Language, Learning, and Creativity

Stephen Pulman

Emeritus Professor,
Department of Computer Science,
Oxford University
stephen.pulman@cs.ox.ac.uk

and Senior NLP Research Scientist, Apple.[1]

30th May 2018

[1] Views expressed here are my own and nothing to do with Apple!

Abstract

It's nearly 70 years since the Turing test was proposed as an operational test for intelligence in computers, and it is still a subject that provokes much discussion. One important aspect of the Turing Test is linguistic ability, and one important aspect of that ability is what Chomsky called "the creative aspect of language use", the ability of language to serve as "an instrument for free expression of thought, unbounded in scope, uncontrolled by stimulus conditions though appropriate to situations".

With every new wave of progress in artificial intelligence, such as that promised by the current "deep learning" paradigm, it's natural to ask the question whether these advances get us any nearer to a machine that could pass the Turing test, or that could use language creatively. In this talk, I'll explore these issues, in particular looking at some parallels between the implications for human learning that we could derive from current deep learning methods, and the intellectual climate of behaviourism and empiricism in language learning and use that Chomsky was reacting against.

Distributional Structure

Zellig Harris: we can discover the grammar of a language from a corpus by purely formal means.

- Noun $= \{X \mid X$ appears in environment "the X is/are ..." $\}$
- Verb $= \{X \mid X$ appears in environment "... is/are X-ing ..." $\}$
- NounPhrase $= \{X \mid X$ appears in environment "X exists." $\}$
- Adjective + Noun $\in$ NounPhrase
- NounPhrase + VerbPhrase $\in$ Sentence
- etc.

These statements make no reference to meaning. The grammar of a language consists of an ordered set of mutually recursive statements like these. Since the discovery procedure needs only the assumption that the sentences in the corpus are grammatical, and notions like "same/different environment" it could in principle be automated.

Empiricist Learning Theory

Chomsky pointed out that such an approach to "learning" the grammar of a language could be construed as a kind of "empiricist" learning theory, since it purports to use only simple notions of similarity and difference, and derives all the "rules" directly from data.

In the empiricist view, the mind is a "tabula rasa" and:

> *"... there is nothing in the intellect that was not first in the senses ..."*

Chomsky argued that if it is construed this way, Harrisian empiricist approaches can be shown to be inadequate, as they are not capable of accounting for some important features found in language.

Structure Dependence

1. The famous pair:

John is easy to please vs. John is eager to please

- while identical in terms of grammatical categories, display differences of interpretation that distributional methods struggle to uncover.

2. If you know a language you know lots of things about the relation between sentences, for example:

Declarative: The man is tall
Yes-no question: Is the man tall?

After lots of examples we might arrive at the generalisation that to make a yes-no question we start with the declarative, look for the first verb from the start of the sentence, and prepose it:

The man is tall $\Rightarrow$ *Is the man tall?*

This will work most of the time, but not always:

The man who was here is tall $\Rightarrow$ *Was the man who here is tall?*

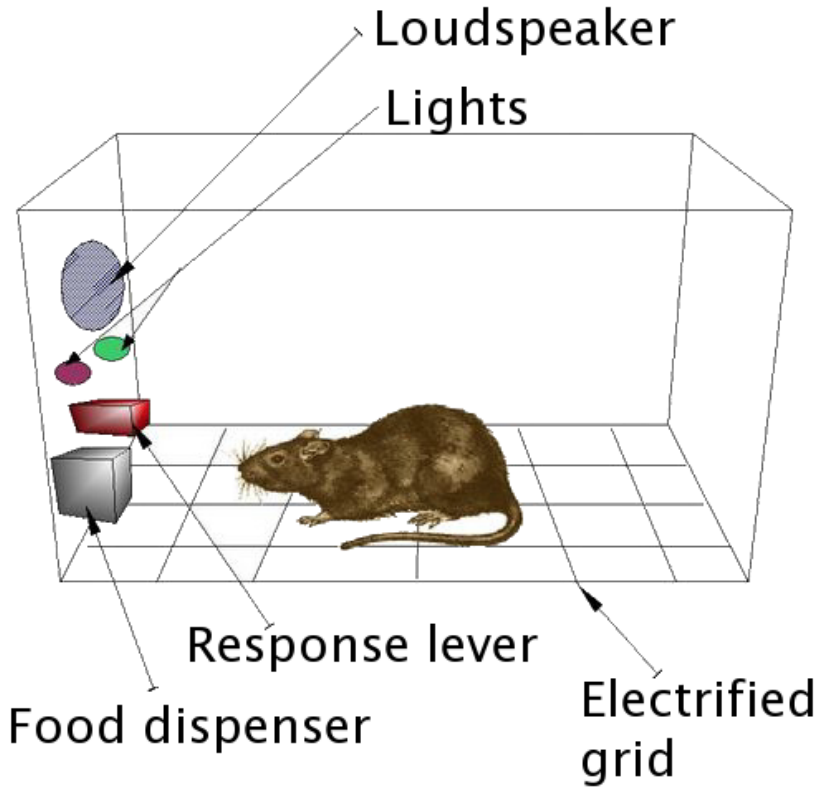
Universal Grammar

The correct hypothesis is: find the first verb after the subject Noun Phrase and prepose that.

[The man who was here] is tall ⇒ Is [the man who was here] tall?

- This "structure dependent" hypothesis requires that at some level a speaker is analysing the sentence hierarchically into abstract phrases. Children learning language converge immediately on the correct hypothesis and do not make structure-independent hypotheses.
- Chomsky argues that the best explanation of these observations is that notions like structure-dependence are hard-wired in us, one of the principles of "Universal Grammar", a species-specific property.
- This is a rationalist theory: the mind is not a blank sheet, but comes equipped with "innate ideas", a set of *a priori* assumptions and biases, that enable learning to be fast, and triggered by relatively small amounts of the relevant data.

Skinner's "Verbal Behaviour"

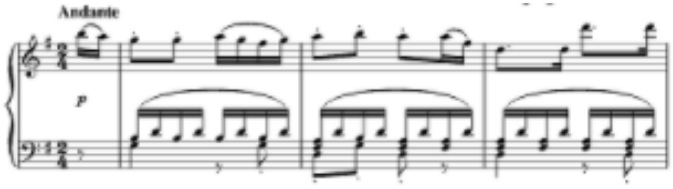


Positive reinforcement: hunger (= "stimulus");
press lever ("response") $\rightarrow$ get food ("reinforcement")
Negative reinforcement: grid electrified;
press lever $\rightarrow$ grid off

Language as stimulus-response

Skinner's aim "to identify the variables that control *verbal behaviour* and specify how they interact to determine a particular verbal response"

"Dutch!"



"Mozart!"

Chomsky's review of Verbal Behaviour

- Chomsky points out that the notions of "stimulus", "response" and "reinforcement" are so extended as to be meaningless in trying to explain a wider range of verbal behaviour:
- "What point is there in saying that the effect of *The telephone is out of order* on the listener is to bring behavior formerly controlled by the stimulus *out of order* under the control of the stimulus *telephone*...by a process of simple conditioning? What laws of conditioning hold in this case? Furthermore, what behavior is *controlled* by the stimulus *out of order*, in the abstract?"
- There is no hope of behaviourism explaining the "creative aspect of language use": any native speaker is capable of producing completely new utterances, not necessarily responding to any stimulus, but still appropriate to the context.

Creative Aspect of Language Use

Language use which is:

1. unbounded: i.e. we can produce a potentially infinite number of new sentences, via the compositional mechanisms of grammar
2. stimulus free: the content of what we say need not be determined by the situation we are in
3. what we say is appropriate to the situation

It's easy to find examples of language use which satisfies one or more of these criteria, but not all three simultaneously.

Descartes had observed in the 17th century that no animal communicative behaviour displayed these properties, which he regarded as criterial for possession of a mind.

Descartes: from A Discourse on Method

(Automata) could never use words or other signs arranged in such a manner as is competent to us in order to declare our thoughts to others: for we may easily conceive a machine to be so constructed that it emits vocables, and even that it emits some correspondent to the action upon it of external objects which cause a change in its organs; for example, if touched in a particular place it may demand what we wish to say to it; if in another it may cry out that it is hurt, and such like; but not that it should arrange them variously so as appositely to reply to what is said in its presence, as men of the lowest grade of intellect can do.

The Turing Test

Turing's operational test for an intelligent machine:

- Human H communicates via textual messages (to abstract away from physical properties) with two agents, one machine and one human.
- If, after a reasonable period of time, H cannot tell which is the machine, then the machine has passed the operational test for intelligence.
- Turing's test requires the machine to display intelligence in the man-in-the-street sense, of being able to do mental arithmetic and solve chess problems.
- But as the following imagined exchange shows, the test also seems to presuppose that the machine can use language creatively in Descartes and Chomsky's sense.

Conversation...

H: In the first line of your sonnet which reads "Shall I compare thee to a summer's day?", would not "a spring day" do as well or better?
M: It wouldn't scan.
H: How about "a winter's day". That would scan all right.
M: Yes, but nobody wants to be compared to a winter's day.
H: Would you say Mr Pickwick reminded you of Christmas?
M: In a way.
H: Yet Christmas is a winter's day, and I do not think Mr Pickwick would mind the comparison.
M: I don't think you're being serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Cultural knowledge

The content of what the machine says relies on highly sophisticated cultural knowledge, in this case partly literary. Turing seems to have shared the empiricist and behaviourist assumptions of the time about how such knowledge is acquired, and proposes to "teach" the machine:

> *"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.)"*

Learning and Teaching

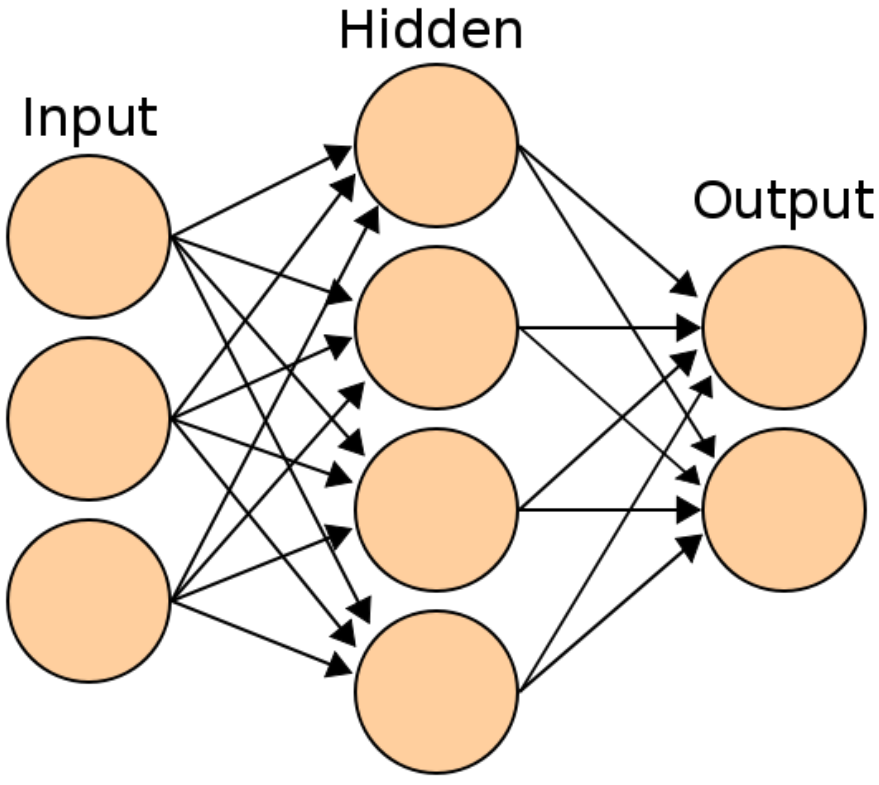
> *"The criterion as to what would be considered reasonable in the way of 'education' cannot be put into mathematical terms, but I suggest that the following would be adequate in practice. Let us suppose that it is intended that the machine shall understand English, and that owing to its having no hands or feet, and not needing to eat, nor desiring to smoke, it will occupy its time mostly in playing games such as Chess and GO and possibly Bridge....I suggest that the education of the machine should be entrusted to some highly competent schoolmaster..."*

The machine will have a memory and whenever a choice as to what do next arises, the machine consults the memory to see what the outcome of similar choices should be. Turing acknowledges that without some idea of what counts as a "favourable outcome" this will not work:

> *"I suggest there should be two keys which can be manipulated by the schoolmaster, and which represent the ideas of pleasure and pain."*

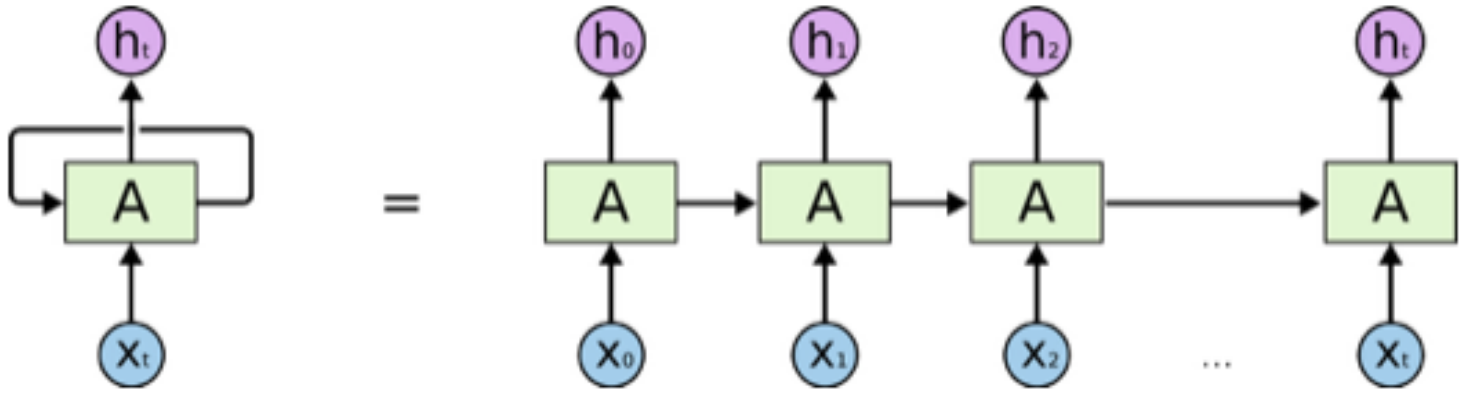
What is deep learning?

Familiar three layer neural network:



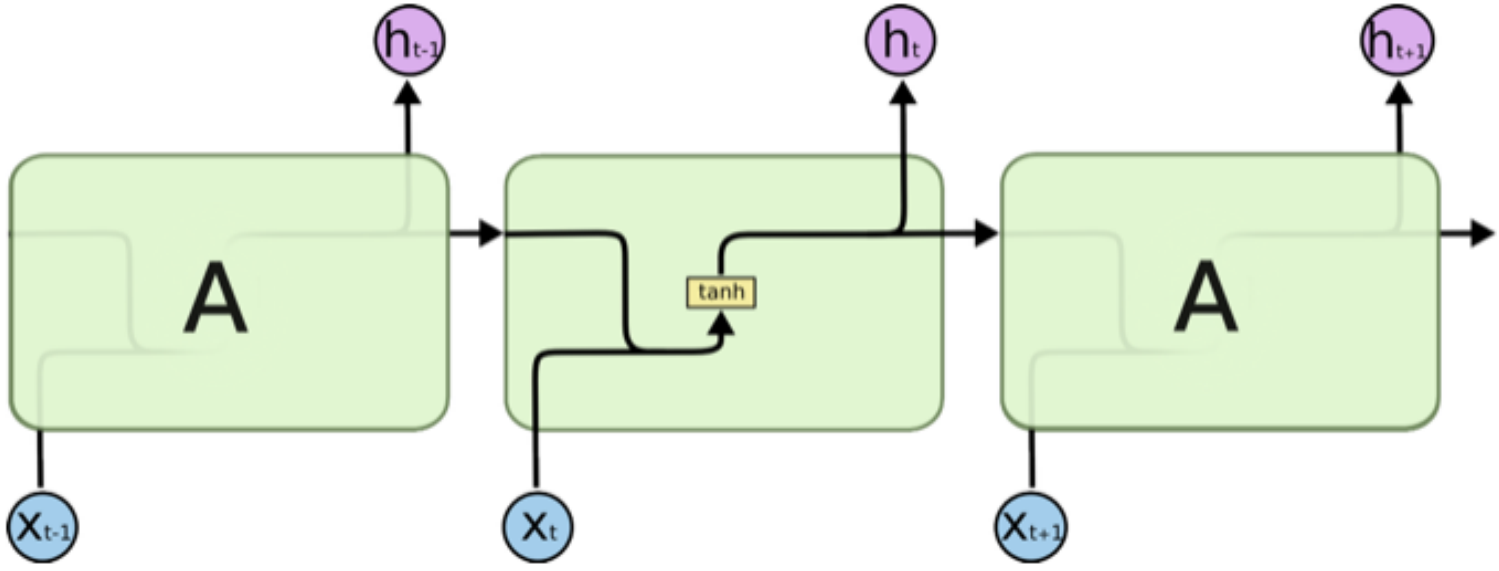
Yann LeCun: "Deep Learning is anything with more than one hidden layer"

Recurrent Neural Networks



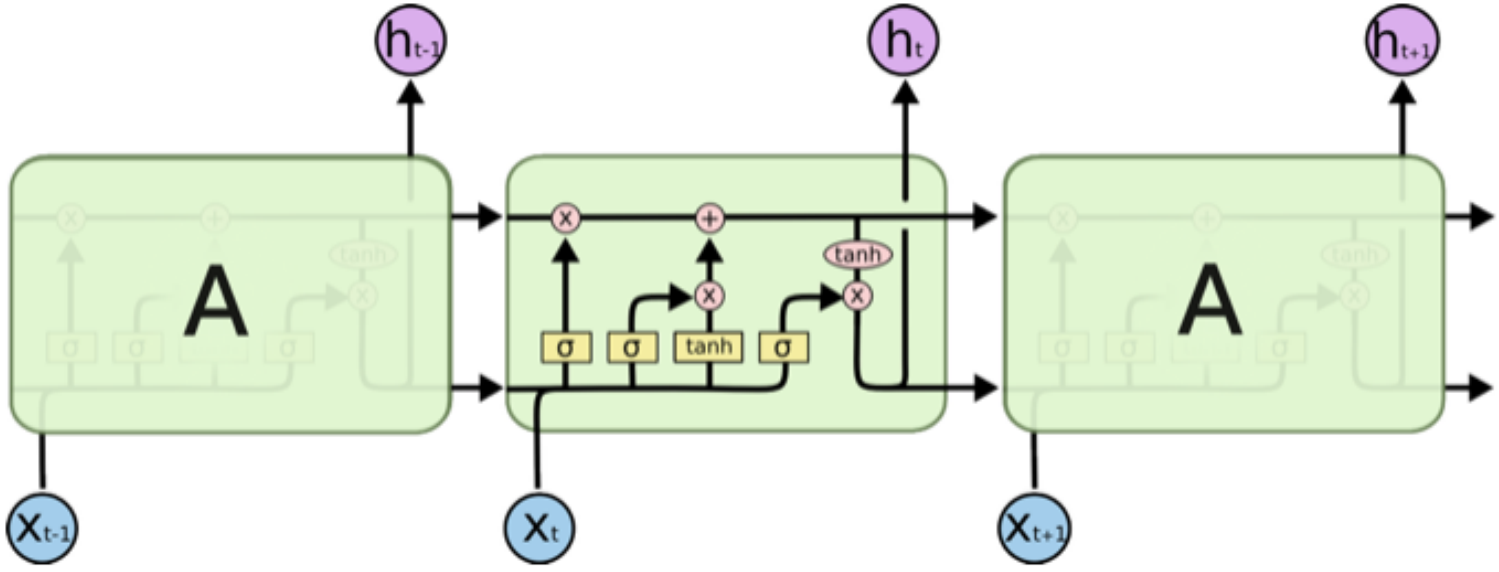
A recurrent neural network, unrolled to fit the length of the input sequence.

Recurrent Neural Networks



An RNN cell takes as input the current item in the sequence and the output of the previous cell.

Long Short Term Memory

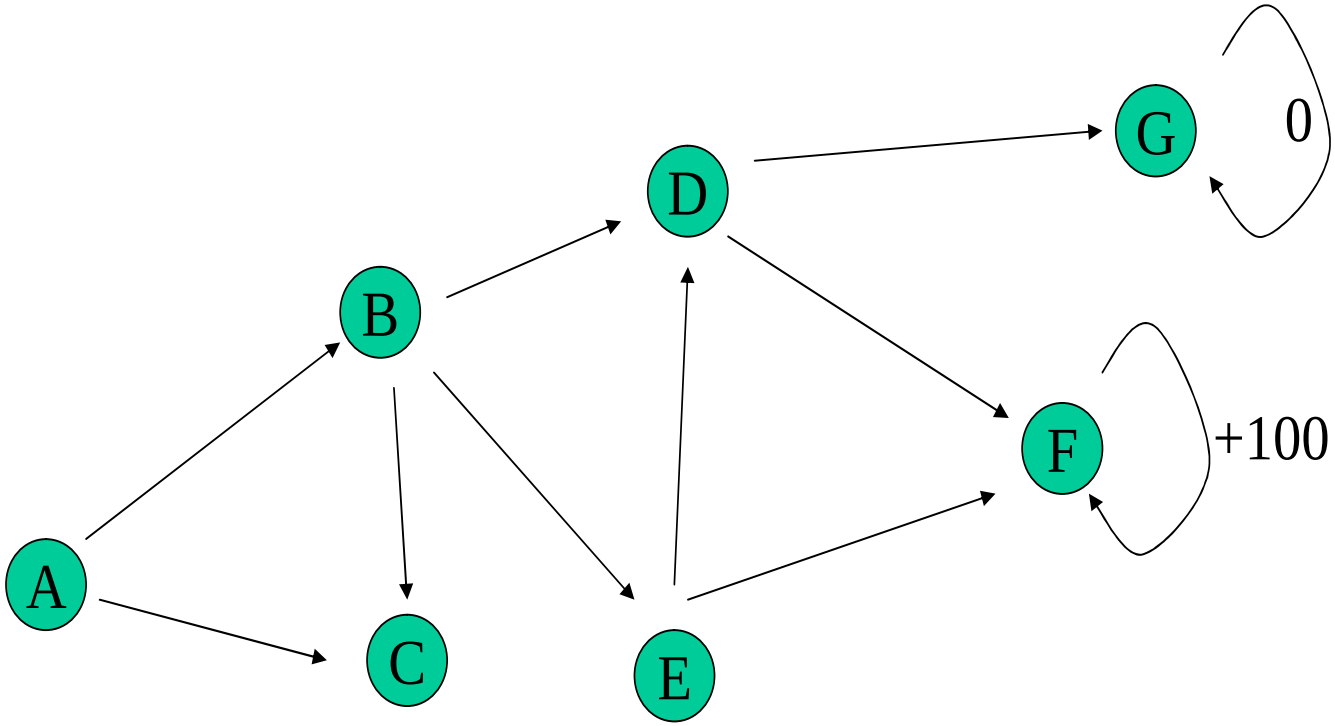


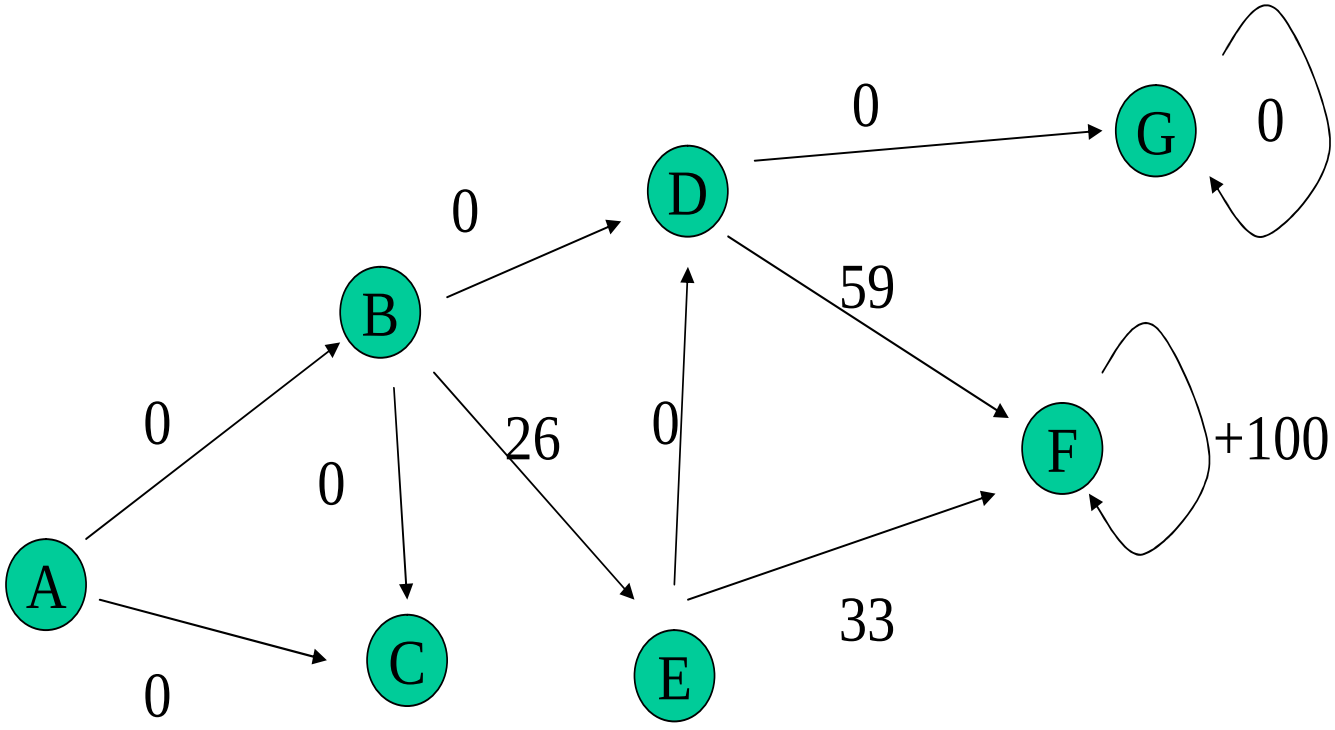
An LSTM cell also learns how much to remember and pass on to the output.

Traditional Reinforcement Learning

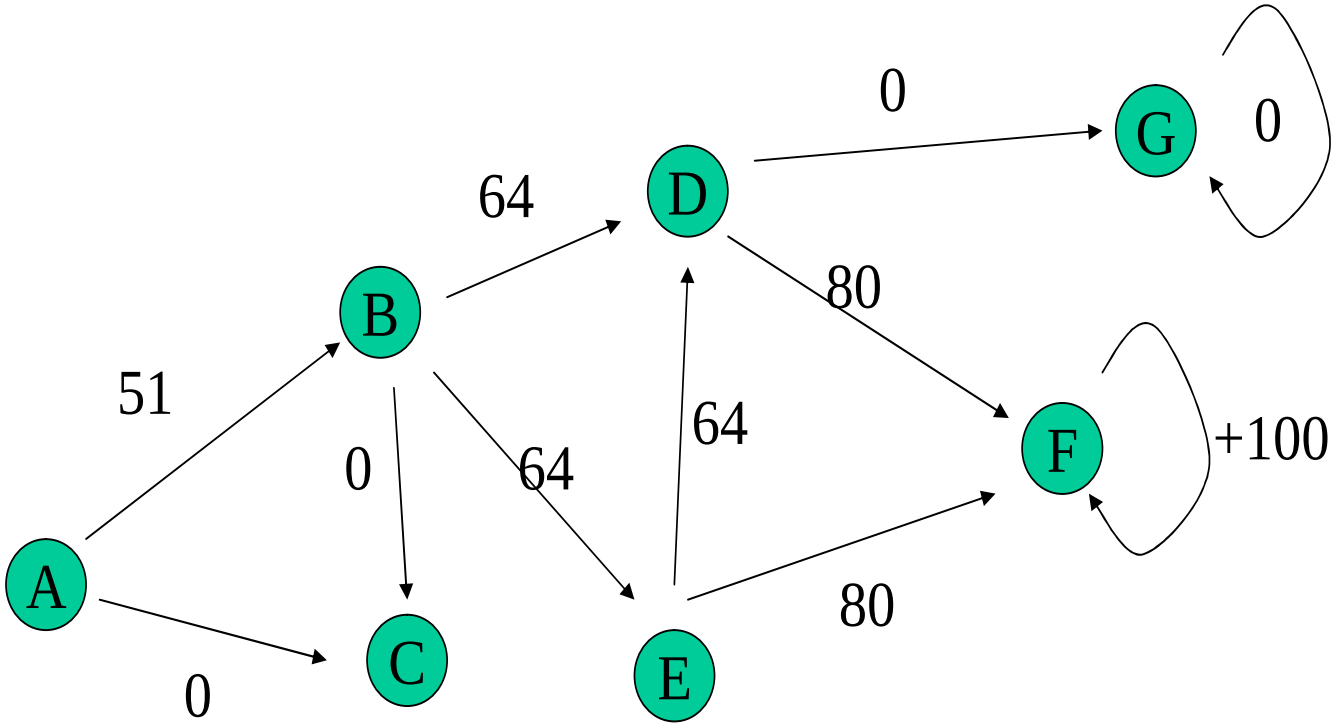
Q-learning is a direct implementation of Skinnerian operant conditioning, and the kind of learning Turing had in mind. The value $Q(s, a)$ is the future reward obtained by taking action a from state s and following an optimal "policy". The Q-values are learned from agent experience following these steps:

1. Pick a state s at random.
2. From the current state s, select an action a. This will get an immediate reward r, and we move to the next state s'
3. Update $Q(s, a)$ based on this experience as follows:
   $$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{b \in A} Q(s', b) - Q(s, a)]$$
   where α is the learning rate and $0 < \gamma < 1$ is a 'discount factor'
4. Go to 1


A
B
C
D
E
F
G
0
+100



after 5 iterations



at convergence

Deep Reinforcement Learning

- represent the Q-value function Q(s,a) by a deep Q-network with weights Q(s,a,w)
- define a loss function using (e.g.) mean squared error in Q-values:

$$\mathcal{L}(w) = \mathbb{E}_{s,a,r,s' \sim \mathcal{D}} \left[\left(r + \gamma \max_{a'} Q(s', a', w) - Q(s, a, w) \right)^2 \right]$$

- train using standard Stochastic Gradient Descent + back-propagation
- recognise states using another NN or nearest neighbour embeddings
- save and replay variant (ϵ-greedy) training experiences

Deep Learning

- No *a priori* knowledge involved - a "blank sheet".
- A completely general approach: no task-specific adaptations
- Everything is learned by multiple exposures to labelled training data.
- ... and lots of it.
- A thoroughly empiricist/behaviourist approach to learning.

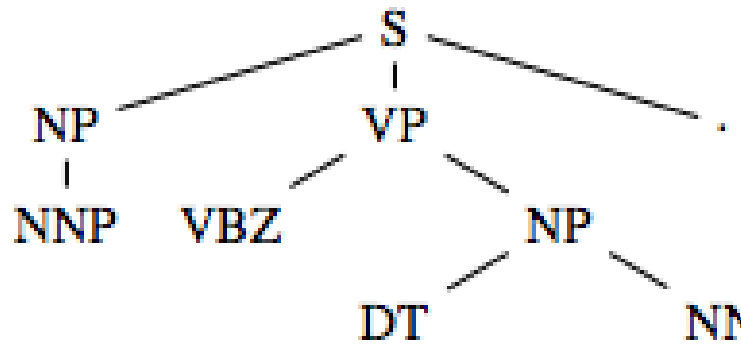
Given Chomskyan doubts about the adequacy of such approaches for grammar learning (or any learning), how likely is it that we could use deep learning methods to build a system capable of learning enough to pass the Turing Test?

Structure Dependence: sequence models

Given that structure dependence is a central part of language, how successful have DL methods been at learning this property?

Vinyals et al 2015 show how an LSTM encoder-decoder sequence-to-sequence model with attention can assign linearised parse trees to input:

John has a dog . →



John has a dog . → (S (NP NNP)$_{NP}$ (VP VBZ (NP DT NN)$_{NP}$)$_{VP}$.)$_{S}$

Structure Dependence: sequence models

```
(S (NP  NNP  NP) (VP  VBZ  (NP  DT  NN  NP)  VP) S)
        |            |       |   |
      John          has      a  dog
```

- The system was trained on over 11m parsed sentences and reached 92.1% F1 score on PTB23: a very impressive result.
- But has this system really learned the notion of hierarchical structure?
- I would say not: it is transducing between two *strings* of symbols, and has no general idea that a left bracket should be matched with a right bracket.
- The errors nearly all involve missing right brackets.

Structure Dependence: modelling subject-verb agreement

Linzen et al. experimented with trying to learn **subject**-*verb* agreement using LSTMs on word embeddings:

1. The **students** *submit* a final project to complete the course.
2. The **students** enrolled in the program *submit* a final project to complete the course.
3. The **students** enrolled in the program in the Department *submit* a final project to complete the course.
4. The **students** enrolled in the program in the Department where my colleague teaches *submit* a final project to complete the course.

Results seem reasonable: with 4 intervening distractors, error rate is only 17.6%.

Learning syntax, or lexical relations?

Bernardy and Lappin (2017) repeated these experiments, with broadly similar results, although accuracy improved with more data and higher dimension embeddings. They also experimented with a reduced vocabulary version intended to encourage learning of abstract syntactic structure (100 most frequent words vocab, with other words represented by their POS tags.) This didn't work well:

> *"DNNs learn better from data populated by richer lexical sequences. This suggests that DNNs are not efficient at picking up abstract syntactic patterns when they are explicitly marked in the data. Instead they extract them incrementally from lexical embeddings through recognition of their distributional regularities. It is also possible that they use the lexical semantic cues that larger vocabularies introduce to determine agreement preferences for a verb."*

This raises the question of what exactly is being learned here: structure dependence, or lexical correlations?

Colourless green ideas sleep furiously

In a recent paper, the Facebook AI group repeated these experiments using both grammatical and nonsensical sentences, and range of agreement constructions in four different languages:

- It *presents* the case for marriage equality and *states* ...
- It *stays* the shuttle for honesty insurance and *finds* ...

They found that although there was a drop in accuracy between grammatical and nonsense sentences, it was about the same for their LSTM system and for people, based on results obtained from human subjects (for Italian). For Italian at least, the LSTM almost reached human performance. They conclude, tentatively, that the LSTM *is* learning grammatical representations rather than lexical dependencies.

Of course, we could just add a stack memory to our LSTM to build in hierarchical structure.

Word embeddings

- Create randomly initialised vectors for each word (varying lengths, say 50).
- Create training data by concatenating vectors for 5-gram (positive example) and same 5-gram but with a randomly different word substituted for one word (negative example).
- Example: ...photographer visits Syrian refugees in...
  vs. ...photographer magenta Syrian refugees in ...
- Use traditional NN to compute score for each, training objective: *score(pos)* $>$ *score(neg)*.
- Simultaneously propagate weight changes to word vectors.
- Resulting vectors implicitly represent contextual and cluster properties of words.

Some interesting properties

Clustering - nearest neighbours within vector space to:

France:	Austria, Belgium, Germany, ...
scratched:	nailed, smashed, punched, scraped, slashed,...
Xbox:	Amiga, Playstation, MSX, iPod,...

Analogy - subtract one vector, then add another, then find nearest neighbour:

king	− man	+ woman	= queen
Paris	− France	+ Italy	= Rome
sushi	− Japan	+ Germany	= bratwurst
but:		+ France	= tapas
and:		+ USA	= pizza!
France	− Sarkozy	+ Berlusconi	= Italy
		+ Merkel	= Germany

So these word embeddings seem to really capture something meaningful.

Word Embeddings and Subcognition

French (1990) argues against the possibility of passing the Turing test, in grounds that some aspects of language understanding involving "subcognitive" processes, in particular associative memory phenomena:

- P(nurse | doctor) > P(teabag | doctor)
- lexical decision: primed with "doctor" you will agree that "patient" is a real word faster than you will for "teabag", and both will be faster than "danbage".
- neologisms: "Flugblogs": (a) a new breakfast cereal (no) (b) a new computer company (no) or (c) large air-filled bags for the feet, used to walk across water? (yes)

It's not impossible that these judgements could be obtained from word and character embeddings.
"Tell me, do you think that doctors have more in common with nurses than they do with teabags?"

Creative Aspect of Language Use vs. the Turing Test

Recall that human utterances are simultaneously:

- Unbounded: i.e. we can produce a potentially infinite number of new sentences, via the compositional mechanisms of grammar.
- Stimulus free: the content of what we say need not be determined by the situation we are in.
- What we say is appropriate to the situation.

Chomsky believes that we are unlikely to develop scientific theories that give us real insight into this ability, since it involves notions like free will, intention, decision making: all "mysteries".

But it is clear that if we could develop a system displaying the creative aspect of language use, that would be sufficient for it to pass Turing's Test.

How far away are we from this?

Unboundedness and novelty?

- We have, via recursive mechanisms of syntax and semantics, the ability to generate an infinite number of new sentences.
- However, these sentences should express some content: our thoughts. We do not know how to guarantee this, nor what factors prompt us (for example) to start a conversation.
- To what extent is this like the Reinforcement Learning setting: given a state, predict the best next action?
- But for Turing's test, we can confine ourselves to the question answering scenario, so that what to say next is determined by the current and previous questions.
- In this respect, passing the Turing Test is a little easier than mastering "the creative aspect of language use".

Stimulus free?

- When Chomsky talks about stimulus he seems to have the strict behaviourist definition in mind: local and immediate.
- But really no utterance or thought is stimulus-free, otherwise it would be random and unconnected to anything.
- It is just that the stimuli may be distant in time or space, complex, and private to the speaker.
- And in principle could be derived from anything in the entire lifetime of experience of the speaker.
- That's a lot of data to train on...

Appropriate to the context?

Wilson and Sperber's Relevance Theory might be a starting point:

- An utterance is relevant to the extent that it yields (positive) cognitive effects.
- "The most important type of cognitive effect achieved by processing an input in a context is a CONTEXTUAL IMPLICATION, a conclusion deducible from the input and the context together, but from neither input nor context alone."
- There is a trade-off between the amount of processing effort needed to derive cognitive effects, and the number and significance of those effects:

Mary: (dislikes most meat, and is allergic to chicken) What are you serving for dinner?
(It's chicken)
John:

(a) We're serving meat
(b) We're serving chicken
(c) Either we're serving chicken or 13 is not a prime number

Computing relevance

- We need to be able to do inference:

  Peter: Did John pay back the money he owed you?
  Mary: No, he forgot to go to the bank.

- bank = money; no bank, no money; no money, no pay...
- We need to be able to rank different inferences, perhaps in terms of information content or utility.
- We need to have some way of comparing processing effort.
- "mind-reading" - we need to be able to model other people's beliefs, desires, and intentions.

Conclusions

- Current RL and DL methods require supervision. Someone has to provide the labels or define the reward function. So we might be able to *teach* a machine to pass the Turing Test, but it will not be able to *learn* for itself.
- Chomsky was essentially correct that many important properties of language (and word meaning) are not manifest in the data. But it is not impossible that the right *a priori* structure could be built in to DL models.
- While mastery of creativity is sufficient to pass the Turing Test, not all aspects may be necessary.
- We could build a machine to pass the Turing Test, but I don't think we could build one that genuinely displayed the creative aspect of language use, again, for essentially the reasons Chomsky argued.
- We don't have the scientific tools to fully understand the mechanisms of choice, free will, and intention that are involved.