U4

Building a Multilingual Ellipsis Corpus

Calvin Josenhans[1], John Phillips[1], Khai Willard[2], Luis Abrego[3], Yuchen Yang[1], Niko Kilo[1] and *Dr. Damir Cavar[1]
[1]Department of Computer Science, [2]Department of Linguistics, [3]Department of Informatics
*corresponding author: dcavariu.edu

Introduction

- Ellipsis is a phenomena where words are omitted from a sentence, but the meaning of the sentence can still be discerned through context [1]
- Ellipsis can be found in a large variety of languages, and in many different forms
- Syntactic parsers often fail with such constructions
  - A lot of raw data contains ellipses, making this a problem
- Being able to reconstruct and parse omitted words allows us to overcome this limitation
- This project involves the collection of data and the training of several different types of models on that data for the purpose of enabling them to parse ellipses

Objective

We aim to collect and categorize a data set of ellipsis constructions that can be used to engineer NLP solutions in multiple languages

Methods

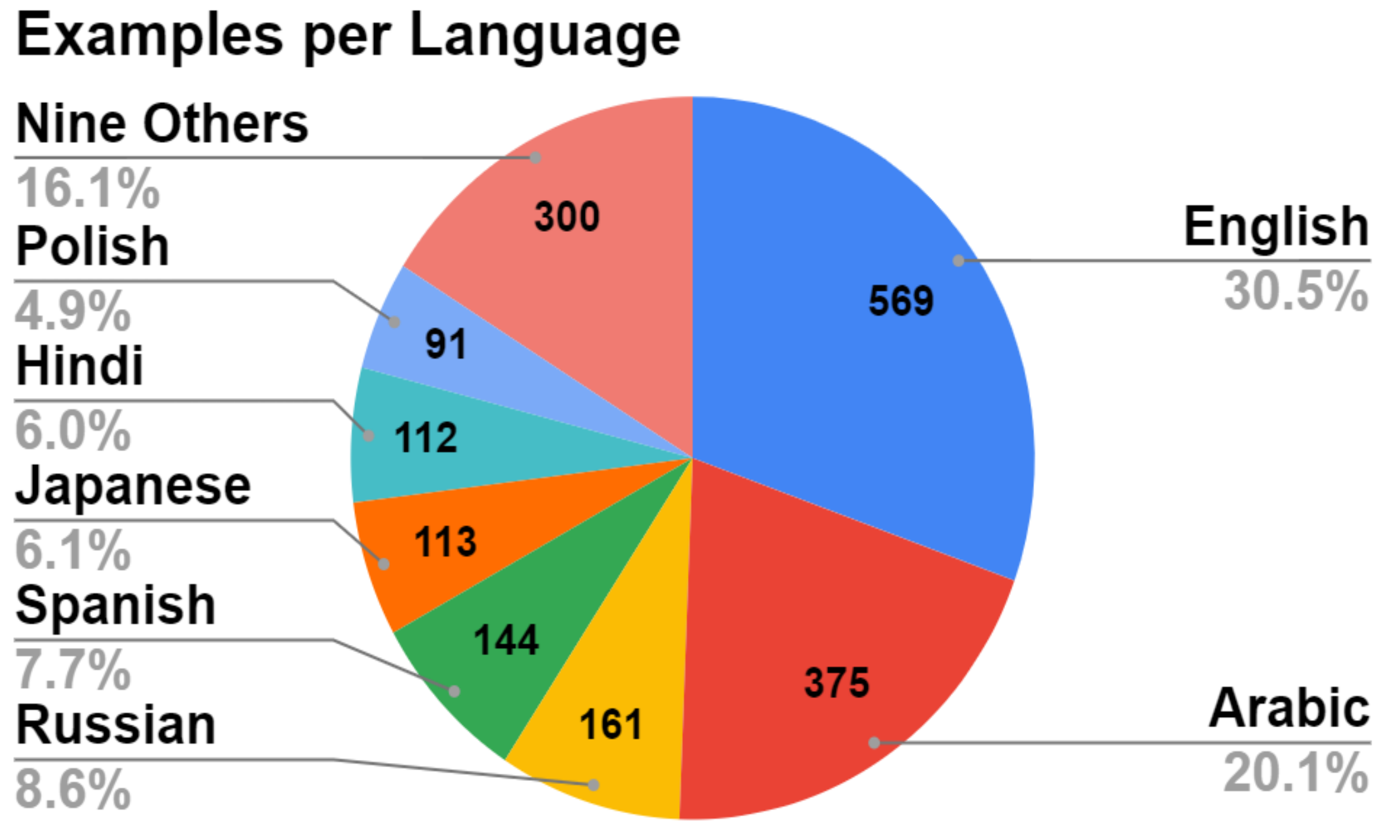
- **Data Collection**
  - Examples from academic papers and news articles
  - Mark the elided positions and words in the data collection.
  - Use collected data to create training and testing data for language models.
- **Data Format**
  - Ellipses data stored in the following format within the corpus:

```
Some ate bread, and others ___ rice.
----
Some ate bread, and others ate rice.
```

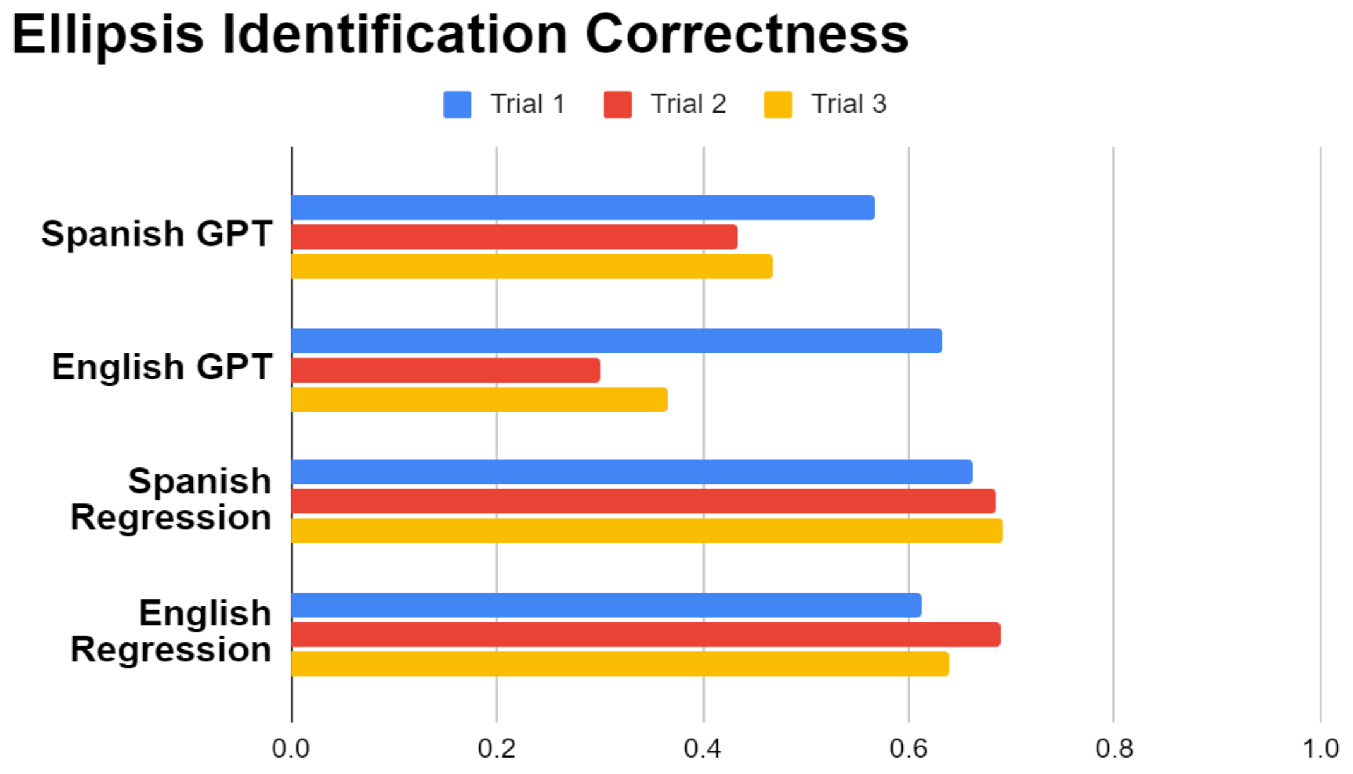
  - The sentence is given with elided material marked by underscores
  - The same sentence with all elided material provided is separated from the former sentence by four dashes
  - This provides the model with an ellipsis and the sentence's intended meaning
  - Any notes or credits are commented with a pound sign
- **Testing**
  - LLM's were given sentences and asked to identify any linguistic ellipsis.
    - ♦ Best performing was GPT-4 with 60% accuracy post-training – not meeting the benchmark
  - Models were given sentences with and without ellipses to train them to distinguish the two
  - Logistic regression model using the frequency of each part of speech as parameters.

Results

- **Corpus Statistics**
  - Data was collected across 16 languages with 1865 ellipsis examples



- **Identifying Ellipsis**
  - Trained logistic regression on parts of speech.
  - Asked ChatGPT 3.5 to identify elliptical constructions.



- Regressions perform at 68% on average, which every LLM failed to meet.
- BERT-type transformer models outperform all other models with 94% accuracy

Discussion and Future Work

- Traditional techniques initially outperform LLMs on identifying ellipses
- A benchmark of 65% for an LLM to meet with sufficient training.
- Debate in linguistics research about what qualifies as ellipsis pose a challenge in data collection.
- Gained experience in NLP research
- Future Work
  - Train models to identify the position of elided words
  - Train models to identify the word that is elided

References

[1] van Craenenbroeck, Jeroen, and Tanja Temmerman (eds), The Oxford Handbook of Ellipsis, Oxford Handbooks (2018; online edn, Oxford Academic, 8 Jan. 2019).

[2] “Ellipsis and Elided Elements in Natural Language: The Hoosier Ellipsis Corpus.” NLP Lab, nlp-lab.org/ellipsis/. Accessed 7 Dec. 2023.

Acknowledgements

Thank you to the NLP lab for their advice and support

LUDDY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

Indiana Natural Language Processing Lab
Undergraduate Research – Fall 2023