Connectionist and Statistical Language Processing

Lecture 5: Learning Phonology and Morphology

Matthew W Crocker

Computerlinguistik
Universität des Saarlandes

References: McLeod et al, Chapter 8 & 9, pages 155-194
Elman et al, Chapter 3, pages 130-147

Overview

- Reading Aloud: Orthography-Phonology
  - S&M, Plaut *et al* models of adult performance
  - Good performance on known and unknown words
  - Models (normal) human behaviour
  - Fails to replicate the double-dissociation (in acquired dyslexics)
  - Importance of input and output representations
- Language Acquisition: how do children acquire language
- English past-tense: Morphology
  - Forming the past tense from the present
  - Similarity: dual-route models to explain a double dissociation
  - Connectionist account: a singe mechanism
- Learning vocabulary: lexical development

Reading Aloud

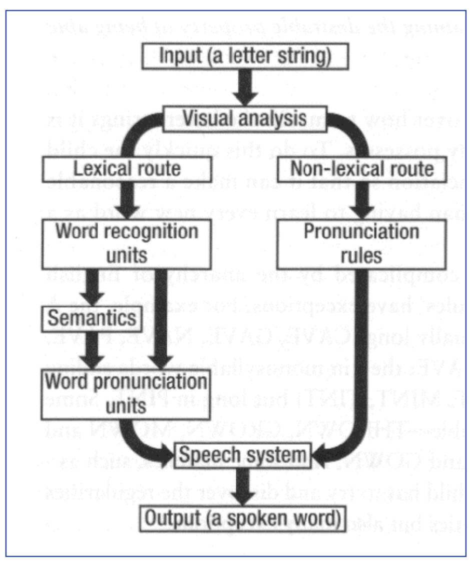
- Task: produce the correct pronunciation for a word, given its printed form

- Suited to connectionist modelling:
  - Since we need to learn mappings from one domain (print) to another (sound)
  - Multi-layer networks are good at this, even when mappings are somewhat arbitrary
  - Human learning is similar to network learning:
    - I.e. learning takes place gradually, over time
    - Incorrect attempts are often corrected

- If a network can't model this linguistic task successfully, it would be a serious blow to connectionist modelling. But …

Dual Route Model

- The standard model of reading posits two independent routes leading to pronunciation of a word, because …
  - People can effortless pronounce words they have never seen:
    - SLINT or MAVE
  - People can pronounce words which break the "rules":
    - PINT or HAVE
- One mechanism uses general rules for pronunciation
- The other mechanism stores pronunciation information with specific words



Behaviour of Dual-Route Models

- Consider: KINT, MINT, and PINT

- KINT is not a word:
  - No entry in the lexicon
  - Can only be pronounced using the “rule-based” mechanism
- MINT is a word:
  - Can be pronounced using the “rule-based” mechanism
  - But exists in the lexicon, so also can be pronounced by the “lexical” route
- PINT is a word, but irregular
  - Can only be correctly pronounced by the lexical route
  - Otherwise, it would rhyme with MINT

Evidence for the Dual-Route Model

- Evidence from neuropsychology show different patterns of behaviour for two types of brain damage (acquired *after* learning):

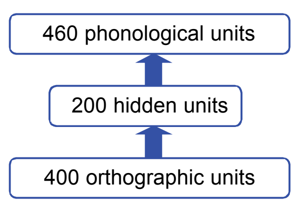
- Phonological dyslexia
  - Symptom: Read words without difficulty, but cannot produce pronunciations for non-words
  - Explanation: Damage to rule-based route; lexical route intact
- Surface dyslexia
  - Symptom: Can pronounce words and non-words correctly, but make errors on irregulars (tendency to regularise)
  - Explanation: Damage to the lexical route; rule-based route intact

- All Dual-Route models share:
  - a lexicon for known words, with specific pronunciation information
  - A rule mechanism for the pronunciation of unknown words

Towards a Connectionist Model

- It is unclear how a connectionist model could naturally implement a dual-route model:
  - No obvious way to implement a lexicon to store information about *particular* words; storage is typically *distributed*
  - No clear way to distinguish “specific information” from “general rules”; only one uniform way to store information, weights of connections
- Examine the behaviour of a standard 2-layer feedforward model
  - Seidenberg & McClelland (1989)
  - Trained to pronounce all the monosyllabic words of English
  - Learning is implemented using the backpropagation algorithm

Seidenberg and McClelland (1989)



- 2 layer feed-forward model:
  - Distributed representations at input and output
  - Distributed knowledge within the net
  - Gradient descent learning
- Input and Output
  - Inputs are activated by the letters of the words
    - 20% activated, on average
  - Outputs represent the phonological features
    - 12% activated, on average
  - Encoding of features does not affect the success
- Processing: $\text{netinput}_i = \sum_j a_j w_{ij} + \text{bias}_i$
  - Activation of a node is calculated using the logistic function

Training the Model

- Learning
  - Weights and bias are initially random
  - Words are presented, and outputs are computed
  - Connection weights are adjusted based on backpropagation of error
- Training
  - All monosyllabic words of 3 or more letters (about 3000) words
  - In each epoch, a sub-set was presented
    - Frequent words appeared more often
  - Over 250 epochs, (THE) was presented 230 times, least common 7 times
    - (THE) is actually 100000 times more likely, but this doesn't change learning
- Performance
  - Outputs were considered correct if the pattern was closer to the correct pronounciation than that of any word
  - After 250 epochs, accuracy was 97%

Results: Seidenberg & McClelland

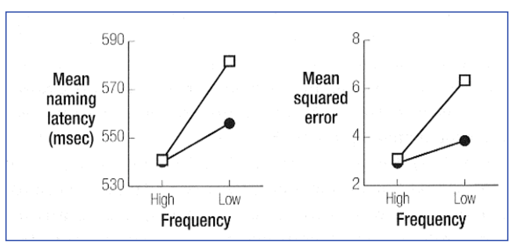
- The model does successfully learn to map most regular and irregular word forms to their correct pronunciation
  - It does this without separate routes for lexical or rule based processing
  - There is no word specific memory
- It does not perform as well as human in pronouncing non-words
- Naming Latency:
  - Experiments have shown that adult reaction times for naming a word is a function of variables such as word frequency and spelling regularity
- The current model cannot directly mimic latencies, since the computation of outputs is constant
- The model can be seen as simulating this observation if we relate the output error score to latency
  - Phonological error score is the difference between the actual pattern and the correct pattern
  - Hypothesis: high error should correlate with longer latencies

Word Frequency Effects

- Common words are pronounced more quickly than uncommon words
  - This is true for most almost all aspects of human information processing
- Conventional (localist) explanation:
  - Frequent words require a lower threshold of activity for "the word recognition device" to "fire"
  - Infrequent words require a higher threshold of activity
- In the Seidenberg & McClelland model, naming latency is modelled by the error:
  - Word frequency is reflected in the training procedure
  - Phonological error is reduced by training, and therefore lower for high frequency words
- The explanation of latencies in terms of error follows directly from the network's architecture and the training regime

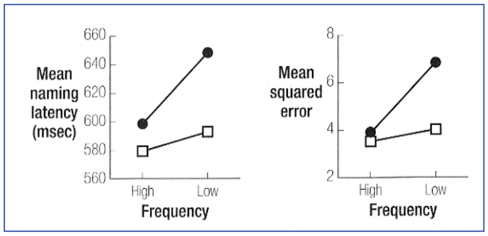
Frequency x Regularity

- In addition to faster naming of frequent words, human subjects exhibit:
  - Faster pronunciation of regular words (e.g GAVE or MUST) than irregular words(e.g. HAVE or PINT)
  - But, this effect interacts with frequency: it is only observed with low frequency words
- For regulars (filled circle) we observe a small effect of frequency
  - It takes slightly longer to pronounce the low frequency regulars
- For irregulars (open square) we observe a large effect of frequency
- The model precisely mimics this pattern of behavior in the error
- 2-route: the confusion of the lexical and rule outcome requires resolution
  - Lexical route wins faster for high frequency words



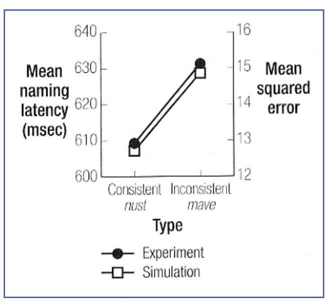
Frequency x Neighborhood Size

- The "neighborhood size" of a word, is defined as the number of words that differ by changing one letter.
- Neighborhood size has been shown to also affect naming latency much the same way as regularity:
  - Not much influence for high frequency words
  - Low frequency words with small neighborhoods (filled circles) are read much more slowly than words with large neighborhoods (open squares)
- Shows "cooperation" of the information learnt in response to different(but similar) inputs
- Again, the connectionist model directly predicts this
- The 2 route model requires a more *ad hoc* explanation, grouping across localist representations of the lexicon



Spelling-to-Sound Consistency

- Consistent spelling patterns: _UST
  - All words have the same pronunciation
- Inconsistent patterns are those with more than one: _AVE
- Observation: adult readers produce pronunciations more quickly for non-words derived from consistent patterns (NUST) than from inconsistent patterns (MAVE)
- This is difficult for 2-route models:
  - Since both are processed by the non-lexical route
  - Consistent and inconsistent rules would need to be distinguished
- The error in the connectionist model predicts this latency effect perfectly



Summary of Seidenberg & McClelland (1989)

- What has the model achieved
  - The model is a single mechanism with no lexical entries or explicit rules
  - Response to an input is a function of the networks entire experience
    - Reflects previous experience on a particular word
    - Experience with words resembling that string
- E.g. specific experience with HAVE is sufficient to overcome the general information that _AVE is usually a long vowel
- The network can produce a plausible pronunciation for MAVE, but error is introduced by experience with inconsistent words like HAVE
- Performance
  - 97% accuracy on pronouncing learned words
  - Models: frequency & interaction with regularity, neighborhood, consistency
- Limitations: It is not as good as humans at
  - Reading non-words (model get 60%, humans 90%)
  - Lexical decision (FRAME is a word, but FRANE is not)

Representations are important

- Position specific: for inputting words of maximum length N:
  - N groups of 26 binary inputs = word
- But consider: LOG, GLAD, SPLIT, GRILL, CRAWL
  - The model needs to learn the correspondence between L and /l/
  - But L always appears in different positions
  - Learning different pronunciations for different positions should be straightforward
  - Allignment: letters and phonemes are not in 1-to-1 correspondence
- Problem: non-position-specific loses important order information:
  - RAT = ART = TAR
- Solution: S&M decompose word and phoneme strings into "triples"
  - FISH = _FI SH_ ISH FIS
  - Each input is associated with 1000 random triples
  - Active is that triple appears in the input word

Wickelfeatures

- S&M still suffer some specific effects
  - Information learned about a letter in one context is not easily generalised

Improving the Model: Plaut et al (1996)

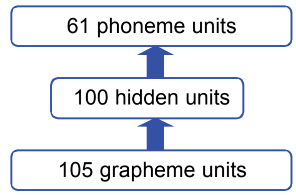
- Plaut et al (1996) solution: non-position-specific + linguistic constraints
  - Monosyllabic word = onset + vowel + coda
  - Strong constraints on order within these clusters
    - E.g, if 't' and 's' are together, 's' always precedes 't'
  - Only one set of grapheme-to-phoneme units is required for the letters in each group
  - Correspondences can be pooled across different words, even when letters appear in different positions
- Input representations:
  - Onset: first letter or consonant cluster (30)
    - y s p t k q c b d g f v j z l m n r w h ch gh gn ph ps rh sh th ts wh
  - Vowel (27)
    - e I o u a y ai au aw ay ea ee ei eu ew ey ie oa oe oi oo ou ow oy ue ui uy
  - Coda: final letter or consonant cluster (48)
    - h r l m n b d g cxf v j s z p t k q bb ch ck dd dg ff gg gh gn ks ll ng nn ph pp ps rr sh sl ss tch th ts tt zz u e es ed
- Monosyllabic words are spelt by choosing one or more candidates from each of the 3 possible groups:
  - THROW: ('th' + 'r'), ('o'), ('w')

Output representations

- Phonology: groups of mutually exclusive members
  - Onset (23)
    - s S C
    - z Z j f v T D p b t d k g m n h
    - l r w y
  - Vowel (14)
    - a e i o u @ ^ A E I O U W Y
  - Coda (24)
    - r
    - l
    - m n N
    - b g d
    - ps ks ts
    - s z
    - f v p k
    - t
    - S Z T D C j
- "Scratch" = 's k r a _ _ _ _ _ _ _ _ C'

The network architecture

- The architecture of the Plaut *et al* network:
  - The are a total 105 possible orthographic onsets, vowels, and codas
  - The are 61 possible phonological onsets, vowels and codas



- Performance of the Plaut *et al* model:
  - Succeeds in learning both regular and exception words
  - Produces the frequency x regularity interaction
  - Demonstrates the influences of frequency and neighbourhood size
- What is the performance on non-words?
  - For consistent words (HEAN/DEAN): model (98%) *versus* human (94%)
  - For inconsistent words (HEAF/DEAF/LEAF): model (72%), human (78%)
    - This reflects production of regular forms: both human & model produced both
- Highlights the importance of encoding … how much knowledge is implicit in the coding scheme

Summary

- Word frequencies:
  - Seidenberg & McClelland presented training materials according to the log frequencies of words
  - People must deal with absolute frequencies which might lead the model to see low frequency items too rarely
  - Plaut *et al* model, however, succeeds with absolute frequencies
- Representations:
  - The right encoding scheme is essential for modelling the findings
    - How much linguistic knowledge is “given” to the network by Plaut’s encoding?
  - They assume this knowledge could be partially acquired prior to reading
    - I.e. children learn to pronounce “talk” before they can read it
  - Doesn’t scale to polysyllabic words
- Doesn’t not explain the double dissociation:
  - ✓ Surface dyslexics (can read exceptions, but not non-words)
  - ✗ Phonological (can pronounce non-words, but not irregulars)

Connectionist models of Acquisition

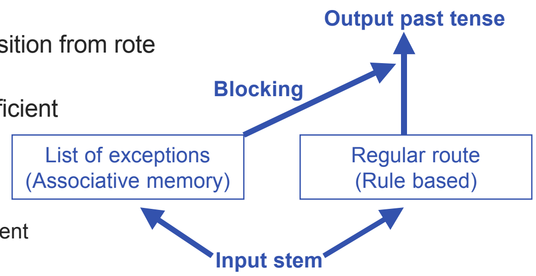
- Symbolic models emphasise the learning of rules and exceptions
- Connectionist models have no direct correlate to such mechanisms
  - Knowledge is stored in a distributed weight matrix
- Models of learning:
  - Start state of the cognitive system
  - Learning mechanism
  - Training environment
  - Acquired skill
- Connectionist models provide a opportunity to model the learning process itself, not just the resulting acquired skill
  - We can test connectionist models against developmental data, at various points during learning
  - Discontinuities in performance (sudden changes in behaviour) can be explained by "emergent properties" of a single, continuous mechanism

Learning the Past Tense

- The problem of past tense formation:
  - Regular formation: stem + 'ed'
  - Irregulars do show some patterns:
    - No-change: hit » hit (all end in a 't' or 'd')
    - Vowel-change: ring » rang,. Sing » sang (rhymes often share vowel-change)
    - Arbitraty: go » went
- Young children often form the past tense of irregular verbs (like GO) by adding ED: <u>overregularisations</u>
  - *"go"+"ed" » "goed"*
- This suggests incorrect application of a <u>learned rule</u>, not just rote learning or imitation
- Overregularisations often occur after the child has already succeeded in producing the correct irregular form: *"went"*
- Thus we need to explain this "U-shaped" learning curve

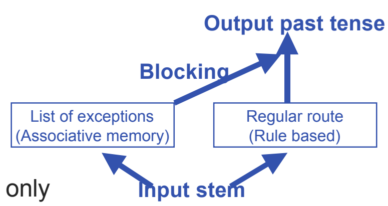
A Symbolic Account: Dual-Route Model

- General pattern of behaviour:
  - Early: children learn past tenses by rote (forms are stored in memory)
  - Later: recognise regularities, add general device to add 'ed' suffix
  - Now: no need to memorise forms, but this leads to incorrect generalisation of the regular rule to irregulars
  - Finally: distinguish which forms can be generated by the rule, and which must be stored (and accessed) as exceptions
- A Dual Route Model:
  - Errors result from the transition from rote learning to rule-governed
  - Recovery occurs after sufficient exposure to irregulars:
    - Increased "strength"
    - Frequency based
    - Faster recovery for frequent irregulars



The Dual-Route Model

- As with reading aloud, this proposal requires two qualitatively different types of mechanism
- Accounts for the observed dissociation:
  - Children make mistakes on irregulars only



- Evidence for double dissociation (Pinker 1994)
  - In some language disorders, children preserve performance on irregulars but not regulars
  - In other disorders, the opposite pattern is observed
- Accounts for the U-shaped learning curve
  - And since irregulars differ in "representational strength" it explains why overregularisation of high frequency irregulars is uncommon
- No explicit account of how the "+ed" rule is learned

Language Acquisition

- Perhaps the notion of inflection is innately specified, and need not itself be learned:
  - The inflectional mechanism is triggered by the environment or maturation
  - Then the exact (language specific) manifestation must be learned

- Criticisms:
  - Early learning tends to be focussed on irregular verbs
  - Irregular sub-classes (hit, sing, ring) might lead to incorrect rule learning
    - These do occurs, but typically late in learning
    - How are good/spurious rules distinguished and selected
  - English is unusual in possessing a large class of regular verbs
    - Only 180 irregulars
  - Only 20% of plurals in Arabic are regular
  - Norwegian has 2 regular forms for verbs: 3 route model ?

Towards a Connectionist Model

- No distinct mechanisms for regular and irregular forms
- No innately specified maturation stage or rules to be triggered

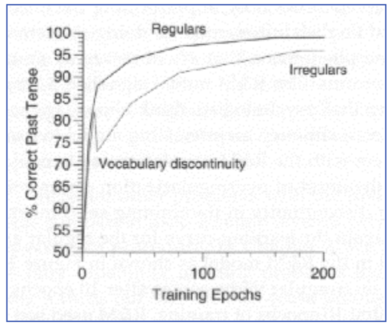
- Parsimonious:
  - Simplifies the structural complexity of the starting state
  - Learning exploits the structure of the learning environment

- Rummelhart and McClelland (1986)
  - 1st attempt to model this problem (or any development system)
  - Modelled U-shaped learning, but heavily criticised (Pinker & Prince 1988)

- Plunkett & Marchman
  - Use a feed-forward network, one hidden layer

Rummelhart and McClelland (1986)

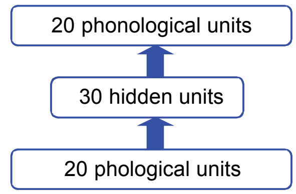
- A single-layer feed-forward network (perceptron)
  - Input: is a phonological representation of the stem (wickelfeatures)
  - Output: is a phonological representation of the past tense (wickelfeatures)
  - Trained using the perceptron learning rule
- Training:
  - First trained on 10 high frequency verbs (8 irregular, 2 regular), 10 epochs
  - Perfect performance
  - Then 420 (medium frequency) verbs (80% regular), 190 epochs
  - Early in training, shows tendency to overregularise, i.e. modelling stage 2
  - End of training, exhibits “adult” (near perfect) performance
  - Generalised reasonably well to 86 low frequency verbs in test set



Performance of R&M (1986)

- Criticisms:
  - Problems with representation using wickelphones/wickelfeatures
  - U-shape performance depends on sudden changes from 10-420 in the training regime
  - Rote learning of first 10 verbs: there was no generalisation to novel stems after 10 epochs
  - Most of the 410 new verbs are regular, overwhelming the network and leading to overregularisation
- Justification: children do exhibit vocabulary spurt at end of year 2
  - But overregularisation errors typically occur at end of year 3
  - Vocabulary spurt is mostly due to nouns
- Single layer Perceptron only works for linearly separable problems
  - Plunkett & Marchman (1991) show residual error remains after extensive training
  - Suggests a hidden-layer network

Plunkett and Marchman (1993)



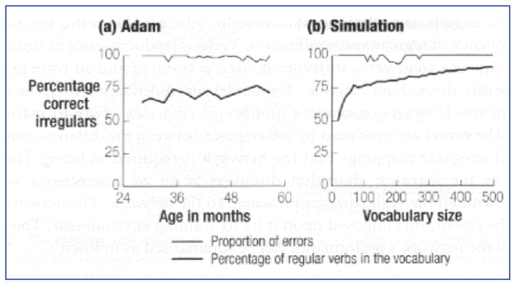
- A standard feed forward network with one hidden layer
- Maps a phonological representation of the stem to a phonological representation of the past tense
- Initially, the model is trained to learn the past tense of 10 regular and 10 irregular verbs
  - Represents currents estimates of children's early vocabulary
- Training proceeds using the standard backprop algorithm, in response to error between actual and desired output
  - Is this plausible?
- Learning must configure the network for both regulars and irregulars
  - Consider: *hit » hit*, but *pit » pitted*
  - We know multi-layer networks can do this, but considerable training may be required

Plunkett and Marchman (continued)

- Training:
  - Initial period of 10 regular and 10 irregular verbs
  - Then vocabulary was gradually increased, to mimic the gradual uptake of words in children
  - Total: 500 word stems, 90% regular (similar to the relative frequency of regulars in English)
  - Higher frequency verbs were introduced earlier in training, and so were also presented to the network more often
    - Irregulars are more frequent, so appear more often in training
    - This is essential, otherwise the regulars swamp the network
    - Arguably more accurately reflects the childs learning environment
- The final model successfully learned the 500 verbs in the training set
  - But errors were made *during* the learning phase
  - Caused by interference between mappings for regulars and irregulars before mature connection weights have been discovered

Performance of P&M

- Early acquisition is characterised by a period of error free performance
- Low overall rate (5-10%) of overregularisation errors
- Overregularisation is not restricted to a particular period of development
- Common irregulars do not exhibit overregularisation (e.g. 'goed' is rare)
- Errors are phonologically conditioned: No change verbs (hit) are robust to overregularisation (e.g 'hitted' is rare)
- Only a very small number of irregularisation errors are observed (e.g. where the network produces 'bat' for 'bite')
- Generally compatible with the results of studies by Marcus *et al* (1992):
  - Early performance is error free, and then low error is more or less random



Discussion

- Performance is close tied to the training environment:
  - Onset of overregularisation is closely bound to a "critical mass" of regular verbs entering the child vocabulary
  - This subsides as the training learns the final solution for the task
- Highly sensitive to training environment:
  - Requires more training on arbitrary irregulars (go/went), which are highly frequent in the language
  - More robust for no-change verbs (hit, put) which are more numerous (type) and less frequent (token)
- Models the frequency x regularity interaction:
  - Faster reaction time for high frequency irregulars than low frequency ones
  - No advantage for regulars
- Differential behaviour for regulars and irregulars result from lesioning
- Suggests it is dangerous to infer dissociations in mechanisms due to observed dissociations in behaviour
  - Critical mass effect can have the appear of a distinct mechanism

Criticism

- We know multi-layered networks can learn such mappings in general; not proof that children use the same type of mechanism
- Pinker & Prasada argue that the (idiosyncratic) statistical properties of English help the model:
  - Regulars have low token frequency but high type frequency: facilitates the generalisation across this class of items
  - Irregulars have low type frequency but high token frequency: facilitates rote learning mechanism for these words
- They argue no connectionist model can accommodate default generalisation for a class which has both low type and token frequency
  - Default inflection of plural nouns in German appear to have this property
- No explanation of the double-dissociation observed by Pinker (1994)

Main conclusions

- Dissociations in performance, do not necessarily entail distinct mechanisms:
  - Reading aloud: a singe mechanism explains regular and irregular pronunciation of monosyllabic rules
  - Past tense: a single model of regular and irregular past tense formation
- But, explaining double dissociations is difficult
  - Has been shown to be possible on small networks, but unclear if larger (more plausible) networks can demonstrate double dissociations
- Connectionist models excel at finding structure and patterns in the environment: "statistical inference machines"
  - The start state for learning may be relatively simple, unspecified
  - Necessary constraints to aid learning come from the environment
- Can such models scale up? Are they successful for languages with different distributional properties?
- Tutorial: The English Past Tense, chapter 11 of Plunkett & Elman