# Neural Language Model Based Attentive Term Dependence Model for Verbose Query (Student Abstract)

**Dipannita Podder, Jiaul H. Paik, Pabitra Mitra**

Indian Institute of Technology Kharagpur, India - 721302
dipannita@iitkgp.ac.in, jiaul@cet.iitkgp.ac.in, pabitra@cse.iitkgp.ac.in

### Abstract

The query-document term matching plays an important role in information retrieval. However, the retrieval performance degrades when the documents get matched with the extraneous terms of the query which frequently arises in verbose queries. To address this problem, we generate the dense vector of the entire query and individual query terms using the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model and subsequently analyze their relation to focus on the central terms. We then propose a context-aware attentive extension of unsupervised Markov Random Field-based sequential term dependence model that explicitly pays more attention to those contextually central terms. The proposed model utilizes the strengths of the pre-trained large language model for estimating the attention weight of terms and rank the documents in a single pass without any supervision.

## Introduction

Understanding the user's intent and providing them with relevant documents becomes challenging when the query contains many extraneous terms. Since the retrieval performance depends on the term matching score of query and document, the search engines need to focus on the central terms and reduce erroneous matching. For decades, several supervised and unsupervised query term weighting approaches (Paik and Oard 2014; Zheng and Callan 2015; Karisani, Rahgozar, and Oroumchian 2016) are proposed to address this problem, but these methods lag behind in capturing the intrinsic meaning of the query. Although several semantic neural retrieval models are proposed in recent times, these models need well-tuned search space for efficient performance (Hammache and Boughanem 2021), especially when the datasets are large, and in these cases, the traditional term-based IR models help.

Thus, we propose an extension of the Markov Random Field (MRF) based sequential term dependence model (SD) (Metzler and Croft 2005), that ranks the documents by giving explicit attention to the central terms. To estimate the attention for each term, we analyze (i) how different query concepts are correlated with the query's intrinsic context, and (ii) how the different query terms are interrelated with each other. Moreover, the dense contextual vectors of the entire query and its individual terms are generated by utilizing the pre-trained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018).



## BERT-based Attentive Term Dependence Model

The MRF-based sequential term dependence model (SD) captures the dependency between the neighboring query terms. It models the joint distribution $P_L(D,Q)$ (parameterized by $L$) using a non-negative potential function over cliques of a graph $G$, in which the nodes correspond to query terms $Q = t_1 t_2 \dots t_n$ and a document $D$. $P_L(D,Q)$ is utilized to estimate $P_L(D|Q)$ as they are rank equivalent. The ranking function of the proposed Bert-based Attentive Term Dependence Model (B-ASD) is shown in Eq. 1, where $g_t(c)$, $g_o(c)$, $g_u(c)$ are feature functions over the three different types of cliques of the query-document graph of SD, i.e., (i) t: a query term (unigram) and the document, (ii) o: the ordered bigrams and the document, and (iii) u: the unordered bigrams. Likewise, $l_t, l_o, l_u$ are the weights given to that feature functions, and $\alpha_{q_t}, \alpha_{q_o}, \alpha_{q_u}$ are the estimated context-aware attention weight of the query terms (unigram, ordered bigram, unordered bigram) that are associated with the cliques. In this work we use Dirichlet language model as the feature functions over the cliques and set $l_t$, $l_o$, $l_u$ to 0.85, 0.1, 0.05 as in (Metzler and Croft 2005).

$$\mathrm{P}_{L,A}(D|Q) = \sum_{c \in t} \alpha_{q_t} l_t g_t(c) + \sum_{c \in o} \alpha_{q_o} l_o g_o(c) + \sum_{c \in o \cup u} \alpha_{q_u} l_u g_u(c) \quad (1)$$

To estimate the attention weight of the terms, we assume, that (i) the central query terms are highly correlated to discourse of the entire query than the other terms, and the contextual correlation of a term $t_i$ and query $Q$, i.e., $C(t_i, Q) = e^{<V_Q, V_{t_i}>d}$ reflects the importance of a term, where $V_Q$ and $V_{t_i}$ are the contextual dense vector of $Q$ and $t_i$, and $d$ is a smoothing parameter. and, (ii) there is contextual dependency between the query terms, thus, the important query terms are strongly interrelated with each other. Moreover, the importance of a term increases if it is correlated to the query more than other important terms. This is formulated as Eq. 2, where $I$ is the initial context aware importance vector and $S$ is the association matrix which is represented as $S(t_i, t_j) = C(t_i, Q)C(t_j, Q), t_i \neq t_j$. The initial context aware importance of every term is assigned to 1. We, then define the importance recursively using Eq. 2.

$$I(t_i) = \sum_{j=1, i \neq j}^{|Q|} S(t_i, t_j).I(t_j), t_i \neq t_j \quad (2)$$

We generate the contextual word embedding of the query using the pre-trained BERT model[1] and use the individual word embeddings as the dense vector of unigrams and embedding of "[CLS]" token as the dense vector of the entire query. Likewise, we average the embeddings of two words involved in unordered bigram to get the dense vector of unordered bigrams. Furthermore, we consider ordered bigrams as a single text piece and generate its embeddings using BERT, and use the "[CLS]" token as the final dense vector for the bigram.

Finally, we incorporate a regulated Inverse Document Frequency ($idf_r(t_i)$) factor into context-aware importance $I(t_i)$ to introduce the additional contribution of a term based on its specificity in the collection into the attention weight. The regulated Inverse Document Frequency is computed as $idf_r(t_i) = idf(t_i)(c + idf(t_i))$. However $idf(t_i) = df(t_i)N$, where $df(t_i)$ is the document frequency of term $t_i$ and $N$ is the total number of documents of the dataset. Hence, the final attentive weights are represented as: $\alpha_{t_i} = I(t_i) \cdot idf_r(t_i)$. We set the two hyper-parameters $d$ and $c$ as 1000 and 10 empirically.

## Experimental Results

We experiment on the two large web datasets, i.e., GOV2 and Clueweb09B, and select the description portions of the topics as queries. The experiments are performed using Indri. We remove the stopwords from all the documents and queries and use porter stemmer to map the words in documents and queries to their root words. The retrieval effectiveness of the proposed BERT-based Attentive Sequential Dependency Model (B-ASD) for the description query topics is measured using recall and utility-based metrics (such as Normalized Discounted Cumulative Gain (NDCG) and Expected Reciprocal Rank (ERR) (Paik et al. 2021), which is shown in Table 1. This table also reflects the comparison of the proposed model with the basic sequential dependence model (Metzler and Croft 2005) and the two query term weighting methods, i.e., (i) term re-weighting with document similarity (DSW) (Karisani, Rahgozar, and Oroumchian 2016) and (ii) term re-weighting with distributed representations (deepTR) (Zheng and Callan 2015).

## Conclusion

In this work, an extension of the Markov Random Field-based sequential term dependence model is proposed. It estimates the attention weights of query terms of the dependence model by leveraging contextual dense vectors, that are generated using the pre-trained contextual BERT model. The correlation of individual query terms with the entire query is analyzed using those dense vectors, and then the attention weight of each query term is estimated by looking at the interrelation of those query terms. The experimental results on the large web collections show the potential of the proposed model.

[1] https://huggingface.co/google/bert_uncased_L-12_H-768_A-12

| **GOV2** #Docs: 2,52,05,179 (topics: 701-850) | | | |
|---|---|---|---|
| | NDCG@10 | NDCG@20 | ERR@10 |
| SD$^{(unsupervised)}$ | 0.455 | 0.452 | 0.436 |
| DSW$^{(unsupervised)}$ | 0.442 | 0.441 | 0.419 |
| DEEPTR-SD$^{(supervised)}$ | 0.470 | 0.466 | 0.433 |
| **B-ASD**$^{(unsupervised)}$*(proposed)* | **0.478** | **0.472** | **0.448** |
| **ClueWeb09B** #Docs: 5,01,84,841 (topics:1-200) | | | |
| | NDCG@10 | NDCG@20 | ERR@10 |
| SD$^{(unsupervised)}$ | 0.189 | 0.206 | 0.174 |
| DSW$^{(unsupervised)}$ | 0.200 | 0.201 | 0.181 |
| DEEPTR-SD$^{(supervised)}$ | 0.204 | 0.212 | 0.186 |
| **B-ASD**$^{(unsupervised)}$*(proposed)* | **0.211** | **0.223** | **0.193** |

Table 1: Comparison of Attentive Term Dependence Model with the baselines. The superscripts denote the nature of the methods (supervised or unsupervised).

## References


Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hammache, A.; and Boughanem, M. 2021. Term position-based language model for information retrieval. *Journal of the Association for Information Science and Technology*, 72(5): 627–642.

Karisani, P.; Rahgozar, M.; and Oroumchian, F. 2016. A query term re-weighting approach using document similarity. *Information Processing & Management*, 52(3): 478–489.

Metzler, D.; and Croft, W. B. 2005. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, 472–479. ISBN 1595930345.

Paik, J. H.; Agrawal, Y.; Rishi, S.; and Shah, V. 2021. Truncated Models for Probabilistic Weighted Retrieval. *ACM Transactions on Information Systems (TOIS)*, 40(3): 1–24.

Paik, J. H.; and Oard, D. W. 2014. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, 131–140.

Zheng, G.; and Callan, J. 2015. Learning to reweight terms with distributed representations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 575–584.