CMPUT 501 Assignment 4

Shehroze Khan and **Afia Anjum**
Department of Computing Science
University of Alberta

1 POS Tagging to detect mislabelled entities

In order to identify the mislabelled entities of each of the sentences from the given json files, we use the default nltk parts of speech (pos) tagger. A program is written using regular expression that would take each of the entities tagged with an entity identifier in a sentence and store them. We also include few steps in the program where we pre-process each of the original sentences, tag a parts of speech to their every words and store the results. We utilize this stored information and the stored entities computed in the early step, that are tagged with the original sentence, in order to find mislabelled entities that are present in that sentence. Here we are having a simple assumption that, if the first entry of an entity gets tagged with any parts of speech other than the following such as, NN(Noun, singular or mass), NNS (Noun, plural), NNP (Proper noun, singular), NNPS (Proper noun, plural), we consider the entity as misidentified entities and the sentence to which this entity belongs as a problematic sentence for a given relation.
In order to produce output, 100 random sentences are selected from each of the given relation(.json) files. The output files are generated having a column of original sentences (without modification or pre-processing), a column containing parts of speech of all the words for that sentence each in one line and a final column containing all the mislabelled entities (if any) for that sentence.

2 Task 1 Findings from the output

Here is a table showing the number of sentences in which filtration is needed with respect to each of the given relation files and the average number of misidentified entities per sentences. This result is generated by selecting 100 sentences randomly from each of the relations(5 relations) and in total 500 sentences over all the relations.

Relation	No. of filt.sent.	Avg. no of mis.ent
awards	5	0.06
business	12	0.15
film	8	0.09
music	29	0.45
people	5	0.05

In the award.json relation, an entity such as ”To Kill A Mockingbird” with the pos tag (TO VB DT NN) refers to the name of a book. Therefore, this type of entity can be disregarded from filtering out.

In the business.json relation, entities such as ”Major Motion Picture”, pos tagged with (JJ NNP NNP) , ”Rising Sun Pictures” tagged (VBG NNP NNPS) , ”British Nuclear Fuels Limited” tagged with (JJ NNP NNP NNP) , are infact names of different organizations or company. Therefore, these type of entities can be disregarded from filtering out. However, entities like ”Publishing Company” having pos tag as (VBG NN), is a bad training example and must be filtered out.

In the music.json relation, the number of filtered sentences is very high. Having a manual inspection on the output file, it can be said that most of the entities tagged in this particular file includes the name of a ’song’, which does not necessarily have to be a Noun or Noun Phrase every time. For example, ”Throw It All Away” is an entity tagged in the original sentence and has the POS tags as (VB PRP DT RB). Therefore, it is not always reasonable to consider such sentences as problematic sentences for this relation. Other examples of entities with similar observations are ”Zero 7” or ”Let It Shine”, where both of them are names of a song.

In the film.json relation, entities such as ”Green Day”, pos tagged with (JJ NNP), refers to the name of a band or ”In The Heat Of The Night”, pos tagged with (IN DT NNP IN DT NNP) , refers to the name of a play. Therefore, these types of

entities can be disregarded from filtering out.

In the people.json relation, entities like "American Director" or "Austro-Hungarian premier" have the POS tags as (JJ NN). Since it can also be expressed as a Noun Phrase, therefore these type of entities can be disregarded from filtering out. However, we should remove any entities which contains only Adjectives(JJ) in it. (e.g. "English", "British", "Swiss").

Most of the problematic sentences have some of the words in upper case letters, even when uppercase is not a correct way to represent the word. This is a serious issue that might confuse the parser. Besides, majority of the entities which are misidentified as entities, have the POS tags Adjective (JJ), Adverb(RB) or different form of Verbs (VB, VBD, VBG, VBN) associated with it where, Adjective(JJ) POS tag is the most common among them.

3 Task 2: Filtering out spurious relational sentences

The table below details the statistics for 100 randomly sampled sentences from each of the five relations. The second column of the table depicts the number of those sentences that correctly identify the relation between the subject and the object. The third column is the average of different verb stems in these sentences.

Relation	Sentences	Avg. Verb Stems
awards	77	0.169
business	49	0.694
film	65	0.354
music	53	0.509
people	36	0.361

The number of verbs are particularly low in the business, music and people relations. This is because we tend to see subject-object relationships expressed through nouns. For example, in the music relation, nouns such as *album*, *offering*, *song*, *mixtape*, and *record* form the lower common ancestor between the subject and object in determining the desired relationship. Similarly, for the people relation, we have relationships expressed through nouns such as *mother*, *brother*, *father*, *son*, and *wife*, which explicitly express the relationship between proper noun subject and object.

The python code provided for this task prints to the console the most common lowest common ancestors seen in the sampled sentences for the five relations. Combining this printed output with our manual analysis of the five output text files (also provided in the submission), we enlist the most frequently used verb stems in the following table:

Relation	Most Frequent Verb Stems
awards	*award; receive*
business	*is*
film	*play; is*
music	*is; feature*
people	*born; is*

Except for the stem *is*, these verb stems in all relations more commonly used in the past tense. Though the verb *is* depicts a direct relationship between subject and object in all five relations, we do have certain cases where this is not true. For instance, for the business relation we have the following sentence:

According to OBJECT , SUBJECT 's ENTITY1 is the newest to add something to lobby charm.

The lowest common ancestor connecting subject and object is *is*. However, this verb does not establish the desired business and business relationship between these two entities.

We also come across verbs that do not express the desired relation between subject and object. For example, consider the following sentence for the people relation:

When ENTITY1 attacked OBJECT , SUBJECT listened

Here the lowest common ancestor connecting subject and object is the verb *listened*. However, it does not signify any people/person/children relationship between these two entities.

References

Online parts of speech tagger
Penn Treebank POS