Proceedings of the Society for Computation in Linguistics

Volume 2 Article 53

2019

Targeted Syntactic Evaluation of Language Models

Rebecca Marvin
Johns Hopkins University, becky@jhu.edu

Tal Linzen
Johns Hopkins University, tal.linzen@jhu.edu

Recommended Citation

Marvin, Rebecca and Linzen, Tal (2019) "Targeted Syntactic Evaluation of Language Models," *Proceedings of the Society for Computation in Linguistics*: Vol. 2 , Article 53.
DOI: https://doi.org/10.7275/p0cq-hv95
Available at: https://scholarworks.umass.edu/scil/vol2/iss1/53

Targeted Syntactic Evaluation of Language Models

Rebecca Marvin
Johns Hopkins University
becky@jhu.edu

Tal Linzen
Johns Hopkins University
tal.linzen@jhu.edu

Language models (LMs) define probability distributions over sequences of words. Most LMs are evaluated using perplexity, a measure related to the probability assigned by the model to a word in the corpus. This measure conflates multiple sources of success (or failure) in predicting the next word; in particular, many words can be predicted based on collocational and semantic factors alone, without a robust representation of the syntactic structure of the sentence.

We argue that more informative syntactic evaluation metrics could accelerate progress towards grammatically sophisticated LMs. Indeed, avoiding ungrammatical predictions may be as important as accurately capturing word collocations, which simple n-gram LMs already excel at. To this end, we propose a metric that assesses whether the probability distribution learned by the LM conforms to the grammar of the language. Concretely, given two sentences that differ minimally from each other, one of which is grammatical and the other is not, it is desirable for the model to assign a higher probability to the grammatical one (Lau et al., 2017; Linzen et al., 2016). We propose to evaluate the LM on sentence pairs that exemplify complex syntactic phenomena; this evaluation strategy provides a fine-grained and interpretable breakdown of the strengths and weaknesses of an LM.

We automatically generated a large number of sentence pairs (∼350,000) using templates. Our data set included three phenomena considered to be sensitive to hierarchical syntactic structure (Everaert et al., 2015; Xiang et al., 2009) — subject-verb agreement, reflexive anaphora and negative polarity items — in the following conditions:

(1) **Simple agreement:**
The farmer smiles/*smile.

(2) **Agreement in a sentential complement:**
The mechanics said the author laughs/*laugh.

(3) **Agreement in short VP coordination:**
The authors laugh and swim/*swims.

(4) **Agreement in long VP coordination:**
The author knows many different foreign languages and enjoys/*enjoy playing tennis with colleagues.

(5) **Agreement across a prepositional phrase:**
The author next to the guards smiles/*smile.

(6) **Agreement across a subject relative clause:**
The author that likes the security guards laughs/*laugh.

(7) **Agreement across an object relative:**
The movies that the guard likes are/*is good.

(8) **Agreement in an object relative:**
The movies that the guard likes/*like are good.

(9) **Simple reflexive anaphora:**
The author injured himself/*themselves.

(10) **Reflexive in sentential complement:**
The mechanics said the author hurt himself/*themselves.

(11) **Reflexive across a relative clause:**
The author that the guards like injured himself/*themselves.

(12) **Simple NPI:**
No/*most authors have ever been famous.

(13) **NPI across a relative clause:**
a. No authors that the guards like have ever been famous.
b. *The authors that no guards like have ever been famous.

All combinations of subject number and local noun number were included in the data set; e.g., for agreement across a prepositional phrase:

(14) a. the farmer near the parent smiles/*smile
b. the farmer near the parents smiles/*smile
c. the farmers near the parent smile/*smiles
d. the farmers near the parents smile/*smiles

We used our challenge to test three LMs: an *n*-gram baseline, a recurrent neural network (RNN)

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 373-374.
New York City, New York, January 3-6, 2019

	RNN	Multitask	*n*-gram	Humans
SUBJECT-VERB AGREEMENT:				
Simple	0.94	1.00	0.79	0.96
In a sentential complement	0.99	0.93	0.79	0.93
Short VP coordination	0.90	0.90	0.51	0.94
Long VP coordination	0.61	0.81	0.50	0.82
Across a prepositional phrase	0.57	0.69	0.50	0.85
Across a subject relative clause	0.56	0.74	0.50	0.88
Across an object relative clause	0.50	0.57	0.50	0.85
Across an object relative (no *that*)	0.52	0.52	0.50	0.82
In an object relative clause	0.84	0.89	0.50	0.78
In an object relative (no *that*)	0.71	0.81	0.50	0.79
REFLEXIVE ANAPHORA:				
Simple	0.83	0.86	0.50	0.96
In a sentential complement	0.86	0.83	0.50	0.91
Across a relative clause	0.55	0.56	0.50	0.87
NEGATIVE POLARITY ITEMS:				
Simple	0.40	0.48	0.06	0.98
Across a relative clause	0.41	0.73	0.60	0.81

Table 1: Overall accuracies for the LSTMs, *n*-gram model and humans on each test case.

LM trained on a 90M word subset of the English Wikipedia, and an RNN LM trained on a multitask objective: language modeling (on the same subset of English Wikipedia) and Combinatory Categorial Grammar (CCG) supertagging (Bangalore and Joshi, 1999), which requires rich syntactic annotations (based on the Penn Treebank).

We also designed a human experiment on Amazon Mechanical Turk that mirrored the task given to the LMs: both versions of a minimal pair were shown on the screen at the same time, and participants were asked to judge which one of them was more acceptable. There is a rich literature showing that humans make mistakes such as subject-verb agreement errors (Bock and Miller, 1991; Phillips et al., 2011); while we would ultimately like to have LMs that do not make any errors (unlike humans), matching human performance would be an impressive first step.

Results of the LMs and humans on our dataset are shown in Table 1. The *n*-gram baseline largely performed at chance, suggesting that good performance on the task requires syntactic representations. The RNN LMs performed well on simple cases but struggled on more complex ones. Multitask training with a supervised syntactic objective improved the performance of the RNN on the challenge set; nevertheless, this model was still much weaker than humans, especially in subject-verb agreement across relative clauses. This suggests that our data set is challenging and can motivate richer language modeling architectures.

References

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Martin B. H. Everaert, Marinus A. C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12):729–743.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, (5):1202–1247.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Colin Phillips, Matthew W. Wagers, and Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey T. Runner, editor, *Experiments at the Interfaces, Syntax and Semantics 37*, pages 153–186.

Ming Xiang, Brian Dillon, and Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108(1):40–55.