Colorless Green Recurrent Networks Dream Hierarchically

Junwei Yang

jy406@cam.ac.uk

January 25, 2019

Outline

- Evaluation of long-distance number agreement predicted by RNNs
  - Definition of problem, previous evaluation schemes
- Details of the approach
  - Four languages, multiple construction types for each language
  - Unsupervised setup – not enforced to learn long-distance agreement

Related Work

Some researches focused either on the morphological and grammatical knowledge, or based on controlled artificial languages.

- Recognise data generated by context-free grammars
- Translate between languages

The closest work to the current paper evaluated the performance of RNNs on the following problem:

- How well can RNNs approximate hierarchical structure?

Related Work

- Tested on predicting English subject-verb agreement
  - e.g., the girl the boys like is/are
- Proved RNNs can handle such constructions

However improvements can be made ...

- Explicit supervision on the target task required to capture agreement
- RNNs may make correct prediction based on semantic/frequency-based information
  - e.g., dogs in the neighbourhood often bark

Methodology

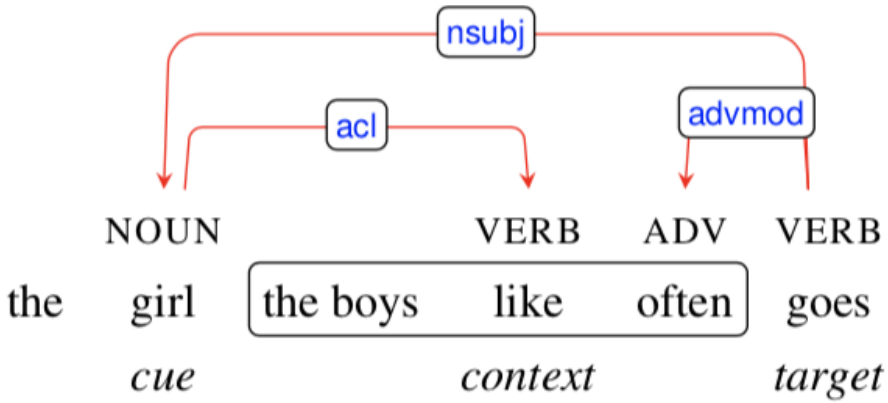
Extensions made to eliminate the potential effect caused by lexical, semantic and frequency-based confounds, and more factors are considered by including the following adjustments:

- Test the agreement with nonce sentences
- Evaluate on more languages with richer morphological systems
- Develop more long-distance number agreement constructions
- Include results from native speakers for comparison
- RNNs only trained to perform generic language modelling tasks

Long-Distance Agreement Constructions

- Agreement relations, e.g., subject-verb agreement
- Any number of elements can be placed between *cue* and *target*
  - The girl thinks
  - The girl you met thinks
  - The girl you met yesterday through her friends thinks
- Only focus on number agreement (singular or plural)

Identification of Constructions



- Collecting pairs of POS tags connected by a dependency arc
- Many possible agreement construction types, e.g., subject-verb, adjective-noun
- Construction consists of the agreement pair and the context

Constructions where only at least three tokens intervened between the cue and the target are considered.

Excluding Non-Agreement Constructions

- Not all cues agree with targets in terms of number for all instances of the construction
- e.g., verb-object construction in English
- Only keep agreement constructions with at least 10 instances of both plural and singular agreement
- 2 constructions remained for English and 21 left for Russian

Sentence Test Set

- Test sets extracted from Universal Dependency treebanks for English, Italian, Hebrew and Russian
- Extraction and tokenisation tools used to process the test sets
- Each original sentence includes all words from the *cue* to the *target*
- Target and its counterpart occurred in the language modelling (LM) vocabulary and the treebank
- Nonce sentences are generated based on original ones

Generating Nonce Sentences

Each content word (noun, verb, adjective, ...) in the sentence is replaced by another random word with the same POS tag from the treebank.

To avoid ambiguity between different POS tags, words that appeared with different POS tags more than 10% of the time in the treebank are removed. (e.g., *target* can either be verb or noun)

Experiments

- Different RNN language models implemented to test the performance
- Model trained on 90 million tokens from Wikipedia
- Three baseline adopted for comparison
- Native speakers of Italian involved to produce results for comparison

Language Models

RNNs:

- The PyTorch implementation of the original RNNs and LSTMs (code)
- Hyperparameters tuned (batch size, learning rate, dropout rate, ...)
- LSTMs outperformed simple RNNs

Baselines:

- Unigram model
- 5-gram model with Kneser-Ney smoothing
- 5-gram LSTM

LSTM Model

Grid search used to search the optimal 2-layer LSTM model trained for 40 epochs

- hidden and embedding size: 200 and 650
- batch size: 20 (only for models with 200 units), 64, and 128
- dropout rate: 0, 0.1, 0.2, 0.4 (only for models with 650 units)
- learning rate: 1, 5, 10, 20

Language	**Hidden size**	**Batch size**	**Dropout rate**	**Learning rate**
Italian	650	64	0.2	20
English	650	128	0.2	20
Hebrew	650	64	0.1	20
Russian	650	64	0.2	20

Results

	IT	EN	HE	RU
#constructions	8	2	18	21
#original	119	41	373	442
Unigram				
Original	54.6	65.9	67.8	60.2
Nonce	54.1	42.5	63.1	54.0
5-gram KN				
Original	63.9	63.4	72.1	73.5
Nonce	52.8	43.4	61.7	56.8
Perplexity	147.8	168.9	122.0	166.6
5-gram LSTM				
Original	81.8 ±3.2	70.2 ±5.8	90.9 ±1.2	91.5 ±0.4
Nonce	78.0 ±1.3	58.2 ±2.1	77.5 ±0.8	85.7 ±0.7
Perplexity	62.6 ±0.2	71.6 ±0.3	59.9 ±0.2	61.1 ±0.4
LSTM				
Original	92.1 ±1.6	81.0 ±2.0	94.7 ±0.4	96.1 ±0.7
Nonce	85.5 ±0.7	74.1 ±1.6	80.8 ±0.8	88.8 ±0.9
Perplexity	45.2 ±0.3	52.1 ±0.3	42.5 ±0.2	48.9 ±0.6

- 5-gram failed to capture number in nonce sentences
- Large improvement on 5-gram LSTM
- LSTM with unlimited history is much better
- English is the hardest language to predict

Difference Between Languages

Two constructions for English extracted:

		N V V	V NP conj V
Italian	Original	$93.3_{\pm 4.1}$	$83.3_{\pm 10.4}$
	Nonce	$92.5_{\pm 2.1}$	$78.5_{\pm 1.7}$
English	Original	$89.6_{\pm 3.6}$	$67.5_{\pm 5.2}$
	Nonce	$68.7_{\pm 0.9}$	$82.5_{\pm 4.8}$
Hebrew	Original	$86.7_{\pm 9.3}$	$83.3_{\pm 5.9}$
	Nonce	$65.7_{\pm 4.1}$	$83.1_{\pm 2.8}$
Russian	Original	-	$95.2_{\pm 1.9}$
	Nonce	-	$86.7_{\pm 1.6}$

- Bad performance of English due to poor morphology and high ambiguity
- Richer morphology and less ambiguity at the POS level leads to better performance and smaller gap (Italian and Russian)
- Russian with the highest accuracy is less prone to human attraction errors
- Largest drop occurred in Hebrew due to incorrect number caused by multiple readings for certain constructions

Human Results

Performance of human and LSTM on the Italian test set:

Construction	#original	Original		Nonce	
		Subjects	LSTM	Subjects	LSTM
DET [AdjP] NOUN	14	98.7	$98.6_{\pm 3.2}$	98.1	$91.7_{\pm 0.4}$
NOUN [RelC / PartP] clitic VERB	6	93.1	$100_{\pm 0.0}$	95.4	$97.8_{\pm 0.8}$
NOUN [RelC / PartP] VERB	27	97.0	$93.3_{\pm 4.1}$	92.3	$92.5_{\pm 2.1}$
ADJ [conjoined ADJs] ADJ	13	98.5	$100_{\pm 0.0}$	98.0	$98.1_{\pm 1.1}$
NOUN [AdjP] relpron VERB	10	95.9	$98.0_{\pm 4.5}$	89.5	$84.0_{\pm 3.3}$
NOUN [PP] ADVERB ADJ	13	91.5	$98.5_{\pm 3.4}$	79.4	$76.9_{\pm 1.4}$
NOUN [PP] VERB (participial)	18	87.1	$77.8_{\pm 3.9}$	73.4	$71.1_{\pm 3.3}$
VERB [NP] CONJ VERB	18	94.0	$83.3_{\pm 10.4}$	86.8	$78.5_{\pm 1.7}$
(Micro) average		94.5	$92.1_{\pm 1.6}$	88.4	$85.5_{\pm 0.7}$

Human Results

Tested whether human and models tend to make similar mistakes, for each sentence the following properties are computed:

- # of times human correct − # of times human incorrect
- Difference in model log probability between correct and incorrect form

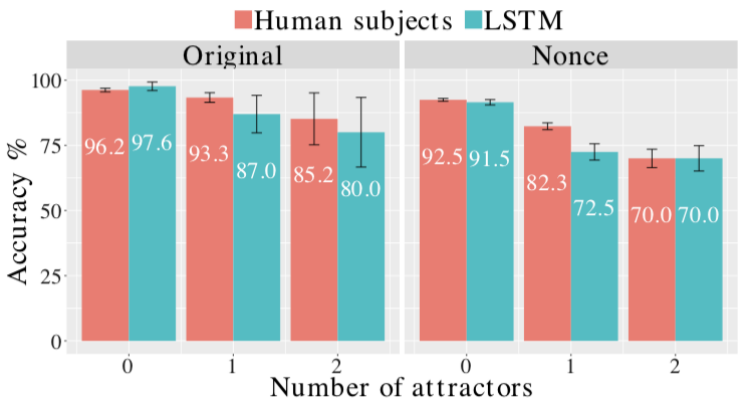
The correlation turned out to be significant, indicating that humans are more likely to choose the correct form that models are more confident about.

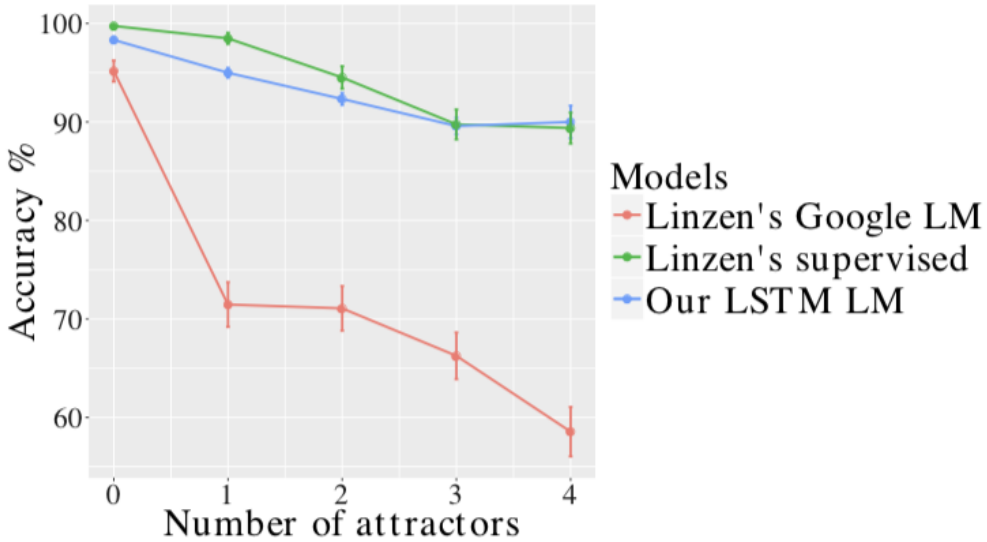
Attractors

Definition

Words with the same POS tag as the cue but the opposite number in the context.

e.g., The girl you met yesterday through her **friends** thinks





Future Work

- Investigate what and how RNNs encode syntactic information
- Extract examples from other long-distance phenomena for analysis
- Extend the current approach to isolate unwanted syntactic phenomena

References

Gulordava, Kristina, et al. (2018)

Colorless green recurrent networks dream hierarchically

arXiv preprint arXiv:1803.11138.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016)

Assessing the ability of LSTMs to learn syntax-sensitive dependencies

arXiv preprint arXiv:1611.01368.

The End