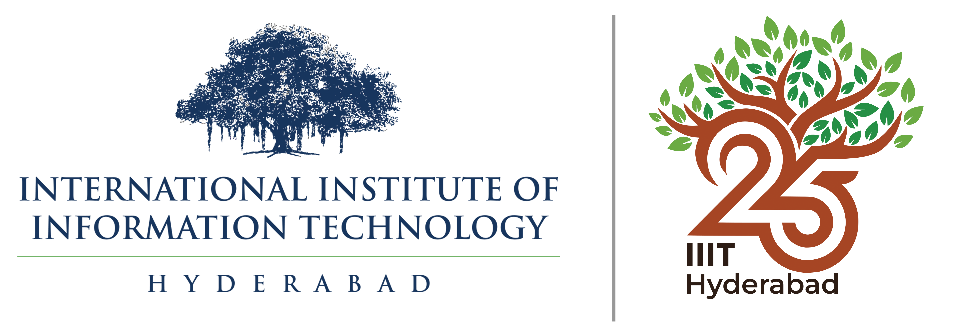
**International Institute of Information Technology, Hyderabad**

**1st Year 2nd Semester**



**Introduction to NLP**

**Interim Submission**

**Team Name:** Semantic Insighters

**Project Mentor:** [Patanjali Bhamidipati](https://courses.iiit.ac.in/user/view.php?id=34461&course=4198)

**Project:** Hypernym Discovery

|  |  |  |
| --- | --- | --- |
| S. No. | Name | Roll No. |
| 1. | Sagnick Bhar | 2023201008 |
| 2. | Soham Ghosh | 2023202011 |
| 3. | Aman Khurana | 2023201017 |

**Problem Statement**

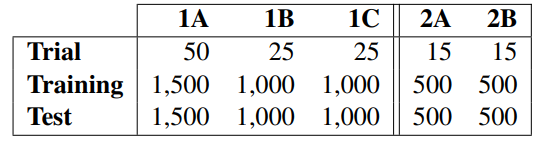
The task of Hypernym Discovery in the paper is defined as finding and extracting suitable hypernyms for a target input term from a large textual corpus. For each input term the expected output is a ranked list of candidate hypernyms (up to 15) drawn from the provided vocabulary.

We are provided with a target term, a source corpus, and a vocabulary, and are required to retrieve a ranked list of candidate hypernyms for the input term. The task is split into two subtasks, general-purpose hypernym discovery and domain-specific hypernym discovery. Each subtask features specific training testing and trial data containing input terms and corresponding gold hypernym lists to evaluate the performance of participating systems.

**About Dataset**

* In this subtask, we consider 3 different languages:
  + English (subtask 1A), with a gold standard of 3,000 labelled terms
  + Italian (subtask 1B) and Spanish (subtask 1C), each with a gold standard of 2,000 labelled terms
* In this subtask we focus on English and consider two different domains of knowledge:
  + Medical (subtask 2A), with a gold standard of 1,000 labelled terms
  + Music (subtask 2B), also with a gold standard of 1,000 labelled terms

All the corpus are divided into train, validation and test splits along with additional vocabulary (upto tri-grams) which contains the subset of primary hypernyms that occurred less than 5-3 times and are over generic in nature. These filtered vocabularies can be used to reduce search space for potential candidates. The hyponyms are divided into two categories- Entity (names, location, etc.) and Concepts (phenomenon, activity, etc.).



From the above table, it is clear that the dataset was divided equally into training and testing sets for each subtask. This division ensures that the hypernym discovery systems are trained on a portion of the data and tested on another portion to evaluate their performance accurately. The trial data, which contains fewer examples compared to the training and testing sets, can also be utilized as a development set. Development sets are often used for fine-tuning models and evaluating performance before final testing.

Sample data of each dataset:

|  |  |  |
| --- | --- | --- |
| **Hyponym** | **Hypernym** | **Corpus** |
| pollution | Atmosphere, windstorm, violent storm, air current, atmospheric state, density current of air, storm damage, atmospheric phenomenon, storm, cyclone, natural phenomenon, tempest wind | English |
| bagpipe | Instrument, musical instrument, pipe, wind, wind instrument, aerophone | Music |
| bone spur | clinical finding, disease | Medical |
| lengüeta | Aleteo, instrumento de viento-madera, instrumento musical | Spanish |
| sesto | Grado, numero ordinale frazione, carica | Italian |

**Our Approach**

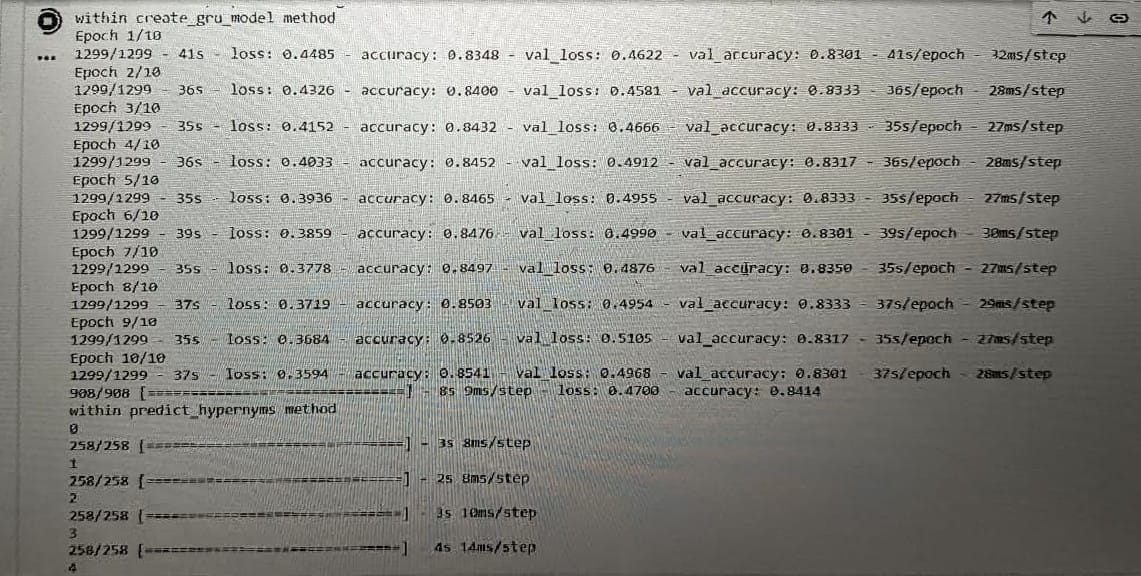
In our base-line strategy, we decided to divide our task into two key subtasks:

1. "Embedding learning" (using Word2vec).
2. Hypernym-hyponym relationship learning.

In the supervised approach, we use neural network models (e.g., GRU and LSTM) to extract the latent representation of hyponym and hypernym embeddings. We then compute the similarity between eq (query's latent representation) and eh (corresponding hypernym's representation) using cosine similarity. The following approaches have been addressed in the literature and will serve as supervised baselines for our progress.

Currently we have generated the embeddings for “1A English data” and “2A medical data” and used the former embeddings for our hypernym prediction task as of yet.

The results are not yet good as this approach needs a lot of improvements which we have stated in the “Future Tasks” section.

**Outputs:**

**Future Tasks**

As per our stated plan, we will currently next start implementing the unsupervised pattern-based approach after which the next immediate task is the implementation of supervised projection-based approach as well as Hybrid approach.

Furthermore, we would conduct a few extra trials in an effort to improve on the findings produced by prior methodologies. We will also investigate sequential pattern mining techniques to automatically extract frequent sequential patterns between hyponym terms and their assigned hypernyms from the corpus. The final evaluation report should compare results from all methodologies and highlight improvements made through subsequent tests. We will also perform all the evaluation tests on each corpus for each method and then present a final comparative analysis of out entire project in the final submission.