INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

Hypernym Discovery

FEBRUARY 2022

COURSE CODE: INTRO TO NLP - CS7.401.S22

Advisor:

Prof. Manish Shrivastava

Team Name:

Natural Barrier

Project Representatives:

Aditya Kumar Singh - 2021701010
Dhruv Srivastava - 2021701021
Nayan Anand - 2021701014

Academic year:

2021-2022

1. Project Overview

Hypernymy, namely “is-a” relation, is a vital lexical-semantic relation in natural languages, which relates general terms to their instances or subtypes. In this relation, we name a specific instance or subtype **hyponym** and its related general term hypernym. For instance, (apple, fruit) is in hypernymy relation, where apple is a hyponym and fruit is one of its hypernyms. Or, for the input word “dog”, some valid hypernyms would be “canine”, “mammal” or “animal”. Due to its general representation ability of semantic relations, hypernymy becomes an essential concept in modern natural-language research, and hypernymy detection becomes a fundamental component in many of the applications, like *Taxonomy construction*, *Semantic search*, *Textual entailment*, and *Question answering*. The **goal** of the **hypernym discovery** is to predict the hypernyms of a query given a large vocabulary of candidate hypernyms. A query can be either a concept (e.g. cocktail or epistemology) or a named entity (e.g. Craig Anderson or City of Whitehorse).

Here our **objective** would be as follows:
Given an input term (or a hyponym **q**), our alogrithm need to retrieves a ranked list (possibly of length 15) of its suitable hypernyms from a large corpus. For training, some hyponyms with their gold hypernym lists are provided.

2. Data Overview:

The competition was divided into 2 major tasks containing 5 subtasks, where first task (i.e., the General-purpose hypernym) consists of **3** subtasks (1A, 1B, and 1C) and the second one (the Domain-Specific Hypernym) have only **2** (2A and 2B).

Further, we’ve been provided with *training*, *trial*, and *test* set each containing both normal data and gold standard data. Further, both normal data and gold data consists of 5 subtasks as mentioned above. The gold standard consists of terms along with their corresponding hypernyms (up to trigrams). For testing (inferencing), systems are provided with terms for which they have to produce a ranked list of their extracted hypernyms. Training and testing data are split evenly (50% training - 50% testing).

Task 1: General Purpose Hypernym Discovery:
For the first subtask we use the 3-billion-word *UMBC corpus*, which is a corpus composed of paragraphs extracted from the web as part of the *Stanford WebBase Project*. This is a very large corpus containing information from different domains. For Italian we use the 2-billion-word *itWac corpus* extracted from different sources of the web, and for Spanish a 1-billion-word *Spanish corpus* which also contains documents from different sources. The dataset had 10000 (term, hypernym) pairs.

Task 2: Domain-Specific Hypernym Discovery:
For the medical domain a combination of abstracts and research papers provided by the MEDLINE (*Medical Literature Analysis and Retrieval System*) repository, which contains academic documents such as scientific publications and paper abstracts, is provided. As regards the music domain, the provided corpus is a concatenation of several music-specific corpora, i.e., music biographies from `Last.fm` contained in ELMD 2.0 (Oramas et al. 2016), the music branch from Wikipedia, and a corpus of album customer reviews from Amazon (Oramas et al. 2017).

3. Implementaion Plan

1. **Supervised Approach**: MLP (Multi-Layer Perceptron) would be our baseline model where our goal would be to extract the embeddings of both *hyponym* and *hypernym*. And based upon the similarity (some sort of distance or metric of embedding space) between hypernym embeddings as

well as between given hyponym and all other hypernym embeddings, a ranked list of candidate hypernyms needed to be drawn out.

2. **Unsupervised Approach**: Here we'll try to cluster the embeddings of hypernym obatined from *BERT* based upon some metric w.r.t to a given query.

3. **CRIM Model**: **C**omputer **R**esearch **I**nstitute of **M**ontreal team has developed this model which exploits a combination of both supervised projection learning and unsupervised pattern-based hypernym discovery. Till date this has proved to be the state-of-the-art model in both tasks (i.e., General-purpose as well as Domain-specific). Plus, we'll try to improvise upon this model so as to have a better *MAP*, *MRR*, and *Precision@N* scores.

References

[1] Gabriel Bernier-Colborne and Caroline Barrière. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 725–731, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[2] CamachoCollados. Semeval-2018 task 9: Hypernym discovery, 2018.

[3] Prakhar Mishra. Automatic extraction of hypernym relations from text using ml.