

COSC 426LA F24 Lab 2

Introduction

The purpose of this lab is to give you hands on experience using the `NLP Scholar` toolkit to answer a question about the linguistic knowledge of a pre-trained transformer language model. By completing this lab, you will demonstrate that you can:

- Develop linguistically motivated evaluation data
- Apply the `NLP Scholar` toolkit to answer a research question
- Develop skills in conducting research

Pre-requisites This lab assumes that you have already cloned the `NLP Scholar` repository and have installed the `nlp` environment by following the instructions in `Install.md`.

Structure

This lab has one part:

1. Go from linguistic phenomenon to model evaluation

Provided files

- `Lab2.py`
- `sample_results.tsv`
- [A google doc template](#) to write responses
- [A google sheet template](#) to structure Part 1

What to submit

- A `tsv` of your data and a `tsv` of your results from Part 1
- A `pdf` of the google doc template with your answers

Part 0

Before starting each lab, get the latest version of the `NLP Scholar` repo by first navigating to the folder on terminal and then executing:

```
git pull
```

Additionally, a package is missing that you need for today. With the `nlp` environment activated run:

```
pip install seaborn
```

Part 1: Ideation (120 minutes)

Consider the following motivating examples:

1. Sally frightened Mary because she was so terrifying.
2. Sally feared Mary because she was so terrifying.

Technically, the pronoun in both 1 and 2 is ambiguous. However, speakers report strong preferences for who **she** should refer to in these sentences. Take a minute to check your judgments.

The core insight is that speakers prefer **she** to refer to the subject Sally in 1 and the object Mary in 2. The sentences are otherwise the same, so it must be the verbs **frightened** and **feared** which modulate preferences. That is, these sentences form a **minimal pair** where the main verb (**frightened** or **feared**) is varied.

In fact, many (possibly all) languages have verbs like this ([Harshorne et al., 2013](#)). These verbs are called **implicit causality** verbs. There are two types: **subject implicit causality** verbs like **frightened** and **object implicit causality** verbs like **feared**. Our research question today is **Do transformer-based language models learn implicit causality?** We will narrow this to a sub-question: **Does distilgpt2 learn the implicit causality bias of verbs?** Your tasks in this lab is to answer this question.

In this first part, think through with your group how can answer this question using the toolkit. Here's some things to keep in mind to help get you started:

- We can't ask what distilgpt2 thinks a pronoun refers to. We need some dependent measure.
- Can we make examples 1 and 2 into templates? Consider what else we could vary to help us see whether the model prefers subject or object referring pronouns.
- What if we varied the stereotypical gender of the subject and object in a sentence? Could we construct meaningful minimal pairs with a pronoun uniquely referring to either the subject or the object?
- distilgpt2 is a causal language model with the name **distilbert/distilgpt2** on HuggingFace. Try out some sentences using **interact** to see if you can use probability as a dependent measure.
- Examples of each type of verb are included below.

Subject IC	Object IC
frightened	feared
bored	believed
frustrated	encouraged
betrayed	cherished
amazed	blamed
confused	divorced
amused	revered
worried	trusted
haunted	liked
upset	valued

In this part, you should use the **interact** mode to test some initial ideas for how to evaluate the model's knowledge. To help scaffold you here, consider this google [sheet](#). It includes the format you should use to organize your experiment on the sheet labeled data. There are columns included to help you think through what information you should included. See the [MinimalPairAnalysis](#) document for more details on these column names.

Using **interact** mode you should fill in the results table with your initial explorations. You should develop by the end of this sentences and an initial result by aggregating over your results table (in the results sheet)