Towards Parameter-Efficient Automation of Data Wrangling Tasks with Prefix-Tuning

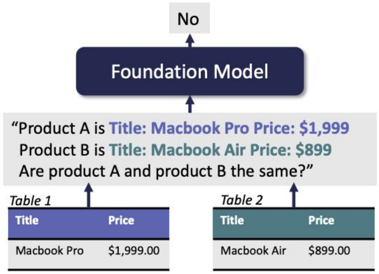
David Vos
TRL @ NeurIPS

INDE Lab, University of Amsterdam
David Vos, Till Döhmen, Sebastian Schelter

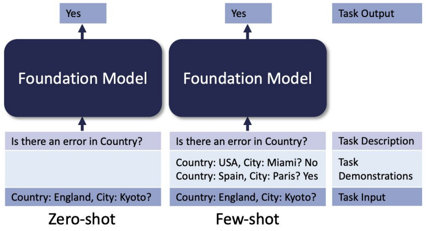
Contents

- Introduction of data wrangling tasks
- Data wrangling and large language models (LLMs)
- Prefix-tuning methodology
- Our contributions
- Experimental setup
- Results, conclusion and future research

Data Wrangling Tasks



Entity Matching (EM)



Error Detection (ED)

Given an entry, infer the missing value(s).

Data Imputation (DI)

The benchmark we use consists of 7 EM, 2 ED and 1 DI datasets

State of the art for data wrangling

Few-shot prompting GPT-3 achieves SoTA performance on 7/10 benchmarks [1].

1. **Format table rows using linearization**

2. **Select suitable prompt-samples (requires manual selection)**

   Randomly selecting results in an average drop of 14.7 F1 points

3. **Format a prompt**

   Minor modifications caused an average variance of 9.4 F1 points

Parameter-efficient and SoTA performance but hard to scale due to lack of automation and privacy concerns.

[1] Avanika Narayan, Ines Chami, Laurel Orr and Christopher Ré. Can Foundation Models Wrangle Your Data? 2022

Is there an automatable and privacy-friendly alternative?

Maybe finetuning?

Finetuning

Finetuning can be performed for smaller LLMs like T5.

1. **Format table rows using linearization**
2. **Update LLM weights with full finetuning procedure**
3. **Duplicate model weights for each new task**

Automatable and more privacy-friendly with similar performance but hard to scale due to expensive model duplications.

Prompting vs. Finetuning

Finetuning

- **Automatable**

  Can be optimized using standard techniques
- **More privacy friendly**

  Can be optimized without looking at data
- **Less scalable**

  Requires a model copy for each new task

Prompting

- **Not easily automatable**

  Requires manual labour and expertise
- **Less privacy friendly**

  Requires manual inspection of training samples
- **Scalable**

  Scales to new tasks without extra parameters

Can we take the best of both prompting and finetuning?

Prefix-tuning?

LLM parameters are frozen.

Prefix-tuning learns a continuous prefix instead of engineering a discrete prompt.

Prefix-tuning

Prefix-tuning learns a continuous prefix instead of defining a discrete prompt [2].

- **Parameter-efficient**

  The prefix generation requires only 0.4% of the parameters required for finetuning.

- **Automatable and privacy friendly**

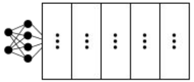
  Prefix parametrization can be trained end-to-end similar to finetuning.

[2] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. ACL, 2021.

Intuition

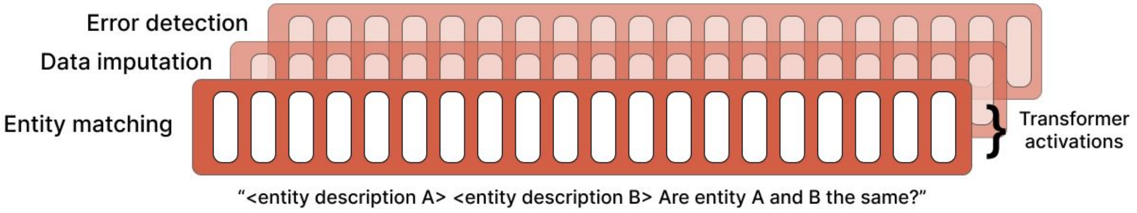
An example for entity matching:

Discrete prompt: Are entity A and Entity B the same? entity A: <entity A>, entity B <entity B>

Prefix-tuning:  entity A: <entity A>, entity B: <entity B>

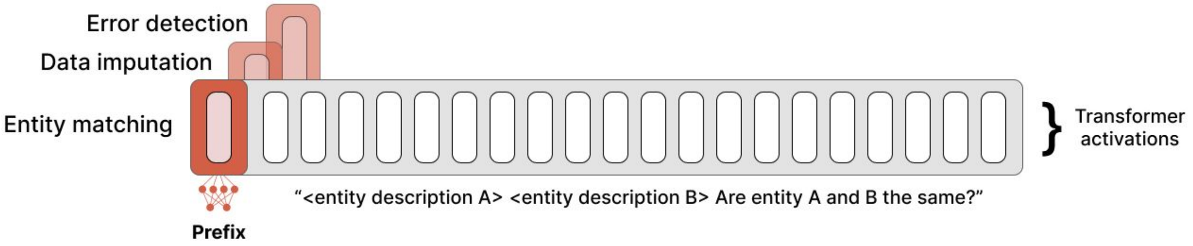
Prefix-tuning

Finetuning (updates all LLM parameters)



$$h_i = \mathrm{LM}_\phi(z_i, h_{<i})$$

Prefix-tuning (keeps LLM parameters frozen and updates the tiny prefix network)



$$h_i = \begin{cases} P_\theta[i,:], & \text{if } i \in \mathsf{P}_{\text{idx}}, \\ \mathrm{LM}_\phi(z_i, h_{<i}), & \text{otherwise.} \end{cases}$$

Figure based on Figure 1 from [2]: Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. ACL, 2021.

We take prefix-tuning to data wrangling

Parameter-efficient and automatable, perfect for big data environments

We compare prefix-tuning to finetuning T5. Prefix-tuning GPT-3 is currently impossible.

Experimental Setup

Table entries are serialized:

Entity matching

"Product A: *serialize(tuple-i).* Product B: *serialize(tuple-j)*. Are product A and product B the same?"

Error Detection

"*serialize(tuple)* Is there an error in *attribute-i: value-i*?"

Data Imputation

"*serialize(tuple) attribute-i*?"

Experimental Setup

- T5-base implementation by Hugging Face
- Trained for 50 epochs or 100 if there is no clear convergence
- Experimented with different learning rates

Table 1: Datasets with their corresponding task, domain and label distribution.

Task	Dataset	Domain	#Samples	Frac. Positive
Entity matching	Beer	food	450	15.1%
	iTunes-Amazon	music	539	24.5%
	Fodors-Zagats	food	946	11.6%
	Walmart-Amazon	electronics	10,242	90.6%
	Amazon-Google	software	11,460	10.2%
	DBLP-ACM	citation	12,363	25%
	DBLP-Google	citation	28,707	18.6%
Error detection	Hospital	healthcare	19,000	2.7%
Data Imputation	Buy	electronics	651	-
	Restaurant	address	864	-

Experimental Setup

We compare prefix-tuning to two other methods:

1. **Zero-shot prompting GPT-3**

   Both methods are automatable, parameter-efficient and privacy friendly.

2. **Finetuning T5**

   Finetuning T5 is less scalable than prefix-tuning.

Results

Table 3: Prefix-tuning drastically outperforms (trainingless) zero-shot prompting across all tasks.

Task	Dataset	Metric	Prefix-tuning T5 (220M params)	Zero-shot prompting GPT-3 (175B params)
Entity matching	DBLP-Google	F1-score	0.9517	0.646
Entity matching	DBLP-ACM	F1-score	0.981	0.935
Entity matching	iTunes-Amazon	F1-score	0.9286	0.659
Entity matching	Fodors-Zagats	F1-score	0.9767	0.872
Entity matching	Beer	F1-score	0.8571	0.786
Entity matching	Walmart-Amazon	F1-score	0.7961	0.606
Entity matching	Amazon-Google	F1-score	0.6642	0.543
Imputation	Buy	Accuracy	0.9231	0.846
Imputation	Restaurant	Accuracy	0.8488	0.709
Error detection	Hospital	F1-score	0.9766	0.069

Results

Table 2: Relative performance of prefix-tuning compared to finetuning on ten data wrangling tasks. In five out of ten cases, prefix-tuning is within 2.3% of the performance of finetuning, even though it leverages only 0.39% of the parameter updates required for finetuning the full model.

Task	Dataset	Metric	Prefix-tuning	Finetuning	Rel. Perf.
Entity matching	DBLP-Google	F1-score	0.9517	0.9552	99.6%
Entity matching	DBLP-ACM	F1-score	0.981	0.9876	99.3%
Error detection	Hospital	F1-score	0.9766	0.9912	98.5%
Entity matching	iTunes-Amazon	F1-score	0.9286	0.9455	98.2%
Entity matching	Fodors-Zagats	F1-score	0.9767	1.000	97.7%
Entity matching	Beer	F1-score	0.8571	0.8966	95.6%
Imputation	Buy	Accuracy	0.9231	0.9692	95.2%
Imputation	Restaurant	Accuracy	0.8488	0.8953	94.8%
Entity matching	Walmart-Amazon	F1-score	0.7961	0.8806	90.4%
Entity matching	Amazon-Google	F1-score	0.6642	0.7436	89.3%

Conclusions and future research

Conclusions

- Performance is within 2.3% of finetuning for five out of ten cases
- Prefix-tuning with T5 outperforms zero-shot prompting with GPT-3
- Prefix-tuning is an excellent option for large enterprise solutions

Future Research

- Scale prefix-tuning approaches to larger models (GPT-3 or alternatives like GPT-J or Bloom)
- Develop more advanced parametrization of the prefix

Any questions?

My contact details

vos.dja@gmail.com

www.davidvos.dev

@dja_vos

davidjavos

Acknowledgements