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.

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

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

System Architecture

Figure 1 summarizes our system into the following 4 steps:
1. Step 1: Separate input data into 3 types of columns (Figure 2)
2. Step 2: Preprocess respectively (e.g. normalize address column w/ Google API)
3. Step 3: Calculate pairwise similarities (Figure 2)
4. 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)

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