Supervised Contrastive Learning for Neural Blocking

Team WBSG – Winner of the ACM SIGMOD Programming Contest 2022
Alexander Brinkmann, Ralph Peeters, Christian Bizer (Advisor)
Data & Web Science Group @ University of Mannheim
{alexander.brinkmann, ralph.peeters, christian.bizer}@uni-mannheim.de

Task Overview

Task: Build a ***blocking system*** for Entity Resolution (ER) on two product datasets to quickly filter out obvious non-matches and obtain a much smaller candidate set of tuple pairs.

DS	Description	Number of Rows	Allowed Candidate Pairs	Training records	Positive Training Pairs	Training Clusters
D1	Notebook Specifications	About 1.000.000	1.000.000	1.661	2.815	718
D2	Multilingual Product Specifications	About 1.000.000	2.000.000	2.006	4.393	680

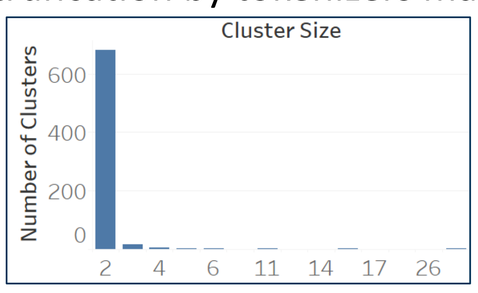
Target Metric: Average recall of both candidate sets

Evaluation Environment: 16 CPU x 2.7 GHz, 32GB main memory, 32 GB storage - no GPU

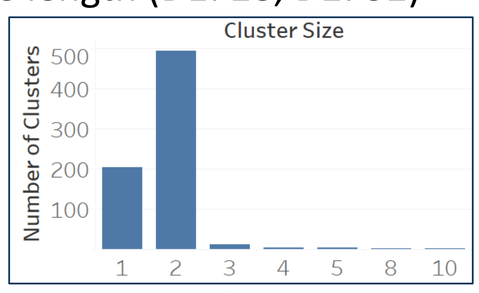
1. Pre-processing

Goal: Reduce ambiguity of the used attributes (D1:title, D2:name)

Techniques: Lowercasing, stop word removal, normalization, truncation by tokenizers max. sequence length (D1: 28, D2: 32)



Records per Cluster – Train D1



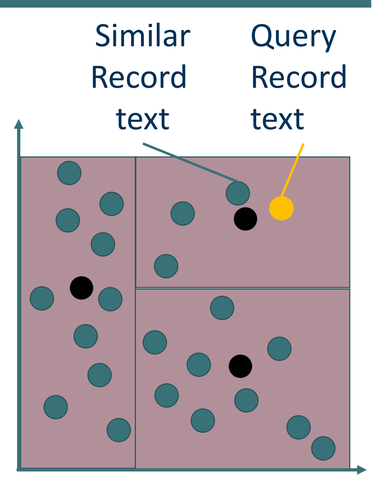
Preprocessed unique record texts per Cluster – Train D1

2. Embedding/ 3. Indexing/ 4. Retrieval

Goal: Embed and index all unique record texts & find top-k neighbors for each record text with min. cosine similarity

Embedding: Transformer with contrastive training using training data plus additional training data from the Web (see right column)

Indexing: Train FAISS index on subset of embeddings to partition embeddings space using Voronoi cells & compress vectors using product quantization, index embeddings



5. Re-ranking

Goal: Increase generalization of top-k similarity rankings

Technique: Calculate Jaccard similarity of candidate pairs & re-rank candidate pairs using the average of Jaccard similarity & max. approx. cosine similarity of the embeddings

6. Pair Generation

Goal: Generate candidate pairs from most similar record texts in cluster until the allowed number of candidate pairs is reached

Unique Record Text	Records
4gb hp elitebook folio 9470m 14 i5 3427u windows	1,2
hp c6z61ut elitebook folio 9470m 14 ultrabook	4,5

1. Records with the same pre-processed record text are assigned as pairs
   - Pairs: (1,2), (4,5) – similarity: 1.0
2. From two neighboring record texts all candidate pairs are created
   - Pairs: (1,4), (1,5), (2,4), (2,5) – similarity of record texts

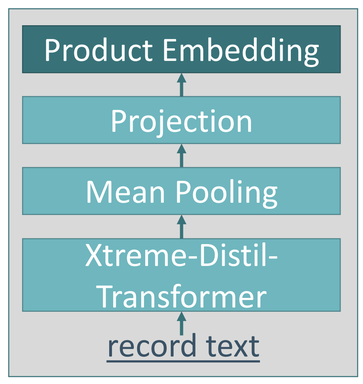
Neural Blocking Pipeline

1. **Pre-processing:** Pre-process records to reduce ambiguity
2. **Embedding:** Embed records using a fine-tuned transformer
3. **Indexing:** Index embeddings using FAISS
4. **Retrieval:** Retrieve similar records
5. **Re-ranking:** Re-rank similar records
6. **Pair Generation:** Generate final candidate pairs

Neural Architecture

Goal: Generate short but distinctive embeddings of records given restricted compute resources
(used attributes D1: title, D2: name)

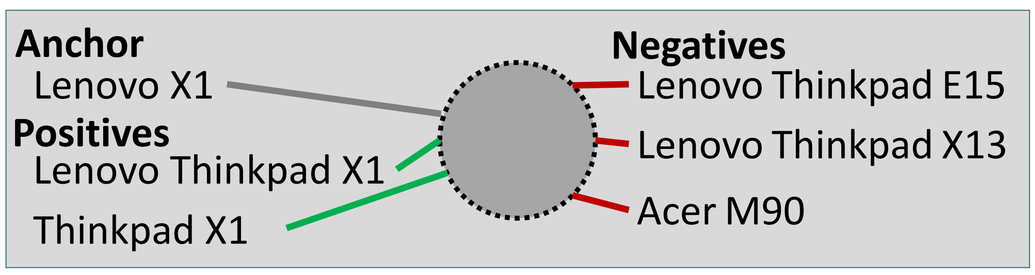
Architecture: Combine an Xtreme-Distil-Transformer (encoder) with mean pooling and a dense projection layer (32 dimensions)



Supervised Contrastive Learning

Goal: Learn representations such that matching records are close to each other and non-matching ones are far apart.

Approach: Large **batch size of 1024** records results in many in-batch distance comparisons during contrastive training.



Additional Training Data

The training sets are extended with additional computer offers from the WDC Product Corpus

- Number of product offers: 437.581
- Number of unique products : 286.356

Results

Average Recall	Recall D1	Recall D2	Runtime (s)
0.529	0.713	0.345	1914.275

Conclusions

- Neural Blocking without GPUs was possible for both SIGMOD Programming Contest 2022 datasets
- Pre-processing & grouping by unique record text reduced the runtime of indexing, retrieval & re-ranking
- Supervised contrastive training had a positive influence on the model's performance
- Additional training data combined with contrastive training strongly improved the model's performance on dataset D1
- Re-ranking with Jaccard Coefficient increased the generalization of the top-k similarity ranking
- The characteristics of products in D2 remain unknown, which makes it difficult to explain the performance difference on the two data sets

SupCon Learning for Neural Blocking:
https://github.com/abrinkmann/acm_sigmoid_2022_challenge

UNIVERSITY OF MANNHEIM
Data and Web Science Group