CMPUT 497 Assignment 4 Report

Yonael Bekele
University of Alberta
Edmonton, Canada
yonael@ualberta.ca

Michael Lin
University of Alberta
Edmonton, Canada
michael.lin@ualberta.ca

OVERVIEW

We use *spaCy* to perform all the tokenization, POS tagging, and dependency parsing. Extraction and substitution of entities, subjects, and objects are done with various regular expressions. **Please note, English is not our first language, mistakes might be made when manually inspecting the samples.**

TASK 1

We extracted the relation sentences from the files and entered them into a dictionary where the key was the relation (file) name. We then cleaned the sentence while keeping track of the entity specified by the freebase token. We then ran tests on a random sample of 100 (per relation). By running the spaCy POS tagger, we were able to identify when entities were not being tagged as nouns. The output format is the sentence, followed by each word with their tag and the tags with "Incorrect" beside them are the misidentified entities (each on one line).

Stats

# of Filtered Sentences	% of misidentified entities per sentence
7	0.07%

Summary of 100 samples from *award*

# of Filtered Sentences	% of misidentified entities per sentence
11	0.11%

Summary of 100 samples from *business*

# of Filtered Sentences	% of misidentified entities per sentence
7	0.07%

Summary of 100 samples from *film*

# of Filtered Sentences	% of misidentified entities per sentence
11	0.14%

Summary of 100 samples from *music*

# of Filtered Sentences	% of misidentified entities per sentence
3	0.03%

Summary of 100 samples from *people*

Summary

The filtered sentences usually only had 1 misidentified entity in each for the relations *award, business, film* and *people*. For the relation *music,* we found that one of the sentences had two misidentified entities. The sentence misidentified "American" and "Canadian" as entities when they are adjectives (JJ).

Pattern

By examining the misclassifications, we find that many errors are adjectives (JJ) that are used to describe where a person is from or even a language of a region. Some of the errors also come from the spaCy POS tagger not recognizing named entities. For example, in the *business* relation, spaCy's POS tagger mislabelled "Zopa", the name of a business, as a determiner. Also, in the *business* relation, any of the descriptors of the regions where businesses operate were mistagged as entities, when they are adjectives. The spaCy tagger does pretty well with recognizing names as seen in the *people* relation, the only mistakes having been adjectives which follows our established pattern. The same holds true for the *film* relation, where describing the actors or films nation, language or roots has been misidentified as an entity. The most unique set of errors was in *music* where we found that in this case, the spaCy POS tagger was unable to properly identify the name of music as a proper noun.

TASK 2

We extract 100 random samples per relation to compile the following statistic and summary. We filter out a list of the lowest common ancestor (LCA) that are verbs and use the lemmatizer from *spaCy* to get the verb stem, then count their occurrence. Then for each relation, we will give one example of verb that fails to represent the relationship between subject and object.

Stats

Correct	Avg Number of different Verb Stems
64	0.156

Summary of 100 samples from *award*

Correct	Avg Number of different Verb Stems
31	0.806

Summary of 100 samples from *business*

Correct	Avg Number of different Verb Stems
35	0.429

Summary of 100 samples from *film*

Correct	Avg Number of different Verb Stems
20	0.83

Summary of 100 samples from *music*

Correct	Avg Number of different Verb Stems
26	0.692

Summary of 100 samples from *people*

Summary

In relation, *award*, out of 100 samples, there are 67 sentences that object and subject are connected by verbs. The word, *awarded* occurs 35 times, following by *received* which occurs 12 times. The majority of connected verbs are in the form of past tense. We found one verb, *struck*, out of 100 samples that fails to represent the relationship, from the sentence below.

To his credit, Mr OBJECT struck a mildly hangdog look as the SUBJECT was placed around his neck

In relation, *business*, there are 41 sentences that object and subject are connected by verbs. The verb *provides*, occurs 3 times, followed by *used* which occurs 2 times. The most common LCA that makes sense is present tense. The verb, *operating*, did not express the relationship.

The SUBJECT has been operating in ENTITY1, ENTITY2 since 1986, and has recently expanded into OBJECT

In relation, *film*, there are 41 sentences that object and subject are connected by verbs. The verb, *plays*, occurs 12 times. *Played* and *appeared* occurs 4 and 2 times. Most common LCAs are present tense. We found the verb, *rebounded*, fail to represent the relation between subject and object.

SUBJECT rebounded his acting career with the portrayal of cab-driver-turned-baby-sitter OBJECT in ENTITY1 's low-budget comedy ENTITY2.

In relation, *music*, there are 21 sentences that object and subject are connected by verbs. The most common verb is *performs*, occurs 3 times. Most common verb LCAs are present tense. The verb, *called*, did not express the relationship between object and subject.

Did Mariah record a song for ENTITY1 called OBJECT with SUBJECT

In relation, *people*, there are 30 sentences that object and subject are connected by verbs. The verb, *born*, occurs 7 times, followed by *managed* which occurs 2 times. Most common LCAs are past tense. The verb, *managed*, failed to represent the relationship.

While SUBJECT was recovering in ENTITY1 after his first stroke, his sons OBJECT and ENTITY2 managed the business affairs

The relations mediated by “is” as the LCA performed very well to express the relation, and appeared frequently in the data. The statistics evaluating the occurrence of “is” as LCA and its expression of the relation are as follows: *awards* 5 out of 5, *business* 11 out of 12, *film* 6 out of 8, *music* 12 out of 13, and *people* 5 out of 6.

TEAM COLLABORATION

Discussion of spaCy with group Delaney & Daniela and Helen & Flora.

REFERENCES

[1] spaCy documentation