Adaptive Latent Entity Expansion for Document Retrieval

Iain Mackie[1], Shubham Chatterjee[2], Sean MacAvaney[3], and
Jeff Dalton[4]

[1] University of Glasgow, UK `i.mackie.1@research.gla.ac.uk`
[2] University of Edinburgh, UK `shubham.chatterjee@ed.ac.uk`
[3] University of Glasgow, UK `sean.macavaney@glasgow.ac.uk`
[4] University of Edinburgh, UK `jeff.dalton@ed.ac.uk`

Abstract. Despite considerable progress in neural relevance ranking techniques, search engines still struggle to process complex queries effectively — both in terms of precision and recall. Sparse and dense Pseudo-Relevance Feedback (PRF) approaches have the potential to overcome limitations in recall, but are only effective with high precision in the top ranks. In this work, we tackle the problem of search over complex queries using three complementary techniques. First, we demonstrate that applying a strong neural re-ranker before sparse or dense PRF can improve the retrieval effectiveness by 5–8%. Second, we propose an enhanced expansion model, Latent Entity Expansion (LEE), which applies fine-grained word and entity-based relevance modelling incorporating localized features. Specifically, we find that by including both words and entities for expansion achieve a further 2–8% improvement in NDCG. Our analysis also demonstrates that LEE is largely robust to its parameters across datasets and performs well on entity-centric queries. And third, we include an "adaptive" component in the retrieval process, which iteratively refines the re-ranking pool during scoring using the expansion model and avoids re-ranking additional documents. We find that this combination of techniques achieves the best NDCG, MAP and R@1k results on the TREC Robust 2004 and CODEC document datasets.

1 Introduction

A fundamental problem in information retrieval is query-document lexical mismatch [2]. A common approach to address this issue is Pseudo-Relevance Feedback (PRF), where a first-pass top-k candidate set of documents is retrieved, and these feedback signals can augment the query for a second-pass retrieval. Early work on PRF focused on term-based query expansion [32,1,33,16], with later work showing entity-based representations can offer improvements on the hardest topics [8]. Recently, this PRF paradigm has also leveraged dense vectors [35,46,53]. However, all these models suffer from the same problem: If the initial query is challenging, the candidate set is unlikely to contain relevant documents in the top ranks, which will cause PRF models to fail.

Meanwhile, neural language models (NLMs) for re-ranking [22] have led to significant advances in effectiveness, particularly precision in the top ranks. In this work, we pull together these research threads on neural re-ranking and entity-based expansion methods to improve the core task of document retrieval. Figure 1 shows how we address the problem of poor pseudo-relevance feedback by applying re-ranking prior to query expansion and re-executing this query. We find that expansion with NLM feedback improves the recall-oriented effectiveness of sparse and dense PRF approaches.

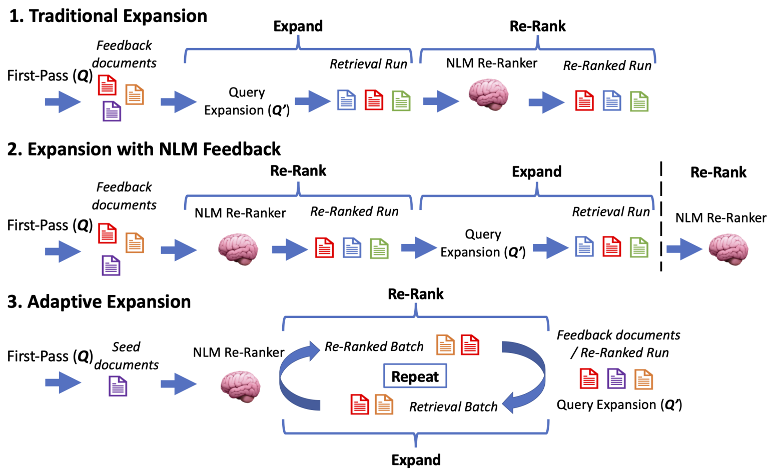


Fig. 1: Rethinking Query Expansion Pipelines Leveraging NLM Feedback

Armed with insights from this analysis, we propose a new model to improve PRF effectiveness further when operating over NLM feedback: Latent Entity Expansion (LEE). LEE is a joint probabilistic term and entity-based expansion model. In contrast with prior work in Latent Concept Expansion, we show that a hybrid expansion model with terms and entities is more effective than comparable individual expansion models. We also demonstrate improved effectiveness from passages based on NLM re-ranking that provide a more fine-grained hybrid relevance model. Furthermore, unlike prior work [33], we find that using dependencies from entity co-occurrence improves effectiveness with passage feedback, but can be harmful with document feedback.

Nonetheless, after our expansion with neural feedback, we find that a second round of neural re-ranking is required to maximize precision. Thus, we draw inspiration from recent adaptive re-ranking work [27] and propose our “adaptive expansion” framework. Specifically, Figure 1 shows how we dynamically refine the re-ranking pool during scoring using the expansion model. This allows us to use NLM feedback for expansion and re-ranking in a single pass and reduces the number of documents scored by around 35%.

Our document test collections, namely TREC Robust and CODEC, focus on challenging "complex" queries. Unlike recent web collections that emphasize "easy" factoid-focused queries, these collections represent challenging topics where existing state-of-the-art methods for sparse and dense retrieval still have significant headroom [29,3]. Through extensive experiments under various conditions, to our knowledge, LEE produces the highest recall ever achieved on these benchmark datasets by 6-12%. Query analysis shows that LEE's hybrid expansion model with terms and entities improves the hardest entity-centric queries, where a fine-grained relevance model and entity dependencies are particularly useful. Furthermore, LEE with adaptive expansion sets a new state-of-the-art for MAP and NDCG without requiring a second round of neural re-ranking, and our model parameters are robust across datasets. Overall, this work demonstrates the potential of probabilistic term-entity expansion models when combined with neural re-ranking. We summarize our contributions below:

- We provide a detailed study of existing probabilistic word and entity expansion models with document and passages feedback from neural re-ranking.
- We propose a new hybrid relevance model for query expansion that incorporates entity dependencies.
- We show that our unsupervised expansion model is state-of-the-art by 6-12% on recall, and when combined with additional neural re-ranking, result in 2-8% improvement on NDCG and MAP.
- We show that our hybrid relevance model with adaptive expansion achieves similar effectiveness additional NLM re-ranking (saving around 35% compute).

2 Related Work

Query Expansion: A common automatic approach for query expansion is *pseudo-relevance feedback* where the top-k documents from an initial retrieval set are assumed relevant. Famous classical methods include Rocchio [41], KL expansion [54], Relevance Modelling [32], and RM3 expansion [1]. Furthermore, recent work, such as CEQE [35], uses query-focused contextualized embedding for expansion. Conversely, this work evaluates expansion models based on NLM feedback. In particular, LEE builds upon Latent Concept Expansion (LCE) [33] to develop a hybrid probability distribution over both words and entities based on re-ranked passage feedback, incorporating entity dependencies.

The rise of dense retrieval has brought variants using vector-based PRF models [20], including ColBERT PRF [46], and ColBERT-TCT PRF [23], and ANCE PRF [53]. Our results shows sparse and entity-based approaches are currently more effective for document retrieval on complex topics. Nonetheless, we do find that using NLM feedback for dense retrieval improves recall.

Entity-Centric Ranking: Our work extends extensive research that incorporates entity-based representations within document ranking [43,31,48,43,49,25,47,24]. This research direction typically uses entity links [10,39,17,13] present in the query or documents to ground the task to an external Knowledge Base (KB).

Prior work has used entity-based query expansion methods to enrich the query with useful concepts [31,48,51]. Furthermore, entity-based language models have been used for document retrieval [8], and EQFE [8] enriches the query with KG entity-based features to improve the hardest topics. Moreover, the Word-Entity Duet [50] framework uses word-based and entity-based representations to embed documents and queries. Lastly, recent work [43] shows that enriching queries and documents using a dense end-to-end entity linking system [17] can provide knowledge-grounded context and improve initial retrieval.

Our work builds on this literature by proposing a hybrid word and entity relevance model derived using NLM passage feedback. Furthermore, we incorporate localization features such as entity dependence. We show that entities are beneficial when used with words (and actually competitive by themselves) given our strong NLM passage ranking, significantly outperforming all prior methods.

NLM Document Ranking: Neural language model (NLM) [22] has shown improvement across information retrieval tasks. However, NLM re-rankers cannot easily ingest the full text when ranking long documents due to the input constraints of these models. Various strategies deal with this problem; for example, BERT-maxp [7,52] and T5-maxp [37] shard long documents into passages that the model can score individually as a proxy for overall document relevance.

In our work, we build upon similar intuition and use NLM passages ranking [37] to identify the most relevant sections of documents to form a more fine-grained relevance model. Additionally, the precision improvements in the top ranks due to NLMs [6,5] provides a more accurate PRF for expansion. We also compare our results against state-of-the-art models fine-tuned on the target datasets. For example, CEDR-KNRM [28], PARADE [18], and MORES [12].

3 Adaptive Expansion

3.1 Rethinking Expansion Pipelines

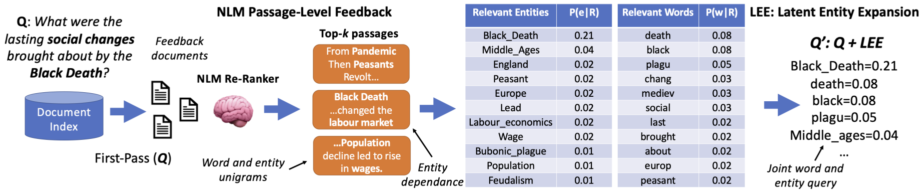
Based on the analysis of current retrieval models, we rethink the standard query expansion pipeline drawing on several research threads. Specifically, NLM reranking models [7,37,36] offer an opportunity to improve the precision of document feedback to form more effective expansion models. We also draw from recent work on adaptive re-ranking [27] to allow our expansion model to use NLM feedback without incurring additional re-ranking cost.

Formally, given a information need (query) Q, we want to return a ranked list of documents $D = [D_1, D_2, ..., D_N]$ relevant to the query Q from a collection C. For generality, documents, D, may also refer to other retrieval units, such as passages. We abstract a document ranking pipeline, and focus on changing the ordering of *query expansion* and *neural re-ranking* components. Figure 1 shows the three expansion pipelines we explore:

- **Traditional expansion**: Our standard document ranking pipeline with expansion [18,28,12]. Specifically, we retrieve an initial set of documents using our PRF retrieval models [1,46,35], before using a neural re-ranker to create our final re-ranked list of documents. The issue with this pipeline is that signals from advanced neural re-rankers are not used to improve initial recall.

- **Expansion with NLM feedback**: We move NLM re-ranking before our expansion model in the pipeline to improve the precision of the feedback set; thus, improving expansion effectiveness. Additionally, a second re-ranking pass could further improve the precision; however, this would also increase computational expense due to extra document scoring.
- **Adaptive expansion**: Instead of having a static run that we re-rank, we propose dynamically updating our document frontier as more documents are scored using our query expansion model. Similar to [27], we alternate our re-ranking of documents between the initial retrieval seed documents and the dynamic frontier based on the expansion model. This iterative batch process of re-ranking and expansion continues, with a batch size of b, until we reach our intended number of documents. Intuitively, updating our query expansion model as more documents are scored is similar to a manual researcher building their understanding of a topic through reading information. Additionally, unlike expansion with NLM feedback with a second re-ranking pass, adaptive expansion does not require additional computation from document re-ranking.

3.2 Latent Entity Expansion


Q: What were the lasting social changes brought about by the Black Death?
Document Index
First-Pass (Q)
Feedback documents
NLM Re-Ranker
NLM Passage-Level Feedback
Top-k passages
From Pandemic Then Peasants Revolt...
Black Death ...changed the labour market
...Population decline led to rise in wages.
Word and entity unigrams
Entity dependance
LEE: Latent Entity Expansion
Q': Q + LEE
Black_Death=0.21
death=0.08
black=0.08
plagu=0.05
Middle_ages=0.04
...
Joint word and entity query

Relevant Entities	P(e|R)
Black_Death	0.21
Middle_Ages	0.04
England	0.02
Peasant	0.02
Europe	0.02
Lead	0.02
Labour_economics	0.02
Wage	0.02
Bubonic_plague	0.01
Population	0.01
Feudalism	0.01

Relevant Words	P(w|R)
death	0.08
black	0.08
plagu	0.05
chang	0.03
mediev	0.03
social	0.03
last	0.02
brought	0.02
about	0.02
europ	0.02
peasant	0.02

Fig. 2: LEE hybrid query expansion using NLM fine-grained passage feedback

Figure 2 depicts our Latent Entity Expansion (LEE) model that incorporates words and entities. Specifically, our query expansion approach uses a strong NLM re-ranked list of documents, which benefits precision in the top ranks (making our feedback more accurate). Thus, we assume top-k documents to be query-relevant feedback R, which we use to construct LEE based on a hybrid relevance model of words ($\{w_1, w_2, ..., w_i\} \in D$) and entities ($\{e_1, e_2, ..., e_N\} \in D$). We use LEE to expand the initial, $Q \rightarrow Q_{LEE}$, and retrieve our documents, $[D_1, D_2, ..., D_N]$.

Deriving Expansion Words Equation 1 shows how we estimate the probability of a word $P(w|R)$ given the assumed relevant documents R. $P(Q|D)$ is obtained by normalizing the NLM scores [37], before we turn it into a probability by dividing the sum of all the normalized scores, $\sum_{D' \in R} P(Q|D')$. The probability of a word given a document, $P(w|D)$, is the term frequency divided by the document length. Following LCE [33], we normalize the distribution using $P(w|C)$ (that we approximate for convenience with $\mathrm{IDF}(w, C)$). Later, in Section 5.1, we show that this feature is important for modelling document relevance.

$$P(w|R) = \sum_{D \in R} \frac{P(Q|D)}{\sum_{D' \in R} P(Q|D')} P(w|D) P(w|C) \tag{1}$$

Deriving Expansion Entities Analogously, we estimate the query-relevance of a document based on the entities contained within that document ($e \in D$). Prior work [33] only uses unigram representations because word dependencies do not improve results. In contrast, LEE incorporates both entity unigrams and dependencies and finds meaningful improvement with passage NLM feedback. The base formulation for entity terms follows how we model word unigrams, providing a unigram estimate of $P(e|R)$. However, we also include entity dependence terms based on co-occurrence to model the relationship between entities.

Estimating relevance of entity dependence, we estimate this as follows:

$$P([e_1, e_2]|R) = \sum_{D \in R} \frac{P(Q|D)}{\sum_{D' \in R} P(Q|D')} P([e_1, e_2]|D) P([e_1, e_2]|C) \tag{2}$$

where $P(Q|D)$ is the normalised NLM score and $P([e_1, e_2]|D)$ is the sum of both entity frequencies divided by the document length. We approximate $P([e_1, e_2]|C)$ as the product of entity IDFs, $\mathrm{IDF}(e_1) \cdot \mathrm{IDF}(e_2)$. Incorporating entity co-occurrence increases the weighting of entities that co-occur with many entities in relevant documents. This helps prioritise the "central entities" that are particularly useful for identifying relevant documents. Unlike [33], results show this entity dependence feature is particularly beneficial with passage feedback, although not meaningful at a document level.

We then combine the unigram and entity dependence models as follows:

$$P(e|R) = \beta \sum_{e_i \in R} P([e, e_i]|R) + (1 - \beta) P_{\mathrm{unigram}}(e|R) \tag{3}$$

where $P([e, e_i]|R)$ is the probability of the entity pair (e, e_i) being in a relevant document, and $P_{\mathrm{unigram}}(e|R)$ is probability of entity e, obtained using a unigram language model.

LEE Duet Representation The final score of a document $D \in R$ is derived from an interpolation of the term-based and entity-based query scores:

$$\mathrm{Score}(D, Q) = \lambda \cdot \mathrm{Score}_{\mathrm{word}}(D, Q) + (1 - \lambda) \cdot \mathrm{Score}_{\mathrm{entity}}(D, Q) \tag{4}$$

where $\mathrm{Score}_{\mathrm{word}}(D, Q)$ is the document score based on our word query expansion, $\mathrm{Score}_{\mathrm{entity}}(D, Q)$ is the document score based on our entity query expansion. For simplicity to execute over large collections, we use BM25 [40] to execute our probabilistic queries over separate document and entity indexes. Furthermore, following work by RM3 [1], we also include the probabilistic interpolations between the terms in the original query and our probability distribution. We then normalise these scores and interpolate using $\lambda \in [0, 1]$. In practise, we find $\lambda = 0.5$ is reasonable across all datasets (see Appendix C for details) .

Adaptive Expansion with LEE We formalise adaptive expansion with LEE. Given the original query Q and the current re-ranked documents, D_{nlm}, we produce our duet representation, Q_{LEE}, to retrieve the next batch of unscored, b-sized documents to be re-score, D_{exp}. Thus, as more documents are scored, and the D_{nlm} set increases in size, our word and entity-based probabilistic query is updated and becomes more representative.

4 Experimental Setup

We release runs and hyperparameters for reproducibility: *link*. Additionally, on paper acceptance we will release all code and data.

4.1 Data Evaluation

Retrieval Corpora We evaluate using two test collections that focus on challenging and complex information needs [29,3]:
TREC Robust04 [45] focuses on poorly performing document ranking topics. This dataset comprises 249 topics, containing short keyword "titles" and longer natural-language "descriptions" of the information needs. We use 5-fold cross-validation with standard folds in previous work [14].
CODEC [30] focuses on the complex information needs of social science researchers (economists, historians, and politicians). This resource contains 42 essay-style topics and encompasses both document ranking and entity ranking tasks. We use the folds outlined within the resource for 4-fold cross-validation.

Entity Linking Entity links provide structured connections between the queries, documents and entities. We use KILT [38] to ground documents, which uses the 2019/08/01 Wikipedia containing around 5.9M entities.

Previous studies show that high-recall information extraction techniques are required for successful usage in ranking tasks [15]. Thus, we use WAT [39] for wikification [34] to ground both concepts and traditional named entities to Wikipedia pages. Additionally, we run a end-to-end entity linker ELQ [17] over the queries, which is optimized to provide entity links for questions.

Indexing and Retrieval We use Pyserini [21] version 0.16.0 for indexing the corpora and datasets for terms and entities. For words, we remove stopwords and use Porter stemming. We store the respective entity mentions using KILT's ids as unique terms for our entity-centric document and passage indexes.

Evaluation We assess the system runs to a run depth of 1,000. We focus on recall-oriented evaluation; thus, the primary measure for this paper is Recall@1000. Additionally, we report MAP and NDCG to understand precision across relevant documents. We use ir-measures for all our evaluation [26]. Lastly, we select a single baseline system for our statistical testing and use a 5% paired-t-test significance using the scipy Python package [44].

4.2 LEE Components and Hyperparameters

Neural Re-ranker (NLM) We use T5-3b [37], a neural re-ranking model (`castorini/monot5-3b-msmarco-10k`) that casts text re-ranking into a sequence-to-sequence setting . Following the paper [37], we shard documents in passages of 10 sentences with a stride length of 5 and use a max-passage aggregation approach. We use the same passage shards to construct query-specific knowledge for efficiency and to align the NLM score for passage expansion methods.

Retrieval and Expansion To avoid query drift, all LEE runs use a tuned BM25 system [40]. We tune LEE hyperparameters using a grid search and cross-validation to optimise R@1000. Specifically, we tune feedback passages (*fb_docs*: 10 to 100 with a step of 10), the number of feedback terms (*fb_terms*: 10 to 100 with a step of 10), the interpolation between the original terms and expansion terms (*original_query_weight*: 0.1 to 0.9 with a step of 0.1). For the entity component, we tune the co-occurrence weighting (β: 0.1 to 0.9 with a step of 0.1), and lastly, the hybrid weighting between word and entity (λ: 0.1 to 0.9 with a step of 0.1) and the run depth (k_{LEE}: 1000, 2000, 3000, and 4000). All hyperparameters are released for reproducibility: *link*

Adaptive Expansion we follow the same experimental setup as [27] to allow a fair comparison. Specifically, we can take an initial BM25 run (R_0) and use a batch size b of 16 to alternate between the initial BM25 run and LEE retrieval with a total re-ranking budget of 1,000 documents. We use same experimental setup for the adaptive re-ranking experiments and the tuned LEE hyperparameters from initial retrieval.

4.3 Comparison Methods

First-Pass Retrieval BM25 [40]: Base retrieval for expansion approaches, we tune $k1$ (0.1 to 5.0 with a step of 0.2) and b (0.1 to 1.0 with a step of 0.1).
BM25 ⇒ Relevance Model (RM3) [1]: We tune *fb_terms* (5 to 95 with a step of 5), *fb_docs* (5 to 100 with a step of 5), and *original_query_weight* (0.1 to 0.9 with a step of 0.1). This is our primary expansion baseline, and we separately tune RM3 expansion parameters on top of the NLM re-ranked run.
Latent Concept Expansion (LCE) [33]: We use the same tuning parameters sweeps as RM3 for both words and entity vocabularies.
SPLADE [11]: We use the first-passage learned sparse runs provided by the author, from checkpoint: `naver/splade-cocondenser-ensembledistil`.
ColBERT-TCT (TCT) [23]: We use TCT-ColBERT-v2-HNP's model in a max-passage approach for document retrieval with same pre-processing as NLM.

ColBERT-TCT with PRF (TCT ⇒ PRF) [19]: We adopt the default dense PRF parameters, i.e. Rocchio PRF depth is 5, α is 0.4 and β 0.6. Furthermore, we also implement a ColBERT-TCT PRF system on top of neural re-ranking.
ENT [43]: We re-implement their best standalone method, "Entities", where we expand queries and documents with the unique names of linked entities. We parameter-tune BM25 in the same manner as our term-based BM25.
ENT ⇒ RM3: We extend [43] to use RM3 expansion and tune parameters in the same manner as BM25 ⇒ RM3.
CEQE [35]: We use CEQE-MaxPool(fine-tuned) for initial retrieval comparison and (BM25+CEDR)+CEQE-MaxPool+CEDR for NLM feedback comparison.

Re-Ranking Entity Query Feature Expansion (EQFE) [8]: We include the best performing EQFE Robust04 run that is provided by the author.
NLM (T5-3B) [37]: We follow the same setup as described in Section 4.2.
CEDR [28]: We use the CEDR-KNRM runs with BERT-base embedding [9].
PARADE [18]: We use the runs from the ELECTRA-Base variant [4].

Adaptive Re-Ranking GAR-BM25 ⇔ NLM [27]: We modify this adaptive passage re-ranking [27] for document ranking. Specifically, we issue the re-ranked document terms as a BM25 query to identify the most similar documents.
GAR-TCT ⇔ NLM [27] We use ColBERT-TCT dense representations to calculate document-to-document similarity. We take the mean of each document's passage vectors as the query vector and do a max-passage exhaustive search.
GAR-ENT ⇔ NLM We extend [27] to represent documents using the WAT entity links and issue a BM25 query to identify similar documents.
RM3 For a fair adaptive comparison to LEE, we use the tuned RM3 model for adaptive expansion, issuing a word-based query after NLM re-ranking batches.

5 Results and Analysis

5.1 RQ1: What is the effectiveness of sparse and dense systems retrieval systems for document retrieval?

Table 1 compares the effectiveness of expansion models on top of an NLM run. Specifically, we compare a BM25 with RM3 expansion and neural re-ranking to our expansion models with NLM [37] feedback. We vary the expansion models (RM3, LCE, and our LEE model), the unit of feedback (documents and passages), and vocabulary (words and entities).

RQ1a: Are passages or documents more effective for NLM-focused expansion? Across both datasets, we see average relative improvement of passages (rows without ^{D}) to particularly improve NDCG (i.e. Robust04 titles +1.8%, descriptions +2.4%, and CODEC +7.2%) and MAP (i.e. Robust04 titles +2.0%, descriptions +3.2%, and CODEC +10.2%), with less relative improvement at R@1000. This shows that passages with NLM scoring provide a more

Table 1: Query expansion varying model (e.g. RM3, LCE, and LEE), NLM feedback (e.g. documents (^{D}) or passages), and vocabulary (e.g. "Ent" or words). Significance testing against BM25 $\Rightarrow$ RM3 $\Rightarrow$ NLM; significantly better ("$^{+}$") and worse ("$^{-}$").

		Robust04 - Title			Robust04 - Description			CODEC		
		NDCG	MAP	R@1k	NDCG	MAP	R@1k	NDCG	MAP	R@1k
1x Re-Rank	BM25 $\Rightarrow$ RM3 $\Rightarrow$ NLM	0.634	**0.377**	0.777	0.652	**0.406**	0.750	0.644	**0.377**	0.816
	BM25 $\Rightarrow$ NLM $\Rightarrow$ RM3-Ent^{D}	0.600^{-}	0.322^{-}	0.779	0.619^{-}	0.343^{-}	0.781^{+}	0.590^{-}	0.292^{-}	0.851^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ RM3-Ent	0.612^{-}	0.331^{-}	0.776	0.643	0.364^{-}	0.792^{+}	0.645	0.331^{-}	0.854^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ RM3^{D}	0.630	0.350^{-}	0.813^{+}	0.616^{-}	0.334^{-}	0.780^{+}	0.615	0.312^{-}	0.865^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ RM3	0.638	0.353^{-}	0.812^{+}	0.625^{-}	0.339^{-}	0.797^{+}	0.641	0.335	0.874^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LCE-Ent^{D}	0.614^{-}	0.335^{-}	0.797	0.640	0.360^{-}	0.806^{+}	0.578^{-}	0.283^{-}	0.849
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LCE-Ent	0.626	0.343^{-}	0.793	0.659	0.377^{-}	0.810^{+}	0.643	0.325^{-}	0.857^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LCE^{D}	0.636	0.353^{-}	0.824^{+}	0.659	0.375^{-}	0.829^{+}	0.606^{-}	0.313^{-}	0.872^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LCE	0.647	0.360^{-}	0.825^{+}	0.668	0.377^{-}	0.843^{+}	0.632	0.326^{-}	0.877^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LEE^{D} (Ours)	0.648	0.366	0.834^{+}	0.673^{+}	0.388	0.845^{+}	0.619	0.321^{-}	0.879^{+}
	BM25 $\Rightarrow$ NLM $\Rightarrow$ LEE (Ours)	**0.660**$^{+}$	0.376	**0.837**$^{+}$	**0.687**$^{+}$	0.401	**0.855**$^{+}$	**0.663**	0.357	**0.883**$^{+}$

fine-grained relevance signal for our query expansion, potentially reducing noise from long documents with less relevant passages.

RQ1b: How does LEE's word-entity expansion compare with existing expansion approaches? Across all datasets, LEE has the best R@1000 and NDCG of any expansion method. In fact, LEE has significantly better R@1000 compared to our NLM re-ranking baseline, with between 7.5-13.7% relative improvement. NDCG significantly improves on Robust04 titles and descriptions and shows relative improvements on CODEC of 2.9% (although not significant). LEE is the only query expansion technique where MAP across all datasets is not significantly worse than the base neural re-ranking pipeline. To the best of our knowledge, these are the best reported first pass R@1000 results across all datasets and highlight the strong recall-oriented effectiveness of word-entity hybrid models that build on neural passage re-ranking. See Appendix A for query-level analysis showing how LEE helps the hardest first-pass queries.

RQ1c: Does entity dependencies help our query expansion model? We find that fine-grained passage signals are important for leveraging entity information, especially when using dependencies to infer relationships between entities. We find that including the entity co-occurrences improves effectiveness versus simply modelling entities based on unigrams; they provide consistent improvements across the datasets increasing MAP by 3.3% on average, NDCG 0.9% and R@1000 0.4%, with no system being negatively affected. This is in contrast to entity co-occurrence at a document level, where MAP reduces on average by 0.2%, with small gains in NDCG and R@1000 of 0.3%, and Robust04 systems being negatively affected.

5.2 RQ2: How does query expansion with a second-pass NLM re-ranking compare to state-of-the-art ranking pipelines?

This research questions explores how our LEE expansion model with passage feedback compares to sparse and dense systems with an additional round of neural re-ranking. Specifically, Table 2 shows LEE (with a second re-ranking phase [37]) compared to current state-of-the-art neural and traditional models

Table 2: Expansion with NLM feedback and second-pass re-ranking; "+" significant improvement over BM25 ⇒ RM3 ⇒ NLM.

		Robust04 - Title			Robust04 - Description			CODEC		
		NDCG	MAP	R@1k	NDCG	MAP	R@1k	NDCG	MAP	R@1k
1xRank	SPLADE [11] ⇒ NLM	0.539	0.309	0.597	0.590	0.357	0.617	-	-	-
	EQFE [8]	0.601	0.328	0.806	-	-	-	-	-	-
	BM25 ⇒ CEQE [35] ⇒ NLM	0.626	0.373	0.764	-	-	-	-	-	-
	BM25 ⇒ RM3 ⇒ CEDR [28]	0.632	0.370	0.776	0.645	0.400	0.758	-	-	-
	BM25 ⇒ RM3 ⇒ PARADE [18]	0.642	0.380	0.776	0.650	0.408	0.758	-	-	-
	ENT ⇒ RM3 ⇒ NLM	0.615	0.366	0.745	0.658	0.407	0.759	0.490	0.373	0.833
	TCT ⇒ PRF ⇒ NLM	0.584	0.345	0.681	0.572	0.364	0.619	0.606	0.351	0.754
	BM25 ⇒ RM3 ⇒ NLM	0.634	0.377	0.777	0.652	0.406	0.750	0.644	0.377	0.816
	BM25 ⇒ NLM ⇒ LEE (Ours)	0.660$^+$	0.376	**0.837**$^+$	0.687$^+$	0.401	**0.855**$^+$	0.663	0.357	**0.883**$^+$
2xRank	CEDR ⇒ CEQE [35] ⇒ NLM	0.644	0.384	0.787	-	-	-	-	-	-
	TCT ⇒ NLM ⇒ PRF ⇒ NLM	0.592	0.349	0.697	0.630	0.390	0.702	0.636	0.369	0.808
	BM25 ⇒ NLM ⇒ RM3 ⇒ NLM	0.656$^+$	0.390$^+$	0.813$^+$	0.674$^+$	0.416$^+$	0.780$^+$	0.659	0.379	0.865$^+$
	BM25 ⇒ NLM ⇒ LEE ⇒ NLM (Ours)	**0.667**$^+$	**0.393**$^+$	**0.837**$^+$	**0.715**$^+$	**0.438**$^+$	**0.855**$^+$	**0.664**$^+$	**0.380**	**0.883**$^+$

on the target datasets. We also compare our system to comparable neural pseudo-relevance feedback techniques that leverage multiple rounds of re-ranking.

These results highlight how gains in effectiveness can be achieved with NLM feedback across standard sparse and dense PRF retrieval models. Specifically, we see relative recall improvements of 5% with RM3 and 8% with ColBERT-TCT-PRF using NLM feedback. Moreover, a second pass neural re-ranker over our LEE initial retrieval run further improves NDCG and MAP. This leads to NDCG and R@1000 being significantly improved compared to the state-of-the-art baseline, with MAP significantly better on Robust04 titles and descriptions. Additionally, LEE with re-ranking achieves the best MAP and NDCG scores when compared to state-of-the-art prior methods and NLM-focused expansion methods (CEQE, ColBERT-TCT-PRF, and RM3).

Although neural re-ranking improves MAP and NDCG results, it is worth highlighting how competitive the LEE unsupervised expansion method (without re-ranking) is when compared to prior work. Specifically, neural re-ranking only increases NDCG between 0.002-0.028 and MAP 0.018-0.037 across the datasets. For example, using the unsupervised LEE query on Robust04 titles, we achieve NDCG@10 of 0.561, which is higher than reported SPLADE [11] results 0.485 (a comparable unsupervised method), better than T5-3b [37] 0.545, and approaching fine-tune PARADE [18] 0.591. Appendix B shows how LEE specifically outperforms on entity-centric queries.

5.3 RQ3: Does adaptive expansion provide effectiveness gains without re-ranking more documents?

Table 3: Adaptive re-ranking effectiveness ("⇔"), with significance testing ("+") against BM25 ⇒ RM3 ⇒ NLM.

		Robust04 - Title			Robust04 - Description			CODEC		
		NDCG	MAP	R@1k	NDCG	MAP	R@1k	NDCG	MAP	R@1k
2xRank	BM25 ⇒ NLM ⇒ LEE ⇒ NLM (Ours)	0.667$^+$	**0.393**$^+$	0.837$^+$	**0.715**$^+$	**0.438**$^+$	**0.855**$^+$	0.664$^+$	0.380	0.883$^+$
1xRank	BM25 ⇒ RM3 ⇒ NLM	0.634	0.377	0.777	0.652	0.406	0.750	0.644	0.377	0.816
	BM25 ⇒ NLM ⇒ LEE (Ours)	0.660$^+$	0.376	0.837$^+$	0.687$^+$	0.401	**0.855**$^+$	0.663	0.357	0.883$^+$
1xRank (Adapt)	BM25 ⇒ GAR-BM25 ⇔ NLM	0.629	0.372	0.768	0.652	0.402	0.747	0.634	0.362	0.797
	BM25 ⇒ GAR-ColBERT ⇔ NLM	0.630	0.374	0.769	0.649	0.402	0.739	0.645	0.368	0.822
	BM25 ⇒ GAR-ENT ⇔ NLM	0.637	0.377	0.781	0.661	0.408	0.758	0.644	0.366	0.821
	BM25 ⇒ RM3 ⇔ NLM	0.655$^+$	0.387$^+$	0.813$^+$	0.675$^+$	0.418$^+$	0.783$^+$	0.653	0.373	0.847
	BM25 ⇒ LEE ⇔ NLM (Ours)	**0.668**$^+$	0.392$^+$	**0.838**$^+$	0.704$^+$	0.435$^+$	0.834$^+$	**0.669**$^+$	**0.382**	**0.887**$^+$

We explore combining LEE with adaptive expansion to improve effectiveness without a second pass re-ranking. Table 3 shows adaptive LEE expansion against LEE with two NLM passes, the adaptive "GAR" systems [27], and an adaptive RM3 expansion system for comparison.

RQ3a: Is adaptive expansion with LEE the most effective adaptive re-ranking method? We find that GAR-based methods that use words (GAR-BM25), entities (GAR-ENT), and dense representation (GAR-ColBERT) are not significantly better than a standard NLM re-ranking pipeline. In fact, even adaptive expansion with RM3 is consistently more effective than all GAR systems, being significantly better on Robust04 over the NLM re-ranking pipeline and better, although not significantly, on CODEC.

Moreover, these results support LEE with adaptive expansion as the most effective adaptive method across all datasets and measures. The significance testing aligns with two re-ranking phases, i.e. being significantly better on Robust04 across all measures and CODEC on R@1000 and NDCG. In fact, LEE with adaptive expansion is nominally better than two re-ranking passes on CODEC across all measures and Robust04 titles on NDCG and R@1000. Furthermore, Appendix C shows these state-of-the-art results can be achieved while being robust to using parameters across datasets.

RQ3b: What are the computational benefits of adaptive expansion? For simplicity, we measure computational expense by the number of documents that require NLM re-ranking, which should be a strong proxy across implementations and hardware. Thus, the computational benefits of adaptive expansion are due to the document set differences between the initial run (i.e. BM25) and the LEE with NLM feedback (i.e. BM25 ⇒ NLM ⇒ LEE). For example, on Robust04 titles, two passes of NLM results in 1,503 unique documents being scored per query, compared to only 1,000 for adaptive re-ranking (i.e. saves 33% scoring cost). We find similar trends in Robust04 descriptions (637 fewer documents to re-score) and CODEC (525 fewer documents to re-score).

6 Conclusions

We show that LEE word-entity expansion using fine-grain passage feedback from NLM re-ranking significantly improves R@1000, with between 8-14% improvement over RM3 expansion. Specifically, the joint modelling of words and entities at a passage level improves relevance modelling, including incorporating entity dependencies. Our method is robust in terms of query-level hurts vs helps, improves recall of the hardest queries by 0.6, and can use parameters across datasets without significantly harming effectiveness. Additionally, we show that our implicit entity ranking is highly effective within the top ranks and helps improve entity-centric queries. Lastly, we demonstrate that LEE with adaptive expansion can avoid two NLM passes and achieve state-of-the-art effectiveness without additional document re-ranking (saving 35% of the re-ranking cost). We believe adaptive expansion can lead to new dynamic expansion models to improve both effectiveness and efficiency.

APPENDIX

A LEE Query Analysis

We conduct a query-by-query analysis to understand why LEE has such significant improvements in R@1000. Focusing on Robust04 and comparing to BM25, we find that LEE helps 166 and hurts 33 title queries, compared to RM3, which helps 139 and hurts 47 queries. These findings are even more evident in description queries, where LEE helps 181 and hurts 30 queries, compared to RM3, which helps 156 and hurts 45 queries. This supports that LEE query expansion that leverages a combination of both words and entities is more robust than simply using words alone.

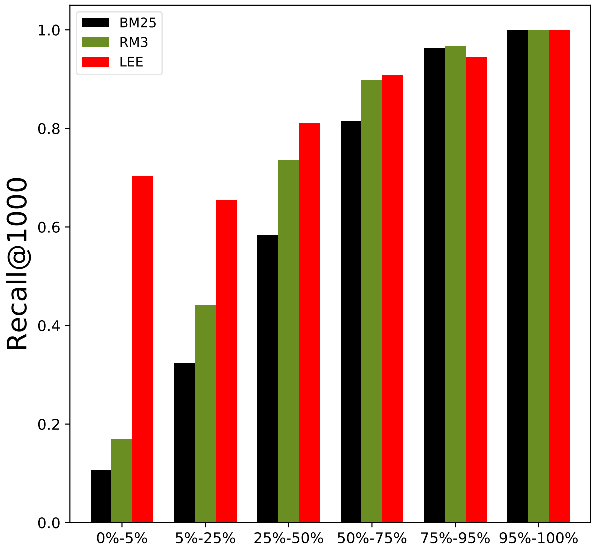


Fig. 3: Query difficulty plot stratified by original BM25 score on Robust04 descriptions with RM3 and NLM ⇒ LEE.

Figure 3 shows that the largest relative gains are on the hardest queries (ordered based on original BM25 retrieval effectiveness). Specifically, LEE improves R@1000 of the hardest 5% of queries by around 0.6 compared to BM25 and 0.55 compared to BM25 with RM3 expansion. Furthermore, substantial gains over RM3 are also observed in 5-25% (+0.2) and 25-50% (+0.1) buckets. This highlights that using top-k passages and joint modeling of terms and entities effectively improves the hardest queries, with minimal drop in effectiveness on the easy queries (i.e. we only slightly reduce effectiveness on 75%-95% band).

Analysing specific queries, we see that joint modelling of words and entities can be beneficial for retrieving relevant documents, when one probability distribution fails. For example, when term-based approaches fail, such as the Robust04 description query, “What impact has the Chunnel had on the British economy and/or the life style of the British?”, where BM25 and BM25 with RM3 expansions both have an R@1000 of 0.061. This failure is driven by vocabulary miss match, where “chunnel” is a less common colloquialism for “Channel Tunnel”. However, LEE achieves R@1000 of 0.862 through strong relevance signals from re-ranked passages due to avoiding vocabulary miss match by weighting [Channel_Tunnel] as the most relevant entity, thus capturing multiple lexical variations, i.e. “Chunnel”, “Channel Tunnel”, “Eurostar”, or “Eurotunnel”.

Additionally, passage feedback is vital for the CODEC query “What were the lasting social changes brought about by the Black Death?”, where using passage-level LEE query increases NDCG from 0.483 to 0.736. Both passage and document feedback methods correctly identify the central entity as [Black_Death], which is contained within 91% of judged relevant documents. However, the document-level feedback only has general topical entities within the higher ranks, such as [Pandemic], [Infection], and [Bubonic_plague]. On the other hand, passage-level relevance signals can identify other important entities, such as [Peasant] (i.e. [Black_Death] leads to [Peasant]s' revolts), and entity co-occurrences such as [Population] and [Wage], (i.e. [Black_Death] leads to changes in workers [Wage] due to [Population] decline). Thus, highlighting the importance of passage feedback for localized features.

B Performance on Entity-Centric Queries

Query analysis shows LEE outperforms on entity-centric queries, where the topical focus is a specific concept or named entity, where dense models struggle with these query types [42]. For example, Robust04 query 376, “World Court” where the user refers to “International Court of Justice”, ColBERT-TCT and ColBERT-TCT-PRF systems only achieve R@1000 of 0.137 and 0.147. Sparse methods also do not perform well, with R@1000 for BM25 of 0.225 and RM3 expansions of 0.235. The sparse models struggle as “world” and “court” are common words with many meanings and instances. However, LEE can use entity mentions to infer the specific instance of the “world court” and model the probability of [International_Court_of_Justice] entity the highest, increasing R@1000 to 0.735. We also find LEE without re-ranking performs better, with a MAP of 0.338 versus 0.250 when re-ranked again, which highlights the benefits of explicit entity modelling with LEE.

On further analysis, LEE's effectiveness within entity-centric information needs is not surprising when we analyse the implicit entity ranking from LEE using CODEC's entity judgements. We compare this to the best baseline systems provided with CODEC, where systems score entity relevance based on the relevance score of Wikipedia pages representing each entity [38]. We can see that LEE is very effective in the top ranks, achieving NDCG@3 of 0.767 and NDCG@10 of 0.554 (much higher than all dataset baselines). This highlights the accuracy of the LEE entity model and explains the improvement of queries requiring explicit entity modeling.

C Discussion of Parameters

As outlined in Section 4.2, we tune our LEE model following the official cross-validation setup outlined for target datasets. However, here we analyse: (1) how effective our method is zero-shot and (2) the impact of λ, i.e., the relative weighting of words and entities.

Table 4: Tuned LEE model vs zero-shot LEE model (CODEC parameters); “+” and “-” are significance testing against tuned.

	Robust04 - Titles			Robust04 - Descriptions		
1xRank	NDCG	MAP	R@1k	NDCG	MAP	R@1k
Tuned	0.660	0.376	0.837	0.687	0.401	0.855
Zero-Shot	0.660	0.374	0.836	0.688	0.401	0.846
2xRank						
Tuned	0.667	0.393	0.837	0.715	0.438	0.855
Zero-Shot	0.667	0.394	0.836	0.710	0.435	0.846
Adaptive						
Tuned	0.668	0.392	0.838	0.704	0.435	0.834
Zero-Shot	0.668	0.393	0.837	0.701	0.432	0.829

To assess LEE expansion in a zero-shot scenario, we use LEE parameters tuned on the CODEC dataset zero-shot on Robust04 titles and descriptions (the exact parameters can be found the released run metadata). Table 4 shows the “Tuned” LEE expansion model against the “Zero-shot” parameters for our unsupervided LEE query, two rounds of NLM re-ranking, and adaptive expansion. We observed no significant differences between the tuned and zero-shot LEE run under these different scenarios, and in some cases zero-shot is the same or marginally better on Robust04 titles. Therefore, this highlights that our proposed method of using NLM passage feedback and combining words and entities with dependencies is robust to its parameter across datasets.

Figure 4 shows the impact of lambda (i.e. relative word-to-entity weighting) on the effectiveness of LEE unsupervised query across our target datasets. Specifically, we see that for R@1000 and MAP, the best weighting is a combination of words and entities. For Robust04 datasets, MAP maximizes around 0.5, which weights word and entity expansions equally. However, for CODEC, precision is maximized around 0.1, favouring weighting entities and showing their precision benefits on domain-specific essay questions. On the other hand, Robust04 shows optimal recall with a relatively even combination of words and entities. However, unlike MAP, R@1000 for CODEC is maximized through a high weighting of words. Overall, this should show the precision-recall tradeoffs for different datasets and confirms that both words and entities are required for robust effectiveness.

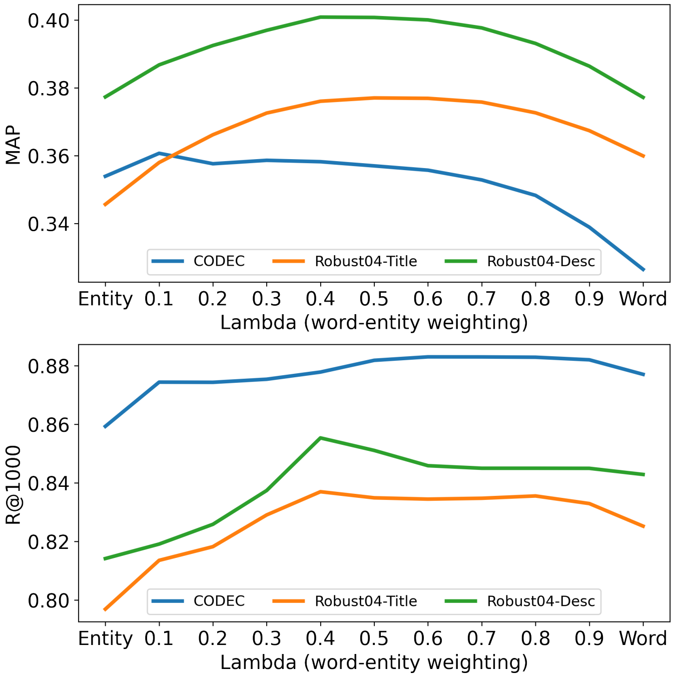


Fig. 4: Lambda (i.e. relative word-to-entity weighting) impact for LEE expansion on CODEC and Robust04 datasets.

References

1. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series p. 189 (2004)
2. Belkin, N.J., Oddy, R.N., Brooks, H.M.: Ask for information retrieval: Part i. background and theory. Journal of documentation (1982)
3. Bolotova, V., Blinov, V., Scholer, F., Croft, W.B., Sanderson, M.: A non-factoid question-answering taxonomy. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1196–1207 (2022)
4. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators (2020)
5. Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. In: Text REtrieval Conference (TREC). TREC (2021)
6. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the trec 2019 deep learning track. arXiv preprint arXiv:2003.07820 (2020)
7. Dai, Z., Callan, J.: Deeper text understanding for ir with contextual neural language modeling. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 985–988 (2019)
8. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 365–374 (2014)

9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423
10. Ferragina, P., Scaiella, U.: Fast and accurate annotation of short texts with wikipedia pages. IEEE software **29**(1), 70–75 (2011)
11. Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansion model for first stage ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2288–2292 (2021)
12. Gao, L., Callan, J.: Long document re-ranking with modular re-ranker (2022)
13. van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: Rel: An entity linker standing on the shoulders of giants. arXiv preprint arXiv:2006.01969 (2020)
14. Huston, S., Croft, W.B.: Parameters learned in the comparison of retrieval models using term dependencies. Ir, University of Massachusetts (2014)
15. Kadry, A., Dietz, L.: Open relation extraction for support passage retrieval: Merit and open issues. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1149–1152 (2017)
16. Kwok, K.L., Chan, M.: Improving two-stage ad-hoc retrieval for short queries. In: SIGIR 1998 (1998)
17. Li, B.Z., Min, S., Iyer, S., Mehdad, Y., Yih, W.t.: Efficient one-pass end-to-end entity linking for questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6433–6441 (2020)
18. Li, C., Yates, A., MacAvaney, S., He, B., Sun, Y.: Parade: Passage representation aggregation for document reranking. arXiv preprint arXiv:2008.09093 (2020)
19. Li, H., Mourad, A., Zhuang, S., Koopman, B., Zuccon, G.: Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. ArXiv **abs/2108.11044** (2021)
20. Li, H., Zhuang, S., Mourad, A., Ma, X., Lin, J., Zuccon, G.: Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In: European Conference on Information Retrieval. pp. 599–612. Springer (2022)
21. Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2356–2362 (2021)
22. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Springer Nature (2022)
23. Lin, S.C., Yang, J.H., Lin, J.: In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). pp. 163–173. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.repl4nlp-1.17, https://aclanthology.org/2021.repl4nlp-1.17
24. Liu, X., Chen, F., Fang, H., Wang, M.: Exploiting entity relationship for query expansion in enterprise search. Information Retrieval **17**(3), 265–294 (2014)
25. Liu, X., Fang, H.: Latent entity space: a novel retrieval approach for entity-bearing queries. Information Retrieval Journal **18**(6), 473–503 (2015)

26. MacAvaney, S., Macdonald, C., Ounis, I.: Streamlining evaluation with ir-measures. In: European Conference on Information Retrieval. pp. 305–310. Springer (2022)
27. MacAvaney, S., Tonellotto, N., Macdonald, C.: Adaptive re-ranking with a corpus graph. In: CIKM (2022)
28. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: Cedr: Contextualized embeddings for document ranking. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 1101–1104 (2019)
29. Mackie, I., Dalton, J., Yates, A.: How deep is your learning: The dl-hard annotated deep learning dataset. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2335–2341 (2021)
30. Mackie, I., Owoicho, P., Gemmell, C., Fischer, S., MacAvaney, S., Dalton, J.: Codec: Complex document and entity collection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2022)
31. Meij, E., Trieschnigg, D., De Rijke, M., Kraaij, W.: Conceptual language models for domain-specific retrieval. Information Processing & Management **46**(4), 448–469 (2010)
32. Metzler, D., Croft, W.B.: 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. pp. 472–479 (2005)
33. Metzler, D., Croft, W.B.: Latent concept expansion using markov random fields. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 311–318 (2007)
34. Mihalcea, R., Csomai, A.: Wikify! linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242 (2007)
35. Naseri, S., Dalton, J., Yates, A., Allan, J.: Ceqe: Contextualized embeddings for query expansion. In: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43. pp. 467–482. Springer (2021)
36. Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
37. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 708–718 (2020)
38. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., et al.: Kilt: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2523–2544 (2021)
39. Piccinno, F., Ferragina, P.: From tagme to wat: a new entity annotator. In: Proceedings of the first international workshop on Entity recognition & disambiguation. pp. 55–62 (2014)
40. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94. pp. 232–241. Springer (1994)
41. Rocchio, J.: Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing pp. 313–323 (1971)

42. Sciavolino, C., Zhong, Z., Lee, J., Chen, D.: Simple entity-centric questions challenge dense retrievers. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6138–6148 (2021)
43. Shehata, D., Arabzadeh, N., Clarke, C.L.: Early stage sparse retrieval with entity linking. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4464–4469 (2022)
44. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)
45. Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004). pp. 52–69. Gaithersburg, Maryland (2004)
46. Wang, X., Macdonald, C., Tonellotto, N., Ounis, I.: Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. ACM Transactions on the Web (2022)
47. Xiong, C., Callan, J.: Esdrank: Connecting query and documents through external semi-structured data. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management. pp. 951–960. CIKM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806456, http://doi.acm.org/10.1145/2806416.2806456
48. Xiong, C., Callan, J.: Query expansion with freebase. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. p. 111–120. ICTIR '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2808194.2809446, https://doi.org/10.1145/2808194.2809446
49. Xiong, C., Callan, J., Liu, T.Y.: Bag-of-entities representation for ranking. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. p. 181–184. ICTIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2970398.2970423, https://doi.org/10.1145/2970398.2970423
50. Xiong, C., Callan, J., Liu, T.Y.: Word-entity duet representations for document ranking. In: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval. pp. 763–772 (2017)
51. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 59–66. SIGIR '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1571941.1571954, https://doi.org/10.1145/1571941.1571954
52. Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying bert to document retrieval with birch. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 19–24 (2019)
53. Yu, H., Xiong, C., Callan, J.: Improving query representations for dense retrieval with pseudo relevance feedback. arXiv preprint arXiv:2108.13454 (2021)
54. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the tenth international conference on Information and knowledge management. pp. 403–410 (2001)