

NEURAL NETWORKS VS GENERATIVE LINGUISTICS

JOE PATER

BY: RITA AMAMOO-OCANSEY

PAPER OUTLINE

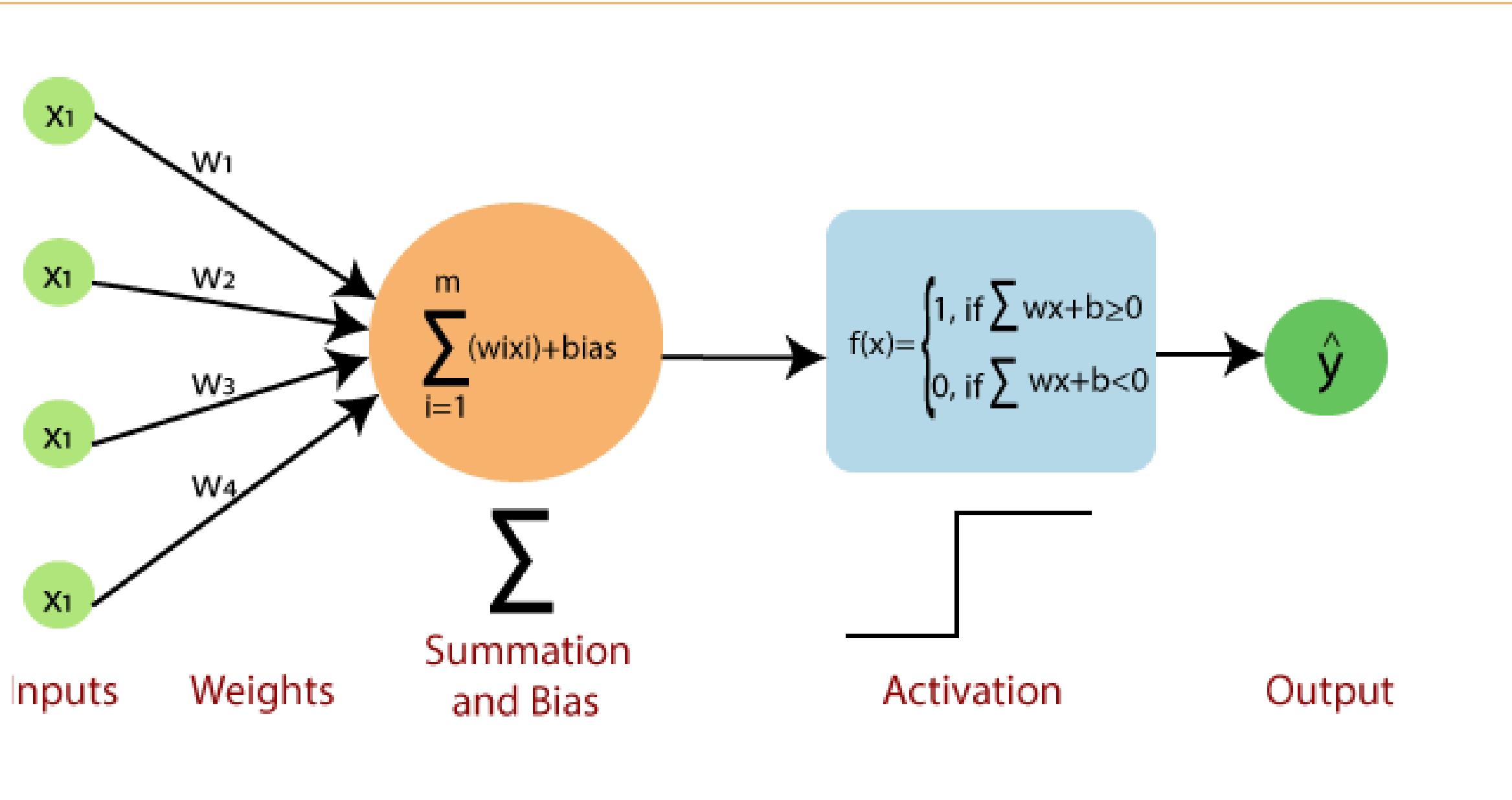
- Section 1 gives general overview and background of authors (introduction).
- Section 2 discusses how Chomsky's and Rosenblatt's proposals each diverged from 'mainstream AI'.
- Section 3 of the article discusses some of that debate and argues that it produced important lessons for future research in both traditions, rather than ending in victory for one or the other side.
- The final section of the article surveys some of the subsequent research over the last thirty years that has integrated aspects of neural network modeling research and generative linguistics.
- **Main Argument:** “I argue that progress on a core goal of generative linguistics, the development of a theory of learning, may well be aided by its integration with neural modeling.”

- Both were born in 1957 by Noam Chomsky and Frank Rosenblatt respectively
- Chomsky took a ‘high-level’ cognitive phenomenon—language, and in particular, syntax—and aimed to show that some reasonably powerful computational machinery was not up to the task of representing it, before going on to propose a more powerful theory that could.
 - Rosenblatt took some very simple computational machinery—mathematical analogues of neural activation and synaptic connections—and aimed to show that it could represent ‘low-level’ cognitive processes involved in object perception and recognition, and that these representations could be learned algorithmically.
 - These two ideas developed separately until later when Rumelhart and McClelland (1986) developed a perceptron-based, or connectionist, model of past-tense formation in English,

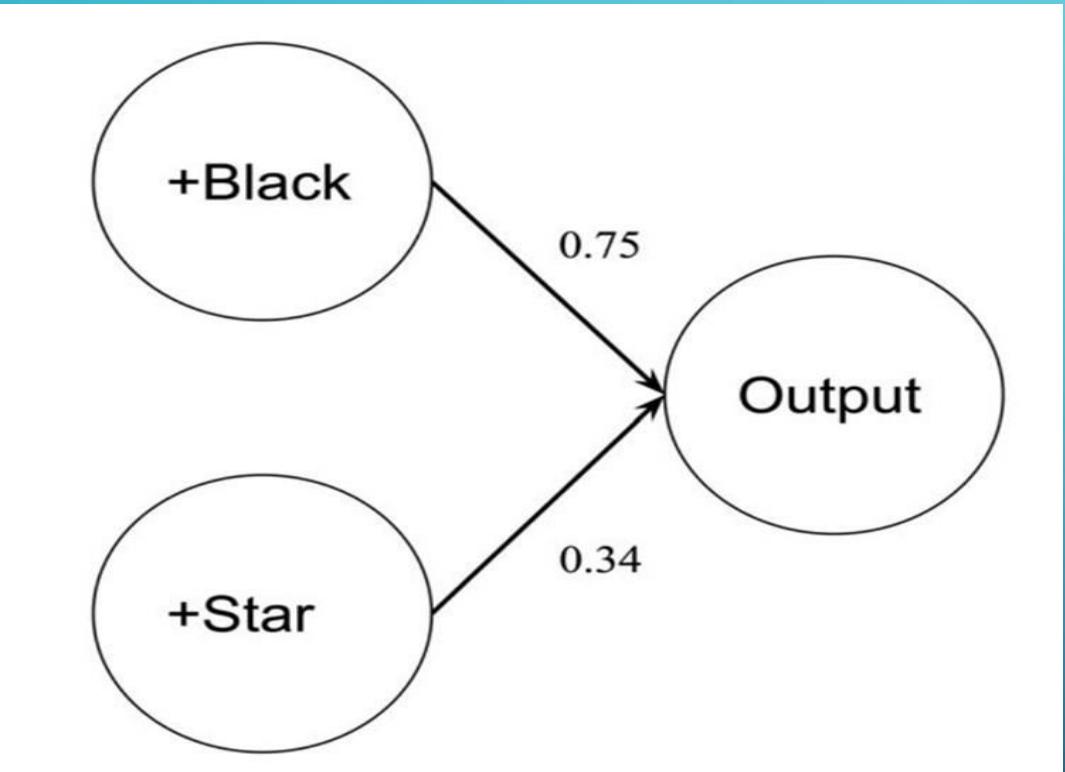


NEURAL NETWORKS (PERCEPTRON)

- The activity of a neuron—also called a node or a unit—is represented as a numerical value, often as 1 or 0, *on* or *off*.
- This activity is passed along synaptic connections to other neurons. The connections are weighted: each one has a real valued number that is multiplied by the signal it receives from an input node.
- A given node becomes active when the sum of incoming weighted signals exceeds a designated threshold

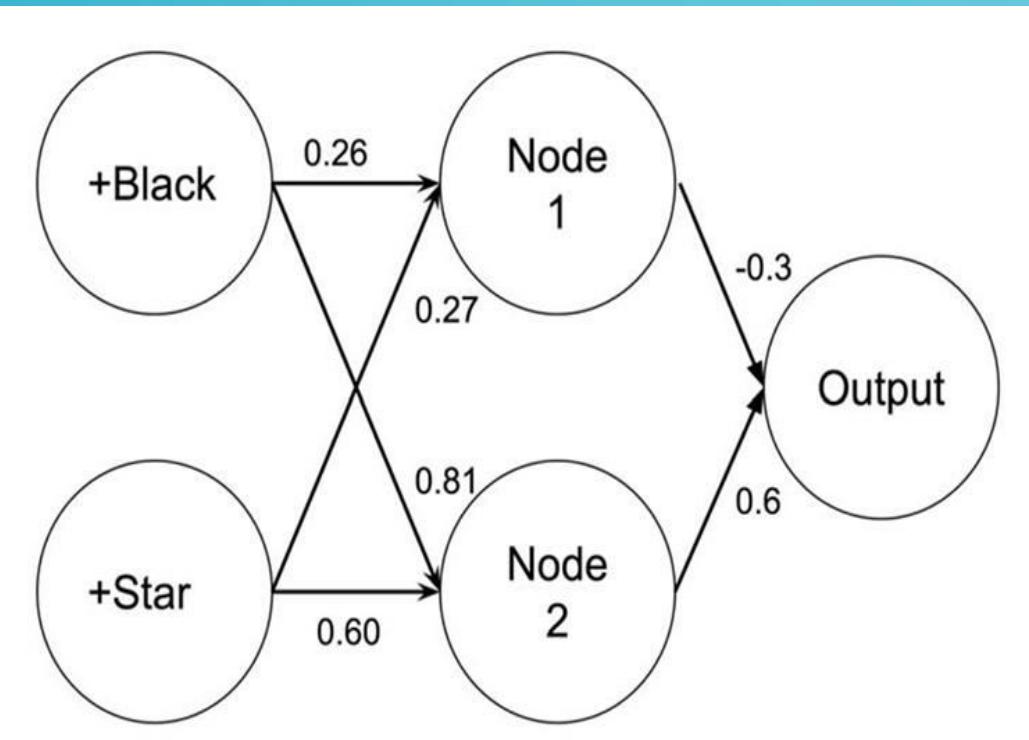


SINGLE LAYER PERCEPTRON



INPUT	+Black	+Star	WEIGHTED SUM	ACTIVATION (> 0.5 input)
★	1	1	1.09	1
☆	0	1	0.34	0
◆	1	0	0.75	1
◇	0	0	0	0

MULTILAYER PERCEPTRON



INPUT	+Black	+Star	WEIGHTED SUM	ACTIVATION (> 0.5)
★	0.26 0.81	0.27 0.60	0.53 1.41	1 1
☆	1 0	1 0.60	1.41 0.60	1 1
◆	0.27 0.60	0.26 0.81	0.53 1.41	0 1
◊	0 0	0 0	0 0	0 0

INPUT	NODE 1	NODE 2	WEIGHTED SUM	ACTIVATION (> 0.5)
★	-0.3	0.6	0.3	0
☆	1	1	0.6	1
◆	0	1	0.6	1
◇	0	0	0	0

Multilayer perceptron part 2: hidden layer to output.

GENERATIVE LINGUISTICS

- “The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences of L and to study the structure of the grammatical sequences.”
- Temporal and Spatial feature relationship.

Eg: *The lion sleeps*

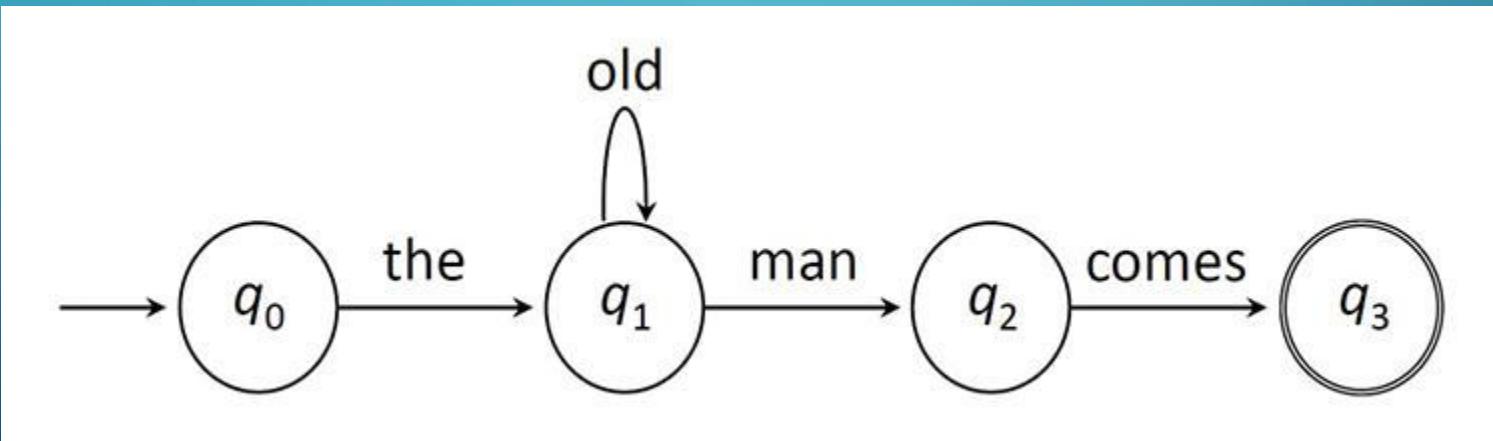
* *Sleeps lion the is not*

- There are also dependencies between the forms of items that occur at different points in time

Eg: *The lion sleeps* and *The lions sleep* vs. **The lion sleep* and **The lions sleeps*.

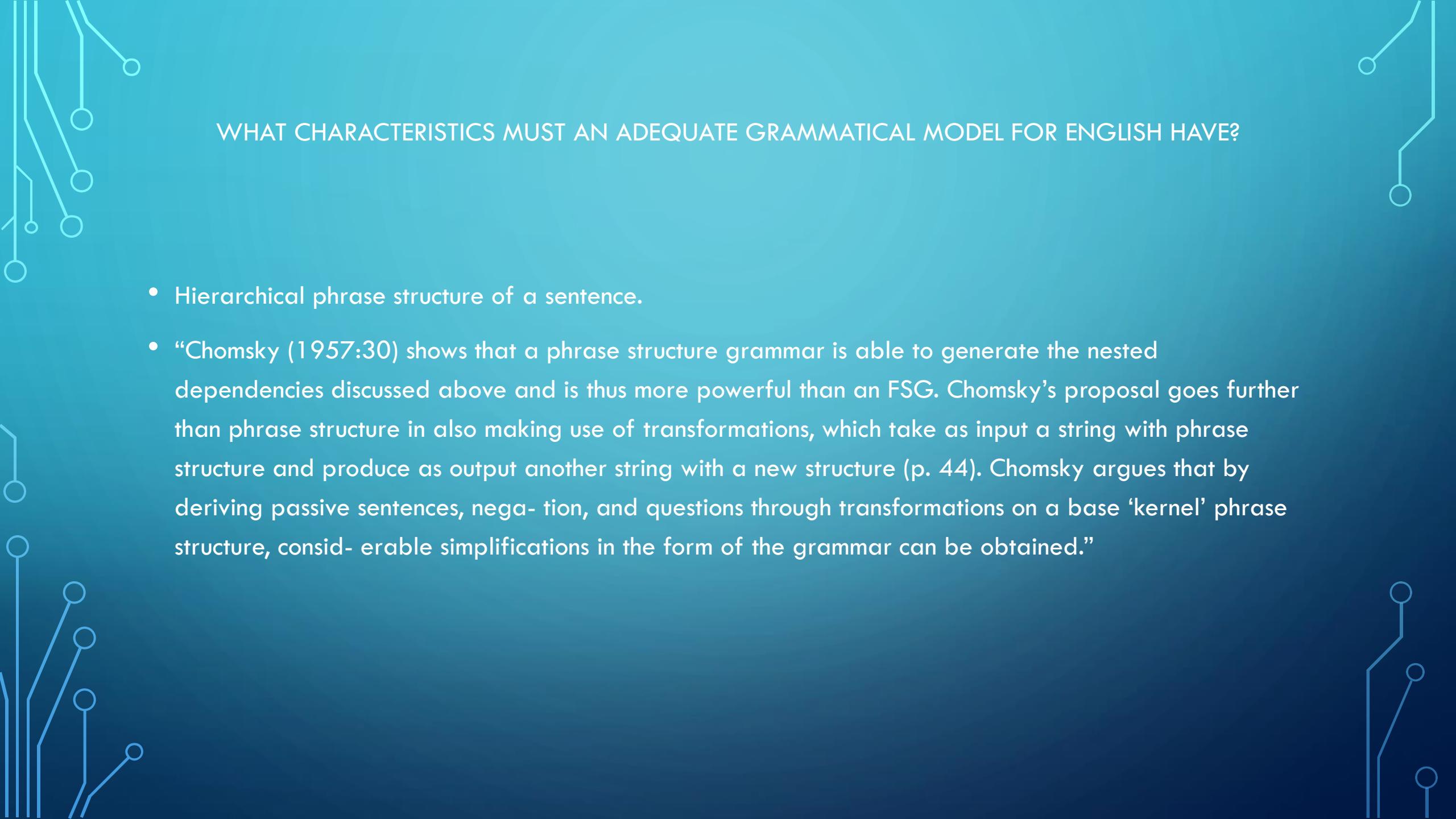
Encoding Language

- Grammatical and ungrammatical sentences can be encoded in terms of allowable transitions between words but not the reversal .
- A finite-state machine, specifies a set of states, along with allowable transitions between them .
- 'The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones'.
- A finite-state grammar (FSG) can generate an infinite number of sentences using finite resources, it allows for loops.



- However Chomsky argues that an FSG cannot represent the full complexity of English
 - Nested dependencies between particles like ifthen, either.... or.
 1. A . If_A John either_B ate or_b drank then_a he couldn't sleep
B * If_A John either_B ate then_a drank or_b he couldn't sleep
Mirror structure(center embedding) however an FSG cannot generate mirrored structures of unbounded size, just as it cannot represent A^nB^n .
 2. a. Colorless green ideas sleep furiously.
b. Furiously ideas sleep green colorless.
- Chomsky claims that since the two sentences, and their subparts, would equally be zero frequency in a corpus, they would be 'equally "remote" from English ... in any statistical model of English'

- Pereira (2000) has shown that this is incorrect: these sentences are in fact distinguished by a bigram model over word categories
- However, Pater still agreed with Chomsky that it would be a ‘dead end’ to try to reduce sentence well-formedness to n -gram probabilities over sequences, or to the more complex distributions over sequences that can be represented by a probabilistic Markov chain instantiation of an FSG, regardless of whether those sequences are of words or categories.



WHAT CHARACTERISTICS MUST AN ADEQUATE GRAMMATICAL MODEL FOR ENGLISH HAVE?

- Hierarchical phrase structure of a sentence.
- “Chomsky (1957:30) shows that a phrase structure grammar is able to generate the nested dependencies discussed above and is thus more powerful than an FSG. Chomsky’s proposal goes further than phrase structure in also making use of transformations, which take as input a string with phrase structure and produce as output another string with a new structure (p. 44). Chomsky argues that by deriving passive sentences, negation, and questions through transformations on a base ‘kernel’ phrase structure, considerable simplifications in the form of the grammar can be obtained.”

- Hierarchical constituents, and underlying forms that are derivationally transformed into the surface structure characteristic of generative analyses of aspects of language other than syntax .
- (Heinz & Idsardi 2011, 2013), argue that phonological restrictions seem to be representable at the segmental string level by FSGs , hierarchical representations are standardly viewed as necessary for an adequate characterization of the phonologies of the world's languages (Selkirk 1981, Yu 2017).

CONTROVERSIAL DEEP DERIVATIONS

- In phonology suspected to pose difficulties for learning .Eg: perceptron hidden layers for XOR example had to be inferred by a learner. And it can pose learning challenges Tesar and Smolensky (2000)
 - Connections Between Linguistics and Neural Networks And Its Usefulness
 - Techniques for learning with hidden layers are potentially useful in learning with explicitly encoded hidden linguistic structure.
 - Hidden layers may be used to learn representations that take the place of explicitly encoded hidden structures.

INNATISM AND EMERGENTISM.

- The inadequacies of (probabilistic) sequential models of language, which can be trained by relatively simple learning algorithms.
- Rejects the structuralist insistence on discovery procedures—algorithms for proceeding from a corpus to an analysis—in favor of a weaker requirement that there be an evaluation procedure, a method for choosing among hypothesized grammars.
- The generative program eventually became one of mapping out the hypothesis space that a learner was claimed to deductively navigate in acquiring a language—of characterizing UG.
- In principles-and- parameters theory (Chomsky 1980:3–4), the hypothesis space is characterized by a set of universal principles, alongside language-specific parameters that are ‘fixed by experience’
- Poverty of the stimulus. Led to the emergence of principles-and-parameters theory
- It is important to emphasize, that a network needs a specified structure for it to represent the effects of learning, just as innate parameters need a specification of how learning works if they are to respond to experience.

Back To Neural Networks

- Single-layer perceptron is an error-correction procedure.
- Rosenblatt's (1957, 1958) learning algorithm for single-layer perceptrons is an error-correction procedure.

Given an input, the network is used to predict an output. If the network's output fails to match the correct output, yielding an error, the weights are changed slightly in the direction of generating the correct activation pattern.

The activation pattern of a hidden layer is not given as part of the training data, so it is not straightforward to update its input weights. Rosenblatt could see that a solution would be to propagate the error signal back through the hidden layer(back propagation).

- According to Terrance Sejnowski (interview cited in Olazaran 1993:398), the crucial step in developing the algorithm known as backpropagation (Werbos 1982, Rumelhart, Hinton, & Williams 1986, LeCun 1988) was to replace the step activation function, which yields discrete activation levels, with a continuous sigmoidal function.

This allowed for the calculation of a gradient, which determines the direction of the weight updates.

Although backpropagation is not guaranteed to find an optimal set of weights, it is (perhaps surprisingly) effective, and a variety of methods exist for increasing the likelihood that it will find a global, rather than a local, minimum of error.

- Because of its focus on learned representations, neural network research is a largely emergentist tradition, and the connectionist linguistic literature often contrasts itself with Chomskyan innatism .

THE PAST-TENSE DEBATE.

- The input to this network is a phonological representation of the uninflected form of an English verb, and its output is the predicted phonological form of its past tense. Rumelhart and McClelland (1986:217) present this network as illustrating an alternative ‘the rules of language are stored in explicit form as propositions, and are used by language production, comprehension and judgment mechanisms’ and that in learning, ‘[h]ypotheses are rejected and replaced as they prove inadequate for the utterances the learner hears’, using a learning mechanism that has ‘innate knowledge of the possible range of human languages’.
- Both the addition of the phonologically appropriate form of -ed in regular past-tense formation and the various forms of irregular past tense such as vowel change (*sing, sang*), no change (*hit, hit*), and suppletion (*go, went*) are all handled by a single network of weighted connections between the atoms of the phonological representation of the uninflected form and those of the inflected one. Learning consists of weight adjustments in response to errors in the prediction of the past-tense form. This may be seen as a form of hypothesis testing in the sense that the current values of the weights represent the network’s current hypothesis, which is modified by a weight update, but it is different from most generative learning algorithms in that the hypotheses are over a continuous rather than a discrete space, and the changes in output are typically gradual, rather than abrupt.

- Pinker and Prince's (1988) critique of Rumelhart & McClelland 1986 is the nature of the phonological representations used in the model. Like the input layer of the object-classification networks there is no temporal order in the input nodes of the past-tense model. To represent phonological contexts, a single node encodes features of both the preceding and following phone, as well as the central phone—a triphone representation called a 'Wickelfeature' (after Wickelgren 1969)(problematic).
- It can represent string as easily as it can represent an identity map, yet no language uses string reversals as a phonological process

CONNECTIONIST MODELS

- Note features for where they appear in a word
- Another is to make use of a recurrent neural network (Elman 1990), in which the phonological string is processed one segment at a time.
- One popular one that ignited further research is that irregular and regular past are the product of separate systems, rather than being produced by a single cognitive module. That is, the irregular past tenses are lexically stored, and anything that looks like a rule-governed regularity is in fact a product of how the words are stored. The regular pattern, by contrast, is the product of a morphosyntactic rule, or rules, that adds the -ed morpheme, and phonological rules of voicing assimilation and vowel epenthesis that yield the contextually appropriate surface forms.

- Although Rumelhart and McClelland's model of the past tense clearly lacks the rules of a standard analysis of the regular past tense, it adopts relatively standard phonological features, and Lachter and Bever (1988:211) argue that choice, as well as the particular configurations of features for the nodes, essentially engineers a rule-based solution into the model.
- The absence of variables in the Rumelhart and McClelland (1986) model is another of the primary targets of Pinker and Prince's (1988) critique.
- According to Paper the debate shows that the space between connectionist and generative models of language is more fluid .
 - there is nothing about a connectionist model that prohibits the use of symbols, including variables, and other representations developed in linguistic traditions.
 - a generative rule-based model can, and often does, have the very specific rules needed to model irregular morphophonology
 - a generative model is not fully innatist in that parameters need to be set by experience
 - a connectionist model is not fully emergentist in that much of its structure must be specified

- Rumelhart and McClelland's (1986) past-tense model uses a probabilistic interpretation of a sigmoid activation function, and thus produces probabilities over different outputs for a given input. Models of generative grammar, from Chomsky 1957 onward, typically use deterministic rules that produce a single output for a given input.
- The past-tense debate also provides a reason to formalize a rule-based model probabilistically: as children acquire the regular -ed past tense, its probability of use seems to increase gradually
- The connectionist models of the past tense, including that of Rumelhart and McClelland, show that simple and explicit models of learning can be combined to good effect with the representational structures developed in linguistics.

FUSION

- OT and HG(with numerical weight rather than ranks, its constraints.
- Prince and Smolensky (2004:Ch. 4) provide an extended argument for constraint interaction in the domain of word stress, comparing the Extrametricality parameter (Hayes 1980) to a violable Nonfinality constraint.

	Weight-to-Stress	Nonfinality	HARMONY
Input: batan	5	2	
Output: batán		-1	-2
bátan	-1		-5

A harmonic grammar tableau

	Weight-to-Stress	Nonfinality	HARMONY
Input: bata	5	2	
batá		-1	-2
Output: báta			0

- The switch from parameters to violable constraints has consequences for the study both of language typology and of learning.
- For language typology, it becomes possible to maintain relatively general formulations of constraints while still accounting for details of individual languages.
- In parametric theory, an analyst's observation of surface violations of constraints in a language in which they are generally active leads to a range of responses, including simply changing a general constraint into a set of more specific ones that can remain inviolable.
- Both OT and HG have provably convergent learning algorithms that will find a constraint ranking, or weighting, for any set of candidate outputs that can be made jointly optimal.

MAXIMUM ENTROPY GRAMMAR

- A probabilistic version of OT and HG
- The probability of an output is proportional to the exponential of its harmony.
- One class of gradual learning algorithm used in this work includes an application of Rosenblatt's perceptron convergence procedure.
- Guarantees of OT and HG learning algorithms apply only when the structure of the learning data is supplied in whole—when all of the constraint violations of each learning datum are known.
- While it might not be a problem for an analyst to supply full structures when studying typology or in modeling some cases of variation, there are many cases of linguistic analysis in which one might not be committed to a particular full structure for each piece of data and would like a learner to find an appropriate grammar.

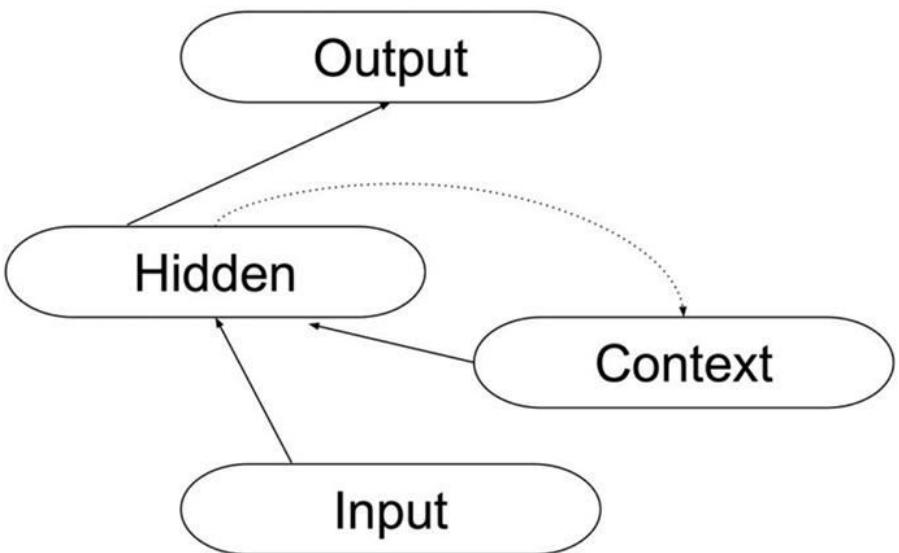
LEARNING WITH HIDDEN STRUCTURES

- A MaxEnt version of this general approach (e.g. Pater et al. 2012, Pater & Staubs 2013, Johnson et al. 2015) creates a single vector (row) of constraint scores for a partially structured learning datum by summing over the probability- weighted vectors of all of the corresponding full structures.

MAXENT AND NEURAL NETWORK

- Both can be learned with gradient-based optimization methods, including gradient descent.
- When the gradient for a neural net with one or more hidden layers is constructed using backpropagation or when a gradient for a MaxEnt model is constructed with hidden structure there is no guarantee that these methods will find the best set of weights for the model, in terms of optimizing the fit of the model's predictions to the data. That is, the learner may not find the global optimum and may instead be trapped in a local minimum of error.

CAN RNN LEARN SYNTAX?



- Can capture some aspect of natural language syntax, including long-distance dependencies, and have recently undergone a resurgence of popularity in AI applications of neural networks to language.
- As illustrated in Figure 4, when moving on to the next element in a sequence, the current hidden layer is copied as a context layer to provide an extra set of inputs to the next computation of the hidden-layer activations, and to the output. The representation encoded in this copied hidden layer provides a basis for the prediction of upcoming elements based on those encountered earlier—that is, it is a type of sequential memory.

- RNN that is simply trained to predict upcoming words generalizes the incorrect rule (contra Lewis & Elman 2001). Also, if the learner is given the task of generating sentences given a meaning representation, it does yield the correct generalization from monoclausal training instances of auxiliary inversion to the structure-dependent rule.

Fitz and Chang

- RNNs are used to map from a sequence of words in one language to a sequence in another, without any intermediate, explicitly encoded linguistic.
- The success of modern RNNs in applied language tasks is due to advances in their architecture, as well as in training methods and computational hardware that jointly allow for training of large networks . However , they tend to also extract incorrect linear regularities.

- Another challenge of RNN is shown in their study of anaphora resolution, Frank et al. (2013) conclude that the representations are not sufficiently abstract, being too tied to particular words rather than to categories.

CONCLUSION

- Neural network and generative linguistic approaches to cognition overlap considerably: they both aim to provide formally explicit accounts of the mental structures underlying cognitive processes, and they both aim to explain how those structures are learned. When viewed more closely, especially with respect to the research practices within each tradition, they may seem to diverge sharply, with the bulk of connectionist practice involving computational learning simulation allied with AI tasks or with psychological experimentation, and with the bulk of generative practice involving grammatical analysis of linguistic systems.