Exploring the Value of Multi-View Learning for Session-Aware Query Representation

Anonymous ACL submission

Abstract

Recent years have witnessed a growing interest towards learning distributed query representations that are able to capture search intent semantics. Most existing approaches learn query embeddings using relevance supervision making them suited only to document ranking tasks. Besides, they generally consider either user's query reformulations or system's rankings whereas previous findings show that user's query behavior and knowledge change depending on the system's results, intertwine and affect each other during the completion of a search task. In this paper, we explore the value of multi-view learning for generic and unsupervised session-aware query representation learning. First, single-view query embeddings are obtained in separate spaces from query reformulations and document ranking representations using transformers. Then, we investigate the use of linear (CCA) and non linear (UMAP) multi-view learning methods, to align those spaces with the aim of revealing similarity traits in the multi-view shared space. Experimental evaluation is carried out in a query classification and session-based retrieval downstream tasks using respectively the KDD and TREC session datasets. The results show that multi-view learning is an effective and controllable approach for unsupervised learning of generic query representations and can reflect search behavior patterns.

1 Introduction

Understanding user's search intent is central in information retrieval (IR). Modeling user's intent inevitably requires to capture search context. Search history is arguably the most salient facet of context that has been widely captured and used in previous work (Teevan et al., 2005; Dehghani et al., 2017; Aloteibi and Clark, 2020; Zhou et al., 2020). It mainly includes the following: (1) the previous user's queries, generally recorded into physical sessions (also called time-based sessions (Lucchese et al., 2011)) or task-based sessions (also called missions (Hagen et al., 2013)); (2) the retrieved documents that the user subsequently selects (e.g., based on clicks), among those retrieved by the IR system in response to her queries. Mining user's search intent from search history is challenging because of phenomena such as vocabulary mismatch between the query and documents, ambiguity issues since two queries even with slight lexical variations may underline different intents (Steiner, 2019; Sanderson, 2008), and topic change in user's search behavior which is particularly prominent while completing complex search tasks (e.g., exploratory and multi-step tasks (Hassan Awadallah et al., 2014; He and Yilmaz, 2017). To address these challenges, recent years have witnessed a growing interest in learning query representations to capture hidden syntactic and semantic relationships (Zamani and Croft, 2016; Grbovic et al., 2016; Bing et al., 2018; Zhang et al., 2019; Zhou et al., 2020). However, learning context-aware query embeddings faces two main issues: (1) user's query formulations included in the search sessions bring word contexts that do not extensively occur at the training phase in web search data (Keller and Lapata, 2003); (2) queries do not exhibit a clear structure as sentences. In most of previous work, query embeddings are learned based on search session contexts modeled from relevant or pseudo-relevant documents returned by the system (Zamani and Croft, 2016, 2017; Zhang et al., 2019). These methods are suited to supervised relevance ranking tasks with sufficient training data. Other methods learn distributed query representations based on user's query reformulations in the search session (Grbovic et al., 2016; Sen et al., 2018; Zhou et al., 2020). These methods are rather unsupervised and applicable to a wide range of downstream language processing tasks making them *generic*.

In this work, we explore the unsupervised problem of learning generic distributed query representations, able to support a wide range of downstream search tasks. As outlined recently, unsupervised representation learning for IR has not received much attention yet (Lin, 2021). This paper attempts to fill this gap by following a query oriented fashion. Specifically, we argue that by considering only one facet of the search session (i.e., documents vs. query reformulations) as done in Sen et al. (2018); Grbovic et al. (2016); Zamani and Croft (2016); Zhang et al. (2019); Zhou et al. (2020), or by considering them both but without relating the semantics underlying between the user's search intent and the system's document results (Bing et al., 2018), we lose valuable mutual information about the *interactive intentions* (Xie, 2002) that could act as a soft supervision during the search task. Based on previous findings (Eickhoff et al., 2014; Liu et al., 2019a) showing how user's query behavior and knowledge change from system's results during the search session, we propose a framework for Session-aware Query rEpresentation learning based on multi-View Learning (SaQuEViL).

SaQuEViL is a two-step architecture that consists of two single-view query encoders, namely user-view and system-view query encoders, and a multi-view query encoder. Each single-view query encoder is based on a bidirectional transformer (Vaswani et al., 2017) at the session level. By investigating the use of unsupervised multi-view based learning algorithms, namely Cross-modal Factor Analysis (CFA) and Uniform Manifold Approximation and Projection (UMAP), the multi-view encoder takes as input the two single-view query embeddings related to the same query and provides a multi-view query representation. The underlying objective functions aim to maximize the alignment of features between both views which leans to reveal the underlying manifold. In the multi-view embedding space, similar queries formulated in the context of similar tasks have spatially close representations.

Our key contributions are: 1) we model generic session-aware query representation as an unsupervised multi-view learning task using a two-step framework architecture, SaQuEViL; 2) we experimentally show the effectiveness of multi-view based representations in query classification and session-retrieval as downstream tasks; 3) we conduct quantitative and qualitative analyses showing the potential of SaQuEViL in understanding user's search behavior.

2 Related Work

2.1 Distributed query representation

A common problem in IR is that queries –the pivotal parts of a retrieval process– are under-specified which is prone to the vocabulary mismatch and thereby, the poor performance of search-related tasks. Recently, much attention has been paid to learning distributed query representations. Previous work following this approach can be organized based on the facet of query context and type of supervision used to learn the distributed representations. In the first line of work, both query context and supervision include user's relevance signals on documents (Zamani and Croft, 2016, 2017; Zhang et al., 2019). The underlying assumption is that the more queries share the same relevant or pseudo-relevant documents among those selected by the retrieval system, the more they have semantically close intent leading to similar embeddings in the latent representation space. Using a probabilistic framework, Zamani and Croft (2016) propose to learn relevance-based query representations based on the embeddings of the query words. Then, the closeness between the probability distribution of the query representation, based on similarity metrics of word embeddings, and the query language model is maximised. Zhang et al. (2019) propose the GEN Encoder which learns distributed representations of queries in two stages. The first stage captures user's intent based on document clicks by using the assumption that queries with similar clicks underline similar intent. The second stage denoises the representations and enhances their generalizability by leveraging human paraphrase labeling in a multi-task learning setting. The second line of work relies on query context held by the search history through query reformulations recorded into physical sessions (Grbovic et al., 2016) or task-based sessions (Mehrotra and Yilmaz, 2017; Sen et al., 2018). Query embeddings are learned based on the assumption that lexically similar queries formulated in similar search sessions across users are semantically related leading to close representations in the embedding space. Mehrotra and Yilmaz (2017) propose task-aware query embeddings by applying the skip-gram model on sequences of queries belonging to the same task-based session. These query representations learned in an unsupervised manner are expected to be generic, thought their evaluation has been limited to specific downstream tasks such as query expansion in sponsored

search (Grbovic et al., 2016) and search task extraction (Sen et al., 2018). A recent line of work uses context built up on query reformulation in a session and documents (Bing et al., 2018; Zhou et al., 2020). For instance, Bing et al. (2018) model an unified graph information where vertices consist of queries in the session, clicked documents and corresponding websites; and edges reflect undifferentiated semantic relationships. The authors propose a supervised model based on an objective function that aims at optimizing, over session data, the log-likelihood of reaching a leaf (i.e., query, URL) in the corresponding Huffman tree.
In contrast to most all the aforementioned works that model query representation as supervised text representation learning based on the core idea of "query sentence", we model query representation as multi-view learning of manifold underlying queries and document results based on the core idea of *interactive intentions* (Xie, 2002) that provide soft supervision during the search session.

2.2 Session-aware query reformulation

Session-aware query reformulation is involved in retrieval-based interactive systems, including dynamic IR systems (Yang et al., 2016), multi-turn Question Answering (QA) (Mensio et al., 2018), and dialogue systems (Cui et al., 2019). Several works studied the connections between search sessions, intentions in query reformulation, and search behavior (Lu et al., 2017; Liu et al., 2019b; Tamine et al., 2020). Among the major findings, we particularly mention the following: (1) query reformulation patterns can be observed in search sessions providing insights on the search process characteristics such as underlying search task stage (Tamine et al., 2020; Eickhoff et al., 2014) and success (Odijk et al., 2015); (2) during the session search, system's results often lead to a change in both user's knowledge and the complexity of subsequent queries (Eickhoff et al., 2014; Liu et al., 2019a); (3) user search process runs into sequential phases, specialization, and intent shift. As user's search intents are gradually satisfied based on system's results, their subsequent queries lean to topically shift (Chen et al., 2021).
The main findings that have been drawn from the literature review strengthen our motivation toward learning single-view query embeddings that capture hidden session-related patterns from the two perspectives of user's sequence of query reformulations in the one hand and system's results in the other hand, and then identify mutual information that can reveal similarities across users' search intents.

3 Background

3.1 Multi-view representation learning

Multi-view representation learning (Li et al., 2019) aims to recover a meaningful latent representation of a target object using data provided by one or multiple sources. The *views* correspond to measurement modalities from such different sources, such as text and images of the same scene (Hwang and Grauman, 2012) but may also be multiple information from the same source such as document text and hyperlinks (Bickel and Scheffer, 2004). Potential applications of multi-view learning include cross-modal retrieval (Hwang and Grauman, 2012; Li et al., 2003) and machine translation (Faruqui and Dyer, 2014). SOTA methods for multi-view feature learning are the Canonical Correlation Analysis (CCA) (Dhillon et al., 2011) and Cross-modal Factor Analysis (CFA) (Li et al., 2003) whose primary goal is to maximize the correlations of features among multiple different views. These methods generally admit global solutions and ignore the non-linearities of multi-view data (Viinikanoja et al., 2010). Unlikely, k-neighbor based manifold learning methods such as Laplacian Eigenmaps (Belkin and Niyogi, 2003), IsoMap (Tenenbaum et al., 2000), and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) recover non-linear dependencies between views. The core of these methods relies upon optimization over a graphical representation of different data sets that are characterized by the same underlying manifold where edges in the graph are computed to preserve the topological structure of this manifold. This optimization yields a shared low-dimensional space where the latent representations of semantically similar data are spatially close to one another. Recently, several proposed methods for multi-view representation learning are based on deep neural networks. For instance, Deep CCA aims to learn complex nonlinear transformations of two views in a shared space (Andrew et al., 2013).

3.2 Definitions and notations

We introduce here some key definitions. Note that we refer the term of *embedding* to either the user-view query vector or system-view query vector and refer the term of *representation* as the final multi-view query vector.

Definition 1. *Search session.* In the literature review, there are two main definitions of search sessions: (1) a *physical session* (Hagen et al., 2013) is a set of consecutive queries automatically delimited using a time-out threshold on user's activities; (2) a *task-based session* which targets an atomic information need through a set of queries that are possibly neither consecutive nor within the same time-based session. SaQuEViL can be readily applied to both definitions of search sessions.
Formally, let $\mathcal{S}$ be the set of users' search sessions. A user's search session $S \subset \mathcal{S}$ consists of: (1) all on-session user's queries $q_{1,S}, q_{2,S}, \ldots, q_{k,S}$ ordered by time where each query $q_{m,S}$, consists of K_m words $q_{m,S} = \{w_{m1}, w_{m2} \ldots, w_{mK_m}\}$; (2) the sets of N top documents returned by the retrieval system as an answer to each query $q_{m,S}$, denoted as $\mathcal{D}^N_{m,S}$.
Definition 2. *User-view query embedding.* Each on-session query $q_{m,S}$ is embedded as a d_1-dimensional user-view query embedding, denoted as $\mathbf{q}^u_{m,S} \in \mathbb{R}^{d_1}$, that captures the user's formulation of his search intent. $\mathbf{q}^u_{m,S}$ is encoded based on its formulation $\{w_{m1}, w_{m2} \ldots, w_{mK_m}\}$ as well as all the formulations of the previous queries in the session $\{q_{m-1,S}, q_{m-2,S} \ldots, q_{1,S}\}$.
Definition 3. *System-view query embedding.* Each on-session query $q_{m,S}$ is embedded as a d_2-dimensional system-view query embedding, denoted as $\mathbf{q}^s_{m,S} \in \mathbb{R}^{d_2}$, that captures the system's understanding of the user's search intent. $\mathbf{q}^s_{m,S}$ is encoded based on document results obtained from the concatenation of the query $q_{m,S}$ along with previous queries in the session.

4 Session-Aware Query Representation By Multi-View Learning

4.1 Problem statement

Let $\mathcal{S} = \{S_1, \ldots, S_K\}$ be a set of sessions such as $S_i = \{q_{1,i}, q_{2,i}, \ldots, q_{ki,i}\}$, including a total of n on-session queries $q_{m,i}$ with $n = (\sum ki)^{i=K}_{i=1}$. The objective of SaQuEViL is twofold: (1) encoding $\Sigma^1 \in \mathbb{R}^{n\times d_1}$ (resp. $\Sigma^2 \in \mathbb{R}^{n\times d_2}$) the vector space embedding and user-view query embeddings $\mathbf{q}^u_{m,i}$ (resp. system-view query embeddings $\mathbf{q}^s_{m,i}$); (2) learn a multi-view latent space $\Sigma^* \in \mathbb{R}^{n\times d}$ (with $d \le min(d_1, d_2)$) and query representations $\hat{\mathbf{q}}_{m,i} \in \Sigma^*$ by jointly achieving pairwise alignments between the user-view embedding $\mathbf{q}^u_{m,i}$ and system-view embedding $\mathbf{q}^s_{m,i}$ and recovering an optimal alignment of manifolds over all the query representations $\hat{\mathbf{q}}_{m,i}$. Final representations are picked to match the downstream task, either when *document* matching is required or session-aware *query* is required.

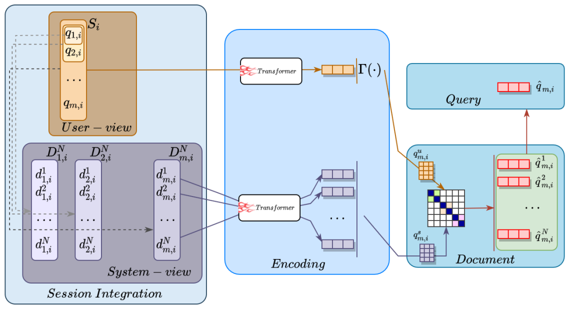


Figure 1: Overview of the SaQuEViL framework.

The two key assumptions of multi-view learning are satisfied (Blum and Mitchell, 1998; Foster et al., 2008): (1) each of the user-view and system-view are independent conditionally to the sessions; and (2) the two single views provide a redundant estimate of the session.

4.2 Multi-view query representation learning

4.2.1 Framework overview

Figure 1 presents an overview of the SaQuEViL framework. For encoding the single-view query embeddings $\mathbf{q}^u_{m,i}$, $\mathbf{q}^s_{m,i}$, we opted for BERT (Devlin et al., 2019) as transformer embedding and followed the standard CLS encoding strategy (BERT_{CLS}). So, $\mathbf{q}^u_{m,i}$ (resp. $\mathbf{q}^s_{m,i}$) is obtained by applying $\Gamma(\text{BERT}_{CLS}([CLS]q_{m,i}))$ (resp. $\otimes_{j=1}^{j=N} \text{BERT}_{CLS}([CLS]head(d^j_{m,i})))$, where $\otimes$ is a vector concatenation operator, $head(\cdot)$ is a function that returns the title and first tokens of a given document, and Γ is an expansion function such as broadcast used to match the dimensions.
Following, we detail the key principles of multi-view query representation learning $\hat{\mathbf{q}}_{m,i}$ using linear (CFA (Dhillon et al., 2011)) and non-linear (UMAP (McInnes et al., 2018)) methods.

4.2.2 CFA-based representation learning

Given the two mean centered matrices $\mathcal{Q}^u \in \mathbb{R}^{d_1\times n}$ and $\mathcal{Q}^s \in \mathbb{R}^{d_2\times n}$, where columns refer respectively to the user-view embeddings $\mathbf{q}^u_{m,i}$ and system-view embedding $\mathbf{q}^s_{m,i}$, CFA learns two linear and orthogonal transformations $A \in \mathbb{R}^{d_1\times d}$ and $B \in \mathbb{R}^{d_2\times d}$ such that the distance between $A^\intercal \mathcal{Q}^u$ and $B^\intercal \mathcal{Q}^s$ is minimized. The CFA objective is:

$$A^*, B^* = argmin_{A,B}(\| A^\intercal \mathcal{Q}^u - B^\intercal \mathcal{Q}^s \|_F) \quad (1)$$

where $A^\intercal A = I$ and $B^\intercal B = I$ and $\|\cdot\|_F$ is the Frobenius norm. The solution of Equation (1) is obtained through the Singular Value Decomposition (SVD) of $Z = (\mathcal{Q}^u)^\intercal \mathcal{Q}^s$, such as $Z = S_z V_z D_z$ and $A^* = S_z, B^* = D_z$ (Krzanowski, 1988). Thus, we obtain the multi-view query representations $\mathbf{q}^u$ and $\mathbf{q}^s$ as the rows of the user-view or system-view transformations $\hat{\mathcal{Q}}^u = (\mathcal{Q}^u)^\intercal A^*$ and $\hat{\mathcal{Q}}^s = (\mathcal{Q}^s)^\intercal B^*$ respectively.

4.2.3 UMAP-based representation learning

Let $G(\mathcal{V}, \xi)$ be the graph where the vertices $\mathcal{V}$ correspond to queries $q_m \in \cup(S_i)_{i=1}^K$ and ξ the edges that reflect a weighted neighborhood relationship $\mathbf{q}_m \sim \mathbf{q}_{m'}$ defined in matrix $\mathcal{W}$ such as $\mathcal{W}(m, m') > 0$ if $\mathbf{q}_m, \mathbf{q}_{m'}$ are neighbors. The two key differences between SOTA graph-based manifold learning algorithms (Belkin and Niyogi, 2003; Tenenbaum et al., 2000) lie in the construction of the k-neighbor edges ξ and the choice of the weights $\mathcal{W}(i, j)$. Specifically, in the multi-view setting of the UMAP method, for each query q_m, there are two induced local graphs: (1) the user graph $G_m^u(\mathcal{V}^u, \xi_m^u)$ where $\mathcal{V}^u$ is the set of k-nearest neighbors of $\mathbf{q}_m^u$ denoted as $Fset^u(q_m)$ and ξ_m is the set of outgoing edges directed from q_m to its set k-nearest neighbors q_{mj}^u thereby inducing the similarity relationship $\mathbf{q}_m^u \sim \mathbf{q}_{mj}^u$ defined in matrix $\mathcal{W}^u(n, n)$; (2) the system graph $G_m^s(\mathcal{V}^s, \xi_m^s)$ where $\mathcal{V}^s$ is the set of k-nearest neighbors of $\mathbf{q}_m^s$ denoted as $Fset^s(q_m)$ and ξ_m^s is the set of outgoing edges directed from q_m to its set k-nearest neighbors $\mathbf{q}_{mj}^s$ thereby inducing the similarity relationship $\mathbf{q}_m^s \sim \mathbf{q}_{mj}^s$ defined in matrix $\mathcal{W}^s(n, n)$. Pairwise alignment between the user-view and system-view of query q_m is ensured by building the graph $G(\mathcal{V}, \xi)$ as a graph intersection between user graph $G_m^u(\mathcal{V}^u, \xi_m^u)$ and system graph $G_m^s(\mathcal{V}^s, \xi_m^s)$ for each query $q_m \in \cup(S_i)_{i=1}^K$. This intersection builds the weighting matrix $\mathcal{W}(n, n)$ based on the weighting matrices $\mathcal{W}^u$ and $\mathcal{W}^s$. Spectral optimization of the multi-view query representations is then achieved by functions $\mathbf{f} : \mathcal{V} \mapsto \mathbb{R}$ that recover the optimal alignment of manifolds underlying queries $q_m \in \cup(S_i)_{i=1}^K$ through the minimization of a cost on graph $G(\mathcal{V}, \xi)$, defined as (McInnes et al., 2018):

$$\mathcal{L}(f) = \sum_{S \in \mathcal{S}; q_m, q'_m \in S} \frac{1}{2}(f_m - f'_m)^2 \mathcal{W}(m, m') \tag{2}$$

subject to scale and translation constraints $f^T f = 1$ *and* $f^T e = 0$.

The optimization process of UMAP is detailed in Belkin and Niyogi (2003); McInnes et al. (2018).

5 Experimental Setting

We address the following research questions:
RQ1) How does the SaQuEViL framework perform in query classification and session-based retrieval as downstream tasks?
RQ2) To what extent the SaQuEViL embedding space preserves the similarities of each of the single-view embedding spaces?
RQ3) Can we use SaQuEViL framework to understand user's search behavior?

5.1 Downstream tasks

5.1.1 Query classification

The goal of query classification consists in assigning an incoming query the most appropriate topic labels (categories). Labels are pre-defined and search-related data are available to train each label.

Data. As previously done by Zamani and Croft (2016); Zamani et al. (2017) to evaluate query embedding performances, we used the KDD 2005 dataset (Li et al., 2005). The dataset consists of 800 queries recorded from MSN search log. The dataset also includes 43 categories -that act as candidate task-based sessions- labeled by human assessors. Accordingly, we assume that the set of queries belonging to each target category c represent a session S_c. To solely measure the quality of the query representations and ensure comparability across query representations, we opted for the classification strategy proposed in Zamani and Croft (2016); Zamani et al. (2017). We first compute the probability of each category (session) $p(Sc_i/q) = \frac{\delta(\vec{Sc_i}, \vec{q})}{\sum_j \delta(\vec{Sc_j}, \vec{q})}$ where $\vec{q}$ is a query vector, $\vec{Sc_i}$ is the centroid vector of category (session) Sc_i. $\vec{Sc_i}$ is computed by averaging the query vectors $\vec{q_{ki}}$ of queries q_{ki} belonging to session Sc_i. Then we select the N top sessions with the highest probabilities as the more likely ones to be assigned to query q.

Evaluation metrics. We consider the evaluation metrics used in the KDD challenge (Li et al., 2005), Recall and F1 measures, and carefully followed their description to implement our evaluation script. Statistical tests are performed using two-tailed paired t-test. We depict a significant increase for p < 0.05 as *.

Baselines and scenarios. We reported traditional SOTA pre-trained embeddings as query encoders *GloVe* (Pennington et al., 2014), *Word2vec* (Mikolov et al., 2013) and *BERT* (Devlin et al., 2019), as well as *RPE* (Zamani and Croft, 2016; Zamani et al., 2017), a SOTA relevance-based query representation model. To show the impact of user-view and system-view alignment, we also compared our multi-view CFA-based and UMAP-based query representations $\hat{\mathbf{q}}$ to the representation vector obtained by concatenation of $\mathbf{q}^u$ and $\mathbf{q}^s$ vectors. The latter scenario is denoted *w/o Align*.

Training and inference. We performed a 5-fold cross-validation over the queries and used the documents rankings provided by the ClueWeb12[1] corpus to learn the SaQuEViL multi-view query representations. The ClueWeb12 corpus was indexed using the respective default configuration of Anserini[2] while the retrieval was done using the default configuration of Pyserini[3] search. With respect to Figure 1, projected $\hat{\mathbf{q}}$ vectors are averaged in order to obtain a unique vector per query. The number of labels assigned to each query was tuned on the training set from 1 to 5.

5.1.2 Session-based retrieval

The goal of session-based retrieval consists in evaluating document rankings over user sessions rather than isolated queries (Carterette et al., 2016).

Data. We use the TREC 2014 session track (Carterette et al., 2016) which provides the following: (1) 1,257 full sessions among which 1,021 of these have at least one reformulation. On average there are 4.33 queries per session, among which the final query in the session is referred to as the *current query*; (2) the ranked list of documents for each past query; and (3) human annotations about type of search for 54 sessions; the latter are labeled using 4 categories of user search behavior w.r.t. the classification designed by Li and Belkin (2008): *known-item*, *interpretive*, *known-subject*, and *exploratory*.

It is worth of noting that we did not use the users' clicks in our experiments since they are considered as weak supervision. The corpus used is the ClueWeb12 collection. The relevance of a document was judged for the results of the current query but judgment is based on the whole session.

[1] https://lemurproject.org/clueweb12.php/

[2] https://github.com/castorini/anserini

[3] https://github.com/castorini/pyserini

Evaluation metrics. We use the TREC session track's official metrics. These are: nDCG@10, ERR@10, nERR@10, and PC@10. All runs are evaluated using the official evaluation script[4].

Baselines and scenarios. We used classical baselines including *Current* and *Aggregated query*. The latter is a concatenation of all the session's queries as suggested in Van Gysel et al. (2016).

Training and inference. In contrary to query classification, projected $\hat{\mathbf{q}}$ vectors are not aggregated as each is used for document ranking. We first compute a neural score by calculating the cosine similarity between the session vector $\sum_{j=1}^{m-1} \hat{\mathbf{q}}_{j,S}$ and the document vector $\hat{\mathbf{q}}_{m,S}^{d^l_{m,S}}$ in the SaQuEViL space. Then we obtain the final score used for document ranking by linearly combining the neural score with the BM25 score as commonly done in neural IR (MacAvaney, 2020).

6 Results and Analysis

6.1 RQ1: Effectiveness evaluation of SaQuEViL in downstream tasks

6.1.1 Query classification

Table 1 presents the performance results in terms of Precision and F1. Note that one strong baseline is obtained by encoding the query with BERT (0.4143) which clearly outperforms a supervised alternative (0.3961), e.g., RPE which is trained on relevance signals (Zamani and Croft, 2017). It can be explained as RPE do not use contextualized embeddings as BERT. We can interestingly see that SaQuEViL, even trained without supervision, outperforms (0.443) both unsupervised (GloVe, Word2vec, and BERT) and supervised encoders (RPE model). This result clearly indicates the value of the alignment to identify relevant mutual information between user's view through query reformulation and system's view through document rankings to enhance the query representation. We can also see that even without alignment, SaQuEViL (0.4274) outperforms BERT (0.4143) indicating that each view information is helpful on this task. Finally, our best scenario corresponds to the SaQuEViL CFA setup that achieves a minimum improvement of 7% in terms of Precision and F1 w.r.t. reported baselines. This result leads us to consider that linear dependencies are revealed from session-based query reformulations and corresponding documents.

[4] https://trec.nist.gov/data/session2014.html

Model	Precision	F1
GloVe	0.3643 (+22.0%)	0.3912 (+28.3%)
Word2vec	0.3712 (+19.7%)	0.4008 (+25.2%)
BERT	0.4143 (+7.2%)	0.4537 (+10.6%)
RPE	0.3961 (+12.2%)	0.4294 (+16.9%)
SaQuEViL		
w/o Align	0.4274	0.4827*
CFA	**0.4443***	**0.5020***
UMAP	0.4246	0.4802*

Table 1: Performance of SaQuEViL query representations and baselines (GloVe (Pennington et al., 2014), Word2vec (Mikolov et al., 2013), BERT (Devlin et al., 2019), and RPE (Zamani and Croft, 2017)) in query classification. The improvements over each baseline of our best scenario, SaQuEViL CFA, are reported in brackets. The highest values are highlighted in bold. Improvement significance w.r.t. BERT is indicated by the superscript '*'.

Model	NDCG@10	ERR@10	nERR@10	PC@10
Current	0.1659	0.1639	0.2332	0.3190
Aggregated	0.1834	0.1952	0.2645	0.3460
SaQuEViL				
w/o Align	0.1841	**0.2021**	**0.2749**	0.3340
CFA	**0.1843**	0.1950	0.2646	**0.3473**
UMAP	0.1835	0.1951	0.2644	0.3450

Table 2: Performance of SaQuEViL query representations and baselines (Aggregated (Van Gysel et al., 2016)) in session-based retrieval. Best results are highlighted in bold.

6.1.2 Session-based retrieval

Table 2 presents the performance scores of SaQuEViL scenarios and baselines in the session-based retrieval downstream task. As expected, including session information outperforms (0.1834) the use of the single query (0.1659) in terms of NDCG@10, but also for all the other metrics. Moreover, we can notice that SaQuEViL slightly improves the Aggregated (Van Gysel et al., 2016) results but none scenario shows a clear wining. SaQuEViL w/o Align setup outperforms in terms of ERR@10 (0.2021) and nERR@10 (0.2749) but SaQuEViL CFA obtains the best scores for NDCG@10 (0.1843) and PC@10 (0.3473). Nerveless, the improvements for the session-based retrieval downstream task are modest[5]. We can also notice that CFA and UMAP methods exhibit the same performance trend.

6.2 RQ2: Analysis of the SaQuEViL multi-View embedding space

Our main objective here is to analyse to what extent the SaQuEViL framework builds a shared embedding space that preserves the structure of the

[5]Note that stronger results on the TREC session 2014 dataset are reported by Aloteibi and Clark (2020), but we only focused on an extrinsic use of SaQuEViL and integration to task specific models is out of the scope of the paper.

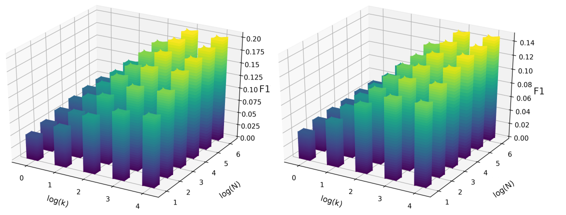


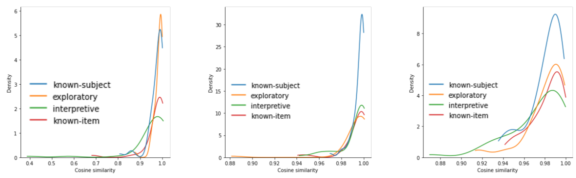
Figure 2: F1 performances when comparing SaQuEViL CFA (left) / UMAP (right) multi-view space and the concatenation of both views embedding. Number of neighbors and ranked documents are in log scale. Better in color as brighter color indicates higher values.

single-view spaces. Grounded with the results obtained above (Section 6.2), we achieve this goal by analysing the discrepancies between the single-view spaces and the shared space obtained with SaQuEViL using the query representations learned in query classification. For each target query q, we consider the k-neighbors of $\hat{\mathbf{q}}$ in the SaQuEViL shared space as the gold standard and the plurality vote of the k-neighbors in each of the single-view spaces, namely, $\mathbf{q}^u$ and $\mathbf{q}^s$, as the prediction. We used the cosine similarity to find neighbors and then compute Precision, Recall and F-measure metrics under a multi-label setup, where each query identifier is considered as a target class. In particular, we analyse the impact of two key parameters of the SaQuEViL framework: number of neighbors (k) and number of top documents (N) used to learn the query representations. Results for different values of k (1, 2, 4, 8, and 16) and N (2, 4, 8, 16, 32, and 64) in log scale are presented in Figure 2. Three main conclusions can be grasped from Figure 2: (1) increasing the number of neighbors increases the similarity between the spaces until 8-16 neighbors then it stabilizes for both methods (CFA and UMAP) in terms of F1; (2) adding extra documents impacts in the same way, e.g. positive at early increments and then stabilizes, but for the two multi-view learning methods; (3) a higher preservation of original similarities in SaQuEViL spaces correlates with higher performances on the downstream task as SaQuEViL CFA obtains a maximum score close to 0.20 of F1 while UMAP is 0.06 points behind (0.14 of F1)[6]. These results might shed light on possible controllable room of improvements of a wide range of downstream tasks including, but not limited to session-based retrieval.

[6]Note that this correlation must have an upper limit lower than 1.0 (F1) as exactly similar spaces may lay on similar performances to our strategy w/o align in downstream tasks.

Model	First k queries into the session				
	1	3	6	9	all
Aggregated	0.373	**0.573**	0.553	0.535	0.535
SaQuEViL					
w/o Align	0.462	0.571	0.607	0.589	0.589
CFA	**0.516**	0.569	**0.625**	**0.625**	**0.625**
UMAP	0.498	0.571	**0.625**	0.589	0.589

Table 3: F1-micro performances of SaQuEViL and baseline (Aggregated (Van Gysel et al., 2016)) encoders in search type classification using TREC session 2014. Highest values of F1-measure are highlighted in bold.



(a) (b) (c)

Figure 3: Distribution of cosine similarities for (a) GloVe, (b) SaQuEViL-w/o Align, and (c) SaQuEViL CFA between adjacent queries per session categorised by search type *known-item*, *interpretive*, *known-subject*, and *exploratory*.

6.3 RQ3: Search behavior understanding

Our aim here is to understand in what extent the SaQuEViL representation space helps understanding behaviors in user session. To do so, we used the *type of search* annotations provided in the TREC session 2014 dataset (*known-item*, *interpretive*, *known-subject*, and *exploratory*). A standard 5-cross fold setup with k-nearest neighbor classifier is used to draw the intrinsic capabilities of the encoders to distinguish user search behavior types. Average results of F1-micro across the 5 folds are presented in Table 3. To perform the classification at the test stage, we used as context the first k queries of sessions (columns 1, 3, 6 and 9) as well as the full session (column *all*). As can be seen from Table 3, SaQuEViL CFA encoder (0.625) clearly outperforms the proposed alternatives, the BERT encoder for the Aggregated queries (0.535) and the SaQuEViL w/o Align (0.589) when considering the full session. Looking at the impact of context length (k) in the classification, we can note that the Aggregated query representation starts with a low performance (0.373) and, when up to 3 queries are used in the session, it achieves the maximal performance (0.573). However, the SaQuEViL w/o Align encoder starts in a higher performance (0.462) and achieves the maximal performance when up to 6 queries are used from the session (0.607). In both cases, the performance drops when the size of the session increases. This also points an advantage of SaQuEViL CFA encoder as it shows a more stable performance (0.516 to 0.625) regardless the number of used queries.
To further our analysis, we plot in Figure 3 distributions of distances between adjacent query pairs for each session w.r.t. corresponding search type and by using different query encoders: GloVe, SaQuEViL w/o Align, and SaQuEViL CFA[7]. We can see that the distribution of CFA encoder significantly differs from the other encoders. Interestingly, we note that CFA better separates the four search types and gradually differentiates the trends of query similarities based on the two dimensions of search namely "goal-quality" and "product" of the search. Indeed, the curves with more spread query similarity values (0.87-0.99) correspond to *interpretive* and *exploratory* sessions which reflect non-factual task products with either specific or amorphous goals leading to issue semantically different queries along the sessions. Unlikely, the curves with high density of narrow and relatively high similarity values (0.93-0.99) reflect factual search as characterised in *known-subject* and *known-item* search.

[7]UMAP exhibits the same distribution trend than CFA and has not been presented for limited space.

7 Conclusion

The paper presented SaQuEViL, a framework that learns query representations that reflect users' intents within a session-based search. By relying on the key finding that system's results affects user's query behavior and knowledge, we advocate the use of unsupervised multi-view learning to capture manifolds in a shared distributed representation space. Through experimental evaluation in two downstream tasks, we show the effectiveness of SaQuEViL over supervised and unsupervised pretrained encoders, though improvements are limited in session-based retrieval that inherently requires relevance supervision. A series of experiments and qualitative analyses also show the potential of SaQuEViL to control the representation space through key parameters that directly influence performance of downstream tasks and additionally, to clearly separate user behaviour patterns in search sessions. We believe that this work opens avenues of research in the design of unsupervised distributed representations able to support search tasks, which has not received much attention yet.

References

Saad Aloteibi and Stephen Clark. 2020. Learning to personalize for web search sessions. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 15–24. ACM.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. ICML'13, page III–1247–III–1255. JMLR.org.

Mikhailand Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Steffen Bickel and Tobias Scheffer. 2004. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, page 19–26, USA. IEEE Computer Society.

Lidong Bing, Zheng-Yu Niu, Piji Li, Wai Lam, and Haifeng Wang. 2018. Learning a unified embedding space of web search from large-scale query log. *Knowledge-Based Systems*, 150:38 – 48.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA. Association for Computing Machinery.

Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The trec session track 2011-2014. SIGIR '16, page 685–688, New York, NY, USA. Association for Computing Machinery.

Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the Web Conference 2021*, pages 743–755.

Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User attention-guided multimodal dialog systems. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 445–454, New York, NY, USA. Association for Computing Machinery.

Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1747–1756.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. 2011. Multi-view learning of word embeddings via cca. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 199–207, Red Hook, NY, USA. Curran Associates Inc.

Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 223–232, New York, NY, USA. Association for Computing Machinery.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Dean P. Foster, Rie Johnson, and Tong Zhang. 2008. Multi-view dimensionality reduction via canonical correlation analysis. Technical report.

Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, Ricardo Baeza-Yates, Andrew Feng, Erik Ordentlich, Lee Yang, and Gavin Owens. 2016. Scalable semantic matching of queries to ads in sponsored search advertising. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 375–384, New York, NY, USA. Association for Computing Machinery.

Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From search session detection to search mission detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 85–92.

Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 829–838.

Jiyin He and Emine Yilmaz. 2017. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 67–76.

Sung Ju Hwang and Kristen Grauman. 2012. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *Int. J. Comput. Vision*, 100(2):134–153.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Comput. Linguist.*, 29(3):459–484.

W. J. Krzanowski. 1988. *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, Inc., USA.

Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, page 604–611, New York, NY, USA. Association for Computing Machinery.

Ying Li, Zijian Zheng, and Honghua (Kathy) Dai. 2005. Kdd cup-2005 report: Facing a great challenge. *SIGKDD Explor. Newsl.*, 7(2):91–99.

Yingming Li, Ming Yang, and Zhongfei Zhang. 2019. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883.

Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management*, 44(6):1822–1837. Adaptive Information Retrieval.

Jimmy Lin. 2021. A proposed conceptual framework for a representational approach to information retrieval. *SIGIR Forum*.

Hanrui Liu, Chang Liu, and Nicholas J. Belkin. 2019a. Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, 56.

Jiqun Liu, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2019b. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 123–132, New York, NY, USA. Association for Computing Machinery.

Kun Lu, Soohyung Joo, Taehun Lee, and Rong Hu. 2017. Factors that influence query reformulations and search performance in health information retrieval: A multilevel modeling approach. *J. Assoc. Inf. Sci. Technol.*, 68(8):1886–1898.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying task-based sessions in search engine query logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 277–286.

Sean MacAvaney. 2020. OpenNIR: A complete neural ad-hoc ranking pipeline. In *WSDM 2020*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Rishabh Mehrotra and Emine Yilmaz. 2017. Task embeddings: Learning query embeddings using task context. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2199–2202, New York, NY, USA. Association for Computing Machinery.

Martino Mensio, Giuseppe Rizzo, and Maurizio Morisio. 2018. Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems. WWW '18, page 1075–1080, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. 2015. Struggling and success in web search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1551–1560, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Mark Sanderson. 2008. Ambiguous queries: Test collections need more sense. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 499–506, New York, NY, USA. Association for Computing Machinery.

Procheta Sen, Debasis Ganguly, and Gareth Jones. 2018. Tempo-lexical context driven word embedding for cross-session search task extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 283–292, New Orleans, Louisiana. Association for Computational Linguistics.

Manuel Steiner. 2019. The influence of backstories on queries with varying levels of intent in task-based specialised information retrieval. In *Advances in Information Retrieval*, pages 375–379, Cham. Springer International Publishing.

Lynda Tamine, Jesús Lovón Melgarejo, and Karen Pinel-Sauvagnat. 2020. What can task teach us about query reformulations? In *Advances in Information Retrieval*, pages 636–650, Cham. Springer International Publishing.

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. SIGIR '05, page 449–456, New York, NY, USA. Association for Computing Machinery.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Christophe Van Gysel, Evangelos Kanoulas, and Maarten de Rijke. 2016. Lexical query modeling in session search. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 69–72.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, page 5998–6008. Curran Associates, Inc.

Jaakko Viinikanoja, Arto Klami, and Samuel Kaski. 2010. Variational bayesian mixture of robust cca models. ECMLPKDD'10, page 370–385, Berlin, Heidelberg. Springer-Verlag.

Hong (Iris) Xie. 2002. Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management*, 38(1):55 – 77.

Grace Hui Yang, Marc Sloan, and Jun Wang. 2016. *Dynamic Information Retrieval Modeling*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational context for ranking in personal search. WWW '17, page 1531–1540, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Hamed Zamani and W. Bruce Croft. 2016. Estimating embedding vectors for queries. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 123–132, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 505–514, New York, NY, USA. Association for Computing Machinery.

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 65–74, New York, NY, USA. Association for Computing Machinery.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Encoding history with context-aware representation learning for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1111–1120, New York, NY, USA. Association for Computing Machinery.