Review paper on deep learning approach for addressing diverse relevance patterns in ad-hoc retrieval

Aditya Upadhyay
Birla Institute of Technology and Sciences,Pilani
Computer Science and Engineering
Email: f20170083@pilani.bits-pilani.ac.in

***Abstract*—This survey paper discusses some of the key concepts and methods for improving the relevance in ad-hoc retrieval for the documents with varying relevance levels, i.e. the documents that are either having document-level relevance or passage-level relevance. For this discussion, one long paper which attempts to model relevance pattern occurring at both document level and passage level, by using a neural network approach.**

1. Introduction

One of the questions in the field of research for ad-hoc retrieval is to learn a generalised function for accessing relevance metrics between a query and a document. There are many methods for accessing relevance which includes traditional ones, like ones based on term-frequency, and extends to deep learning-based methods that include layers that make feature selection and then tries to provide a subset of the corpus as a relevant document for the retrieval task. It can be through relevance feedback and its variants, including pseudo relevance feedback, probabilistic relevance feedback. We consider a document to be relevant, if any part, even as small as a sentence, is considered as relevant for a user query. Hence for a query, the entire document may be relevant(verbosity hypothesis), or it can be partially relevant(space hypothesis). Based on the methods adopted for relevance matching, the existing methods for ad-hoc retrieval fall broadly under three categories:

Document wide approaches
These methods focus on document-wide relevance signal to compare the relevance of a particular document to the user query. A relevance signal is any metrics for assessing the relevance of a given data to the query posted for it. There are traditional retrieval models as well as machine learning-based retrieval model. The traditional retrieval models rely on computing similarity of the query-document pair based on the mathematical and statistical approaches, for example, considering term-frequency and inverse document frequency for finding the similarity and the ranking function used in BM25 which is a bag-of-words based probabilistic model. The Machine learning-based retrieval model uses a feature extraction stage to take out the relevance signal that can be query dependent, document independent or query-document independent.

Moreover, in later stages, these features are used to learn linear/non-linear retrieval models. Similarly, deep learning-based methods are of two types: representative focused model, that creates an abstract representation for both query and document to access for relevance matching, and interaction focused model, that find histogram to get document-wide relevance signals. This model aggregates local relevance signals for getting global relevance signals, but the two are not competitive. The main problem with these methods is their inherent bias towards long documents owing to strong relevance signal present in them.

Passage level approaches
Here, the author has considered the methods, that take passage-level relevance signal into account, rather than those that use a passage as retrieval unit. These methods adopt a pre-segmentation at passage level. Various researches have shown that the results from passage-based retrieval are more reliable than document-level relevance matching.

Hybrid approaches
Researches found that using a hybrid approach, i.e. by combining the signals from passage level with document-wide signals is better than using either of the two independently. However, a linear combination of the two signals does not lead to a far improvement. These methods generally fail to capture a diverse variety of relevance signals found in the document due to unified combination strategies, hence giving a mixed performance on different datasets.

As discussed above, while the current approaches either consider the document-wide relevance metrics or passage-level relevance metrics. Even,when a model tries to combine the two types of relevance signals, still it is limited to linear combination of the two, leaving diverse relevance patterns existing in documents of a corpus unaddressed. Hence,it requires a more adaptive model that can give better retrieval performance in these systems also. This paper tries to summarize the key points and observation from the [1],while simultaneously discussing the model architecture.These paper is based on problem to increase the retrieval metrics(namely Mean Average Precision,Precision@K, NDCG@K) to serve

the user with a more appropriate and effective retrieval system and provides a model that is better than current state-of-the-art methods in ad-hoc retrieval. Here, we will be discussing the a deep learning model in which a data-driven approach is applied to find document-level and passage-level relevance signal and comparing these two for the final output. The Neural Model proposed for this is HiNT(HIerarchial Neural maTching model). It consists of two layers: a local matching layer and a global decision layer. The former layer produces a set of local relevance signal between the query and document by using deep matching networks to learn the passage level-relevance signals wheres the latter one, accumulates these local signal at different granularities and generates final global relevance assessment by letting these signals compete with each other for representation.

2. HINT model: Architecture

The model broadly consists of two layers as discussed in the subsequent sections:

2.1. Local matching layer

This layer is responsible for generating passage-level relevance signal, using which global relevance layer, provides relevance measures.Each document D can be represented as D = $[P_1, P_2..., P_K$ where P_i represents the i^{th} passage present in the document, and K represents the total number of passages in the document. These passages are combined with a query using a relevance matching function f to get the final relevance evidence for the documents.Formally,

$$e_i = f(P_i, Q)$$

,where i varies from 1 to K(the number of passages in document).The Window passages are passages marked by a fixed number of words in them,(obtained as a result of fixed-sized sliding window). Window passages are most widely adopted, owing to their simplicity and effectiveness for document retrieval. Since the final assessment of relevance depends heavily on the chosen relevance matching model, hence it is required that chosen model must incorporate multiple heuristics and statistical functions including exact and semantic matching, proximity and term importance. For this model, a fixed size window is used to define passages and an existing spatial GRU (Gated Recurrent Unit) model for relevance matching between query and document. The implementation of the model is discussed in the upcoming sections:

2.1.1. Input Layer. Query and document are represented as term vectors to that a relevance can be measured easily and with high effectiveness. Moreover, the passage is formed by adopting a fixed-size sliding window approach for segmenting documents into passages.i,e,

$$Q = [\ w_1^{(Q)}, w_2^{(Q)}, w_3^{(Q)}..., w_M^{(Q)}],$$

$$D = [\ w_1^{(D)}, w_2^{(D)}, w_3^{(D)}..., w_N^{(D)}],$$

$$P = [\ w_1^{(P)}, w_2^{(P)}, w_3^{(P)}..., w_L^{(P)}]$$

where M, N, L represents query length, document length and passage length, respectively.

2.1.2. Deep Relevance Matching network. Two matching matrix M^{cos} and M^{xor} are considered that incorporates semantic matching and exact matching, respectively. These two are used to distinguish exact matching signal and semantic matching signal as an exact matching signal are more important for ad-hoc retrieval task. Entries of matrix M^{cos} are filled by the cosine similarity between the corresponding query and document words whereas for the matrix M^{xor}, these are filled by the xor of the corresponding query and document words, as the name suggests.These 2D tensors are later converted to a 3D tensor

$$S_{ij} = [x_i, y_j, M_{ij}]$$

where the $x_i = w_i^Q * W_s$ and $y_j = w_j^P * W_s$, where W_s is the transformation parameter to be learnt. spatial GRU is applied to these two tensors to get the passage-level relevance signal. GRU is applied in two directions,s i.e. from top left to bottom right and from top right to bottom left to enrich the relevance signals generated at this layer i.e.

$$H_{ij} = g(H_{i-1,j}, H_{i,j-1}, H_{i-1,j-1}, S_{ij})$$

where H denotes the hidden state for exact matching and semantic matching layer. The two hidden states are aggregated to form the local relevance evidence by considering the last value(i.e. when i=M(query size) and j=L(last passage)).

$$e = [H_{ij}^{cos}, H_{ij}^{xor}]$$

The local relevance evidence from the two directions is then aggregated to give final local relevance signal.

$$e^{final} = [e, e^{rev}]$$

where e^{rev} represents the evidence vector in reverse direction.

2.2. Global Decision layer

The document-relevance pattern can be diverse across all the documents in the corpus, i.e. a document may be fully or partially relevant to a user query. Hence, it is required that the global decision layer does not stick to static rules but should evaluate by competing for the document -relevance signal at different granularity. A hybrid of independent decision model and accumulative decision model, the namely hybrid decision model, is taken into account. An independent decision model uses the top-k relevance signal to provide the final judgement, and an accumulative decision model accumulates the passage level signal provided by LSTM and uses top-k accumulated signal for final relevance judgment. In the hybrid model, the passage level relevance signals are allowed to compete, after undergoing a non-linear transform for scale uniformity between two, with

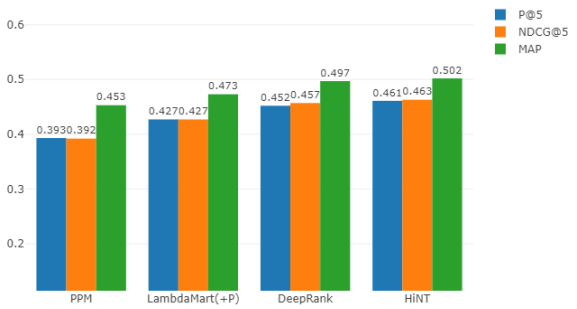

Figure 1. Plot of performance metrics against model name

an accumulated signal to select out top-k relevance signal among two, which is then fed to a multi-layered perceptron for final output. This ensures that both document-level and passage-level relevance signal gets representation in the final assessment.

2.3. Model Training

The model uses pair-wise ranking loss function along with Adam optimization for training.

3. Experimentation and Outcomes

For evaluation of performance, the model used two datasets: MQ2007 and MQ2008. These datasets are primarily used because of having a large number of queries and availability of original content. For data-preprocessing, all the documents and queries are white-space tokenised,lowercased and stemmed. Along with this stopwords and words appearing infrequently are also removed. Moreover, the documents are further, segmented into passages using the fixed-window size sliding window.

For comparing the outcomes, three types of models are incorporated viz. traditional retrieve models, learning to rank models and deep matching models.

In the traditional retrieval model, PPM(probabilistic passage model) is considered as it is a discriminative probabilistic model in capturing passage-level signals and uses a linear interpolating function to combine document retrieval scores with passage retrieval scores. Also, despite document-wide retrieval model, BM25 performed nice and served as a strong baseline for comparing other models. LambdaMart is a representative list-wise model. It uses a gradient boosting approach and is currently, the state-of-the-art learning to rank algorithm. As the HiNT model utilises passage-level features, hence some feature-based comparison features are introduced into the LambdaMart to create LambdaMart(+P) which will be used for discussion related to learning to rank algorithms.

Here, the deep learning model considered for comparison is DeepRank, which is considered a state-of-the-art deep matching model. Because of the inherent depth of the deep learning model, they are prone to overfitting. To avoid overfitting, the number of parameters is reduced in Convolutional network and fully connected network to adapt to the relatively small size of datasets.

To understand the impact of exact matching signal and semantic matching signal,authors considered different variants of HiNT model including,M^{cos}+MLP, M^{xor}+MLP. These models represent the configuration of HiNT model, hence throwing light on how different signals compete with each other with different strength. By choosing different types of models and IR heuristic, it is evident that the exact matching signals(based on the performance of M^{xor}+MLP) are critical for ad-hoc retrieval tasks. Also, if the inherent model is unable to distinguish the exact matching signal from the semantic matching signal, it can lead to a significant drop in the performance. (As the $M^{cos} + MLP$ model inherently carries both exact matching and semantic matching signals). However, if this is not the case(e.g. for spatial GRU), then, semantic matching relevance signal provides better performance, than exact matching signals. It is also shown that encoding a variety of IR heuristic a local relevance matching layer can improve the performance significantly. The plot of different variants of the local matching layer is shown against the MAP in figure

As compared against the independent $HiNT^{ID}$ and the accumulative decision model $HINT^{AD}$, the hybrid model $HINT^{HD}$ used in the paper outperformed in the performance parameter. Moreover, the accumulative model, outperformed the independent model, telling that accumulated signal from a variety of text can give better results than signals from different texts. By comparing the performance of PLM(passage language model) and PPM against the BM25 model, it is observed that a simple combination produces mixed results on the corpus with diverse relevance pattern documents. It is also, observed that learning to Rank models outperforms the traditional models owing to access to a more significant and rich set of features. Finally, in deep learning models, DSSM performed worse than even traditional retrieval models(since using only similarity matrix), emphasising the need for exact matching models in ad-hoc retrieval tasks.

To provide an overview on the experimental outcomes as proposed in the paper, a plot in figure 1 is provided that compares the best model among the chosen ones in each class viz. traditional models, learning to rank models, deep learning models and HiNT model. In each of these types, the original paper considered different models; however for succinctness, we will be considering only those models which performed best in their respective types. These models are compared against the precision@5, non-discounted cumulative gain@5 and Mean Average Precision(MAP).

For the relation of passage size,used for discriminating passages,and performance, it has been observed that upon increasing the passage size, the performance over benchmark dataset MQ2007, first increases till it reaches the peak performance and then starts to deteriorating .The reason proposed for this,assumes that small passage size,can lead to poor relevance matching signals whereas a large passage size, produces limited amount of passages, which

are relevant hence affecting at the global layer.

4. Conclusion

In this paper, we have summarised the proposal of the research paper [1], to provide the reader with an overview of the shortcomings in the current-state-of-art methods and the need to develop a new Neural Retrieval Model, namely HiNT model. As discussed by the authors in the original paper, this model is open to future research as it can include other implementations for the two layers used in the model. Also, this model can be extended to include features like PageRank to further improve retrieval performance. Moreover, current research paper uses cosine similarity only, for the semantic matching matrix. However, it can be further modified by trying other distance metrics like Euclidean distance, to further increase the model retrieval performance. However, this model suffers from the limitations faced by classical retrieval models, i.e. user utility may not be captured by the relevance parameters. In such cases, it is better to go with relevance feedback and query expansion measure by involving user more deeply in the retrieval task. Applying relevance feedback(can be positive feedback or negative feedback) based on the relevant and non-relevant passages [2], can also be useful.Moreover, One can also consider using NPRF [3],as a solution to increase the relevant documents returned.

Acknowledgments

The author would like to thank the administration for supporting and providing support and guidance for effectively completing this term paper.

References

[1] Y. Fan, J. Guo, Y. Lan, J. Xu, C. Zhai, and X. Cheng, "Modeling diverse relevance patterns in ad-hoc retrieval," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 375–384. [Online]. Available: https://doi.org/10.1145/3209978.3209980

[2] E. Brondwine, A. Shtok, and O. Kurland, "Utilizing focused relevance feedback," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1061–1064. [Online]. Available: https://doi.org/10.1145/2911451.2914695

[3] C. Li, Y. Sun, B. He, L. Wang, K. Hui, A. Yates, L. Sun, and J. Xu, "NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4482–4491. [Online]. Available: https://www.aclweb.org/anthology/D18-1478