6, p. 414-417 (1983).

[41] Newell A., Unified Theories of Cognition. Harvard University Press, Cambridge, Massachusetts, (1990).

[42] Olga Parsons , Gail A. Carpenter, ARTMAP neural networks for information fusion and data mining: map production and target recognition methodologies. Neural Networks, v.16 n.7, p.1075-1089 (2003).

[43] Piaget J., The Origins of Intelligence in Children. International Universities Press, Madison, New York (1952).

[44] Price C.J., The anatomy of language: contributions from functional neuroimaging. J Anat. 197(Pt 3):335–359, (2000).

[45] Roelfsema P.R., van Ooyen A., Attention-Gated Reinforcement Learning of Internal Representations for Classification. Neural Computation, Vol. 17, p.2176-2214 (2005).

[46] Rogers T.T., McClelland J.L., Precis of Semantic Cognition: A Parallel Distributed Processing Approach. Behavioral and Brain Sciences, Vol. 31, p.689-749 (2008).

[47] Rumelhart D.E., McClelland J.A., Parallel Distributed Processing - Explorations of the Microstructure of Cognition, The MIT Press, Cambridge, MA (1986).

[48] Sharkey N.E., Implementing soft preferences for structural disambiguation. KONNAI. Journal of Psycholinguistic Research, Vol. 23, No. 4, p.295-322 (1990).

[49] Smolensky P, On variable binding and the representation of symbolic structures in connectionist systems. Tech Report CU-CS-355-87, Dept of Computer Science, University of Colorado, Boulder, CO (1987).

[50] Stemmer, Nathan. The Role of Innate and Acquired Generalization Classes in Language Acquisition. Paper presented at the Interdisciplinary Conference, "Perspectives on Language", University of Louisville, May 6-8 (1976).

[51] Weng J., Luciw M.D., Optimal In-Place Self-Organization for Cortical Development: Limited Cells, Sparse Coding and Cortical Topography, Proc. 5th International Conference on Development and Learning (ICDL'06), Bloomington, IN, May 31 - June 3, p.+1-7, (2006).

[52] Weng J., Luciw M., Dually optimal neuronal layers: Lobe component analysis. IEEE Trans. Autonomous Mental Development, 1(1):6885, (2009).

[53] Weng J., A 5-Chunk Developmental Brain-Mind Network Model for Multiple Events in Complex Backgrounds International Joint Conference on Neural Networks, July 18-23, Barcelona, Spain, p. +1-8, (2010).

[54] Weng J., Zhang Q., Chi M., and Xue X., Complex Text Processing by the Temporal Context Machines. Proc. IEEE 8th International Conference on Development and Learning, Shanghai, China, pp. 1-8, June 4-7, 2009.

[55] Weng J., Zhang Q., Chi M., Xue X., Complex Text Processing by the Temporal Context Machines. Proceedings of the 10th Annual Conference of the Cognitive Science Society, Montreal (1988).IEEE 8th International Conference on Development and Learning, Shanghai, China, June 4-7 (2009).

[56] Yeung H.H., Werker J.F., 'Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. Cognition, Vol. 113, No. 2, p.234-243 (2009).

[57] Zhang Y., Weng J., Task Transfer by a Developmental Robot. IEEE Transactions on Evolutionary Computation, Vol. 11, No. 2, p.226-248, April (2007).

[58] Zwaan R.A., Radvansky G.A., Situation Models in Language Comprehension and Memory. Psychological Bulletin, Vol 123, No. 2, p.162-185 (1998).