Green: main sales point
Yellow: main technical point
Red: hypothesis
Orange: important prior work
Light orange: justification
Light cyan: hypothesis strengthener
Light pink: hypothesis weakener

Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking

Wei Yang,[1] Kuang Lu,[2] Peilin Yang, and Jimmy Lin[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of Electrical and Computer Engineering, University of Delaware

ABSTRACT

Is neural IR mostly hype? In a recent SIGIR Forum article, Lin expressed skepticism that neural ranking models were actually improving *ad hoc* retrieval effectiveness in limited data scenarios. He provided anecdotal evidence that authors of neural IR papers demonstrate "wins" by comparing against weak baselines. This paper provides a rigorous evaluation of those claims in two ways: First, we conducted a meta-analysis of papers that have reported experimental results on the TREC Robust04 test collection. We do not find evidence of an upward trend in effectiveness over time. In fact, the best reported results are from a decade ago and no recent neural approach comes close. Second, we applied five recent neural models to rerank the strong baselines that Lin used to make his arguments. A significant improvement was observed for one of the models, demonstrating additivity in gains. While there appears to be merit to neural IR approaches, at least some of the gains reported in the literature appear illusory.

1 INTRODUCTION

In a recent SIGIR Forum opinion piece, Lin [12] criticized the state of information retrieval research, making two main points. First, he lamented the "neural hype" and wondered that for "classic" *ad hoc* retrieval problems (limited relevance judgments and no behavioral data), whether neural ranking techniques represented genuine advances in effectiveness. As anecdotal evidence, he discussed two recent papers that demonstrated improvements over weak baselines, but in absolute terms, the reported results were no better than a well-tuned bag-of-words query expansion baseline.

In this paper, we attempt a rigorous evaluation of the claims made by Lin. Following his general arguments and focusing specifically on the test collection from the TREC 2004 Robust Track, a meta-analysis of the literature shows no upward trend in reported effectiveness over time. The best reported results on the collection are from a decade ago, and no recent paper (using neural approaches or otherwise) has reported effectiveness close to those levels. Analysis of over one hundred papers confirms that the baseline comparison conditions are often not as strong as they should be. Thus, Lin's criticism that comparisons to weak baselines still pervade the IR community rings true.

As a follow up, we applied a number of recent neural ranking models from the MatchZoo toolkit [5] to rerank the strong baselines that Lin used to make his arguments. Out of five neural models, one was able to significantly improve upon Lin's results. In other words, the effectiveness gains from one neural model is additive with respect to a strong baseline—which provides evidence that neural IR *can* lead to "real" improvements. Nevertheless, four out of the five models examined were not able to significantly beat the baseline, suggesting that gains attributable to neural approaches are not as widespread as the literature suggests. The absolute average precision values we report represent the highest values for neural models that we are aware of, although in absolute terms they are still much lower than the best known results.

2 META-ANALYSIS

The broader context of Lin's article is a recent series of papers that reflects a general angst (at least by some researchers) about the state of machine learning and its applications, in particular regarding empirical rigor and whether genuine advances are being made [13, 17]. These issues are not new, and similar discussions have been going on in IR for a while. The landmark study by Armstrong et al. [3] in 2009 found that comparisons to weak baselines pervade the literature. A decade later, is this still the case?

We began by conducting a meta-analysis to rigorously examine Lin's criticism. His argument specifically focused on document ranking models that could be trained with commonly-available evaluation resources; specifically, such models should not require log data. As he argued, the test collection from the TREC 2004 Robust Track (Robust04 for short) is the best exemplar of such data. In order to restrict the scope of our meta-analysis, we followed his line of reasoning and compiled a list of all papers that have reported experimental results on Robust04.

We exhaustively examined every publication from 2005 to 2018 in the following venues to identify those that reported results on the TREC 2004 Robust Track: SIGIR, CIKM, WWW, ICTIR, ECIR, KDD, WSDM, TOIS, IRJ, IPM, and JASIST. This was supplemented by Google Scholar searches to identify a few additional papers not in the venues indicated above. We applied a few exclusion criteria, best characterized as corner cases—for example, papers that only used a subset of the topics or papers where the focus was a task other than *ad hoc* retrieval, such as query difficulty prediction. In total, we arrived at 108 papers, summarized in Table 1. All raw data are documented and publicly available.[1]

For each paper, we extracted the highest average precision score achieved on Robust04 by the proposed methods, regardless of experimental condition (ignoring oracle conditions and other unrealistic setups). We further categorized the papers into neural (16) and non-neural (92) approaches. Methods that used word embeddings but not neural networks directly in ranking were considered "neural" in our classification. From each paper we also extracted the authors' baseline: in most cases, these were explicitly defined; if multiple were presented, we selected the best. If the paper did not explicitly mention a baseline, we selected the best comparison condition using a method *not* by the authors.

[1] https://github.com/lintool/robust04-analysis

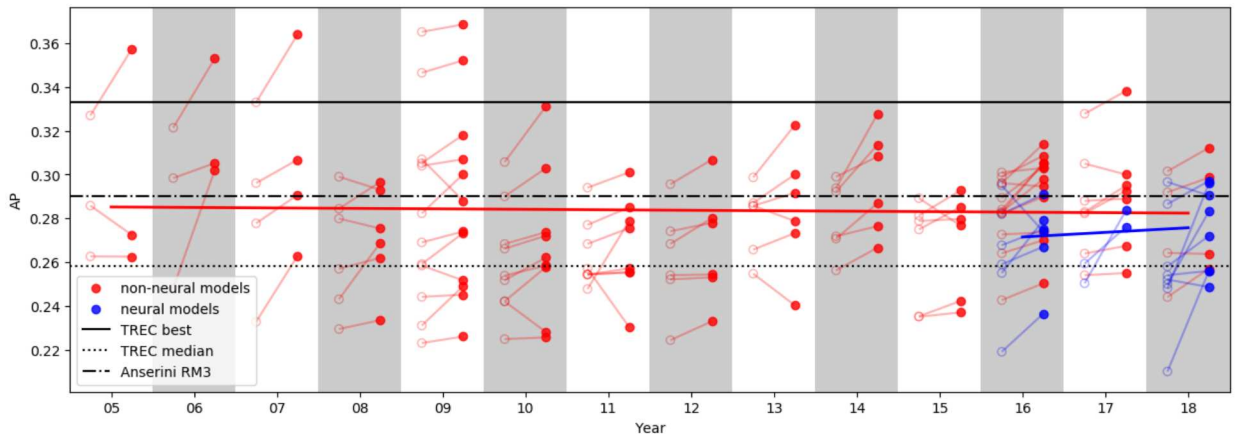

Figure 1: Visualization results on Robust04, where baseline and best AP scores are represented by empty and filled circles.

Year	Count	Venues
2005	3	SIGIR (1), CIKM (2)
2006	3	SIGIR (2), ECIR (1)
2007	4	SIGIR (1), CIKM (1), TOIS (1) , IRJ (1)
2008	6	SIGIR (4), ECIR(1), IPM (1)
2009	12	SIGIR (7), CIKM (1), ECIR (1), ICTIR (2), IPM (1)
2010	9	SIGIR (3), CIKM (4), WSDM (1), ECIR (1)
2011	8	SIGIR (2), CIKM (4), ICTIR (1), IRJ (1)
2012	6	SIGIR (2), CIKM (1), WSDM (1), ECIR (1), IRJ (1)
2013	6	SIGIR (1), CIKM (1), ICTIR (1), TOIS (1), IPM (2)
2014	6	SIGIR (2), TOIS (1), IRJ (2), JASIST (1)
2015	6	SIGIR (1), CIKM (1), ICTIR (2), TOIS (1), IRJ (1)
2016	18	SIGIR (3), CIKM (7), ICTIR (3), WWW (1), ECIR (2), TOIS (1), JASIST (1)
2017	9	SIGIR (5), WSDM (1), ECIR (1), ICTIR (2)
2018	12	SIGIR (2), CIKM (3), ECIR (2), ICTIR (2), IPM (1), EMNLP (2)

Table 1: Summary of papers examined.

A visualization of our meta-analysis is shown in Figure 1. For each paper, we show the baseline and the best result as an empty circle and a filled circle (respectively), connected by a line. All papers are grouped by their publication year. Neural approaches are shown in blue, and non-neural approaches in red. We also show two regression trendlines, for non-neural as well as neural approaches. The best submitted run at the TREC 2004 Robust Track (TREC best) at AP 0.333 is shown as a solid black line, and the median TREC run under the "title" condition at AP 0.258 is shown as a dotted black line (TREC median). Finally, we show the effectiveness of an untuned RM3 run (i.e., default parameters) from the Anserini system (used in the experiments described in Section 3). Its score of AP 0.2903 is higher than 65 papers (60.2%).

Our meta-analysis shows that researchers still frequently compare against weak baselines: In 36 papers (33.3%), the baseline lies below the TREC median. In fact, 25 papers (23.1%) report best results that are below the TREC median. Across all 108 papers, only six reported scores higher than the TREC best. The highest AP we encountered was by Cormack et al. [4] in 2009, at 0.3686. Across over a decade's worth of publications, we see no obvious upward trend in terms of effectiveness on this test collection.

Focusing specifically on the neural approaches, 8 out of 16 papers (50.0%) used a baseline that is below the TREC median. The highest reported score we encountered was by Zamani et al. [25] in 2018, at 0.2971. While there does appear to be an upward trend in the effectiveness of the neural approaches, the best reported score is still quite a bit lower than the TREC best and nowhere close to the best known results.

We are aware that not all papers purport to advance retrieval effectiveness (for example, papers about efficiency, proposing different frameworks, etc.). Nevertheless, these papers are in the minority, and we believe that our visualization provides an accurate high-level snapshot of the state of the field on this test collection. It appears that Lin's admonishments about continued use of weak baselines and skepticism about neural ranking models are warranted.

3 EXAMINING ADDITIVITY

Beyond revealing comparisons to weak baselines as widespread, Armstrong et al. [3] further examined why exactly this was methodologically problematic. Such comparisons lead to improvements that "don't add up" because of non-additive gains. The prototypical research paper on *ad hoc* retrieval proposes an innovation and compares it to a baseline that does not include the innovation; as expected, the innovation leads to increases in effectiveness. In this way, researchers collectively introduce dozens of different innovations, all of which improve on their respective baselines.

The key question, however, is whether the effectiveness gains of these innovations are additive. This might not occur, for example, if they exploit the same relevance signals. To put more precisely, does an improvement over a weak baseline still hold if applied to a strong baseline? If the answer is *no*, then gains over weak baselines may be illusory, and from a methodological perspective, we should not accept gains as "real" and "meaningful" unless they improve over strong baselines. Armstrong et al. [3] presented some evidence that many improvements are not additive, a finding which has been confirmed and expanded on by Kharazmi et al. [11]. However, the

debate is not fully settled, as Akcay et al. [2] demonstrated additivity in search result diversification after better parameter tuning.

In the second part of our study, we explicitly examine the additivity hypothesis with respect to recent neural ranking models. Specifically, we applied neural ranking models on top of the strong baselines that Lin used to make his arguments, which showed that a well-tuned implementation of query expansion based on RM3 [1] beats the average precision reported in two recent neural IR papers, anonymously referred to as "Paper 1" and "Paper 2".

3.1 Experimental Setup

We began by replicating Lin's results with the Anserini toolkit [22], using exactly the same experimental settings (tokenization, stemming, etc.) described in an online guide.[2] These runs used exactly the same cross-validation splits as Paper 1 (two-fold) and Paper 2 (five-fold), thus supporting a fair comparison.

On top of Lin's runs, we applied a number of neural ranking models from MatchZoo (version 1.0) [5]: DSSM [10], CDSSM [18], DRMM [7], KNRM [20], DUET [15]. These models were selected because they were specifically designed for *ad hoc* retrieval; other models available in MatchZoo, such as ARC-I [9], MV-LSTM [19], and aNMM [21] were mainly designed for short texts and not geared towards handling documents (which are much longer). MatchZoo is implemented in Keras, using the TensorFlow backend.

The neural models were deployed in a reranking setup, where the output of the models were linearly interpolated with scores from the RM3 baseline: $\text{score} = \alpha \cdot \text{score}_{\text{NN}} + (1 - \alpha) \cdot \text{score}_{\text{RM3}}$ Note that this design allows the possibility of disregarding the RM3 scores completely, with $\alpha = 1$. In our architecture, Anserini passes the raw text (minus markup tags) of the retrieved documents to MatchZoo, which internally handles document processing (tokenization, embedding lookup, etc.) prior to inference.

Following established practice, all models were trained using only the documents in the baseline RM3 runs that appear in the Robust04 relevance judgments (i.e., as opposed to all judgments in the qrels file). Word vectors were pre-trained on the Google News corpus (3 billion words).[3] The entire test collection has 249 topics (with relevance judgments). For the two-fold cross-validation condition to match Paper 1, we randomly sampled 25 topics from the training fold as the validation set; the other fold serves as the test set. For the five-fold cross-validation condition to match Paper 2, we selected three folds for training, one fold for validation, and used the remaining fold for testing. In all cases, we selected model parameters to maximize average precision on the development test. The weight α for score interpolation with RM3 was selected in the same manner. We set the maximum training epochs to five and used early stopping with five patience iterations. The batch size was set to 100 and all "title" queries were padded to ten tokens. Other hyperparameters were tuned using the validation set. All models were trained on an NVIDIA GeForce GTX 1080 GPU; it takes about one minute to train the DRMM model and a few hours for the others.

[2] github.com/castorini/Anserini/blob/master/docs/experiments-forum2018.md
[3] github.com/mmihaltz/word2vec-GoogleNews-vectors

Condition	AP	NDCG@20
BM25 [7]	0.255	0.418
DRMM [7]	0.279	0.431
2-fold results from Paper 1		
Paper 1	0.2971	-
BM25+RM3	0.2987	0.4398
+ DSSM	0.2993	0.4467
+ CDSSM	0.2988	0.4455
+ DRMM	0.3126†	0.4646†
+ KNRM	0.3033	0.4423
+ DUET	0.3021	0.4471
5-fold results from Paper 2		
Paper 2	0.272	-
BM25+RM3	0.3033	0.4514
+ DSSM	0.3026	0.4491
+ CDSSM	0.2995	0.4468
+ DRMM	0.3152†	0.4718†
+ KNRM	0.3036	0.4441
+ DUET	0.3051	0.4502

Table 2: Experimental results applying neural models to rerank a strong baseline; † indicates statistical significance.

3.2 Results

Our experimental results are shown in Table 2. Of all the neural models we examined in MatchZoo, only the original DRMM paper evaluated on Robust04; the first two rows show the DRMM results and their BM25 baseline (both copied from the original paper [7]). The paper reported a fairly substantial gain in AP, but based on our meta-analysis, the baseline is right around the TREC median and the DRMM score is still below Anserini RM3 (see Figure 1).

The second and third blocks of Table 2 report results from the two-fold and five-fold cross-validation conditions to match Paper 1 and Paper 2. Results from Paper 1 and Paper 2 are provided for reference (neither report NDCG@20). Note that our BM25+RM3 results are slightly higher than the results reported by Lin [12] because of code improvements after the publication of the article. We see that our "baseline" already beats the best results reported in Paper 1 and Paper 2. Based on our meta-analysis, an AP score of 0.3033 (five-fold) beats all published neural results and is better than the best results in 86 out of 108 papers (79.6%).

Experiments show that reranking our strong baseline with neural models yields small improvements in many cases.[4] Statistical significance of metric differences was assessed using a paired two-tailed t-test: the only observed significant difference is with DRMM ($p = 0.0032$). Even correcting for multiple hypothesis testing (e.g., Bonferroni correction), this difference remains statistically significant. Our five-fold cross-validation result of 0.3152 with DRMM is the highest reported AP using a neural approach that we are aware of; nevertheless, there are still 10 out of 92 non-neural papers (10.9%) from our meta-analysis that beat this.

[4] The reader might wonder how it is possible that a neural model actually makes results worse, since a setting of $\alpha = 1.0$ would ignore the neural model scores. However, due to cross-validation, this may not be the learned parameter.

3.3 Discussion

We specifically tackle a number of shortcomings and limitations of our study. First, only the five models implemented in MatchZoo were examined, and the quality of those implementations might be questioned. We concede this point, and so our findings apply to only the *MatchZoo implementations* of the various neural models. Nevertheless, MatchZoo has gained broad acceptance in the community as a solid experimental platform on which to explore neural ranking tasks.

The next obvious objection is that we've only examined these particular five neural ranking models. This, of course, is valid criticism, but an exhaustive study of *all* models would be impractical. We argue that the models selected are representative of the types of approaches pursued by researchers today, and that these results suffice to support at least some tentative generalizations.

The next criticism we anticipate concerns our evidence combination method, simple linear interpolation of scores. While there are much more sophisticated approaches to integrating multiple relevance signals, this approach is commonly used [6, 8, 16, 23, 24]. In a separate experiment where we explicitly ignored the retrieval scores, effectiveness was significantly lower. We leave open the possibility of better evidence aggregation methods, but such future work does not detract from our findings here.

Another criticism to our study is the limited data condition—we are training with only TREC judgments. Surely, the plethora of training data that comes from behavioral logs must be considered. While we do not dispute the effectiveness of neural approaches given large amounts of data, exploring the range of data conditions under which those models work is itself interesting. We note a stark contrast here: for NLP tasks, researchers have been able to extract gains from neural approaches with only "modest" amounts of data (as a rough definition, datasets that can be created outside an industrial context without behavioral logs). If it is the case that IR researchers cannot demonstrate gains *except with data only available to large companies*—this in itself is an interesting statement about neural IR. Mitra and Craswell [14] classified DRMM as a lexical matching modeling (in fact, the model explicitly captures *tf* and *idf*). DUET is a hybrid lexical/semantic matching model, while the others are semantic matching primarily. One possible interpretation of our findings is that TREC judgments alone are not sufficient to train semantic matching models.

Finally, there is a modeling decision worth discussing: In our experiments, all models except for DRMM truncate the length of the input document to the first K tokens (the `text2_maxlen` parameter in MatchZoo). Somewhat surprisingly, this is a practical issue that does not appear to be discussed in previous papers, but has a direct impact on model training time. We performed a coarse-grained sweep of the parameter and discovered that a value of K above 200 appears to be sufficient and doesn't seem to alter effectiveness substantially (one contributing factor might be the writing style of news articles). The results reported here use a K value of 500, which is longer than most documents, but still yields reasonable model training times. We believe that document truncation can be ruled out as a reason why four of the five neural ranking models do not yield additive improvements.

4 CONCLUSIONS

We believe that our study supports the following conclusions: At least with respect to the Robust04 test collection, it does not appear that the IR community as a whole has heeded the admonishments of Armstrong et al. [3] from a decade ago. Our meta-analysis shows that comparisons to weak baselines still pervade the literature. The high water mark on Robust04 in terms of average precision was actually set in 2009, and no reported results (neural or otherwise) come close. Focusing specifically on the implementations of five neural ranking models in MatchZoo, we find that only one is able to significantly improve upon a well-tuned RM3 run in a reranking setup on this collection. While neural networks no doubt represent an exciting direction in information retrieval, we believe that at least some of the gains reported in the literature are illusory.

REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, M. Smucker, T. Strohman, H. Turtle, and C. Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *TREC*.

[2] M. Akcay, I. Altingovde, C. Macdonald, and I. Ounis. 2017. On the Additivity and Weak Baselines for Search Result Diversification Research. In *ICTIR*. 109–116.

[3] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements That Don't Add Up: Ad-hoc Retrieval Results Since 1998. In *CIKM*. 601–610.

[4] G. Cormack, C. Clarke, and S. Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *SIGIR*. 758–759.

[5] Y. Fan, L. Pang, J. Hou, J. Guo, Y. Lan, and X. Cheng. 2017. MatchZoo: A toolkit for deep text matching. *arXiv:1707.07270* (2017).

[6] D. Ganguly, D. Roy, M. Mitra, and G. Jones. 2015. Word Wmbedding Based Generalized Language Model for Information Retrieval. In *SIGIR*. 795–798.

[7] J. Guo, Y. Fan, Q. Ai, and W. Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM*. 55–64.

[8] C. Van Gysel, M. de Rijke, and E. Kanoulas. 2018. Neural Vector Spaces for Unsupervised Information Retrieval. *TOIS* 36, 4 (2018), Article 38.

[9] B. Hu, Z. Lu, H. Li, and Q. Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *NIPS*. 2042–2050.

[10] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *CIKM*. 2333–2338.

[11] S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. 2016. Examining Additivity and Weak Baselines. *TOSI* 34, 4 (2016), Article 23.

[12] J. Lin. 2018. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (2018), 40–51.

[13] Z. Lipton and J. Steinhardt. 2018. Troubling Trends in Machine Learning Scholarship. *arXiv:1807.03341v2* (2018).

[14] B. Mitra and N. Craswell. 2017. Neural Models for Information Retrieval. *arXiv:1705.01509v1* (2017).

[15] B. Mitra, F. Diaz, and N. Craswell. 2017. Learning to Match using Local and Distributed Representations of Text for Web Search. In *WWW*. 1291–1299.

[16] J. Rao, W. Yang, Y. Zhang, F. Ture, and J. Lin. 2019. Multi-Perspective Relevance Matching with Hierarchical ConvNets for Social Media Search. In *AAAI*.

[17] D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. In *ICLR Workshops*.

[18] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. 2014. Learning Semantic Representations using Convolutional Neural Networks for Web Search. In *WWW*. 373–374.

[19] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng. 2016. A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations.. In *AAAI*. 2835–2841.

[20] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power. 2017. End-to-end Neural Ad-hoc Ranking with Kernel Pooling. In *SIGIR*. 55–64.

[21] L. Yang, Q. Ai, J. Guo, and W. Croft. 2016. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *CIKM*. 287–296.

[22] P. Yang, H. Fang, and J. Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *JDIQ* 10, 4 (2018), Article 16.

[23] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. *arXiv:1902.01718* (2019).

[24] H. Zamani and W. Croft. 2016. Embedding-based query language models. In *ICTIR*. 147–156.

[25] H. Zamani, M. Dehghani, W. Croft, E. Learned-Miller, and J. Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *CIKM*. 497–506.