Domain Independent Entity Resolution - Team A

Jan-Felix Schneider, Serena Zhang, Julie Yu, Rakshita Nagalla, Charissa Ding
Andre Bittencourt , Felipe Penha (Neoway)
Andreas Mueller (Columbia University)

Entity Resolution – An Ubiquitous Task in Data Management

Entity resolution (ER), the task of identifying data entries that refer to the same real-world entities, is essential in data integration. Our team aims to develop a domain-agnostic, accurate, and scalable ER system to be offered as a service by Neoway to its business customers.

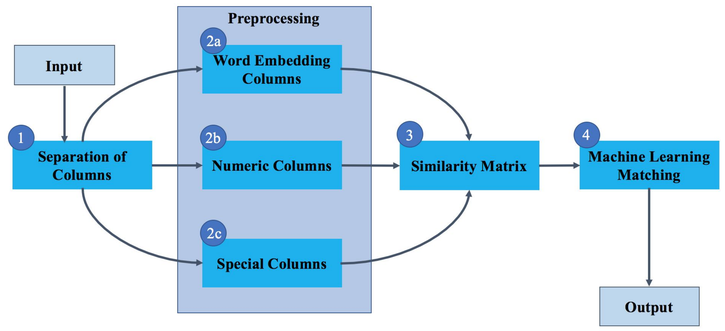

Figure 1. Overview of System Architecture

Challenges:

- Domain independent
- Faulty entries:
  - misspelling
  - abbreviations
  - missing values
- Large data size (20B rows)
- Language independent

System Architecture

Figure 1 summarizes our system into the following 4 steps:

Step 1: Separate input data into 3 types of columns (Figure 2)

Step 2: Preprocess respectively (e.g. normalize address column w/ Google API)

Step 3: Calculate pairwise similarities (Figure 2)

Step 4: Train a classification model (e.g. Random Forest) to obtain match likelihood

Word Embedding Approach

To extract valuable information from text fields such as product description, our team leveraged the word embedding approach and experimented several aggregation methods to obtain the embedding feature. (Figure 3)

	Firmographic (Neoway)	Publication (ACM-DBLP)	E-Commerce (Google-Amazon)	Methods and Similarity Metrics
Special Columns	Company Name ; Street Address; Phone Number; Email	Author Name; Conference Name; Paper Title	Product Name; Manufacturer	Levenshtein; Jaccard; Jaro-Winkler
Word Embedding Columns	Company Name; Street Address; CIty; State	Author Name; Conference Name; Paper Title	Product Name; Product Description; Manufacturer	Glove; Word2Vec; fastText Cosine Similarity
Numeric Columns	Zip Code	Publish Year	Price	Scaled Gaussian; Min-Max Ratio

Figure 2: Examples of Column Types and Similarity Metrics

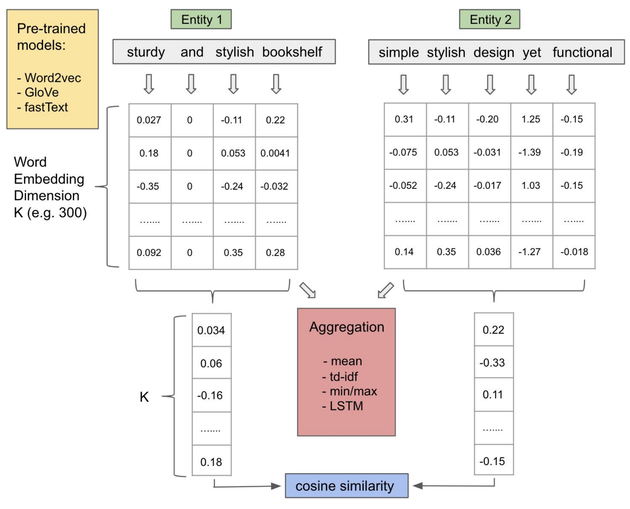

Figure 3: Word Embedding Approach for Long Texts

Results on different domains (Firmographics, E-Commerce, Publications)

- ML Models: Random Forest, Decision Tree, and SVM
- Word Embedding Models: Word2Vec, GloVe, fastText
- Word Embedding Aggregation: Average, tf-idf, Min and Max
- Best Results: Random Forest + Word2Vec + tf-idf

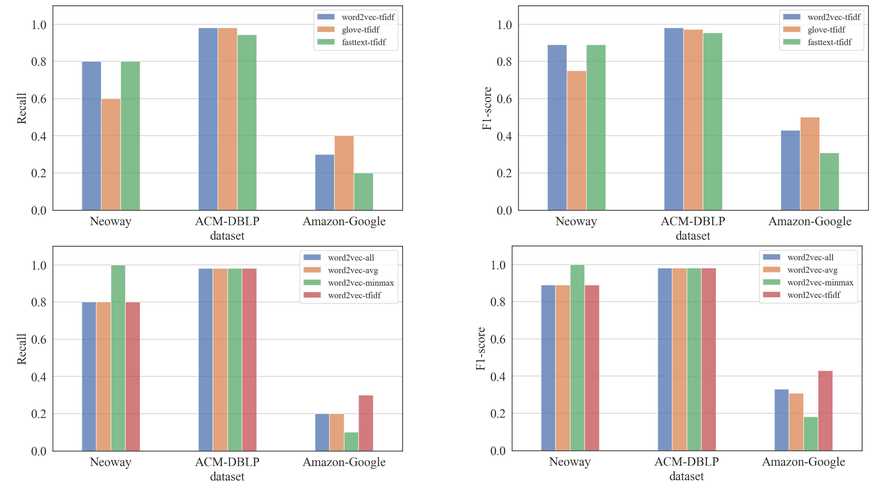

Figure 4. Evaluations on Datasets from Different Domains

Note: Results above were achieved on a representative subsample of the datasets.

Engineering

From an engineering perspective, we implemented our system with modular design and the team members collaborated on development leveraging git and unit testing.

Next Steps

- Deep Learning Approach (Figure 5):
  - We have a working model implemented in PyTorch and it's in development for evaluation
- Scalability:
  - We are in development for a Pyspark implementation for the prediction step.

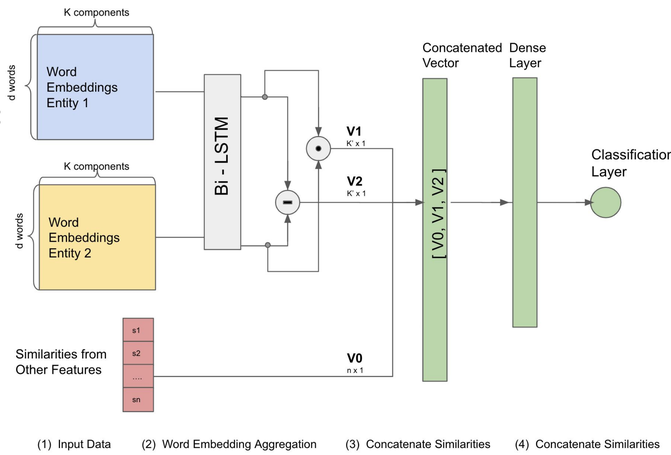

Figure 5. Deep Model Architecture (LSTM as Aggregation of Word Embeddings)

Acknowledgments

We would like to thank Neoway for providing their data and advice to our project, as well as Andreas Mueller for his helpful feedback and suggestions on our work.

References

[1] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems", 2010

[2] Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration", 2018