Master thesis proposals

For more information, please contact Professor Sonia Bergamaschi

sonia.bergamaschi@unimore.it

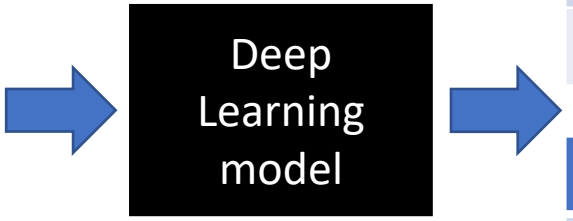
Master thesis proposal: Deep Learning for Data Transformation

- Create a deep learning model by using pre-trained transformers such as BERT to standardize dirty records to create golden records;
- Standardized records improve the performance of entity resolution algorithms

Dirty records

ID	Name	Surname	Location
1	Thomas	Jones	Hills street
2	Richie	William	Main street 9

ID	Nominative	Address
3	Tom Jones	Hill St
4	Rick Williams	Main street 9



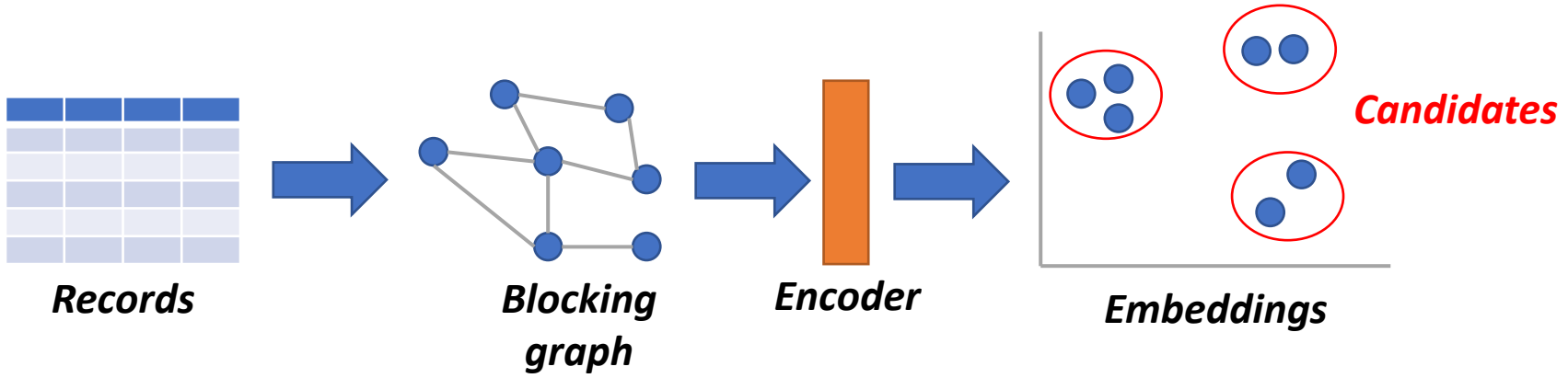
Golden records

ID	Name	Surname	Location
1	Thomas	Jones	Hills street
2	Richard	Williams	Main street 9

ID	Nominative	Address
3	Thomas Jones	Hills street
4	Richard Williams	Main street 9

Master thesis proposal: using graph neural networks to create embeddings for finding similar entity profiles

- Meta-blocking generates a graph by using the blocks information;
- We can use this graph to generate embeddings for every entity profile by using graph embeddings techniques;
- The embeddings can be used to find similar entity profiles (i.e. candidate matching pairs) in an efficient way;
- The approach can be extended by enriching the graph with other information.
- References:
  - Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2020, June). Creating embeddings of heterogeneous relational datasets for data integration tasks. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 1335-1349).
  - Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD*



More topics for master thesis

- Data Preparation
- Active Learning
- Transfer Learning
- Single-Label Classifiers
- Entity Resolution On-Demand

(see the dedicated descriptions in the following slides)

If interested, feel free to contact us for ideas or more details

State-of-the-art ER frameworks

State-of-the-art frameworks for Entity Resolution rely on:

- **Machine Learning (ML)**
  (e.g., *Magellan*)
  - A. Doan et al. (2020). Magellan: toward building ecosystems of entity matching solutions. *Communications of the ACM*, 63(8), 83-91.
- **Deep Learning (DL)**
  (e.g., *DeepMatcher*)
  - S. Mugdal et al. (2018). Deep Learning for Entity Matching: A Design Space Exploration. SIGMOD 2018: 19-34.

How to deal with the **need for labeled data** (requiring a significant **human effort**) to train the models?

New directions for ER

- **Transfer Learning**
  Storing knowledge gained while solving a problem and applying it to a different but related problem (**pre-trained EM models**)
  - Y. Li et al. (2020). Deep Entity Matching with Pre-Trained Language Models. *PVLDB* 14(1), 50-60.
- **Active Learning**
  The learning algorithm interactively queries a user or a source (**oracle**) to label dynamically collected ambiguous examples in order to refine the learned model (classifier) upon them
  - V. Meduri et al. (2020). A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching. SIGMOD 2018: 1133-1147.
  - **G. Simonini et al. (2021). The Case for Multi-task Active Learning Entity Resolution. SEBD 2021: 363-370.**
  - …and many more papers
- **Single-Label Classifiers**
  Train the classifier using only matching/non-matching pairs
  - R. Wu et al. (2020). ZeroER: Entity Resolution using Zero Labeled Examples. SIGMOD 2020: 1149-1164.

Limitations of batch ER

Traditional (batch) ER pipeline:

- Perform ER offline on the entire data
- Run the queries on the clean data

This represents a significant **waste of time, resources, and money** (e.g., pay-as-you-go contracts in the cloud) if our data changes frequently and/or if the data scientist is interested in just a portion of the data (this interest can be expressed by using a query; e.g., data exploration)

New approaches to ER

- **Query-Driven ER**: ER on the portion of dataset effectively useful to answer the query
  - H. Altwaijry et al. (2017). QDA: A Query-Driven Approach to Entity Resolution. *TKDE* 29(2), 402-417.
- **Progressive ER**: maximize the retrieval of matches when limited time and/or computational resources are available (driven by matching probability)
  - **G. Simonini et al. (2018). Schema-Agnostic Progressive Entity Resolution. ICDE 2018: 53-64.**

Entity Resolution On-Demand: perform ER only on the portion of dataset useful to answer the query and return the entities appearing in the result as soon as they are obtained according to the priority defined through the query itself (*early stopping*)