Connectionist Sentence Processing in Perspective

Review by Eric Auer – Original Paper by Mark Steedman

January 31, 2001

What we will look at

Comparing symbolist and connectionist approaches

- Simple Recurrent Networks as Graded State Machines
- Associative Nets/Memories storing structure
- Conceptual Structures vs. Innate Knowledge

1. Both sides of the story

Modular syntactic processing

- Rule-based approach: Grammar (rules), Algorithm/Automaton, Oracle
- Connectionist approach: May be modular, but often with distributed repn
- Will NN develop rules or are they just useful as oracle?
- Most NN do not scale well due to problems with back propagation

Various kinds of grammar

- Competence grammar: Modelling the theoretic intuition
- Performance grammar: Modelling human performance
- Covering grammar: Accepting the same strings as a given grammar ($\Rightarrow$ weak equivalence)

Embedding and performance

On human performance with embedding

- Chomsky: Recursive embedding $\Rightarrow$ Natural Language at least context free
- Maybe humans only use a limited weakly equivalent FSM covering grammar?
- Possible reasons: Memory limitations (FSM too simple), incomplete algorithm, something misleads oracle

The preferred interpretation

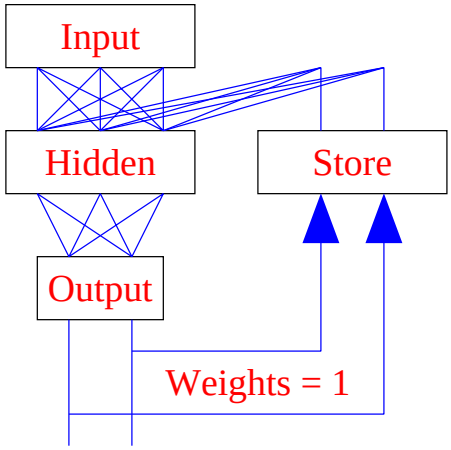
Interpretation problems

- Syntactic derivations should give extractable interpretation (structure, logical forms)
- But interpretation may also build up in flow of control only
- Model theory tells us about things like truth of a proposition for both
- But for practical use, we want to know WHY the proposition is true

2. Recurrent networks

Jordan's RN

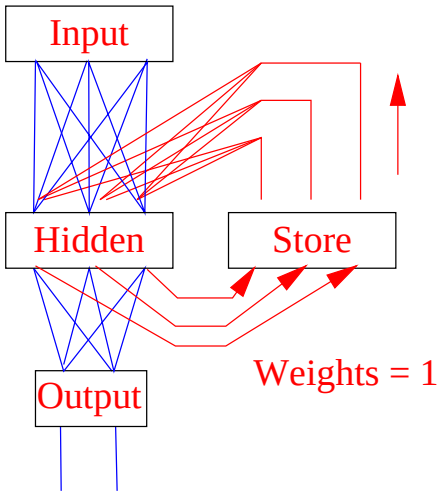
- Used to model motor control problems
- Example: coarticulation in speech
- Feedback of copied past output
- Quite limited flexibility



Elman's Simple Recurrent networks

Elman's SimpleRN

- Used to predict next word/category . . .
- Feedback of copied past hidden unit state
- Copes with abstract properties of sequences and distant dependencies
- Able to learn limited CFG (with center embedding) and Trans-CFG (crossing deps like in verb raising)
- Seems to implement a "graded state" automaton (maybe ≈ FSM)



All categories ≠ meaning

Category prediction is only part of the problem

- Predicting all categories does not mean grammar understanding
- Even after disambiguation of each word, global ambiguity may remain. Example: "Put the block in the box on the table"

So what do we use the SRN for?

- POST performance does not account for full human performance
- The SRN is good at tagging with sub-categorization
- If SRN could use BTT, it may capture some aspects of grammar
- Todays SRN are useful for POST, but no emergent grammar

3. Psychological relevance of RN

Simulating human performance

- Tabor and Tanenhaus: Added approximation of BTT
- PCA and "gravitational" analysis: The SRN acts similar to a n-gram based stochastic POST
- T. and T. predicted processing effort / reading times involving thematic fit

Simulating human performance

- The cop arrested by the detective . . .
- There is the misleading idea that the cop might be the subject
- Symbolist simulations of this "garden path" effect involve structure
- The SRN seems to capture this without structure, but a FSM might do, too.

Plausibility checking ⇒ inference?

Limitations of SRN predictions

- Semantic and pragmatic plausibility needs more processing than n-gram analysis can do
- Context may even come from "between the lines"
- Crain, Altman and Steedman: Mentioning zero or one cop before supports the cop-is-subject interpretation of "The cop arrested by the detective was guilty"
- . . . but mentioning some cops supports the cop-is-object interpretation (being arrested as a distinguishing feature)
- If context can have that impact, inference rather than n-gram frequency knowledge must be involved

Syntactic priming and a new trend

The study of Dell et al.

- A SRN and a production network have their hidden layers connected
- Presenting a sentence in active/passive voice biases the PN . . .
- So there seems to be syntactic priming without involving rules
- But this effect may even be not syntactic at all but involves only changing transition probs in a FSM

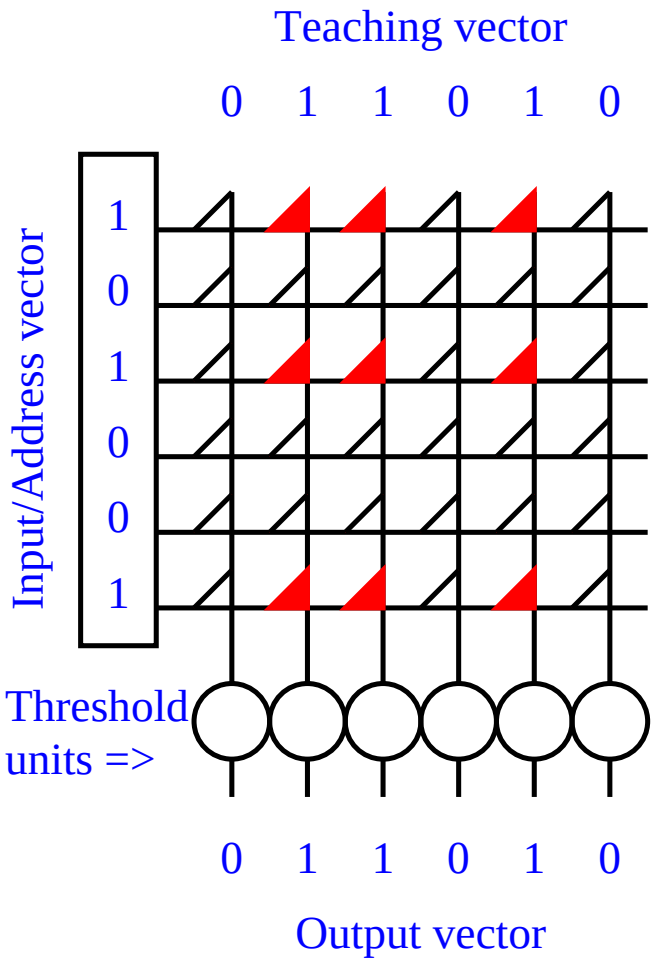
A new way to use SRNs

- A SRN may act like a symbolic POST or a HMM
- There is a trend in symbolic stochastic language processing towards carrying probs into the grammar
- So the SRN can act as a special POST or be part of the lexicon

4. Old school associative memory

Willshaw et al.: The 70s way

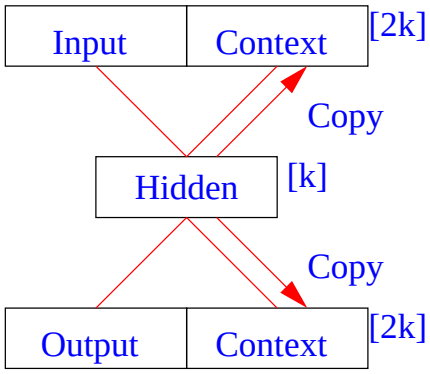
- Very simple concept
- Efficient distributed storage of vectors/pointers
- Interesting noise- and damage-resistance properties



The LRAAM – another AM

Pollack 1990: Recursive Auto-AM

- Structure similar to standard feed-forward NN
- Additional units may store labels (LRAAM) or content
- Stores a parse tree or other recursive structure
- Encoding may be optimized by approximating the underlying CFG
- It is not likely that the RAAM learns grammar from trees



The RAAM and beyond

Suggested improvements

- Poor scaling properties due to back propagation may be solved by Plate's Holographic Reduced Repns
- A recurrent network parser could be added
- Adding of stack or finite state control would introduce more symbolist concepts . . .
- . . . but the wild romance of pure SRN and the error tolerance of the AM will get lost

Do we need innate knowledge?

Grammar induction from strings alone won't work

- Grammar induction from strings would require a huge corpus
- Therefore, symbolists claim there must be innate knowledge of language

Universal Conceptual Structures

- UCS may emerge from how the sensory-motor-apparatus, memory etc. are structured (biol. plausible)
- As Chomsky pointed out, we cannot observe the nature of UCS. UCS could emerge "Universal Grammar".
- words plus UCS $\Rightarrow$ logical forms, plus categories $\Rightarrow$ language specific lexicon!?
- AM are a useful way to store the lexicon in a distributed form

5. The lexicon – language specific

A trend to do things with the lexicon

- Language specific phenomena get stored in the lexicon today for LFG, CCG, HPSG, LTAG and some versions of GB
- Advantage: Close integration of lexicon / syn / sem / phon
- E.g. CCG associates directional syn type, logical form and phon type with each word/constituent
- Advantage: Non-standard constituents for coordination and intonation modelled in an elegant way, grammar acquisition is reduced to finding the syn type for each word (eg. S/NP vs. S\NP)
- AM and RN can do grammar acq. that way, FSM or SRN (used as graded SM) used as distributed category lexicon
- Problem: If SRN stores categories in a very distributed way, the mapping between meaning and category will be hard to find

The Optimality Theory approach

Neural Networks for "soft" constraints

- Grimshaw: categorization vs. constraint-satisfaction problem
- Elegant solution with OT (implements ordered and "soft" constraints), may be similar to human lexicon (not parser)
- OT constraint systems are similar to finite state transducers, so an AM based lexical acq. device seems plausible

Probabilistic guidance

- Another approach (Collins and Charniak 1997): guide grammar by probs (from POST . . .)
- Fits quite nice to lexicalized grammars like CCG, HPSG and LTAG
- Use NN to find the guiding probs?

6. Semantics: hard for symbolists?

Parsers want semantic information

- Decisions during parsing sometimes involve semantic information
- So we need early semantic analysis for disambiguation
- Classic approaches use statistical methods for that (not really good)
- Using Knowledge Representation Systems is not feasible for most domains

NN solving semantic problems?

Problems with decomposition of meaning

- There is very little decomposition below morpheme level
- Example: kill only decomposes to CAUSE and DIE, and CAUSE is not even the same as cause
- So we would need to feed lots of concepts into a KRS

NNs trade "efficiency for obscurity"

- NN with their distributed repn seem well suited to capture concepts and their relations
- We are not likely to understand the internal repn, but we get efficiency and learnability in return to this problem

7. Conclusion: A new model

Having UCS emerge – an alternative to innate knowledge

- As we have seen before, UCS are a plausible basis for language learning
- UCS can be seen as an emergent feature of the sensory-motor development
- Therefore a promising research program would be simulating that process

First phase: Conceptualizing the prelinguistic world

- Primary bodily actions and sensations
- Coordinating and primary actions like reaching
- conceptualizing identity, permanence and location of objects . . .

Building UCS continued

First phase continued: Identity / permanence / location of objects . . .

- . . . independent of their percepts and later independent of the actions they are involved in (not involving child's actions)
- Objects involved in events: Intrinsic events like falling
- Events with multiple/intermediate participants, tools and goals

Now we have learned most of the prerequisites for language acquisition during the sensory-motor development phase, next is language learning itself.

Learning language from UCS

Second phase (UCS now learned): Language learning

- deictic terms (this/that . . .)
- markers of topic, comment and contrast
- common nouns
- spatial and path terms
- causal verbs
- modal and propositional attitude verbs
- temporal verbs

To get this project working, techniques that scale better than back propagation seem quite important.

If the project succeeds, the emerging semantics could make us view phenomena like quantification, modality, negation and variable-binding in new ways — so the challenge seems really worth it.