Project Overview

For this project, I investigated how does style of language used for training effect the ability of a model to learn specific grammatical paradigm. I trained two small BabyLMs on texts from different domains and evaluated them on the same grammatical paradigm to compare their performance.

Both models, training data, test dataset, scraping script and the code for both training and evaluating models are on github: https://github.com/jitka1997/Lab-NN-project (its a link)

The github repo has a small README to know where to find everything (folders for models for example).

Training Data

For training data I scraped texts from Project Gutenberg and created two 2.5MB datasets representing different writing domains:

Formal/Academic Texts:

- Declaration of Independence
- Federalist Papers
- Darwin's Origin of Species
- Adam Smith's Wealth of Nations

Fiction Texts:

- Pride and Prejudice
- The Great Gatsby
- Dracula
- Frankenstein
- Sherlock Holmes stories

I selected these texts because they represent fundamentally different writing styles - formal writing versus narrative, creative writing.

Evaluation

I tested both models on the minimal pair set called "animate subject trans" from the BLiMP datasets [A. Warstadt et al., 2020]. It consists of 1000 minimal pairs. This paradigm evaluates how well models understand that only animate agents can perform certain actions. Example below show a pretty good explanation.

- ✅ (Good) "Danielle visited Irene" (people can visit)
- ❌ (Bad) "The eye visited Irene" (body parts cant visit)

Used measurements

I measured surprisal scores from minicons on the minimal pair dataset for both models. From this I determined for each pair if the model was correct or wrong by saying it was correct if it had higher surprisal score on the bad sentence. Then I calculated the overall accuracy of a model as `(#correct_pairs/#all_pairs) * 100%`.

I also calculated the differences of surprisal scores between good and bad sentences for each pair separately for correct and wrong pairs.

Model Configuration

I trained the models using the notebook from lecture with these hyperparameters:

- Hidden size: 128
- Number of hidden layers: 6
- Intermediate size: 512
- Attention heads: 8

I trained both models for 5 epochs. After 5 epochs the validation loss started to grow again, so I suspect the model was only getting more overfitted after this.

Hypothesis and results

I chose this paradigm because I expected it could show different performance on the models. My hypotheses was that the fiction-trained model (fiction model) would outperform the formal-trained model (formal model) because narrative texts contain a lot of examples of characters performing actions specifically in a form [subject] [verb] [object]. On the other hand formal or academic text usually use pasive form, in which [subject] is/was [verb]ed by [agent].

The results contradicted my hypothesis:

- **Formal model**: 62.9% accuracy
- **Fiction model**: 51.8% accuracy

Both models are above the baseline of random guess (accuracy 50%), so we can say they learned this paradigm at least at some level. The suprisal score differences are arguably big enough to mean something. The means are:

- **Formal model, correct pairs**: 11.84

- **Formal model, wrong pairs**: 8.63

- **Fiction model, correct pairs**: 9.96

- **Fiction model, wrong pairs**: 7.69

Analysis

My explanation for this unexpected outcome has two parts:

Advantages of formal training data:

Academic and political texts follow strict grammatical conventions where verbs are used in their literal, conventional meanings. And the goal is for the agent-action relationships to be clearly defined with as little ambiguity as possible.

Disadvantages of fiction training data:

Fictional texts have frequent use of metaphorical language and fictional constructs like personification assign agency to inanimate objects. Apart from this, generally creative language can violate conventional grammatical restrictions.

As an example from the dataset (pair 114):

- ✅ (Good) "Jesus hugged Paul."
  - Formal model surprisal: 73.17
  - Fiction model surprisal: 62.30
- ❌ (Bad) "The snake hugged Paul"
  - Formal model surprisal: 78.94
  - Fiction model surprisal: 55.48

Formal model had correctly lower surprisal on the good sentence compared to the bad one and the fiction model had it the other way around. So the formal model was correct and the fictional model was wrong. But in a fictional world we can easily imagine a snake hugging someone.

Conclusion

This project demonstrates that training domain impacts learning of a specific grammatical paradigm in BabyLMs. While fiction provides numerous examples of animate agents performing actions, the metaphorical and creative language characteristic of fictional texts appears to interfere with learning strict grammatical constraints. Formal texts, despite their different stylistic properties, offer more consistent patterns for learning fundamental argument structure knowledge.

These findings suggest that the consistency of training examples may be more important than their frequency for learning specific grammatical paradigms.

Reference

WWarstadt, Alex and Parrish, Alicia and Liu, Haokun and Mohananey, Anhad and Peng, Wei and Wang, Sheng-Fu and Bowman, Samuel R. - BLiMP: The Benchmark of Linguistic Minimal Pairs for English. (2020) *Transactions of the Association for Computational Linguistics, vol 8, pages 377-392*