CMPUT 497: Assignment 4 – Task 1

Shouyang Zhou
University of Alberta
Edmonton, Alberta, Canada
shouyang@ualberta.ca

1 Methodology

My method is to suppose each entity marked in a sample is a noun phrase and to decipher the POS tags associated to that noun phrase to decide as to whether it is correct or not. Naturally, noun phrases must contain at least one noun, if none of the POS tags assigned to the phrase are nouns then it is likely that the entity is mislabeled (from the prospective of the POS tagger). My script examines the POS tags associated to each entity phrase and checks if there is overlap between the noun POS tag variants and that of the entity. ("NN", "NNS", "NNP", "NNPS"). This rule may be too general, as such an alterative reduced version will also be considered.

My script uses the spaCy pretrained model "en_core_web_sm" model based on the OntoNotes dataset to assign POS tags. I choose this model as it was a recommended multipurpose model. It is suitable for this assignment's data since the both the training material and the assignment material are both web sourced data.

My script generates an object per JSON-sentence section. Upon initialization, each object extracts entities from the sentence, creates an entity markup removed version of each sentence, and uses spaCy to assign POS tags to the preprocessed sentence. After, each object examines the POS tags associated with each entity phrase for nouns.

Please note, my script's output processes the entire relation file. I sample 100 sentences per relation after from the output file.

2 Filter Analysis

The first 100 sentences were examined for each relation. I sample the first 100 sentences as they appear to be from an already randomized source. The summary results are as follows. Extrapolating from the following, artistic subjects such as film and music have far higher numbers of possible misidentified entities by POS tagging. On the other hand, awards appear to have been adequately recognized.

Intuitively, certain topics such as awards and business divisions / operations must convey their subject matter directly, hence they may easier to recognize. Artistic matters do not require this level of direct communication and may often choose an expressive albeit indirect name.

	Sentences.	# of Entities.	Average.
Award	0	0	0
Business	4	5	1.25
Film	11	11	1
Music	11	11	1
People	8	8	1

Common suspected errors are predominately, single word adjectives often of a nationality. Across categories, nationality adjectives are often identified as a separate entity. Similarly, the religious adjectives, "biblical" and "Raelian" follow suit.

In music and film relations, verbs and adverb phrases are candidate mislabeled entities. Some verb entities are "Scream", "Scream 2", and "Shining". Some phrases are "Live at Last", "Bite Down Hard", and "Up To Here". The phrases are tagged to either start or end in an adverb and are likely to contain a conjunction (IN tag).

The mislabeling detection in music and film may over-represent the true number of mislabeling. Examining these phrases and verbs, when considering the subject matter and sentence details, they suggest to me that they are genuine film or album/song names/entities. However, given the contextless or common usage of the words/phrases alone, they mis-tagging of POS tagger is excusable.

3 Alternative Filter Analysis

I also considered using only the tags "NNP" and "NNPS" as candidate anchor tags for noun phrases. This filter is more specific than what the assignment asks for (since the selection criterion is pronouns rather than nouns). This filter generates more true positives (non-entities) at the cost of significantly more false positives.

	Sentences.	# of Entities.	Average.
Award	3	3	1
Business	7	8	1.14
Film	16	17	1.06
Music	26	26	1
People	12	12	1

This method was beneficial in the award and business relations since it pruned generic terms such as "blue room", 'Groceries", "Flim", "Newpaper", which to me doesn't point towards a specific entity.

However, this was significantly disruptive in the latter relations due to the nature of the expressive names of characters, films, and people. For example, "Yuma" and "Wolverine" are clearly distinct entities but they are (rightly) tagged to be singular nouns. Likewise, in music, the phrase naming of some songs is an issue, for example "His Name Is Alive".

4 Conclusion

Using POS tagging and then supposing each entity is a noun phrase can be a convenient albeit conservative form of entity validation. Using general rules, this method identifies around 5%-10% of the sampled datasets to contain spurious entities, a non-trivial amount.

This method is dependent on the POS tagger used, its training data, and the nature of the entities the user wishes to validate. Some entities are harder to validate than others such as names of artistic creations such as film and music via this method.

Using a more specific set of POS tag, it may be possible to increase the accuracy of this method. For example, filtering out abstract nouns vs concrete nouns may be helpful.