Abstract

Knowledge workers (such as healthcare information professionals, patent agents and recruitment professionals) undertake work tasks where search forms a core part of their duties. In these instances, the search task is often complex and time-consuming and requires specialist expert knowledge to formulate accurate search strategies. Interactive features such as query expansion can play a key role in supporting these tasks. However, generating query suggestions within a professional search context requires that consideration be given to the specialist, structured nature of the search strategies they employ. In this paper, we investigate a variety of query expansion methods applied to a collection of Boolean search strategies used in a variety of real-world professional search tasks. The results demonstrate the utility of context-free distributional language models and the value of using linguistic cues to optimise the balance between precision and recall.

Introduction

Many knowledge workers rely on the effective use of search applications in the course of their professional duties (Verberne et al., 2019). For example, healthcare information professionals perform systematic reviews of published literature sources as the foundation of evidence-based medicine (Russell-Rose and Chamberlain, 2017). Likewise, patent agents rely on prior art search as the foundation of their due diligence process (Lupu et al., 2011). Similarly, recruitment professionals use Boolean search as the foundation of the candidate sourcing process (Russell-Rose and Chamberlain, 2016a).

However, systematic literature reviews can take years to complete (Bastian et al., 2010), and new research findings may be published in the interim, leading to a lack of currency and potential for inaccuracy (Shojania et al., 2007). Likewise, patent infringement suits have been filed at a rate of more than 10 a day due to the later discovery of prior art which their original search missed (Gibbs, 2006). And recruitment professionals report that finding candidates with appropriate skills and experience continues to be their primary concern (Russell-Rose and Chamberlain, 2016b).

The traditional solution to structured search problems is to use form-based query builders such as that shown in Figure 1. The output of these tools is typically a series of Boolean expressions consisting of keywords, operators and ontology terms, which are combined to form a multi-line artefact known as a *search strategy* (Figure 2).

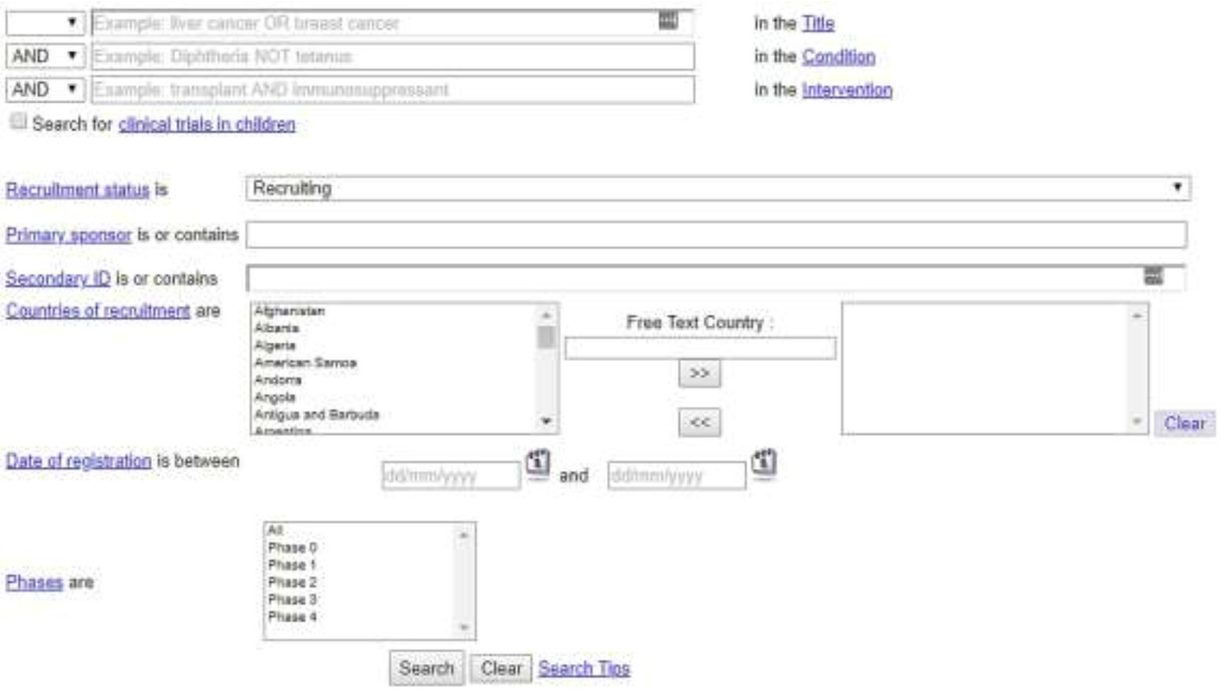

Figure 1. The World Health Organisation's clinical trials search portal.

```
1   A01N0025-004/CPC
2   RODENT OR RAT OR RATS OR MOUSE OR MICE
3   BAIT OR POISON
4   2 AND 3
5   1 OR 4
6   AVERSIVE OR ADVERSIVE OR DETER? OR REPEL?
7   NONTARGET OR (NON WITH TARGET) OR HUMAN OR DOMESTIC OR PET
OR DOG OR CAT
8   6 AND 7
9   8 AND 5
10   BITREX OR DENATONIUM OR BITREXENE OR BITTERANT OR BITTER
11   10 AND 5
12   9 OR 11
```

Figure 2. An example patent search strategy.

In this paper, we review the role of query expansion within the context of professional structured search applications. We investigate a number of techniques for generating interactive query suggestions, and evaluate them using a variety of real-world data.

Background

Professional search

The term ‘professional search’ refers to search for information in a work context which often involves complex information needs, the use of multiple repositories and the incorporation of domain-specific taxonomies or vocabularies (Verberne et al., 2018). Various authors have provided descriptive and behavioural definitions of the term (see (Russell-Rose et al., 2018) for an overview). One of the earliest definitions was proposed by Koster et al. (Koster et al., 2009), whereby professional search:

- Is performed by a professional for financial compensation;
- Is within a particular domain and/or area of expertise;
- Has a specified brief, which is typically well defined but complex;
- Has a high value outcome where the results will reduce risk, provide assurances, etc.;
- Has budgetary constraints such as time and money.

A key distinction between professional search tasks and other kinds of search tasks, such as casual search (Elsweiler et al., 2012) and web search[1] (Broder, 2002) is that the latter:

- Are typically performed on a discretionary basis;
- Are not necessarily performed by an expert searcher or domain expert;
- And do not place at stake the professional reputation of the searcher.

Query expansion

Given the complexity of professional search tasks and their reliance on specialist terminology, query expansion offers a natural approach to assist the searcher (Liu et al., 2011). Query expansion is the process of reformulating or augmenting a user’s query in order to increase its effectiveness (Manning et al., 2008).

The primary methods for query expansion are referred to as either *local* (based on documents retrieved by the query) or *global* (using resources independent of the query). Selection of suggested expansion terms can be either *automated* (applied without explicit user interaction) or *interactive* (guided by the user).

Global methods involve the use of resources such as thesauri, controlled vocabularies or ontologies to identify related terms in the form of synonyms, hypernyms, hyponyms, etc. (Aggarwal and Buitelaar, 2012). Such resources may be either *manually curated* or generated from text corpora using *distributional methods*. Automated global methods can increase recall significantly but may also reduce precision by adding irrelevant or out-of-domain terms to the query (Manning et al., 2008).

ologies are more useful for query expansion when they are specific to the task domain. Generic ources such as WordNet are considered less useful and may not distinguish class concepts from

instances (Bhogal et al., 2007). Some ontologies offer an additional source of related terms in the form of words occurring in the term definitions (Navigli and Velardi, 2003). In the biomedical domain, expanding queries with related MeSH terms has been shown to be useful (Rivas et al., 2014), while adding synonyms from the more comprehensive UMLS has been found to improve recall (Griffon et al., 2012), at the expense of precision (Zeng et al., 2012).

The development of efficient distributional methods has revolutionised natural language processing techniques for finding related terms (Collobert et al., 2011; Mikolov et al., 2013a). Consequently, a number of researchers have considered the utility of *word embeddings* for query expansion. Kuzi (Kuzi et al., 2016), Roy (Roy et al., 2016) and Diaz (Diaz et al., 2016) all used local embeddings trained on TREC corpora, with differing results. While Kuzi (Kuzi et al., 2016) found that local word embeddings outperformed the standard RM3 relevance model, Roy (Roy et al., 2016) found the opposite. More recently, we have seen that contextual embeddings, such as those based on BERT, have transformed the state of the art not only in natural language processing (Devlin et al., 2019) but also in information retrieval (Lin, 2019; Mitra and Craswell, 2018). Given the nature of our investigation where we expand query terms on an individual basis, we focus on context-free embeddings.

A fundamental problem with most query expansion techniques is that queries may be harmed as well as improved (Xiong and Callan, 2015). In addition, with fully automated techniques the user may be unable to control how the expansion terms are applied. We address these issues by treating query expansion as a *recommendation* task, i.e. given a query term entered by the user, can we recommend further relevant terms. Framing the task in this way is significant, since the use of an interactive approach allows the user to exercise a more informed judgement regarding both term selection and application within a structured search strategy.

Application context

Query suggestions are a common feature of many web search engines, and have served as the focus of many research studies e.g. (Tahery and Farzi, 2020). Since search queries on the web typically consist of short sequences of keywords with little or no linguistic structure (Kumar et al., 2020), term suggestions can offer immediate value as either an addition to the current query or as a wholesale replacement (Kruschwitz et al., 2013).

Although there have been studies investigating query expansion within a professional search context, e.g. Verberne et al (Verberne et al., 2016), examples of commercial systems in production are relatively rare. This may be due in part to the challenges presented by the structured nature of the queries themselves. For example, when sourcing candidates for a client brief, recruiters might use a structured query such as that shown in Figure 3.

```
Java AND (Design OR develop OR code OR
```

```
Program) AND ("* Engineer" OR MTS OR "*
Develop*" OR Scientist OR technologist) AND
(J2EE OR Struts OR Spring) AND (Algorithm OR
"Data Structure" OR PS OR "Problem Solving")
```

Figure 3. An example recruitment search strategy.

For a query such as this, it is not sufficient simply to offer suggested terms as additions or as wholesale replacements. Instead, term suggestions must be both relevant and specific to the individual subexpressions it contains. In the above example, query suggestions relevant to the first subexpression would be quite inappropriate for the second subexpression.

We have therefore structured our investigation using an approach based on previous query suggestion studies (Albakour et al., 2011), in which existing, human-generated resources are treated as a 'gold standard'. In our case, a gold standard exists in the form of published search strategies. In this context, the evaluation process measures the extent to which terms found in those strategies can be predicted.[2] For example, given the term *rodent* in line 2 of the strategy of Figure 2, we measure the extent to which the related terms *rat*, *rats*, *mouse*, and *mice* can be predicted. This particular example contains five such disjunctions (lines 2, 3, 6, 7 and 10), so it offers five opportunities for evaluation. Moreover, since we use publicly available sources our experiments can be more easily replicated by others.[3]

Arguably, an ideal test collection for such an evaluation would contain search strategies curated specifically for the purpose. However, an ideal test collection should also include:

- Search strategies from more than one domain
- Search strategies which are actively maintained and updated by the professional community.

For our test collection we therefore aggregated samples from the following resources:

1. The CLEF 2017 eHealth Lab (Goeuriot et al., 2017) which includes a curated set of 20 topics for Diagnostic Test Accuracy (DTA) reviews. Each of these topics includes a manually constructed search strategy created by subject matter experts. The 20 search strategies in this collection yielded 102 disjunctions containing 898 terms (i.e. a mean of 8.80 terms per disjunction). Each term consists of a mean of 1.40 tokens.
2. The SIGN search filters[4] is an actively maintained collection of '*pre-tested strategies that identify the higher quality evidence from the vast amounts of literature indexed in the major medical databases*'. We also consulted the InterTASC Information Specialists' Sub-Group.[5] On their advice [Glanville, personal communication], we augmented our collection with two further strategies (Glanville, 2017). This resulted in a total of eight actively maintained strategies, consisting of 47

(Gianville, 2017). This resulted in a total of eight actively maintained strategies, consisting of 47 disjunctions containing 355 terms (i.e. a mean of 7.55 terms per disjunction). Each term consists of a mean of 1.70 tokens.

3. A collection of recruitment search strategies. There is no standard test collection for recruitment search, but there are various community initiatives to collect Boolean strings for recruitment, notably:

a. The Boolean Search Strings Repository[6]: a communal collection of recruitment search strings curated by Irina Shamaeva
b. The Boolean Search String Experiment[7]: a collection of Boolean strings collected by Glen Cathey to address a specific recruitment brief.

After deduplication, these two sources in combination yielded a total of 46 search strategies, containing 80 disjunctions with 571 terms (a mean of 7.15 terms per disjunction). Each term consists of a mean of 1.38 tokens.

In aggregate, these three sources represent data that is curated, actively maintained, and specific to more than one domain. In sum they contain a total of 74 search strategies consisting of 229 disjunctions and 1,824 individual query terms. To the best of our knowledge, our experiments represent the first study of this scale and coverage.

Research questions

In this paper, we investigate the following research questions:

1. To what extent can methods based on manually curated ontologies provide suitable query suggestions for professional search?
2. To what extent can methods based on context-free distributional language models provide suitable query suggestions for professional search?
3. To what extent can combining the above methods improve on the performance of either method in isolation?

Materials and methods

As discussed above, in our experimental setup we investigate the extent to which different methods can predict gold standard data in the form of human-generated search strategies. We consider a variety of methods, as follows:

1. Related terms extracted from manually curated ontologies
2. Terms generated using context-free distributional language models
3. Combinations of the above resources in a variety of configurations.

Manually curated ontologies

Query suggestions can be generated by querying manually curated ontological resources to identify related terms in the form of hypernyms, hyponyms etc. Many such resources are hosted on the web

related terms in the form of hypernyms, hyponyms etc. Many such resources are hosted on the web as Linked Open Data,[8] and support access via structured query languages such as SPARQL. We investigated a variety of such resources, of which the first two may be considered general-purpose, and the latter four specific to healthcare:

1. **DBpedia** is a project aiming to extract structured content from Wikipedia (Gangemi et al., 2018). The DBpedia data set describes 4.58 million entities, out of which 4.22 million are classified in a consistent ontology.
2. **WebISA** (Seitner et al., 2016) is a publicly available database containing hypernymy relations extracted from the CommonCrawl web corpus.[9] The LOD version contains 11.7 million hypernymy relations, each provided with rich provenance information and confidence estimates.
3. **Medical Subject Headings**[10] (MeSH) is a controlled vocabulary for the purpose of indexing documents in the life sciences. It contains a total of 25,186 *subject headings*, which are accompanied by a short description or definition, links to related descriptors, and a list of synonyms or very similar terms.
4. **RxNorm**[11] is a terminology that contains all medications available on the US market. It has concepts for drug ingredients, clinical drugs and dose forms.
5. The **British National Formulary** (BNF)[12] is a pharmaceutical reference that contains information about medicines available on the UK National Health Service (NHS).
6. **The DrugBank** database[13] is an online database containing information on drugs and drug targets. The latest release of DrugBank contains 11,683 drug entries, 1,117 approved biotech drugs, 128 nutraceuticals and over 5,505 experimental drugs.

We created SPARQL queries to their respective endpoints to retrieve related terms, and set the maximum number of results to the default of 100. In cases where querying a particular resource returned more than one type of related term (e.g. both 'broader' and 'narrower' terms), these were aggregated and returned as a single list.

Context-free distributional language models

Word embeddings have become the de facto representation standard in many NLP applications (Jurafsky and Martin, 2020), and can be used to generate query suggestions in the form of related terms. Word embeddings can be learned from text corpora using a variety of techniques, e.g. word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), BERT (Devlin et al., 2019) etc. A number of publicly available, pre-built embedding models are available, trained on sources such as Wikipedia (Pennington et al., 2014), GoogleNews (Mikolov et al., 2013a), and PubMed (Chiu et al., 2016). We investigate the following *context-free* embeddings:

- Word2vec trained on Google news (Mikolov et al., 2013b)
- GloVe trained on Wikipedia + Gigaword5 (Pennington et al., 2014)
- FastText trained on Wikipedia (Bojanowski et al., 2017)
- Word2vec trained on PubMed articles, with different window sizes (2 and 30) (Chiu et al., 2016)

We also built bespoke models using an PubMed Open Access full text snapshot which consisted of 944,672 full-text articles. Using an initial test set we identified the optimal parameter settings as dimensions = 300, window size = 5, min word count = 10. We created two bespoke Word2vec models: one which consisted solely of unigrams, and a second model which also included bigrams and trigrams.

Results

Our overall evaluation approach was as follows: for every strategy in our test collection, we iterate over each disjunction and calculate precision, recall and F score for each term, based on the overlap between the suggested term set and the gold standard. We then repeat this process for each method, and report performance in terms of average (arithmetic mean of) precision, recall and F score.[14] We test for significance using one-way ANOVA, and report values where $p < 0.01$.

Manually curated ontologies

Table 1 shows the arithmetic mean of precision (P) and recall (R) and the F score (F) for the manually curated resources with the highest F value highlighted in bold. Comparing F scores for the general purpose resources (DBpedia vs. WEBISA) shows a significant difference in favour of the former on all three data sets, particularly Recruitment $F(1, 1140) = 59.20$, $p < 0.01$.

Table 1. Precision, recall and F for manually curated resources.

	CLEF 2017 (n = 898)			**SIGN (n = 355)**			**Recruitment (n = 571)**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DBpedia	0.026	0.046	**0.033**	0.024	0.034	**0.028**	0.019	0.043	**0.026**
WebISA	0.013	0.010	0.011	0.014	0.009	0.011	0.005	0.004	0.004
MeSH	0.065	0.017	0.027	0.148	0.015	0.027	n/a	n/a	n/a
RxNorm	0.000	0.000	0.000	0.000	0.000	0.000	n/a	n/a	n/a
BNF	0.002	0.001	0.001	0.000	0.000	0.000	n/a	n/a	n/a
DrugBank	0.000	0.000	0.000	0.000	0.000	0.000	n/a	n/a	n/a

Note. Bold values represent highest F values.

The source of suggested terms has a significant effect on performance for both CLEF, $F(5, 5382) = 109.53$, $p < 0.01$ and SIGN $F(5, 2124) = 62.03$, $p < 0.01$. The use of a specialist resource appears to be beneficial in terms of precision, with relatively high values shown by MeSH (0.148 for SIGN data). This reflects the highly specialised nature of this resource. However, the best performing resource overall (in terms of F measure) remains DBpedia.

Context-free distributional language models

The results for the language models are shown in Table 2, with the highest F values highlighted in bold. Overall, these scores are generally higher than those of the ontological relations. The choice of model has a significant effect on performance, although the pattern is inconsistent: the bespoke PubMed unigram model performs best on CLEF $F(6, 6279) = 27.49$, $p < 0.01$, while the bespoke PubMed trigram model performs the best on SIGN $F(6, 2478) = 6.19$, $p < 0.01$. Their performance is comparable to that of Word2vec+PubMed (win30) (Chiu et al., 2016), which provides some evidence for the reproducibility of these results. Comparing the three generic models on recruitment data, GloVe+Wikipedia performs best $F(2, 1710) = 19.78$, $p < 0.01$. These results illustrate the value of using domain-specific models (the lower half of the table) rather than generic models (the upper half).

Table 2. Precision, recall and F for distributional models.

	CLEF 2017 (n = 898)			**SIGN (n = 355)**			**Recruitment (n = 571)**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Word2vec+Google News	0.033	0.037	0.035	0.027	0.025	0.028	0.041	0.035	0.038
GloVe+Wikipedia	0.044	0.047	0.045	0.026	0.030	0.028	0.057	0.047	**0.051**
FastText+Wikipedia	0.024	0.038	0.029	0.019	0.016	0.017	0.024	0.018	0.021
Word2vec+PubMed (win2)	0.057	0.062	0.059	0.026	0.028	0.027	n/a	n/a	n/a
Word2vec+PubMed (win30)	0.069	0.073	0.071	0.028	0.033	0.030	n/a	n/a	n/a
Bespoke word2vec+PubMed, unigrams	0.071	0.075	**0.073**	0.038	0.040	0.039	n/a	n/a	n/a
Bespoke									

Bespoke word2vec+PubMed, trigrams	0.069	0.072	0.072	0.042	0.040	**0.041**	n/a	n/a	n/a

Note. Bold values represent highest F values.

Combining sources

It may be possible to improve performance by combining results from two or more sources. Evidently, the nature of that improvement will depend on the particular sources being combined and the way in which their respective result sets intersect. In this section we investigate the effects of combining the best performing curated resources with the best performing language models.

Simple aggregation

The simplest form of aggregation is to combine two term suggestion sets as a 'bag of words'. Table 3 shows the results of applying a combination of the DBpedia ontology and the GloVe+Wikipedia language model to recruitment data (also showing the results for each method in isolation), with the highest values highlighted in bold. Combining two sources improves recall, but at the expense of precision, with a decrease in F score (compared to GloVe in isolation). Comparing F scores shows that aggregation has a significant effect on performance $F(2, 1710) = 20.14$, $p < 0.01$.

Table 3. Precision, recall and F for simple aggregation of terms from DBPEDIA and GloVe.

	Recruitment (n = 571)		
	P	**R**	**F**
DBpedia (alone)	0.019	0.043	0.026
GloVe+Wikipedia (alone)	**0.057**	0.047	**0.051**
Aggregated	0.030	**0.081**	0.044

Note. Bold values represent highest values.

Table 4 shows the results of combining the MeSH ontology with the word2vec PubMed trigram language model for healthcare (also showing the results for each method in isolation), with the highest values highlighted in bold. The combination offers improvements in both recall and F score for both data sets. Moreover, the use of aggregation has a consistently positive and significant effect

for both data sets. Moreover, the use of aggregation has a consistently positive and significant effect on performance on both CLEF $F(2, 2691) = 78.57$, $p < 0.01$ and SIGN $F(2, 1062) = 5.36$, $p < 0.01$.

Table 4. Precision, recall and F for simple aggregation of terms from MeSH and PubMed trigram model.

	CLEF 2017 (n = 898)			SIGN (n = 355)		
	P	**R**	**F**	**P**	**R**	**F**
MeSH (alone)	**0.065**	0.017	0.027	**0.148**	0.015	0.027
Bespoke PubMed trigram (alone)	0.071	0.075	0.073	0.042	**0.040**	0.041
Aggregated	0.082	**0.081**	**0.081**	0.075	0.035	**0.048**

Note. Bold values represent highest values.

Back-off approaches

One possible explanation for the positive effect of aggregation is that language models tend to learn robust representations for frequent terms, which tends to favour unigrams. By contrast, manually curated ontologies tend to provide better coverage of higher order ngrams (bigrams and above), which reflects their focus on named entities and other specialist terminology. To test this hypothesis, we implemented two further combinations which exploited the ngram order in finding related terms:

'Loose pipelining':

1. Tokenise the query term (based on whitespace)
2. If number of tokens >1, look up term (ngram) in curated ontology
3. Look up term (unigram or ngram) in language model
4. Combine results and return as a unified list

'Strict pipelining':

1. Tokenise the query term (based on whitespace)
2. If number of tokens >1, look up term (ngram) in curated ontology
    a. If no results from curated ontology, look up term (ngram) in language model
3. Else look up term (unigram) in language model
4. Combine results and return as a unified list

What these approaches have in common is that curated resources are only used for higher order ngrams (bigrams and above). Where they differ is that in the second variation the language model is only used if the curated ontology returned no results or if the term is a unigram. Table 5 shows the

results of this approach, along with the results from the approaches above (repeated here for convenience), with the highest values highlighted in bold:

Table 5. Precision, recall and F for combinations using backoff approaches.

	CLEF 2017 (n = 898)			SIGN (n = 355)			Recruitment (n = 571)		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Curated ontology	0.065	0.017	0.027	0.148	0.015	0.027	0.019	0.043	0.026
Language model	0.071	0.075	0.073	0.042	0.040	0.041	0.057	0.047	0.051
Simple aggregation	0.082	**0.081**	0.081	0.073	**0.074**	0.035	0.030	**0.081**	0.044
Loose pipelining	0.083	0.081	0.082	0.075	0.035	0.048	0.061	0.069	0.065
Strict pipelining	**0.100**	0.076	**0.086**	**0.135**	0.032	**0.052**	**0.065**	0.068	**0.066**

Note. Bold values represent highest values.

The results show that simple aggregation consistently produces the highest recall, which reflects the undifferentiated, broader nature of a combined suggested terms list. Conversely, 'strict pipelining' consistently produces the highest precision, which supports the hypothesis that ngram order can be exploited when finding related terms. Moreover, the F scores show that it is possible to combine suggestions from different sources using strict pipelining to deliver a more effective balance of precision & recall.

Discussion

It is important to recognise that although the use of query expansion has been the subject of many studies, relatively few have focused explicitly on the professional search context. To the best of our knowledge this is the first study of this scale to evaluate interactive expansion within the context of structured queries using publicly available, human-generated search strategies.[15]

Turning to the results themselves, we may make a few general observations. First, although some of the results may appear low in absolute terms, the key observation is that relative differences are statistically significant and generalisable. Moreover, the potential impact on professional search practice could be significant: with patent search tasks taking a median of 12 hours to complete

(Russell-Rose et al., 2018), even a 10 per cent saving due to improved query formulation would translate to 1.2 hours of billable time per task. Likewise, librarians spend an average aggregated time of 26.9 hours on systematic reviews, most of which is spent on search strategy development and translation (Bullers et al., 2018). Query expansion is known to be highly valued by healthcare information professionals, so the potential for adoption of even imperfect query suggestion techniques could lead to considerable impact.

Comparing the different techniques, we see that the use of language models outperforms methods based on manually-curated resources. It is possible of course that other human-curated resources may offer improved performance, e.g. ConceptNet,[16] Wikidata,[17] etc. However, the six sources investigated in this study offer a reasonable basis for comparison, and the investigation of additional resources is suggested as an area for further work.

In addition to the above, the practice of combining sources offers the prospect of further improvement, with simple aggregation having a consistently positive and significant effect on recall across all data sets. Moreover, it is possible to deliver a better balance between precision & recall by utilising ngram order in the combination, e.g. using strict pipelining to optimise for precision.

It is important also to recognise that the results represent a lower bound on potential performance, since some of the terms identified as false positives may transpire to be true positives in a live task scenario. For example, the first disjunction in the recruitment data set contains the terms:

['analyst', 'business analyst', 'business process analyst', 'data analyst', 'reporting analyst']

When DBPEDIA is queried using the second of these terms ('business analyst'), it returns the following suggestions:

```
['BA', 'Business occupations', 'Business terms', 'Systems analysis',
'Functional analyst', 'Software Business Analyst', 'Business analysis',
'Computer occupations', 'Business systems analyst', 'Analyst']
```

Arguably, the terms 'BA', 'Software business analyst', 'Business systems analyst' and 'Analyst' are all true positives. However, due to the offline evaluation process they are all labelled as false positives apart from 'Analyst', resulting in a precision of 0.1 instead of 0.4. Moreover, had the term 'BA' (a common abbreviation for 'business analyst') been included in the original disjunction, the recall would be 0.333 instead of 0.2.

This observation brings us naturally onto the limitations of this study. Although the test data represents a sizable collection of search strategies, there is no guarantee that they are optimal, i.e. they represent an 'ideal' articulation of the information needs they represent. Indeed, the very fact that they were created without access to the type of query formulation techniques proposed in this paper would imply that they are less than 'perfect'. However, this does not mean they are without value: the majority are drawn from hand-curated, published and publicly maintained sources, and represent the work of trained experts. They may not be ideal, but they are representative of a

broader population, and in this respect we believe they are a valid approximation of professional search behaviour.

Evidently, to accurately evaluate how real users would react in a real task scenario, it is necessary to set up a user study involving representative human participants. This is of course more expensive and time consuming, and user studies can be more challenging to scale and replicate. In this respect the value of this study is in investigating a diverse set of techniques using human generated search strategies as a proxy for human behaviour. As such it offers a scalable and reproducible approach which allows more expensive online studies to be better focused on specific issues and tasks.

Conclusions and further work

In this paper, we review the role of query suggestions within the context of professional search strategies used in real-world search tasks. We investigate a number of techniques for generating query suggestions, and evaluate them using a variety of data sources. We now draw conclusions in relation to our original research questions:

1. *To what extent can methods based on manually curated ontologies provide suitable query suggestions for professional search?*

We found that the source of suggested terms has a significant effect on performance, with the use of a specialist resource being beneficial in terms of precision, with relatively high values shown by MeSH. However, the best performing resource overall remains DBpedia.

2. *To what extent can methods based on context-free distributional language models provide suitable query suggestions for professional search?*

We found that context-free distributional language models outperformed the use of manually-curated resources. We also found that our own bespoke Pubmed model outperformed the best of the third party pre-built models on healthcare data. The best performing model on recruitment data was found to be GloVe+Wikipedia.

3. *To what extent can combining the above methods improve on the performance of either method in isolation?*

We found that simple aggregation consistently produced higher recall than any method in isolation. Moreover, the use of aggregate methods showed that it is possible to exploit ngram order in finding related terms. ‘Strict pipelining’ consistently produced the highest precision and highest overall F score, which demonstrates that it is possible to combine suggestions from different sources to deliver a better overall balance of precision & recall.

Future work

This work provides a benchmark set of results (in an under-explored area) for future experiments. A valuable next step would be to scale the work horizontally, e.g. to other curated resources (such as

ConceptNet[18] and Wikidata[19]) or to other distributional models and frameworks. A suitable next step may be to explore contextual embeddings such as BERT (Devlin et al., 2019), for example using neighbouring disjunction terms as context.

A further form of scaling is to investigate other domains: in this study we focused on healthcare and recruitment, aligning with two professions known to be among the heaviest users of complex, Boolean queries. It would be interesting to extend this work to other professions such as patent search, competitive intelligence, and media monitoring (Russell-Rose et al., 2018).

Finally, a further area for future work is to compare these findings with human judgements as might be elicited via a user study. This work could explore the degree to which our findings align with that of naturalistic use, and determine the extent to which false positives identified in our study may actually transpire to be true positives in live, interactive usage.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Innovate UK Open Competition R&D grant 102975, 'Intelligent Search Assistance'. Innovate UK had no involvement in the study design, data analysis, report writing or decision to submit for publication.

ORCID iD

Tony Russell-Rose 

Footnotes

Code availability Test data is publicly available via Github. Evaluation code is hosted on BitBucket and can be made available on demand.

---

1. However, some professional search could be mediated via the web, and conversely, not all work-based searching is professional in nature.

Go To Footnote