Fine-grained Entity Typing

Sharvi Tomar
stomar2@illinois.edu

November 12, 2022

1 Overview

This document reports on the fine-grained entity typing system that I developed. Given a list of extracted entity mentions, the system classifies each mention into one or more of the types defined in an ontology. The ontology consists of approximately 200 fine-grained entity types that are representative of news data.

2 Introduction

2.1 Entity Extraction

Entity extraction (named entity recognition) is an information extraction task that seeks to locate and classify named entities mentioned in unstructured text into predefined categories (e.g., person names, organizations, locations). For example, let's consider the following simple sentence:

"Bill Gates founded Microsoft in 1975"

Then an entity extraction system needs to produce the following annotation:

"[Bill Gates]PERSON founded [Microsoft]ORGANIZATION in [1975]DATE"

We encourage you to try out AllenNLP's NER demo to see more examples of the task.

Note that entity extraction has two sub-tasks: the "localization" and "classification" steps. In this assignment, you will need to only focus on the "classification" step (also known as

entity typing). We will give you a list of entity mentions extracted from some documents. And your task is to determine the type of each mention.

For example, let's suppose your system needs to process the simple sentence shown above. Then, your system will be given the mentions "Bill Gates", "Microsoft", and "1975"; and it will need to determine the types of mentions (i.e., PERSON, ORGANIZATION, and DATE).

2.2 Recognizing Ultra Fine-grained EntitieS (RUFES)

Recognizing Ultra Fine-grained EntitieS (RUFES) is a shared task that extends entity extraction to a new fine-grained entity ontology that consists of approximately 200 fine-grained entity types that are representative of news data. Each entity type in the ontology has a one-sentence definition along with some examples. Some sample types can be seen in the table below.

Given an input document, the system is required to automatically identify an entity as a cluster of name, nominal, and/or pronominal mentions, and classify the entity into one or more of the types defined in the ontology.

In fact, the full RUFES task can be divided into 3 sub-tasks: mention extraction, coreference resolution, and entity typing. Again, in this assignment, you will need to only focus on the entity typing subtask.

3 Data

I have used the data provided by RUFES 2020.

- For training, I have 50 annotated documents provided by RUFES 2020.
- For testing, we have 106 annotated documents.

For building an entity-typing system, to the folder data/ltf and the tab file in the folder annotation:

- LTF files consist of tokenized documents. A simple script for reading ltf files is available at: ACE_ERE_Scripts
- The tab file consists of ground-truth annotations.

Since the data provided by RUFES 2020 is not that large, I made use of external data for fine-grained typing Wikipedia downsamples data

A file containing mention extraction results on the test set produced by a model is provided. The file is a tab-separated file. Each row corresponds to one extracted mention:

- The first column is the mention string.
- The second column is the mention justification.
- The third column is the mention type: "NAM" (for name mentions), "NOM" (for nominal mentions), or "PRO" (for pronominal mentions).

Finally, the ontology of RUFES 2020 is available at NIST ontology

4 Methodology

4.1 Data Preparation

Since the RUFES training data was not large with the number of unique entity types in:

RUFES train set: 210

RUFES test set: 225

NIST Ontology: 265

I used the Wikipedia downsampled data to augment the training data. The Wikipedia downsampled data had entity types specified in a different format. I used the ontology list provided by NIST to modify the entity types of Wikipedia downsampled data to make this data suitable for the task by performing the following steps:

1. Removed digits from the entity types
2. If the entity type of observation is from Level3 of the ontology list, I replaced the label with a list[Level1, Level1.Level2, Level1.Level2.Level3].
3. If the entity type of observation is from Level2 of the ontology list, I replaced the label with a list[Level1, Level1.Level2].
4. I removed the observations with entity types not present in the ontology list.

   The python notebook could be found here Jupyter Notebook

4.2 Model Training

I used an LSTM-based model for entity typing by treating it as a multi-label classification task.

The training data I supplied to the model was RUFES train data along with the modified Wikipedia-downsampled data. I ran the model for about 50 epochs as the model seemed to have stabilized in the batch-wise Precision, Recall and F1-score metrics.

The testing data was the set of sentences for the mention types provided in the mentions.tab.

4.3 Results Discussion

The evaluation for the task was performed by using the RUFES Scorer