Heidelberg University
Institute of Computer Science
Summer Term 2023
Seminar: Modern Information Retrieval
Instructors: Prof. Dr. Michael Gertz
Nicolas Reuter
Ashish Chouhan
Jayson Salazar
John Ziegler

Seminar Paper

Dense Retrieval with Entity Views

Name:	Johannes Gabriel Sindlinger
Student ID:	3729339
Study Programme:	Master, Computer and Data Science (3rd semester)
Email:	johannes.sindlinger@stud.uni-heidelberg.de
Date of Submission:	02/08/2023

I, **Johannes Gabriel Sindlinger**, hereby certify that I have written the paper entitled **Dense Retrieval with Entity Views** in the seminar **Modern Information Retrieval** in the **Summer Term 2023** with **Prof. Dr. Michael Gertz** and **Nicolas Reuter**, **Ashish Chouhan**, **Jayson Salazar** and **John Ziegler** independently and only with the aids indicated within the work. I have clearly marked citations as well as the use of external sources, texts and aids according to the rules of scientific practice.

I am aware that I am not allowed to claim other people's texts and text passages as my own and that a violation of this basic rule of scientific work is considered an attempt to deceive and cheat, which will result in appropriate consequences. These consist of the grading of the examination performance with „not sufficient“ (5,0) as well as further measures if necessary.

Furthermore, I confirm that this work has not been presented in the same or similar form in any other seminar.

Heidelberg, 02/08/2023 ________________

Contents

1 Introduction

In the field of information retrieval, finding relevant information efficiently and effectively from large collections has been a challenge for a long time. This task involves finding documents or passages that are most relevant to a particular query or topic. Conventional methods of information retrieval often rely on sparse representations such as BM-25 (Robertson and Zaragoza [17]), which can miss subtle semantic connections and fail to fully capture the context of the information being searched for.

To address these limitations, dense retrieval has emerged as an alternative approach, utilizing dense vector embeddings of documents and queries. These embeddings, often generated using large language models such as BERT (Devlin et al. [5]), encode richer semantic information, enabling a more comprehensive understanding of content. Two main approaches to dense retrieval are the bi-encoder and cross-encoder methods. The bi-encoder approach creates separate vector representations for queries and documents, while the cross-encoder approach directly measures the similarity between the query and document, potentially yielding more accurate results at a higher computational cost.

Recently, information retrieval has evolved from solely ranking matching documents to a 'rich search' or 'Entity-oriented search' (Balog [2]). 'Rich search' implies that queries often require specific information about entities, facts, or structured data for more sophisticated tasks.

However, large language models like BERT suffer from an issue within this context: entities such as people, places, or organizations are not entirely represented in the models (Heinzerling and Inui [8]). Thus, large language models are likely to represent socially relevant entities that occur in frequent documents, but at the same time entities that are rather rare not.

As an example, the two authors refer to the training process of BERT model, which the well-known masked language model approach. During training phase, the model is charged to predict randomly masked tokens based on the context of the surrounding words. In the example, 'Paris is the capital of [MASK]' (Heinzerling and Inui [8]), 'France' is correctly predicted with a high probability, demonstrating the model's ability to recognize frequent entities. However, prediction might be ambiguous for rarer entities like 'Sesame

Street' in the case of 'Bert is a character on [MASK]' (Heinzerling and Inui [8]). In particular, entities that were not present during the training process and only emerged afterwards cannot be captured within language models at all.

This limitation motivated Tran and Yates [20] to explore solutions in their paper, which is the subject of this seminar report. The subsequent sections will present and critically examine Tran and Yates' work. We will contextualize the paper in current research in section 2, explain the methods used in section 3, present the results in section 4, and conclude with a critical assessment and personal opinion in section 5.

2 Related Work

Entities have been a subject of interest in the field of information retrieval since the emergence of knowledge repositories like Wikipedia. Research in this context can be divided into the time before and after the advent of large language models such as BERT [5].

2.1 Entities in Sparse Retrieval

Prior to the advent of large language models such as BERT [5], research focused on sparse retrieval methods like BM25, where entities from knowledge bases were used to enhance query understanding. Efforts were made to expand queries using Wikipedia descriptions of entities [24] and extract additional features like synonyms and relationships from linked knowledge bases [4]. Balog [2] provided a comprehensive meta-analysis of related research.

2.2 Entities in Dense Retrieval

While some research has already addressed the consideration of entities in sparse retrieval, the amount of work in the field of dense retrieval in this context is rather limited.

Some work focuses on integration of entities within interaction-based methods, treating textual and entity information as common inputs for ranking models. An example of this approach is the work of Xiong et al. [21, 23]: They developed a method to build ranking features by incorporating an attention mechanism of word embeddings and entity embeddings and could significantly outperform baselines for word-based and entity-based learning to rank systems.

Interaction-based methods such as these can be considered as extensions of the cross-encoder approach of dense retrieval and therefore potentially face the same issue of high computational complexity as described in section 1. All documents and query pairs must be processed at runtime, which can lead to significant time and resource overhead. This leads to a slow overall retrieval process and can be inefficient, especially with large data sets.

In the context of the broad field of natural language processing tasks, entities have also been considered in several research papers and have been shown to have a significant impact on its tasks (e.g.). The most prominent example is the ERNIE model (Sun et al. [19]), which incorporates knowledge from both pre-training tasks and external knowledge graphs, enabling it to achieve better contextual understanding and knowledge integration in natural language processing tasks.

The existing concepts of dense retrieval, including interaction-based methods and models like ERNIE, do not fully explore entities independently from the underlying large language model for the traditional search retrieval task. This represents the main novelty of Tran and Yates' work [20], which will be elaborated in the following sections.

3 Methodology

Tran and Yates' primary contribution lies in their approach to consider entities independently of the underlying large language model. To achieve this, they combine embeddings from an arbitrary large language model with embeddings of entities extracted from documents and queries. The embeddings of entities offer multiple views on the same information, with different sets of entities representing various perspectives on the queries and documents. Depending on the displayed view, different relevant documents or sections of documents can be identified.

Tran and Yates do not incorporate separate embeddings of queries and documents into a learned framework. Instead, they merge these embeddings into a joint vector space. This joint vector space results in a final embedding used to calculate similarities between vectors. Thus, Tran and Yates adopt the bi-encoder model (as discussed in section 1), allowing for indexing of document embeddings and enabling fast ranking computations through ANN search.

3.1 General Model

Tran and Yates' proposed method involves independent embeddings of documents and entities. The process entails merging the embeddings of both documents and entities for each query and document to achieve a joint vector space, as depicted in Figure 1. Following the bi-encoder approach, the embeddings for queries and documents are created independently using a pre-trained large language model.

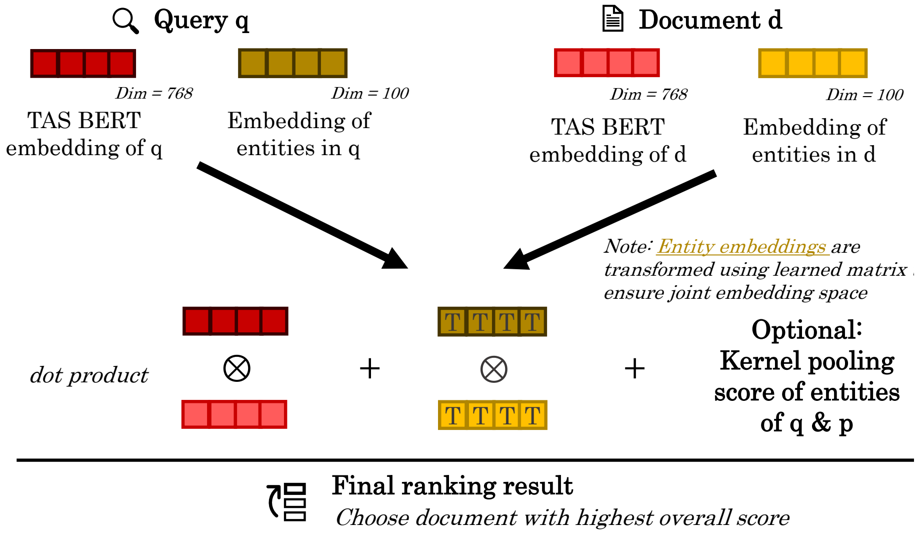

Query q
Dim = 768
TAS BERT embedding of q
Dim = 100
Embedding of entities in q
Document d
Dim = 768
TAS BERT embedding of d
Dim = 100
Embedding of entities in d
Note: Entity embeddings are transformed using learned matrix
ensure joint embedding space
T T T T
dot product
⊗
+
⊗
+
T T T T
Optional: Kernel pooling score of entities of q & p
Final ranking result
Choose document with highest overall score

Figure 1: General model

To maintain low computational costs without compromising effectiveness, Tran and Yates use a distilled version of BERT that was fine-tuned via the TAS BERT approach for information retrieval by Hofst"atter et al. [9].

The TAS BERT approach employs topic-aware sampling for fine-tuning on information retrieval task, thus queries in the training dataset are clustered based on different topics. During training, query samples are extracted exclusively from clusters with identical topics, aiming to enhance the sensitivity of the information retrieval model to different topics.

In the dense retrieval setting, embeddings of tokens are compared using similarity measurements to rank the best documents concerning a query. Tran and Yates follow a similar principle but enhance it by incorporating embeddings of entities. For each query and document, a single entity embedding is generated and concatenated with the textual embedding to create an embedding that captures both the semantic context and entity information.

In a usual dense retrieval setting, the generated embeddings of tokens are then compared using similarity measurements. When a query is submitted to a system, the ranking of the best documents with respect to the query is carried out, based on the calculated similarity score between the duets of the given query and all documents.

Experimental results demonstrate that concatenation produces superior results for combining text and entity embeddings compared to other methods as max pooling and sum pooling. Table 1 displays the respective analysis of Tran and Yates based on their experimental setup, which is introduced in detail in section 4. Apart from the analytical point of view, the approach of concatenation offers the advantage that it can be understood quite intuitively.

Operators	**MS MARCO Dev**		
	nDCG	**MRR**	**MAP**
Sum	0.393	0.335	0.339
Max	0.388	0.330	0.334
Concat	0.396	0.341	0.343

Table 1: Varying Aggregation Operators of Embedding Concatenation

However, concatenation introduces a challenge as the word embeddings and entity embeddings are generated in different vector spaces with varying dimensions and magnitudes. To address this issue, Tran and Yates introduce a transformation matrix $W \in \mathbb{R}^{100\times100}$ to transform the entity embeddings into the joint vector space.

So let $\mathbf{E}(t) \in \mathbb{R}^{100}$ be the embedding of entities of text t, the transformed entity embedding is given as:

$$\mathbf{R}_{entity}(t) = W^T \cdot \mathbf{E}(t) \tag{1}$$

The values of the matrix are determined during training process and thus reflect a meaningful transformation of the embeddings into a joint vector space.

Given a text t with its corresponding word embedding $\mathbf{R}_{text}(t)$ derived from the pre-trained language model, the final vector representation of t is obtained by concatenating the transformed entity embedding $\mathbf{R}_{entity}(t)$ and the word embedding:

$$\mathbf{R}_{final}(t) = \mathbf{R}_{text}(t) \oplus \mathbf{R}_{entity}(t) \tag{2}$$

The ranking score value of a pair of query q and document d is then obtained using the dot product $\otimes$ as a similarity measure:

$$\begin{aligned} \mathbf{Score}(q, d) &= (\mathbf{R}_{final}(q) \otimes \mathbf{R}_{final}(d)) && (3) \\ &= (\mathbf{R}_{text}(q) \otimes \mathbf{R}_{text}(d)) \oplus (\mathbf{R}_{entity}(q) \otimes \mathbf{R}_{entity}(d)) && (4) \end{aligned}$$

As an optional addition to their model, Tran and Yates include an external scoring source called KNRM signal, which measures the relationship between query entities and documents. This interaction-based approach requires knowledge of the query and document at runtime. For details, consult following subsection 3.2.

Incorporating the KNRM signal S_{knrm} into the final ranking score yields the score with KNRM signal, denoted by $\mathbf{Score}_{knrm}(q, d)$:

$$\mathbf{Score}_{knrm}(q, d) = (\mathbf{R}_{final}(q) \otimes \mathbf{R}_{final}(d)) + S_{knrm} \tag{5}$$

3.2 KNRM Signal

Tran and Yates introduce the KNRM signal, originally developed by Xiong et al. [22], as an additional scoring mechanism to extend their basic approach by a separate framework. Unlike its original use, Tran and Yates adapt the KNRM model to specifically capture the interaction between entities in queries and documents. The calculation of the KNRM signal follows these steps:

1. Let $X(q)$ be the set of all entity embeddings within query q, $X(d)$ any

set of entity embeddings which occur in document d. For EVA Single and EVA Single-QA models (see subsubsection 3.3.3) $X(d)$ contains all entities with the given document, for EVA Multi (see subsubsection 3.3.3) $X(d)$ contains only entities of specific subsets of entities of d. The entity interaction matrix is defined as:

$$T_{i,j} := sim(X_i(q), X_j(d)) \ ,$$

where $X_i(q)$ and $X_j(d)$ are the i-th and j-th embedding of q and d. The similarity function is given as cosine similarity. Embeddings are generated via Wikipedia2Vec (Yamada et. al [25]), as described in subsubsection 3.3.1.

2. Build k kernels using radial basis function, which creates differentiable histograms around given μ and σ^2.

$$K_l(X_i(q)) = \sum_{j=1}^{|X(p)|} \exp\left(-\frac{(T_{i,j} - \mu_i)^2}{2\sigma_i^2}\right)$$

3. Pool / Summarize the k results into a k-dimensional feature vector:

$$\overrightarrow{K(X_i(q))} = [K_1(X_i(q)), \ldots, K_k(X_i(q))]$$

4. Build kernel-pooled representation $\phi(T)$ by calculating log-sum for each query entity:

$$\phi(T) = \sum_{i=1}^{|X(q)|} \log \overrightarrow{K(X_i(q))}$$

5. Get final kernel pooling score by applying a learned ranking layer. Note that $\tanh(\cdot) \in (-1, 1)$ and therefore $\sup S_{knrm} = 1$:

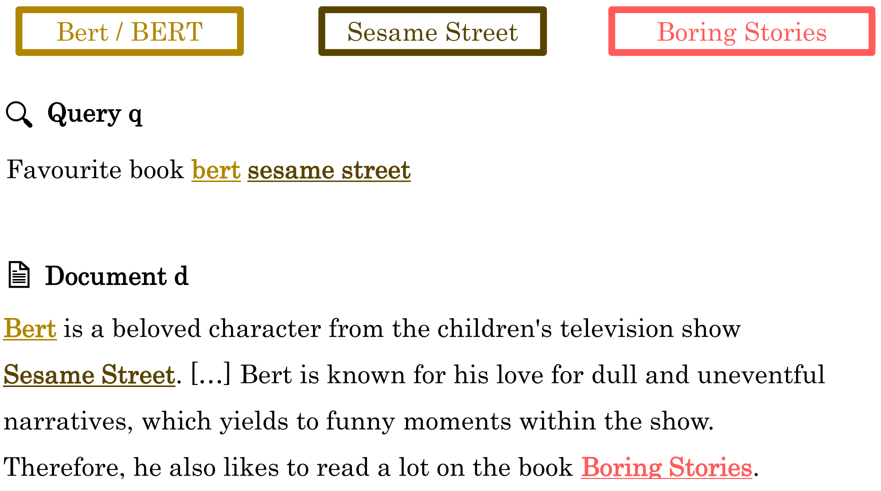
$$S_{\text{knrm}} = \tanh(w^T \phi(T) + b)$$

Despite being an interaction-based model that requires scoring during runtime for all query-document pairs, the computational complexity of the KNRM approach remains limited. The computations involved are relatively straightforward, and the additional learned layer does not significantly increase the computational overhead. Furthermore, these computations can be performed in parallel with the other components of Tran and Yates' approach. Empirical results concerning the latency of the models, with and without the KNRM signal, validate this assertion (see section 4).

3.3 Generating Entity Embeddings

As described in subsection 3.1, Tran and Yates create embeddings for both word tokens and entities in queries and documents. While the word token embeddings are generated using pre-trained TAS BERT, the primary focus of the contribution of Tran and Yates lies in the creation of entity embeddings. In particular, as queries and documents often contain multiple entities, these entities need to be aggregated into a single embedding.

To illustrate this process, consider the example depicted in Figure 2. In this example, the query 'Favourite book bert sesame street' corresponds to a document that mentions three entities: Bert, Sesame Street and Boring Stories.



Favourite book bert sesame street

Bert is a beloved character from the children's television show Sesame Street. [...] Bert is known for his love for dull and uneventful narratives, which yields to funny moments within the show. Therefore, he also likes to read a lot on the book Boring Stories.

Figure 2: Example query and example document

3.3.1 Extracting Entities

To aggregate multiple entities in a document or query, Tran and Yates first extract these entities from the text using external frameworks Dexter (Ceccarelli et al. [3]) and Wikipedia2Vec (Yamada et. al [25]).

- The Dexter framework is an entity linkage system that resolves entity mentions in text to corresponding entities in a knowledge base. In particular, Dexter employs a combination of methods, including named entity recognition and pattern matching, to perform this task.
- Wikipedia2Vec leverages the structure of Wikipedia to generate dense vector representations (embeddings) for words, articles, and entities. It utilizes the Word2Vec algorithm (Mikolov et al. [14]) to capture semantic relationships from Wikipedia, producing high-dimensional embeddings.

The extraction procedure involves submitting a document or query to Dexter, which extracts entity mentions from the given text. These entity names are then passed on to the knowledge base, which transforms them into embeddings. Wikipedia2Vec provides vectors in dimension 100 as default, Tran and Yates keep this value in their model. Figure 3 visualizes this process.

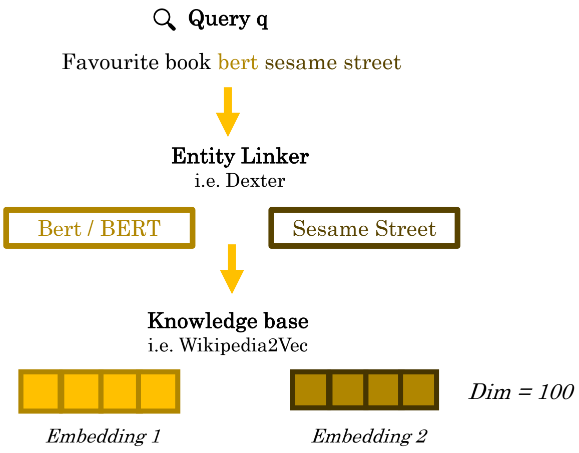


Figure 3: Process of entity extraction

3.3.2 Combining Entities for Queries

For now, multiple embeddings for each entity within a query or a document are generated by applying the procedure outlined in subsubsection 3.3.1. The aggregation of these generated embeddings differs based on whether a query or a document is considered, given the usual brevity of queries compared to documents.

For queries, Tran and Yates adopt a straightforward approach. They aggregate the embeddings of entities by averaging all entity embeddings across all dimensions. Figure 4 visualizes this procedure.

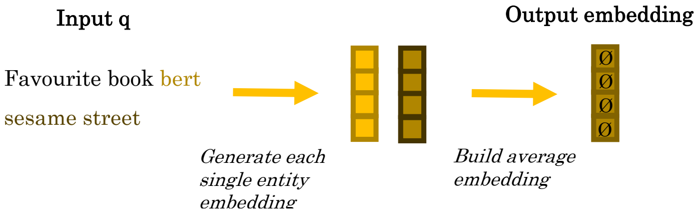


Figure 4: Process of generating a single entity embedding for a query

3.3.3 Combining Entities for Documents

Documents usually contain many different entities, as the example in Figure 2 indicates. Additionally, documents might also cover different aspects of a topic, so entities from very different areas might appear within them. To handle this additional complexity, the authors proposed three successive methods for entity aggregation in documents, namely:

- Single Entity Representation (EVA Single)
- Query-Aware Single Entity Representation (EVA Single-QA)
- Multiple Entity View Representation (EVA Multi)

The term 'EVA' stands for Entity Views in Dense Retrieval. The concept of 'Entity Views' which gives the paper its name, is particularly relevant in the third method, EVA Multi, and will be elaborated in the following sections.

EVA Single The key idea behind the Single Entity Representation is to extract all entities present in the document and subsequently generate a single output embedding by computing the average of these entity embeddings. This process is analogous to the one depicted in Figure 4 for queries, and a visual representation for documents is illustrated in Figure 5.

The initial approach for aggregating entities in a document, called EVA Single, is similar to the one used for queries. The key idea behind the Single

Entity Representation is to extract all entities present in the document and subsequently generate a single output embedding by computing the average of these entity embeddings. This process is analogous to the one depicted in Figure 4 for queries, and a visual representation for documents is illustrated in Figure 5.

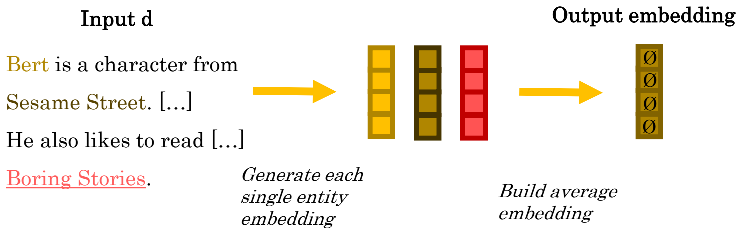


Figure 5: Process of generating an entity embedding for a document following the EVA Single approach

Despite its simplicity, this approach overlooks the fact that documents may cover diverse topics. By discarding query information, the EVA Single approach considers all entities, including those that may be only partially relevant to the document's main topic. Consequently, this disregards the relevance of entities within the ranking process, leading to biased results, as observed in section 4.

EVA Single-QA To address this problem, Tran and Yates introduce the Query-Aware Single Entity Representation, which creates embeddings that are tailored to the specific needs of a given query. However, this improvement comes with increased computational complexity since it assumes knowledge of the query beforehand. As a result, calculations for all query-document pairs must be performed during runtime, eliminating the possibility of precomputing document embeddings and indexing, leading to higher latency, as observed in section 4.

The underlying idea of the EVA single QA model is to filter the entities of a document based on the information of a given query and select only entities with high similarity to a query entity. Figure 6 visualizes this process.

The filtering is conducted using Algorithm 1, which is applied to each query-document pair (q, d). For every entity within q, the algorithm searches for the corresponding entity in d that possesses the maximum cosine similarity.

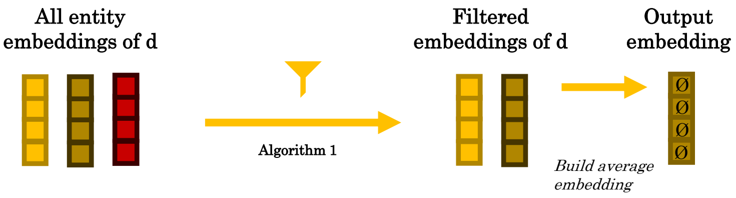


Figure 6: Process of generating an entity embedding for a document following the EVA Single-QA approach

If the similarity exceeds a specified threshold α, the respective entity in d is added to the filtered list $X_{focus}(d)$. Subsequently, the single output embedding for document d is computed as the average of all entity embeddings within $X_{focus}(d)$.

Algorithm 1 Query-aware document entity representation

```
Input: Query $q$ and document $d$, threshold $\alpha$
Output: Filtered entity embedding list $X_{focus}(d)$ of $d$
$X(q) \leftarrow$ set of embeddings of entities in $q$
$X_{focus}(d) \leftarrow \{\}$
for $e$ in $X(q)$ do
    $e^* \leftarrow$ entity embedding in $d$ having the maximum cosine similarity with $e$
    if cosine similarity$(e^*, e) > \alpha$ then
        $X_{focus}(d) \leftarrow X_{focus}(d) \cup \{e^*\}$
    end if
end for
return $X_{focus}(d)$
```

EVA Multi To address the issue of known queries required by the EVA Single-QA approach, Tran and Yates propose the EVA Multi method, which introduces multiple entity views as a solution with minimal negative consequences.

Analysis training data revealed that in the vast majority of instances, the number of entities in queries does not exceed two. For instance, in the MS Marco Dev dataset, 99.6 % of the 300,000 training instances and 99.5 % of the test instances contain two or fewer entities. Based on this observation, Tran and Yates focused their research on the assumption that it is sufficient to consider a maximum of two entities in queries.

Entities	Training Queries		Testing Queries	
	Count	Fraction	Count	Fraction
0	130,353	0.435	3,442	0.483
1	149,073	0.497	3,232	0.454
2	19,207	0.064	416	0.058
3+	1,367	0.004	37	0.005
Total	300,000		7,127	
Average	0.640		0.587	

Table 2: Summary statistics of the queries.

Under this assumption, when applying Algorithm 1, at most two entities remain in the filtered embedding list X_{focus}. This limitation stems from the algorithm iterating over the set of all entities in the given query once, resulting in X_{focus} containing either zero, one, or at most two items. The number of all possible sets eligible for X_{focus} is restricted to $|\{\}| + |X(d)| + \binom{|X(d)|}{2}$, where $|X(d)|$ corresponds to the number of entities in document d.

Leveraging this observation, Tran and Yates introduce clusters of entities, which represent different views of a document. In the EVA Multi approach, all possible single itemsets and sets of pairs of the entities are generated. However, sets of pairs are considered only if the cosine similarity between the two entities within a pair exceeds a predefined threshold β, ensuring only relevant entity views are generated. The final output embeddings are then calculated by averaging the embeddings of all items within each set. When applying the optional KNRM signal (see subsection 3.2) to this approach, the entity interaction matrix T is built only upon the set of entity embeddings within a single cluster, rather than considering all entities within the document.

This allows all calculations to be performed independently of query information, enabling indexing. Unlike the previous methods, EVA Single and EVA Single-QA, the EVA Multi approach generates several entity output embeddings per document.

Figure 7 visualizes the process of generating the embeddings of multiple entity views or clusters. For the given example of three entities within an input document, six different views on entities and document entity representations are generated.

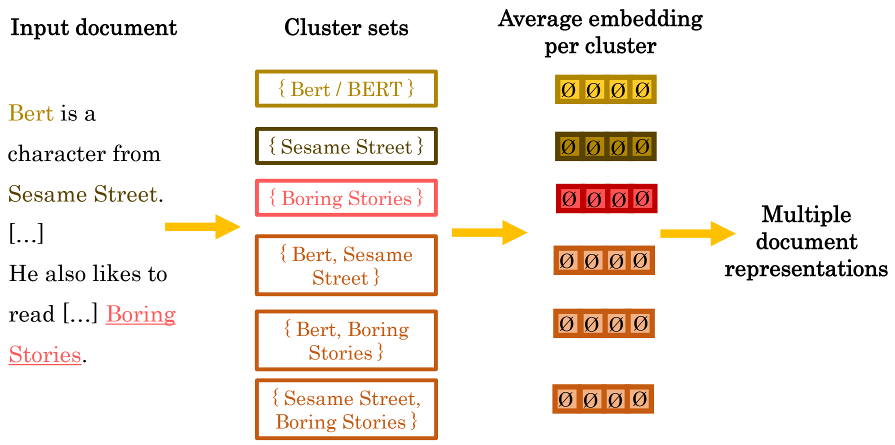

Input document
Cluster sets
Average embedding per cluster
Bert is a character from Sesame Street. [...] He also likes to read [...] Boring Stories.
{ Bert / BERT }
{ Sesame Street }
{ Boring Stories }
{ Bert, Sesame Street }
{ Bert, Boring Stories }
{ Sesame Street, Boring Stories }
Ø Ø Ø Ø
Multiple document representations

Figure 7: Process of generating entity embeddings for a document following the EVA Multi approach

4 Results

In this section, the concrete implementation of the described models in section 3 and the resulting results are presented. The focus of the elaborations is on the major findings of the work by Tran and Yates; accordingly, not all results are listed herein. Details can be studied in the paper by Tran and Yates [20].

4.1 Experimental Setup

Tran and Yates implemented their models in a TensorFlow (Abadi et al. [1]) setup: For the pretrained language model, they chose a distilled BERT (Sanh et al. [18]), fine-tuned it using TAS BERT approach (Hofstätter et al. [9]), as described in subsection 3.1. To determine the embeddings, they used Dexter (Ceccarelli et al. [3]) and Wikipedia2Vec (Yamada et. al [25]), as explained in subsection 3.3.

Training was conducted using pairwise hinge loss on four Quadro RTX 8000 GPUs in parallel, employing 300,000 training samples from the MS MARCO dataset (Nguyen et al. [15]). For the models allowing indexing (EVA Single and EVA Multi, see subsubsection 3.3.3), the end-trained model was used to index embeddings for documents. Evaluation was performed on 7,127 test samples from the MS Marco dataset and the datasets TREC Deep Learning

(DL) Track 2019 (MacAvaney et al. [12]), TREC DL 2020 (MacAvaney et al. [13]), TREC DL HARD (Yates et al. [26]). Evaluation metrics were chosen to be nDCG@10, MRR@10, MAP@1000.

The models proposed by Tran and Yates (EVA Single, EVA Single-QA, and EVA Multi, as described in subsubsection 3.3.3) were compared against various baselines as BM-25, TAS BERT, ERNIE and others. Details can be examined within the paper by Tran and Yates [20].

4.2 Efficiency

The evaluation of different models by Tran and Yates included both efficiency and effectiveness. Complete results for both aspects can be found in the paper by Tran and Yates [20]. In terms of efficiency, particular attention was paid to latency, and the results for all models under examination are presented in Table 3. The table displays the average search time per query across all evaluation datasets for each model, all executed on the same server.

Methods	**Latency (ms)**
Low latency (<100 ms)	
BM25	13
ANCE	25
ERNIE Tuned	29
ERNIE Multi	70
TAS BERT	28
EVA Single	40
EVA Multi	76
EVA Multi-KNRM	74
Higher latency (>100 ms)	
EVA Single-QA	2,039
EVA Single-QA-KNRM	3,839
BM25 + T5 (Zero-Shot)	5,052
Best Reported	-

Table 3: Analysis of effectiveness of EVA models and baselines

4.3 Effectiveness

The effectiveness results of the various models consistently hold across all combinations of evaluation metrics and datasets. To avoid redundancy, this report will provide exemplary results based on the nDCG@10 metric and the TREC DL Track 2019 dataset [12], which serve as representatives of the overall results. These exemplary results are illustrated in a visual format in Figure 8. Further details on the main outcomes can be found in the subsequent subsection 4.4.

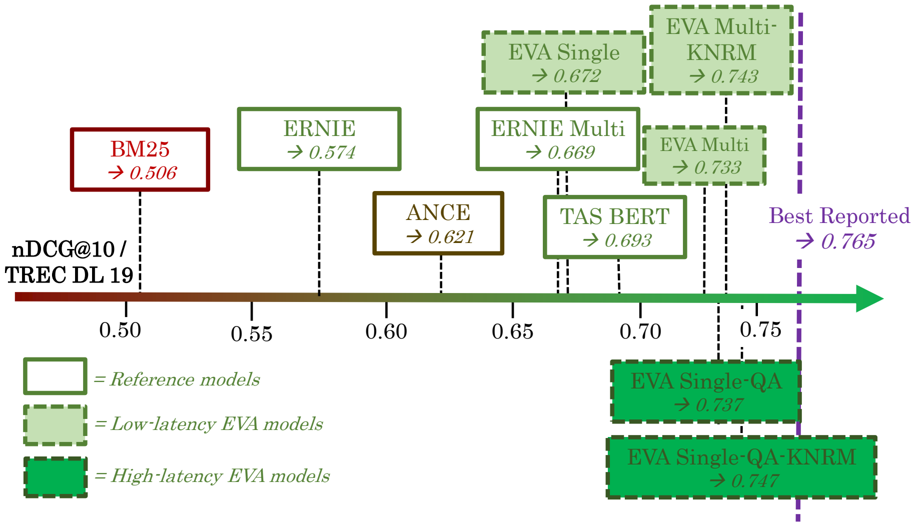


Figure 8: Exemplary results of EVA models and baselines for TREC DL 19 dataset and evaluation metric nDCG@10

4.4 Impact

The outcomes of Tran and Yates' work, based on the results from subsection 4.3 and subsection 4.2, can be summarized as follows:

1. **Enriching pre-trained language models with entity embeddings improve effectiveness significantly**: The effectiveness results demonstrate that the models proposed by Tran and Yates, namely EVA Multi and EVA Single-QA, both with and without KNRM signal, outperform the various baselines. In the example presented in Figure 8, the nDCG@10 value for the best performing baseline, TAS BERT, is 0.693. In contrast, the values for EVA Multi and EVA Single-QA are 0.733 and 0.737, respectively, representing an increase in effectiveness of 5.8 % (EVA Multi) and 6.3 %. Moreover, the models only marginally deviate from the best reported results.

2. **Multiple entity views increase performance to single view**:

The initial approach, EVA Single, which incorporates entities in documents without a specific focus on the query information, exhibits poor results and even underperforms the baseline TAS BERT. This is due to the inclusion of entirely irrelevant entities within documents during the retrieval process, leading to biased results (see subsubsection 3.3.3). However, when considering only entities relevant to the respective query, as in the cases of EVA Single-QA and EVA Multi, a significant increase in effectiveness is observed compared to all baselines, as explained earlier.

3. **KNRM signal provides slight improvement of effectiveness**: The comparison of the results for the EVA models with and without the additional KNRM signal shows slightly better performance for the models with the KNRM signal. This is evident in the exemplary results of Figure 8, where the EVA Multi KNRM model achieves an nDCG@10 value of 0.748, slightly higher than the value of 0.733 for the EVA Multi model. A similar observation can be made for the EVA Single-QA model. However, the effect is modest, as the EVA Multi approach outperforms the same approach with the additional KNRM signal in the case of the TREC DL 2020 dataset. In conclusion, the impact of introducing entity embeddings is more significant than that of introducing the KNRM signal.

4. **Removing known query assumptions has minor impact on effectiveness, but increases efficiency drastically**: The best effectiveness results are achieved by EVA Multi and EVA Single-QA, as described above. The crucial distinction between these models lies in the assumption of knowing queries at runtime for EVA Single-QA and the need for large language model inference at runtime (see subsubsection 3.3.3). This leads to substantial differences in latency compared to the EVA Multi model. As shown in Table 3, the latency for the EVA Multi models is 74 ms with KNRM signal and 76 ms without, whereas the latency for both EVA Single-QA models exceeds two seconds. In real-world scenarios, such prolonged latency times are impractical for an information retrieval system, as users typically expect faster results. However, given that the effectiveness results of EVA Multi and EVA Single-QA only differ marginally, the EVA Multi approach provides a

good compromise between efficiency and effectiveness.

5 Discussion & Criticism

Tran and Yates propose an approach that demonstrates the significant improvement of classical dense retrieval methods through the incorporation of entity information. The results, surpassing the baselines TAS BERT and ERNIE, support the effectiveness of their method. Several positive aspects and some limitations of their work can be identified.

5.1 Positives

- Satisfying results: As demonstrated in subsection 4.4, Tran and Yates' approach outperforms the baselines, justifying the introduction of their method.

- Simple and intuitive approach: The ideas and explanations provided by Tran and Yates are straightforward to follow. The introduced methods, from the general model (see subsection 3.1) to the algorithms used (e.g., algorithm 1), have a clear structure. For instance, the choice of aggregation method for embeddings of word tokens and entities, i.e. concatenation of embeddings (see Figure 1), is easier to understand than alternative approaches like max pooling or sum pooling. Based on that, the system could be extended with additional embeddings without much effort.

- Consideration of both effectiveness and efficiency: While many research efforts focus solely on achieving top results on leader boards, Tran and Yates' approach appears to be more oriented towards practical use and sustainability. The EVA Multi-approach offers both high effectiveness and efficiency.

5.2 Negatives

- Entities aren't universally relevant: Tran and Yates concentrate solely on the impact of entities in their work, but their analyses reveal that for many queries, entities play no significant role. As shown in Table 2, 43.5 % of all queries in the training data do not contain entities, causing the EVA models being ineffective for such queries.

- Lack of originality: The ideas presented by Tran and Yates are based on existing frameworks, and their contribution mainly lies in the composition of ideas from other authors. As a result, the approach, while intuitive, cannot be classified as groundbreaking. This is evident from the limited citations of their work so far, with only one citation in Kamphuis et al. [10].

5.3 Possible Extensions

The granular structure of Tran and Yates' models allows for possible extensions through the exchange or addition of individual components. For instance, the pre-trained language model used to calculate word-level embeddings could be extended beyond the two baseline models TAS BERT and ERNIE to other, more sophisticated models. In domain-specific use cases, a tailored choice of the pre-trained model might further enhance effectiveness. For example, in a biomedical context, BioBERT (Lee et. al [11]) could be considered.

The embedding component for entities could also be replaced or supplemented with other choices. For example, keyword embeddings, as presented by Gab'ın et al. [6], or structural information, such as that presented by Raman et al. [16], could serve as alternative sources for external embeddings. Furthermore, in the context of HTML files, Guo et al.'s approach [7] could be explored for generating embeddings.

These extensions could address some limitations of the current approach and potentially lead to further improvements in effectiveness and applicability in different domains.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Krisztian Balog. *Entity-oriented search.* Springer Nature, 2018.

[3] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, pages 17–20, 2013.

[4] Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 365–374, 2014.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Jorge Gabín, M Eduardo Ares, and Javier Parapar. Keyword embeddings for query suggestion. In *European Conference on Information Retrieval*, pages 346–360. Springer, 2023.

[7] Yu Guo, Zhengyi Ma, Jiaxin Mao, Hongjin Qian, Xinyu Zhang, Hao Jiang, Zhao Cao, and Zhicheng Dou. Webformer: Pre-training with web pages for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1502–1512, 2022.

[8] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. *arXiv preprint arXiv:2008.09036*, 2020.

[9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International*

ACM SIGIR Conference on Research and Development in Information Retrieval, pages 113–122, 2021.

[10] Chris Kamphuis, Aileen Lin, Siwen Yang, Jimmy Lin, Arjen P de Vries, and Faegheh Hasibi. Mmead: Ms marco entity annotations and disambiguations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2817–2825, 2023.

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[12] Sean MacAvaney, Andrew Yates, Sergey Feldman, Wei Guo, Yixing Hua, Tom Kenter, Federico Nanni, Bhaskar Mitra, Navid Rekabsaz, and Hamed Zamani. Overview of the trec 2019 deep learning track. In *Proceedings of The 28th Text REtrieval Conference (TREC 2019)*, 2019.

[13] Sean MacAvaney, Andrew Yates, Sergey Feldman, Wei Guo, Yixing Hua, Tom Kenter, Bhaskar Mitra, Federico Nanni, Navid Rekabsaz, and Hamed Zamani. Overview of the trec 2020 deep learning track. In *Proceedings of The 29th Text REtrieval Conference (TREC 2020)*, 2020.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119, 2013.

[15] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660, 2016.

[16] Natraj Raman, Sameena Shah, and Manuela Veloso. Structure and semantics preserving document representations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 780–790, 2022.

[17] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, March 2009. ISSN 1554-0669.

[18] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[20] Hai Dang Tran and Andrew Yates. Dense retrieval with entity views. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1955–1964, 2022.

[21] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 763–772, 2017.

[22] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.

[23] Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279, 2017.

[24] Yang Xu, Gareth JF Jones, and Bin Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66, 2009.

[25] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *arXiv preprint arXiv:1812.06280*, 2018.

[26] Andrew Yates, Sean MacAvaney, Bhaskar Mitra, Navid Rekabsaz, Hamed Zamani, Chenyan Li, Xiang Xu, Zhuyun Dai, Saptarshi Pal, Hui Fang, et al. Overview of the trec 2020 deep learning for hard informa-

tion retrieval (dlhard) track. In *Proceedings of The 29th Text REtrieval Conference (TREC 2020)*, 2020.