Spectacular successes and failures of recurrent neural networks applied to language

Marco Baroni

Facebook AI Research
Recurrent neural networks

external input

state of the network at the previous time step

output
Recurrent neural networks
The "unfolded" view

Recurrent connections
Language modeling

cat \rightarrow sat \rightarrow on \rightarrow the \rightarrow mat

the \rightarrow cat \rightarrow sat \rightarrow on \rightarrow the \rightarrow mat
What are RNNs learning about language?

I’ve seen cat, I bet mat will follow.
What are RNNs learning about language?

I’m inside a PP, I’m waiting for the embedded NP
Do RNNs uncover the structure-dependent nature of language in raw linguistic input?

• Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen and Marco Baroni. *Colorless green recurrent networks dream hierarchically*. To be presented at NAACL 2018

• Inspired by: Tal Linzen, Emmanuel Dupoux & Yoav Goldberg TACL 2016.
Long-distance number agreement
Structures, not strings!
Evaraert et al. 2015

• The **boy** is jumping
• The **boy** [that I saw] **is** jumping
• The **boy** [that I saw yesterday] **is** jumping
• The **boy** [that I saw yesterday on the rocks] **is** jumping
Abstracting away from semantic and frequency effects

Dogs [in captivity] bark

"Grammar is best formulated as a self-contained study independent of semantics. In particular, the notion of grammaticalness cannot be identified with meaningfulness" (Chomsky 1957)

• Colorless green ideas sleep furiously
• The colorless green ideas [that I ate] sleep furiously
• The colorless green ideas [that I ate yesterday] sleep furiously
• The colorless green ideas [that I ate yesterday with the chair] sleep furiously
Building long-distance number agreement test sets

- Examples automatically extracted from Universal Dependencies treebanks (http://universaldependencies.org/)
- At least 3 tokens between cue and target
- Each original example transformed into a nonce sentence by replacing content words with morphologically-matched random words
- Languages studied: Italian, English, Hebrew, Russian
Extracting examples (1)

• NOUN [relative clause/participial phrase] VERB
  • la domanda [che vi viene rivolta] è
    the question that you-cl is asked is

• NOUN [relative clause/participial phrase] clitic VERB
  • il trono [che la guerra fredda aveva concesso alla Bomba] si scopre
    the throne that the cold war had granted to-the Bomb CL discovers

• NOUN [adjectival phrase] relpron VERB
  • Ci sono momenti [importanti ,] che finiscono
    there are moments important , that end
Extracting examples (2)

• NOUN [pp] VERB (participial)
  • la politica [di isolamento dall' estero] seguita
    the policy of isolation from abroad followed-sg

• NOUN [pp] adverb ADJ
  • l' ultimo palazzo [del potere] ancora intatto
    the last palace of power still intact

• DET [adjectival phrase] NOUN
  • una [vera e propria] avventura
    a real and proper adventure
Extracting examples (3)

- ADJ [conjoined adjectives] ADJ
  - Sono soprattutto tedeschi [, austriaci e] italiani
    are especially German-pl, Austrian-pl and Italian-pl

- VERB [verb complements] conj VERB
  - ha estratto [la pistola] e sparato
    has pulled-out-sg the gun and shot
Long-distance construction extraction stats

- Italian: 8 constructions, 119 sentences
- English: 2 constructions, 82 sentences
- Hebrew: 19 constructions, 746 sentences
- Russian: 22 constructions, 884 sentences
Generating colorless green sentences

• For each sentence prefix in treebank matching one of the selected constructions, replace each content word with a random word having the same morphological features:
  
  • L’ articolata disciplina [prevista dall’ art. 15] serve OR servono
    The articulate discipline envisaged by-the art. 15 serves OR serve (the purpose)
  • L’ opaca bomba [sottoposta dall' alloggio 200] pensa OR pensano
    The opaque bomb submitted by-the lodging 200 thinks OR think

• 9 nonce sentences generated for each example of each construction found in treebank
Training the RNN

• RNN trained on 80M words of Wikipedia, using language modeling method
  • Predict the next word given context
• Words are presented as unanalyzed primitives
  • RNN has no access to morphological structure
• RNN is not explicitly exposed to syntactic constituents
• None of the target constructions occurs in more than 1% of sentences in the relevant language
Training the RNN
The gory details

• Training with minibatch stochastic gradient descent for 40 epochs
  • Dividing learning rate by 4 if validation performance (perplexity) goes up at end of epoch

• Hyperparameters selected based on model performance on validation set, not long-distance agreement accuracy:
  • Italian: LSTM, 2 layers, 650 hidden units, batch size 64, dropout 0.2, learning rate 10
  • English: same as Italian
  • Hebrew: same as Italian, except dropout 0.1, learning rate 20
  • Russian: same as Italian, except dropout 0.2, learning rate 20
Eliciting grammaticality judgments from the trained RNN

At test time, RNN is fed sentence prefix, and it assigns probability to singular/plural continuations: it is said to be "correct" if it assigns higher probability to correct continuation.
## Number prediction accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>RNN original</th>
<th>RNN nonce</th>
<th>5-gram LM original</th>
<th>5-gram LM nonce</th>
<th>most freq original</th>
<th>most freq nonce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>95%</td>
<td>87%</td>
<td>64%</td>
<td>53%</td>
<td>55%</td>
<td>54%</td>
</tr>
<tr>
<td>English</td>
<td>82%</td>
<td>76%</td>
<td>63%</td>
<td>43%</td>
<td>66%</td>
<td>42%</td>
</tr>
<tr>
<td>Hebrew</td>
<td>95%</td>
<td>82%</td>
<td>72%</td>
<td>61%</td>
<td>68%</td>
<td>63%</td>
</tr>
<tr>
<td>Russian</td>
<td>97%</td>
<td>90%</td>
<td>73%</td>
<td>57%</td>
<td>60%</td>
<td>54%</td>
</tr>
</tbody>
</table>
# Number prediction accuracy: RNN vs human subjects in Italian

<table>
<thead>
<tr>
<th></th>
<th>RNN original</th>
<th>RNN nonce</th>
<th>subjects original</th>
<th>subjects nonce</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN [relC/partP] VERB</td>
<td>89%</td>
<td>95%</td>
<td>97%</td>
<td>92%</td>
</tr>
<tr>
<td>NOUN [relC/partP] clitic VERB</td>
<td>100%</td>
<td>96%</td>
<td>93%</td>
<td>95%</td>
</tr>
<tr>
<td>NOUN [adjP] relpron VERB</td>
<td>100%</td>
<td>88%</td>
<td>96%</td>
<td>89%</td>
</tr>
<tr>
<td>NOUN [pp] VERB (participial)</td>
<td>78%</td>
<td>73%</td>
<td>87%</td>
<td>73%</td>
</tr>
<tr>
<td>NOUN [pp] adverb ADJ</td>
<td>100%</td>
<td>79%</td>
<td>92%</td>
<td>79%</td>
</tr>
<tr>
<td>DET [adjP] NOUN</td>
<td>100%</td>
<td>92%</td>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>ADJ [conjoined adjectives] ADJ</td>
<td>100%</td>
<td>99%</td>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>VERB [verb complements] conj VERB</td>
<td>94%</td>
<td>76%</td>
<td>94%</td>
<td>87%</td>
</tr>
<tr>
<td><strong>average</strong></td>
<td><strong>95%</strong></td>
<td><strong>87%</strong></td>
<td><strong>95%</strong></td>
<td><strong>89%</strong></td>
</tr>
</tbody>
</table>
The problem with NOUN [pp] VERB participial

• Verb participle could in principle be part of prepositional phrase

• Original (RNN gets it right):
  • *l’ uomo [di 67 anni] massacrato*
  • the man of 67 years butchered

• Nonce (RNN makes wrong prediction):
  • *l’ apprendista [di 5.000 ingegneri] celebrato*
  • the apprentice of 5,000 engineers celebrated-sg
Dealing with attractors

Le colline che si intendono sul negozio del raffreddore devono

The hills that cl agree on-the store of-the cold must-pl

![Bar chart showing accuracy comparison between original and generated texts for subjects and LSTM models.](chart.png)
Better language models, better syntactic skills
Colorless green RNNs: summary

• Although they are not equipped with prior knowledge about syntax, RNNs are not simply good rote learners of shallow patterns
• Just by processing large amounts of raw text (language modeling), they acquire advanced structure-dependent syntactic competence
  • ... distinct from semantic and lexical competence
Systematic compositionality with modern RNNs

In collaboration with Brenden Lake

Lots of earlier work on neural networks and systematicity, main novelty here is that we test latest-generation, state-of-the-art architectures!

• https://github.com/brendenlake/SCAN/
• https://arxiv.org/abs/1711.00350
Systematic compositionality
Fodor and Pylyshyn 1988, Marcus 2003, 2018...

• Walk
• Walk twice
• Walk three times
• Run
• Run twice
• Run three times
• Dax
Systematic compositionality
Fodor and Pylyshyn 1988, Marcus 2003, 2018...

• Walk
• Walk twice
• Walk three times
• Run
• Run twice
• Run three times
• Dax
• Dax twice
• Dax three times
Systematic compositionality
Fodor and Pylyshyn 1988, Marcus 2003, 2018...

- Walk
- Walk twice
- Walk three times
- Run
- Run twice
- Run three times
- Dax
- Dax twice
- Dax three times

\[[X \text{ twice}]\] = \[[X][X]\]
\[[X \text{ three times}]\] = \[[X][X][X][X]\]
\[[\text{dax}]\] = perform daxing action
Systematic compositionality
Fodor and Pylyshyn 1988, Marcus 2003, 2018...

• Walk
• Walk twice
• Walk three times
• Run
• Run twice
• Run three times
• Dax
• Dax twice three times

\[
[[X \text{ twice}]] = [[X]][[X]]
\]
\[
[[X \text{ three times}]] = [[X]][[X]][[X]]
\]
\[
[[\text{dax}]] = \text{perform daxing action}
\]
SCAN: systematic compositionality in a simple grounded environment

walk and turn left!
Testing generalization

TRAINING PHASE

walk
WALK

walk and turn left
WALK LTURN

run thrice
RUN RUN RUN

look right and walk left
RTURN LOOK LTURN WALK

jump after walk
WALK JUMP

run around
RUN RUN RUN RUN RUN

walk and run
RUN WALK

TEST TIME

jump around and turn left
The SCAN commands: examples

• Primitive commands:
  • run -> RUN
  • walk -> WALK
  • turn left -> LTURN

• Modifiers:
  • walk left -> LTURN WALK
  • run twice -> RUN RUN

• Conjunctions:
  • walk left and run twice -> LTURN WALK RUN RUN
  • run twice after walk left -> RUN RUN LTURN WALK

• Simplifications:
  • No scope ambiguity ("walk and [run twice]"")
  • No recursion ("walk and run" vs "walk and run and walk")
Sequence-to-sequence RNNs for SCAN

- jump
- twice
- and
- walk
- <EOS>
- <SOS>
- JUMP
- JUMP
- WALK
- <EOS>
General methodology

• Train sequence-to-sequence RNN on 100k commands and corresponding action sequences
• At test time, only new composed commands presented
• Each test command presented once
• RNN must generate right action sequence at first try

• Training details: ADAM optimization with 0.001 learning rate and 50% teacher forcing
• Best model overall:
  • 2-layer LSTM with 200 hidden units per layer, no attention, 0.5 dropout
Experiment 1: random train/test split

• Included in training tasks:
  • look around left twice
  • look around left twice and turn left
  • jump right twice
  • run twice and jump right twice

• Presented during testing:
  • look around left twice and jump right twice
Random train/test split results

Accuracy on new commands (%) vs Percent of commands used for training.
Experiment 2: split by action length

- Train on commands requiring shorter action sequences (up to 22 actions)
  - jump around left twice (16 actions)
  - walk opposite right thrice (9 actions)
  - jump around left twice and walk opposite right twice (22 actions)

- Test on commands requiring longer actions sequences (from 24 to 48 actions)
  - jump around left twice and walk opposite right thrice (25 actions)

A grammar must reflect and explain the ability of a speaker to produce and understand new sentences which may be longer than any he has previously heard (Chomsky 1956)
Length split results

Accuracy on new commands (%) vs. Ground-truth action sequence length.
Experiment 3: generalizing composition of a primitive command (the "dax" experiment)

- Training set contains all possible commands with "run", "walk", look", "turn left", "turn right":
  - "run", "run twice", "turn left and run opposite thrice", "walk after run", ...

- but only a small set of composed "jump" commands:
  - "jump", "jump left", "run and jump", "jump around twice"

- System tested on all remaining "jump" commands:
  - jump twice
  - jump left and run opposite thrice
  - walk after jump
  - ...

38
Composed-"jump" split results

Accuracy on new commands (%) vs Number of composed commands used for training.
Proof-of-concept replication in Machine Translation

• Training: 100k sentences including:
  • I am daxy -> je suis daxiste
  • ... and many more simple sentences illustrating the paradigm below with other adjectives

• Test set includes:
  • you are daxy -> tu es daxiste
  • he is daxy -> il est daxiste
  • I am not daxy -> je ne suis pas daxiste
  • you are not daxy -> tu n'es pas daxiste
  • he is not daxy -> il n'est pas daxiste
  • I am very daxy -> je suis très daxiste
  • you are very daxy -> tu es très daxiste
  • he is very daxy -> il est très daxiste
Proof-of-concept replication in Machine Translation

• Out best RNN model gets only 1/8 daxy translation right ("he is daxy")

• For comparison:
  • "tired" occurred in 80 separate constructions in training
  • Model correctly translated equivalent "tired" sentences with 8/8 accuracy
A caveat: SCAN from a RNN's point of view...

orp mamp!

plek!

plek mamp!

??????
Conclusion

• RNNs impress and disappoint in surprising ways:
  • L'opaca bomba sottoposta dall'alloggio 200 ... pensa
  • "run", "run twice", "walk", "walk twice", "jump"... ???

• What's the difference?
  • Simplest hypothesis: choosing between two forms is an easier task than generating output from scratch
  • Can we detect traces of systematic compositionality in the SCAN RNN?
thank you

grazie mille