Neural Document Expansion with User Feedback

Yue Yin*
Beijing Normal University
bnuyinyue@outlook.com

Chenyan Xiong
Microsoft Research AI
chenyan.xiong@microsoft.com

Cheng Luo
Tsinghua University
chengluo@tsinghua.edu.cn

Zhiyuan Liu
Tsinghua University
liuzy@tsinghua.edu.cn

ABSTRACT

This paper presents a neural document expansion approach (NeuDEF) that enriches document representations for neural ranking models. NeuDEF harvests expansion terms from queries which lead to clicks on the document and weights these expansion terms with learned attention. It is plugged into a standard neural ranker and learned end-to-end. Experiments on a commercial search log demonstrate that NeuDEF significantly improves the accuracy of state-of-the-art neural rankers and expansion methods on queries with different frequencies. Further studies show the contribution of click queries and learned expansion weights, as well as the influence of document popularity of NeuDEF's effectiveness.

KEYWORDS

Neural Information Retrieval; Document Expansion; User Feedback.

ACM Reference Format:
Yue Yin, Chenyan Xiong, Cheng Luo, and Zhiyuan Liu. 2019. Neural Document Expansion with User Feedback. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, San Francisco, USA, 4 pages. https://doi.org/10.1145/3341981.3344213

1 INTRODUCTION

Neural information retrieval (Neu-IR) methods have shown promising results in various search scenarios. These neural ranking models leverage distributed representations (embeddings) to conduct soft matches between query-documents, and at the same time, leverage large model capacity to revive "classic" IR intuitions, such as translation model [8], phrase matches [1], and multi-field evidence combination [9]. This paper revisits the document expansion technique and develops NeuDEF, a Neural Document Expansion method that explicitly expands documents using user Feedback signals.

NeuDEF first harvests candidate expansion terms for a document from queries lead to user clicks on it (click queries). Then its *attention mechanism* weights the expansion terms based on both self-attention and the matches between the document and click queries. The weighted expansion terms form an additional document representation and can be integrated into various neural ranking models. During learning, the attention mechanism on expansion terms and the neural ranking model are end-to-end trained using document ranking labels; the integrated system learns how to expand and rank documents jointly.

In our experiments on two search log samples from a commercial search engine, NeuDEF significantly improves the ranking accuracy of its base ranker: K-NRM [8] and outperforms previous state-of-the-art neural ranking methods and document expansion techniques by large margins. Our additional studies show that NeuDEF's attention mechanism assigns higher weights to novel expansion terms and NeuDEF generalizes user feedback signals to unseen queries.

2 NEURAL DOCUMENT EXPANSION

This section presents K-NRM [8], the base ranker, NeuDEF, our expansion model, and the joint learning of the two.

2.1 Base Ranker Recap

K-NRM is an interaction-based neural ranking model that uses density estimation kernels to soft match term pairs [8]:

$$\text{K-NRM}(q,d;w) = w^T \sum_{t_j \in q} \log \Phi(t_j, d), \tag{1}$$

$$\Phi(t_j, d) = \{\phi_1(t_j,d), \ldots \phi_k(t_j,d) \ldots, \phi_K(t_j,d)\}, \tag{2}$$

$$\phi_k(t_j, d) = \sum_{t_d \in d} \exp(-\frac{(\cos(\vec{t}_j, \vec{t}_d) - \mu_k)^2}{2\sigma_k^2}). \tag{3}$$

ϕ_k is a Gaussian (RBF) kernel. It "soft counts" the number of document terms t_d that match the query term t_j near its kernel region (μ_k, σ_k). The match is calculated by the cosine similarity, $\cos(\vec{t}_j, \vec{t}_d)$, of their word embeddings, which are learned end-to-end for the whole vocabulary. w is the kernel weight to be learned.

2.2 Expansion Term Selection and Weighting

NeuDEF leverages user feedback in search logs to find expansion terms. It first considers the click queries C_d of the document d:

$$C_d = \{c_i | click(c_i, d) = \text{True}\}. \tag{4}$$

$click(c, d)$ is True if there is any click on d from the query c.

Terms from clicked queries form the candidate expansion set T_d:

$$T_d = \{t_j | \exists c_i : t_j \in c_i, c_i \in C_d\}. \tag{5}$$

A clicked query is likely to be related to the document as a user has used the query to search and click on the document.

[*] This work was done when the first author was visiting Tsinghua University.

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344213

Given the expansion terms T_d and clicked queries C_d of document d, NeuDEF calculates the weights $A(T_d, d) = \{a(t_j, d) | t_j \in T_d\}$ of the expansion terms T_d by an attention mechanism.

The attention first uses Multi-Head Self Attention [7] to weight document and clicked queries independently. To capture the cross-queries information and generate better word-level attention weights, we concatenate all the clicked queries for a specific document and feed into the transformer (Self-ATT). Then the attention matches the clicked queries to the document:

$$m(c_i, d) = \text{K-NRM}(\text{Self-ATT}(c_i), \text{Self-ATT}(d); w_c). \qquad (6)$$

It uses the same architecture as the base ranker but different parameters. Then it calculates the attention weight of term t_j by summing match scores from the clicked queries it appears in:

$$a(t_j, d) = \sum_{c_i, t_j \in c_i} m(c_i, d). \qquad (7)$$

The more related clicked queries the term t_j appears in, the more expansion weights, $a(t_j, d)$, it receives.

NeuDEF provides the expansion terms T_d, which introduce user feedback, and their weights $A(T_d, d)$ learned by self-attention and document-query matches.

2.3 Joint Learning with Neural Rankers

NeuDEF is integrated to the base ranker f by providing an additional expansion field, which includes expansion terms T_d and attention weights $A(T_d, d)$ and is linearly combined with the base ranker (α and β are the combine weights):

$$f_{\text{NeuDEF}}(q, d) = \alpha f(q, d) + \beta f'(q, de), \qquad (8)$$

$$f(q, d) = \text{K-NRM}(q, d; w_r). \qquad (9)$$

We use a modified K-NRM on the expansion field:

$$f'(q, de) = w_{de}^T \sum_{t_q \in q} \log \Phi'(t_q, de), \qquad (10)$$

$$\phi'_k(t_q, de) = \sum_{t_j \in T_d} a(t_j, d) \exp(-\frac{(\cos(\vec{t_q}, P\vec{t_j}) - \mu_k)^2}{2\sigma_k^2}). \qquad (11)$$

The modified version weights expansion term t_j by the attention $a(t_j, d)$, and adds a projection P on the expansion word embeddings to distinguish them from the original words. The three K-NRM models used in $f(q, d)$, $f'(q, de)$, and $m(c_i, d)$ (Eq. 6) share the same word embeddings and kernel hyper-parameters.

NeuDEF is then trained with the neural ranker using standard pairwise hinge loss:

$$\sum_{d^+, d^- \in D^{+,-}} \max(0, 1 - f_{\text{NeuDEF}}(q, d^+) + f_{\text{NeuDEF}}(q, d^-)). \qquad (12)$$

d^+, d^- are the relevant and irrelevant document pairs of the query. Instead of conducting the expansion and ranking separately, NeuDEF learns the document expansion and document ranking jointly from ranking labels using back-propagation.

3 EXPERIMENTAL METHODOLOGY

Dataset. Sogou, a commercial search engine based in China, released a sample of search log with queries, documents, and user clicks to various academic partners. Our experiments use two training datasets sampled from Sogou log: Sogou-KNRM, the one used by K-NRM [8], for a fair comparison, and Sogou-QCL, the public release of Sogou search log [10]. We follow the same setting with prior research [8, 10] and refer to their papers for more details in the datasets due to space limitation [8, 10].

Our evaluations on Sogou-KNRM dataset omit the Testing-SAME setting which is prune to overfitting [1] and add torso (50-1000 appearances) and tail (less than 50) queries. We use four evaluation scenarios for Sogou-KNRM dataset: (1) **Testing-Raw Head**, **Torso**, and **Tail**: User clicks as relevance labels and evaluate on head, torso, and tail queries; (2) **Testing-DIFF**: Click model (TACM [3]) inferred relevance labels and evaluate on head queries. For Sogou-QCL dataset, we use TACM inferred relevance labels to train and test our model on head queries, following the exact same setting for the Table 5 in Sogou-QCL original paper [10].

To study the effectiveness of document expansion with body field, The body texts are from our crawled HTMLs and parsed by Boilerpipe. They are combined linearly with the titles following standard multi-field ranking setup.

Expansion Candidates. All expansion approaches harvest the expansion terms *solely using the queries and clicks in the training split.* As there is no overlap in the training and testing queries, the expansion approaches use no information from the testing data.

Evaluation Metrics. Testing-DIFF is evaluated by NDCG@{1, 10} and Testing-Raw is evaluated by MRR [8]. Statistical significance is tested by permutation test with $p < 0.05$.

Model performances of the ranking model (f_2) w.r.t the baseline model (f_1) at *per document level* are compared by ΔRR:

$$\Delta\text{RR}_{f_1 \rightarrow f_2}(d) = \sum_q y(q, d)\{RR_{f_2}(q, d) - RR_{f_1}(q, d)\}. \qquad (13)$$

$y(q, d) = \{+1, -1\}$ is the relevance label. RR(q, d) is the reciprocal rank of d under q. Better ranking models have positive ΔRRs.

Baselines. Using the same experimental setup with prior research [8, 10] makes our method directly comparable with their baselines: CDSSM [4], DRMM [2], K-NRM [8] and Conv-KNRM [1]. We use their shared implementations to obtain baseline results on torso and tail queries. We also implemented NRM-F [9], the fielded version of CDSSM, using the same fields as in NeuDEF. All neural ranking baselines leverage user feedback following Eq. 8.

We implemented and compared with many document expansion baselines. DELM [6] is a traditional document expansion language model via document neighbors. We selected the top 5 words as document expansion fields according to their frequency in all the neighbor documents. ExpaNet [5] is a neural text expansion model with memory network generated via document neighbors. We combined the expanded document features from ExpaNet with soft match features in original K-NRM's dense layer. We treated all the documents under a specific query as neighborhoods.

We also compared with expanding using other meta-data: (1) DocFreq: the number of times the document appears in search log; (2) CQCount: the number of queries lead to clicks in the document;

Two simpler versions of NeuDEF are compared too: (1) NeuDEF-TF, which weights expansion terms only by their frequency in the clicked queries; (2) NeuDEF-NoTrans, which does not use transformer's self attention.

Table 1: Ranking Accuracy on Sogou-KNRM dataset. Relative performances over K-NRM are in percentages. †, ‡, § and ¶ indicate statistically significant improvements over K-NRM†, NRM-F‡, NeuDEF-TF§, and NeuDEF-NoTrans¶.

	Testing-RAW, MRR						Testing-DIFF			
Model	**Head**		**Torso**		**Tail**		**NDCG@1**		**NDCG@10**	
DRMM [2]	0.2335	-30.1%	0.3102	-16.0%	0.2951	-6.4%	0.2126	-27.5%	0.3592	-14.3%
CDSSM [4]	0.2501	-25.1%	0.3184	-13.7%	0.2928	-7.1%	0.2017	-31.2%	0.3500	-16.5%
K-NRM [8]	0.3339	-	0.3691	-	0.3152	-	0.2931	-	0.4190	-
Conv-KNRM [1]	0.3382	+1.3%	0.3645	-1.2%	0.3218	+2.1%	0.2988	+1.9%	0.4204	+0.3%
K-NRM+DELM+TF [6]	0.3351	+0.4%	0.3701	+0.3%	0.3121	-1.0%	0.2901	-1.0%	0.4203	+0.3%
K-NRM+ExpaNet [5]	0.3402	+1.9%	0.3702	+0.3%	0.3234	+2.6%	0.3004	+2.5%	0.4212	+0.5%
K-NRM+DocFreq	0.3501†	+4.9%	0.3714	+0.6%	0.3297†	+4.6%	0.3223†	+10.0%	0.4289	+2.4%
K-NRM+CQCount	0.3604†	+7.9%	0.3785	+2.5%	0.3386†	+7.4%	0.3345†	+14.1%	0.4398†	+5.0%
NRM-F [9]	0.3747†	+12.2%	0.4094†	+10.9%	0.3545†	+12.5%	0.3419†	+16.6%	0.4776†	+14.0%
NeuDEF-TF	0.3947†,‡	+18.2%	0.4246†	+15.0%	0.3430†	+8.8%	0.3672†,‡	+25.3%	0.4896†,‡	+16.8%
NeuDEF-NoTrans	**0.4054**†,‡,§	+21.4%	0.4688†,‡,§	+27.0%	0.3584†	+13.7%	0.3785†,‡	+29.1%	0.5023†,‡,§	+19.9%
NeuDEF	0.4038†,‡,§	+20.9%	**0.4730**†,‡,§	+28.1%	**0.3675**†,‡,§,¶	+16.6%	**0.3858**†,‡	+31.6%	**0.5056**†,‡,§	+20.7%

Table 2: Performance on title (T) and body (B) field individually on Sogou-KNRM. Relative performances compared to K-NRM and the significant improvements over K-NRM†, NRM-F‡, NeuDEF-TF§ and NeuDEF-NoTrans¶ are compared in each field group.

	Testing-RAW, MRR						Testing-DIFF			
Model	**Head**		**Torso**		**Tail**		**NDCG@1**		**NDCG@10**	
K-NRM(T)	0.3440	-	0.3747	-	0.3244	-	0.3132	-	0.4288	-
NRM-F(T)	0.3835†	+11.5%	0.4174†	+11.4%	0.3448	+6.3%	0.3706†	+18.3%	0.4651†	+8.5%
NeuDEF-TF(T)	0.3905†	+13.5%	0.4523†,‡	+20.7%	0.3322	+2.4%	0.3678†	+17.4%	0.4864†	+13.4%
NeuDEF-NoTrans(T)	**0.4104**†,‡,§	+19.3%	0.4692†,‡	+25.2%	0.3608†,§	+11.2%	0.3759†	+20.0%	0.4985†,‡	+16.3%
NeuDEF(T)	0.4017†,‡,§	+16.8%	**0.4775**†,‡	+27.4%	**0.3706**†,‡,§,¶	+14.2%	**0.3763**†	+20.1%	**0.5003**†,‡	+16.7%
K-NRM(B)	0.2728	-	0.3226	-	0.2539	-	0.2275	-	0.3744	-
NRM-F(B)	0.3486†	+27.8%	0.3907†	+21.1%	0.3147†	+23.9%	0.3092†	+35.9%	0.4601†	+22.9%
NeuDEF-TF(B)	0.3608†	+32.3%	0.3945†	+22.3%	0.3164†	+24.6%	0.3335†	+46.6%	**0.4730**†	+26.3%
NeuDEF-NoTrans(B)	0.3605†	+32.1%	0.4140†,‡,§	+28.3%	0.3279†	+29.1%	0.3304†	+45.2%	0.4729†	+26.3%
NeuDEF(B)	**0.3633**†	+33.2%	**0.4309**†,‡,§,¶	+33.6%	**0.3397**†,‡,§,¶	+33.8%	**0.3338**†	+46.7%	0.4715†	+25.9%

Implementation Details. All baselines use the same setting in prior research [8]: 300-dimension embedding layer; 165,877-word vocabulary; one exact match kernel ($\mu = 1, \sigma = 10^{-3}$); and ten kernels equally distributed in (-1, 1) ($\mu \in \{0.9, 0.7, ..., -0.9\}, \sigma = 0.1$). 1 multi-head attention layer with 4 heads is used in NeuDEF. Documents with no clicked query are not expanded. The learning of all methods use the same training data. All neural methods use Adam optimizer with learning rate 0.001, batch size 64 and $\epsilon = 1e{-}5$, and early stopping on 20% random selected validation data.

4 EVALUATION

This section presents evaluation results.

4.1 Overall Ranking Accuracy

Table 1 lists the results on Sogou-KNRM. NeuDEF outperforms all baselines. It improves NRM-F, the previous state-of-the-art, on Head, Torso, and Tail. NeuDEF outperforms it base ranker (K-NRM) on tail queries by (13.7%). The attention mechanism learns effective expansion weights: NeuDEF significantly outperforms NeuDEF-TF. The transformer helps on the tail, as compared with NeuDEF-NoTrans.

Table 2 lists the results of K-NRM, NRM-F, and NeuDEFs on the title and body individually. NeuDEF performs the best on each field. Note that it has been observed that adding body fields does not contribute much [8, 9]. How to better model long body text is a future research direction. The performances of NeuDEF and main baselines on Sogou-QCL [10] are in Table 3. The trends are similar.

4.2 Learned Expansion Weights

This experiment analyzes the expansion weights learned by NeuDEF's attention mechanism on three groups of expansion terms: those from clicked queries that have *No Overlap* with the document content, those have *Partial Overlaps*, and those *Contained* by the document. Figure 1 shows the distribution of terms in these three groups and their average expansion weights normalized on each document.

About 10% clicked queries have no term overlap with the document title or body (they might be retrieved by some query expansion-alike techniques). NeuDEF assigns about one thirds of its learned attention weights on clicked queries that have no overlap with the document; with document title, 10% of expansion terms from the No Overlap group received 30% of expansion weights. NeuDEF

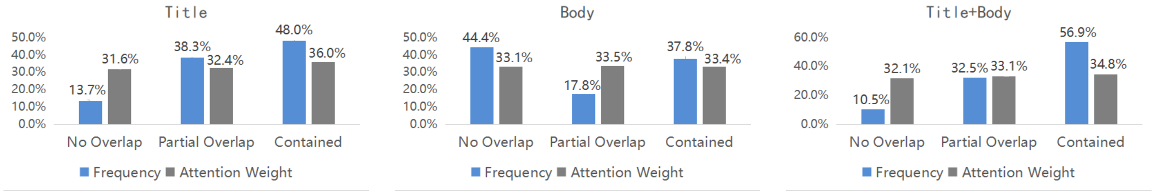

Figure 1: Frequency Distribution and normalized Attention Weights on the expansion terms from three clicked query groups: those have No Overlap with, Partial Overlap with, or are Contained by the corresponding document fields.

Table 3: Accuracy on Sogou-QCL dataset. Relative performances are compared to K-NRM. Statistical significance is marked by †(K-NRM), ‡(NRM-F), §(NeuDEF-TF) and ¶(NeuDEF-NoTrans).

	Testing-TACM			
Model	**NDCG@1**		**NDCG@10**	
K-NRM [8]	0.2409	-	0.3888	-
NRM-F [9]	0.2546†	+5.7%	0.4438†	+14.1%
NeuDEF-TF	0.2734†,‡	+13.5%	0.4606†,‡	+18.5%
NeuDEF-NoTrans	0.2801†,‡,§	+16.3%	**0.4651**†,‡	+19.6%
NeuDEF	**0.2804**†,‡,§	+16.4%	0.4643†,‡	+19.4%

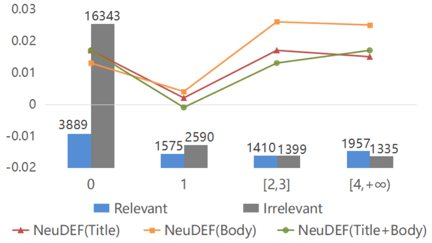

Figure 2: Performance on documents with different number of clicked queries. X-axis is the number of clicked queries. Histograms are the number of documents. Plots and Y-axis are the average ΔRR compared to K-NRM; higher is better.

learns to favor novel expansion terms that are related to but do not appear in the document, which may bring in extra information.

4.3 Document Level Performances

In our experiments, all testing queries are "unseen" queries as they never appear in the training split nor used in the document expansion. The advantage of NeuDEF is that it leverages the feedback signals at document level: a query might never appear before in the query log but the candidate documents may have seen before.

This experiment studies NeuDEF's document level performance w.r.t. different amounts of user feedback signals. It groups documents based on their number of clicked queries, then it evaluates the ΔRR of NeuDEF over K-NRM on each combination of document fields. The results on head queries are shown in Figure 2. Results on torso and tail are similar and omitted due to space constraints.

The number of clicked queries per document follows a long tail distribution. The user preferences also heavily favor popular documents; documents with more click queries are more likely to be relevant. NeuDEF performs better than K-NRM on all groups and with all types of document content. Even on documents with no clicked queries where NeuDEF withdraws to the base K-NRM model, adding expansion terms provides extra information in training and helps NeuDEF learn better parameters than its base ranker.

5 CONCLUSIONS AND FUTURE WORK

This paper presents NeuDEF, a neural document expansion approach that enriches document representations for neural ranking models using user feedback signals. Experiments demonstrate NeuDEF's effectiveness and its ability to better utilize user feedback signals and generalize them to unseen queries through document expansions.

Future work includes bringing expansion terms from external resources and developing more advanced neural expansion models.

6 ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Program of China (No. 2018YFC0831901). We thank Sogou for providing access to the search log.

REFERENCES

[1] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of WSDM 2018.*

[2] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W.Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of CIKM 2016.*

[3] Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2017. Time-aware click model. *ACM Transactions on Information Systems* (2017).

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

[5] Jian Tang, Yue Wang, Kai Zheng, and Qiaozhu Mei. 2017. End-to-end learning for short text expansion. In *Proceedings of KDD 2017.* ACM, 1105–1113.

[6] Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of ACL 2006.* Association for Computational Linguistics, 407–414.

[7] 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.

[8] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of SIGIR 2017.*

[9] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural ranking models with multiple document fields. In *Proceedings of WSDM 2018.*

[10] Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-QCL: A New Dataset with Click Relevance Label. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* ACM, 1117–1120.