# CS372 homework #4 report

20160632 Janghyun Cho

We were to find triples, [A, x, B], where A and B are noun or noun phrases and x is a verb among 5 verbs that we chose. I chose positive and negative verbs with accelerate and block, leaving following 5 verbs to search: ['activate', 'inhibit', 'bind', 'accelerate', 'block']. By using Entrez.efetch, I randomly picked up few hundred articles from medline and parsed their abstracts to meet the requirements. Give priority to recent abstract, year limit of 30 sentences, 10 sentences per journal, 2 sentences per organization, and 20 sentences to each five distinct words. If the sentence has two verbs, I gave the sentence to verb that appears earlier.

After getting sentences was the making my answer sheets of triples. To get basic aid, I applied position tag to each word so that making decision of which nouns the verb is connecting can be seen easier. It's my mistake but I forgot to using learning module and just implemented code for finding triples my goal was to put the problem to its simplest shape. My triples and answers are only restricted to one word each for A, x, and B. Making post processing easier.

My idea was to make a tree of sentence with given sentence and first find for verb with its position tag and lemmatized form. Then starting from the verb's position, it looks front and back to find A and B it connects to. Sadly I forgot to think about by, which makes up passive form of A, x, and B. Like A x by B should been B x A but I didn't look for it in both my answers and training. I made tree with three parts, NP, PP, and VP. They are just as given in lecture, NP: DT or JJ or NN.*, PP: IN + NP, VP: VB.* + NP or PP So verbs only exist without any parent or just inside a VP. So I made cases for each of them. If the verb stands by itself, A should be before the verb and B should be after the verb. And it was done accordingly by searching nearest NPs. In other case, if the verb is inside a tree of VP, I looked for A before the VP and B just inside the VP, either NP or PP since both has nouns. Appearance of CC such as and or or is a special case. If CC appears right next to NP, then if CC is in between two NPs, then both NPs can be A or B. Other case, if CC appears to verb before NP does, verb must look for another verb before it can get NP for A or B. There were some cases that I don't think either appropriate A or B exists. in that case I put

the word None for the empty case.

With these ideas, I made up my code and random 80 and 20 integers which adding them makes 0~99. By comparing my results with 80 training datas, I made some adjustments. And with that value, I ran testing datas. Result is like following. Precision is with model's results. Precision = actual truth that model said True / all truth that model said True. Recall = True that model managed to predict / all truth given by answer. F-score is harmonic mean of Precision and Recall. So for training and testing results, my performance is like following. We can see all results of training is slightly better than that of testing due to changes made by comparing answers with results.

| | Training | Testing |
|---|---|---|
| Precision | 0.46 | 0.4 |
| Recall | 0.575 | 0.5 |
| F-score | 0.5111 | 0.4444 |

I think there are many changes that can be made. Also, performance is just around half-half, which means it still has much ways to go. First, I wasn't able to manipulate commas correctly. Using comma can mean many things. Either it can be used as NounA, NounB to represent A and B are equal or noun verbA object, verbB object to give two verbs A and B to same noun at the front. Second I think prepositions are not taken care of fully, especially 'by'. As mentioned earlier, I didn't do much for passive forms of sentence, which is error prone. Also, having words right next to prepositions have chance of connecting it to far away noun but not much work's done here. Third, I can't fully trust the answers of triples that I gave to be both trained and tested. Many sentences are hard and of complex form, which makes me to understand it fully. Lastly I think there are some problem with position tagging with science terminologies. Some words I thought as noun or verb was not with that tag. Perhaps this is also with third reason.