Hu et al., 2020
Sinha et al., 2019

Greta Tuckute & Kamoya K Ikhofua

MIT Fall 2020, 6.884
Symbolic Generalization
Motivation

Natural language understanding systems to generalize in a systematic and robust way

- Diagnostic tests - how can we probe these generalization abilities?
  - Syntactic generalization (Hu et al., 2020, “SG”) and logical reasoning (Sinha et al., 2019, “CLUTRR”)

- Evaluation metrics for language models?
Perplexity is not sufficient to check for human-like syntactic knowledge:

- It basically measures the probability of seeing some collection of words together
- However some words which are rarely seen together are grammatically correct
- *Colorless green ideas sleep furiously* (Chomsky, 1957)
- Need a more fine-grained way to assess learning outcomes of neural language models
SG: Paradigm

Assess NL models on custom sentences designed using psycholinguistic and syntax literature/methodology

- Compare critical sentence regions NOT full-sentence probabilities.
- Factor out confounds (e.g. token lexical frequency, n-gram statistics)
SG: Paradigm

- Cover the scope of syntax phenomena: 16/47 (Carnie et al., 2012)
- Group syntax phenomena into 6 circuits based on processing algorithm
SG: Circuits

1. Agreement
2. Licensing
3. Garden-Path Effects
4. Gross Syntactic Expectation
5. Center Embedding
6. Long-Distance Dependencies
SG: Agreement

(A) The farmer that the clerks embarrassed knows$_{V_{sg}}$ many people.

(B) *The farmer that the clerks embarrassed know$_{V_{pl}}$ many people.

(C) The farmers that the clerk embarrassed know$_{V_{pl}}$ many people.

(D) *The farmers that the clerk embarrassed knows$_{V_{sg}}$ many people.

\[ P_A(V_{sg}) > P_B(V_{pl}) \land P_C(V_{pl}) > P_D(V_{sg}) \]

Chance is 25% (or up to 50%)
SG: NPI Licensing

- The word “any” is a negative polarity item (NPI)
- The word “no” can license an NPI when it structurally commands it, such as in A

A) **No** managers that respected the guard have had **any** luck

B) *The managers {that respected **no** guard} have had **any** luck

(Reflexive Pronoun Licensing was also included in sub-class suites)
SG: NPI Licensing

(A) No managers that respected the guard have
NPI
had any luck. [+NEG, −DISTRACTER]

(B) *The managers that respected no guard have
NPI
had any luck. [−NEG, +DISTRACTER]

(C) *The managers that respected the guard have
NPI
had any luck. [−NEG, −DISTRACTER]

(D) No managers that respected no guard have
NPI
had any luck. [+NEG, +DISTRACTER]

Acceptable orderings:

- ADBC
- ADCB
- DABC
- DACB
- ACDB (?)

Chance: 5/24
SG: Reflexive Pronoun Licensing

(A) The author that the senators liked hurt herself$_{R_{sg\text{-fem}}}$.

(B) *The authors that the senator liked hurt herself$_{R_{sg\text{-fem}}}$.

(C) The authors that the senator liked hurt themselves$_{R_{pl}}$.

(D) *The author that the senator liked hurt themselves$_{R_{pl}}$.

\[ P_A(R_{sg}) > P_B(R_{sg}) \land P_C(R_{pl}) > P_D(R_{pl}) \]  

Chance: 25%
SG: NP/Z Garden-Paths

(A)  As the ship crossed the waters remained blue and calm. [TRANS,NO COMMA]

(B)  As the ship crossed, the waters remained blue and calm. [TRANS,COMMA]

(C)  As the ship drifted the waters remained blue and calm. [INTRANS,NO COMMA]

(D)  As the ship drifted, the waters remained blue and calm. [INTRANS,COMMA]
SG: Main-Verb Reduced Relative Garden-Paths

(A) The child kicked in the chaos found her way back home. [REDUCED, AMBIG]

(B) The child who was kicked in the chaos found her way back home.

(C) The child forgotten in the chaos found her way back home.

(D) The child who was forgotten in the chaos found her way back home.

Chance is 25%
SG: Gross Syntactic Expectation (Subordination)

(A) The minister praised the building.

(B) *After the minister praised the building.

(C) ??The minister praised the building, it started to rain.

(D) After the minister praised the building, it started to rain.

\[ P_A(\text{END}) > P_B(\text{END}) \land P_D(\text{MC}) < P_C(\text{MC}) \]
SG: Center Embedding

(A) The painting_{N_1} that the artist_{N_2} who lived long ago painted_{V_2} deteriorated_{V_1}. [correct]

(B) #The painting_{N_1} that the artist_{N_2} who lived long ago deteriorated_{V_1} painted_{V_2}. [incorrect]

\[ P_A(V_2 V_1) > P_B(V_1 V_2) \]

\[ P(\text{painted deteriorated} | \text{The painting that the artist}) > P(\text{deteriorated painted} | \text{The painting that the artist}) \]
(A) I know that our uncle grabbed the food in front of the guests at the holiday party. [THAT, NO GAP]

(B) *I know what our uncle grabbed the food in front of the guests at the holiday party. [WH, NO GAP]

(C) ??I know that our uncle grabbed in front of the guests at the holiday party. [THAT, GAP]

(D) I know what our uncle grabbed in front of in front of the guests at the holiday party. [WH, GAP]
SG: Pseudo-Clefting

(A) What the worker did was \textcolor{green}{board the plane}.

(B) ?What the worker did was \textcolor{red}{the plane}.

(C) What the worker repaired was \textcolor{green}{the plane}.

(D) *What the worker repaired was \textcolor{red}{board the plane}.

\[ S_D(VP) > S_A(VP) \land S_B(NP) > S_C(NP) \]
SG: Assessment

accuracy_per_test_suite = correct predictions / total items

- Test for stability by including syntactically irrelevant but semantically plausible syntactic content before the critical region
  - E.g:
  - The keys to the cabinet on the left are on the table
  - *The keys to the cabinet on the left is on the table

- Compare model class to dataset size
SG: Score by Model Class

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

- BLLIP-XS: 1M tokens
- BLLIP-S: 5M tokens
- BLLIP-M: 14M tokens
- BLLIP-LG: 42M tokens
SG: Perplexity and SG Score
SG: Perplexity and Brain-Score

$r = .44$

Schrimpf et al., 2020
SG: The Influence of Model Architecture
SG: The Influence of Model Architecture

- Architectures as priors to the linguistic representation that can be developed
- Robustness depends on model architecture
SG: The Influence of Dataset Size
SG: The Influence of Dataset Size

![Graph showing the influence of dataset size on SG score]
The Influence of Dataset Size

- Increasing amount of training data yields diminishing returns:
  - “(...) require over 10 billion tokens to achieve human-like performance, and most would require trillions of tokens to achieve perfect accuracy – an impractically large amount of training data, especially for these relatively simple syntactic phenomena.” (van Schijndel et al., 2019)

- Limited data efficiency

- Structured architectures or explicit syntactic supervision

- Humans? 11-27 million total words of input per year? (Hart & Risley, 1995; Brysbaert et al., 2016)
SG: The Influence of Dataset Size

Figure 5: Evaluation results on all models, split across test suite circuits.
CLUTRR: Motivation and Paradigm

- **Compositional Language Understanding and Text-based Relational Reasoning**
- Kinship inductive reasoning
- Unseen combinations of logical rules
- Model robustness

Kristin and her son Justin went to visit her mother Carol on a nice Sunday afternoon. They went out for a movie together and had a good time.

Q: How is Carol related to Justin?
A: Carol is the grandmother of Justin
CLUTRR: Motivation and Paradigm

- **Productivity**
  - mother(mother(mother(Justin))) ~ great grandmother of Justin

- **Systematicity**
  - Only certain sets allowed with symmetries: son(Justin, Kristin) ~ mother(Kristin, Justin)

- **Compositionality**
  - son(Justin, Kristin) consists of components

- **Memory (compression)**

- **Children are not exposed to systematic dataset**
CLUTRR: Dataset Generation & Paradigm

Step 1

Step 2

Step 3

KB:
1. \( g(X,Y) \) \( :- w(X,Z), g(Z,Y) \)
2. \( g(X,Y) \) \( :- d(X,Z), m(Z,Y) \)

\[
g(B,G) = [w(B,A), g(A,G)] = [w(B,A), [d(A,D), m(D,G)]]
\]

Step 4

Story

B is the wife of A
D is the daughter of A
D is the mother of G

QA

(B, G) \( \rightarrow \) B is the \textit{grandmother} of G
CLUTRR: Model Robustness
CLUTRR: Systematic Generalization

Systematic Generalization - Trained on k=2 and k=3

Systematic Generalization - Trained on k=2, 3 and 4
# CLUTRR: Model Robustness

<table>
<thead>
<tr>
<th>Models</th>
<th>Training</th>
<th>Testing</th>
<th>BiLSTM - Attention</th>
<th>BiLSTM - Mean</th>
<th>RN</th>
<th>MAC</th>
<th>BERT</th>
<th>BERT-LSTM</th>
<th>Structured model (with graph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Clean</td>
<td></td>
<td>0.58 ± 0.05</td>
<td>0.53 ± 0.05</td>
<td>0.49 ± 0.06</td>
<td>0.63 ± 0.08</td>
<td>0.37 ± 0.06</td>
<td>0.67 ± 0.03</td>
<td>1.0 ± 0.0</td>
</tr>
<tr>
<td>Supporting</td>
<td></td>
<td></td>
<td>0.76 ± 0.02</td>
<td>0.64 ± 0.22</td>
<td>0.58 ± 0.06</td>
<td>0.71 ± 0.07</td>
<td>0.28 ± 0.1</td>
<td>0.66 ± 0.06</td>
<td>0.24 ± 0.2</td>
</tr>
<tr>
<td>Irrelevant</td>
<td></td>
<td></td>
<td>0.7 ± 0.15</td>
<td>0.76 ± 0.02</td>
<td>0.59 ± 0.06</td>
<td>0.69 ± 0.05</td>
<td>0.24 ± 0.08</td>
<td>0.55 ± 0.03</td>
<td>0.51 ± 0.15</td>
</tr>
<tr>
<td>Disconnected</td>
<td></td>
<td></td>
<td>0.49 ± 0.05</td>
<td>0.45 ± 0.05</td>
<td>0.5 ± 0.06</td>
<td>0.59 ± 0.05</td>
<td>0.24 ± 0.08</td>
<td>0.5 ± 0.06</td>
<td>0.8 ± 0.17</td>
</tr>
<tr>
<td>Supporting</td>
<td>Supporting</td>
<td></td>
<td>0.67 ± 0.06</td>
<td>0.66 ± 0.07</td>
<td>0.68 ± 0.05</td>
<td>0.65 ± 0.04</td>
<td>0.32 ± 0.09</td>
<td>0.57 ± 0.04</td>
<td>0.98 ± 0.01</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Irrelevant</td>
<td></td>
<td>0.51 ± 0.06</td>
<td>0.52 ± 0.06</td>
<td>0.5 ± 0.04</td>
<td>0.56 ± 0.04</td>
<td>0.25 ± 0.06</td>
<td>0.53 ± 0.06</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>Disconnected</td>
<td>Disconnected</td>
<td></td>
<td>0.57 ± 0.07</td>
<td>0.57 ± 0.06</td>
<td>0.45 ± 0.11</td>
<td>0.4 ± 0.1</td>
<td>0.17 ± 0.05</td>
<td>0.47 ± 0.06</td>
<td>0.96 ± 0.01</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>0.61 ± 0.08</td>
<td>0.59 ± 0.08</td>
<td>0.54 ± 0.07</td>
<td>0.61 ± 0.06</td>
<td>0.30 ± 0.07</td>
<td>0.56 ± 0.05</td>
<td>0.77 ± 0.09</td>
</tr>
</tbody>
</table>
## CLUTRR: Model Robustness (noisy training)

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>BiLSTM - Attention</th>
<th>BiLSTM - Mean</th>
<th>RN</th>
<th>MAC</th>
<th>BERT</th>
<th>BERT-LSTM</th>
<th>GAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporting</td>
<td>Clean</td>
<td>0.38 ±0.04</td>
<td>0.32 ±0.04</td>
<td>0.45 ±0.03</td>
<td>0.19 ±0.06</td>
<td>0.39 ±0.06</td>
<td>0.92 ±0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.67 ±0.06</td>
<td>0.66 ±0.07</td>
<td>0.65 ±0.04</td>
<td>0.32 ±0.09</td>
<td>0.57 ±0.04</td>
<td>0.98 ±0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.44 ±0.03</td>
<td>0.39 ±0.03</td>
<td>0.51 ±0.08</td>
<td>0.2 ±0.06</td>
<td>0.36 ±0.05</td>
<td>0.5 ±0.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.31 ±0.21</td>
<td>0.25 ±0.16</td>
<td>0.47 ±0.05</td>
<td>0.2 ±0.08</td>
<td>0.32 ±0.04</td>
<td>0.92 ±0.06</td>
<td></td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Clean</td>
<td>0.57 ±0.05</td>
<td>0.56 ±0.05</td>
<td>0.46 ±0.13</td>
<td>0.24 ±0.06</td>
<td>0.46 ±0.08</td>
<td>0.92 ±0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.38 ±0.22</td>
<td>0.31 ±0.16</td>
<td>0.61 ±0.07</td>
<td>0.27 ±0.06</td>
<td>0.46 ±0.04</td>
<td>0.77 ±0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.51 ±0.08</td>
<td>0.52 ±0.04</td>
<td>0.56 ±0.04</td>
<td>0.25 ±0.06</td>
<td>0.53 ±0.00</td>
<td>0.93 ±0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.44 ±0.26</td>
<td>0.54 ±0.27</td>
<td>0.55 ±0.05</td>
<td>0.61 ±0.06</td>
<td>0.45 ±0.08</td>
<td>0.85 ±0.25</td>
<td></td>
</tr>
<tr>
<td>Disconnected</td>
<td>Clean</td>
<td>0.45 ±0.02</td>
<td>0.47 ±0.03</td>
<td>0.53 ±0.09</td>
<td>0.22 ±0.09</td>
<td>0.44 ±0.05</td>
<td>0.75 ±0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supporting</td>
<td>0.47 ±0.03</td>
<td>0.46 ±0.05</td>
<td>0.54 ±0.03</td>
<td>0.22 ±0.06</td>
<td>0.38 ±0.08</td>
<td>0.78 ±0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irrelevant</td>
<td>0.47 ±0.03</td>
<td>0.48 ±0.03</td>
<td>0.52 ±0.04</td>
<td>0.17 ±0.04</td>
<td>0.38 ±0.05</td>
<td>0.56 ±0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disconnected</td>
<td>0.57 ±0.07</td>
<td>0.57 ±0.06</td>
<td>0.45 ±0.11</td>
<td>0.17 ±0.05</td>
<td>0.47 ±0.06</td>
<td>0.96 ±0.01</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.47 ±0.08</td>
<td>0.46 ±0.08</td>
<td>0.52 ±0.07</td>
<td>0.23 ±0.07</td>
<td>0.43 ±0.05</td>
<td>0.82 ±0.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Testing the robustness of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.
Future work & Perspectives

- Sub-word tokenization
- Active attention and reasoning
- Generalization across tasks
- Abstractions as probabilistic
- Architecture and dimensionality reduction


Supplementary
Figure 6: Systematic Generalizability of different models on CLUTRR-Gen task (having 20% less placeholders and without training and testing placeholder split), when **Left**: trained with $k = 2$ and $k = 3$ and **Right**: trained with $k = 2, 3$ and $4$.
## CLUTTR, Table 5

<table>
<thead>
<tr>
<th>Relation Length</th>
<th>Human Performance</th>
<th>Reported Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time Limited</td>
<td>Unlimited Time</td>
</tr>
<tr>
<td>2</td>
<td>0.848</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.773</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.477</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.424</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.406</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5: Human performance accuracies on CLUTTR dataset. Humans are provided the Clean-Generalization version of the dataset, and we test on two scenarios: when a human is given limited time to solve the task, and when a human is given unlimited time to solve the task. Regardless of time, our evaluators provide a score of difficulty of individual puzzles.
Table 4: Testing the robustness on toy placeholders of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.
CLUTTR, Fig. 7

Figure 7: Systematic Generalization comparison with different Embedding policies
Figure 2: Lines depict number of training tokens needed for LSTMs to achieve human-like (left) or 99.99% accuracy (right) in each syntactic agreement condition, according to our estimates. Bars depict the amount of data on which each model was trained.