Database Systems

Academic Year 2022 - 2023

Group

Christos Morfopoulos

George Balaskas

Selected Paper :

Annotating columns with pre-trained language models

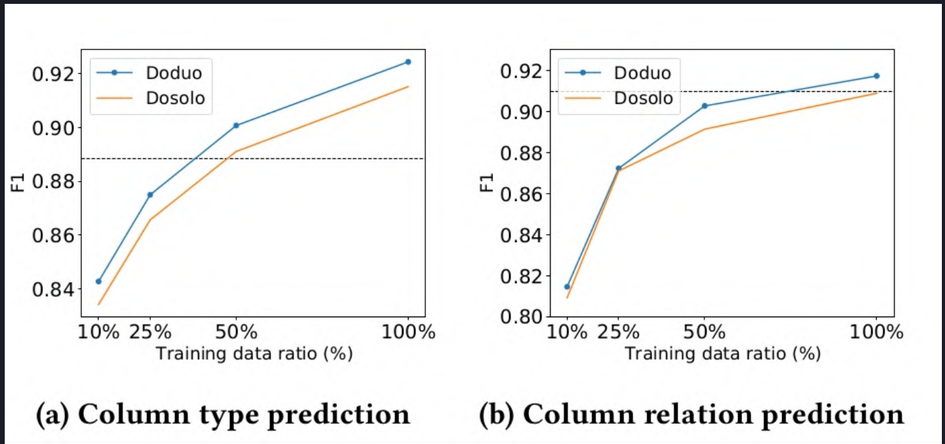
https://arxiv.org/abs/2104.01785

Quick Recap

- The main goal is to predict the name as well as the relation of the column(s) based on its concatenated column values.
- The authors utilized the BERT model, a powerful LLM.
- It can be considered as a sequence Multi-label Classification Task.
- The tasks of Annotating the Columns and Predicting the Column Relations can be combined via Multi Task Learning
- Therefore, the authors introduced DoDUO , a unified table-wise model which is trained for both tasks, and DoSOLO a single column model.
- DoDUO can be considered as SOA, outperforming all the previous models including DoSOLO.
- The training Datasets were the WikiTable and Viznet.

DoDUO & DoSOLO Performance

Fig. Performance of DoDUO - DoSOLO



- From the figures we can notice that the DoDUO outperforms DoSOLO in both tasks, suggesting that taking the whole table context and predicting its respective columns, plays an important role in contrast to predicting single column name independently.

Fig. F1 scores of DoDUO - DoSOLO in the respective task

Table 6: Ablation study on the WikiTable dataset.

Method	Type prediction	Relation prediction
DODUO	**92.50**	**91.90**
w/ shuffled rows	91.94	91.61
w/ shuffled cols	92.68	91.98
DOSOLO	91.37 (1.23% ↓)	91.24 (0.7% ↓)
$\text{DOSOLO}_{\text{SCOL}}$	82.45 (21.9% ↓)	83.08 (9.6% ↓)

- In this way, we can also notice that the Multi Task Learning in those relative tasks, improved remarkably the performance of the model, pointing how beneficial it is to achieve the greatest performance.

Main Approaches & Dataset

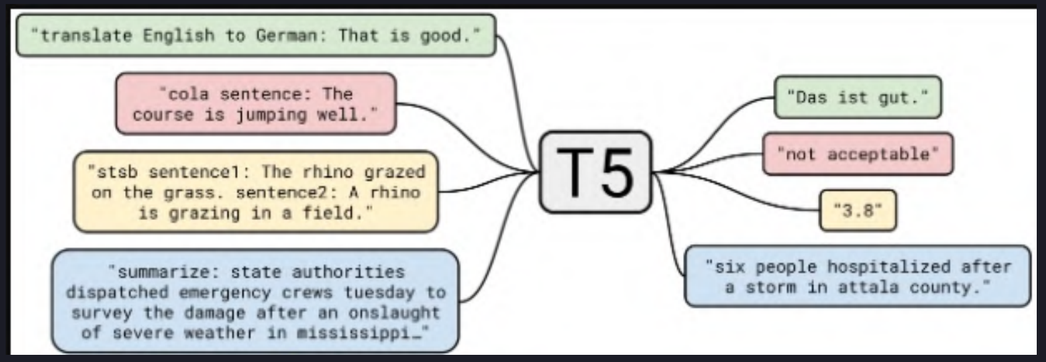
Approaches

Masked Column Prediction

Inspired by the training of the TaBERT model,
we employed the technique of Masked Column Prediction, in order to improve the performance of the pretrained DoDUO model. Similar to MLM, in the MCP the goal is to mask certain columns so the model can learn the table context by predicting the column values of a table.

T5 architecture/model

Fig. Illustrative example of T5 architecture



Due to errors in the original code, we took the initiative to experiment with a different architecture and explore any common pattern between those approaches.
We utilized the T5 model , a transformer-based sequence to sequence model, which “models” every problem in text-to-text format.
The applications of the T5 model have a great variety, such as classification, summarization, translation or even regression!

WikiTable Dataset

We chose to focus on the WikiTable Dataset, although is much larger than VizNet.

For column type prediction, the dataset provides 628.254 columns from 397.098 tables annotated by 255 column types. For column relations, the dataset provides 62.954 column pairs annotated with 121 relation types from 52.943 tables.

Since it is a multi-label classification , in the column type we have 255 number of classes whereas in the column relation we have 121 number of classes

Fig. A WordCloud of the Column Values

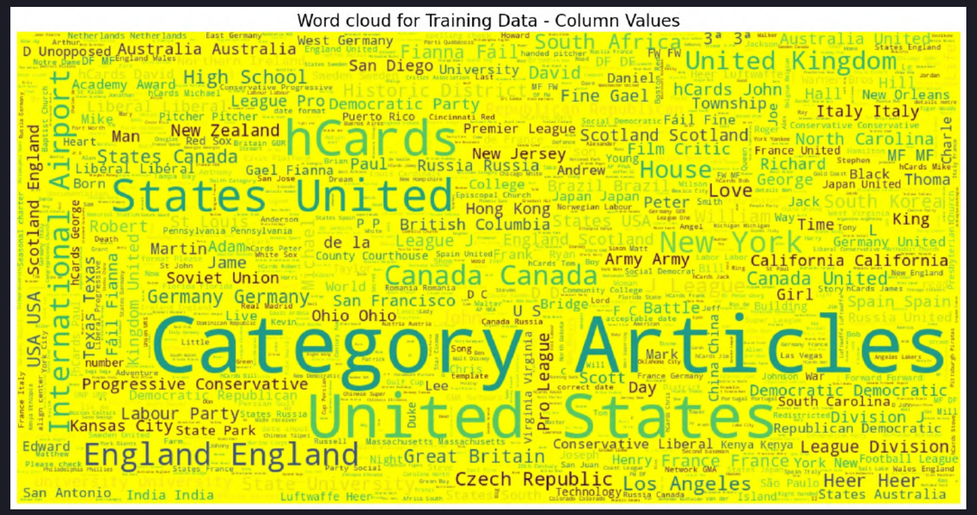
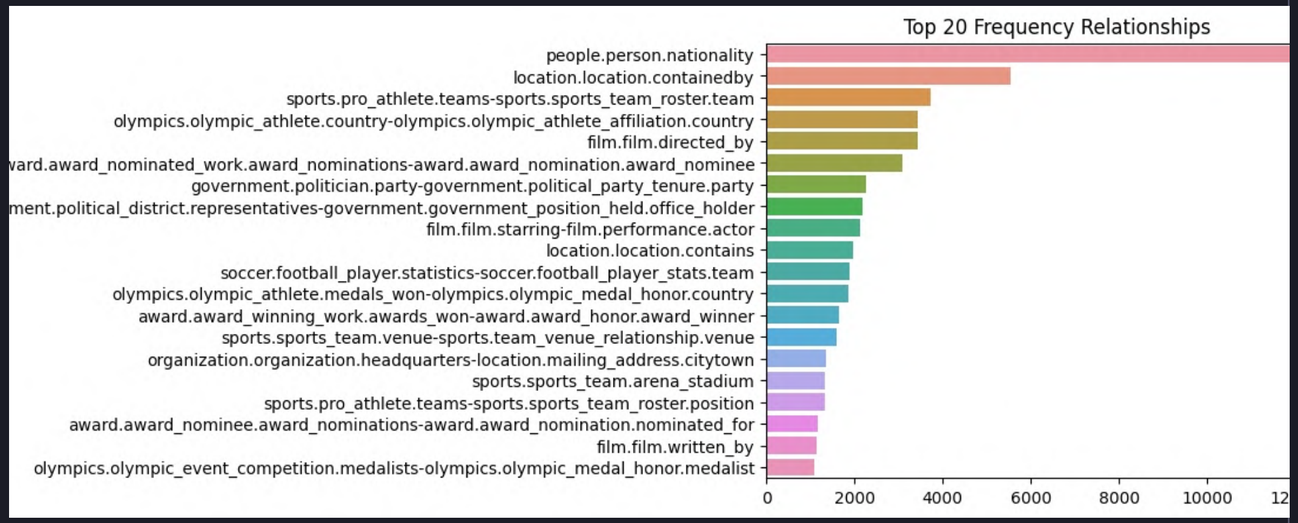


Fig. The Top 20 most frequent relations in the dataset



Masked Column Prediction (MCP)

Fig. Training example of a table from WikiTable dataset with data& labels

	table_id	labels	data	label_ids
248689	10535445-4	[film.actor, tv.tv_personality, people.person]	Stacy Keibler Giselle Fernandez Tatum O'Neal S...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
248690	10535445-4	[tv.tv_actor, film.actor, tv.tv_personality, p...	Master P Lisa Rinna Tatum O'Neal Master P Mast...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

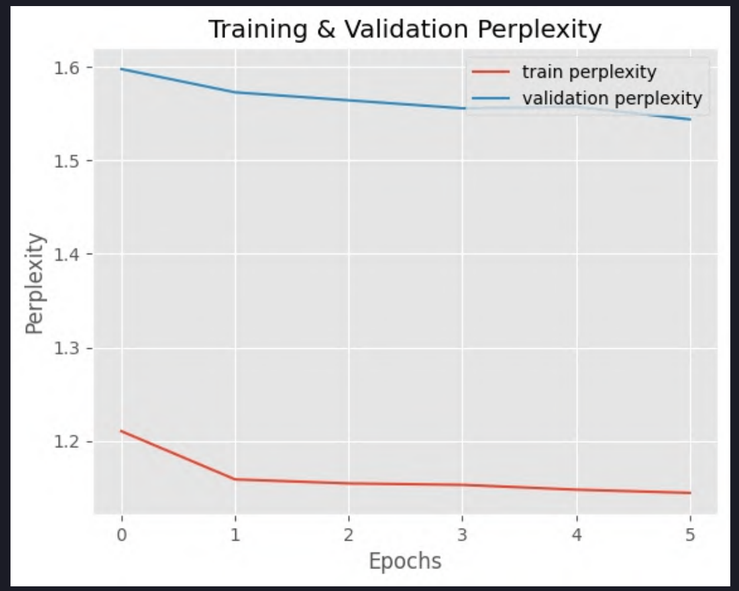
Fig. Training example of a table, in which the special tokens (MASK) are applied

```
[ ]  1 tokenizer.decode(example_MSP_dataset[2350].get('input_ids'))

'[CLS] stacy keibler giselle fernandez tatum o'neal stacy keibler drew lachey drew lachey lisa rinna stacy keibler stacy keibler drew lachey stacy keibler drew lachey stacy keibler lisa rinna drew lachey jerry rice drew lachey [SEP] [MASK] [MASK] [MASK] [MAS
K] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]'

[ ]  1 tokenizer.decode(example_MSP_dataset[2350].get('labels'))

'[CLS] stacy keibler giselle fernandez tatum o'neal stacy keibler drew lachey drew lachey lisa rinna stacy keibler stacy keibler drew lachey stacy keibler drew lachey stacy keibler lisa rinna drew lachey jerry rice drew lachey [SEP] master p lisa rinna tatum
o'neal master p master p jerry rice tia carrere master p jerry rice stacy keibler [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PA
D] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]'
```

Finetune DoDUO via MCP :

- The MCP technique was first introduced in TaBERT and is very similar to Masked Language Modelling
- It is mainly an **attempt/experiment** to improve the remarkable performance of DoDUO model.
- We trained a vanilla Bert model and only for the column type task, and then we switch the encoder to DoDUO's.
- We masked randomly 20% of the total columns.
- Due to computational resources, we have a threshold of 250.000 training examples and we trained the model for 6 epochs

Fig. Visualization of the Perplexity of the model



MCP - Performance

- A common practice to evaluate the performance of the MLMs, is the perplexity evaluation metric, which is the exponential of the training & validation loss.
- The lower the perplexity the better the performance
- We applied inference on the test dataset and the micro F1 score was 90,86% in contrast to the previous 90,71% F1 score.
- The improvement may be small, but the positive outcome is the integration of MCP.

T5 model variants

T5 model variants

Fig. The length of the tokenized column values

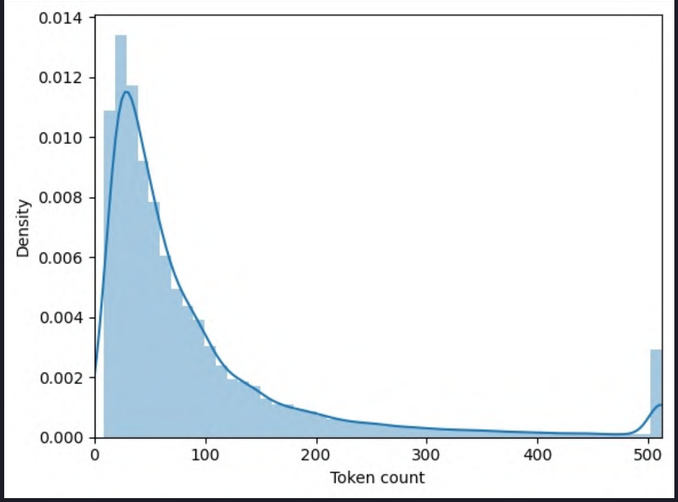


Fig. A partition of the dictionary with its classes

```
0 american_football.football_coach
1 american_football.football_conference
2 american_football.football_player
3 american_football.football_team
4 amusement_parks.park
5 amusement_parks.ride
6 architecture.architectural_structure_owner
7 architecture.building
8 architecture.structure
9 architecture.venue
10 astronomy.asteroid
11 astronomy.astronomical_discovery
12 astronomy.celestial_object
13 astronomy.constellation
14 astronomy.orbital_relationship
15 astronomy.star_system_body
16 automotive.company
17 automotive.model
18 aviation.aircraft_model
19 aviation.aircraft_owner
20 aviation.airline
21 aviation.airport
22 award.award
23 award.award_category
24 award.award_ceremony
25 award.award_discipline
26 award.award_presenting_organization
27 award.competition
28 award.hall_of_fame_inductee
29 award.recurring_competition
30 baseball.baseball_league
31 baseball.baseball_player
32 baseball.baseball_position
```

- We took the initiative after the single column model failures.
- It is based on T5-small, an encoder-decoder transformer model.
- As it is text-to-text model, we should define the Source & Target max length.The bigger is the length, the more computational expensive(180-50).
- We trained a total of five T5 models (*single-col type, single-col relation , single-col MTL, tablewise col type, tablewise col relation*).
- To save computation cost, we defined a training threshold of 150000 examples .
-
- To perceive homogeneitywe trained every model for 10 epochs.
- We defined a dictionary for both tasks (col type & col relation) with 255 classes and 121 classes respectively, mainly to help the decoder !
- We followed the same approach as DoDUO, concatenating the column values and applying multi label classification.
- More or less all the T5 models follow the same rationale, apart from the MTL model, where we use 2 different dataloaders and a single Loss function (with equal weights).
- The overall goal is to explore an alternative way of a non Bert-based model and investigate as well as confirm any similar pattern between DoDUO & T5

Performance of T5 models

Fig. Single Col Annotation model

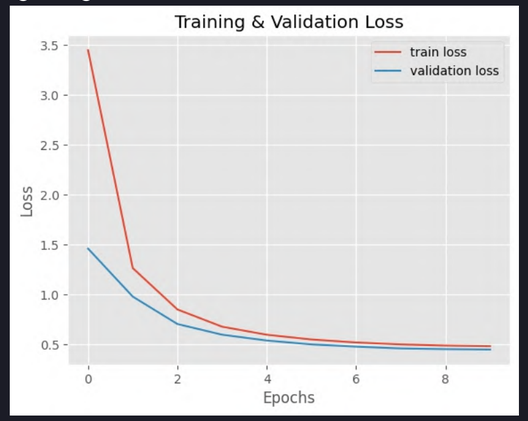


Fig. Single Col MTL model

Fig. Tablewise Relation model

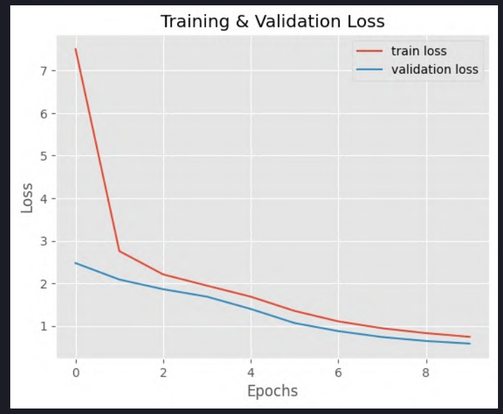


Fig. Single Col Relation model

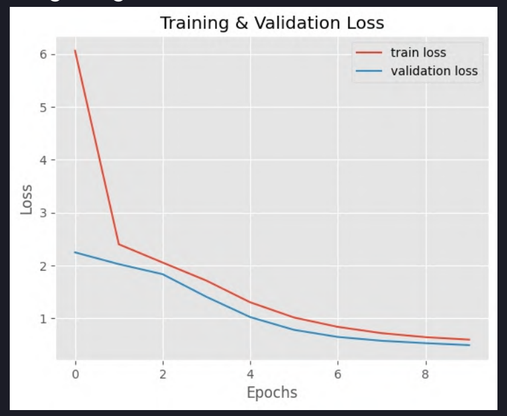


Fig. Tablewise Annotation model

Summary:

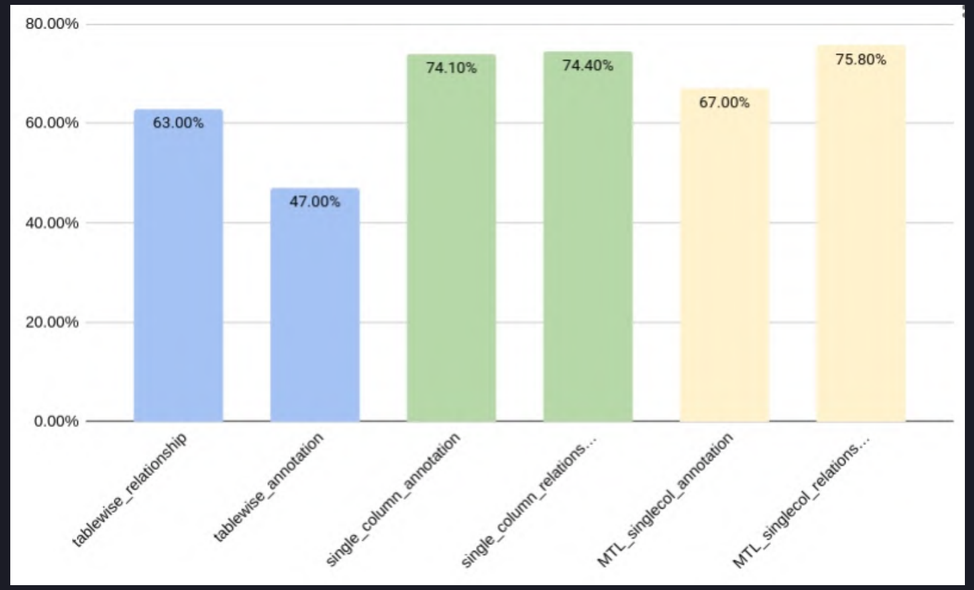
- The single column annotation model achieved a F1 score of 74,1% and accuracy of 50%
- The single column relation model achieved a F1 score of 74,4% and accuracy of 67%
- The single column MTL model achieved F1 scores of 67% and 75.8% in the col type and relation task respectively.
- The table-wise annotation model achieved a F1 score of 47%.
- The table-wise relation model achieved a F1 score of 63%.

*all the F1 scores come from the validation set.

Overall Performance

Overall Performance

Fig. An overall graph with all the F1 scores of the T5 models



At a quick glance from the graph we can verify that:

- The single column models outperform the table-wise models in each task.
- The table-wise models did not succeed to have a competitive performance, although they captured the whole table context.
- The Multi task single column model is the best in terms of performance, which achieved a 75.8% F1 score outperforming the single column relation model (74.4%) .

Conclusion

Conclusion

To sum up, after the T5 experiments as well as the MCP integration we can easily conclude that ：

- All T5 model variants have a similar pattern in their training phase.A key point is that validation losses were decreasing, therefore the models have potential for even more training/learning from their tasks.

- The MCP is an alternative technique in order the model to learn the training data.

- The single column (T5) models outperform the table-wise models, where the DoDUO & DoSOLO have different behavior.

- The main novelty of the paper, we verified it as well, is that via Multi Task Learning the model can actually learn from its relative tasks and have a better performance after all.

Thank you

Any Questions?