A Multilingual Approach for Unsupervised Search Task Identification

Luis Eduardo Lugo Martinez, Jose G. Moreno, Gilles Hubert

A Multilingual Approach for Unsupervised Search Task Identification

Luis Lugo
IRIT UMR 5505 CNRS, U. de Toulouse
Toulouse, France
luis.lugo@irit.fr

Jose G. Moreno
IRIT UMR 5505 CNRS, U. de Toulouse
Toulouse, France
jose.moreno@irit.fr

Gilles Hubert
IRIT UMR 5505 CNRS, U. de Toulouse
Toulouse, France
gilles.hubert@irit.fr

ABSTRACT

Users convert their information needs to search queries, which are then run on available search engines. Query logs registered by search engines enable the automatic identification of the search tasks that users perform to fulfill their information needs. Search engine logs contain queries in multiple languages, but most existing methods for search task identification are not multilingual. Some methods rely on search context training of custom embeddings or external indexed collections that support a single language, making it challenging to support the multiple languages of queries run in search engines. Other methods depend on supervised components and user identifiers to model search tasks. The supervised components require labeled collections, which are difficult and costly to get in multiple languages. Also, the need for user identifiers renders these methods unfeasible in user agnostic scenarios. Hence, we propose an unsupervised multilingual approach for search task identification. The proposed approach is user agnostic, enabling its use in both user-independent and personalized scenarios. Furthermore, the multilingual query representation enables us to address the existing trade-off when mapping new queries to the identified search tasks.

KEYWORDS

Search task; Graph clustering; Multilingual semantic space

ACM Reference Format:
Luis Lugo, Jose G. Moreno, and Gilles Hubert. 2020. A Multilingual Approach for Unsupervised Search Task Identification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401258

1 INTRODUCTION

To support users during the different steps of the information seeking process, it is crucial to correctly group queries in search logs according to their search tasks. Mining query logs from search engines enable the automatic modeling of search tasks. Precise modeling of search tasks is required for user supporting applications like query term prediction, query recommendations, user modeling based on tasks, personalization in e-commerce, and results ranking [8, 11, 17], which enhance search engine support for helping users to fulfill their information needs.

Most unsupervised search task identification methods rely on custom training of word embeddings using search collections or external indexed collections to provide semantic similarities for search queries [10, 11, 15]. Unfortunately, these methods cannot support user queries in multiple languages. Other supervised approaches for search task identification require collections of manually labeled data to train the models [4, 18], which are challenging to create because of the cost of manual labeling [18] and the long-tail nature of search queries [20].

We propose a multilingual method for unsupervised search task identification. The proposed approach combines graph clustering methods [10, 15] with recent general language models [1, 19] for obtaining query representations in a multilingual semantic vector space. The proposed search task identification approach is independent of user identifiers, enabling the modeling of search tasks in user agnostic or personalized applications. We also address the existing trade-off between accuracy and query time [17] that arises when mapping new incoming queries to identified search tasks.

2 RELATED WORK

Search logs provide fine-grained details at the query level, enabling the characterization and classification of individual queries according to the information need they are related to [8]. Hence, it is possible to cluster related queries in the search log to model the tasks that users perform on search engines to fulfill their information needs.

The QC-WCC approach is a clustering method based on graphs [10], where nodes correspond to queries and edges are weighted according to the lexical similarities between the queries. QC-HTC [10] is a computationally simpler alternative, although less accurate. It exploits the sequential nature of query logs to decrease the computational complexity of the QC-WCC method. Heuristics based methods [5, 7] also exploit the sequential nature of query logs, establishing a cascade of rules to group queries into search tasks. Another graph based method, QRY-VEC [15], uses custom trained word embeddings to compute the similarity between queries, outperforming methods that rely on lexical similarities [10]. In a similar way to QC-HTC, Bestlink SVM [18] leverages the chronological structure of the log by establishing links between the analyzed query and queries in the past. The chronological structure of the log is also exploited in CA-LSTM [4], a recurrent neural network (RNN) that determines if adjacent queries issued by the same user

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401258

are part of the same task or not. Once the pairs of adjacent queries are segmented into sessions representing the same task, the graph based QC-HTC [10] method performs the task identification.

Nevertheless, Bestlink SVM [18] and CA-LSTM [4] require a supervised component to perform task identification. CA-LSTM also [4] employs user identifiers to determine the adjacent queries needed to provide context to the recurrent architecture. However, user agnostic scenarios do not have user identifiers available. Also, the context required from adjacent queries could not be available, especially when dealing with simple search tasks like fact-finding, which tend to have a single query [8]. BRTs [11] depend on user identifiers and time sessions to compute the user/time affinity. Similarly, some graph clustering methods [10] use time and user sessions to group the queries before the clustering. Heuristics task identification methods [5, 7] rely on user information and timestamps to identify the search tasks, along with several manually set thresholds. Furthermore, relying on time sessions to identify tasks [5, 7, 11] could be misleading. According to multiple analyses of search query logs [10, 11, 15], users tend to multitask during single time sessions and some complex tasks extend during multiple sessions.

Likewise, both QRY-VEC [15] and BRTs [11] use a custom training word embedding model. The custom training performed using the tempo-lexical context [15] can avoid topic shifting [14], but unfortunately, there are not enough labeled collections to train multilingual word embeddings using such context. Also, the use of an external index [10, 15] requires time-consuming index access at the retrieval model [7] and similar to the issue present in custom training word embeddings, there are not enough corpus for a multilingual clustering of queries. Furthermore, several methods use cosine similarity between query vectors to prune edges in the clustering graph [15] or compute the query affinities [7, 11]. However, the angular similarity has a better performance than the cosine similarity in semantic textual similarity (STS) between sentence pairs [1].

3 TASK IDENTIFICATION APPROACH

The proposed unsupervised search task identification approach uses a multilingual query representation and a graph based clustering method to group queries related to the same search task. In contrast with previous methods [4, 7, 10, 11, 15, 18], it supports queries in multiple languages. Also, it does not utilize user identifiers [4, 7, 11] and has no supervised components [4, 18]. In this section, we cover the multilingual query representation and explain the graph based clustering method.

3.1 Multilingual query representation

Pretrained word embeddings have been released in several languages [6, 12]. However, using pre-trained word vectors can generate topic shifting [14, 20] and requires an additional phase to detect the language of the query to correctly select the pre-trained model to compute the multilingual query vector. Instead, Multilingual Universal Sentence Encoder (MUSE) [19] provides universal representations for sentence embeddings suited to information retrieval tasks [20].

MUSE has models based on transformers [16] or convolutional neural networks (CNNs). We use the transformer-based model. It is more computationally expensive than the CNN based model; however, it is more accurate on several tasks, including sentence retrieval, bitext retrieval, and retrieval question answering. The transformer-based model relies on the encoder part of the transformer architecture. It takes into account context-aware embeddings and averages together the results from the encoder to produce one vector per sentence. Training is based on question-answer pairs, translation pairs, and the Stanford Natural Language Inference (SNLI) corpus. Furthermore, MUSE utilizes SentencePiece, a language-independent subword tokenizer, to process the input query text. SentencePiece covers above 99% of possible tokens in all languages. Likewise, MUSE supports queries in sixteen languages: Arabic (ar), Chinese PRC (zh), Chinese Taiwan (zh-tw), Dutch(nl), English(en), German (de), French (fr), Italian (it), Portuguese (pt), Spanish (es), Japanese (ja), Korean (ko), Russian (ru), Polish (pl), Thai (th), and Turkish (tr).

Algorithm 1 MGBC algorithm

Input: Query log Q **Output**: Task labels L

$V \leftarrow \{\}, E \leftarrow \{\}, G(V, E) \leftarrow (V, E)$
for all $q_i \in Q$ **do**
  $v_i \leftarrow multilingual\_vector(q_i)$
  $V \leftarrow V \cup \{v_i\}$
end for
for all $v_i, v_j \in V$ **do**
  $\mathbf{e}_k \leftarrow S_{ang}(v_i, v_j)$
  $E \leftarrow E \cup \{\mathbf{e}_k\}$
end for
for all $\mathbf{e}_k \in E$ **do**
  if $\mathbf{e}_k < \eta$ **then**
    $E \leftarrow E \setminus \{\mathbf{e}_k\}$
  end if
end for
for all $C_i \in G(V, E)$ **do**
  $task_i \leftarrow i$
  for all $v_k \in C_i$ **do**
    $L[v_k] \leftarrow task_i$
  end for
end for

3.2 Graph Based Clustering

Existing clustering methods for search task identification rely on lexical similarities [10] or cosine distances between word embeddings [11, 15]. However, the angular similarity [1] has been used in recent research [3, 19] to better discriminate text representations in natural language processing. In particular, the angular similarity has been found to perform better in STS between sentence pairs than the cosine similarity [1]; thus, we use it to compute the similarity between query pairs. Given two queries q_i, q_j, with multilingual vector representations v_i, v_j, the angular similarity S_{ang} is defined as follows [1]:

$$S_{ang}(v_i, v_j) = -\arccos\left(\frac{v_i v_j}{|v_i|\,|v_j|}\right) \tag{1}$$

The Multilingual Graph Based Clustering (MGBC) relies on the multilingual query representation to encode queries in the search logs (Algorithm 1). Once the queries are converted into vectors in the multilingual semantic space, a weighted undirected graph $G(V, E)$ is created, where query vectors are nodes in the graph and S_{ang} is the weight for the edges connecting the queries. After creating the fully connected graph, the graph is pruned by filtering out edges with $S_{ang} < \eta$, where η is a threshold optimized during the clustering process: $\eta = k/10, 0 < k \leq 10, k \in \mathbb{N}$. The connected components C in the graph after the pruning process represent the search tasks in the query log. Every connected component receives a unique task label $task_i$, which becomes the label for all the nodes pertaining to the connected component C_i [2, 10, 13, 15].

4 EXPERIMENTS AND RESULTS

Following recent work [4], we use the F_β score, with $\beta = 1$ for the balanced metric, and $\beta = 0.6$, which gives more weight to the precision of the search task identification. Formally, $F_\beta = (1 + \beta^2) * p * r / (\beta^2 * p + r)$, where p is precision and r is recall. The Student's paired t-test provides the test for statistical significance [20]. For evaluating the effectiveness of the proposed approach, we consider a user agnostic search task identification. We also address the existing trade-off when mapping queries to identified search tasks.

4.1 User agnostic search task identification

The dataset for evaluating the clustering approach in a user agnostic scenario is the Cross-Session Task Extraction (CSTE) dataset [15]. The CSTE dataset has 1424 entries with 224 labels corresponding to cross-session search tasks, without grouping queries by user information or query timestamps. As a baseline, we use the state-of-the-art QRY-VEC [15] method, an unsupervised task identification method that uses custom trained tempo-lexical embeddings, averaging the embeddings for each word in the query to compute a single vector per query. We also include results from QC-WCC [10]. For a fair comparison, we do not include methods with supervised components [4, 18] or methods depending on user identifiers or query timestamps [7, 11].

Clustering method	α	η	F_1	$F_{0.6}$
QC-WCC	0.8	0.4	0.471	0.428
QRY-VEC word2vec	0.6	0.5	0.473	0.441
QRY-VEC tempo-lexical	0.6	0.7	0.538	0.488
MGBC	0.4	0.3	**0.624**	**0.695**

Table 1: Clustering performance for the CSTE dataset with search task labels. Results are statistically significant against the baseline with $p \leq 0.05$.

To compare to the QRY-VEC method, we use the same index similarity S_{ind} than the baseline [15], which is based on the ClueWeb12B dataset. We adjust the angular similarity S_{ang} in equation 1 by the use of a convex combination of both angular and index similarities. The similarity between queries q_i, q_j with multilingual vectors v_i, v_j becomes [10, 15]:

$$S_{ang}(v_i, v_j) = -\alpha * \arccos\left(\frac{v_i v_j}{|v_i|\,|v_j|}\right) + (1 - \alpha) * (S_{ind}) \quad (2)$$

where α, η are parameters to be optimized during the process of clustering for search tasks. The optimization uses a grid search with $\alpha = k/10, \eta = k/10$, where $0 < k \leq 10, k \in \mathbb{N}$ [15], selecting the model with the best F_1 metric.

Results show that MGBC outperforms the baseline method in search task identification (Table 1). It gets better performance than both lexically based (QC-WCC) and monolingual query vectors based (QRY-VEC) methods for identifying tasks, highlighting the ability of the multilingual semantic vector space to encode queries for the modeling of search tasks.

Using Google Cloud Translation API, we translate the CSTE dataset to all the languages supported by MGBC. Running the search task identification method on automatically translated queries enables the assessment of the method in multilingual task identification (Table 2). For multilingual tests, MGBC uses the angular similarity S_{ang} in equation 1. F_1 metrics varies from 0.429 with the Turkish language to 0.484 with the French language, which are located around the F_1 metric of 0.456 for the English language, the original language of the dataset. These results reflect the quality of the multilingual semantic space to represent the search tasks. Overall, no drop in performance is observed despite the use of automatic translation, suggesting an adequate performance in the sixteen languages for the search task identification approach. Although there exist variations in results for the different languages, they are explained by the expected differences in the automatic translation results.

4.2 Mapping queries to search tasks

The same multilingual semantic space for query representation and the S_{ang} similarity in Equation 1 enable us to address the existing trade-off when mapping new incoming queries to the identified search tasks. Previously analyzed methods for mapping queries face a trade-off between accuracy and execution time. The most accurate method uses an inverted index approach based on a BM25 retrieval model; however, its average time per query is much slower than a Trie data structure implementation, which is the fastest method [17]. To address this trade-off, we utilize the Neighborhood Graph and Tree (NGT) approximate nearest neighbor method [9], along with S_{ang} and multilingual query vectors.

Three datasets have been proposed to evaluate the mapping of new incoming queries: the AOL query log (AOLQL), the TREC query topics (TRECQT), and the WikiHow questions (WIKIHQ) based datasets [17]. For comparison, we use publicly available implementations for the Trie data structure[1] and the BM25 retrieval model[2] with default parameters [17, 19]. Experiments run on a virtual machine instance with 8 CPUs of 3GHz and 60GB of RAM. We compute time per query as the average time for mapping 10^4 queries, while accuracy is measured using a leave-one-out evaluation, independently selecting 100 random queries from the dataset and repeating the evaluation during 50 runs [17].

[1] https://github.com/google/pygtrie
[2] https://github.com/nhirakawa/BM25

ISO 639-1	ar	zh	zh-tw	nl	en	de	fr	it	pt	es	ja	ko	ru	pl	th	tr
η	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8	0.8	0.8
F_1	0.447	0.480	0.482	0.449	0.456	0.450	0.484	0.452	0.458	0.450	0.453	0.451	0.449	0.460	0.444	0.429
$F_{0.6}$	0.395	0.473	0.476	0.431	0.437	0.432	0.547	0.434	0.438	0.432	0.436	0.396	0.429	0.524	0.427	0.401

Table 2: Search task identification results for the supported languages of the MGBC task identification approach. Results are statistically significant between languages with $p \leq 0.05$

Dataset	Method	Accuracy	F_1	$F_{0.6}$	Query time
AOLQL	Trie	0.693	0.543	0.543	0.029ms
	BM25	**0.809**	**0.689**	**0.689**	0.947s
	NGT	0.751	0.608	0.607	0.308ms
TRECQT	Trie	0.650	0.519	0.518	0.030ms
	BM25	0.791	0.688	0.688	2.532s
	NGT	**0.804**	**0.705**	**0.704**	0.299ms
WIKIHQ	Trie	0.471	0.310	0.311	0.032ms
	BM25	0.621	0.453	0.454	6.572m
	NGT	**0.648**	**0.481**	**0.481**	0.368ms

Table 3: Metrics for mapping queries to search tasks. Results are statistically significant against Trie with $p \leq 0.05$.

We test several values of k nearest neighbors for NGT, finding $k = 9$ as the best performing setup. NGT is several times faster than the inverted index based on BM25 (Table 3), keeping average times per query below half a millisecond. The speedup obtained with NGT does not imply a deterioration in the accuracy metrics for TRECQT and WIKIHQ datasets. Also, AOLQL differences are much lower than the Trie data structure drop in metrics. Similarly, NGT is more accurate than the Trie data structure in all the three datasets; nonetheless, the latter continues to be faster in terms of average time per query.

5 CONCLUSION

The MGBC multilingual search task identification approach enables the modeling of search tasks from query logs, supporting queries in sixteen languages. Experiments show that the proposed approach outperforms baseline identification methods. Also, MGBC is user-independent, enabling its use in both user agnostic and personalized search task identification applications. Moreover, the same multilingual semantic space and query similarity of MGBC can be used with NGT nearest neighbor method to address the existing trade-off when mapping new queries to identified search tasks. NGT provides metrics at the same level of the BM25 retrieval model results; however, it is several times faster, keeping query response times below half a millisecond, a crucial aspect for running on the fly applications for supporting search engine users. For future work, we want to extend the number of languages supported by MGBC. Also, we plan to explore additional unsupervised approaches to improve search task clustering performance.

REFERENCES

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 169–174.

[2] Zheng Chen and Heng Ji. 2010. Graph-based clustering for computational linguistics: A survey. In *Proceedings of the 2010 workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 1–9.

[3] Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836* (2018).

[4] Cong Du, Peng Shu, and Yong Li. 2018. CA-LSTM: Search Task Identification with Context Attention based LSTM. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1101–1104.

[5] Daniel Gayo-Avello. 2009. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences* 179, 12 (2009), 1822–1843.

[6] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[7] 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*. 85–92.

[8] Marti Hearst. 2009. *Search user interfaces*. Cambridge University Press, Cambridge, CB2 8BS, UK.

[9] Masajiro Iwasaki and Daisuke Miyazaki. 2018. Optimization of indexing based on k-nearest neighbor graph for proximity search in high-dimensional data. *arXiv preprint arXiv:1810.07355* (2018).

[10] 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*. ACM, 277–286.

[11] Rishabh Mehrotra and Emine Yilmaz. 2017. Extracting hierarchies of search tasks & subtasks via a bayesian nonparametric approach. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 285–294.

[12] 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*. 3111–3119.

[13] Maria CV Nascimento and Andre CPLF De Carvalho. 2011. Spectral methods for graph clustering–a survey. *European Journal of Operational Research* 211, 2 (2011), 221–231.

[14] Navid Rekabsaz, Mihai Lupu, Allan Hanbury, and Hamed Zamani. 2017. Word Embedding Causes Topic Shifting; Exploit Global Context!. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1105–1108.

[15] 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*. 283–292.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[17] Michael Völske, Ehsan Fatehifar, Benno Stein, and Matthias Hagen. 2019. Query-Task Mapping. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 969–972.

[18] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. 2013. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1353–1364.

[19] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual Universal Sentence Encoder for Semantic Retrieval. *arXiv preprint arXiv:1907.04307* (2019).

[20] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic Intent Representation in Web Search. In *The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.