510 project proposal

Research Question

Entity matching (EM) over textual data is an important problem in data cleaning and analytics. However, existing EM systems leveraging pre-trained language models (e.g., BERT) to approximately process EM tasks but does not provide statistical guarantees on the results. Can we approximately process EM tasks while maintaining statistical guarantees such as recall targets on the results?

Significance

Statistical guarantees on the approximate results of data analytics is important for real-world applications. On resource-intensive tasks, practitioners are interested in approximate execute the task while have guarantees on error semantics or accuracy semantics.

Novelty

There are two main related field:

1. EM with pre-trained models: [1-2]
2. Approximate selection query processing: [3]
   Our work is different from (1) because existing EM systems does not provide guarantees on the results. Our work is different (2) because existing AQP systems for text data do not support operations on two tables.

Approach

We propose to use the two-stage stratified sampling + pilot sampling to efficiently tackle the EM tasks with pre-trained models, providing statistical guaranteed results.

Evaluation

We find three real-world datasets that can be used to evaluate our approach:

1. Quora
2. Company
3. Stackoverflow
   We provide guarantees on the recall target of EM tasks. Therefore, we will evaluate the following performance:
4. Does the approach achieves statistical guarantees?
5. How efficient the approach is to achieve a guaranteed recall target?
6. Is the approach sensitive to pre-trained model?

Timeline

By 04/13: we collect data and models for evaluation
By 04/20: we finish the algorithm and artifact for evaluation
By 04/27: we finish the evaluation on the algorithm

Task division

Yuxuan Zhu: algorithm + evaluation
Chengsong Zhang: deliverable

Reference

[1] Thirumuruganathan, Saravanan, et al. "Deep learning for blocking in entity matching: a design space exploration." *Proceedings of the VLDB Endowment* 14.11 (2021): 2459-2472.
[2] Li, Yuliang, et al. "Deep entity matching: Challenges and opportunities." *Journal of Data and Information Quality (JDIQ)* 13.1 (2021): 1-17.
[3] Kang, Daniel, et al. "Approximate Selection with Guarantees using Proxies." *Proceedings of the VLDB Endowment* 13.11.