Assignment Report : Distant Supervision

Arnob Mallik and **Arif Hasnat**
Department of Computing Science
University of Alberta

1 Task 1

For task 1, our program considers any entity with at least one non-Noun POS tag (Nouns-only rule) as potentially misidentified entity. It takes sample of 100 sentences randomly. From these samples, the program identifies any sentence that have at least one potentially misidentified entity as a potentially incorrectly-tagged sentence and outputs it for further investigation.

1.1 Findings

Following table shows a summary of the output by our program for each relations using the sample of 100 sentences:

Relation	# of filtered sentences	# of misidentified entities	Avg # of misidentified entities per sentence
Award.. Winner	90	94	1.04
Business.. Industry	25	30	1.2
Actor.. Character	21	24	1.14
Artist.. Album	63	84	1.33
Person.. Children	24	33	1.38

The average number of misidentified entities per sentence is calculated as the number of misidentified entities divided by the number of filtered sentences.

Using the output of our program, we manually analyzed each sentence to identify actually incorrectly-tagged sentences. Our findings are:

- We found that most of the entities with at least one noun POS tag should be correctly tagged entities. These entities are most of the time award names, artist band names, album names, business names or film names etc. For example, "Presidential Medal of Freedom", "Victor Company of Japan", "Quantum of Solace" etc.

- In some cases, combination of POS tags with at least one Noun represents adjective or adverb. For example, "This week" is not a correctly tagged entity.

- Some of the entities are given POS tags other than Noun by our POS tagger. But, we think they are correctly tagged entities. For example, "Bond" (VB) as film, "Milly" (RB) as Person.

- Most of the truly incorrectly-tagged entities we found are given Adjective (JJ) tags by our POS tagger. Examples are "Canadian", "Swedish", "American", "French" etc. The relation "Actor..Character" contains these adjectives most that's why the number of actually incorrectly tagged entities is highest for this relation.

- Very few of the incorrectly tagged entities are of other POS tags. For example, "Ai" is tagged as entity but it is tagged as Verb (VBP) by our tagger. Actually, "Ai" is part of an entity "I Ai n?t Got You" which is a music album. So, the whole album name should be tagged as entity. Another example, "Shining" is tagged as Verb (VBG) by our tagger and it should not be an entity. Both of these are from "Artist..Album" relation.

Following table shows the statistics of actually misidentified entities and number of adjective (JJ) POS tags for those entities per relation after our manual inspection:

Relation	# of actually filtered sentences	# of actually misidentified entities	# of Adjectives (JJ)
Award.. Winner	2	2	2
Business.. Industry	3	3	2
Actor.. Character	11	11	11
Artist.. Album	8	9	6
Person.. Children	4	4	4

1.1.1 Summary of Findings

In summary, our findings about **identifiable pattern related to filtered sentences** are:

- Our program filters entities with combination of Noun, Preposition and Determinant POS tags most. Such as "Medal of Freedom".
- Sometimes, a part of an actual entity is incorrectly tagged as an entity instead of the whole entity. Such as "Ai" in "I Ai n?t Got You" music album.
- For some cases, a group of words consists of Noun and also other POS tags and represents adjective or adverb as a whole. But it is misidentified as entity. Such as "This Week".
- Most of the cases, an adjective that is used to express nationality or information about country of an entity is tagged incorrectly as an entity. Such as "American", "French" etc.

For **most common POS tags associated with misidentified entities**, our findings are:

- Most common POS tag associated with truly misidentified entities is Adjective (JJ).
- For a very few cases, the POS tag is Verb (VBP or VBG).

2 Task 2

For Task 2, we sampled 100 sentences from each of the 5 relations and inspected their output. The following table shows for each relation, the number of sentences where the LCA of OBJECT and SUBJECT is a verb, the number of sentences which actually express the relation through a verb and the number of different verb stems in those correctly expressed sentences.

Relation	# sentences where LCA is verb	# correctly expressed sentences through verbs	# different verb stems in correctly expressed sentenes
Award..Winner	91	63	10
Business ..Industry	63	34	14
Actor..Character	70	44	9
Artist..Album	57	33	10
Person..Children	54	27	6

The following table shows the average number of different verb stem in correctly expressed sentences :

Relation	avg # different verb stems in correctly expressed sentences
Award..Winner	0.159
Business ..Industry	0.412
Actor..Character	0.205
Artist..Album	0.303
Person..Children	0.222

We calculate the average number of different verb stems by diving the number of distinct verb stems by the number of total verb stems in the correctly expressed sentences.

2.1 Award..Award_Winner Relation

2.1.1 Suitable Verbs

In case of this relation, we found 63 sentences (out of 100) that correctly express the relation through a verb and among these sentences we found 10 different verb stems. The following table shows the details of these verb stems :

Verb Stems	Verb Forms (Counts)
award	awarded (26)
receive	received(11), receives(2)
win	won (8), wins (1)
present	presented (7)
share	share(1), shared(1)
give	gave(1), given(1)
earn	earned (1)
bestow	bestowed (1)
honor	honored (1)
be	is (1)

From this table, we can see that the most common verb stems are *"award"*, *"receive"*, *"win"* and *"present"* which are mostly used in their past form. This is because we found patterns like the following quite frequently :

- "OBJECT was awarded the SUBJECT."
- "OBJECT received the SUBJECT."
- "OBJECT won the SUBJECT."
- "OBJECT was presented with the SUBJECT."

Hence, the verbs *"awarded"*, *"received"*, *"won"* and *"presented"* can be deemed as **good choices** for expressing the Award..Award_Winner relation.

2.1.2 Error Cases

In the following cases, the Award..Award_Winner relation was not expressed correctly by a sentence :

- The most common erroneous scenario that we found is that the person giving the award is often wrongly tagged as the OBJECT. Such as,
  - "OBJECT (George HW Bush) awarded ENTITY1 (Brinkley) the SUBJECT (Presidential Medal of Freedom) , the nation's highest civilian honor".
- There are some cases where the verb does not express the Award..Award_Winner relation. Such as,
  - "OBJECT (Kennedy) **inaugurated** the SUBJECT (Presidential Medal of Freedom) for musicians". Here the verb *"inaugurated"* expresses the relation between the award and someone who inaugurated the award, not the award winner.
  - "Two-time SUBJECT author OBJECT **makes** history come alive in this accessible story of ENTITY1 's birthday."

  Hence, the verbs *"inaugurated"* and *"makes"* can be deemed as **bad choices** for the given relation.
- The verb *"is"* often shows up as the lowest common ancestor of SUBJECT and OBJECT, but does not express the Award..Award_Winner relation. Such as,
  - "Mr OBJECT **is** very popular in the ENTITY1 and was awarded the SUBJECT". Here, the relation is best expressed through the verb *awarded* but the LCA of SUBJECT and OBJECT is the verb *is*.
  - "OBJECT **is** the third president of ENTITY1 to be honored with the SUBJECT". Here, the relation is best expressed through the verb *honored* but the LCA of SUBJECT and OBJECT is the verb *is*.

.

2.2 Business..Industry Relation

2.2.1 Suitable Verbs

In case of this relation, we found 34 sentences (out of 100) that correctly express the relation through a verb and among these sentences we found 14 different verb stems. The most important ones are shown in the following table :

Verb Stems	Verb Forms (Counts)
be	is (15), was(2), are(1), been(1)
provide	provide (1), provides(1)
offer	offers(1)
deliver	delivers(1)
involve	involved(1)

Among these verbs, only the verb *"is"* has been frequently used to mediate the Business..Industry relation. For example,

- "SUBJECT **is** a global OBJECT leader."
- "SUBJECT **is** a OBJECT solution."
- "SUBJECT **is** the marketplace for OBJECT."

Hence, the verb *"is"* can be a **good choice** to express this relation.

Other verb stems that express the Business..Industry relation well include *"provide"*, *"deliver"*, *"involve"*, *"offer"*. Such as,

- "SUBJECT also **provides** OBJECT services."
- "SUBJECT got **involved** in OBJECT traffic in 1998."
- "SUBJECT **offers** superior OBJECT."
- "SUBJECT **delivers** on OBJECT 's Promise of ROI with New Solution."

Hence, the present forms of the verb stems *"provide"*, *"involve"*, *"offer"* and the past form of the verb stem *"involve"* can also be deemed as **good choices** for this relation.

2.2.2 Error Cases

There are some verbs in the dataset that appear as LCA's of the SUBJECT and OBJECT, but do not express the Businees..Industry relation. Such as,

- "OBJECT publisher SUBJECT **says** it will lay off 500 workers."
- "OBJECT firm SUBJECT has **named** ENTITY1."
- "SUBJECT was **followed** by ENTITY1 , which recently released a downloadable OBJECT version of its popular software."

Hence, the verbs *"says"*, *"named"* and *"followed"* can be considered bad choices for this relation.

2.3 Actor..Character Relation

2.3.1 Suitable Verbs

In case of this relation, we found 44 sentences (out of 100) that correctly express the relation through a verb and among these sentences we found 9 different verb stems. The following table shows the details of these verb stems :

Verb Stems	Verb Forms (Counts)
play	plays (14), played(6)
star	star (4)
portray	portrays (2), portrayed (1)
be	is (6), was(2)
return	returns(3)
give	given (2), gives(1)
voice	voiced (1)
make	makes (1)
channel	channeled (1)

From this table, we can see that the verb *"play"* occurred frequently in both past and present forms. As the verbs *"play"*, *"star"* and *"portray"* are synonymous, we can safely say that both present and past forms of these three verb stems are **good choices** with respect to the Actor..Character relation. We found the following examples in our dataset :

- "SUBJECT played OBJECT on the long-running sitcom ENTITY1."
- "SUBJECT portrays the character of OBJECT in ENTITY1."
- "SUBJECT portrayed young OBJECT in the original ENTITY1 (1980)."
- "SUBJECT stars as OBJECT."

Also, the present and past form of the verb stem *"be"* was used quite regularly to express the relation. Such as,

- "SUBJECT is hilarious as OBJECT."
- "SUBJECT is also amusing as OBJECT."
- "I believe SUBJECT was the ultimate OBJECT"

Hence, the verbs *"is"* and *"was"* can also be considered as **good choices** for the Actor..Character relation.

2.3.2 Error Cases

There are some cases, where the verb which is the LCA of the SUBJECT and OBJECT, does not express the Actor..Character relation. Such as:

- "OBJECT played by SUBJECT **appears** in ENTITY1 and ENTITY2." Here, the relation should be expressed by the verb *played*, but the LCA is *appears*.

- "I also **loved** SUBJECT as ENTITY1 yuppie OBJECT."
- "Meanwhile, ENTITY1 **has** a scoop that states SUBJECT will play The OBJECT."
- "As a matter of fact, ENTITY1 **recommended** SUBJECT for the role of OBJECT"

Hence, these verbs can be considered **bad choices** for this relation.

2.4 Artist..Album Relation

2.4.1 Suitable Verbs

In case of this relation, we found 33 sentences (out of 100) that correctly express the relation through a verb and among these sentences we found 10 different verb stems. The following table shows the details of the most frequent verb stems :

Verb Stems	Verb Forms (Counts)
release	released (7), release(3), releasing(1), releases(1)
be	is (9), was(3)

In our samples, we found that the verb *"release"* was frequently used in different forms to mediate the Artist..Album relation. Such as,

- "In 1995, SUBJECT **released** OBJECT."
- "SUBJECT will be **releasing** a 25th anniversary collection."
- "In 1995, SUBJECT **releases** the classic OBJECT."

Hence, the present, past and present participle forms of the verb *"release"* can be considered **good choices** for this relation.

Also, the present and past forms of the verb stem *"be"* was used quite often to express the Artis..Album relation. For example,

- "OBJECT **is** a folk/rock album by SUBJECT."
- "OBJECT **was** SUBJECT 's last studio release."

Hence, the verbs *"is"* and *"was"* are also **good choices** for this relation.

2.4.2 Error Cases

There are some cases, where the verb which is the LCA of the SUBJECT and OBJECT, does not express the Actor..Character relation. Such as:

- "ENTITY1 has **joined** SUBJECT on two previous albums, ENTITY2 and OBJECT."
- "SUBJECT 's album OBJECT was **awarded** a ENTITY1 for Best Latin/Tropical Performance."
- "SUBJECT then **switched** to ENTITY1 and released OBJECT , named for a track on the album."
- "When I **purchased** SUBJECT 's newest record OBJECT , these memories came back in deluge."

Hence, the verbs *"joined"*, *"awarded"*, *"switched"* and *"purchased"* are **bad choices** for the Arist..Album relation.

2.5 Person..Children Relation

2.5.1 Suitable Verbs

In case of this relation, we found 27 sentences (out of 100) that correctly express the relation through a verb and among these sentences we found 6 different verb stems. The following table shows the details of these verb stems :

Verb Stems	Verb Forms (Counts)
be	was(16), is(3)
share	shares(2)
succeed	succeeded(2)
give	gave(2)
pop	popped(1)
born	born(1)

The verbs *"was"* and *"is"* appeared frequently and expressed the Person..Children relation correctly in some case. Such as,

- "OBJECT **was** a son of SUBJECT."
- "OBJECT **was** a daughter of the ENTITY1 emperor SUBJECT."
- "SUBJECT **is** the father of ENTITY1 laureate OBJECT."

Hence, these verbs can be considered as **good choices** for this relation.

2.5.2 Error cases

There are some cases, where the verb which is the LCA of the SUBJECT and OBJECT, does not express the Person..Children relation. Such as:

- "She **married** OBJECT , son of SUBJECT , ENTITY1 and ENTITY2 , and known as Philip the Catholic."
- "Succeeding SUBJECT , OBJECT **persecuted** ENTITY1 and his followers under pressure from ENTITY2."
- "OBJECT , son of SUBJECT , is **seen** as the culmination of utilitarianism."

Hence, the verbs *"married"*, *"persecuted"*, and *"seen"* are **bad choices** for the Person..Children relation.

2.6 Task 2 Summary

The following table shows a summary of our findings:

Relation	**Good Choices (Verbs)**
Award.. Award_Winner	awarded, presented, won, received
Business.. Industry	is, provides offers, delivers
Actor.. Character	plays, played, stars, starred, portray, portrayed, is, was
Artist.. Album	released, release, releases, releasing, is, was
Person.. Children	was, is