Informatics 1 Cognitive Science

Lecture 7: A Neural Network Model of the Past Tense

Frank Keller

30 January 2024

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Slide credits: Frank Mollica, Chris Lucas, Mirella Lapata

Overview

Recap: Words & Rules vs. Neural Nets

- Words & Rules theory explains the dichotomy between regular and irregular verbs. But issues remain, e.g., blocking needs to be stipulated.
- Maybe rules are not necessary to explain the past tense.
- Maybe children simply analogize from verbs they already know (e.g., from correct forms like *folded, molded, scolded* to over-regularization's like *holded*).
- All-rules vs. all-memory approach; rationalism vs. empiricism.
- Neural network: computer modeling approach inspired by biological networks of neurons (perceptrons, feed-forward networks).
- A neural net model should pick up both regular and irregular past temse patterns from the training data.

Recap: U-Shaped Learning

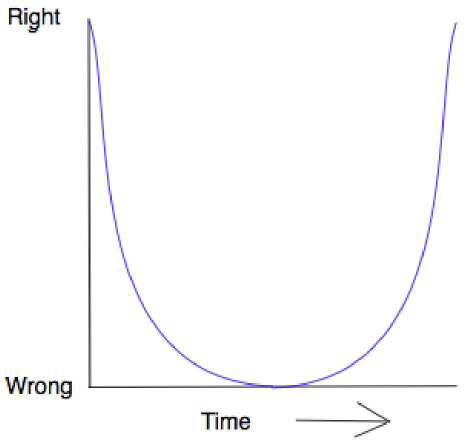
Children's performance gets better as they get older. With inflectional morphology they get worse before getting better. This is what child psychologists call U-shaped development.

Stage 1 children produce both regular and irregular past tense forms with very few errors.

Stage 2 after a certain amount of time, the error rate appears to increase significantly; children add regular past tense suffix *-ed* to irregular verb stems even with verbs whose past tense forms they had previously mastered.

Stage 3 the error rate slowly decreases, as the child gets older, until almost no errors are made.

Recap: U-Shaped Learning



- U-shaped learning in early childhood cognitive development.
- Child uses *spoke*, then *speaked*, and later again *spoke.*

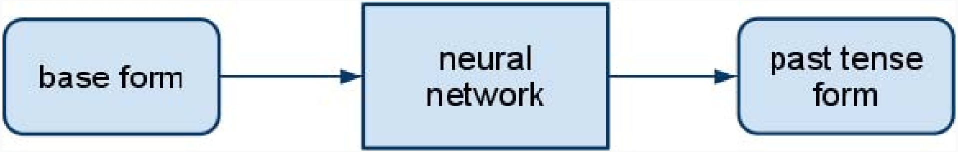
The Rumelhart & McClelland Model

The Rumelhart & McClelland Model

- In the 1980s, Rumelhart and McClelland promoted feed-forward multilayer perceptrons as basis for cognitive modeling.
- They called their approach Parallel Distributed Processing (PDP).
- The aim was to simulate children's three-stage performance in the acquisition of the past tense.
- Not a full-blown language processor that learns past tense from full sentences heard in everyday experience.
- Model is trained with pairs of inputs: (a) the phonological structure of the stem and (b) the phonological structure of the correct past tense.
- Model is tested by giving it the stem and examining what it generates as the corresponding past tense form.

The Rumelhart & McClelland Model

- Their model was pretty radical: no lexicon of words, no rules.
- Two-level fully-connected feed-forward perceptron network.
- And they didn't even use hidden units (later versions did).
- Input: a verb's base form, e.g., /dans/, /sink/
- Output: the past tense form, i.e., /danst/, /sank/



It's all in the Features

- The design of the input and output of the model is crucial.
- R&M assume input and output are represented as phonemes: *came* $\rightarrow$ /kAm/.
- There about 35 phonemes in English.
- We need a representations that represents the context of a phoneme, and allows us to generalize, e.g., for subregularities such as *sing-sang*, *ring-rang*, *spring-sprang*.
- So we want to encode the context of a phoneme, and its relative position, but not the exact phoneme sequence.
- Use triples of phonemes: *Wickelphones:* *sing* $\rightarrow$ /siN/ $\rightarrow$ {#si, siN, iN#}; here # denotes word boundary.
- So iN# can become aN# independent of word length.

It's all in the Features

- Wickelphones take up too much space: there are 35^3 of them.
- Instead, R&M use phonological features to represent phonemes:

$$\textit{came} \rightarrow /\text{kAm}/ \rightarrow \begin{bmatrix} \textit{Interrupted} \\ \textit{Back} \\ \textit{Stop} \\ \textit{Unvoiced} \end{bmatrix} \begin{bmatrix} \textit{Vowel} \\ \textit{Front} \\ \textit{Low} \\ \textit{Long} \end{bmatrix} \begin{bmatrix} \textit{Interrupted} \\ \textit{Front} \\ \textit{Nasal} \\ \textit{Voiced} \end{bmatrix}$$

- Only four features are required to represent all English phonemes; these correspond to 11 binary combinations.
- We again use triples to encode the input: but of features, not of phonemes: *Wickelfeatures.*
- We now have 11^3 possible combinations; after eliminating some redundancy, 460 are left. So we need 460 binary units.

It's all in the Features

- The input and output layer of fixed sized (460 units each).
- For a given word, the Wickelfeatures of all its phonemes are activated together. So *sing* activates the Wickelfeatures for #si, siN, and iN#.
- There is no representation of the order of the phonemes (beyond the phoneme triples).
- $460 \times 460 = 211{,}600$ connections (and weights) to be learned (no hidden layer).
- Initially, these connections are all set to 0.
- Then the model was trained with with 420 input/output pairs (verb baseforms paired with their past-tense forms).

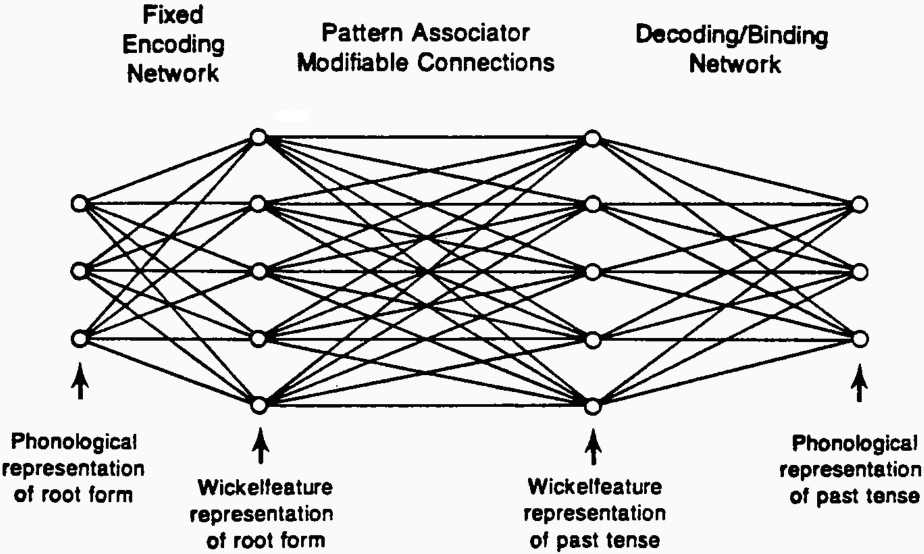
It's all in the Features

Time for a short quiz on Wooclap!

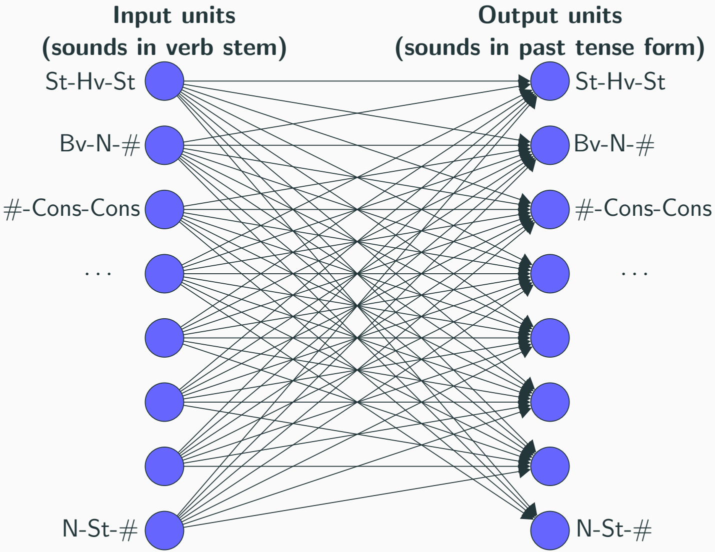
https://app.wooclap.com/MIMHYA

Model Structure

Model Structure



Model Structure



Results

And the Result Was . . .

- After the 84,000 training iterations, the network worked well for almost all the 420 verbs in the training set
- Performed adequately for separate test set of 86 other verbs
- 3/4 of regular verb stems were assigned the correct past tense
- Most irregular verbs stems were assigned overgeneralized regular past tense forms (e.g., *digged*, *catched*)

Childlike Behavior

- U-shaped learning: after a period of outputting *gave* correctly, the network shifted to the incorrect *gived.*
- Was reluctant to stick *-ed* on a stem ending in /t/ or /d/.
- Made lots of childlike errors, e.g., *cling-clang*, *sip-sept.*

Discontinuous Training

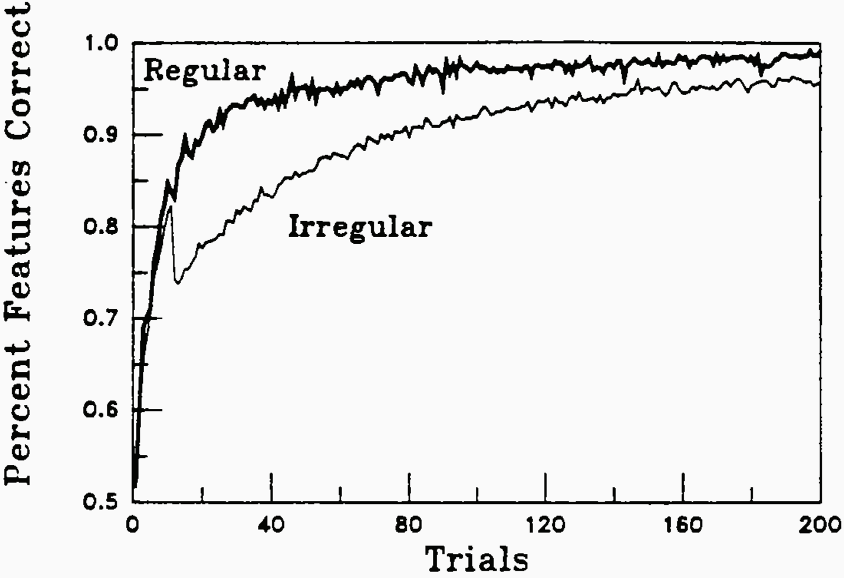
R&M's empirical observations

- children learn common verbs first, and rarer verbs later
- they tend to learn irregular verbs before regular ones
- children's vocabulary grows very quickly all of a sudden, a few months after they start learning words, i.e., at some point they get a huge spurt of regular verbs.

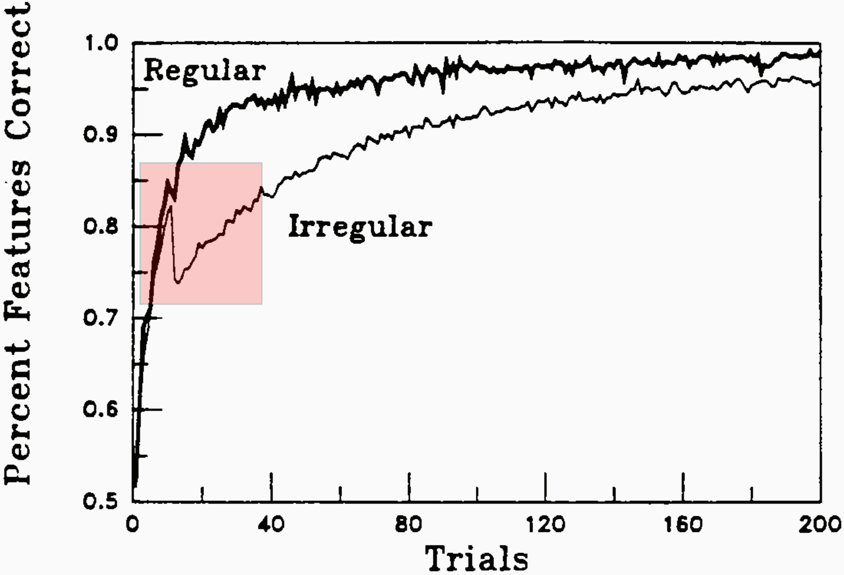
R&M's network training

- They first trained it on just 10 verbs, all at once, 8 irregular.
- And then trained it on 410 verbs, all at once, 80% regular
- Error rates increase dramatically at the start of the second training phase, before recovering gradually.
- Model started to make errors such as *breaked.*

U-shaped Learning Curve



U-shaped Learning Curve



What Does the Model Learn?

Q_1: Do parents start at some point using more regular verbs when talking to their children?

A_1: Data from spontaneous conversations involving children shows no evidence for this.

Q_2: Is there a vocabulary spurt and thus a richer mixture of regular verbs when children begin to over-apply *-ed*?

A_2: Children's vocabularies explosion starts in the mid-to-late ones, not in the mid-to-late twos (when children start to make over-regularization errors, new regular verbs are actually coming in more slowly than they were previously).

Q_3: What if we change the network's training scheme?

A_3: The training regime is fragile. But human language learning is robust (e.g., children exhibit similar learning curves, even with vastly different data to learn from).

Criticism of R&M's Model

Problem 1

R&M's model only produces past-tense forms; you cannot run the model backwards and recognize past-tense forms. Obviously, people do both!

Problem 2

The model computes every detail of the pronunciation of the past-tense form. Many details common in other parts of the language system. Should they be duplicated in different networks?

Problem 3

The representation in terms of a single block of Wickelfeatures is overly simplified, missing out important aspects of phonology.

Criticism of R&M's Model

What Wickelfeature representations can't do:

- They can't tell the difference between words that contain the same triples but in a different order.
- They can't deal with reduplication: /algal/ 'straight' and /algalgal/ 'ramrod straight' (Oykangand) have the same Wickelphones.
- They can't tell difference between words that sound alike (e.g., *break-broke*, *brake-braked*).
- Phonologically similar words such as /slit/ and /silt/ have completely different Wickelphones (but swapping of sounds is a common phonological process).

Criticism of R&M's Model

Time for a short quiz on Wooclap!

https://app.wooclap.com/MIMHYA

The Normal Course of Science

R&M's model illustrates the no rules, all memory extreme:

- it was sufficiently explicit to make testable predictions;
- researchers did experiments which appeared to conflict with those predictions;
- criticism led to the design of revised experiments;
- model also changed to fix flaws (e.g., different input representation, addition of hidden layer, rule-like mechanism).

We will briefly look at the Kiros & Cotterell (K&C) model, which uses state-of-the-art deep learning to model the past tense.

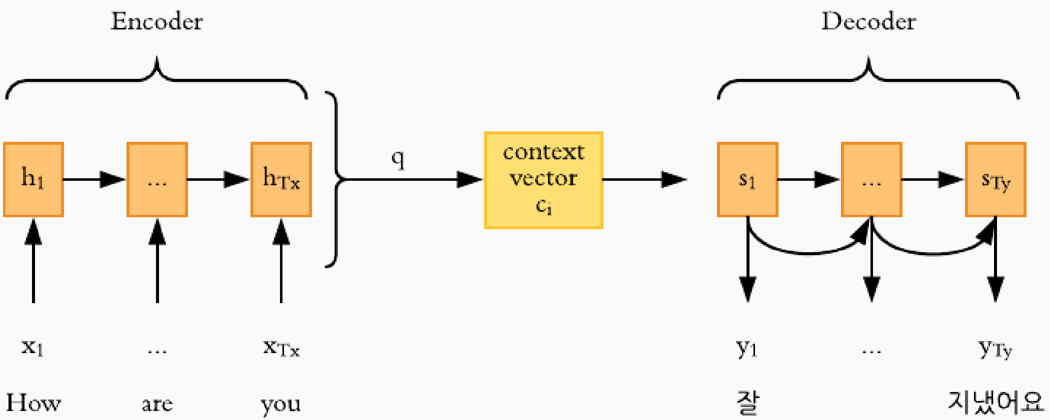
The Kiros & Cotterell Model

The Kiros & Cotterell Model

What Kiros & Cotterell do differently:

- use a *recurrent neural network* (RNN) instead of a perceptron;
- RNNs can represent input of variable length: great for words or sentences!
- no Wickelfeatures: the RNNs models sequences naturally; similar sequences get similar representations;
- use an *encoder-decoder* architecture, essentially two RNNs put together, one for the input and one for the output;
- use *attention,* which is a way of learning with parts of the input and output matter most;
- use *multitask learning,* i.e., train a single network to learn multiple phonological and morphological processes.

Encoder-Decoder Architecture



Kiros & Cotterell's Model

K&C compare to the Minimal Generalization Learner (MGL) of Albright and Hayes, the best available rule-based model:

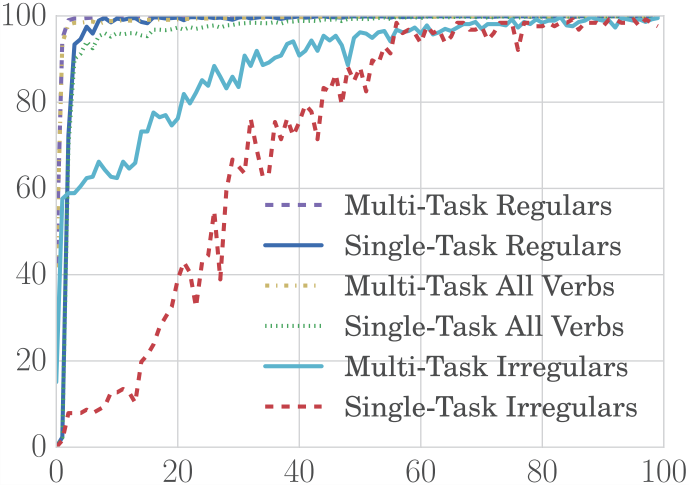
	all			regular			irregular		
	train	dev	test	train	dev	test	train	dev	test
Single-Task (MGL)	96.0	96.0	94.5	99.9	100.0	100.0	0.0	0.0	0.0
Single-Task (Type)	99.8†	97.4	95.1	99.9	99.2	98.9	97.6†	53.3†	28.6†
Multi-Task (Type)	100.0†	96.9	95.1	100.0	99.5	99.7	99.2†	33.3†	28.6†

MGL completely fails on irregular verbs.

Multitask learning improves performance slightly.

Kiros & Cotterell's Model

No evidence of U-shaped learning:



Kiros & Cotterell's Model

Evidence of oscillating development for individual verbs:

CLING		MISLEAD		CATCH		FLY	
#	output	#	output	#	output	#	output
5	[klɪŋd]	8	[mɪsliːdɪd]	7	[kætʃ]	6	[flaɪd]
11	[klʌŋ]	19	[mɪslɛd]	31	[kætʃ]	31	[fluː]
13	[klɪŋ]	21	[mɪslɛd]	43	[kɔt]	40	[flaɪd]
14	[klɪŋd]	23	[mɪslɛd]	44	[kætʃ]	42	[fleɪ]
18	[klʌŋ]	24	[mɪsliːdɪd]	51	[kætʃ]	47	[flaɪd]
21	[klɪŋd]	29	[mɪslɛd]	52	[kɔt]	56	[fluː]
28	[klʌŋ]	30	[mɪsliːdɪd]	66	[kætʃ]	62	[flaɪd]
40	[klʌŋ]	41	[mɪslɛd]	73	[kɔt]	70	[fluː]

After 40 epochs, these verbs are learned correctly.

Summary

- Simple learning model shows the characteristics of young children learning the past tense.
- Generates U-shaped learning curve for irregular verbs and exhibits tendency to overgeneralize similar to young children.
- Makes empirical predictions that can be tested.
- Manipulates actual data and can simulate rather than describe behavior; specific representations (e.g., Wickelfeatures).
- Is neural networks the right approach to learning? The jury is still out; it certainly challenges our understanding of how linguistic information is acquired and applied.

Next lecture: speech segmentation.