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

Greta Tuckute & Kamoya K Ikhofua

MIT Fall 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?

SG: Man shall not live by perplexity alone

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 test human-level understanding of syntax

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$_{\mathrm{V}_{sg}}$ many people.

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

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

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

$$P_{\mathrm{A}}(\mathrm{V}_{\mathrm{sg}}) > P_{\mathrm{B}}(\mathrm{V}_{\mathrm{pl}}) \wedge P_{\mathrm{C}}(\mathrm{V}_{\mathrm{pl}}) > P_{\mathrm{D}}(\mathrm{V}_{\mathrm{sg}})$$

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 had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [+NEG,–DISTRACTOR]

(B) *The managers that respected no guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [–NEG,+DISTRACTOR]

(C) *The managers that respected the guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [–NEG,–DISTRACTOR]

(D) No managers that respected no guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [+NEG,+DISTRACTOR]

$$P_{\text{A}}(\text{NPI}) > P_{\text{C}}(\text{NPI}) \wedge P_{\text{D}}(\text{NPI}) > P_{\text{B}}(\text{NPI}) \wedge$$
$$P_{\text{A}}(\text{NPI}) > P_{\text{B}}(\text{NPI})$$

Acceptable orderings:

ADBC
ADCB
DABC
DACB
ACDB (?)

Chance: 5/24

SG: NP/Z Garden-Paths

(A) !As the ship crossed the waters $\overbrace{\text{remained}}^{\mathrm{V}^*}$ blue and calm. [TRANS,NO COMMA]

(B) As the ship crossed, the waters $\overbrace{\text{remained}}^{\mathrm{V}^*}$ blue and calm. [TRANS,COMMA]

(C) As the ship drifted the waters $\overbrace{\text{remained}}^{\mathrm{V}^*}$ blue and calm. [INTRANS,NO COMMA]

(D) As the ship drifted, the waters $\overbrace{\text{remained}}^{\mathrm{V}^*}$ blue and calm. [INTRANS,COMMA]

$$S_{\mathrm{A}}(\mathrm{V}^*) > S_{\mathrm{B}}(\mathrm{V}^*) \wedge S_{\mathrm{A}}(\mathrm{V}^*) > S_{\mathrm{C}}(\mathrm{V}^*) \wedge$$
$$S_{\mathrm{A}}(\mathrm{V}^*) - S_{\mathrm{B}}(\mathrm{V}^*) > S_{\mathrm{C}}(\mathrm{V}^*) - S_{\mathrm{D}}(\mathrm{V}^*)$$

(Main Verb / Reduced Relative Clause paths were also included in sub-class suites)

SG: Gross Syntactic Expectation

(A) The minister praised the building $\overbrace{\text{.}}^{\textsc{End}}$

(B) *After the minister praised the building $\overbrace{\text{.}}^{\textsc{End}}$

(C) ??The minister praised the building, $\overbrace{\text{it started to rain.}}^{\text{MC}}$

(D) After the minster praised the building, $\overbrace{\text{it started to rain.}}^{\text{MC}}$

$$P_{\mathrm{A}}(\textsc{End}) > P_{\mathrm{B}}(\textsc{End}) \wedge P_{\mathrm{D}}(\mathrm{MC}) < P_{\mathrm{C}}(\mathrm{MC})$$

SG: Center Embedding

The paintings that the artist painted deteriorated

>

*The paintings that the artist deteriorated painted

SG: Long Distance Dependencies

The **keys** to the cabinet **are** on the table

>

*The **keys** to the cabinet **is** on the table

SG: Cleft

The **keys** to the cabinet **are** on the table

>

*The **keys** to the cabinet **is** on the table

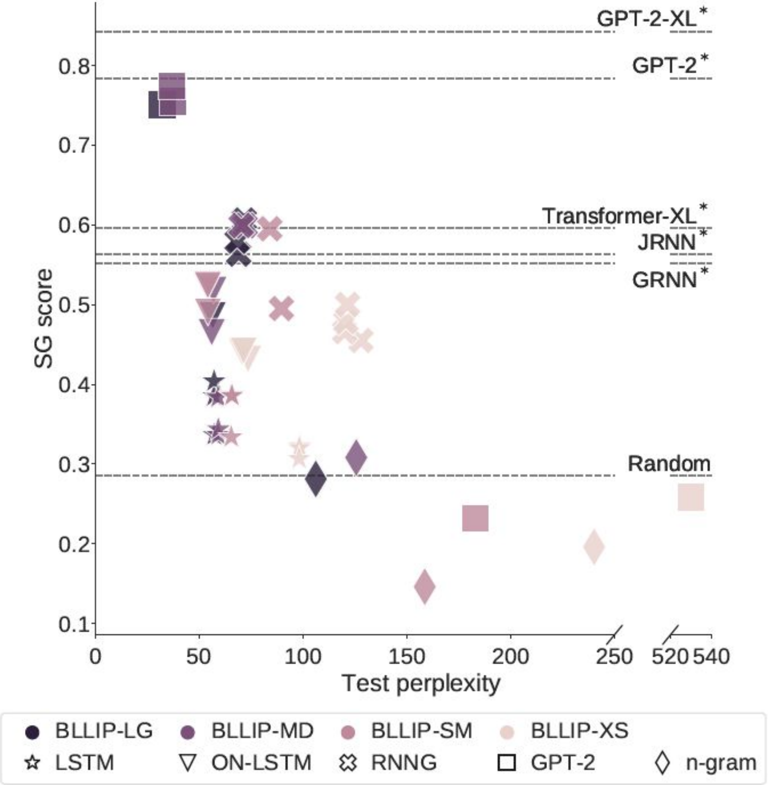
Syntactic Generalization

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

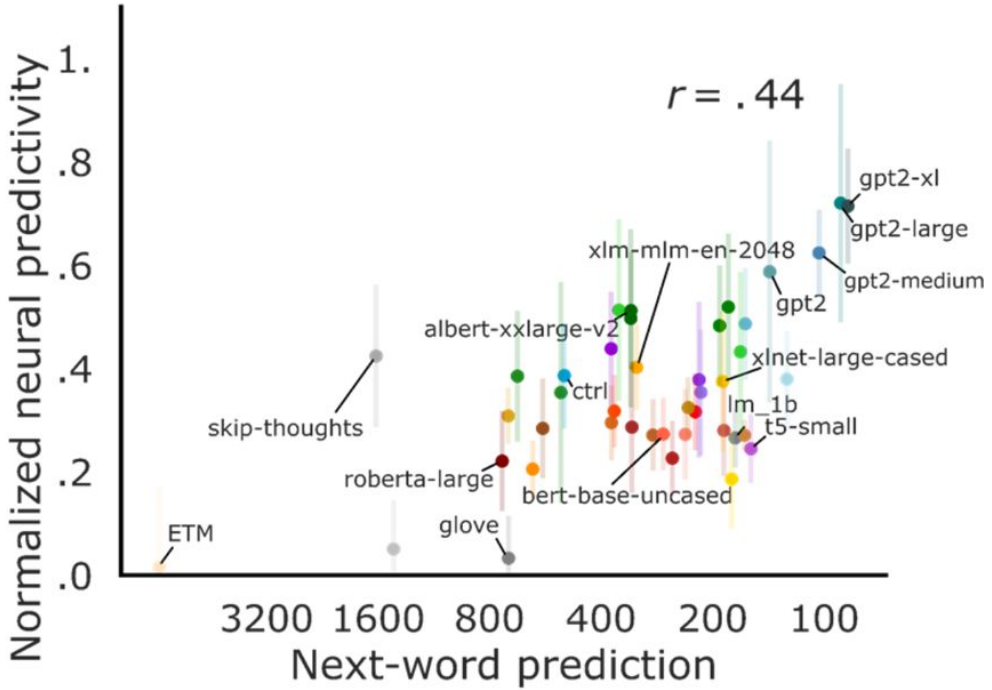
- Test for stability by including syntactically irrelevant but semantically plausible syntactic content before the critical region
    - E.g:
    - The keys to the cabinet on the left are on the table
    - *The keys to the cabinet on the left is on the table

- Compare model class to dataset size

SG: Perplexity and SG Score

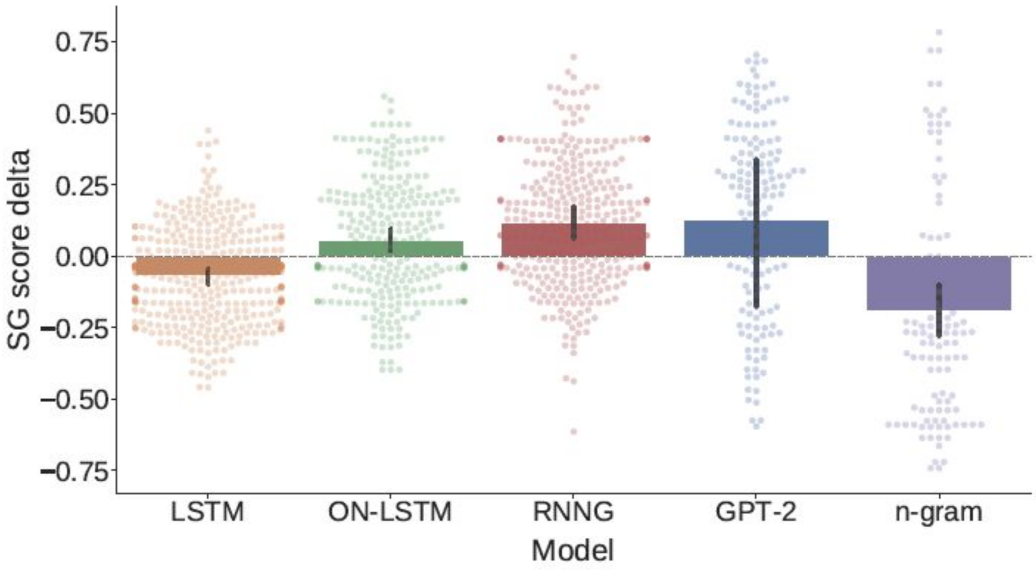


(SG:) Perplexity and Brain-Score



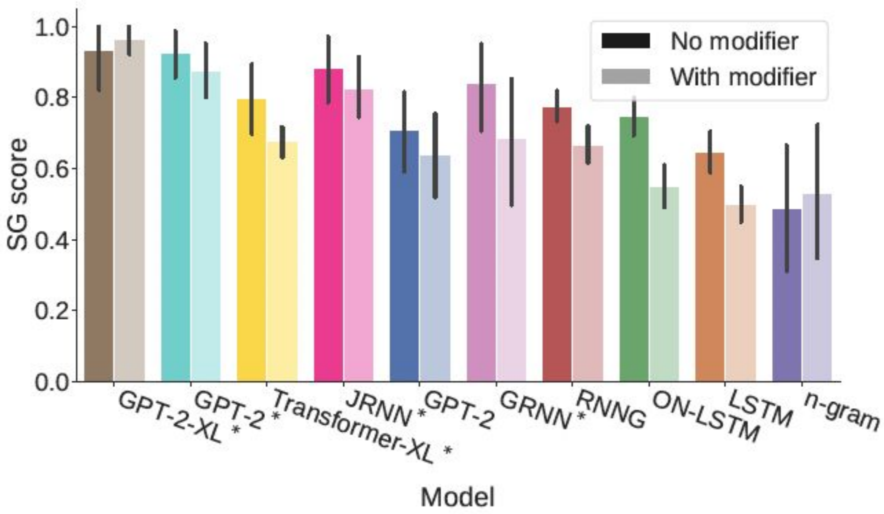
Schrimpf et al., 2020

SG: The Influence of Model Architecture

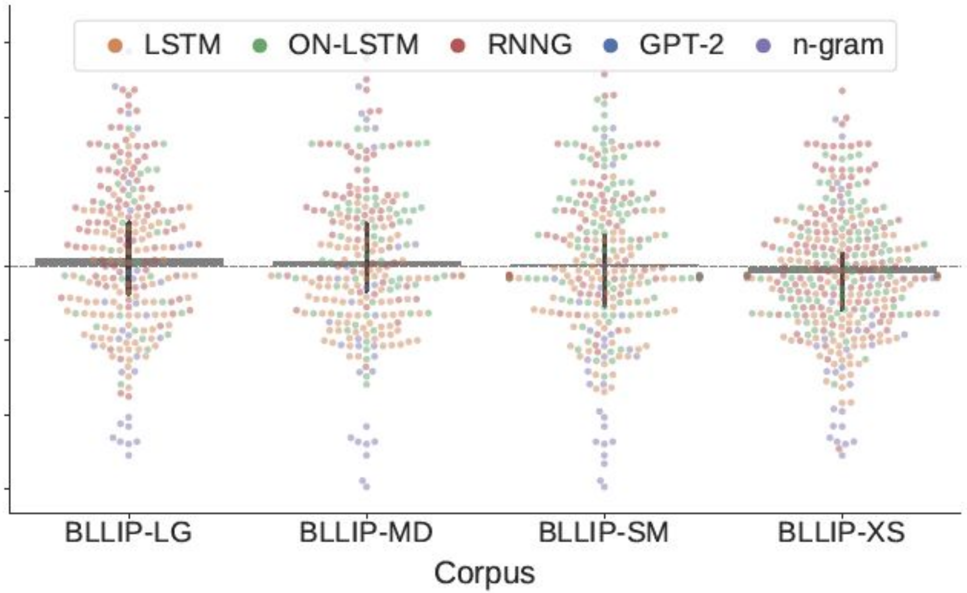


SG: The Influence of Model Architecture

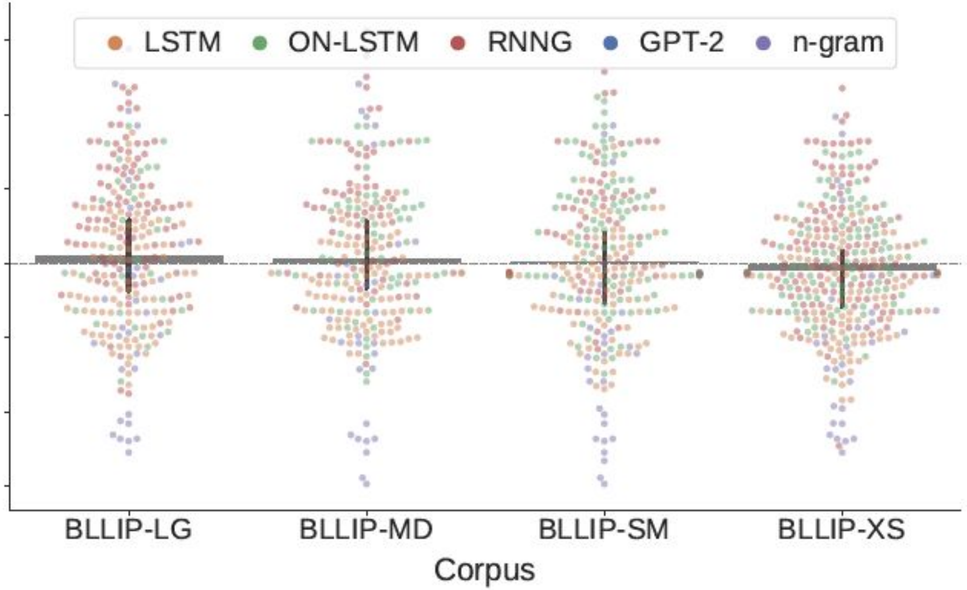
- Architectures as priors to the linguistic representation that can be developed
- Robustness depends on model architecture

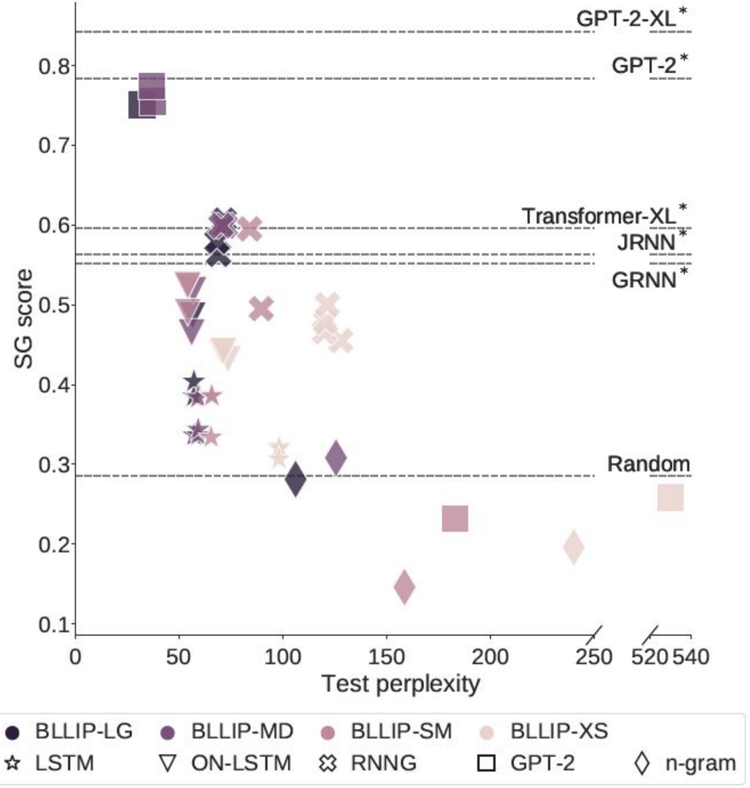


SG: The Influence of Dataset Size



SG: The Influence of Dataset Size





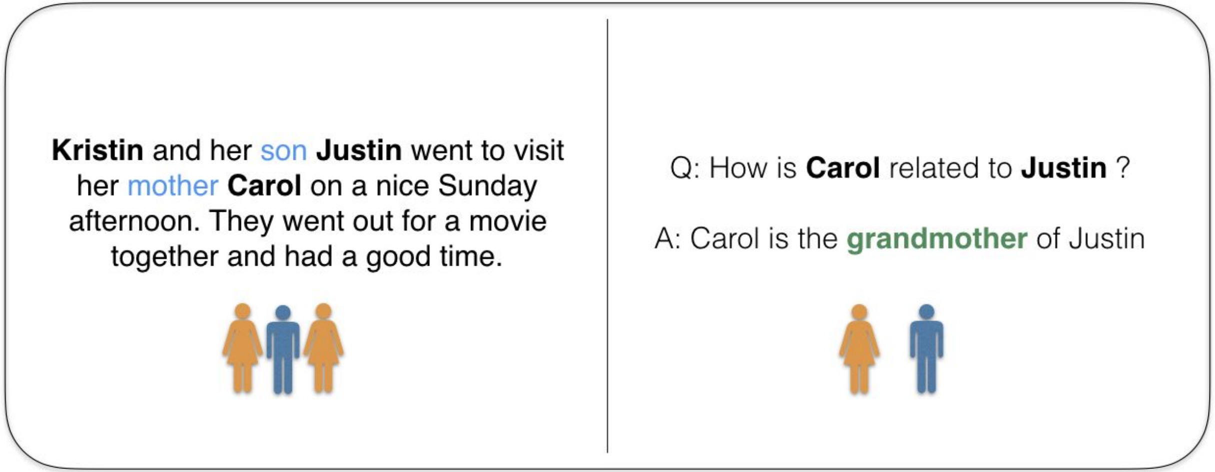
SG: The Influence of Dataset Size

- 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)

CLUTRR: Motivation and Paradigm

- **C**ompositional **L**anguage **U**nderstanding and **T**ext-based **R**elational **R**easoning
- Kinship inductive reasoning

- Unseen combinations of logical rules
- Model robustness

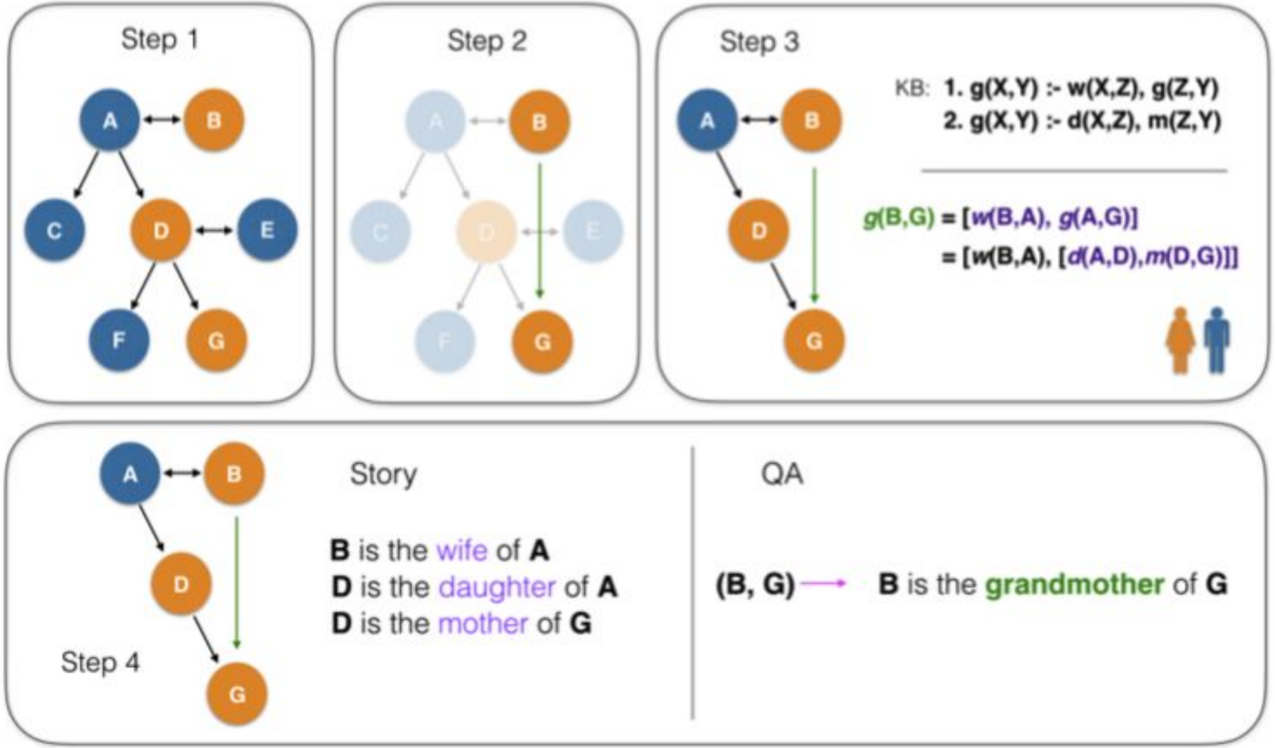


CLUTRR: Motivation and Paradigm

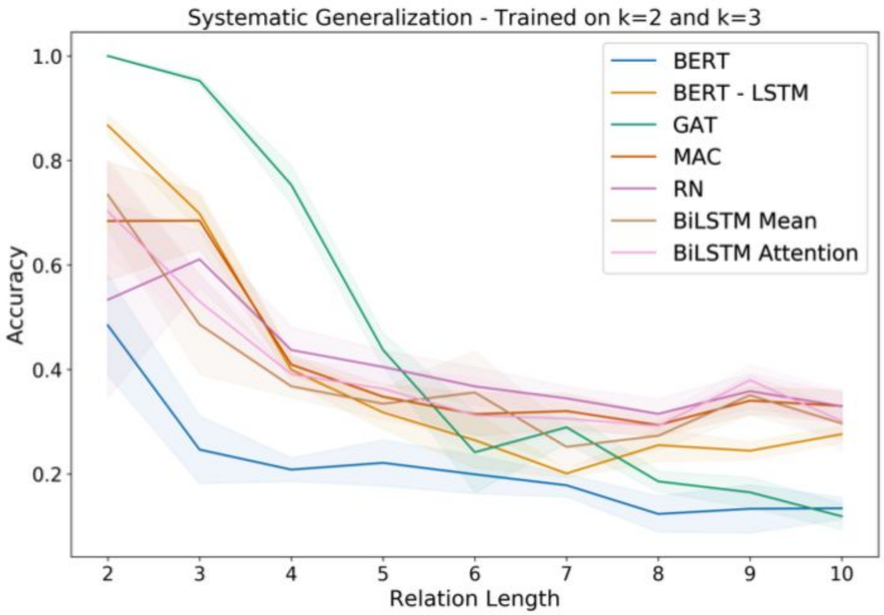
- Productivity
  - mother(mother(mother(Justin))) ~ great grandmother of Justin
- Systematicity
  - Only certain sets allowed with symmetries: son(Justin, Kristin) ~ mother(Kristin, Justin)
- Compositionality
  - son(Justin, Kristin) consists of components

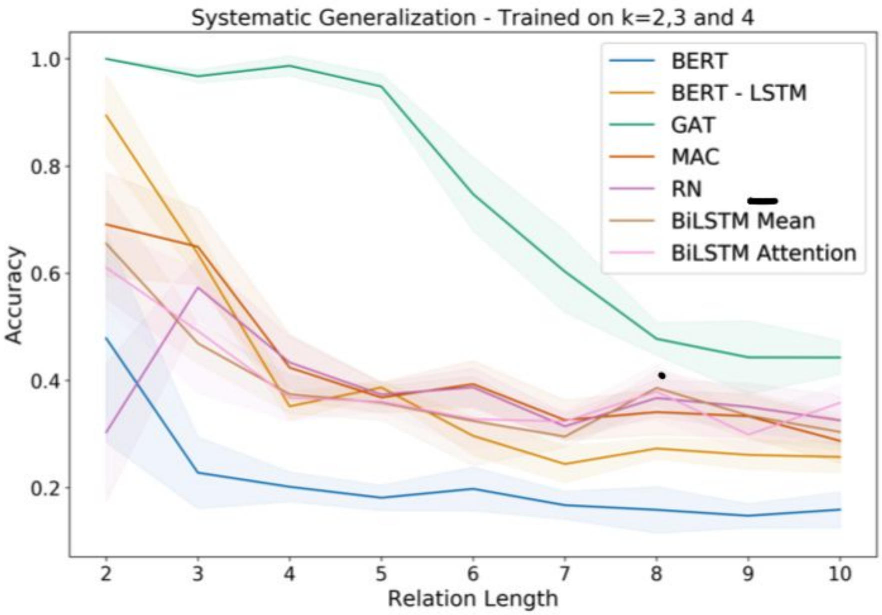
- Memory (compression)
- Children are not exposed to systematic dataset

CLUTRR: Dataset Generation & Paradigm



CLUTRR: Experiment Results

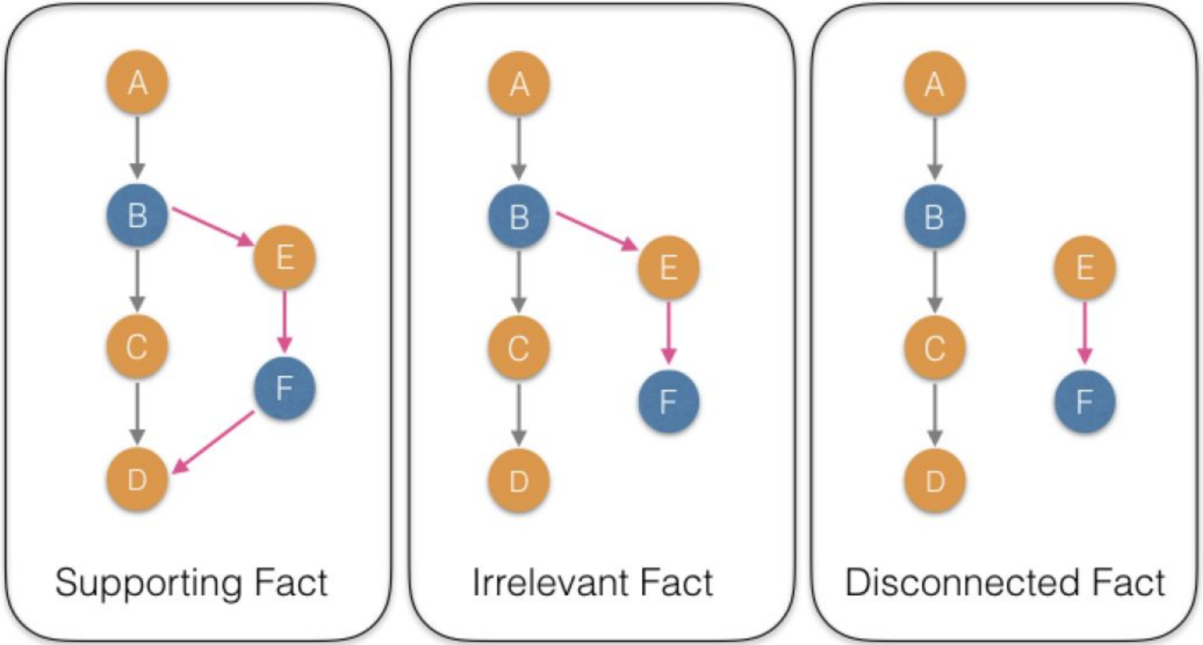




CLUTRR: Experiment Results

Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Clean	Clean	0.58 ±0.05	0.53 ±0.05	0.49 ±0.06	0.63 ±0.08	0.37 ±0.06	0.67 ±0.03	**1.0** ±0.0
	Supporting	**0.76** ±0.02	0.64 ±0.22	0.58 ±0.06	0.71 ±0.07	0.28 ±0.1	0.66 ±0.06	0.24 ±0.2
	Irrelevant	0.7 ±0.15	**0.76** ±0.02	0.59 ±0.06	0.69 ±0.05	0.24 ±0.08	0.55 ±0.03	0.51 ±0.15
	Disconnected	0.49 ±0.05	0.45 ±0.05	0.5 ±0.06	0.59 ±0.05	0.24 ±0.08	0.5 ±0.06	**0.8** ±0.17
Supporting	Supporting	0.67 ±0.06	0.66 ±0.07	0.68 ±0.05	0.65 ±0.04	0.32 ±0.09	0.57 ±0.04	**0.98** ±0.01
Irrelevant	Irrelevant	0.51 ±0.06	0.52 ±0.06	0.5 ±0.04	0.56 ±0.04	0.25 ±0.06	0.53 ±0.06	**0.93** ±0.01
Disconnected	Disconnected	0.57 ±0.07	0.57 ±0.06	0.45 ±0.11	0.4 ±0.1	0.17 ±0.05	0.47 ±0.06	**0.96** ±0.01
Average		**0.61** ±0.08	0.59 ±0.08	0.54 ±0.07	**0.61** ±0.06	0.30 ±0.07	0.56 ±0.05	**0.77** ±0.09

CLUTRR: Model Robustness



CLUTRR: Model Robustness

Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Clean	Clean	0.58 ±0.05	0.53 ±0.05	0.49 ±0.06	0.63 ±0.08	0.37 ±0.06	0.67 ±0.03	**1.0** ±0.0
	Supporting	**0.76** ±0.02	0.64 ±0.22	0.58 ±0.06	0.71 ±0.07	0.28 ±0.1	0.66 ±0.06	0.24 ±0.2
	Irrelevant	0.7 ±0.15	**0.76** ±0.02	0.59 ±0.06	0.69 ±0.05	0.24 ±0.08	0.55 ±0.03	0.51 ±0.15
	Disconnected	0.49 ±0.05	0.45 ±0.05	0.5 ±0.06	0.59 ±0.05	0.24 ±0.08	0.5 ±0.06	**0.8** ±0.17
Supporting	Supporting	0.67 ±0.06	0.66 ±0.07	0.68 ±0.05	0.65 ±0.04	0.32 ±0.09	0.57 ±0.04	**0.98** ±0.01
Irrelevant	Irrelevant	0.51 ±0.06	0.52 ±0.06	0.5 ±0.04	0.56 ±0.04	0.25 ±0.06	0.53 ±0.06	**0.93** ±0.01
Disconnected	Disconnected	0.57 ±0.07	0.57 ±0.06	0.45 ±0.11	0.4 ±0.1	0.17 ±0.05	0.47 ±0.06	**0.96** ±0.01
Average		**0.61** ±0.08	0.59 ±0.08	0.54 ±0.07	**0.61** ±0.06	0.30 ±0.07	0.56 ±0.05	**0.77** ±0.09

Future work & Perspectives

- Sub-word tokenization
- Common-sense reasoning
- Abstractions as probabilistic
-

References

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes Publishing Company.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., Fedorenko, E (2020): Artificial Neural Networks Accurately Predict Language Processing in the Brain, *bioRxiv* 2020.06.26.174482; doi: https://doi.org/10.1101/2020.06.26.174482.