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.

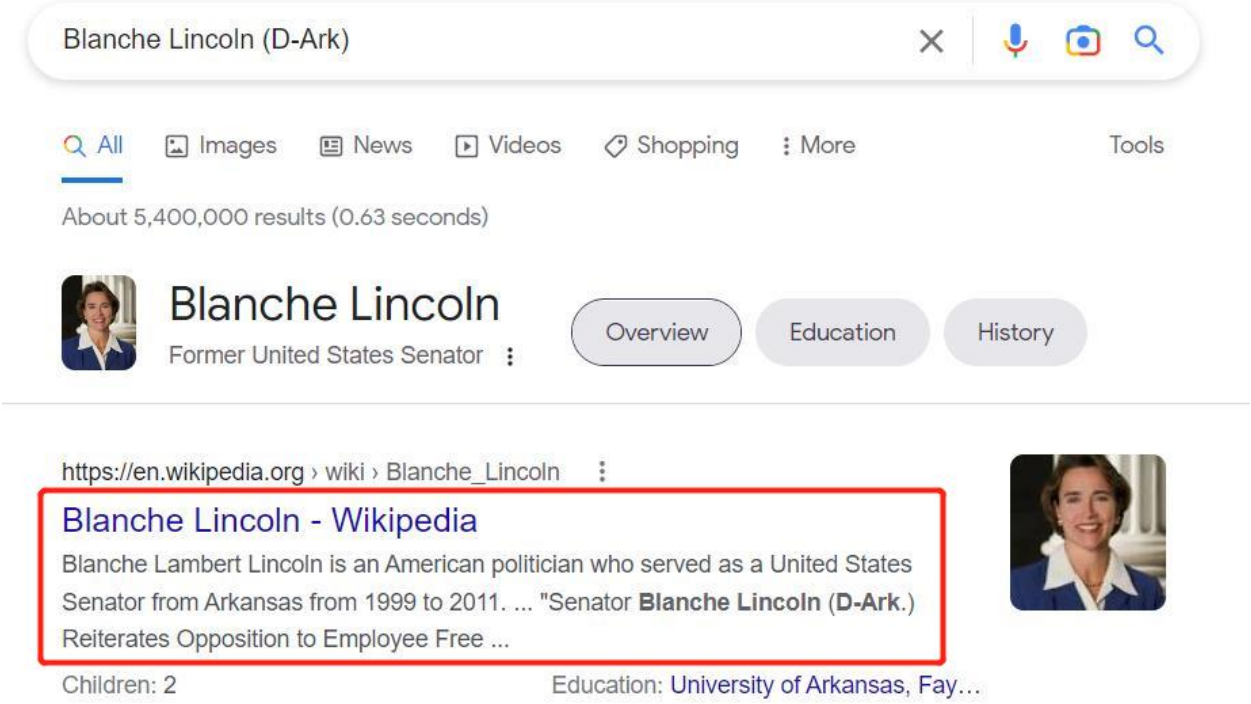
Sato

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.

#data	confusion Mat																	
5	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	1	0	0	1	0	0	12	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	7	1	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
10	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
33	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
67	0	0	0	1	0	0	1	0	0	0	0	1	2	4	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

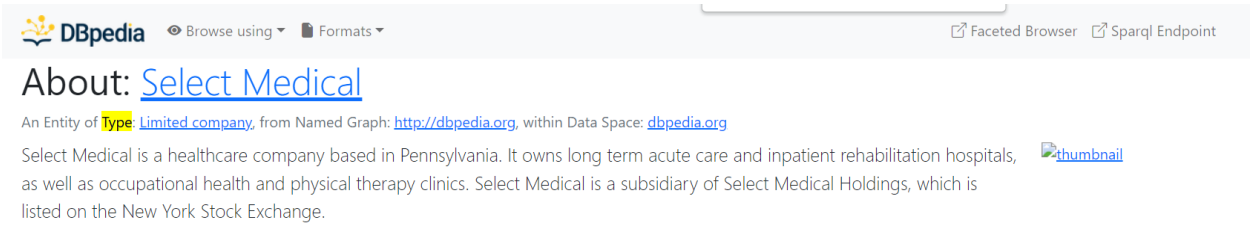
This is an example of the type affiliate and this column only has one data and I tried to find it on the Google and it turned out that it's a politician and I strongly doubt that's mistake in the dataset. Which makes the Doduo unable to correctly annotate the type and it's also impossible for us to manually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

```
91    H_YG6XAIGMQWK3ZAGM,0,category,11,3/4 Women WOMEN 60km Women C
92    H_YG6XAIGMQWK3ZAGM,1,teamName,74,CRCA/NYVelocity CRCA/NYVELOC
93    as_BCZ64R6SHHJEUI3E,0,affiliate,1,Blanche Lincoln (D-Ark)
94    .e_YJPMI4KRRDZ5PIYI,0,name,46,AUD / USD AUD / GBP AUD / EUR
95    m_R3TV743IONCDZZEO,0,year,77,2007,2006,2006,2006
```



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.

```
7X,1,position,56,Visitor & Member Services Associate Executive Director Devel
J,0,affiliate,1,Select Medical Corp Select Medical Corp Select Medical Corp S
JQ,0,type,75,Movie TV OVA TV TV TV TV DVD Special DVD Special TV TV OVA TV TV
```



rdf:type	• owl:Thing • dbo:Company • yago:Institution108053576 • yago:Organization108008335 • yago:Abstraction100002137 • yago:Company108058098 • yago:Group100031264 • dbo:Organisation • yago:WikicatCompaniesBasedInHarrisburg,Pennsylvania • yago:WikicatCompaniesEstablishedIn1996 • yago:WikicatCompaniesListedOnTheNewYorkStockExchange • yago:WikicatPrivateEquityPortfolioCompanies • yago:SocialGroup107950920 • dul:Agent • schema:Organization • dul:SocialPerson • dbo:Agent • wikidata:Q24229398 • wikidata:Q43229 • wikidata:Q4830453 • yago:YagoLegalActor • yago:YagoLegalActorGeo • yago:YagoPermanentlyLocatedEntity

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.

```
on.gz_2032-9 Center Place, Toms Rive_27DMUPKR3PCGLF4F
n.gz_1080--48_QFKDWJQBLGSY4I7N,0,brand,9,Ages:
Y4,0,format,36,PDF (240K) Complete Source PDF (7.3M)
OS,0,brand,9,"Be Amazing! Toys Toys R Us Easy Aces Fred Meyer Radiant Exports, Noida, India Smal
QP,0,class,13,2nd Class 1st Class

Heshan Congtin Technological Development Co. Ltd., of China Ultimate Products (HK) Ltd.,
```

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.

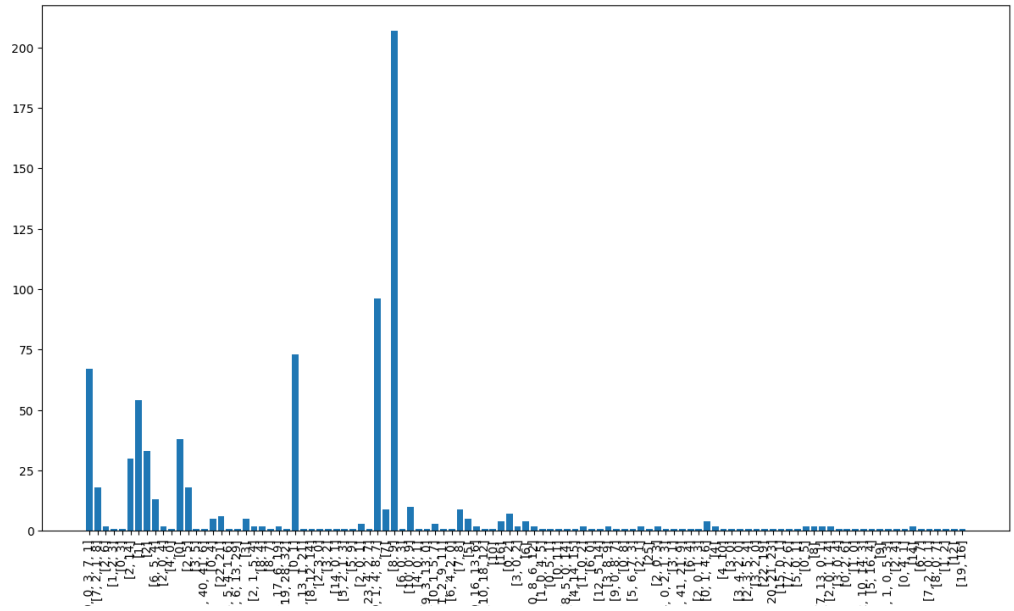
WikiTables

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.

	table_id	labels	data	abel_ids
4018	10015132-12	['sports.pro_athlete', 'basketball.basketball_player', 'people.person']	Jamaal Magloire Sean Marks Shawn Marion Donyell Marshall Darrick Martin Roger Mason Tony Massenburg Bc	0, 0, 0, 0, (
4019	10015132-12	['location.country', 'location.location']	Canada New Zealand United States United States United States United States United States United States Unit	0, 0, 0, 0, (
4020	10015132-12	['sports.sports_position']	Center Forward-Center Forward Forward Guard Guard Forward Forward Forward-Center Guard-Forward Forwa	0, 0, 0, 0, (

GitTables



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.

	col0	col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12	col13	col14	col15	col16	col17
0	0	21600280	1.61E+09	MIA	Miami	202337	Luke Babb	F		7:52	0	0	0	0	0	0	1	2
1	1	21600280	1.61E+09	MIA	Miami	201177	Josh McRc	F		23:32	5	10	0.5	0	1	0	0	0
2	2	21600280	1.61E+09	MIA	Miami	202355	Hassan Wl	C		36:16:00	4	11	0.364	0	0	0	1	2
3	3	21600280	1.61E+09	MIA	Miami	203585	Rodney Mı	G		36:03:00	2	4	0.5	1	2	0.5	4	4
4	4	21600280	1.61E+09	MIA	Miami	201609	Goran Dra	G		36:50:00	10	21	0.476	2	3	0.667	5	7
5	5	21600280	1.61E+09	MIA	Miami	204020	Tyler Johnson			29:43:00	3	11	0.273	0	4	0	2	2
6	6	21600280	1.61E+09	MIA	Miami	201961	Wayne Ellington			33:32:00	6	9	0.667	3	5	0.6	2	2
7	7	21600280	1.61E+09	MIA	Miami	201949	James Johnson			24:28:00	11	15	0.733	2	4	0.5	0	0
8	8	21600280	1.61E+09	MIA	Miami	203186	Willie Reed			11:44	3	5	0.6	0	0	0	0	0
9	9	21600280	1.61E+09	MIA	Miami	2617	Udonis Haslem		DNP - Coach's Decision									
10	10	21600280	1.61E+09	MIA	Miami	1626196	Josh Richardson		NWT - Sprained right ankle									
11	11	21600280	1.61E+09	MIA	Miami	202682	Derrick Williams		DND - Back spasms									
12	12	21600280	1.61E+09	MIA	Miami	1626159	Justise Winslow		NWT - sore left wrist									
13	13	21600280	1.61E+09	UTA	Utah	2207	Joe Johnsc	F		32:35:00	6	9	0.667	4	6	0.667	2	2
14	14	21600280	1.61E+09	UTA	Utah	2564	Boris Diaw	F		13:34	0	4	0	0	2	0	0	0
15	15	21600280	1.61E+09	UTA	Utah	203497	Rudy Gobe	C		35:23:00	3	4	0.75	0	0	0	0	0
16	16	21600280	1.61E+09	UTA	Utah	202330	Gordon He	G		40:19:00	12	22	0.545	3	7	0.429	5	6
17	17	21600280	1.61E+09	UTA	Utah	203957	Dante Exu	G		18:27	1	6	0.167	1	4	0.25	2	2
18	18	21600280	1.61E+09	UTA	Utah	1626168	Trey Lyles			24:40:00	5	8	0.625	4	4	1	0	0
19	19	21600280	1.61E+09	UTA	Utah	202714	Shelvin Mack			32:47:00	4	9	0.444	1	2	0.5	2	4
20	20	21600280	1.61E+09	UTA	Utah	204060	Joe Ingles			22:26	6	8	0.75	3	4	0.75	0	0
21	21	21600280	1.61E+09	UTA	Utah	203481	Jeff Withey			12:37	2	2	1	0	0	0	2	3
22	22	21600280	1.61E+09	UTA	Utah	203526	Raul Neto			7:12	1	1	1	1	1	1	0	0
23	23	21600280	1.61E+09	UTA	Utah	1627762	Joel Bolomboy		DNP - Coach's Decision									
24	24	21600280	1.61E+09	UTA	Utah	201588	George Hill		DND - Sprained left big toe									
25	25	21600280	1.61E+09	UTA	Utah	203918	Rodney Hood		DND - Right hamstring strain									