Tempo-Lexical Context driven Word Embedding for Cross-Session Search Task Extraction

Procheta Sen
ADAPT Centre
School of Computing
Dublin City University
Dublin 9, Ireland
procheta.sen2@mail.dcu.ie

Debasis Ganguly
IBM Research Labs
Dublin, Ireland
debasis.ganguly1@ie.ibm.com

Gareth J.F. Jones
School of Computing
ADAPT Centre
Dublin City University
Dublin 9, Ireland
Gareth.Jones@dcu.ie

Abstract

Search task extraction in information retrieval is the process of identifying search intents over a set of queries relating to the same topical information need. Search tasks may potentially span across multiple search sessions. Most existing research on search task extraction has focused on identifying tasks within a single session, where the notion of a session is defined by a fixed length time window. By contrast, in this work we seek to identify tasks that span across multiple sessions. To identify tasks, we conduct a global analysis of a query log in its entirety without restricting analysis to individual temporal windows. To capture inherent task semantics, we represent queries as vectors in an abstract space. We learn the embedding of query words in this space by leveraging the temporal and lexical contexts of queries. To evaluate the effectiveness of the proposed query embedding, we conduct experiments of clustering queries into tasks with a particular interest of measuring the cross-session search task recall. Results of our experiments demonstrate that task extraction effectiveness, including cross-session recall, is improved significantly with the help of our proposed method of embedding the query terms by leveraging the temporal and temp-lexical contexts of queries.

1 Introduction

A complex search task is defined as a "a multi-aspect or a multi-step information need consisting of a set of related subtasks, each of which might recursively be complex" (Hassan Awadallah et al., 2014). For example, a task of making arrangements for travel to a conference qualifies as a complex search task because there are several choices that a user needs to make in order to plan his entire trip, e.g. selecting flight, hotel, making arrangements for local transport, finding the conference venue, finding good places to eat around, finding local sight-seeing options after the conference etc. All these sub-tasks are likely to take place within their own search sessions, where a session is defined as a set comprised of queries executed during a time period of a specific length, usually about half-an-hour (Lucchese et al., 2013).

In this paper, we address the problem of automatically predicting whether search sessions, focused on specific activities, are a part of a broader complex search task, which we refer to as the *cross-session search task extraction problem*. Cross-session search task extraction can potentially find applications in designing more proactive search engines, which may suggest relevant information about specific subsequent subtasks along a timeline, e.g. suggesting places to eat around a conference venue without the user needing to execute these queries.

To see why cross-session search task extraction is a challenging problem, firstly, note that it is likely that a query session for flight booking and one for local sightseeing around a conference venue may be far apart in time, as a result of which simple approaches of grouping queries by their timestamps, e.g. (Lucchese et al., 2013), are not likely to yield satisfactory outcomes. Secondly, the term overlap between the queries of these two sessions is also likely to be low, indicating that using lexical similarity for clustering cross-session queries into a single group, e.g. (Lucchese et al., 2013; Wang et al., 2013), is unlikely to be effective.

As an illustrative example of term mismatch, consider the two queries 'Eric Harris', 'Reb Vodka' from the AOL query log[1]. Although these two queries do not share any common terms between them, they refer to the task of finding infor-

[1] https://archive.org/download/AOL_search_data_leak_2006

Proceedings of NAACL-HLT 2018, pages 283–292
New Orleans, Louisiana, June 1 - 6, 2018.

mation on the Columbine high school massacre, the first query referring to the name of the first murderer while the second one refers to their nickname.

Our Contributions. To alleviate the identified problems with attempting to group queries by their timestamps or lexical similarities, we propose to embed queries in a task-based semantic space in a manner that will give two similar queries in this space a high likelihood of pertaining to the same underlying task. Word embedding algorithms, such as 'word2vec' (Mikolov et al., 2013), make use of the lexical context in learning vector representations of words. We propose to *transform* these word vectors into a task-oriented semantic space with the objective of making two words that are likely to be a part of the same search task closer to each other.

To learn the transformation function, we make use of average session duration and lexical similarities between within-session queries. Our method thus provides a unifying framework for addressing tempo-lexical similarity, in contrast to previous approaches that treat these two separately. Another important contribution of our proposed method is that we are able to empirically demonstrate that our proposed method is more effective than existing algorithms (Lucchese et al., 2013; Mehrotra and Yilmaz, 2017) in extracting cross-session search tasks without the application of any external information for estimating task relatedness. For instance, the work in (Lucchese et al., 2013) relies on Wikipedia to contextualize queries, while the one in (Verma and Yilmaz, 2014) uses Wikipedia-based entity recognition to estimate task relatedness.

The rest of the paper is organized as follows. In Section 2 we overview previous work in task extraction and query embedding. In Section 3, we introduce our semantic context driven transformation-based word vector embedding algorithm to enhance cross-session query similarity matching. Section 4 then describes how the transformed query vectors are clustered into search tasks. Section 5 describes our experimental setup. Section 6 presents the results of our experiments. Section 7 concludes the paper with suggestions for future work.

2 Related Work

In this section we review existing work in search task extraction and query embedding and contrast this with the method introduced in this paper. We look first at work on unsupervised task extraction and then consider work on supervised methods, and final briefly consider a study introducing a query embedding method.

2.1 Unsupervised Task Extraction

A method for extracting tasks within each search session is proposed in (Lucchese et al., 2013). A session is defined with fixed length time windows. After investigating a wide range of time length values, the optimum is reported to be 26 minutes, which is what we also use in our work. The study reported in (Lucchese et al., 2013) also investigated a number of clustering techniques to group together related queries from each session into tasks. A wide range of features were investigated to define the similarity between a pair of queries, e.g. edit distance, cosine-similarity and Jaccard coefficient of character level trigrams. In contrast to (Lucchese et al., 2013; Wang et al., 2013), we investigate the use of embedded query vectors to compute similarity, rather than depending on character and word level lexical similarity features, e.g. edit distance, term overlap, trigram character overlap etc. Another difference of our method from (Lucchese et al., 2013) is that instead of restricting clustering to each session, we cluster the entire dataset globally, which implies that our method is not limited by variations in session duration. We also evaluate the effectiveness of clustering the entire dataset rather than on aggregating clustering effectiveness separately for each session as in (Lucchese et al., 2013).

Extraction of task hierarchies was investigated in (Mehrotra et al., 2016). Given a set of task related queries, they composed query vectors as weighted combinations of the constituent query term vectors, the weights being the maximum likelihood estimates from query-task relationships. A Chinese Restaurant Process (CRP) based posterior inference process was then used to extract the tasks from individual queries. In an extension of this work (Mehrotra and Yilmaz, 2017), the authors proposed a Bayesian non-parametric approach for extracting task hierarchies. The main difference between our approach and (Mehrotra et al., 2016; Mehrotra and Yilmaz, 2017) is that

our focus is on finding cross-session tasks from a query log, rather than finding hierarchies of tasks. Further, instead of using similarities between embedded query vectors as one of the features to estimate the relatedness between two queries, we propose a task semantics driven embedding technique to transform a query in close proximity to its task-related counterpart.

An entity extraction method was applied in (Verma and Yilmaz, 2014) to estimate similarities between queries for the purpose of task extraction. In contrast to this, our method does not rely on an entity extractor to extract cross-session tasks.

2.2 Supervised Task Extraction

A supervised approach for automatically segmenting queries into task hierarchies was proposed in (Jones and Klinkner, 2008). They trained logistic regression models to determine whether two queries belong to same task or not. According to (Wang et al., 2013), the disadvantage of using a classifier based approach for extracting tasks is that with the binary predictions of the classifier it is difficult to model the transitive task dependence between the queries, e.g. if query pairs (q_1, q_2) and (q_2, q_3) are predicted to be part of the same task, the classifier may not predict that q_1 and q_3 are also a part of the same task. Graph-based clustering on the binary adjacency matrix between query pairs (obtained from logistic regression output) is also likely to introduce noise during clustering (Wang et al., 2013). The limitations of (Jones and Klinkner, 2008) were alleviated in the work reported in (Wang et al., 2013), which employs a structural SVM framework for estimating the weights of different lexical features to measure the similarity between two queries.

The difference between the studies reported in (Jones and Klinkner, 2008; Wang et al., 2013) and our work is that we propose a completely unsupervised approach for clustering queries. This implies that our method does not rely on the availability of training data, the construction of which requires considerable manual effort.

2.3 Query Embedding

A relevance-based word embedding technique was developed in (Zamani and Croft, 2017). This method uses the top documents retrieved for each query to learn the association between the query terms and those occurring in the retrieved documents. In contrast to retrieving ranked lists for every query as in (Zamani and Croft, 2017), we capture the semantic context of query words with the help of other useful cues for task-relatedness, e.g. the time-gap between queries.

3 Embedding Query Words

'Word2vec' is a standard approach to obtain embedded words vectors (Mikolov et al., 2013). The word2vec approach aims to create similar vector representations of words that have similar context, and are thus assumed to be significantly semantically related. In this section, we explain why the standard word2vec method may not be suitable for embedding queries in an abstract space of task-semantics for the purpose of using these vectors to extract cross-session search tasks. To address this problem, we propose a method of word embedding that is able to capture larger semantic contexts for better estimation of the word vectors.

3.1 Problems with Short Documents

Let $\mathbf{w} \in \mathbb{R}^d$ denote the vector representation of a word $w \in V$, V and d being the vocabulary and the dimension of the embedded vectors, respectively. Let W be a $d \times V$ matrix, where each d dimensional column vector represents a word vector. Let D be an indicator random variable denoting semantic relatedness of a word with its context. Given a pair of words, (w, c), the probability that the word c is observed in the context of word w is given by $\sigma(\exp(-(\mathbf{w} \cdot \mathbf{c})))$. Word embedding for a given corpus is obtained by sliding a window along with its context through each word position in the corpus maximizing the objective function shown in Equation 1.

$$J(\theta) = \sum_{w_t, c_t \in D^+} \sum_{c \in c_t} log(P(D = 1 | \mathbf{w_t}, \mathbf{c_t})) - \sum_{w_t, c'_t \in D^-} \sum_{c \in c'_t} log(P(D = 1 | \mathbf{w_t}, \mathbf{c'_t})) \quad (1)$$

In Equation 1, w_t is the word in the t^{th} position in a training document corpus, c_t is the set of observed context words of word w_t within a word window, c'_t is the set of randomly sampled words from outside the context of w_t. D^+ denotes the set of all observed word-context pairs (w_t, c_t), whereas D^- consists of pairs (w_t, c'_t).

The word2vec algorithm respects document boundaries by not extending the context vector

across them. In the context of our empirical study, we aim to learn word vector embedding from a query log, where each document in the 'word2vec' terminology refers to a single query. In the case of keyword-based search queries, often comprising 2-3 words, the average number of context vectors is much lower than the number of contexts available for the standard word embedding scenario of full length documents, e.g. news articles and web pages. Consequently, this may result in ineffective estimation of the word-context semantic relations for the queries.

3.2 Word Vector Transformation with Semantic Contexts

To alleviate the problems of short contexts when embedding queries, we propose to learn a transformation matrix to transform a set of word vectors to, generally speaking, another abstract space. The aim is to transform a word vector $\mathbf{w}$ so that it is close to a set of other words that *respects* the characteristics of this abstract space. In the context of our problem, the abstract space refers to the embedding space of task-relatedness with the characteristics that queries that are a part of the same search task should be embedded close to each other.

We adopt a general terminology of referring to the desired similarity in the abstract space as *semantic similarity*, which in the context of our problem refers to task-relatedness and is not to be confused with linguistic semantics. Formally, the set of words similar to the word w is represented by the set $\Phi(w)$ shown in Equation 2,

$$\Phi(w) = \{v : (w, v) \in S\}, \tag{2}$$

where S denotes the semantic relation between a pair of words. In particular, the set $\Phi(w)$ depends on the definition of the semantic relation S between two words, which we will describe in Section 3.3.

Assuming the existence of a pre-defined semantic relation S between word pairs, we define the loss function for a word vector $\mathbf{w}$ as shown in Equation 3.

$$l(\mathbf{w};\theta) = \sum_{\mathbf{v}:v\in\Phi(w)} \sum_{\mathbf{u}:u\notin\Phi(w)} \max\big(0, m - ((\theta\mathbf{w})^T\mathbf{v} - (\theta\mathbf{w})^T\mathbf{u})\big) \tag{3}$$

Equation 3 defines a hinge loss function with margin m (set to 1 in our experiments). The loss function is parameterized by the transformation matrix $\theta \in \mathbb{R}^{d\times d}$, and is learned by iterating with stochastic gradient descent. The word vectors used in learning the parameter matrix θ are obtained by the word2vec skip-gram algorithm. After training, each word vector $\mathbf{w}$ is transformed to $\mathbf{w}'$ in Equation 4.

$$\mathbf{w}' = \theta \cdot \mathbf{w} \tag{4}$$

Informally speaking, the objective function aims to maximize the similarity between two word vectors $\mathbf{w}$ and $\mathbf{v}$ that are members of the same semantic context. On the other hand, it minimizes the similarity between the word vector $\mathbf{w}$ and a word vector $\mathbf{u}$ randomly sampled from outside its context, as defined by the semantic relation S of Equation 2. In principle, the objective function of Equation 3 is similar to the word2vec objective function of Equation 1, the difference being in the definition of the context vector. While the word2vec algorithm relies on an adjacent sequence of words to define a context, in our proposed approach, we rely on a pre-defined set of binary relations between words.

Another analogy of Equation 3 can be drawn with the multi-modal embedding loss function proposed in (Frome et al., 2013), where the words from the caption of an image constitute the notion of the 'semantic context' of the image vector used to transform it. For our problem, we make use of this context to associate the task-specific relationship between query words.

3.3 Temporal Semantic Context

In the particular context of query logs, temporal similarity is likely to play an important role in topically grouping queries. This is because queries in the same search session are usually related to the same topic, as observed in previous studies (Lucchese et al., 2013; Wang et al., 2013). For example, it can be observed from the AOL query log that the words 'reb' and 'vodka' belong to the same search session as the words 'eric' and 'harris' (see example in Section 1). In this case, the semantic relationship S, as described in Section 3.2, considers terms u (e.g. 'vodka') and v (e.g. 'harris') from the same query session to be semantically related. To define the semantic relation S, we take into account a temporal context specified by a time window of 26 minutes as reported in (Lucchese et al., 2013). Specifically, if two queries belong to the same search session, as defined by a fixed length time window, then each

constituent word pair within them is considered to be members of the set S.

3.4 Tempo-Lexical Semantic Context

In real-life settings, even within a session of a specified time length, users often multi-task their activities (possibly by using multiple browser tabs or windows) (Lucchese et al., 2013; Wang et al., 2013; Mehrotra and Yilmaz, 2017). To address this issue, we further cluster the queries of each search session into mutually disjoint groups. Our hypothesis is that this grouping of the queries of a single search session into multiple clusters may improve word embedding of the query terms further by restricting the semantic relationship S (Equation 2) to consider terms from related queries within each cluster separately.

Our clustering approach is based on a weighted graph of query similarities computed by a linear combination of content-based similarity (Sim_r) and retrieved document list based similarity (Sim_r) as described in Section 4 and Section 5.3. The clusters then provide the tempo-lexical contexts which are subsequently used to improve the quality of embedding of the query words.

4 Clustering of Embedded Queries

In this section, we describe our unsupervised approach to identifying cross-session tasks by clustering the query vectors, where the constituent query word vectors are obtained using the word embedding approaches described in Section 3.

4.1 Query Vector Embedding

We hypothesize that the modified word vector embedding approach of Section 3.2 will be more effective in capturing the session specific semantics of query terms since it takes into account the temporal context of query session information from query logs. We adopt a standard word vector combination method to form embedded query vectors. Because of the compositionality property of word vectors (Mikolov et al., 2013), the simple method of averaging over the constituent word vectors has been reported to work well for various tasks such as term re-weighting and query reformulation (Zheng and Callan, 2015; Grbovic et al., 2015).

4.2 Clustering of Query Vectors

Unlike previous approaches of grouping together queries according to fixed time windows, and then clustering the queries within each time window separately (Lucchese et al., 2013; Wang et al., 2013), we take a more general approach of clustering the overall set of query vectors.

Since the number of query clusters cannot be known a priori, the number of clusters is estimated by adopting a clustering approach that does not require the number of clusters to be specified. We adopt the best performing clustering method identified in (Lucchese et al., 2013) referred to as QC-$_{WCC}$. This is a graph based clustering algorithm that extracts the weighted connected components of a graph after constructing a complete graph and then pruning off the edges that are below a predefined threshold, η.

In QC-$_{WCC}$, the weights between the graph edges are defined by a linear combination of two types of similarities: i) content-based (Sim_c), and ii) retrieval based (Sim_r), as shown in Equation 5, in which the overall similarity is controlled by the linear combination parameter α.

$$Sim(q_i, q_j) = \alpha Sim_c(q_i, q_j) + (1 - \alpha) Sim_r(q_i, q_j) \quad (5)$$

- *Content-based* similarity: Measured with the help of character trigrams and normalized Levenshtein similarity between query pairs.
- *Retrieval-based* similarity: Each query is contextualized with a Wikipedia collection. More specifically, two queries are considered similar if the top 1000 documents retrieved by them are also similar.

In contrast to the experimental setup of (Lucchese et al., 2013), a) we conduct clustering globally instead of clustering each individual query session separately; and b) the edge weights of the graph-based clustering in our case refers to the cosine-similarity values computed between the embedded query vectors and cosine similarity values between the vectors obtained from top 1000 documents retrieved from Clueweb12B, a publicly available web collection[2].

5 Experimental Setup

In this section we describe the setup for our experimental study. We begin with an overview of our datasets, then introduce the experimental baselines used and the objectives of our experiments, finally

[2] http://boston.lti.cs.cmu.edu/clueweb12/

we set out the parameter settings used in our experiments.

5.1 Dataset

Similar to previous reported studies (Lucchese et al., 2013; Mehrotra and Yilmaz, 2017; Mehrotra et al., 2016; Verma and Yilmaz, 2014), we use the AOL query log for our experiments. In order to compare our results with these studies, we use the same subset of 1424 queries from the AOL query log for the evaluation of task extraction effectiveness as used in these earlier studies. However, since the purpose of these studies was only to extract tasks from a single session, in its annotation scheme two queries only qualified as part of the same task if they appeared within the same session.

In contrast, since we investigate cross-session task extraction, the time length threshold is not applicable to our annotation scheme, as a result of which, we re-annotated the task labelled dataset of (Lucchese et al., 2013). In particular, our annotation scheme was solely based on the underlying search intent of the query.

While re-annotating the dataset of (Lucchese et al., 2013), the annotators were instructed not to change the task labels within each session. Instead, the annotators were asked to re-label task identifiers spanning across different query sessions. For example, the annotation of (Lucchese et al., 2013) considered 'robert f kennedy jr' and 'robert francis kennedy' to belong to two different tasks since these queries were executed during different sessions. However, our annotation scheme considers them to be a part of the same task.

Two persons were employed to carry out our annotation step of the set of 1424 queries in two different batches. They were asked to come to a consensus when trying to merge the task labels across their individual batches. The annotators were instructed to use a commercial search engine (e.g. Google), if required, to determine if two queries from different search sessions could potentially relate to the same underlying task. Table 1 provides an overview of our annotated task labels; this shows that there are a considerable number of sessions that contain queries spanning across session boundaries. It can be seen from Table 1 that after post-processing the single session task labels, the total number of distinct tasks is reduced. This is indicative of the fact that the modified dataset is able to consider queries from different search sessions as a part of the same search task (there are 36,768 of them as shown in Table 1). The post-processed dataset with cross-session task labels that we use for our experiments is publicly available[3].

	Task label granularity	
Item	Within-session	Cross-session
#Queries	1424	1424
#Tasks annotated	554	224
#Sessions	307	307
#Sessions with cross-session tasks	0	239
#Query pairs across sessions judged in the same task	0	36768

Table 1: Dataset statistics of task annotated queries from the AOL query log. Cross-session task labels are post-processed annotations of the dataset prepared by (Lucchese et al., 2013).

5.2 Baselines and Experiment Objectives

Since our proposed task extraction method is unsupervised, for a fair comparison we only employ unsupervised approaches as baselines. More specifically, we did not consider the supervised approaches reported in (Jones and Klinkner, 2008; Wang et al., 2013) as our baselines.

As our first baseline, we re-implemented QC-WCC, the best performing approach (Lucchese et al., 2013) (briefly described in Section 4.2). This study investigated a wide range of features, clustering methods and parameter settings. We adopt the same linear combination of similarities in our study as shown in Equation 5.

Our re-implementation of this work involves a slight change to the original one. Instead of using a Wikipedia document collection, we employ a much larger collection of crawled web documents, namely the ClueWeb12B collection, comprising of nearly $52M$ documents[4]. Our reasons for using the ClueWeb collection are as follows. Firstly, our study is carried out using queries from a Web search log and hence it is reasonable to expect that a web collection will provide better estimates of semantic similarities between the queries. Secondly, a number of our queries in our dataset are not of expository type, and hence the num-

[3] https://github.com/procheta/AOLTaskExtraction/blob/master/Task.csv

[4] http://boston.lti.cs.cmu.edu/clueweb12/

ber of matching Wikipedia articles is expected to be low for them due to vocabulary mismatch. On the other hand, the web collection, being diverse, is expected to retrieve more matching articles for these types of queries. To compare the performance of our implementation of QC-$_{WCC}$ with ClueWeb12B with the original one, we adopted an experimental and evaluation setup identical to that of (Lucchese et al., 2013), the only difference being in the collection used for deriving the semantic similarities. For the retrieval model we used the LM-JM (Language Model with Jelineck-Mercer smoothing) with the smoothing parameter set to 0.6 as suggested in (Lavrenko and Croft, 2001).

To demonstrate the potential benefits of our proposed tempo and tempo-lexical context-driven word embedding based approaches for the query terms, we employed the following three baselines.

1. **Qry vec skip-gram**: In this approach, query vectors were obtained by summing over the constituent word vectors obtained using the standard skip-gram (Mikolov et al., 2013).
2. **Qry vec (All-in-one Session Context)**: We hypothesized that additional context is likely to capture task-specific semantics of the query terms. A boundary condition arises when the entire query log is assumed to belong to one session. To show that the temporal context needs to be focused, in this approach, we investigate the effect of setting the context set S to the entire vocabulary of query terms.
3. **Qry vec (Pre-trained Google news vectors)** We hypothesized that additional context is likely to be useful to learn the vector representations of constituent words of short documents (in this case, queries). To see if pre-trained word vectors from an external generic corpus can be useful to alleviate the problem of short documents, we employ pre-trained word vectors from the Google news corpus to obtain the vector representation of the queries.

The objective of the experiments is to show that our proposed query term embedding method can outperform the above mentioned baselines, thus indicating that within-session adjacency information can be useful to learn task specific semantics.

5.3 Parameters and Evaluation Metrics

Parameters. In our method, we use the cosine similarity between the embedded query vectors instead of using character 3-grams and Levenshtein similarity, as used in (Lucchese et al., 2013)), to compute $Sim_c(q_i, q_j)$ between any two query pairs q_i and q_j. We employ three different embedding strategies for our experiments: i) standard word2vec, ii) transformed vectors with temporal context, and iii) transformed vectors with tempo-lexical contexts. In all our experiments, we tune α from 0 to 1 in steps of 0.1. The second parameter common to all the methods, the threshold η, which is used in QC-$_{WCC}$ clustering to prune off edges from the weighted similarity graph between query pairs. We tuned η in the range 0.1 to 1 in steps of 0.1 for each method separately.

For our word vector based experiments, we used the skip-gram model to train the word vectors using the entire AOL query log comprising over 6M queries. The dimensionality of the word vectors was set to 200. The initially obtained word vectors were used as starting inputs to learn the temporal and tempo-lexical transformations. For the tempo-lexical based transformation method, we used the optimal value of η as obtained from the QC-$_{WCC}$ baseline, to cluster the queries in each temporal window of 26 minutes.

Evaluation Metrics. Since we use weighted clustering to extract cross-session search tasks, we used standard clustering evaluation metrics to evaluate the effectiveness of the task extraction. Clustering is typically evaluated with the effectiveness of the pair-wise decisions of assigning data points to the same or different clusters. In our case, the number of true positives was given by the number of query pairs in the ground-truth that were judged to belong to the same task and were also predicted by the system to be a part of the same task. Similarly, we computed the false positives and the true negatives. Based on these counts, we computed the standard metrics of precision, recall, and F-score (similar to (Lucchese et al., 2013)).

Additionally, to measure how many of the total number of cross-session queries that were part of the same search tasks were discovered by these approaches, we computed the cross-session recall (denoted as 'CS-Recall'). This metric was computed as the ratio of the number of correctly identified cross-session similar-task query pairs against

Query Similarity	Parameters		Metrics			
	α	η	F-score	Prec	Recall	CS-Recall
QC-$_{WCC}$(3gram+ Levenestine) (Lucchese et al., 2013)	0.8	0.4	0.471	0.387	0.603	0.1930
Qry vec skip-gram	0.7	0.8	0.524*	**0.465***	0.602	0.7161*
Qry vec (All-in-one Session Context)	0.7	0.5	0.499	0.430	0.595	0.6400
Qry vec (Pre-trained Google news vectors)	0.6	0.5	0.473	0.410	0.558	0.6400
Qry vec with temporal context	1.0	0.7	0.536*†	0.461*	0.643*†	0.7393*†
Qry vec with tempo-lexical context	0.6	0.7	**0.538***†	0.441*	**0.691***†‡	**0.7395***†‡

Table 2: Comparison between the best results obtained after parameter tuning on different unsupervised approaches of task extraction. For all methods, $1-\alpha$ represents the weight of the semantic similarity estimated from ClueWeb12B. *†‡ indicates statistical significance (paired t-test with 95% confidence) with respect to (Lucchese et al., 2013), 'Qry vec skip-gram' and 'Qry vec with temporal context' respectively.

Task extraction method	Session-F-socre
QC-$_{WCC}$ on trigram+Levenshtein with Wikipedia (Lucchese et al., 2013)	0.812*
QC-$_{WCC}$ on trigram+Levenshtein with ClueWeb12B	0.834
Non-parametric clustering on average query term vectors (Mehrotra et al., 2016)	0.845†
QC-$_{WCC}$ on average of baseline skip-gram query word vectors	0.837
QC-$_{WCC}$ on transformed word vectors with temporal context	**0.847**
QC-$_{WCC}$ on transformed word vectors with tempo-lexical context	0.840

Table 3: Within-session Task Extraction effectiveness.

the total number of them (36,768 as reported in Table 1).

In order to extend evaluation of our proposed approach to within-session task extraction, for comparison with existing studies, we computed the clustering metrics for each individual session and then computed the weighted average of these values over each session as reported in (Lucchese et al., 2013; Mehrotra and Yilmaz, 2017). Although these earlier studies refer to this weighted measure as F-score, we refer to this version of F-score as 'Session-F-score'.

6 Results

In this section, we report the results of our investigations of our proposed query vector based cross-session search task extraction. We first investigate the effectiveness of our proposed approach on the cross-session search task extraction and then report and compare results with existing approaches for within-session task extraction.

6.1 Cross-Session Task Extraction

Table 2 shows the results of weighted clustering QC-$_{WCC}$ with optimal α and η settings for each individual method. It can be seen that the first baseline approach QC-$_{WCC}$ performs poorly because trigram and Levenshtein similarities lack the semantic information required to effectively cluster task-related queries into the same cluster. Clustering effectiveness improves considerably when weighted clustering is conducted using the cosine similarities between the query vectors, i.e. the 'Qry vec skip-gram' approach. This suggests that the word vectors are better able to capture the semantic relatedness between the task-related query terms.

It can be seen that using the entire query log as one context, i.e. the approach 'Qry vec (All-in-one Sesion Context)' yields worse results than the baseline skip-gram approach, which shows that a focused context is required for effectively embedding the query terms. Results with pre-trained word vectors on a large news corpora, i.e. the approach 'Qry vec (Pre-trained Google news vectors)', show that additional out-of-domain and generic context is not helpful for improving the quality of the embedded query term vectors.

Transformation of the word vectors leveraging the semantic contexts (i.e. our proposed method in Section 3) outperforms the clustering effectiveness obtained with the baseline approaches. The most important observation is that the use of temporal context in learning word vectors results in best performance for $\alpha = 1$, i.e. when no retrieval-based similarity is used (see Equation 5). This suggests that optimally trained word vectors can produce effective task clusters without the use of external collections in contextualizing the queries. The use of tempo-lexical contexts, i.e. when the semantic context used to learn the transformation matrix for the word vectors is restricted to sim-

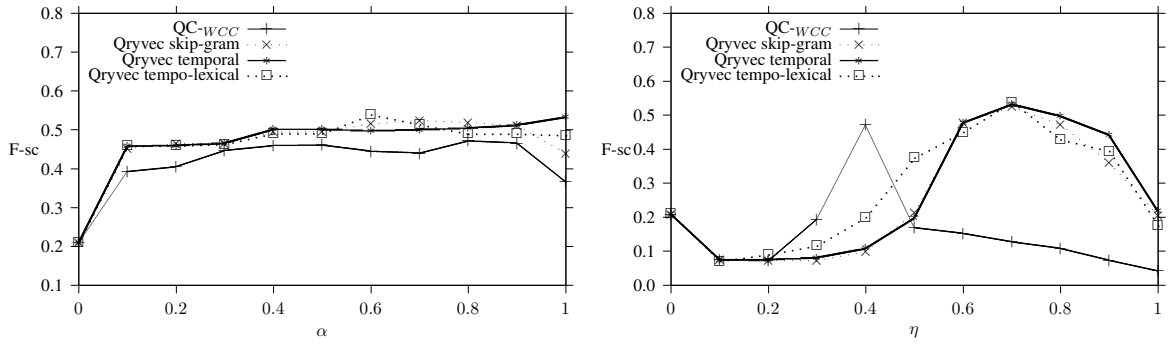

Figure 1: Sensitivity of task clustering with variations in α (left) and η (right).

ilar queries within search sessions, the effectiveness improves further. In particular, Table 2 shows that both tempo and tempo-lexical transformations are able to improve recall significantly suggesting that the transformation helps to group more truly task-related queries into the same cluster.

Next, we show the effect of varying the parameters α and η separately in Figure 1. The values of η for each corresponding method in the left graph of Figure 1 are those reported in Table 2. Similarly, for the plot on the right of Figure 1, the α values correspond to those reported in Table 2. A value of $\alpha = 1$ considers only the content based similarity (see Equation 5). It can be observed from Figure 1 (left) that at $\alpha = 1$, the F-score values for all the query embedding based approaches are higher than the baseline method of QC-$_{WCC}$. This indicates that the query embedding based approaches perform well without relying on similarity-based retrieval using an external collection. In general, it can be observed that over a wide range of α and η settings, the F-score values of the embedding based methods outperform the QC-$_{WCC}$ method.

6.2 Within-session task extraction

In this experimental setup, we make use of the session duration span of 26 minutes, similar to (Lucchese et al., 2013), to restrict query clustering to each individual session. Similar to (Lucchese et al., 2013; Mehrotra and Yilmaz, 2017), we employ the session averaged clustering metrics for measuring the effectiveness of the different approaches (see Section 5.3). We use the within-session ground-truth of (Lucchese et al., 2013) to evaluate the task extraction effectiveness.

Table 3 reports the results for various within-session task clustering approaches. The results with * and † are taken from the results reported in (Lucchese et al., 2013) and (Mehrotra et al., 2016). The following observations can be made with regard to Table 3. Firstly, the use of ClueWeb12B contributed to an improvement in task extraction effectiveness, thus demonstrating that our re-implementation of (Lucchese et al., 2013) is comparable with that of the original. Secondly, an important observation is that the use of average query term vectors along with contextual information from ClueWeb12B outperforms the approach of trigram and Levenshtein based similarity computation of (Lucchese et al., 2013). Thirdly, it can be observed that results improve with the application of transformation based word vector embedding of the query terms. The temporal context proves more effective than the tempo-lexical one.

7 Conclusions and Future Work

In this paper, we studied the problem of cross-session task extraction. We proposed a transformation based word embedding approach that takes into account the temporal and tempo-lexical contexts of queries to learn task-specific semantics. Our experiments on the AOL query log indicate that the proposed temporal and tempo-lexical query embedding method significantly outperform the baseline word2vec embedding. As future work, we would like to investigate supervised methods for cross-session task extraction.

Acknowledgement

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant No. 13/RC/2106) (www.adaptcentre.ie).

References

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Proc. of NIPS'13*. pages 2121–2129.

Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context- and content-aware embeddings for query rewriting in sponsored search. In *Proc. of SIGIR '15*. pages 383–392.

Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proc. of CIKM'14*. pages 829–838.

Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proc. of CIKM '08*. pages 699–708.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proc. of SIGIR '01*. pages 120–127.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2013. Discovering tasks from search engine query logs. *ACM Trans. Inf. Syst.* 31(3):14:1–14:43.

Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Deconstructing complex search tasks: a bayesian nonparametric approach for extracting sub-tasks. In *Proc. of NAACL HLT '16*. pages 599–605.

Rishabh Mehrotra and Emine Yilmaz. 2017. Extracting hierarchies of search tasks and subtasks via a bayesian nonparametric approach. In *Proc. of SIGIR'17*. pages 285–294.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS '13*. pages 3111–3119.

Manisha Verma and Emine Yilmaz. 2014. Entity oriented task extraction from query logs. In *Proc. of CIKM'14*. pages 1975–1978.

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W. White, and Wei Chu. 2013. Learning to extract cross-session search tasks. In *Proc. of WWW '13*. pages 1353–1364.

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *Proc. of SIGIR'17*. pages 505–514.

Guoqing Zheng and Jamie Callan. 2015. Learning to reweight terms with distributed representations. In *Proc. of SIGIR '15*. pages 575–584.