Enhancing Semantic Search Accuracy for Pharmaceutical Invoice Data Mapping: A Two-Stage Engine with Cross Encoders, Bi-Encoders, and Filtering Techniques

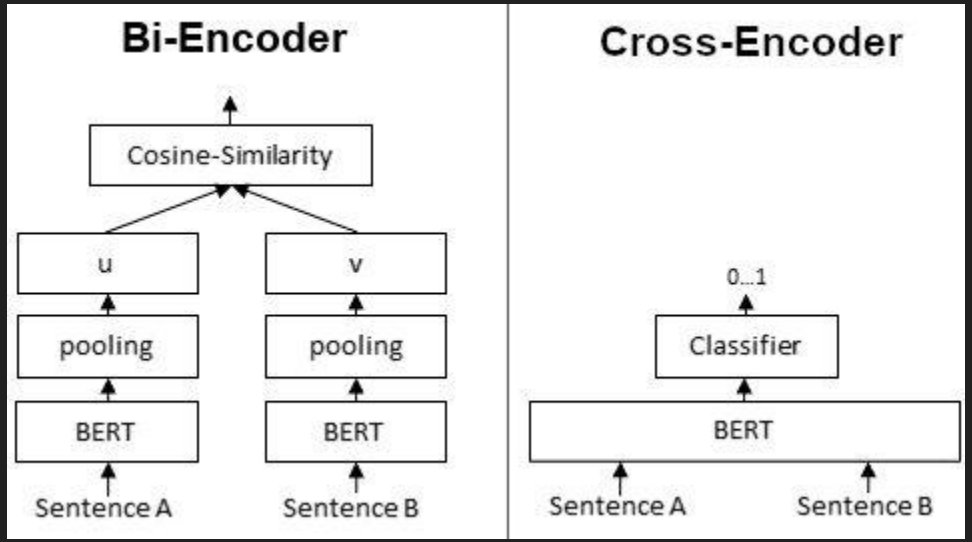
Ishan Pandey, Mehul Singh, Keshav Goel, Aman Sharma, Ayan Pandey, and Ronit Aryan

Introduction

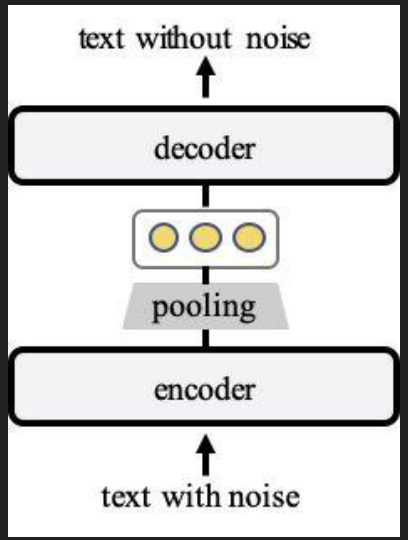
The pharmaceutical industry generates a vast amount of data, including invoice data that requires accurate mapping for various purposes, such as inventory management, financial accounting, and regulatory compliance.

Mapping pharmaceutical invoice data is a challenging task due to the complexity of the data, including varying product names, units of measurement, and numerical values. The accuracy of the mapping process is crucial as false positives can have severe consequences.

This paper proposes a two-stage methodology using deep learning architectures, including cross encoders, Bi-encoders, and SpanBERT, to enhance the accuracy of semantic search in pharmaceutical invoice data mapping. The methodology involves a base semantic search engine and a filtering engine, aiming to minimize false positives and improve overall accuracy.


Bi-Encoder
Cosine-Similarity
u
v
pooling
pooling
BERT
BERT
Sentence A
Sentence B
Cross-Encoder
0...1
Classifier
BERT
Sentence A
Sentence B

- The proposed methodology uses a two-stage process involving a base semantic search engine and a filtering engine, leveraging deep learning architectures like cross-encoders, Bi-encoders, and SpanBERT to minimize false positives and improve accuracy.
- **Bi-encoders** use two separate encoders to encode the query and document, while cross-encoders combine them into a single input sequence to capture more fine-grained interactions.
- **TSDAE** is a self-supervised pre-training method that leverages the denoising autoencoder framework with a transformer architecture to learn meaningful representations of text.
- **SpanBERT** is a variant of the BERT model designed for tasks that involve predicting or extracting spans of text from a given input. It incorporates additional pre-training objectives, such as span-based masked language modeling and span boundary prediction, to learn better representations of text spans.


text without noise
decoder
pooling
encoder
text with noise

TSDAE

NOVELTY

- Novelty: Addresses unique challenges in mapping pharmaceutical product data, focuses on minimizing false positives

- Contextual Understanding: Uses cross encoders and Bi-encoders, leverages Augmented SBERT technique for limited data

- Minimizing False Positives: Incorporates SpanBERT model in filtering engine, refines base semantic search engine results

- Adaptable & Scalable: Easily adapted to different datasets and scenarios, versatile for various applications in pharmaceutical industry and other domains

- Conclusion: Combines advanced deep learning architectures and specialized filtering engine for accurate and reliable mapping

Methodology

Dataset Preparation

1.1. Small Dataset Mapping

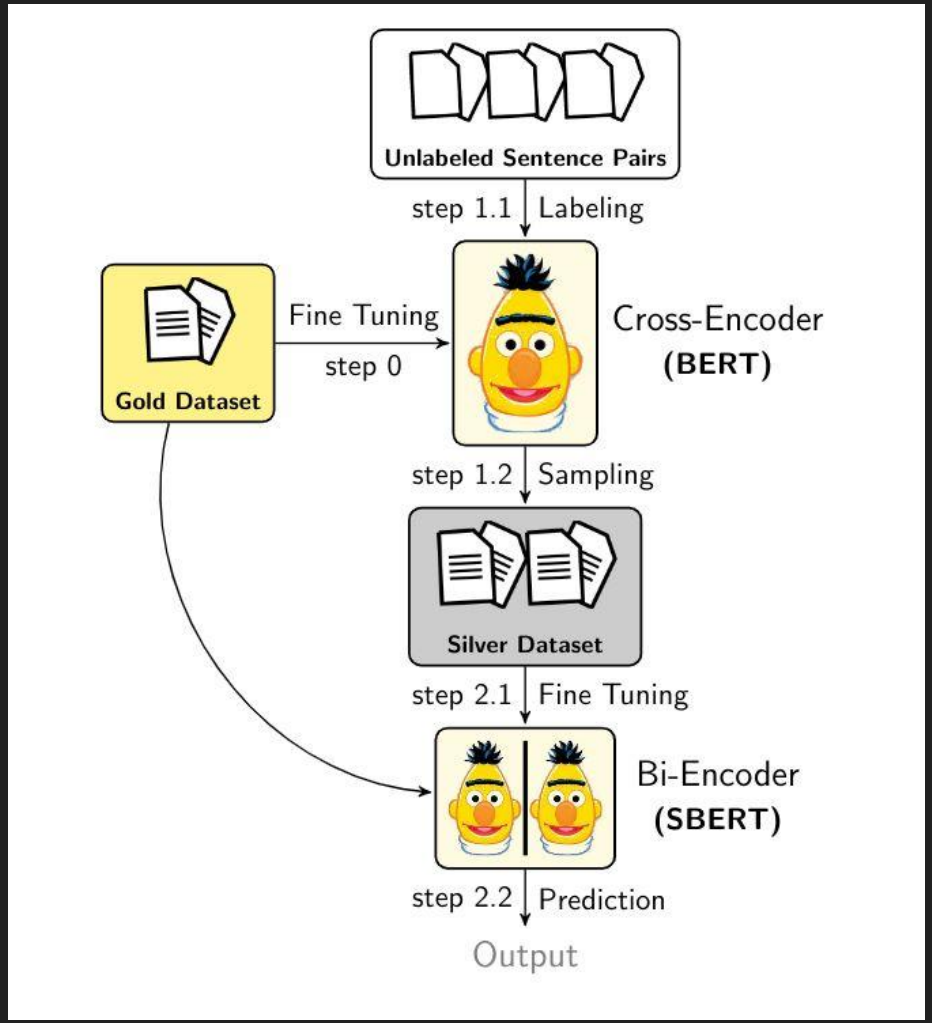
The initial dataset is composed of a small collection of semantically rich data, which is mapped to create meaningful relationships between the data points.

1.2. Augmented SBERT for Large Dataset Labeling

To generate a larger dataset, we employ an effective data-augmentation strategy known as Augmented SBERT. This approach utilizes a high-performing and slow cross-encoder (BERT) to label a larger set of input pairs, thereby augmenting the training data for the Bi-encoder (SBERT).

Methodology

The research proposes a two-stage process, including a base semantic search engine and a filtering engine, to improve the accuracy of semantic search in pharmaceutical invoice data mapping. The approach involves a dataset preparation, base semantic search engine development, and filtering engine design, utilizing state-of-the-art deep learning architectures such as cross encoders, Bi-encoders, and SpanBERT. The filtering engine filters false positives by identifying the overlap of words, training a SpanBERT model, and comparing extracted features from the label and target during the inference process.


Unlabeled Sentence Pairs
step 1.1 Labeling
Gold Dataset
Fine Tuning
step 0
Cross-Encoder
(BERT)
step 1.2 Sampling
Silver Dataset
step 2.1 Fine Tuning
Bi-Encoder
(SBERT)
step 2.2 Prediction
Output

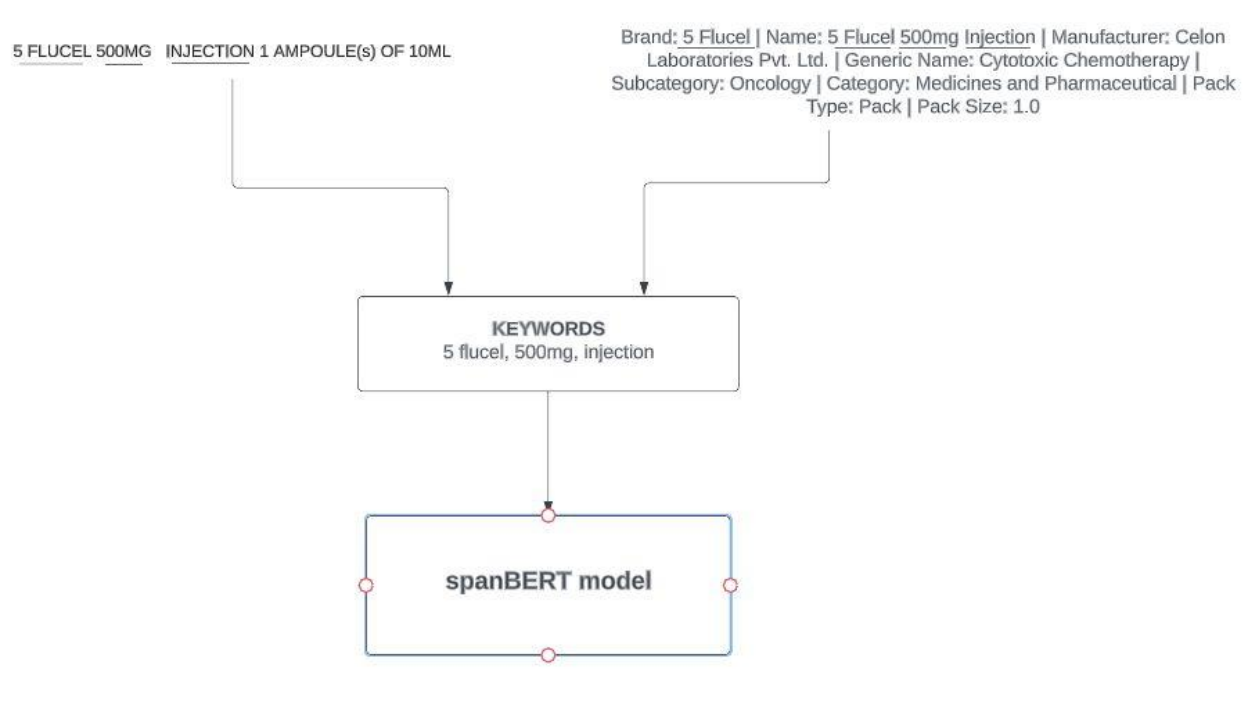
Base Semantic Search Engine

2.1. Cross Encoder Training

A cross encoder is trained with the small dataset using Multiple Negatives Ranking (MNR) Loss to improve the model's ability to rank semantically similar pairs. The deduplicate data loader is utilized to avoid redundancies in the dataset during the training process.

2.2. MPNet Model Fine-tuning

The MPNet model is initially fine-tuned using the Transformer-based Denoising AutoEncoder (TSDAE) method to pretrain the model with a denoising objective. Following this, the model is fine-tuned in a supervised manner using the labeled large dataset and MNR Loss to enhance its semantic search capabilities.


5 FLUCEL 500MG INJECTION 1 AMPOULE(s) OF 10ML
Brand: 5 Flucel | Name: 5 Flucel 500mg Injection | Manufacturer: Celon Laboratories Pvt. Ltd. | Generic Name: Cytotoxic Chemotherapy | Subcategory: Oncology | Category: Medicines and Pharmaceutical | Pack Type: Pack | Pack Size: 1.0
KEYWORDS
5 flucel, 500mg, injection
spanBERT model

3.1. Overlap Identification

The filtering engine identifies the overlap of words present in both the label and target, such as quantities (ml, l, mcg, mg) and associated numerical values. This process aids in the extraction of relevant features from the modified dataset.

3.2. SpanBERT Model Training

A SpanBERT model is trained to extract terms from the modified dataset
based on the identified overlaps. This model is utilized during inference to extract features from both the label and target.

3.3. Feature Matching and Filtering

The filtering engine compares the extracted features from the label and target during the inference process. A prediction is passed only when the extracted features match, minimizing the occurrence of false positives in the search results.

By integrating the base semantic search engine with the filtering engine, this methodology aims to provide a comprehensive solution for enhancing the accuracy and relevance of semantic search results.

Advantage of contextualized sentence representations:

- Best performance: Contextualized embedding methods for labels

- Comparison: BERT+USE and BERT+Sent BERT outperformed BERT+Sent2Vec and BERT+GloVe

- Reason: Transformer-based encoding with self-attention generates superior representations

- Ideal models: Sentence BERT and Universal Sentence Encoder for retrieval-based tasks, pre-trained on semantic text similarity (STS) tasks

Conclusion

In our evaluation, we compared our proposed method with several baseline models, including traditional information retrieval-based approaches, semantic search approaches using pre-trained sentence embeddings, and supervised models combining pre-trained sentence embeddings with additional fine-tuning on a small dataset. Our proposed method outperformed all baseline models across all evaluation metrics, with a Recall of 0.96 and a Precision of 0.93.

We also found that the inclusion of a specialized filtering engine designed to minimize false positives significantly improved the performance of our method, highlighting its potential as a valuable solution for accurate and reliable data mapping in the pharmaceutical industry.