A Systematic Assessment of Syntactic Generalization in Neural Language Models

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy
(MIT, Harvard University)

2020年9月25-26日 最先端NLP勉強会

読んだ人：篠田 一聡 (東大 相澤研 D2)

読んだ論文

A Systematic Assessment of Syntactic Generalization in Neural Language Models

Jennifer Hu[1], **Jon Gauthier**[1], **Peng Qian**[1], **Ethan Wilcox**[2], and **Roger P. Levy**[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Linguistics, Harvard University

- ACL2020 (long)
- 選んだ理由：
  - 投票数が多いのに選ばれていなかったから
  - 結果の図が面白そうだったから

何をやったか

- 背景
  - ニューラルネットが言語モデルタスクでPerplexityを下げてきたが、それが人間のようなsyntactic knowledgeの獲得につながっているかは未知だった
- やったこと
  - ニューラル言語モデルが持つ文法的な知識についての体系的な評価を行った

何がわかったか

- 言語モデルのPerplexityとsyntactic generalizationに相関はある？
  - ~~ない~~
  - 特定のモデル同士ではみられなかった
- 訓練データのサイズやモデルのアーキテクチャがsyntactic generalizationに与える影響は？
  - アーキテクチャの方が影響が大きい
- 文法に関する評価を体系的に行うとどんな結果が得られるのか？
  - 文法知識の種類によって、良いモデルの構造が異なることがわかった

(Sec 2.1) Perplexityとは

- 定義
  - テストコーパスCに含まれるn単語の同時確率の(-1/N)乗

$$\mathrm{PPL}(C) = p(w_1, w_2, \ldots w_N)^{-\frac{1}{N}}$$

- 確率的言語モデルの評価指標として広く使われている

(Sec 2.1) 単語の確率と読み時間

- 主に認知科学分野において、**読み時間と単語が出現する確率には関係がある**ことが知られていた。(Levy, 2008; Smith & Levy, 2013; Bicknell & Levy, 2010)
- “Predictive power of word surprisal for reading times is a linear function of language model quality” (Goodkind and Bicknell, 2018)
  - **Perplexityが良い言語モデル**ほど、（視線の動きから計測した）**人の読み時間を精度よく予測できる**

縦軸：読み時間の予測精度のよさ

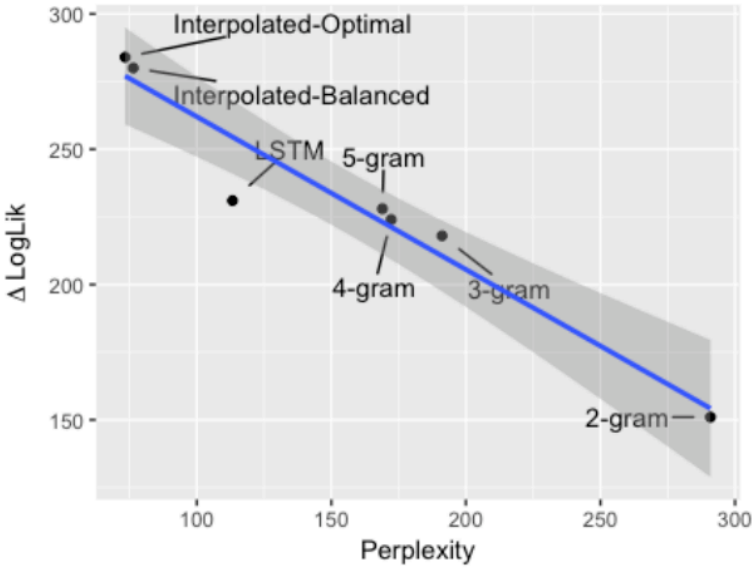

Figure 1: Improvements in log likelihood for linear models, charted against decreases in perplexity. Distance from the central trend line is indicative of larger departures in log likelihood as a function of perplexity. The blue line represents a linear best fit, with a coefficient of -1.66 and $R^2 = 0.94$

(Sec 2.1) Perplexityについて著者が思うこと

- Perplexityはbroad-coverageで、言語モデルのsyntactic knowledgeを評価することはできないのではないか？

- 例えば、確率はかなり低いが文法的に正しい文は存在する
  - *Colorless green ideas sleep furiously* (Chomsky, 1957)

(Sec 2.2) Targeted tests for syntactic generalization

- Perplexityの代わりに、特定のsyntactic phenomenaごとに言語モデルが人のように汎化できるかどうかで評価されてきた。（Linzen et al., 2016; Lau et al., 2017; Gulordava et al., 2018）
- **The targeted syntactic evaluation paradigm** (Marvin and Linzen, 2018; Futrell et al., 2019) … 元は心理言語学の手法
  - 例 (subject–verb number agreement)
    - The keys to the cabinet are on the table
    - The keys to the cabinet is on the table
    - 言語モデルはどちらの文により大きい確率を割り当てられるか？

(Sec 2.2) garden-pathing example

POSに曖昧性がある例

- The child kicked in the chaos found her …

左から順にkickedまで読んだ人
「多分kickedはmain verbだな」

(Sec 2.2) garden-pathing example

POSに曖昧性がある例

- The child kicked in the chaos found her …

左から順にfoundまで読んだ人
「やっぱりfoundがmain verbで、
kickedは過去分詞だった」

(Sec 2.2) garden-pathing example

POSに曖昧性がない例

- The child forgotten in the chaos found her …

左から順にforgottenまで読んだ人
「forgottenは過去分詞以外ありえない」

曖昧性があると読みづらいし、脳内の処理が大変
これを利用して・・・

(Sec 2.2) garden-pathing example

(A) The child kicked in the chaos found . . .
(B) The child forgotten in the chaos found . . .
(C) The child who was kicked in the chaos found . . .
(D) The child who was forgotten in the chaos found . . .

- この4つの文の「読みづらさ」を、言語モデルに判断させる
- どうやって？
  - 言語モデルがそれぞれの文脈から予測したfoundの確率を比較
  - foundの確率が高い/低いなら、モデルにとって読みやすい/にくい文

(Sec 2.2) garden-pathing example

(A) The child kicked in the chaos found . . .
(B) The child forgotten in the chaos found . . .
(C) The child who was kicked in the chaos found . . .
(D) The child who was forgotten in the chaos found . . .

- 「読みづらさ」の予測はどうあるべきか
  - 人間と同じようになっていて欲しい

(Sec 2.2) garden-pathing example

(A) The child kicked in the chaos found . . .
(B) The child forgotten in the chaos found . . .
(C) The child who was kicked in the chaos found . . .
(D) The child who was forgotten in the chaos found . . .

- 「読みづらさ」の予測はどうあるべきか
  - (A) よりも (B) の方が読みやすいはず
  - (A) よりも (C) の方が読みやすいはず
  - (A) と (B) の読みづらさの差は、(C) と (D) のそれよりも大きいはず

(Sec 3) 提案評価手法

- モデルのデザイン(アーキテクチャ, データのサイズ)と性能の指標(perplexity, syntactic generalization)の2x2の関係を体系的に調べる！

- 34個のテスト(test suites)を用意した
  - 既存研究からかき集めた＋オリジナルに2個設計した

- これらは16個のsyntactic phenomenaをカバーする

(Sec 3.1) Test suites

- 心理言語学の研究から知見を借りて、様々な種類のテストをする
- 各テストは、Sec 2.2のgarden-pathing exampleのように、モデルが予測した確率を比較してそれが要件を満たすかどうかを評価する

(Sec 3.1) Syntactic coverage

- 提案評価指標がカバーしているのは、Carnie (2012)の教科書で紹介されている47個のsyntactic phenomenaのうち、16個
- 16個でも割と広範囲をカバーできている（と主張している）

(Sec 3.1) Test suitesの分類 (6種類)

- Agreement
  - 例えばsubject-verb number agreementなど
- Licensing
  - 前置詞や否定語など、特定の単語がないといけないもの
- Garden-Path Effects
  - 文をLocalに見るとcoherentでもglobalに見ると曖昧性があるもの
- Gross Syntactic Expectation
  - 長いチャンク（名詞節など）に関わるもの
- Center Embedding
  - 再帰的に入れ子構造になっている文をstackのように処理できるか
- Long-Distance Dependencies
  - 長距離の依存関係に関するもの

(Sec 3.2) Model training data

- BLLIP (Charniak et al., 2000)
  - English newswire corpus

BLLIP sizes:	XS	SM	MD	LG
# sentences	40K	200K	600K	1.8M
# tokens	1M	4.8M	14M	42M
# non-UNK types	24K	57K	100K	170K
# UNK types	68	70	71	74

Table 1: Statistics of training set for each corpus size.

(Sec 3.3) Model classes (architechtures)

- LSTM (Hochreiter and Schmidhuber, 1997)
- ON-LSTM (Ordered Neuron; Shen et al., 2019)
  - 言語の階層構造をinductive biasとして盛り込んだモデル
- RNNG (RNN grammar; Dyer et al., 2016)
  - ON-LSTMと同じように階層構造を表現できるが、constituency parseのラベルが必要なので、off-the-shelf 構文解析器(Kitaev and Klein, 2018)を使用
- Transformer (Vaswani et al., 2017)
  - GPT-2をscratchから訓練
- n-gram
- off-the-shelf language models (pretrained GPT-2, etc.)

- ※左から順に処理して行った時に生じる曖昧性を利用したテストなので、モデルもそのようなものに限る

(Sec 4) Results

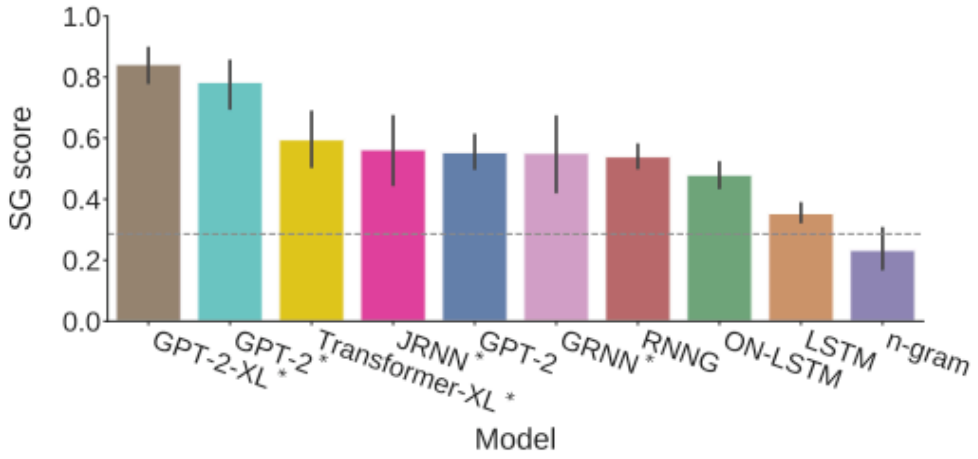

Figure 1: Average SG score by model class. Asterisks denote off-the-shelf models. Error bars denote bootstrapped 95% confidence intervals of the mean.

- Neural modelは全てrandom baselineよりはSG scoreが高い
- Neural model内での差が顕著。最大で2倍違う。

(Sec 4) Results

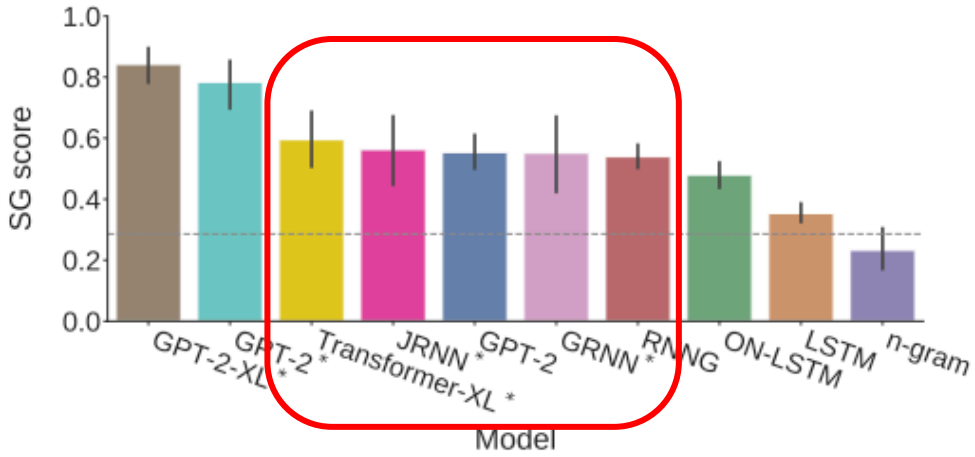

Figure 1: Average SG score by model class. Asterisks denote off-the-shelf models. Error bars denote bootstrapped 95% confidence intervals of the mean.

- さらに、赤枠の中のモデルは、訓練データのサイズが違うのに近いSGスコアを出している

左から順に訓練データのtokenの数は103M, 800M, 42M(?), 90M, 42M(?)

(Sec 4.1) Syntactic generalization (SG) and perplexity

SG scoreとperplexityの関係は？

- (GPT-2以外のニューラルモデルでは)**2つのスコアに相関はない**
- SG scoreの振れ幅を説明できるperplexity以外の要因があるはず
- n-gramのSG scoreが悪いのは想像通り
- Neural modelの中では、各訓練データのサイズでSG scoreを比べるとGPT-2が最も高いスコア(MD, LG)と低いスコア(XS, SM)を出している。最も高くなっているのはsub wordの影響もあり得る。
- 同じ色同士で比べると、モデルによってSG scoreが結構違う

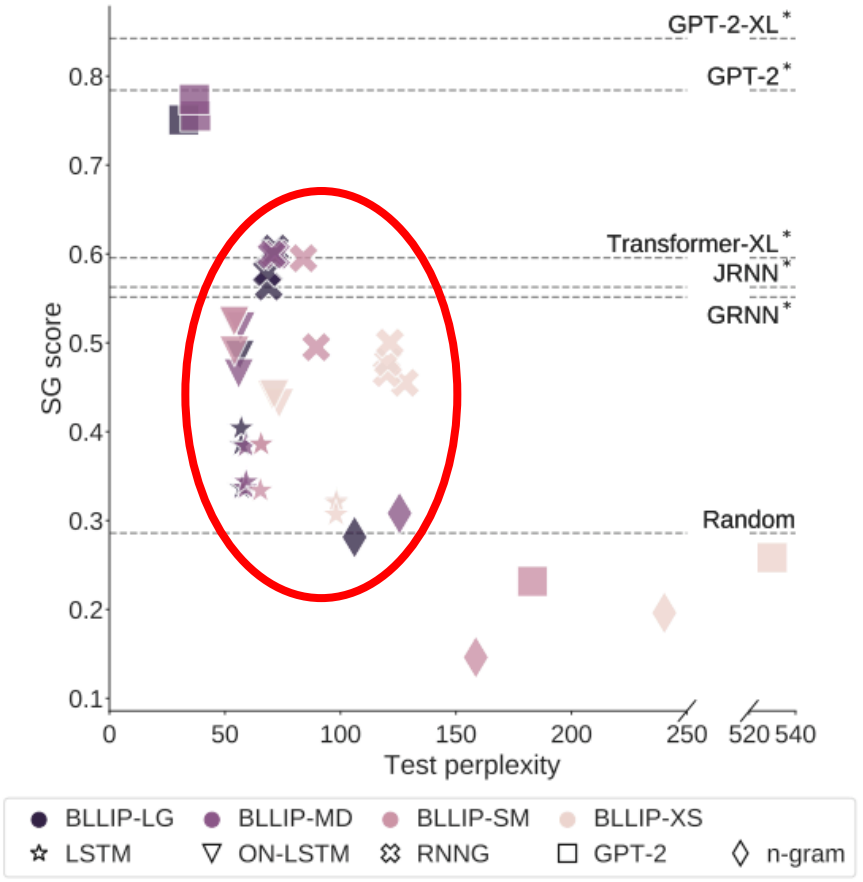

Figure 2: Relationship between SG score and perplexity on our held-out BLLIP test set for each model.

(Sec 4.2) Inductive bias and data scale

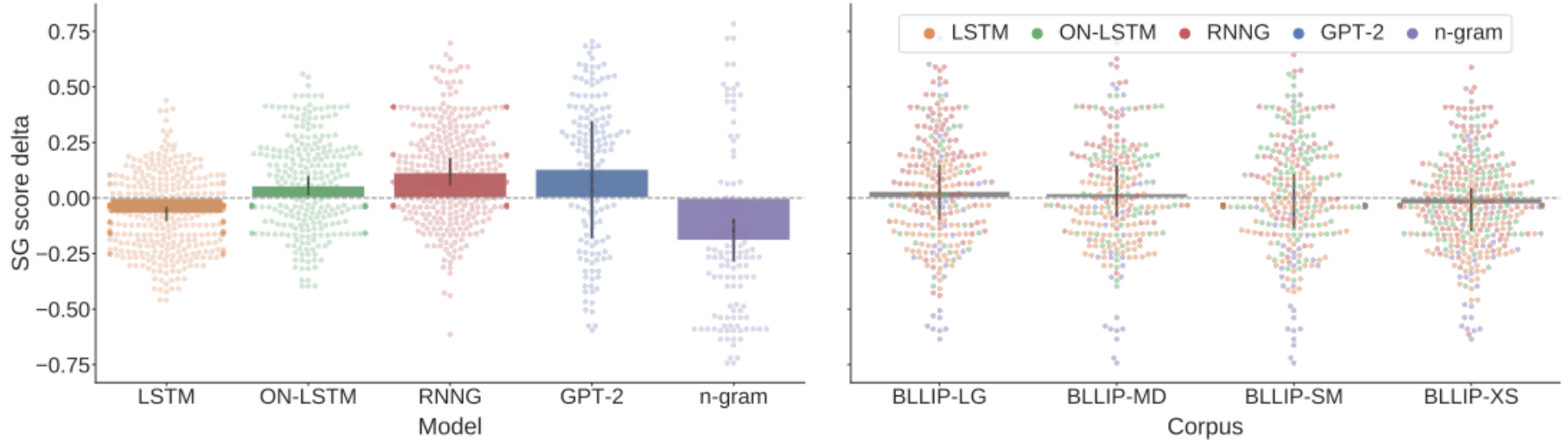

Figure 3: Main results of our controlled evaluation of model class and dataset size. SG score varies more by model class (left) than by training dataset size (right).

- Test suiteごとの難易度の影響を排除して、モデルのアーキテクチャとデータサイズの影響だけを見るために、**各テストで全モデルの平均点を出して、そこからの差を見る。**
- **SG scoreへの影響は、モデルのアーキテクチャ＞訓練データのサイズ**

(Sec 4.2) Inductive bias and data scale

補足

- GPT-2は公開されているものと自前で訓練したもので、訓練データの数は100:1くらいなのに、ほとんどSG scoreが変わらなかった。
  - やはりデータサイズのSG scoreへの影響はないとは言えないものの小さそう。
- linear mixed-effects regression modelを使ってSG score deltaの予測にどれくらい使えるかという観点から影響を調べたが、やはりアーキテクチャ(inductive bias)の方がデータサイズよりSG scoreへの影響が大きそう

(Sec 4.3) Circuit-level effects on SG score

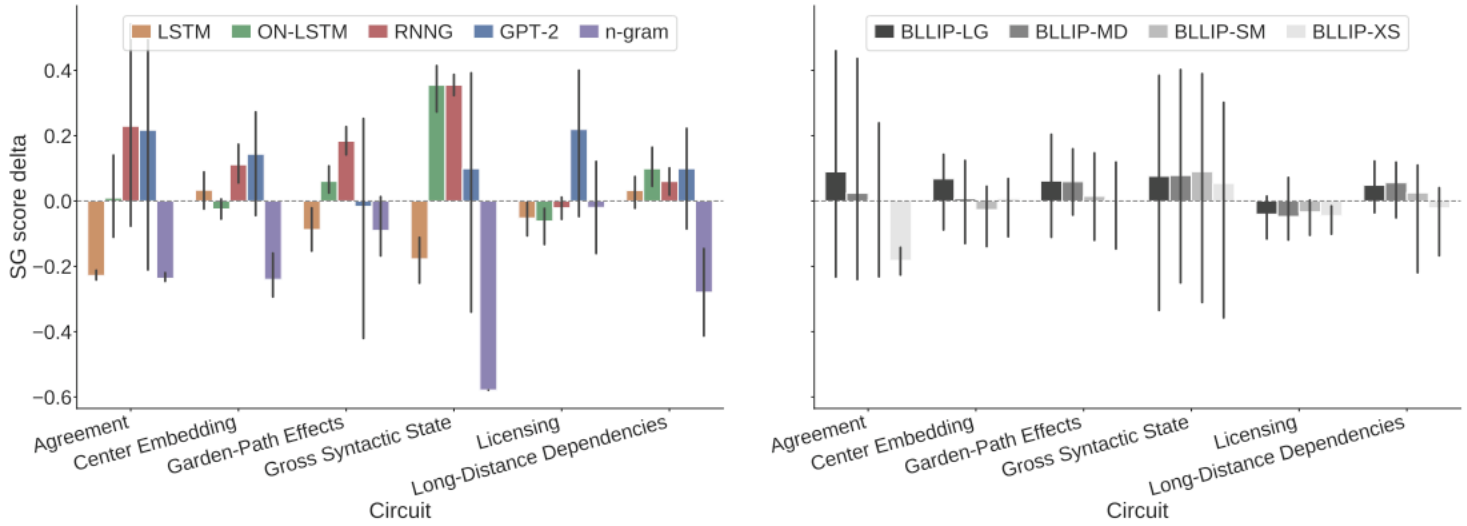

Figure 4: Controlled evaluation results, split across test suite circuits. Circuit-level differences in SG score vary more by model class (left) than by training dataset size (right).

- (right) データサイズはSG score deltaに(Agreement以外では)影響をあまり与えない
- (left) しかしアーキテクチャは影響が大きそう
- (left) Licensingについては言語モデル普遍の独立なsyntactic processが関わっているのでは
- (left) **circuitごとにどのモデルがより得意であるが異なる**

(Sec 4.3) Circuit-level effects on SG score

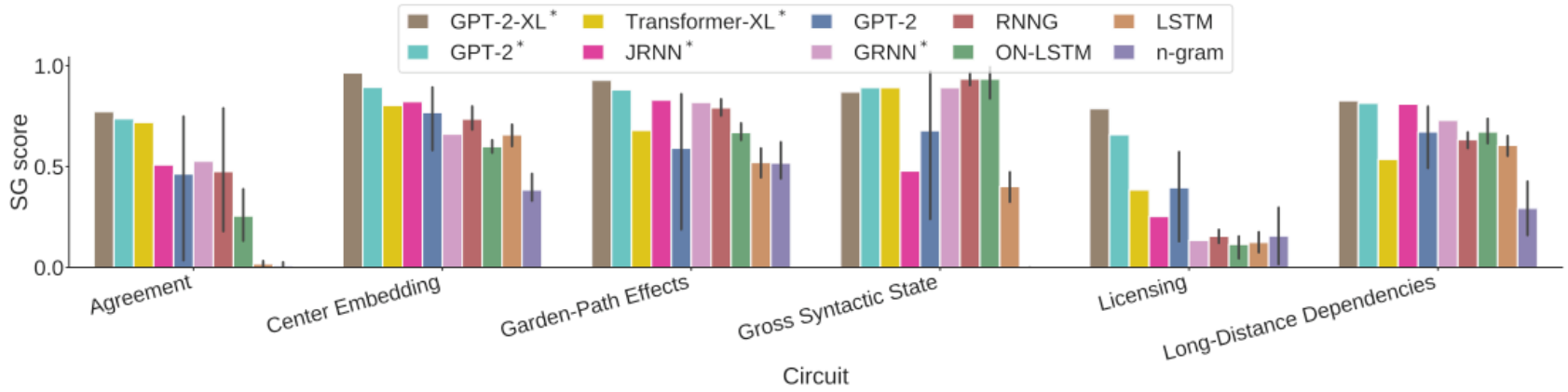

Figure 5: Evaluation results on all models, split across test suite circuits.

- RNNGがoff-the-shelfのモデルと比べて、訓練データ数が少ないのに同等のSG scoreを出せているのはすごい

(Sec 4.4) Stability to modifiers

- 修飾語の挿入に対して頑健かどうか
  - モデルによって明らかに違いがある
  - RNNGの方がON-LSTMよりも比較的頑健
  - GPT-2-XLは全く悪影響を受けない

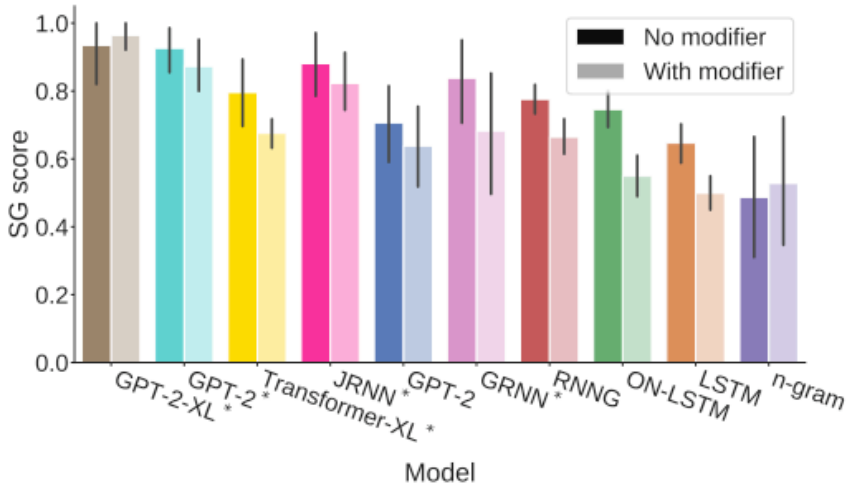

Figure 6: SG score on the pairs of test suites with and without intervening modifiers: Center Embedding, Cleft, MVRR, NPZ-Ambiguous, and NPZ-Object.

(Sec 4.5) Effects of model preprocessing

- 自前で試したモデルの中では、GPT-2だけBPEによるサブワードを使っている
- そのため、GPT-2の結果は、Transformerを使っていることとサブワードを使っていることの双方が影響していることを留意

- 実際、high resourceの時、サブワードを使わずに単語レベルでGPT-2を訓練した時、perplexityは良かったが、SG scoreは劇的に下がった

- Low resourceの時は逆に、語彙のセットを訓練データ内の頻度で足切りせずにサブワードの総数は固定。よってlow resourceでGPT-2を使うのはoverparameterizationなのでは。

(Sec 5) Discussion

- PerplexityとSG scoreに相関はなく、これらは互いにモデルの違う側面を評価しているのでは。
- 訓練データのサイズよりもモデルのアーキテクチャの方がSGに寄与している
- Licensingではモデルによらず一貫して失敗していたが、他のテストはモデルによって得意不得意が一貫していない
  - Syntactic circuitごとに違う処理をモデルに要求しているのかも
- 現時点では最も大規模なテストだが、まだ始まりにすぎず、もっとtest suiteのバリエーションを豊かにすると言語モデルのsyntactic capabilityのより詳しい理解につながる

まとめ

- いろいろな言語モデルのアーキテクチャ・訓練データのサイズを試してみた結果、syntactic generalization (SG) スコアとPerplexityには相関がないことがわかった。

- 訓練データのサイズよりも、モデルのinductive biasの方が SG スコアに与える影響が大きかった。

- アーキテクチャによって得意なsyntactic testに違いがある

感想

- ここで使われているtargeted syntactic evaluation (Marvin and Linzen, 2018)のようにテストケースを用意する評価手法は、モデルのfine tuningを必要とせず、いろいろなタスクに応用できると感じた
- 評価手法レベルでの新規性は既存の研究をスケールした、ということにとどまりそう
- とは言え、いろんなモデルとデータのサイズを体系的に比較して、新たにわかったこともあるという感じ
  - LSTMは大体他より悪い
- 個人的にはそれぞれのsyntactic knowledgeを獲得するのに必要なモデルの構造(inductive bias)に興味があったのだが、この研究ではこれについて得られる洞察は限定的。なぜなら…
  - GPT-2はサブワードの効果もありRNN vs self-attentionの公平な比較はできていなさそう
  - RNNGはマルチタスク学習をしているのでちょっとずるい。URNNGも使ってみて欲しい。RNNG, GPT-2も比較に入れてinductive biasの違いが大きいというのは少しover claimな気がする。
- 詳細な分析は今後の研究に期待