Retrieving Information from Multiple Sources

Anurag Roy
IIEST Shibpur, India

Kripabandhu Ghosh
IIT Kanpur, India

Moumita Basu
UEM Kolkata; IIEST Shibpur, India

Parth Gupta
Amazon, India

Saptarshi Ghosh
IIT Kharagpur; IIEST Shibpur, India

ABSTRACT

The Web has several information sources on which an ongoing event is discussed. To get a complete picture of the event, it is important to retrieve information from multiple sources. We propose a novel neural network based model which integrates the embeddings from multiple sources, and thus retrieves information from them jointly, as opposed to combining multiple retrieval results. The importance of the proposed model is that no document-aligned comparable data is needed. Experiments on posts related to a particular event from three different sources - Facebook, Twitter and WhatsApp - exhibit the efficacy of the proposed model.

CCS CONCEPTS

• **Information systems → Information retrieval**;

KEYWORDS

Multi-view retrieval; word embedding; deep learning

ACM Reference Format:
Anurag Roy, Kripabandhu Ghosh, Moumita Basu, Parth Gupta, and Saptarshi Ghosh. 2018. Retrieving Information from Multiple Sources. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186920

1 INTRODUCTION AND MOTIVATION

The Web contains several information sources, including social media (e.g., Facebook, Twitter, WhatsApp), news sites, personal blogs, and so on. An ongoing event is discussed on all these channels; however, there are usually qualitative differences in the information obtained from different sources. As a result, to get a complete picture of an ongoing topic or event, it is necessary to retrieve information from multiple information sources. Further, given the real-time nature of online sources, the retrieval model for the multiple sources needs to be learned quickly.

The existing methodologies for retrieval from multiple sources depend upon a central issue – *availability of document-aligned comparable data.* If such data is available, then common topic models or word embeddings can be learned; e.g., this approach was used by Vulic *et al.* [5] for bilingual retrieval. However, preparing document-aligned comparable data requires a lot of human involvement and time. Hence such approaches are not suitable when information about an ongoing event has to be retrieved quickly. If document-aligned comparable training data is *not* available, then the two most intuitive approaches are –

(Approach 1) Learning a single topic model or word embedding across all sources taken together: This approach ignores the fact that different information sources have their own inherent characteristics which vary from once source to another.

(Approach 2) Retrieving separately from different sources and then combining: Retrieval models are used to retrieve results separately for each source, and then the results are combined, e.g., using data fusion techniques [1].

Proposed approach: In this work, we investigate the modeling of text across different information sources (or views) in a unified framework. We propose a novel deep learning-based multi-view retrieval model which attempts to learn document embeddings on a common space, where differences among the various data sources would not exist. Importantly, the proposed model does not require document-aligned training data.

We analyse the performance of our proposed model over posts related to a common event, from three distinct online sources – Facebook, Twitter and WhatsApp. We demonstrate that the proposed model enables significantly better retrieval compared to the two approaches described above (in absence of document-aligned comparable data).

2 PROPOSED APPROACH

The proposed multi-view model is structurally a feed-forward neural network that maps an input document vector to a multi-view space that would have the contextual information of multiple domains. We assume that for a small set of queries, the relevant documents from the different sources (views) are known. Note that this training data is much simpler to build, as compared to document-level comparable data. The proposed model, trained over this training data, allows efficient retrieval for many other queries.

Let $e_i \in \mathbb{R}^n$ be the embedding of a document from source i relevant to a query q. Let e_j be the embedding of another document, which is also relevant to query q, from any other source j. We wish to learn a generic space $e_o \in \mathbb{R}^n$ such that e_o normalises the differences between the information sources and helps better retrieval. The idea is to obtain e_o such that it exhibits the characteristics of both e_i and e_j. To this end, we use a feed-forward neural network with a single hidden layer that takes e_i and e_j as input and gives e_o as output. The transformation of e_i into e_o can be explained as: $h = f(W_1 * e_i + b_1)$ and $e_o = f(W_2 * h + b_2)$ where W_i and b_i represent the i^{th} layer weights and biases respectively; h represents the hidden layer and f is the non-linear activation function. The model is trained to minimize the objective function $J(\theta) = \|e_o - e_i \circ e_j\|$

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186920

Query	#Twitter	#Facebook	#WhatsApp
money donation	1076	39	1072
available hospitals	50	15	280
blood donation	130	0	138
medicine medical equipment need	41	49	101

Table 1: Examples of some queries and the number of relevant documents from each dataset.

where $\circ$ denotes the Hadamard Product, i.e., element-wise product between two vectors, and $||X||$ denotes the $L2$ norm.

Once the multi-view model is trained, we generate a generalized document embedding for each document. For retrieval, the query will also be passed through this multi-view network and matching will take place in the same space.

3 EXPERIMENTS AND RESULTS

This section shows the efficacy of our proposed multi-view model, by comparing its performance with those of baseline models.

Dataset: For the present work, we considered a specific event – the Nepal earthquake in April 2015 – and the messages posted after the event on three social media: Twitter, Facebook and WhatsApp. The Twitter posts were collected using the Twitter Search API, and the Facebook posts were collected using the Radian6 tool (https://www.marketingcloud.com/products/social-media-marketing/radian6), both using the search keywords 'nepal' and 'quake', and for the duration of two weeks following the earthquake. We also collected WhatsApp chat-logs of members of a medical NGO (Doctors For You) who were engaged in relief operations after the earthquake. After de-duplication, we obtained (i) 50,018 tweets, (ii) 85,483 Facebook posts, and (iii) 3,438 WhatsApp messages. All experiments reported here were carried out on these three datasets.

Queries and gold standard relevance judgements: Based on the feedback of NGOs, we identified a set of *30 queries* or information needs specific to the event. Next, we employed human annotators to develop the gold standard relevance judgements for the queries. Table 1 shows some of the queries, and the number of relevant documents found for each query from the three data-sets.

Train-test setup: We performed a 3-fold cross-validation on the set of 30 queries. Query numbers 1-10, 11-20 and 21-30 were used as training sets in turn for the three folds, while the rest of the queries served as the test sets.

Baseline retrieval models: We compare the retrieval of our proposed multi-view model with that of the following baselines:

(1) BM25 [3]: We combined all documents from all three sources into a single corpus, and applied Okapi-BM25 ranking (with $k = 0.5$).

(2) Language model with Dirichlet smoothing (LM) [6]: For this baseline also, we rank the documents from all the three different sources taken together (parameter μ taken as 2000).

(3) Single view embeddings: For this baseline, a single word embedding is learned over all the documents (in the training sets) from all the sources. Subsequently, retrieval was done on the test set of queries using the learned embedding.

(4) Data fusion: Three different word embeddings were learned over the training sets for the three sources, and were used for retrieval on the test sets (from the same source). The results were fused using a standard data fusion algorithm *CombSUM* [1].

Embeddings and retrieval setup: All experiments utilize word embeddings learned using Word2vec [2] with the following parameters – skip-gram model, vector size: 400, context size: 3, learning rate: 0.01, negative sample size for negative subsampling = 5. For both a query and a document, we construct a vector by averaging the vectors of the constituent words generated by the underlying embedding model. The posts in the corresponding datasets are arranged in the decreasing order of the *cosine similarity* score of each document-vector with the associated query-vector.

Algorithm	MAP	Precision@20	Recall@100	Bpref
Single view	0.0087	0.0450	0.0044	0.0469
Data Fusion	0.0169	0.0683	0.0206	0.0835
BM25	0.0026	0.0367	0.0059	0.0128
LM	0.0051	0.0350	0.0193	0.0189
Proposed	$\mathbf{0.0280}^{SDBL}$	$\mathbf{0.1367}^{SDBL}$	$\mathbf{0.0287}^{SDBL}$	$\mathbf{0.0942}^{SBL}$

Table 2: Comparison of retrieval performance, averaged over 3-fold cross-validation. Bold font shows the best value, which the proposed method always achieves. Super-scripts S, D, B and L indicate that the proposed method is statistically significantly better at 95% confidence interval ($p < 0.05$) than Single view, Data fusion, BM25 and LM respectively.

Evaluation measures: We report the retrieval performance of all models in terms of Precision@20, Recall@100, Mean Average Precision (MAP) and Bpref. For all models, the retrieval is performed and evaluated once for each test query (via 3-fold cross validation), and the average across all queries is reported.

Retrieval results: Table 2 reports the retrieval results of the proposed methodology and the baselines, averaged over all the queries in the test set. We see that retrieval using the proposed methodology numerically outperforms all the baseline methodologies. The proposed approach gives statistically significant performance improvements computed at 95% confidence interval ($p < 0.05$) by Wilcoxon signed-rank test [4] over Single view, BM25 and LM in all the measures and over Data Fusion in all the measures except Bpref. It should also be noted that Data Fusion performs better than single view approach, which highlights the importance of handling different sources separately.

4 CONCLUSION AND FUTURE DIRECTIONS

We proposed a novel neural network architecture for retrieval from multiple sources. The proposed architecture does not need expensive document-aligned training data, which makes the proposed model attractive for quick retrieval across multiple online sources.

The low performance scores achieved by all the methods indicate that the problem is challenging and necessitates better methods. We also look to use the proposed architecture in retrieval across data sources varied in length, languages, scripts and modalities.

REFERENCES

[1] Edward A. Fox and Joseph A. Shaw. 1993. Combination of Multiple Searches. In *Proceedings of TREC 1993*. http://trec.nist.gov/pubs/trec2/papers/txt/23.txt.

[2] T. Mikolov, W.T. Yih, and G. Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *NAACL HLT 2013*.

[3] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[4] S. Siegel. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

[5] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *Proc. ACM SIGIR*. 363–372.

[6] Chengxiang Zhai and John Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proc. ACM SIGIR*.