Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments

Authors: Alex Warstadt, Samuel R. Bowman

Models: 3 Pretrained Encoders:

- BERT LARGE
  - Trained using masked language modeling and next sentence prediction tasks.
- GPT
  - Trained using a standard language modeling task
- BiLSTM baseline - (what is our baseline if we do not have this model at our disposal?)
  - This transformer might not work in the NLPScholar toolkit. Ask Grusha whether we should use this model in our replication.

  These are all publicly available and should be in the toolkit, minus BiLSTM

Goal:

- Evaluate whether models such as BERT and GPT understand general grammatical knowledge.

Dataset

- Corpus of Linguistic Acceptability (CoLA)
  - A dataset of over 10k example sentences labeled for acceptability and sampled from linguistics publications discussing numerous linguistic phenomena.
    - The base form of the CoLA dataset does not distinguish between these linguistic phenomena, which is an issue.
  - The authors augment CoLA with a new syntactically annotated evaluation set that combines fine-grained and domain genera.
    - 13 Total syntactic phenomena are being tested.
      - There are more specific phenomena under these categories.
- CoLA & Acceptability Classification
  - The adapted dataset with grammatical annotations

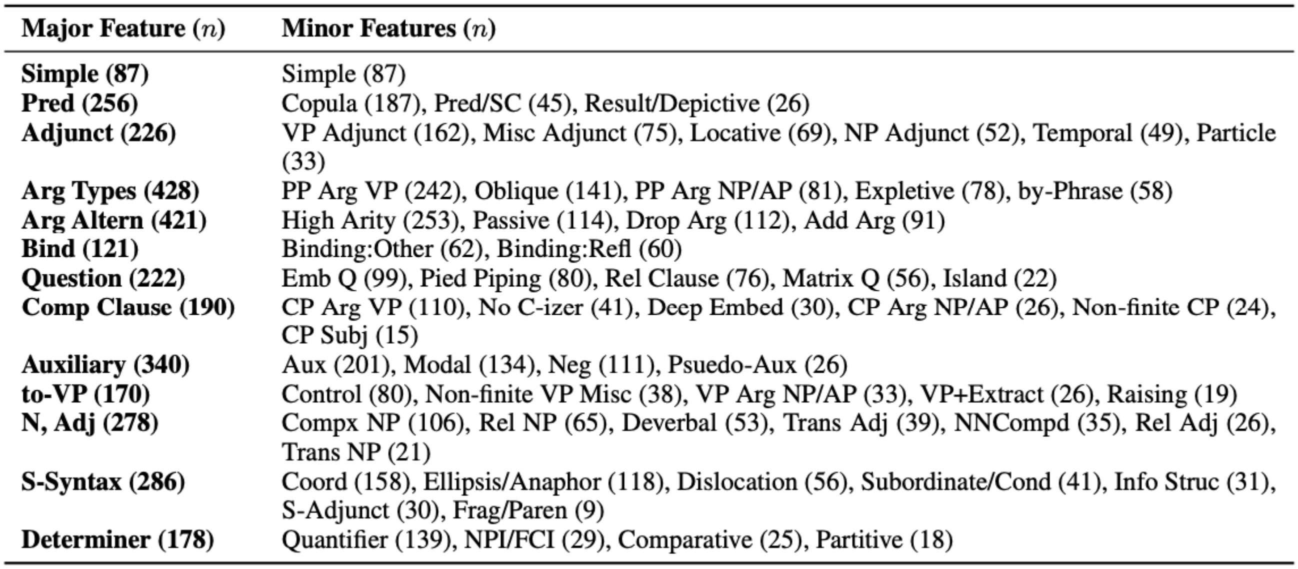
Analysis Set

- 1043 sentences labeled with 13 major features, further divided into 59 minor features. Shown in TABLE 5.
- Each of the 59 minor features belongs to a single major feature.
- A sentence belongs to a major feature if it belongs to one or more relevant minor features classified under that major feature.

The average sentence is positively labeled with 3.22 major features (SD=1.66) on average, and the average a major feature is present in 224 sentences (SD=112).

The average sentence is positively labeled with 4.31 minor features (SD=2.59). The average minor feature is present in 71.3 sentences (SD=54.7). Every sentence is labeled with at least one feature. Sentences without any obvious phenomena of interest are labeled SIMPLE.

Grammatical Features



Major Feature (*n*)	**Minor Features** (*n*)
Simple (87)	Simple (87)
Pred (256)	Copula (187), Pred/SC (45), Result/Depictive (26)
Adjunct (226)	VP Adjunct (162), Misc Adjunct (75), Locative (69), NP Adjunct (52), Temporal (49), Particle (33)
Arg Types (428)	PP Arg VP (242), Oblique (141), PP Arg NP/AP (81), Expletive (78), by-Phrase (58)
Arg Altern (421)	High Arity (253), Passive (114), Drop Arg (112), Add Arg (91)
Bind (121)	Binding:Other (62), Binding:Refl (60)
Question (222)	Emb Q (99), Pied Piping (80), Rel Clause (76), Matrix Q (56), Island (22)
Comp Clause (190)	CP Arg VP (110), No C-izer (41), Deep Embed (30), CP Arg NP/AP (26), Non-finite CP (24), CP Subj (15)
Auxiliary (340)	Aux (201), Modal (134), Neg (111), Psuedo-Aux (26)
to-VP (170)	Control (80), Non-finite VP Misc (38), VP Arg NP/AP (33), VP+Extract (26), Raising (19)
N, Adj (278)	Compx NP (106), Rel NP (65), Deverbal (53), Trans Adj (39), NNCompd (35), Rel Adj (26), Trans NP (21)
S-Syntax (286)	Coord (158), Ellipsis/Anaphor (118), Dislocation (56), Subordinate/Cond (41), Info Struc (31), S-Adjunct (30), Frag/Paren (9)
Determiner (178)	Quantifier (139), NPI/FCI (29), Comparative (25), Partitive (18)

Table 2: Major features and their associated minor features (with number of occurrences *n*).

The data is available to be downloaded at the end of the first page

- Both the original CoLA dataset and the grammatically annotated CoLA datasets.

Model Performance

Long-distance dependencies:

- What do you think I ate?

  Training data weakness vs model weakness

Correlations in Categories

Reasons for Overlap

1. Some features have overlapping definitions.
   - e.g., `Expletive` is a strict subset of `ADD ARG`

2. Grammatical facts of English drive the correlation between, for instance, `Question` and `AUX`.
3. An unusually high correlation between `EMB-Q` and `ELLIPSIS/ANAPHOR` can be attributed mainly to a bias in a particular source in CoLA.