**Dear professor Iwaihara,**

According to the analysis I made in these days I think continue to do in the column type task is possible as long as we didn't seek to outperformed the state of art model because it's really hard to exceed no matter the micro or macro F1 score without using its model structure. But I'm not sure if I should just switch to another task now. I would like to prefer your suggestion if you could spare me some time for another meeting recently.

**Sato**

For the wikitables oh below shows analysis on the subtle data set the first column represents the number of data which is the type in the test data set well as we can see the sota actually make little mistake on the types except the type with very little number since the total number of the test data set is over 20,000 the mistake on this data is very minor to see overall dataset.



This is an example of the type affiliate and this column only has one data and I tried to find it on the Google and it turned out that it's a politician and I strongly doubt that's mistake in the dataset. Which makes the Doduo unable to correctly annotate the type and it's also impossible for us to manually annotate it by following some certain rules or patterns.

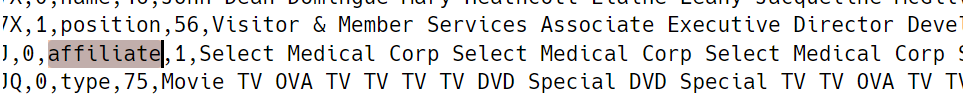
Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

and the same thing happened here too and it's also an affiliate data type and we got this select medical crop. We can look up it on the Dbpedia but there is no clue about the affiliates and I try to add the affiliates as keywords and Google it and there is also no result.



Graphical user interface, text, application

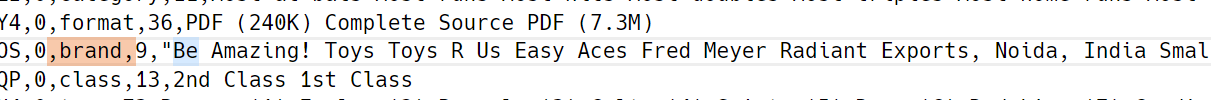
Description automatically generated

A picture containing table

Description automatically generated

and next it's about the type brand you can see here there be amazing toys and I searched on Google and it turned out it's a company and it has amazing toy brand and the amazing toy product and I think it's also very difficult for us to annotate it manually even following some rules. And if we want to use the majority vote the third picture below shows the following data, which are almost all limited company, having nothing to do with brand.







I also checked the other types in the sato dataset and it shows that almost all the incorrect part have some connection with the weakness of the dataset itself which cannot be fixed with some certain rules or the majority vote. So I'm afraid that all the improvements on the sato dataset is impossible.

**WikiTables**

For the wikitables I found some points that I missed in the past pyramid experiment, and I was using Li Peining’s method on the wikitables. Actually, the wikitables dataset is fine grained and the number of answer result label is more than one on some occasions Despite the large scale of wikitables dataset, the content in the table is very specified, which means maybe lookup method over the knowledge base can do some help. However, since the performance of state of art is really good, I'm not sure if we can exceed it on this dataset. And I'm not doing experiments on it. Maybe I will send you another email attached with the experiments about wikitables in two days later.

Below is a snapshot of the wikitables dataset. As you can see the label contains the roughly labeled type and the fine grained specified type and for one column data The data could be any of the roughly annotated label or the specified label. And this property makes the wikitables dataset have 255 types, which is far away beyond the number of Sato datasets which is 78.



**GitTables**

Chart, histogram

Description automatically generated

Above is the figure of the gift table that set, The X axis is a bit crowd. but it can be ignored, every point in the X axis represent table pattern there are only 2500 columns in the data set and the number of table is 800 but we can see through the figure that the table with the same pattern holds the maximum number of 200 which means there are a lot of duplicates in the data set. And I checked the data set manually and found the data is not really that good. below is of the example of the table pattern with the maximum number. End the column eight and column nine is the column we want to annotate and there is a lot of empty cell in the table which makes it not a really an ideal data set for column typing task. So for now I think the gate table data set can be excluded from the target data set.

Calendar

Description automatically generated