CMPUT 497: Assignment 4 – Task 2

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

1 Methodology

My script generates an object per JSON-sentence section. It processes each entire file. Upon initialization, each object preprocesses each sentence replacing entities with their entity placeholder tokens. After, a spaCy document is generated, and the dependency trees of the subject and object are extracted to find the lowest common ancestor.

My script uses the spaCy pretrained model “en_core_web_sm” model based on the OntoNotes to generate the dependency tree. This model is trained on web data hence it should be suitable for this dataset.

Please note, my script processes the entire relation file. My analysis is done from sampling the output files per relation.

2 Filter Analysis

The following summary statistics were generated via sampling the first 100 sentences per relation from the output files. The relations in the files appear to be largely random, barring some pairs of relations which were extracted from a text. As such, I see no major reason to implement a more sophisticated sampling scheme. This analysis was done in Excel, see the file “task_2.xlsx”. I will describe my methodology in judging verb mediated relations then analyze the relations separately.

	Verb supported Relations.	Distinct Verb-Stems.	Stems. Per Sent.
Award	73 (82)	9 (15)	0.12 (0.18)
Business	35 (57)	17 (31)	0.48 (0.54)
Film	40 (65)	11 (15)	0.28 (0.23)
Music	28 (41)	15 (17)	0.53 (0.41)
People	20 (42)	6 (10)	0.30 (0.24)

* Non-verbal relations in brackets.

Judgement Methodology

I consider the lowest common ancestor of the dependency tree of the subject and object as generated by spaCy. Note that this does not have to be a verb, which in that case I will consider it as a non—verbally mediated relation. For verbs such as “is” and “of”, I will consider the path of the subject and object to the verb to validate the relation. Some relations may chain the subject to the object or vice versa, in these cases, I will consider these as non-verbal relations. In these, I will try to consider the next lowest common ancestor to determine a non-verbal relation. For example, a dependency path to the root may be “Subject -> Object -> launched” representing the phrase “ … ENTITY 1 launched SUBJECT’s OBJECT today …”, since the verb launched is actually mediating the phrase, I aim to consider that instead and the text as a whole.

Award Winners Relation

This relation had the highest occurrence of relations mediated by a verb in the relations surveyed. Most frequently, relations were mediated with the words “awarded”, “received”, “won”. The verbs in this relation were predominately past tense and concentrated amongst the three verbs. The verbs were used in a mixture of passive and active phrases, generally active phrases were more likely to contain ambiguously attributed relations. For instance, a phrase like “The president awarded John the medal of honor”, is more likely to demonstrate misattribution (the president is marked to be the recipient) than “John was awarded the medal of honor by the president”. I will consider these as verb supported relations in the spirit of the assignment (to consider verbal relations).

The verb “is” is generally (count: 8) a correct mediator of this relation. This occurs when the award is a sort of title. For example, “{person} is Playboy Playmate of the month”.

Business Operating Industry Relation

This relation had a moderate number of verb mediated relations. This relation had a diverse set of verbs and non-verbs mediating relations, outside of the verbs “is” and “was”, many verbs were only used once. Generally, the verbal mediators convey some notion of participation. The verb “is” is used to mediate 17 of the relations. This relation demonstrates a mixture of active, present tense, and passive past tense verbs. Present tense is more common than that of the awards relation. It also demonstrates a significant amount of relations mediated via nonverbal lowest common ancestors.

The business news article phrasing of some samples convey relations with terms such as “coordinates”, “provides”, and “practices”. Likewise, when articles refer to a historic firm, they often use a passive past tense form verb such as “provided”, “launched”, “added”.

The verb ‘is” often directly used to mediate the subject to the object or is used to mediate a phrase noun to connect the two. For example, “Air Zimbabwe is an airline” or “MKC Networks is the supplier of Voice over IP for …”.

A significant amount of nonverbal relations exists in the sample. Non-verbally many relations were direct where the object or subject was the lowest common ancestor. Also, concise introductions of a business firm often take the form of “OBJECT makers SUBJECT, …” using the words “makers”, “manufacturers”, “publishers” etc.

Spurious relations often noted some action taken by a business with respect to another industry. Verbs such as “sponsored”, “announced”, “reached”, and “subcontracted” are examples of this. Typically, these denote some past tense action that a firm had undertaken.

Actors To Performances Relation

This relation was generally characterized by relatively few distinct mediating terms actors to roles. The most common forms were active present tense relations mediated by the word “plays” (count: 20) or “stars’ (count: 3) and active past tense versions “played” (count: 5). Most relations expressed used some variant of the word “play” and “star”.

The verb “is” (count: 5) can mediate relations between actors to roles often as an expression of their performance, for example “SHU QI is effective as SCARFACE”. Sometimes this relation is direct such as “OBJECT is SUBJECT in ENTITY …” or “SUBJECT is the voice of OBJECT in ENTITY …”.

Non-relation expressing verbs appear to be from reviewers or people expressing their reaction to an actor’s performance. Some of these indirectly express the actor to a role but do so in such a way that the LAC word and path do not directly convey the relation. For example “SUBJECT does indeed sizzle in this role” or “From this point on, Owen can understand OBJECT , who has the voice of SUBJECT”. These generate a significant amount of relations where either the object or the subject are the lowest common ancestor (count: 22).

Spurious relations appear to be speculative comments on actors to roles. For example, some relations are mediated by the words “amusing”, “ideal”, “brainchild”, “makes”, and “tempted”. The phrasing of these sentences appears to be commentary on which actor to role pairs may be preferable.

Music Artists To Album Relation

This relation is expressed as a mixture of active present and past tense verbs. The most common verb was forms of the words “released”, “recorded”, “produced’, and “performed”.

The verb “is” at times mediates the artist to either the album via the word “album” or indirectly via some description of the album. For example, ” Kammel Kalamak SUBJECT 's latest album, OBJECT is the long anticipated follow up to ENTITY1” and “SUBJECT 's brief OBJECT is a powerful psychological experience”.

In non-verbal relations, often the subject or the object were the lowest common ancestor. Often these were mediated by a higher verb or a direct relation such as “SUBECT’s OBJECT is …”

Some connective words are “released”, ‘by”, “sings”, and “from”.

Spurious verbal relations appear to be news segments expression the actions of musicians or reactions to albums. For instance “That same year, ENTITY1 began recording with SUBJECT , and after becoming a regular on their OBJECT”.

Persons To Children Relation
Relations in terms of persons were the least likely to be expressed as verbs. Being generous in my consideration, the verbs “mother”, “father”, and “is” consisted of 9 of the 20 verb supported relations. Considering the actual use of the word’s “mother” and “father”, they are used more as noun that verbs.

Verbally and non-verbally substantiated relations are often in the form of active present tense phrases such as “OBJECT 's mother SUBJECT was the guiding force in OBJECT 's life”. Generally, these phrases predominately use contractions to like the subject and object to the LAC term. In this category are noun mediated relations often in the form of “X’s son/daughter” or “son/daughter of X”.

Non-verbal relations were common in this dataset where many relations were expressed via the words “son” and “daughter’. Also, the word “born” is frequently used. Mediations using “born”, if considered, must be treated like the word “is”, where the path linking subject and object to “born” is of more importance than the word itself.

The verb ‘is” at times mediates a relation in the form of “**X** of ” where **X** is in {son, daughter, father, mother}.

Spurious verb mediated relations appear to be past tense accounts of other familial relations. Words such as “married”, “emigrated”, “wife”, ‘grew” are exemplary mediators.

3 Conclusion

In this task, we see that differing relations are characterized differently by the term used to mediate relations. Some relations are expressed heavily in terms of one style (active vs passive, past tense vs present tense) and in terms of verbal and non-verbally mediated relations. Furthermore, the use of “is” and similar generic terms are a common albeit indirect mediation method. Such terms require careful consideration of the path to term between subject and object. The relations here also demonstrated significant differences in the number of distinct terms (verbal and non-verbal) used to mediate relations across genres / domains.

This task also demonstrates the data quality and model quality difficulties when working on real world applications of natural language processing. The dataset contains some samples which were not in English, contain misspellings, odd punctuation, or other inconsistencies. The dependency trees generated via spaCy is also a confounding source of error.