Targeted Syntactic Evaluation of Language Models

Rebecca Marvin and Tal Linzen

EMNLP (2018)

Summary

- Tested the grammaticality of paired sentences, as given by a pre-trained LM

- Their conclusions are that LSTMs:

  → Have difficulties with non-local constructs

  → Struggle with object relative clauses

  → Huge performance gap compared to humans

Similarities to previous paper

- Same corpus and hyper-parameters to train LMs
- Baselines:
  - → 5-gram LM with Kneser-Ney smoothing
  - → LSTM LM (single-task)
- Both compare to human judgements

Differences from previous paper

	Gulordava et al. (previous paper)	Marvin and Linzen (this paper)
Sentence pair	original and nonce	grammatical and ungrammatical
Syntactic structures	just number agreement	number agreement + NPI
Prediction task	target word	whole sentence
Conclusion	LSTMs succeed	LSTMs fail

Corpus

- Three challenging syntactic structures (manually constructed with templates):
  - → subject-verb agreement
  - → reflexive anaphora
  - → negative polarity items (NPI)
- This allows for greater coverage and control than in the naturally occurring setting.

Multi-task LSTM LM

- Combine two objective functions:
  - → One for the usual LM
  - → One for classifying the CCG supertag
- Sum them, with equal weight
- Makes the model more syntax-aware

Combinatory categorial grammar (CCG)

Until	(S/S)/(S[adj]\NP)
recently	S[adj]\NP
,	,
national	N/N
governments	N
in	(NP\NP)/NP
Europe	N
controlled	(S[dcl]\NP)/NP
most	N
of	(NP\NP)/NP
the	NP[nb]/N
air	N/N
time	N

Results

- *n*-gram baseline performs close to random
- Simple RNN struggles on complex examples
- Multi-task RNN is still weaker than humans
- Particularly hard:
  - → Relative clauses without “that”
  - → Reflexive anaphora in across a relative clause
  - → Predicting “herself” (gender bias, corpus-based)
  - → Greater overall probability for ungramm. NPIs

Discussion

- In these settings, did LSTMs learn syntax?

Discussion

- In these settings, did LSTMs learn syntax?
- Which stronger architectures would do better?

Assessing BERT's Syntactic Abilities
(Yoav Goldberg, 2019)

	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no *that*)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no *that*)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

Table 3: Results on the Marvin and Linzen (2018) stimuli. M&L results numbers are taken from Marvin and Linzen (2018). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

Discussion

- In these settings, did LSTMs learn syntax?
- Which stronger architectures would do better?
- What can linguistics and deep learning contribute to each other?

Discussion

- In these settings, did LSTMs learn syntax?
- Which stronger architectures would do better?
- What can linguistics and deep learning contribute to each other?