Local and Global Query Expansion for Hierarchical Complex Topics

Jeffrey Dalton[1], Shahrzad Naseri[2], Laura Dietz[3], and James Allan[2]

[1] School of Computing Science, University of Glasgow, Glasgow, UK
[2] Center for Intelligent Information Retrieval, College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, USA
[3] Department of Computer Science, University of New Hampshire, Durham NH, USA
jeff.dalton@glasgow.ac.uk
{shnaseri, allan}@cs.umass.edu
dietz@cs.unh.edu

Abstract. In this work we study local and global methods for query expansion for multifaceted complex topics. We study word-based and entity-based expansion methods and extend these approaches to complex topics using fine-grained expansion on different elements of the hierarchical query structure. For a source of hierarchical complex topics we use the TREC Complex Answer Retrieval (CAR) benchmark data collection. We find that leveraging the hierarchical topic structure is needed for both local and global expansion methods to be effective. Further, the results show that entity-based expansion methods show significant gains over word-based models alone, with local feedback providing the largest improvement. The results on the CAR paragraph retrieval task demonstrate that expansion models that incorporate both the hierarchical query structure and entity-based expansion result in a greater than 20% improvement over word-based expansion approaches.

1 Introduction

Current web search engines incorporate question answer (QA) results for a significant fraction of queries. These QA results are a mixture of factoid questions ["Who won the James Beard Award for best new chef 2018?"] that can be answered from web results or from entity-based knowledge bases. However, many questions require more than short fact-like responses. In particular, topics like ["What are the causes of the Civil War?"] require multifaceted essay-like responses that span a rich variety of subtopics with hierarchical structure: Geography and demographics, States' rights, The rise of abolitionism, Historical tensions and compromises, and others. These 'complex' and multifaceted topics differ significantly from simple factoid QA information needs.

Hierarchical complex topics have rich structure that could (and should) be leveraged for effective retrieval. The first type of structure is the hierarchical nature of the topics. They start with a root topic and contain more specific subtopics in a hierarchy. We propose fine-grained methods that perform expansion both at the overarching level as well as for each of the subtopics individually.

One of the fundamental issues is the well-studied problem of vocabulary mismatch. Retrieval from a paragraph collection exacerbates this problem because the content retrieved is short passages that may be taken out of context. In particular, the discourse nature of language in related series of paragraphs on the same topic makes extensive use of language variety and abbreviations. Neural ranking models address the vocabulary mismatch problem by learning dense word embedding representations [15],[11], [14]. In this work we use learned 'global' embedding models that use joint embeddings of both words and entities. We also experiment with 'local' models derived from pseudo-relevance feedback expansion approaches.

Our proposed expansion methods incorporate expansion features that can be used in a first-pass retrieval model. The state-of-the-art neural network models [15],[11] for complex topics re-rank a candidate set of paragraphs retrieved from a simple and fast first pass baseline model, most commonly BM25. However, these simple models often fail to return relevant paragraphs in the candidate pool. The consequence is that multi-pass reranking models are limited by the (poor) effectiveness of the underlying first-pass retrieval. As a result, Nanni et al. [15] find that the improvement over the non-neural baseline models is marginal for retrieving paragraphs for complex topics. Our proposed expansion methods address this fundamental problem. In contrast to neural approaches, we perform expansion natively in the search system and combine the features using linear Learning-to-Rank (LTR) methods.

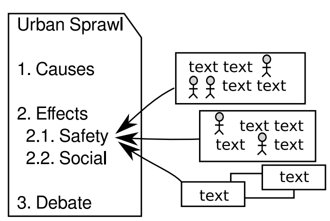

Fig. 1. Example of a complex topic from the TREC Complex Answer Retrieval track

Complex multifaceted queries (topics) of this nature are currently being studied in the TREC Complex Answer Retrieval (CAR) track. In TREC CAR, the task is **given a complex topic composed of different subtopics - a skeleton of a Wikipedia article, the system should retrieve: 1) relevant paragraphs and 2) relevant entities for each subtopic.** The first CAR task is passage (paragraph) retrieval to find relevant paragraphs for each subtopic. The second task is entity retrieval, to identify important people, locations, and other concepts that should be mentioned in the synthesis. An example of a complex topic is given in Figure 1.

For TREC CAR, the nature of complex topics is particularly entity-centric because the subtopics are based around entities from a knowledge base. Further, recent test collections based on Wikipedia paragraphs include rich text and entity representations [7]. As a result, this topic collection is an interesting domain for models that incorporate both text and entity representations of queries and documents [4], [23],[24].

In this work we make several contributions to methods and understanding for expansion in complex, multifaceted, and hierarchical queries:

- We develop entity-aware query expansion methods for passage retrieval. We use probabilistic retrieval approaches and entity embedding vectors with entity-aware indicators including entity identifiers, entity aliases, and words. Entity-aware models for different levels of the topics are combined with an LTR approach.
- The experimental evaluation demonstrates that our entity-aware approach outperforms a learned combination of probabilistic word-based models by 20%. It further outperforms the best performing approach from the TREC CAR year one evaluation.

The remainder of the paper is structured as follows. First, we provide background and related work on TREC CAR as well as the broader area of entity-focused query expansion. Next, we introduce the existing and newly proposed expansion model for complex hierarchical topics. Finally, we perform an empirical study evaluation on the TREC CAR paragraph retrieval task evaluating the effectiveness of a variety of local and global expansion methods.

2 Background and Related Work

Question answering Although they may not be not strictly formulated as questions, retrieval for complex topics is related to approaches that answer questions from web content. Retrieval techniques for effective question answering is undergoing a resurgence of research, with a particular interest in non-factoid QA [2], [3]. These works are similar to complex answer retrieval in that they perform question answering by retrieving relevant passages, in particular paragraphs from Wikipedia. The key difference with the current work is that their topics are a single question with one answer. In contrast, the complex topic retrieval addressed in this work focuses on comprehensive complex answers with topics that have explicit multifaceted hierarchical relationships.

TREC CAR The TREC Complex Answer Retrieval track was introduced in 2017 to address retrieval for complex topics. For a survey of approaches, see Nanni et al. [15] as well as the overview [7]. Nanni et al. evaluate a variety of models, including a leading neural ranker (Duet model) [14]. They find that while the neural network gives the best performance, the gains over leading retrieval approaches are only modest. Another neural model, PACRR, by MacAvaney at al. [11] is consistently shown to improve effectiveness on CAR, we use this as one of our baseline methods. In all cases in the 2017 evaluation, BM25 is used to create candidate sets for reranking. However, BM25 fundamentally limits the effectiveness that reranking runs by constraining the candidate pool. In this work, we study methods for expansion that may be used for feature-based reranking and that incorporate entity-based representations.

Structured queries Complex queries in retrieval is not a new problem. In fact, some of the earliest uses of retrieval focused on boolean retrieval. Users constructed complex boolean expressions with complex subqueries [20]. This was later followed up with more complex query capability [21]. Follow-up query languages that support rich query expressions include: INQUERY, Lucene, Terrier, and Galago. However, these languages are usually only used internally to rewrite simple keyword queries, possibly using some inferred structure from natural language processing. In contrast, CAR query topics contain explicit multifaceted hierarchical structure. We test various ways of using this structure in expansion models.

Relevance Feedback Expansion Models One of the fundamental challenges in retrieval is vocabulary mismatch and one of the primary mechanisms to address this problem is relevance feedback that takes a user judgment of a document and uses this to build an updated query model. Pseudo-relevance feedback (PRF) [9], [1] approaches perform this task automatically, assuming the top documents are relevant. We build on previous work that uses mixtures of relevance models [12], but apply it to creating fine-grained expansions from complex hierarchical topic headings. Further, as the results in this work demonstrate, PRF is most effective when there is a high density of relevant documents in top ranked results. In contrast for CAR, there are few relevant documents that are often not retrieved in first-pass retrieval. To overcome this issue we propose using a fine-grained score-based fusion approach and we utilize entity-based expansion features. The results demonstrate that our approaches using external entity-based features is more robust than word-based approaches.

Embedding-based Expansion Models Another approach to overcome the word mismatch problem is using global collection word embeddings. Word embedding techniques learn a low-dimensional vector (compared to the vocabulary size) for each vocabulary term in which the similarity between the word vectors captures the semantic as well as the syntactic similarities between the corresponding words. Word embeddings are unsupervised learning methods since they only need raw text data without other explicit labels. Xiong et al. propose a model for ad-hoc document retrieval that represents documents in queries in both text and entity spaces, leveraging entity embeddings in their approach [23]. In this work we use joint entity-word embedding models to perform global term expansion.

Knowledge-base Expansion Models Recent previous work demonstrates that query expansion using external knowledge sources and entity annotations can lead to significant improvements to a variety of retrieval tasks [4], including entity linking of queries [8], and using entity-derived language models for document representation [18]. There is also recent work on determining the salience of entities in documents [24] for ranking. Beyond salience, research focused on identifying latent entities [10], [22] and connecting the query-document vocabularies in a latent space. We build on these entity-centric representations and utilize entity query annotations, explicit entity links, and related entities from entity feedback and entity embedding models. We study the differences between these different elements for the CAR task. For an overview of work in this area we refer the reader to [6].

3 Methodology

3.1 Complex Hierarchical Topics

A complex topic T consists of heading nodes constructed in a hierarchical topic tree, an example is shown in Figure 1. Each heading node, h, represents the subtopic elements. For example, a complex topic with subtopics delimited by a slash would be: "Urban sprawl/Effects/Increased infrastructure and transportation cost". This consists of three heading nodes - the leaf heading is "Increased infrastructure and transportation cost" with the root heading "Urban sprawl" and intermediate heading "Effects". The tree structure provides information about the hierarchical relationship between subtopics. In particular, the most important relationship is that the root heading is the main focus of the overall topic.

Given a complex topic tree T, the outline consists of a representation for each of the subtopic heading nodes $h \in H$. At the basic level, each heading contains its word representation from text, $W : \{w_1, ..., w_k\}$, a sequence of words in the subtopic. Beyond words, each heading can also be represented by features extracted by information extraction and natural language processing techniques, for example part of speech tags and simple dependence relationships.

In particular, we hypothesize that another key element of effective retrieval with complex topics going beyond words to include entities and entity relationships. Therefore, we propose representing the topic as well as documents with entity mentions, T_M and D_M respectively, where each has $M : \{m_1, ..., m_k\}$ with m_k a mention of an entity e in a knowledge base. Given an entity-centric corpus and task along with rich structure, the mix of word and entity representation offers significant potential for retrieval with complex topics. The result is sequence of ordered entities within a heading with provenance connecting the entity annotations to free text. In TREC CAR as well as adhoc document retrieval, this representation is (partially) latent - it must inferred from the topic text.

3.2 Topic Expansion Model

In this work, we study use of different expansion methods over diverse types of representations, based on words and entities. To specify the representations we use different term vocabularies, $v \in V$, for example:

- Words, $W : \{w_1, ..., w_k\}$ are the unigram words from the collection vocabulary.

– Entities, $E : \{e_1, ..., e_k\}$ are entities from a knowledge base, matched based on their entity identifiers.

Note that entities may have multiple vocabularies that interact with one another. We can match entities to word representations using the entity names and aliases $A : \{a_1, ..., a_k\}$ derived from their Wikipedia name, anchor text, redirects, and disambiguation pages.

For expansion, we study two approaches: the expansion based on local query-specific relevance feedback and the global word-entity embedding similarity. We elaborate more about these approached in section 3.3 and 3.4, respectively.

To perform effective expansion, our goal is to estimate the probability of relevance for an entry in the vocabulary with respect to the complex topic, T. In other words, regardless of the underlying expansion method, the overarching goal is to identify the latent representation of the topic across all vocabulary dimensions: $p(V|T)$. However, a single expansion model for an entire complex topic is unlikely to be effective. For both expansion methods we also build a mixture of fine-grained expansions for each subtopic node that are combined. For every type in the vocabulary V, and for every heading node $h \in H$, we create a feature, $f(h, D)$.

In table 1 we illustrate different approaches for expansion that include three dimensions of the expansion: the expansion method, the representation type, and which subtopic to expand. An example is, [Antibiotic use in livestock/Use in different livestock/In swine production]. In this case, $R =$ [Antibiotic use in livestock] is the root , $I =$ [Use in different livestock] is an intermediate node, and $H =$ [In swine production] is the leaf heading. We vary the topic representation using differing combinations of these tree elements. The most common approach by participants in TREC CAR is to simply concatenate the RIH context into one query and to ignore the heading relationships or boundaries. In contrast, our fine-grained method preserves these elements and handles them separately.

We use a simple and effective method for combining heading evidence up to the topic-level. Features are combined using a log-linear model with parameters, θ. The number of these features is limited to approximately 10. This scale allows it to be learned efficiently using coordinate ascent to directly optimize the target retrieval metric. All of the score-level features, both heading derived and feedback, correspond to queries that can be expressed natively in the first pass matching phase of a search system.

3.3 Relevance Model Expansion

Lavrenko and Croft introduce relevance modeling, an approach to query expansion that derives a probabilistic model of term importance from documents that receive high scores, given the initial query [9]. In our model, we derive a distribution over all types of the vocabulary. In this case, $p(D = d|T)$ is the relevance of the document to the topic, derived from score for the document under the query model. The $p(V|d)$ is the probability of the vocabulary under the language model of the document using that representation.

Table 1. Examples of topic expansion features across word and entity vocabularies. All features are for R, I, and H nodes separately. The example topic is: [Antibiotic use in livestock/Use in different livestock/In swine production]. The entities identified in the topic are: [Antibiotics, Livestock/ Livestock/ Domestic pig, Pig farming]

Name	Description	Feature Example
RIH-QL	Representing words from the root, intermediate, and leaf subtopics	(antibiotic use livestock different swine production)
RIH-IDs-Embed	Representing expanded entities from global embeddings from the root, intermediate, and leaf subtopics using their IDs	Antibiotics → Tetracycline.id Livestock → Cattle.id Pig farming → (Animal husbandry).id
H-Names-Embed	Expansion of entity names within the leaf subtopic using global embeddings	Pig farming → (animal husbandry dairy farming poultry ubre blanca)
R-Aliases-Embed	Expansion of aliases of entity within the root subtopic using global embeddings	Tetracycline → (tetracyn sumycin hydrochloride) Cattle → (cow bull calf bovine heifer steer moo)

3.4 Embedding-based Expansion

In this section, we first elaborate how we learn the global embeddings. We then explain how we use the learned model for expanding complex queries.

Joint Entity-Word Embeddings Motivated by the vocabulary mismatch problem, we learn a joint entity-word embedding following the approach presented by Ni et al. [16]. We learn a low dimensional vector representation for entities and words based on the Mikolov Skip-gram model [13] using term co-occurrence information within a text. Each entity mention is considered as a single "term". The Skip-gram model aims to maximize the probability of current term based on its surrounding terms using a neural network. We thus model entities using their word context (and vice versa).

The following excerpt shows the transformation of text with entity mentions using special placeholders for each entity mention:

> `The_World_Health_Organization_(WHO)` is a `specialized_agency_of_the_United_Nations` that is concerned with international `public_health`. It was established on 7 April 1948, and is headquartered in `Geneva`, `Switzerland`.

We build a mixture of fine-grained expansions for each subtopic in a complex topic. We compute embedding-based similarity for both explicit entity mentions as well as words, two types from in the vocabulary. For the global similarity between dense embedding vectors we use the cosine similarity. In addition to expanding each subtopic node individually, we also perform expansion of the complete topic tree as a whole. The embedding vector of node (or entire query tree) is represented as the average (mean) of the embedding vector of each element within it.

4 Experimental Setup

4.1 Data

The primary dataset used for experiments is from the TREC Complex Answer Retrieval (CAR) track, v2.1 [5], released for the 2018 TREC evaluation. The CAR data is derived from a dump of Wikipedia from December 2016. There are 29,678,367 paragraphs in the V2 paragraph collection.

Each outline consists of the hierarchical skeleton of a Wikipedia article and its subtopics. Each individual heading is a complex topic for which relevant content (paragraphs) needs to be retrieved. In 2017, the test topics are chosen from articles on open information needs, i.e., not people, not organizations, not events, etc. The benchmark consists of 250 topics, split equally (roughly) into train and test sets.

The TREC CAR setup includes two types of heading-level judgments, *automatic* and *manual*. The automatic (binary) judgments are derived directly from Wikipedia and the manual judgments are created by NIST assessors. A key outcome from 2017 was that the automatic benchmark data is useful for differentiating between systems and not subject to the pooling bias in manual judgments (it's also much larger) [7]. In this work, we use the automatic judgment to evaluate our methods because the original retrieval methods were not in the pool and we found that the manual judgments had a high degree of unjudged results even for the baselines.

Knowledge base For the experiments here we use the non-benchmark articles from Wikipedia as a knowledge base. These include the full article text, including the heading structure. It does not include the infobox and other data that was excluded in the CAR pre-processing. In addition to the text, we use anchor text, redirects, and disambiguation metadata derived from the article collection and provided in the data.

Evaluation measures We use the standard measures reported in TREC CAR evaluations. The primary evaluation measure is Mean Average Precision (MAP). We report R-Precision, because the number of relevant documents in TREC CAR varies widely across topics. The NDCG@1000 metric is included following standard practice in the track. For statistical significance, we use a paired t-test and report significance at the 95% confidence interval.

4.2 System details

In this section we provide additional details of the systems used for the implementation. The TREC CAR paragraph collection is indexed using the Galago [4] retrieval system, an open-source research system. The query models and feedback expansion models are all implemented using the Galago query language. The paragraphs are indexed with the link fields to allow exact and partial matches of entity links in the paragraphs. Stopword removal is performed on the heading queries using the 418 INQUERY stop word list. Stemming is performed using the built-in Krovetz stemmer.

In our score fusion model we use a log-linear model combination of different features for ranking. The model parameters, θ are optimized using coordinate ascent to directly optimize the target retrieval measure, Mean Average precision (MAP). The implementation of the model is available in the open-source RankLib learning-to-rank library.

Parameter settings For the experiments we use the provided train / test topic splits. We tune the retrieval hyper-parameters on the training data using grid search. For the Sequential Dependence Model (SDM) baseline parameters are $mu = 1200$, $uww = 0.02$, $odw = 0.10$, and $uniw = 0.82$. We observe that these parameters differ from the default settings which are optimized for short adhoc TREC queries and longer newswire documents. In contrast, the paragraph content in the CAR collection are much shorter. For relevance feedback, we use the SDM model as the baseline retrieval. The expansion parameters are tuned similarly and we find that 10 expansion documents with 20 feedback terms and an interpolation weight of 0.8 is most effective.

Query Entity Annotation The topics in TREC CAR do not have explicit entity links. To support matching paragraph entity documents, we annotate the complex topic headings with entities. Entity linking is performed on each heading for both the train and test benchmark collections. We use the open-source state-of-the-art SMAPH entity linker [5]. Although not the main focus of the paper, we observe that the entity linker suffers from significant recall issues, missing a large fraction of the entities in the complex topic headings, which are directly derived from Wikipedia entity titles. As a result, the utility of explicit query entity links is lower than we expected.

Document entity annotations For entity mentions in documents we use the existing entity links provided in Wikipedia. We note that the entity links in Wikipedia are sparse and biased. By convention only the first mention of an entity in an article is annotated with a link. This biases retrieval based on entity identifiers towards paragraphs that occur early in a Wikipedia article. An area for future work is to perform entity annotation on the documents to improve mention recall. For example one known issue in the current setup is that many mentions that use abbreviations are not currently linked, thereby limiting the effectiveness of entity link approaches.

[4] http://www.lemurproject.org/galago.php
[5] https://github.com/marcocor/smaph

Table 2. Text-based baselines and expansion methods. * indicates significance over the RH-SDM run.

Model	MAP	R-Prec	NDCG
RIH-QL	0.110	0.088	0.228
RH-SDM	0.132	0.109	0.248
RH-SDM-RM3	0.127	0.102	0.243
L2R-SDM-RM3	0.142*	0.107	0.257*
Embedding-Term	0.143*	0.119*	0.261*
GUIR (neural)	0.137	0.112	0.237
GUIR-Exp (neural)	0.142*	0.117	0.242

Table 3. Baseline: Combinations of SDM and RM3 over different outline levels combined with L2R. Learned feature combination weights displayed.

Model	Weight
RIH-QL	0.288
R-SDM	0.153
H-SDM	0.340
RH-SDM	0.108
RH-SDM-RM3	0.110

Learning Embeddings The joint entity-word embeddings are learned from the DBpedia 2016-10 full article dump. To learn the entity embeddings we use the Word2Vec implementation in gensim [19] version 3.4.0 with parameters as follow: window-size = 10, sub-sampling = 1e-3, and cutoff min-count = 0. The learned embedding dimension is equal to 200 and we learned embeddings of 3.0M entities out of 4.8M entities available in Wikipedia.

5 Results

In this section we present our main experimental results. We start with proven word-based retrieval and expansion methods. This includes state-of-the-art neural baselines. We then build on these methods and experiment with local and global entity-based expansion.

5.1 Word-based retrieval and expansion

We first evaluate standard text retrieval methods for heading retrieval. The results are shown in Table 2. The baseline model, RIH-QL, is a standard bag-of-words query-likelihood model [17] on all terms in the topic. All other runs are statistically significant gains over this simple baseline. The table also shows results for an Sequential Dependence Model (SDM) that uses the root and leaf subtopics of the heading. We also experimented with other variations (H-QL, RIH-SDM, RH-QL, etc...), but these are all outperformed by RH-SDM. RH-SDM was the best performing unsupervised model for this collection in TREC 2018. We also evaluate using a relevance model term-based expansion on top of the best SDM run. We find that the RM3 performance is insignificantly worse than the SDM baseline, demonstrating the PRF based on words is challenging in this environment. We attribute this to the sparseness of relevant paragraphs to the topics, an average of 4.3 paragraphs per topic, with baselines retrieving on average about half of those, 2.2.

We experimented with combining the baseline systems with additional fine-grained SDM components from each part of the query (subtopic) separately and weighting and combining them into a linear model, the L2R-SDM-RM3 method. The features and learned weights are given in Table 3. We observe that the H-SDM feature is the most important, putting greater emphasis on the leaf subtopic (approximately 2x the root topic). Combining these baseline retrieval and subtopic heading components results in significant gains over all the models individually, including RH-SDM. The Embedding-Term method is L2R-SDM-RM3 with addition of global word expansion. The results show a small, but insignificant improvement to the model effectiveness.

The bottom of Table 2 shows a comparison with one of the leading neural ranking models from the Georgetown University IR group (GUIR). It uses the PACRR neural ranking architecture modified with heading independence and heading frequency context vectors [11]. The second row (Exp) adds expansion words of the topic's query terms. Interestingly, the learned GUIR neural run does do not improve significantly over the RH-SDM baseline, the SDM model even slightly outperforms it on NDCG. The learned word-based expansion methods L2R-SDM-RM3 and Embedding-Term are both statistically significant over the GUIR base run for MAP, but not statistically significantly different from the Exp run. This indicates that our methods are comparable to state-of-the-art word-based expansion models using deep learning for this collection.

Table 4. Entity-based expansion with varying latent entity models. * indicates significance over the L2R-SDM-RM3 Baseline.

Model	**MAP**	**R-Prec**	**NDCG**
L2R-SDM-RM3 Baseline	0.142	0.107	0.257
Entity_Embedding	0.154	0.127	0.277
Entity_Retrieval	0.160*	0.133	0.284*
Entity_Collection_PRF	0.172*	0.146*	0.297*

5.2 Entity expansion

In this section we study combining the previous word-based representations with entity representations. We use entities annotated in the query as well as inferred entities from local and global sources: global embeddings, local entity retrieval, and local pseudo-relevance feedback on the paragraph collection. Each of the entity expansion models is a learned combination of subtopic expansions across the different entity vocabularies (identifiers, names, aliases, and unigram entity language models).

The results are shown in Table 4. The baseline method is L2R-SDM-RM3, the learning to rank combination of all word-based expansion features. Each entity model adds additional entity features to this baseline. The results show that adding entity-based features improves effectiveness consistently across all entity inference methods. There are benefits to using global entity embeddings, but they are not significant over the baseline. The local retrieval and collection PRF expansion models both result in significant improvements over the baseline. In particular, the collection entity representation shows the largest effectiveness gains. Additionally, all of the entity-based expansion methods show statistically significant improvements over the GUIR-Exp word-based expansion run.

We find that all entity-expansion methods consistently improve the results. When compared with the baseline word model they have a win-loss ratio varying from 2.6 up to 4.6. The best method based on collection feedback has 281 losses, 1300 wins, with a win-loss ratio of 4.6. In contrast, the win-loss ratio for the GUIR-Exp model is 1.1, hurting almost as many queries as it helps. Consequently, we conclude that entity-based expansion methods more consistently improve effectiveness for complex topics when compared with word-based expansion methods.

6 Conclusion

In this work we study local and global expansion methods that utilize word-based and entity-based features for retrieval with hierarchical semi-structured queries. We propose a method that performs a mixture of fine-grained (subtopic level) feedback models for each element of the structured query and combines them using score-based fusion. On the TREC CAR paragraph ranking task, we demonstrate that entity-centric subtopic-level expansion models constitute the most effective methods - even outperforming established neural ranking methods. Further, the entity-based expansion results show significant and consistent effectiveness gains over the word-based expansion methods, resulting in a greater than 20% improvement in mean average precision.

The new proposed expansion methods build on proven probabilistic expansion methods and combine multiple feature representations to create more robust retrieval for complex topics. As search evolves to support more complex tasks the nature of complex topics will continue to develop. We envision more complex topic structures that will grow in size. This work presents an important first step in leveraging structure effectively. We anticipate that additional modeling of the complex hierarchical relationships across diverse vocabularies (words, entities, etc...) will lead to further improvements in the future.

7 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1617408. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

1. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 243–250. SIGIR '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1390334.1390377
2. Cohen, D., Croft, W.B.: End to end long short term memory networks for Non-Factoid question answering. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. pp. 143–146. ACM (Sep 2016)
3. Cohen, D., Yang, L., Croft, W.B.: WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018. pp. 1165–1168 (2018). https://doi.org/10.1145/3209978.3210118, https://doi.org/10.1145/3209978.3210118
4. 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. SIGIR '14, ACM, New York, NY, USA (2014)
5. Dietz, L., Gamari, B., Dalton, J.: TREC CAR 2.1: A data set for complex answer retrieval" (2018), http://trec-car.cs.unh.edu
6. Dietz, L., Kotov, A., Meij, E.: Utilizing knowledge graphs for text-centric information retrieval. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1387–1390. ACM (2018)
7. Dietz, L., Verma, M., Radlinski, F., Craswell, N.: Trec complex answer retrieval overview. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (2017), https://trec.nist.gov/pubs/trec26/papers/Overview-CAR.pdf
8. Hasibi, F., Balog, K., Bratsberg, S.E.: Exploiting entity linking in queries for entity retrieval. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. pp. 209–218. ICTIR '16, ACM, New York, NY, USA (2016)
9. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 120–127. SIGIR '01, ACM, New York, NY, USA (2001). https://doi.org/10.1145/383952.383972
10. Liu, X., Fang, H.: Latent entity space: a novel retrieval approach for entity-bearing queries. Inf. Retr. Journal **18**(6), 473–503 (2015), https://doi.org/10.1007/s10791-015-9267-x
11. MacAvaney, S., Yates, A., Cohan, A., Soldaini, L., Hui, K., Goharian, N., Frieder, O.: Characterizing question facets for complex answer retrieval (May 2018)
12. Metzler, D., Diaz, F., Strohman, T., Croft, W.B.: Umass robust 2005: Using mixtures of relevance models for query expansion. In: Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005 (2005)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

14. Mitra, B., Diaz, F., Craswell, N.: Learning to match using local and distributed representations of text for web search. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1291–1299. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017)
15. Nanni, F., Mitra, B., Magnusson, M., Dietz, L.: Benchmark for complex answer retrieval. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. pp. 293–296. ICTIR '17, ACM, New York, NY, USA (2017)
16. Ni, Y., Xu, Q.K., Cao, F., Mass, Y., Sheinwald, D., Zhu, H.J., Cao, S.S.: Semantic documents relatedness using concept graph representation. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 635–644. ACM (2016)
17. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 275–281. SIGIR '98, ACM, New York, NY, USA (1998)
18. Raviv, H., Kurland, O., Carmel, D.: Document retrieval using Entity-Based language models. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 65–74. SIGIR '16, ACM, New York, NY, USA (2016)
19. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
20. Salton, G., Fox, E.A., Wu, H.: Extended boolean information retrieval. Commun. ACM **26**(11), 1022–1036 (1983)
21. Turtle, H., Croft, W.B.: Evaluation of an inference network-based retrieval model. ACM Trans. Inf. Syst. Secur. **9**(3), 187–222 (Jul 1991)
22. Xiong, C., Callan, J.: EsdRank: Connecting query and documents through external Semi-Structured data. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 951–960. CIKM '15, ACM, New York, NY, USA (2015)
23. 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. ACM (2017)
24. Xiong, C., Liu, Z., Callan, J., Liu, T.Y.: Towards better text understanding and retrieval through kernel entity salience modeling. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 575–584. ACM (June 2018)