MLWIKIR: A Python toolkit for building large-scale Wikipedia-based Information Retrieval Datasets in Chinese, English, French, Italian, Japanese, Spanish and more

Jibril Frej
jibril.frej@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG
* Institute of Engineering Univ. Grenoble Alpes

Didier Schwab
didier.schwab@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG
* Institute of Engineering Univ. Grenoble Alpes

Jean-Pierre Chevallet
jean-pierre.chevallet@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG
* Institute of Engineering Univ. Grenoble Alpes

ABSTRACT

Deep learning allowed for new state-of-the-art performance on *ad-hoc* information retrieval (IR). This approach usually requires large amounts of annotated data to be more effective than traditional baselines such as BM25. However, most standard *ad-hoc* IR datasets publicly available for academic research (e.g. *Robust04*, *ClueWeb09*) have at most 250 annotated queries and are usually in English only. Deep learning models for IR (e.g. DUET, Conv-KNRM) perform poorly on such datasets as they are trained and evaluated on large scale datasets collected from commercial search engines, not publicly available for academic research. This is a problem for reproducibility and the advancement of research. Moreover, most datasets are in English or Chinese only and deep learning models for *ad-hoc* IR are not evaluated on other languages. In this paper, we propose *MLWIKIR*: an open-source toolkit to automatically build large-scale information retrieval datasets based on *Wikipedia* in 10 different languages that can be adapted to any Wikipedia language given a tokenizer.

KEYWORDS

Information Retrieval, Open Source, Dataset, Deep Learning

1 INTRODUCTION

Recent deep learning (DL) models enable significant progresses in several fields of natural language processing (NLP) such as language modeling, question answering and natural language inference [5, 23]. However, textual *ad-hoc* information retrieval (IR) did not benefit from DL as much as other NLP tasks [4]. This difference can be explained by: **(1)** the lack of publicly available datasets *ad-hoc* IR with large amount of labelled data; **(2)** the complexity of the ranking problem that makes difficult the use of unsupervised learning for *ad-hoc* IR [4]. Consequently, most DL models that have been proposed for *ad-hoc* IR use one of the following approach: (1) training and evaluation on private large scale collection built from commercial search engine in English [16] or Chinese [21] that are not publicly available. Such experiments are expensive, time consuming and not reproducible; (2) training and evaluation on MQ2007 and MQ2008 datasets that are publicly available [6, 18]. However, such datasets do not have large amount of labelled queries (see Table 1) which can restrain the DL model design; (3) Using weak supervision [4, 25] that consists in pre-training a model using label produced by an unsupervised method (e.g., BM25). This approach can bias large models to rank similarly as the unsupervised method.

Related Work. Recently, several large scale datasets for *ad-hoc* IR have been made available for public research. Zheng et. al [26] proposed *Sogou-QCL*, a publicly available dataset in Chinese made from a commercial search engine. In 2016, Microsoft AI & Research released MS MARCO: a collection of datasets focused on deep learning in search [17]. MS MARCO is regularly updated with new tasks such as question answering, KeyPhrase Extraction or passage ranking, however the document ranking task (that was used at the TREC 2019 Deep Learning Track [2]) has not been released to the public yet and is in English only. To our knowledge, the MS MARCO document ranking dataset will be made available at the end of the month of march 2020. Finally, Frej et. al [7] developed *WIKIR*: a toolkit for building English IR datasets from Wikipedia and made two datasets available: wikIR78k and wikIRS78k. However most datasets for *ad-hoc* IR are in English or Chinese (see Table 1). Therefore, DL models developed for *ad-hoc* IR have not been evaluated on most languages. Consequently there is no empirical evidences that such models will be effective for most languages, especially models that makes explicit assumptions about relevance, such as DRMM [10].

Following the work of Frej et. al, we propose *MLWIKIR*: a python toolkit to build *ad-hoc* IR datasets from Wikipedia in 10 different languages. *MLWIKIR* can also be used to train and evaluate DL models for *ad-hoc* IR on the datasets it constructed.

In short, our contributions are the following:

- We provide *MLWIKIR*: a toolkit to build *Wikipedia*-based Information Retrieval datasets in 10 different languages;
- We make available 20 datasets built with *MLWIKIR*;
- We train and evaluate several deep learning models on our datasets and study their effectiveness with respect to the language.

https://github.com/getalp/wikIR

Dataset	#Query	#Doc	Avg #d^+/q	Language
GOV2	150	25M	181.51	English
ClueWeb09	200	1B	74.62	English
Robust04	250	0.5M	63.28	English
MQ2008	784	14k	3.82	English
MQ2007	1,692	65k	10.63	English
wikIRS78k	78k	2.4M	39.02	English
wikIR78k	78k	2.4M	39.02	English
Sogou-QCL	537k	5.4M	14.40	Chinese

Table 1: Statistics of several publicly available *ad-hoc* IR Dataset. Avg #d^+/q denotes the average number of relevant document per query.

Language	#Queries	#Documents	Avg #d^+/q
Swedish	4,996	158k	38.94
Dutch	10k	317k	43,31
Chinese	14k	282k	30.13
Russian	16k	591k	45.83
Italian	20k	509k	28.60
Spanish	22k	648k	43.42
French	24k	739k	50.42
Japanese	29k	657k	56.48
German	36k	1.1M	40.61
English	78k	2.4M	39.02

Table 2: Statistics of the datasets for each languages. Avg #d^+/q denotes the average number of relevant document per query

2 DATASET CONSTRUCTION

In this section, we describe the process of dataset construction made by *MLWIKIR* toolkit. The construction process is similar to the one made by Frej et. al [7]:

Query construction. Queries are extracted using either the title of the first sentence of Wikipedia articles. These two options are proposed in order to be able to build either short and well defined queries or long an noisy queries.

Document extraction. The set of documents consists of the set of Wikipedia articles. We remove title and first sentence from articles to avoid having documents that start with the exact formulation of queries. We do so to avoid favoring models that take into account exact matching signals and word order because of a bias in the data.

Relevance label construction. We propose 3 relevance levels: highly relevant (2); relevant (1) and non relevant (0). We consider that a document is highly relevant with respect to a query if they are constructed from the same Wikipedia article. We consider that a document is relevant with respect to a query if there is an internal Wikipedia link from the first sentence of the article related to the document to the article related to the query. For example, if we consider the query *"Continent"*, the most relevant (relevance = 2) document is *"... up to seven regions are commonly regarded as continents ... "* because they are built from the same the article. The document *"It comprises the westernmost part of Eurasia ... "* is relevant (relevance = 1) because the article *Europe* contains a link to the *Continent* article in it's first sentence. All query-document pairs that are not associated with a relevance level of (2) or (1) are considered as non relevant.

3 DATASETS DESCRIPTION

In this section, we describe the datasets created with *MLWIKIR*.

We created 20 datasets in 10 different languages. Each language is associated with two datasets: one with queries constructed from the title of Wikipedia's articles and one with queries constructed from the first sentence of Wikipedia's articles. We offer a data set with short and clear queries from titles and a data set with long and noisy queries for first sentences to study the resistance of IR models to noisy queries. Datasets are built using the full set of Wikipedia article associated to the considered language. Queries and associated qrels are randomly split into training, validation and test sets of size 80%, 10%, 10% respectively. Statistics of all collections are displayed in Table 2.

4 EXPERIMENTAL SETTINGS

4.1 IR Models

We evaluated 4 models on our datasets: Okapi BM25 [19], DUET [16], DRMM [11] and Conv-KNRM [3].

BM25. Okapi BM25 [19] is a stat-of-the-art probabilistic model for IR that uses exact matching between query and document terms.

DUET. DUET [16] is a deep learning model for *ad-hoc* IR. It uses both local (exact matching signals) and distributed (word embeddings) representations of text as input to asses relevance between a query and a document. Representations are processed by convolutional, fully connected and pooling layers.

DRMM. DRMM [11] consists of a multi layer perceptron that takes as input a set of matching histograms (one for each query term). Bins of the histograms correspond to the count of local interaction (cosine similarity between embeddings) withing a given range (e.g., [0.5, 1)). Exact matching signals have their own bin. Because DRMM has few parameters (455) and it takes explicitly into account exact matching signals it usually requires few labelled data to outperform traditional baselines such as BM25.

Conv-KNRM. Conv-KNRM [3] is an interaction-focused model. It uses several convolutional filters to build multiple representations of query n-grams and documents n-grams. These representations are then compared using the cosine similarity in order to form several interaction matrices between the query and the document. Kernel pooling is applied to each of these interaction matrices to produce learning-to-rank features that are finally processed by a linear layer and a non linear activation function to produce the final matching score.

4.2 Implementation details

Tokenization. With the exception of Chinese and Japanese, we used a simple white space and punctuation-based tokenizer. We use Jieba tokenization system for Chinese articles and TinySegmenter for Japanese articles.

https://github.com/fxsjy/jieba

http://chasen.org/~taku/software/TinySegmenter/

Language	Model	ndcg@5		ndcg@10		ndcg@20	
		title	fist sentence	title	fist sentence	title	fist sentence
Swedish	BM25	0.4532	0.3487	0.4119	0.3155	0.4120	0.3176
	DUET	0.3437^{-}	**0.4027**$^{+}$	0.3166^{-}	**0.3670**$^{+}$	0.3271^{-}	**0.3597**$^{+}$
	DRMM	**0.4670**	0.3838^{+}	**0.4197**	0.3482^{+}	**0.4183**	0.3466^{+}
	Conv-KNRM	0.3453^{-}	0.3173	0.3242^{-}	0.2970	0.3362^{-}	0.2988
Dutch	BM25	0.3952	0.3186	0.3600	0.2902	0.3609	0.2886
	DUET	0.2625^{-}	0.3293	0.2472^{-}	0.3012	0.2570^{-}	0.3005
	DRMM	**0.4084**$^{+}$	**0.3370**$^{+}$	**0.3682**	**0.3065**$^{+}$	**0.3682**	**0.3060**$^{+}$
	Conv-KNRM	0.3550^{-}	0.2883^{-}	0.3308^{-}	0.2767	0.3348^{-}	0.2834
Chinese	BM25	**0.4269**	0.2965	**0.3917**	0.2741	**0.3960**	0.2759
	DUET	0.3777^{-}	**0.3696**$^{+}$	0.3473^{-}	**0.3330**$^{+}$	0.3494^{-}	**0.3255**$^{+}$
	DRMM	0.4211	0.3445^{+}	0.3901	0.3097^{+}	0.3954	0.3082^{+}
	Conv-KNRM	0.2928^{-}	0.2756	0.2850^{-}	0.2611	0.2993^{-}	0.2665
Russian	BM25	0.3495	0.1562	0.3194	0.1429	0.3236	0.1463
	DUET	0.3125^{-}	0.1149^{-}	0.2885^{-}	0.1085^{-}	0.2932^{-}	0.1170^{-}
	DRMM	**0.3592**$^{+}$	**0.1698**$^{+}$	**0.3286**$^{+}$	**0.1532**$^{+}$	**0.3309**$^{+}$	**0.1561**$^{+}$
	Conv-KNRM	0.3295^{-}	0.1501	0.3064^{-}	0.1388	0.3107^{-}	0.1450
Italian	BM25	0.2661	0.1988	0.2500	0.1835	0.2569	0.1883
	DUET	0.1090^{-}	0.1805^{-}	0.1200^{-}	0.1704^{-}	0.1413^{-}	0.1741^{-}
	DRMM	**0.2666**	**0.2157**$^{+}$	**0.2512**	**0.1985**$^{+}$	**0.2573**	**0.2009**$^{+}$
	Conv-KNRM	0.2398^{-}	0.1855	0.2292^{-}	0.1772	0.2415^{-}	0.1822
Spanish	BM25	0.2792	0.2476	0.2655	0.2292	0.2779	0.2340
	DUET	0.1286^{-}	0.2030^{-}	0.1373^{-}	0.1968^{-}	0.1571^{-}	0.2041^{-}
	DRMM	**0.2830**	**0.2523**	**0.2683**	**0.2353**	**0.2789**	**0.2373**
	Conv-KNRM	0.2786	0.2335	0.2653	0.2202	0.2723	0.2269
French	BM25	0.3139	0.2730	0.2928	0.2510	0.2973	0.2524
	DUET	0.2155^{-}	0.2342^{-}	0.2095^{-}	0.2207^{-}	0.2215^{-}	0.2262^{-}
	DRMM	**0.3183**	**0.2962**	**0.2964**	**0.2700**$^{+}$	**0.3022**$^{+}$	**0.2694**$^{+}$
	Conv-KNRM	0.3040	0.2673	0.2860	0.2507	0.2924	0.2557
Japanese	BM25	**0.3444**	0.2788	**0.3263**	0.2626	**0.3310**	0.2670
	DUET	0.3318	**0.3385**$^{+}$	0.3137^{-}	**0.3162**$^{+}$	0.3177^{-}	**0.3166**$^{+}$
	DRMM	0.3412	0.3030^{+}	0.3206	0.2818^{+}	0.3252	0.2839^{+}
	Conv-KNRM	0.2953^{-}	0.2806	0.2851^{-}	0.2705	0.2948^{-}	0.2791^{+}
German	BM25	0.3664	0.2874	0.3354	0.2611	0.3372	0.2638
	DUET	0.2306^{-}	0.3423^{+}	0.2319^{-}	0.3126^{+}	0.2517^{-}	0.3109^{+}
	DRMM	0.3686	0.3256^{+}	0.3378	0.2956^{+}	0.3417	0.2951^{+}
	Conv-KNRM	**0.3886**$^{+}$	**0.3500**$^{+}$	**0.3570**$^{+}$	**0.3200**$^{+}$	**0.3571**$^{+}$	**0.3176**$^{+}$
English	BM25	0.3269	0.2944	0.3045	0.2673	0.3098	0.2695
	DUET	0.3323	0.3252^{+}	0.3044	0.2964^{+}	0.3082	0.2951^{+}
	DRMM	**0.3462**$^{+}$	0.3188^{+}	**0.3189**$^{+}$	0.2872^{+}	**0.3227**$^{+}$	0.2868^{+}
	Conv-KNRM	0.3080^{-}	**0.3253**$^{+}$	0.2906^{-}	**0.3004**$^{+}$	0.2992^{-}	**0.3010**$^{+}$

Table 3: Performance comparison of different models on the test set of datasets constructed with *MLWIKIR*. Significant improvement/degradation with respect to BM25 is denoted as +/- with p-value < 0.01.

Stop words removal and stemming. With the exception of Chinese and Japanese text is stemmed and stop words are removed with the python *nltk* toolkit [15]. Chinese and Japanese are not stemmed and their stop words are removed based on the list provided by the *many-stop-words* library. **Evaluation metrics.** We use the normalized discounted cumulative gain [13] (nDCG) to evaluate the performance of IR models. We use a two-tailed paired t-test with Bonferroni correction to measure statistically significant differences between the evaluation metrics [8, 20].

Embeddings. For each language, we used the publicly available word embeddings pre-trained on Wikipedia and common crawl [9] with the *fasttext* [1] algorithm. Using *fasttext* allowed us to be able to associate embeddings to terms that were not occurring in the training data.

Neural Networks. We trained and evaluated the neural models for IR with *MatchZoo* [12] deep text matching library. We use the Adam optimizer [14] and the cross entropy loss function for ranking provided by *MatchZoo* to train neural models. Hyperparameters

https://pypi.org/project/many-stop-words/

are tuned to maximize the nDCG@5 on the validation set using random search. For efficiency reasons and as commonly done in *ad-hoc* IR [24], neural networks for IR are evaluated using re-ranking with BM25 as a first stage ranker.

5 RESULTS AND DISCUSSION

As we can see on Table 3, models react very differently to languages and query type.

Title. For most languages (with the exception of German, Japanese and Chinese) the DRMM model has the best performances on short and well defined queries. However it does not consistently achieves statistical significance against BM25. Moreover, in most cases DUET and Conv-KNRM do not even manage to achieve performances similar to BM25. Moreover, BM25 performs the best with respect to all metrics on Chinese and Japanese with short and well defined queries. The only exception is the German language: the Conv-KNRM produces the best results with statistical significance against BM25. These results suggest that BM25 is still a strong baseline when considering short and well defined queries. In other terms, neural approaches for *ad-hoc* IR may be more useful for noisy and ambiguous queries.

First sentence. As we might have expected, BM25 on long and noisy queries achieves worst performances than when using short and well defined queries. This is also the case for DRMM and Conv-KNRM but interestingly that's not the case for the DUET model. With the exception of Russian, the DUET model performs similarly or better when it is trained and evaluated on queries based on first sentences of articles rather than queries bases on titles. Further investigation is required to explain this behavior. DRMM outperforms BM25 with statistical significance on most languages. The DUET produces the highest quality results on Swedish, Chinese and Japanese and Conv-KNRM performs the best on English and German. Such results empirically confirm that neural approaches for *ad-hoc* IR are especially useful for long and noisy queries. However, depending on the language, the gain obtained by neural approaches varies greatly when considering deep architectures such as DUET and Conv-KNRM. On the one hand, DRMM always improves the performances of BM25, which is consistent with the work of Yang et.al [22]. However, because DRMM has very few parameters and relies on the limited information from the matching histograms, it does not benefit from large amounts of training data and cannot outperform BM25 with a large margin. On the other hand, DUET is less consistent than DRMM in the sense that, depending on the language considered, it can outperform BM25 with a large margin (German, Japanese, Swedish) or show statistically significant degradation compared to BM25 performances (Russian, Italian, Spanish, French). Further experiments with datasets comparable in size are required in order to study the effectiveness of IR models on different languages.

6 CONCLUSIONS AND FUTURE WORK

In this work, we propose *MLWIKIR*: a publicly available toolkit for building Wikipedia-based IR datasets in multiple languages. We used *MLWIKIR* to build 20 large scale *ad-hoc* IR datasets in 10 different languages. These datasets will be made publicly available for research and reproducibility purposes. We also trained, evaluated and compared several neural networks for *ad-hoc* IR on each of these datasets in order to empirically study the effectiveness of deep leaning approaches on different languages. The scripts to train and evaluate IR models on our datasets will also be made available for reproducibility purposes.

As future work, we plan to use *MLWIKIR* to produce datasets in different languages that are comparable in size in order to study the effect of language on the performance of deep learning models. We will also use our datasets to pre-train deep learning models and fine tune them on traditional IR datasets as a form of weak supervision [4].

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2019. Overview of the trec 2019 deep learning track. *TREC (to appear)* (2019).

[3] 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 the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. ACM, New York, NY, USA, 126–134. https://doi.org/10.1145/3159652.3159659

[4] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* https://arxiv.org/abs/1704.08803

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling Diverse Relevance Patterns in Ad-hoc Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018.* 375–384. https://doi.org/10.1145/3209978.3209980

[7] Jibril Frej, Didier Schwab, and Jean-Pierre Chevallet. 2019. WIKIR: A Python toolkit for building a large-scale Wikipedia-based English Information Retrieval Dataset. *arXiv preprint arXiv:1912.01901* (2019).

[8] Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (Feb. 2018), 32–41. https://doi.org/10.1145/3190580.3190586

[9] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[10] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016.* 55–64.

[11] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/2983323.2983769

[12] J. Guo, Y. Fan, X. Ji, and X. Cheng. 2019. MatchZoo: A Learning, Practicing, and Developing System for Neural Text Matching. In *SIGIR 2019, Paris, France, July 21-25, 2019.* 1297–1300.

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[14] D. P. Kingma and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.* http://arxiv.org/abs/1412.6980

[15] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). http://arxiv.org/abs/cs.CL/0205028

[16] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match Using Local and Distributed Representations of Text for Web Search. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1291–1299. https://doi.org/10.1145/3038912.3052579

[17] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[18] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. DeepRank: A New Deep Architecture for Relevance Ranking in Information Retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. ACM, New York, NY, USA, 257–266. https://doi.org/10.1145/3132847.3132914

[19] Stephen E. Robertson and Steve Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*. 232–241. https://doi.org/10.1007/978-1-4471-2099-5_24

[20] Julián Urbano, Mónica Marrero, and Diego Martín. 2013. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*. 925–928. https://doi.org/10.1145/2484028.2484163

[21] 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 the 40th International ACM SIGIR conference on research and development in information retrieval*. ACM, 55–64.

[22] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. 1129–1132. https://doi.org/10.1145/3331184.3331340

[23] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR* abs/1906.08237 (2019). arXiv:1906.08237 http://arxiv.org/abs/1906.08237

[24] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 497–506.

[25] H. Zamani, M. Dehghani, W. B. Croft, E. G. Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *CIKM 2018, Torino, Italy, October 22-26, 2018*. 497–506.

[26] 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 &#38; Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. ACM, New York, NY, USA, 1117–1120. https://doi.org/10.1145/3209978.3210092