

Tagging Issues and Hypothesis Building for Doc Level

2013_6_9_st-535.txt Wrong Tagging

A bomb was blasted targeting a house at Harijan Street [LOC_Event] of Gosaninuagaon [LOC_Event] area of Berhampur [LOC_Event] on Saturday . According to Gosaninuagaon [LOC_Others] police [ORG_Others] , three persons of the targeted family including two women received minor injuries . As per eyewitnesses , two youths hurled bomb at the backdoor of home of Upendra Hati [PER_Victim][LOC_Event] at around 11 a.m. . Past rivalry between neighbours was suspected to be the reason behind this incident , police sources said .

2012_9_20_st-36.txt Karachi

Together with Tuesday twin blasts in Karachi [LOC_Event] aimed at the Bohra community , bombs have taken 16 lives in 24 hours . Six of the seven killed in Karachi [LOC_Event, LOC_Victim] , were from the community. Dawn reported that a large number of Bohras from across the country and India were visiting Karachi [LOC_Event] [LOC_Others] in connection with the visit.

[Neelesh] The last Karachi is tagged as LOC_Others and as per sentence level but It should be LOC_Event

ev_064_st_014.txt: Mumbai

Intelligence Bureau sources say the massive attack on Mumbai [LOC_Event] on Tuesday was to disturb communal harmony. Raman believes there is a strong possibility of Al Qaeda being behind the blasts and they could have been executed by fugitive gangster Dawood Ibrahim -- who was behind the March 12 , 1993 blasts in Mumbai [LOC_Others]

[Neelesh] The label given at sentence level is correct. So here sentence level assignment should given weight age over considering other mentions too.

ev_081_st_008: ABVP

NEW DELHI [LOC_Others] : After the issue of involvement of Hindu outfits in the Malegaon [LOC_Event] [LOC_Others] blasts came up in Rajya Sabha [ORG_Others] [LOC_Others] , ABVP [ORG_Accused][ORG_Others] leaders organized a press meet to clear its name from the controversy . The members of BJP 's [ORG_Others] student outfit insisted that the party had nothing to do with the blasts in Malegaon [LOC_Event] [LOC_Others] in September this year and it was just a political conspiracy to defame ABVP [ORG_Accused][ORG_Others] and Hindu [O] [ORG_Others] outfits . First ATS [ORG_Others] mentioned the possibility of Vidyarthi Parishad [ORG_Accused] and VHP [ORG_Accused] being involved in the blasts and then the matter was also raised in Rajya Sabha [ORG_Others] [LOC_Others] by a Congress [ORG_Others] MP [ORG_Others] . But ABVP [ORG_Accused][ORG_Others] is in no way involved in the blasts , " said Sunil Bansal [PER_Others] , regional organizing secretary , ABVP [ORG_Accused]. He added , ` ABVP [ORG_Accused] has even demanded that there should be a proper inquiry and ATS [ORG_Others] should come out with an evidence and not talk without any basis . They do n't have any proof against any ABVP [ORG_Accused] activist . " According to Bansal [PER_Others], the name of ABVP [ORG_Accused] came up during the investigations after a bike was found near the blasts site in Malegaon [LOC_Event]] [LOC_Others] . ` The bike belonged to Pragya Singh [PER_Accused] , who was an ABVP [ORG_Accused] activist in 1997 and is now a sadhvi . She had sold off the bike before she took to spirituality , " Bansal [PER_Others] said.NSUI , meanwhile , is holding a protest march against ABVP 's [ORG_Accused] alleged role in the Malegaon [LOC_Event]] [LOC_Others] blasts on Friday .

[Neelesh] In above, sentence level (local context) is not apt for assigning label to respective mentions. All these mentions are linked and so should be given the label ORG_Accused

The bike belonged to Pragya Singh [PER_Accused] , who was an ABVP [ORG_Accused] activist in 1997 and is now a sadhvi [This statement itself will not give proper info whether accused or not.]

Other Example

Mumbai was attacked on 28 Nov, 2008. The trial of the accused will be in Mumbai.

[Neelesh] [First Mumbai will be treated as LOC_Event and later one as LOC_Others]

Mumbai was attacked on 28 Nov, 2008. Few suspicious people have visited Mumbai 2 days back.

[Neelesh] [First Mumbai will be treated as LOC_Event and later one as LOC_Event]

Thoughts

- Correcting the tags. Two cases
 - Erroneous tags
 - If an entity is accused or victim, and if that person or location is being referred wrt that event, it should be tagged with the same label.
 - The annotator should not tag only taking sentence as reference point, It should be purpose of event as reference.
- Some sentence may give a correct label to the entity at sentence level and other sentence locally couldn't give the correct label [May need to refer other mentions to correct their labels.] Example: Karachi Attack example.
- Sometimes sentence level label is correct but other labels may give wrong info. Here sentence level (local context) is more suitable. Example: Mumbai Attack example.
- Both above points are contradictory, causing dilemma in building hypothesis.
- Somehow the relationship between labels being assigned by local context [Sentence Level] and global context(doc or event) [Labels of Other Mentions] need to be considered.
- Type embedding will just learn all the words. So for this it's a global learning. Combining local context as well as global. So Type representation should not have this flaw.
- Since we have opted for ranking model instead of classification, we are not interested in correcting all the labels but we want top entities in each label to be right.
- Considering above point, Type Representation has the flaw that it is learning the words also which are both present in Others as well as Accused/Victim lowering the correctness of these labels.
- How to decide which decision should be taken and how to model it in Ranking?

Methods

Doc2Vec: Consider all the entity mentions present in a doc/event and learn representation of an entity at corpus level.

True Type Assumption and Type Hierarchy: Assuming that an entity could have one prominent role (True Type) at doc/event level and don't consider it while ranking for rest of the types. For example first Rank for PER_Accused and exclude top@k persons from Rank list of PER_Victim and PER_Others. Second is how to define this type hierarchy (which type should be evaluated first. Domain Expert decision or Try all the combinations).

Combining the Ranking Score of an Entity for a Role (Doc/Event Level): The same entity could appear multiple times in a document and it will be in Ranked list. Can we combine these score to improve precision????

[Type Representation]

Identify candidate sentences to learn Type Representation:

Since we are not interested in labelling all the mentions correctly and looking into dataset, sentence level context may not just represent the label as the label is due to linkage to other mention. If we could identify the sentences which actually describe the role we may have better results.

[One Good Motivation for Ranking Model vs Classification Model]

Sequence labeling task is focused on assigning each mention certain label based on sequence, which will certainly fail here. As it will assign the LOC_Others in looking the local context.

Where Entity Ranking is not interested in correcting each mention. It is interested in finding top candidates for each Type.

. In sequence modeling learning is joint for all labels and entities. In ranking we can focus on One Role at a time and try to work on that, giving more flexibility