Task Embeddings: Learning Query Embeddings using Task Context

Rishabh Mehrotra[†] and Emine Yilmaz[†*]
[†]University College London, London, United Kingdom
[*]The Alan Turing Institute, British Library, London, United Kingdom
{r.mehrotra,e.yilmaz}@cs.ucl.ac.uk

ABSTRACT

Continuous space word embedding have been shown to be highly effective in many information retrieval tasks. Embedding representation models make use of local information available in immediately surrounding words to project nearby context words closer in the embedding space. With rising multi-tasking nature of web search sessions, users often try to accomplish different tasks in a single search session. Consequently, the search context gets polluted with queries from different unrelated tasks which renders the context heterogeneous. In this work, we hypothesize that task information provides better context for IR systems to learn from. We propose a novel task context embedding architecture to learn representation of queries in low-dimensional space by leveraging their task context information from historical search logs using neural embedding models. In addition to qualitative analysis, we empirically demonstrate the benefit of leveraging task context to learn query representations.

1 INTRODUCTION

Users tend to seek information by issuing queries to a search engine. The need for search often resides within an external context that prompts the user to formulate their information needs as search queries. When an information need, or task, requires multiple searches, the sequence of queries form a context which influences interaction behavior for the duration of the search process. Search context plays an important role in understanding user's needs and can be leveraged to develop better representations and ranking models. While a major portion of existing work have investigated user behavior using search sessions as the fundamental focus of search activity, search tasks are emerging as a competing perspective in this space with recent studies suggesting that users seek to complete multiple search tasks within a single search session, while also taking multiple sessions to finish a single task at times [12]. As a result, search tasks have steadily emerged as accurate units to capture searcher's goals and seeking behavioral insights.

A direct result of users being engaged in multitasking and task switching behaviors is that the resulting search context is heterogeneous, composed of interleaved search goals and tasks. Recent advancements in task extraction techniques have made it possible to segregate search activity logs into a set of interleaved tasks [13, 23]. For example, while multi-tasking in their search sessions, users pursue many different tasks at once, often switching between them. Figure 1 shows an example of a user multitasking in her search session, with her task of finding information on GRE exams interleaved with finding music videos of Coldplay band. Such heterogenous context makes it difficult for the retrieval system to use to create localized user interest models, provide contextual result rankings, query suggestions, and other user support offerings.

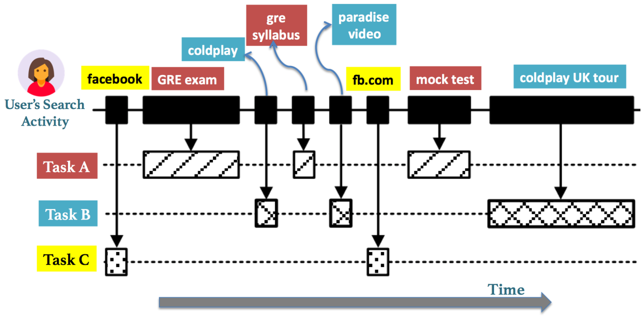

Figure 1: Exemplar user interaction with search engine.

In this work, we aim at mitigating the ill effects of heterogeneous contexts by leveraging task information while learning representations of users' information needs. Learning meaningful and accurate representations of queries is an important problem in web search, with most retrieval, ranking, query expansion and query suggestion methods heavily relying on informative ways of representing search queries. Beyond traditional one-hot vectors and TF-IDF approaches, the distributed semantic representations based on dense vectors of vocabulary terms, also known as word embeddings, have been shown to be highly effective in many natural language processing and information retrieval tasks [4, 18]. In general, these approaches provide global representations of words; each word has a fixed representation, regardless of any discourse context. While a global representation provides some advantages, search context can vary dramatically by task. Most word embedding learning techniques rely on a window-based training which uses local information in immediately surrounding words. Given the multi-tasking nature of search sessions, the resulting query context is rendered heterogeneous and might consist of queries from different unrelated tasks.

In this work, we aim at leveraging task context while learning query representations. Specifically, we propose a novel task based embedding architecture to learn distributed semantic representation of query terms which prefers task context over local

CIKM'17, November 6–10, 2017, Singapore.
ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3133098

information in immediately surrounding words. We propose that embeddings be learned on a task-constrained context instead of the traditionally used global or session context. The proposed task embedding model is able to extract improved query representations which capture task context. In addition to qualitative analysis, we demonstrate the benefit of learning task based embeddings over traditional query representation techniques by showing enhanced performance when generating query suggestions.

2 RELATED WORK

We leverage insights from a number of research areas including recent advances in leveraging search context, task extraction and word embeddings. We cover these different areas of related work and discuss how our work relates to and extends prior work.

Search Context:
There is a growing body of work examining how knowledge of a searcher's interests and search context can be used to improve various aspects of search. The information retrieval (IR) community has theorized about context [17], developed context-sensitive search models [19], leveraged context to predict user interests [25] and performed user studies investigating the role of context in the search process [8]. Using the context of user activities within a search session has also been used to improve query analysis. Cao et al. [2, 3] represented search context by modeling sessions as sequences of user queries and clicks and applied the models to query suggestion, query categorization, and URL recommendation. Mihalkova and Mooney [15] used similar search session features to disambiguate the current query.

Search Tasks:
While a major share of prior work have considered search sessions as the focal unit of analysis for seeking behavioral insights, search tasks are emerging as a competing perspective in this space. There have been attempts to extract in-session tasks [7, 11, 21], and cross-session tasks [10, 24] from query sequences based on classification and clustering methods. Kotov et al [10] and Agichtein et al [1] studied the problem of cross-session task extraction via binary same-task classification, and found that different types of tasks demonstrate different life spans. More recent efforts have focussed on extracting subtasks from a complex task [13] as well as extracting task-subtask hierarchies [14].

Distributional Semantics for IR:
Word embeddings have been studied in various IR contexts such as term reweighting [26], cross-lingual retrieval [6, 22] and short text similarity [9]. Beyond word co-occurrence, recent studies have also explored learning text embeddings from clickthrough data [20], session data [5] and for query prefix-suffix pairs [18]. Finally, Diaz *et al.* [4] highlight the value of locally-training word embeddings in a query-specific manner.

3 TASK EMBEDDINGS

Our goal in this work is to learn richer embeddings and explore the use of task context to learn more contextual query representations. In this section, we propose a novel embedding architecture based on task context. As a precursor to generating relevant task context, we first need to extract the set queries which belong to the same overall task given a sequence of queries issued by a user over a period of time.

3.1 Extracting "*On-Task*" Queries

In order to extract *on-task queries*, we make use of the Latent Structural SVM framework [24] for task identification. Given query sequences, search tasks are identified by clustering queries into tasks by find the strongest link between a candidate query and queries in the target cluster (*bestlink*). This is achieved by making use of the structural learning method with latent variables, i.e., latent structural SVMs, to utilize the hidden structure of query interdependencies to explore the dependency among queries within the same task.

Given a query sequence $Q = q_1, q_2, ..., q_M$, a feature vector for the task partition y is specified by the hidden best-link structure h as $\psi(Q, y, h)$. Based on $\psi(Q, y, h)$, the bestlink SVM is a linear model parameterized by w, and predicts the task partition by,

$$(y*, h*) = argmax_{y \in Y, h \in H} w^T \psi(Q, y, h) \tag{1}$$

where Y and H represent the sets of possible structures of y and h respectively. $y*$ becomes the output for cross-session tasks and $h*$ is the inferred latent structure. Based on the best-link structure, $h(q_i, q_j) = 1$ if query q_i and q_j are directly connected in h; and otherwise, $h(q_i, q_j) = 0$, with the added clause that a a query can only link to another query in the past, or formally, $\sum_{i=0}^{j-1} h(q_i, q_j) = 1 \; \forall j \geq 1$. The feature vector for any particular task partition y is defined over the links in h as,

$$\psi(Q, y, h) = \sum_{i,j} h(q_i, q_j) \sum_{s=1}^{S} \phi_s(q_i, q_j) \tag{2}$$

where a set of symmetric pairwise features $\phi_s(q_i, q_j)$ is given to characterize the similarity between query q_i and q_j. Given a set of query logs with annotated tasks, the feature vector design and the directed linkage structure of h can be inferred in an SVM setting. A detailed overview of the approach can be found in Want *et al.* [24].

Given the above formulation, we run the task extraction algorithm on search logs to extract all queries belonging to the same task. Such a query collection is henceforth referred to as "*on-task queries*".

3.2 Task Context Embedding Architecture

Estimating accurate query representations plays a crucial role in many information retrieval tasks and past work have relied on a number of different ways of building such representations from simple term frequency based approaches to the recent word embeddings. While generically learnt word embedding models have performed well in various NLP tasks, we hypothesize that incorporating task context while learning query embeddings would result in more accurate representation. In this section we describe the propose task based embedding architecture which leverages the task information as described in Section 3.1.

Given a search log comprising of a set $\boldsymbol{S}$ of $||S||$ query sequences obtained from online users, where each query sequence $S = (q_1, ..., q_{M_s})$ $\in \mathbf{S}$ is defined as an uninterupted sequence of M_s queries, and each

Query: london		Query: usps	
Global	**Task Context**	**Global**	**Task Context**
birmingham	weather	postal_service	track
nyc	time	fedex	hours
england	tube	track	delivery

Table 1: Qualitative comparison of similar words fetched using global embeddings and task embeddings.

query $q_m = (w_{m1}, w_{m2}, ...w_{mT_m})$ consists of T_m words, our objective is to find D-dimensional real-valued representation $v_{q_m} \in R^D$ of each query q_m. We begin by tagging task membership information for each query t_{q_m} using the task extraction module and casting a query sequence from a given user as a *sentence* fed into the neural embedding model.

Traditionally, embedding based models learn query representations using the skip-gram model [16] by maximizing the objective function over the entire set S of search sessions, defined as:

$$L = \sum_{s\in S}\sum_{q_m \in s}\sum_{-b\le i\le b, i\neq 0} logP(q_{m+i}|q_m) \qquad (3)$$

where v_q and v_q^i are the input and output vector representations of query q, b is defined as length of the context for query sequences, and V is the number of unique queries in the vocabulary. To incorporate the task context, we modify the objective function and incorporate a selective task context window selection function in the likelihood objective:

$$L = \sum_{s\in S}\sum_{q_m \in s}\sum_{-b\le i\le b, i\neq 0} \mathbb{1}(t_{q_{m+i}} = t_{q_m}) \times log\, P(q_{m+i}|q_m) \qquad (4)$$

The objective only considers surrounding queries which belong to the same task as the current query and disregards other non-task queries from consideration for a query's context. Probability $P(q_{m+i}|q_m)$ of observing a neighboring query q_{m+i} given the current query q_m is defined using soft-max,

$$P(q_{m+i}|q_m) = \frac{exp(v_{q_m}^T v'_{q_{m+i}})}{\sum_{q=1}^{|V|} exp(v_{q_m}^T v'_{q_m})} \qquad (5)$$

From these equations we can see that the model considers the temporal and task context of query sequences, where queries with similar contexts (i.e., with similar neighboring queries which belong to the same task) will have similar vector representations in the projected semantic space.

The proposed objective is optimized using stochastic gradient ascent, suitable for large-scale problems. However, computation of gradients ∇L for the likelihood function equation above is proportional to the vocabulary size V , which is computationally expensive in practical tasks as V could easily reach hundreds of millions. As an alternative, we used negative sampling approach proposed in [16], which significantly reduces the computational complexity.

4 EXPERIMENTAL EVALUATION

In this section we demonstrate the benefit of incorporating task context while learning query representations. We consider the task of query suggestions and provide empirical comparisons of the proposed method against various baselines.

Dataset:
In order to extract query embeddings, we use a random sample of 1 week of search log data from May 2016 of a commercial web search engine comprising of user ID information along with session identifier and query text. The dataset composed of 24M search impressions spread over 8M search sessions, issued by over 200K users, resulting in a vocabulary size of over 5M words. We train the embeddings model of Word2Vec using all the queries in this corpus based on the different search contexts considered. As per the free parameters, the dimension of the word vectors was set to values in 100, 300, the number of negative examples is in 5. Since query text is used to learn embedding, we keep the window size as 2 which totals to 4 words as context per query term. Sub-sampling of frequent terms was not performed and all other parameters were set to default values.

Baselines:
We consider a number of baselines, including embedding based approaches composed of various different contexts.

(1) Global Embeddings: We use a word2vec model trained on the document collection retrieved by the queries as global embeddings.
(2) Session embeddings: We use search sessions as context while learning embeddings.
(3) Random: To validate the usefulness of considering query sequence information, we randomly shuffle queries issued by a user before inputting to embedding learning model.

Qualitative Analysis
We begin with a qualitative analysis of the extracted representations by showing *nearby* query terms. Table 1 shows the top 3 query terms which are most similar to two randomly chosen queries. We observe that the suggestions shown using task embeddings are more coherent and related than the ones from global embeddings. In web search context, suggestions like '*weather*' and '*tube*' are more contextually relevant to be one of the aspects a user might be looking for rather than suggestions comprising of similar city name suggestions.

Query Suggestions
To evaluate our approach we make use of the query representations obtained to generate lists of query suggestions. Specifically, we make use of the TREC Tasks track data from two years (2015 & 2016) to rank query suggestions that would help users fulfil their information need. Tasks track data provides test collections for evaluating the usefulness of retrieval systems in terms helping people achieve their search tasks. The dataset comprise of 100 different tasks embodied by a query each, with each task containing a list of possible candidate queries that represent the set of all tasks a user who submitted the query may be looking for. Each of these candidate queries are judged for relevance labels by human assessors. Overall, the dataset spans over 2 years (2015 and 2016).

For each query, we consider the entire pool of candidate queries and use the query representations to find the similarity of the query with each candidate query based on which they are rank. The relevance labels provided with each candidate query are used to compute the average relevance and NDCG@k metrics.

	TREC Tasks 2015			TREC Tasks 2016		
Method	**AvgRel@3**	**AvgRel@5**	**AvgRel@10**	**AvgRel@3**	**AvgRel@5**	**AvgRel@10**
Random	0.613	0.56	0.542	1.13	0.99	0.926
Global	0.631	0.572	0.558	1.15	1.01	0.932
Session	0.633	0.573	0.564	1.14	1.02	0.934
Task	**0.662**$^{*\&}$	**0.591**$^{*\&}$	**0.571**$^{*\&}$	**1.16**$^{\&}$	**1.04***	**0.944**$^{*\&}$

Table 2: Average Relevance results. * and & indicate statistical significant ($p \leq 0.05$) using paired t-tests compared to the Global and Session context baselines respectively.

	TREC Tasks 2015			TREC Tasks 2016		
Method	**NDCG@3**	**NDCG@5**	**NDCG@10**	**NDCG@3**	**NDCG@5**	**NDCG@10**
Random	0.289	0.3	0.302	0.511	0.509	0.522
Global	0.292	0.301	0.309	0.524	0.513	0.52
Session	**0.314**	0.309	0.318	0.510	0.514	0.527
Task	**0.314***	**0.311**$^{*\&}$	**0.321**$^{*\&}$	**0.526**$^{\&}$	**0.529**$^{*\&}$	**0.535**$^{*\&}$

Table 3: NDCG@k results. * and & indicate statistical significant ($p \leq 0.05$) using paired t-tests compared to the Global and Session context baselines respectively.

Tables 2 presents the average precision scores for the different baselines considered. We observe that though global representations perform better than traditional and simpler representation techniques, they perform worse than session based and task based embeddings. This highlights the importance of considering local context when learning representations, since generic contexts are usually heterogeneous and ill fitted to retrieval problems. Among the neural local context models, task based context performs better than session based contexts. This confirms our hypothesis that sessions are usually polluted with queries from various tasks, and as a result the resulting context isn't informative enough.

While relevance scores are important, often system designers have a constraint to rank top-k suggestions. To this end, in addition to average relevance scores, we make use of the candidate ranking to compute NDCG scores and present results in Table 3. Similar to our previous observation, we observe that neural representation methods generally perform better than non-neural models. Amidst session based and task context based, task based representation performs better than the corresponding session context.

5 CONCLUSION

Search context has played an important role in solving various retrieval tasks. In this work, we leveraged task context to learn query representations. Experimental evidence suggests tasks context enriched representations perform better than traditional representations, and at the same time, task context is more informative than session context. These findings have implications in designing better personalization and recommendations techniques aimed at exploiting task context for enhanced user support.

REFERENCES

[1] Eugene Agichtein, White, Dumais, and Bennet. Search, interrupted: understanding and predicting search task continuation. In *SIGIR 2012*.

[2] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. Context-aware query classification. In *SIGIR 2009*.

[3] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *KDD 2008*.

[4] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of ACL 2016*.

[5] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. Context-and content-aware embeddings for query rewriting in sponsored search. In *SIGIR 2015*.

[6] Parth Gupta, Kalika Bali, Rafael E Banchs, Monojit Choudhury, and Paolo Rosso. Query expansion for mixed-script information retrieval. In *SIGIR 2014*.

[7] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM 2008*.

[8] Diane Kelly, Vijay Deepak Dollu, and Xin Fu. The loquacious user: a document-independent source of terms for query expansion. In *SIGIR 2005*.

[9] Tom Kenter and Maarten de Rijke. Short text similarity with word embeddings. In *CIKM 2015*.

[10] Alexander Kotov, Paul N Bennett, Ryen W White, Susan T Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *SIGIR 2011*.

[11] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *WSDM 2011*.

[12] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. Characterizing users' multi-tasking behavior in web search. In *CHIIR 2016*.

[13] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Deconstructing Complex Tasks. In *Proceedings of NAACL*.

[14] Rishabh Mehrotra and Emine Yilmaz. 2017. Extracting Hierarchies of Search Tasks & Subtasks via a Bayesian Nonparametric Approach. In *Proceedings of SIGIR 2017*. ACM, 285–294.

[15] Lilyana Mihalkova and Raymond Mooney. 2009. Learning to disambiguate search queries from short sessions. *ML-KDD* (2009).

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*.

[17] Bhaskar Mitra. Exploring session context using distributed representations of queries and reformulations. In *SIGIR 2015*.

[18] Bhaskar Mitra and Nick Craswell. Query auto-completion for rare prefixes. In *SIGIR 2015*.

[19] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive information retrieval using implicit feedback. In *SIGIR 2005*.

[20] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In *WWW 2014*.

[21] Amanda Spink, Sherry Koshman, Minsoo Park, Chris Field, and Bernard J Jansen. Multitasking web search on vivisimo. com. In *ITCC 2005*.

[22] Ivan Vulić and Marie-Francine Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR 2015*.

[23] Chong Wang and David M Blei. Variational inference for the nested Chinese restaurant process. In *NIPS 2009*.

[24] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. Learning to extract cross-session search tasks. In *WWW 2013*.

[25] Ryen W White, Paul N Bennett, and Susan T Dumais. Predicting short-term interests using activity-based search context. In *CIKM 2010*.

[26] Guoqing Zheng and Jamie Callan. Learning to reweight terms with distributed representations. In *SIGIR 2015*.