Document and Query Expansion for Information Retrieval on Building Regulations

Ruben Kruiper[1,2*], Ioannis Konstas[1], Alasdair J.G. Gray[1], Farhad Sadeghineko[2], Richard Watson[2], and Bimal Kumar[2]

[1]Heriot-Watt University, United Kingdom
[2]Northumbria University, United Kingdom
[*]r.kruiper@northumbria.ac.uk

Abstract Regulations and test criteria for building products are captured in hundreds of interrelated documents. It can be daunting to figure out which of these documents contain information that is relevant to your building project or product. In this paper, we describe work on an Information Retrieval (IR) system that aims to search through the contents of building regulations. Based on practitioner interviews we develop a small dataset of user-queries for which we would like to return relevant passages of documents. We explore several approaches to Query Expansion (QE) and Document Expansion (DE), taking into account the scarcity of openly available knowledge sources in our small technical domain. We show that IR performance can be greatly improved using QE and DE, and retrieve a top-3 relevant result for up to 85% of out queries. We share our IR dataset and the code to replicate our approach.

Keywords: Information Retrieval, Building Regulations, Query Expansion, Document Expansion

1 Introduction

Building regulations are put in place to ensure that buildings are safe, efficient, accessible, and so on (Meijer, Visscher, and Sheridan 2014) – we use '*building regulations*' to refer to any statutory regulations, guidance and associated standards. Most building work requires approval, hence regulations are frequently accessed by professionals in the construction industry, from architects to building inspectors and contractors (McKechnie, Shaaban, and Lockley 2001). However, it can be both challenging and time-consuming to identify relevant regulations, despite the existing hierarchical organisation of documents (Lau, Law, and Wiederhold 2005; Cheng et al. 2008; Zhong et al. 2020).

To exemplify this, consider that the British Standards Institute (BSI) online portal BSOL holds 2K regulatory documents that are classified as [GBM48 '*Construction In General*'][1]. But searching through these documents on BSOL often returns no results, e.g., no documents are returned for the query '*fire requirements of rafters*'. Even the query '*rafters*' returns only 2 currently active documents, with the following titles and International Classification for Standards classification(s):

[1]`https://bsol.bsigroup.com/` [Accessed April 2023]

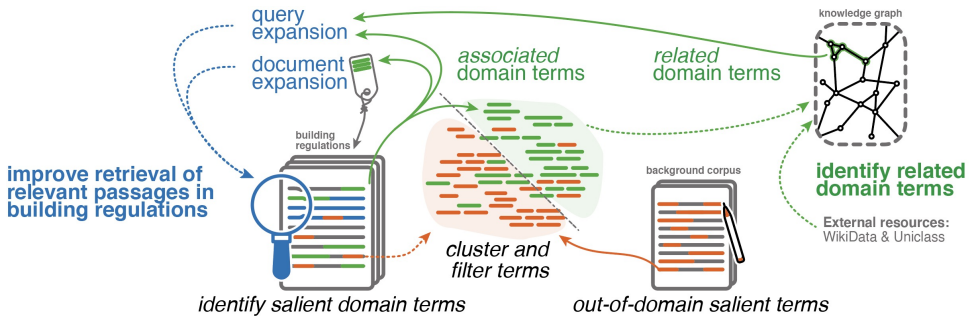

Figure 1: Our Information Retrieval (IR) system enables search through the contents of building regulations. To improve the matching potential of the queries and passages we implement Document and Query Expansion, respectively. Candidate terms are computed relying on associations between salient terms that were identified in the regulatory texts, as well as using relations found in a Knowledge Graph that links to external domain knowledge.

1. Structural design of low-rise buildings. Code of practice for timber floors and roofs for housing. [ICS: *91.040.30 - Residential buildings, 91.060.20 - Roofs, 91.060.30 - Ceilings. Floors. Stairs*]
2. Brackets for eaves gutters. Requirements and testing. [ICS: *91.060.20 - Roofs*]

Because little information is provided beyond the title and ICS classification, it can be complicated to determine the relevance of results. Unsurprisingly, construction professionals tend to search for relevant regulations using alternative methods, such as inspecting the approval documentation for related products on the market, or contacting trade organisations and consultants (Cerovsek 2009).

In this paper we present an open-source tool for Information Retrieval (IR) over building regulations. The system searches directly through the contents of regulations, retrieving passages of text within the documents that better demonstrate the relevance of documents. Because user queries may contain terminology that simply is not used in regulations, we explore Query Expansion (QE) and Document Expansion (DE) to improve the chances of identifying relevant results. As such, our evaluation is based on ad hoc, one-off user-initiated queries inspired by the following scenario: "*A manufacturer of building component X is looking to identify which standards, tests or criteria X should comply with.*". We develop a small IR dataset of 42 queries – notably BSOL does not return results for any of our queries. We share this dataset and the code of our system[2]. We show that the system returns a relevant result in the top-3 results for up to 85% of our queries, on a limited set of 420 building regulations in PDF format.

The paper is organised as follows. Section 2 describes related work. Section 3 describes our IR approach, which is visualised at a high-level in Figure 1, Section 4 describes our data and queries. Section 5 describes our experimental setup and results.

2 Related work

2.1 Sparse and dense Information Retrieval

Much of the work on IR over building regulations relies on sparse IR algorithms, such as BM25 (Cerovsek 2009; McGibbney and Kumar 2011; Lin, Chi, and Hsieh 2012). Sparse retrievers are fast and achieve good performance, but they rely on keyword-matching and term counts (Karpukhin et al. 2020). A common caveat of these methods is that if a search keyword does not occur in the text-to-be-retrieved, the text will not be returned – despite the presence of related or even synonymous terms. Our work aims to enable retrieval using Out-Of-Vocabulary (OOV) query keywords.

Recently Dense Passage Retrieval (DPR) has been shown to outperform strong sparse retrievers within the overarching task of Question Answering (QA) (Karpukhin et al. 2020). Here, queries and passages are encoded using a pretrained BERT-based embedder (Devlin et al. 2018). Encoding text with BERT limits the size of passages and queries, typically to 512 tokens. In comparison to indexing long texts as single documents, a benefit of dividing the text into smaller passages is that it enables a more fine-grained search (Sannier and Baudry 2012). And in comparison to keyword search, a benefit of encoding queries is the ability of handling long structured queries, such as sentences (Dai and Callan 2019). Finally, the encoding represents text as subword-units[3] that are better suited at handling rare, OOV and misspelled terms (Sennrich, Haddow, and Birch 2016). This is particularly useful in our setting of IR in a small technical domain, as we do not have access to enough clean domain text to pre-train a domain-specific BERT model.

In contrast to existing work on QA over building regulations (Zhong et al. 2020), we are not interested in answering informational queries along the lines of '*What is an extension joint?*'. We aim to retrieve documents that describe, e.g., the '*structural requirements of extension joints*'. And in comparison to existing work on IR over building regulations, we do not rely solely on domain ontologies to identify key terms and partition documents into passages.

2.2 Building Regulations terminology

Highlighting salient concepts in text helps identifying relevant building regulations (McKechnie, Shaaban, and Lockley 2001), e.g., based on the occurrence of these concepts in domain-specific controlled vocabularies (Cerovsek 2009; Cheng et al. 2008) and ontologies (McGibbney and Kumar 2011). Domain

[2]We share our code and data at: https://github.com/rubenkruiper/IRReC

[3]Subword-units refer to a finite set of character combinations that represent common snippets and words, e.g., a word like '*units*' may be broken down into '*unit*' and '*s###*' with the three hashtags indicating the end of a word. They enable computing unique representations for unseen words by combining the representations of the snippets that make up such new words.

knowledge has also been used to inform the division of regulations into smaller passages that revolve around a specific topic (Lin, Chi, and Hsieh 2012). Instead of relying on external sources of domain knowledge, one could use heuristics or off-the-shelf tools to identify an open-ended set of concepts in texts (Xu and W. B. Croft 1996; Lau, Law, and Wiederhold 2005; Lin, Chi, and Hsieh 2012). We explore both options: we automatically generate a Knowledge Graph (KG) from our set of regulatory documents relying on a recently developed tool (Kruiper et al. 2023b), which contains links to the prominent domain taxonomy Uniclass (Gelder 2015). Furthermore, we rely on the SPaR.txt tool for the discovery of Multi-Word Expressions (MWE) in the domain of building regulations (Kruiper et al. 2021), and filter general domain terms that overlap with concepts extracted from a set of medical device regulations – inspired by (Meyers et al. 2018).

2.3 Document and Query Expansion

In IR a query is matched to documents in a collection. One can try to improve results by increasing the matching potential of (1) the query, and/or (2) the documents. To achieve the latter one can enrich representations of the text with associated terms, this is sometimes called Document Expansion (DE) (Bai et al. 2005). Query Expansion (QE) methods aim to improve results for poorly formulated queries, by automatically adding related terms to a query (M. Mitra, Singhal, and Buckley 1998). Identifying candidate terms for expansion can be based on domain knowledge sources (Bouchoucha, He, and Nie 2013; Xiong and Callan 2015), or based on co-occurrence of the query term and candidate term (Qiu and Frei 1993; Gao et al. 2004; Cao, Nie, and Bai 2005; Diaz, B. Mitra, and Craswell 2016; Kuzi, Shtok, and Kurland 2016).

Often, queries do not capture a user's information need very well, e.g., a query may consist of only a few words (W. Croft, Cook, and Wilder 1995; Spink et al. 2001), the query words may not match the words used in any of the relevant documents (Furnas et al. 1987; Xu and W. B. Croft 1996), and/or a user may simply not know how to express the information need (Azad and Deepak 2019). QE usually appends synonyms – or otherwise relevant terms – to the query, effectively emphasising recall over precision (Carpineto and Romano 2012). A common approach to QE is to first retrieve documents using the non-expanded query. Then, from these documents relevant terms are selected and used to expand the query – the user is only shown documents retrieved by the expanded query (Rocchio 1971). Here, relevance of terms inside the document can be manually labelled, derived from user interactions with retrieved results, or simply assumed – the latter is known as Pseudo-Relevance Feedback (PRF) (Azad and Deepak 2019).

While QE has a long history, e.g., (Maron and Kuhns 1960), only in recent decades QE was shown to improve IR without trading off precision for recall (Carpineto and Romano 2012). QE candidates can '*damage*' a query (Xiong and Callan 2015), so selection of terms strongly influences how well QE works (Willett and Peat 1991). Recent work on QE for clinical terms relies on domain dictionaries to identify precise and relevant candidates (Kim et al. 2021), but licensing restrictions complicate this approach in the building regulations domain. Our automatically generated KG contains notions of term relatedness through definitions from WikiData (Vrandečić and Krötzsch 2014), a large general domain and open source knowledge base. As such, we consider DE and QE based on co-occurrence associations between automatically identified domain concepts, as well as QE based on external knowledge sources and PRF.

3 Methodology

Few studies explore the use of practitioner's terminology to enable search over building regulations (Cheng et al. 2008). We conduct interviews with participants from various roles related to architecture, which corroborate the literature's notion that queries often do not capture a user's information need very well – also see section 4.3 and appendix B. Our hypothesis is that IR accuracy and recall will be improved by emphasising the distributional similarity between salient domain-specific terms found in queries and passages. To this end we develop an IR system that can perform sparse and dense retrieval, see section 3.1. Sections 3.2 to 3.4 describe how we identify domain terms and emphasise their presence in passages and queries.

3.1 Information Retrieval

We implement our IR system using the Haystack framework (Pietsch et al. 2022) and retrieve passages of text found within documents. Sparse retrieval relies on the ElasticSearch (ElasticSearch 2022) implementation of BM25 and BM25F-multimatch over the passages. Dense indices and retrieval rely on DPR

(Karpukhin et al. 2020) encoding and FAISS (Johnson, Douze, and Jegou 2021). We also consider a weighted combination of both sparse and dense retrieval results, which we refer to as '*hybrid*'. Notably, we evaluate retrieval based on the ranking of single passages. For practical use of the system we actually score whole documents based on the scores of individually retrieved passages from the same document. Intuitively, a higher rank will be assigned to documents that contain many relevant passages according to one or more retrievers.

3.2 Identify domain terminology

To identify key terms in queries and passages we rely on a Named Entity Recognition tool. The system currently relies on SPaR.txt (Kruiper et al. 2021), which was developed to process building regulations and is able to identify MWEs. SPaR.txt expects single sentence inputs, but the sentences that we provide contain many processing errors, such as mistakes in detecting sentence boundaries. The reason is that sentences are directly extracted from the PDF documents without cleaning. This degrades the performance of term identification, and we filter out many messy and non-useful candidate terms using a handful of regular expressions, e.g., to remove email addresses and section numbers.

SPaR.txt does not discriminate between domain-specific and generic spans of texts. Therefore, we explore filtering general domain MWEs based on IDF-weights in comparison to a background corpus – following (Kruiper et al. 2023b). We rely on Phrase-BERT (Wang, Thompson, and Iyyer 2021) for tokenization and embedding of spans. A k-NN classifier is used to compare term embeddings to it's 2 nearest neighbours from a training set. The training set is automatically generated from the all terms that SPaR.txt finds in the fore and background corpora. Extracted terms are considered out-of-domain based on (1) how many of a term's 500 NNs can be found only in the foreground corpus, and (2) based on a modified Term Frequency-Inverse Document Frequency (TF-IDF) metric:

$$TF\text{-}IDF(t) = \log(1 + \frac{fc_t}{fc_t + bc_t}) * \log(avgIDF_t)$$

with fc_t the number of times term t occurs in the foreground corpus, bc_t the background corpus count, and $avgIDF$ the averaged IDF weight over the subword tokens of term t.

3.3 Document Expansion

We expand passages found in the building regulations by providing terms as labels. The aim is to emphasise the domain terms that occur and improve the matching potential. We provide our sparse BM25F retriever with additional fields and dense retrieval is composed of multiple indices, these fields and indices capture:

1. terms found by SPaR.txt in the passage: unfiltered, filtered and domain-classified filtered spans.
2. the nearest neighbours for the filtered set of terms identified by SPaR.txt in the passage.

Dense DE consists of running multiple retrievers and summing the scores for retrieved passages – a weight is applied to reflect the desired contribution of each retriever. Similarly, a weight is applied to the fields in sparse BM25F.

3.4 Query Expansion

We explore three methods of candidate identification for QE, resulting in three sets of candidates. The first set of candidates is derived through PRF, where the query is run through a sparse retriever and filtered SPaR.txt labels from the results are considered to be relevant terms. We count how often each of the SPaR.txt labels occur in all the results, and select the top k most common terms.

The second set of candidates are the Nearest Neighbours (NN)s for spans found in the query itself. We rely on SPaR.txt to identify spans in a query. Based on the normalised embeddings for these spans, we compute the top k most closely related terms out of all terms found in our corpus. We rely on the k-NN algorithm and aim to avoid morphologically similar terms, such a the plural form of a span, based on Levenshtein edit distance.

Third, we use the query-derived object spans to identify related nodes in our KG. To this end we represent the KG as a weighted undirected graph, and compute a relatedness score R:

$$R = distance_{span-candidate} * \log(degree_{candidate}) * AvgNeighbourhoodDegree_{candidate}$$

Where the $distance_{span-candidate}$ captures the types of relations that exist between the query span and the candidate span in the KG. The $degree_{candidate}$ captures how common a span is in terms of the number

Table 1: Overview of the number of documents, sentences, words and labels. *Word length is character-based, other lengths are counted in number of words.

	Total	**Unique**	**avg. len. (std. dev)**
Sentences	741K	214K	25.24 (17.52)
Words	18.7M	121K	5.14* (3.38)
NER labels	5.8M	569K	2.58 (3.52)
Filtered NER	1.6M	46K	1.27 (0.53)
Domain-specific NER	1.3M	42K	1.27 (0.53)
Neighbours	3.3M	46K	1.48 (0.62)

of relations to other spans in the KG. Similarly, $AvgNeighbourhoodDegree$ is a measure of how common a span's neighbours are. The intuition is that we favour closely related, yet common candidate spans.

The final candidates for QE are scored using the average IDF weights for each candidate, and a weight based on their origin – PRF, NN, KG. The expanded query is a concatenation of the initial query, the spans identified in the query through SPaR.txt, and the top k highest scoring expansion candidates.

4 Data and pre-processing

4.1 Document collection

In collaboration with the Building Research Establishment (BRE) and BSI, we collect a dataset of 420 British Standards in PDF format that our universities have a license for. To discriminate between generic and domain terms, we contrast the noun-based SPaR.txt spans that occur in our set of Building Regulations, against a set of European regulations on the design of medical devices – following (Kruiper et al. 2023b).

Text is extracted from our corpus using pdftotext[4]. Without additional pre-processing we split texts into sentences and using the Penn Treebank Tokenizer and use PunkTokenizer (Kiss and Strunk 2006) to count words – see Table 1 for an overview of counts. We find a total of 288K passages, each being a concatenation of whole sentences with a maximum length of 100 words. We do not distinguish between section titles and main body text, nor do we currently retain structural information from tables, lists, and multi-column documents. Nevertheless, we assume that within our 100-word passages it is feasible to find combinations of terms that are related to the terms within a query.

4.2 Domain knowledge

We would like to expand queries and documents with relevant domain terminology. We consider how to identify such domain terms (1) from the building regulations themselves, and (2) from external resources. We rely on (Kruiper et al. 2023b) to automatically generate a KG of terms found in our set of regulation documents, and related terms found in WikiData. The nodes in the KG are domain-specific terms found in our corpus of building regulations; as identified by SPaR.txt, then filtered and classified. Edge types in the KG include semantic similarity, morphological similarity and whether a term occurs in the WikiData definition of another term – we refer the reader to (Kruiper et al. 2023b) for details.

As existing sources of domain terminology we consider: NRM3[5] the IFCowl ontology (Pauwels, Zhang, and Lee 2017), and Uniclass (Gelder 2015). We find that most terminology in these resources is not readily useful for QE– within the scope of IR over building regulations. To exemplify this, we describe some of the issues that we encountered with the terminology found in the largest of the three, Uniclass.

Uniclass is used by various manufacturers of building products (Alani et al. 2020), and has been proposed as the terminological standard for the Platform Design for Manufacture and Assembly framework, e.g., for tracking and recording critical data throughout design and maintenance of building work (Bryden Wood Technology Limited 2017). However, Uniclass is organised following ISO standard 12006-2:2015 to ensure interoperability with ICS (Gelder 2015), and, generally, an issue with classification schemes like ICS is that they represent the needs of the agencies that issue and enforce the regulations, rather than the needs of users (Cheng et al. 2008). Of the 15K Uniclass labels only 1.662 (11,1%) occur verbatim in our 407 documents. On the one hand, Uniclass has a far wider coverage than our set of building

[4] `https://github.com/jalan/pdftotext` [Accessed October 2022]

[5] `https://www.rics.org/uk/upholding-professional-standards/sector-standards/construction/nrm/` accessed October 2022

Table 2: Overview of the query lengths before and after QE – candidates were drawn from the KG and NNs.

	Queries	**Expanded queries**
Avg. length (words)	5.86	20.19
Std. deviation	2.10	4.37
Shortest	3	12
Longest	13	34

regulations. On the other hand, many leaf nodes are amalgamations of properties that are unlikely to appear verbatim in text, such as '*Fibre cement profiled sheet self-supporting cladding systems*'.

Similar to Uniclass, the terminology found in IFCowl and NRM3 requires additional processing before being useful for QE. Specifically, the terminology does not align well with the terminology used in the regulations (Kruiper et al. 2023a), nor does it align particularly well with the types of queries we may expect. In case such terms are found in a query they are expected to require expansion, by identifying similar terms that are easier to match with relevant regulations – the terms themselves are not good candidates for expansion.

4.3 User queries

We conducted two sets of interviews with participants from various roles related to architecture – for details see Appendix A. The aim was to identify how users of regulations formulate their queries and the type of terminology they use. The information need in queries is found to be both *navigational*, e.g., searching for a specific document, as well as *informational*, e.g., searching for the information content regardless of source (Hedin et al. 2009). We also find that queries are significantly longer than the average ±2 words of web-queries[6] (Xu and W. B. Croft 1996). One reason is that users often combine specific facets to their information need, e.g., *'fire requirements' of 'external cladding' on 'flats'*. On top of this, many (±50%) of the domain terms in building regulations are MWEs.

Two domain experts provide a set of 42 queries that were curated to express a specific information need – see Appendix B. The aim is to avoid vague queries that complicate the evaluation of the IR results, e.g., a query like '*bitumen felt, roof covering*' is preferred over simply '*felt*'. BSOL does not return results for any of our queries. Table 2 provide some insights in the lengths of our queries. The longest query is 12 words and the shortest 2, with an average query length of 5.85 words and standard deviation of 1.80. We use SPaR.txt (Kruiper et al. 2021) to identify key terms in the queries. To help SPaR.txt identify key terms in our queries, we ensure that our queries are placed in context of a sentence or as a comma-separated termlist.

[6] `https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/` [Accessed April 2023]

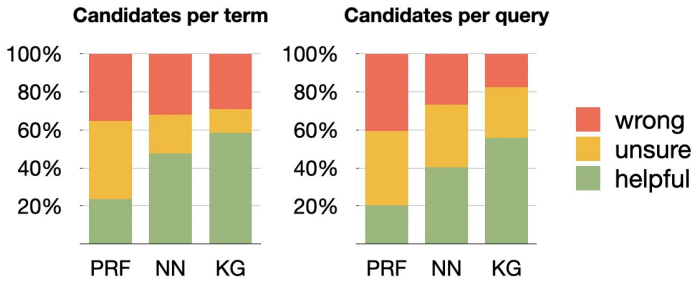

Figure 2: Overview of QE annotation results. The graph on the left side indicates how often a candidate was annotated as a useful expansion term for each setting. The graph on the right indicates how often a query was improved or damaged as a result of the candidate.

Table 3: Overview of results for each of our IR settings, all measures are based on the top 3 results for each query. Bold-faced values indicate best performance. MRR stands for Mean Reciprocal Rank, MAP for Mean Average Precision, the F1 score is a harmonic mean between precision and accuracy.

	sparse	**sQ**	**sD**	**sQD**	**dense**	**dQ**	**dD**	**dQD**	**hybrid**	**hQ**	**hD**	**hQD**
avg.F1	47.62	48.10	**59.76**	59.05	44.05	48.10	59.29	56.90	44.05	48.10	59.29	56.90
MRR	56.35	50.00	**66.27**	65.48	49.21	51.98	61.51	63.89	49.21	51.98	61.51	63.89
MAP	54.96	50.00	**65.28**	64.48	47.82	52.78	61.71	63.10	47.82	52.78	61.71	63.10
Total/129	50	50	**65**	63	47	50	**65**	60	47	50	**65**	60

5 Evaluation

5.1 Query Expansion settings

We manually investigate the QE candidates for the unique spans that occur in our 42 queries. The top 3 QE candidates are selected from the NNs, from PRF, and from the KG. Annotation results for the quality of QE candidates are visualised in Figure 2. An example of a NN-derived QE candidate that hurts performance is '*flat roof*' when searching for '*pitched roof*'. For KG candidates it is possible to better control which candidates are retrieved, e.g., by avoiding the expansion of common terms like '*roof*'. An example of a candidate where annotators are not sure of its effects is '*girder joists*' for the span '*timber joist*', as the former are often not made of timber. Based on the annotation results we omit PRF candidates from our IR experiment. The weight for NN and KG candidates is both set to 1.

PRF candidates are derived from both relevant and irrelevant documents, the latter is known to cause query drift – where an expanded query misrepresents the initial *information need* (M. Mitra, Singhal, and Buckley 1998). We find that the quality of our PRF candidates is relatively low and indeed many queries get '*damaged*' – see Figure 2. It is worth considering that passages of 100 words are relatively short, which may decrease the potential of finding strong candidates through PRF. The candidates from the KG and especially NN provide are more often thought to be helpful. Notably, controlling KG-based expansion of query terms is more easily implemented, e.g., based on the degree of a term in the KG or specific relations between terms.

5.2 Information Retrieval settings

We run several ablation experiments with our system, where the resulting passages have to be manually annotated. The ablations aim to compare sparse against dense and hybrid retrieval, each with and without QE, DE and both. The top 3 ranked results are compared for each of the 12 experimental settings, leading to total of 1,512 query and result combinations. Due to the number of annotations and time constraints, we do not experiment with different settings for weighting DE labels. Our annotation guidelines can be found in Appendix C. The weight for passage-texts is set to 2, all other retrieval fields for the passages are set to 1: document title, unfiltered SPAR.TXT labels, filtered SPAR.TXT labels, domain-specific SPAR.TXT labels, and nearest neighbours for the SPAR.TXT labels.

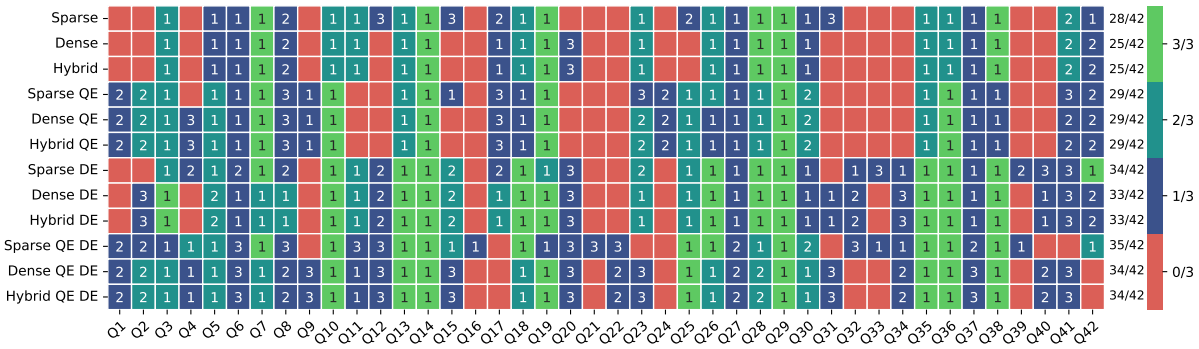

Figure 3: Overview of annotation results per query for each of our IR settings. Colours reflect the number of positive results, out of a possible total of 3 for each query. The numbers inside the cells indicate the best rank for a positive result. On the right-hand side an indication is given of how many queries were answered out of the total 42 queries.

5.3 Information Retrieval results

Table 3 provides an overview of metrics for evaluating IR system performance. For the domain of building regulations it seems that dense retrieval, using out-of-the-box DPR encoding, performs worse than the basic sparse retriever – disproving the distributional similarity aspect of our hypothesis. This is likely because DPR is trained on general domain text, and may have trouble finding a good representation for domain terminology, such as '*u-value*' or '*timber joist*'.

Following standard IR metrics, such as Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), QE slightly decreases performance on sparse retrieval and improves dense retrieval. Figure 3 provides some more detailed insight in the results per query. QE slightly improves the ability of retrievers to find an answer to queries within the top 3 results. We expect that by tweaking the settings and QE approaches, it is possible to find a better balance between precision and recall. In all settings DE improves performance significantly, both in terms of metrics and the number of queries answered. While sD performs best in terms of metrics, sQD is able to provide a relevant result in the top-3 for more queries. These findings provide evidence in favour of our hypothesis, that domain-specific IR accuracy and recall may both be improved by emphasising salient terms in queries and passages.

6 Conclusions

This study investigates passage level IR over the relatively small technical domain of building regulations. Traditional sparse retrieval can be complicated when query terms do not occur in the building regulations at all. We explore how DE, QE and dense retrieval can enable OOV search and improve IR in a domain where there exist few openly available resources. Our results indicate that retrieval performance can be greatly improved by highlighting salient terms in passages, without relying on manually curated external sources of domain knowledge. Nevertheless, the significance of our quantitative evaluation is limited due to (1) our relatively small evaluation dataset and (2) our non-comprehensive set of input documents.

Our approach and findings may benefit the development of IR systems in the domain of building regulations. We find that industry practitioners often use terminology that differs from the terms found in regulatory documents – as suggested by literature. Improving IR based on such terms can also help querying regulations directly from CAD software, e.g., based on a selection of Building Information Model (BIM) components. The reason is that the terminology that can be found in BIM models, such as the terms found in IFCowl and Uniclass, differs significantly from the terminology used in the regulations.

We share our code and IR dataset, hoping to encourage readers to extend our methods and apply the system in novel ways or in other domains. The current system is a proof-of-concept and we can think of several ways to try and improve performance. E.g., domain classification performance is limited as we currently rely on a background corpus that is a few orders of magnitude smaller than our foreground corpus. Where the system relies on existing tools, e.g., for NER and KG generation, once could consider trying alternative solutions to provide the same functionality.

Acknowledgments

This research was funded by the Building Research Establishment (BRE) as part of the Construction Innovation Hub (CIH). The authors are grateful to BSI for sharing some of their standards. We are also grateful to BIMacademy, specifically Murillo Piazzi for conducting stakeholder interviews and Lee Maguire for developing a web-based user interface. Furthermore, we thank xBIM, where Steve Lockley and Andrew Ward helped us explore querying our system directly from BIM-models.

References

Alani, Yasir et al. (Aug. 2020). "Whole Life Cycle Construction Information Flow using Semantic Web Technologies: A Case for Infrastructure Projects". In: *Proc. 37th CIB W78 Information Technology for Construction Conference (CIB W78)*, pp. 141–155. DOI: 10.46421/2706-6568.37.2020.paper011. URL: https://itc.scix.net/paper/w78-2020-paper-011.

Azad, Hiteshwar Kumar and Akshay Deepak (2019). "Query expansion techniques for information retrieval: A survey". In: *Information Processing and Management* 56.5, pp. 1698–1735. ISSN: 03064573. DOI: 10.1016/j.ipm.2019.05.009.

Bai, Jing et al. (2005). "Query expansion using term relationships in language models for information retrieval". In: *International Conference on Information and Knowledge Management, Proceedings*, pp. 688–695. ISBN: 1595931406. DOI: 10.1145/1099554.1099725.

Bouchoucha, Arbi, Jing He, and Jian Yun Nie (2013). "Diversified query expansion using ConceptNet". In: *International Conference on Information and Knowledge Management, Proceedings*, pp. 1861–1864. ISBN: 9781450322638. DOI: 10.1145/2505515.2507881.

Bryden Wood Technology Limited (2017). *Delivery Platforms for Government Assets*. Tech. rep., p. 146. URL: https://www.cdbb.cam.ac.uk/system/files/documents/DigitalBuiltBritainbook_screen.pdf%0Ahttps://www.brydenwood.co.uk/filedownload.php?a=9969-5d78cbf52e90d.

Cao, Guihong, Jian Yun Nie, and Jing Bai (2005). "Integrating word relationships into language models". In: *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, New York, USA: ACM Press, pp. 298–305. ISBN: 1595930345. DOI: 10.1145/1076034.1076086.

Carpineto, Claudio and Giovanni Romano (2012). "A survey of automatic query expansion in information retrieval". In: *ACM Computing Surveys* 44.1. ISSN: 03600300. DOI: 10.1145/2071389.2071390.

Cerovsek, Tomo (Dec. 2009). "Advancing regulation retrieval with profiling, controlled vocabularies and networked services". In: *2009 2nd International Conference on Adaptive Science & Technology (ICAST)*. IEEE, pp. 257–264. ISBN: 978-1-4244-3522-7. DOI: 10.1109/ICASTECH.2009.5409716. URL: http://ieeexplore.ieee.org/document/5409716/.

Cheng, Chin Pang et al. (2008). "Regulation retrieval using industry specific taxonomies". In: *Artificial Intelligence and Law* 16.3, pp. 277–303. ISSN: 09248463. DOI: 10.1007/s10506-008-9065-5.

Croft, W, Robert Cook, and Dean Wilder (1995). "Providing government information on the Internet: Experiences with THOMAS". In: *Proceedings of Digital Libraries Conference*. ISBN: 2027079629.

Dai, Zhuyun and Jamie Callan (2019). "Deeper text understanding for IR with contextual neural language modeling". In: *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 985–988. DOI: 10.1145/3331184.3331303.

Devlin, Jacob et al. (Oct. 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. URL: https://github.com/tensorflow/tensor2tensor%20http://arxiv.org/abs/1810.04805.

Diaz, Fernando, Bhaskar Mitra, and Nick Craswell (2016). "Query expansion with locally-trained word embeddings". In: *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers* 1, pp. 367–377. DOI: 10.18653/v1/p16-1035.

ElasticSearch (2022). *Free and Open Search: The Creators of Elasticsearch, ELK & Kibana | Elastic*. URL: https://www.elastic.co/.

Furnas, G W et al. (1987). "The Vocabulary Problem Hen y Ledgard Editor in Human-System Communication". In: *Communications of the ACM* 30.11, pp. 964–971.

Gao, Jianfeng et al. (2004). "Dependence language model for information retrieval". In: *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 170–177. DOI: 10.1145/1008992.1009024.

Gelder, John (2015). "The principles of a classification system for BIM: Uniclass 2015". In: *Proceedings of the 49th International Conference of the Architectural Science Association* 1, pp. 287–297. URL: https://anzasca.net/wp-content/uploads/2015/12/028_Gelder_ASA2015.pdf.

Hedin, Bruce et al. (2009). "Overview of the TREC 2009 legal track". In: *NIST Special Publication*, pp. 1–9. ISSN: 1048776X.

Johnson, Jeff, Matthijs Douze, and Herve Jegou (2021). "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data* 7.3, pp. 535–547. ISSN: 23327790. DOI: 10.1109/TBDATA.2019.2921572. URL: https://github.com/facebookresearch/faiss.

Karpukhin, Vladimir et al. (Nov. 2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 6769–6781. ISBN: 9781952148606. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://aclanthology.org/2020.emnlp-main.550.pdf%20https://www.aclweb.org/anthology/2020.emnlp-main.550.

Kim, Bosung et al. (Oct. 2021). "Query Reformulation for Descriptive Queries of Jargon Words Using a Knowledge Graph based on a Dictionary". In: *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, pp. 854–862. ISBN: 9781450384469. DOI: 10.1145/3459637.3482382.

Kiss, Tibor and Jan Strunk (2006). "Unsupervised multilingual sentence boundary detection". In: *Computational Linguistics* 32.4, pp. 485–525. ISSN: 15309312. DOI: 10.1162/coli.2006.32.4.485.

Kruiper, Ruben et al. (2021). “SPaR.txt, a Cheap Shallow Parsing Approach for Regulatory Texts”. In: pp. 129–143. DOI: 10.18653/v1/2021.nllp-1.14.
— (2023a). “Don’t shoehorn, but Link Compliance Checking Data”. In.
— (2023b). “Taking stock: a Linked Data inventory of Compliance Checking terms derived from Building Regulations”. In.
Kuzi, Saar, Anna Shtok, and Oren Kurland (2016). “Query expansion using word embeddings”. In: *International Conference on Information and Knowledge Management, Proceedings* 24-28-Octo, pp. 1929–1932. DOI: 10.1145/2983323.2983876.
Lau, Gloria T, Kincho H Law, and Gio Wiederhold (2005). “Legal information retrieval and application to E-rulemaking”. In: *Proceedings of the International Conference on Artificial Intelligence and Law*. New York, New York, USA: ACM Press, pp. 146–154. ISBN: 1595930817. DOI: 10.1145/1165485.1165508. URL: http://www.westlaw.com..
Lin, Hsien Tang, Nai Wen Chi, and Shang Hsien Hsieh (2012). “A concept-based information retrieval approach for engineering domain-specific technical documents”. In: *Advanced Engineering Informatics* 26.2, pp. 349–360. ISSN: 14740346. DOI: 10.1016/j.aei.2011.12.003. URL: http://dx.doi.org/10.1016/j.aei.2011.12.003.
Maron, M. E. and J L Kuhns (1960). “On Relevance, Probabilistic Indexing and Information Retrieval”. In: *Journal of the ACM (JACM)* 7.3, pp. 216–244. ISSN: 1557735X. DOI: 10.1145/321033.321035.
McGibbney, L.J. and Bimal Kumar (2011). “A Knowledge-directed Information Retrieval and Management Framework for Energy Performance Building Regulations”. In: *International Workshop on Computing in Civil Engineering*.
McKechnie, John, Sameh Shaaban, and Stephen Lockley (2001). “Computer Assisted Processing of Large Unstructured Document Sets: A Case Study in the Construction Industry”. In: *Proceedings of the ACM Symposium on Document Engineering*, pp. 11–17. ISBN: 1581134320.
Meijer, Frits, Henk Visscher, and Lilly Sheridan (2014). *Building regulations in Europe Part I: A comparison of the systems of building control in eight European countries*. Delft: Delft University Press Science. ISBN: 9040723737.
Meyers, Adam et al. (2018). “The termolator: Terminology recognition based on chunking, statistical and search-based scores”. In: *Frontiers in Research Metrics and Analytics* 3:19.January, pp. 1–14. ISSN: 16130073. DOI: 10.3389/FRMA.2018.00019/FULL.
Mitra, Mandar, Amit Singhal, and Chris Buckley (1998). “Improving automatic query expansion”. In: *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pp. 206–214. ISSN: 01635840. DOI: 10.1145/290941.290995.
Pauwels, Pieter, Sijie Zhang, and Yong Cheol Lee (2017). “Semantic web technologies in AEC industry: A literature overview”. In: *Automation in Construction* 73, pp. 145–165. ISSN: 09265805. DOI: 10.1016/j.autcon.2016.10.003. URL: http://dx.doi.org/10.1016/j.autcon.2016.10.003.
Pietsch, M. et al. (2022). *Haystack*. URL: https://github.com/deepset-ai/haystack.
Qiu, Yonggang and H P Frei (1993). “Concept based query expansion”. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Infofmation Retrieval*, pp. 160–169. ISBN: 0897916050. DOI: 10.1145/160688.160713.
Rocchio, J (1971). “Relevance feedback in information retrieval”. In: *The SMART Retrieval System—Experiments in Automatic Document Processing*, pp. 313–323. URL: https://ci.nii.ac.jp/naid/10000074359/.
Sannier, Nicolas and Benoit Baudry (2012). “Toward multilevel textual requirements traceability using model-driven engineering and information retrieval”. In: *2012 2nd IEEE International Workshop on Model-Driven Requirements Engineering, MoDRE 2012 - Proceedings*, pp. 29–38. DOI: 10.1109/MoDRE.2012.6360072.
Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). “Neural machine translation of rare words with subword units”. In: *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*. Vol. 3, pp. 1715–1725. ISBN: 9781510827585. DOI: 10.18653/v1/p16-1162.
Spink, Amanda et al. (2001). “Searching the Web: The Public and Their Queries”. In: *Journal of the American Society for Information Science and Technology* 52.3, pp. 226–234. ISSN: 15322882. DOI: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R. URL: http://www.excite.com.
Vrandečić, Denny and Markus Krötzsch (2014). “Wikidata: A free collaborative knowledgebase”. In: *Communications of the ACM* 57.10, pp. 78–85. ISSN: 15577317. DOI: 10.1145/2629489.
Wang, Shufan, Laure Thompson, and Mohit Iyyer (2021). “Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration”. In: Figure 1, pp. 10837–10851. DOI: 10.18653/v1/2021.emnlp-main.846.

Willett, Peter and Helen Peat (1991). "The limitations of term co-occurence data for query expansion in document retrieval systems". In: *Journal of the American Society for Information Science* 42.5, pp. 378–383. ISSN: 0002-8231.

Xiong, Chenyan and Jamie Callan (2015). "Query expansion with Freebase". In: *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval*, pp. 111–120. DOI: 10.1145/2808194.2809446.

Xu, Jinxi and W Bruce Croft (1996). "Query expansion using local and global document analysis". In: *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pp. 4–11. DOI: 10.1145/243199.243202.

Zhong, Botao et al. (2020). "A building regulation question answering system: A deep learning methodology". In: *Advanced Engineering Informatics* 46.April, p. 101195. ISSN: 14740346. DOI: 10.1016/j.aei.2020.101195. URL: https://doi.org/10.1016/j.aei.2020.101195.

A Query collection: interviews

The aim of interviewing domain practitioners is to identify how users of regulations formulate their queries and the type of terminology they use. The scope of our interviews is limited to fire safety and structural requirements for the active/passive roof subassembly, as defined by the EU harmonised standards. In a **first round** of eight 30-minute interviews participants were shown an image of one of the components of the active roof subassembly, an example can be seen in Figure 4.

Figure 4: Example of one of the figures shown to interviewees during the first round of interviews.

Given an image of a building component, participants were asked 12 questions that aim to identify the terminology one might use to search for relevant standards:

- Could you list 5 terms you would use to name the labelled element in the picture?
  - And are there any other terms you know that are used to describe this element?
- When designing the element(s) in the picture, what aspects would you need to consider to ensure that your designs and the work on site were compliant with building standards?
- Are there any aspects you would need to consider regarding the **structure**?
  - With regards to the element labelled in the picture and the considerations you have described, which words or phrases would you use to **search for standards** related to compliance of the structural performance (i.e **mechanical resistance** and **stability**)?
  - Can you think of any other words or phrases you would try?
- Are there any aspects you would need to consider regarding **the safety in case of fire**?
  - With regards to the element labelled in the picture and the considerations you have described, which words or phrases would you use to **search for standards** related to **compliance for safety in case of fire**?
  - Can you think of any other words or phrases you would try?
- With regards to the element labelled in the picture, **which section(s) of which standard(s)** do you find to be relevant to achieve compliance against **the mechanical resistance and stability requirements** of the search term(s) you mentioned?

  - Does the standard mentioned in the previous question describe a test?
  - Do you know any (other) standard(s) that describe relevant tests?
- With regards to the element labelled in the picture, **which section(s) of which standard(s)** do you find to be relevant to achieve compliance against **safety in case of fire requirements** of the search term(s) you mentioned?
  - Does the standard mentioned in the previous question describe a test?
  - Do you know any (other) standard(s) that describe relevant tests?
- Do you know any other people who would be available to answer these questions?

We collect a total of 1K queries. Participants are often unable to suggest relevant standards, indicating that it is unlikely that users of regulations know which documents to use by from memory. The few suggested standards were more generic documents, such as Approved Document B, that refer to other documents for specific requirements.

A **second round** of six 60-minute interviews emphasised the search for relevant standards. Participants were provided with 5 queries that were randomly selected from the first round of interviews. Participants were asked to search for relevant standards using conventional search engines, and reformulate the query if no satisfactory results could be identified. Reformulated queries were recorded, as well as 41 unique documents that were thought to be relevant to about half of the queries.

B Query selection

The final selection of queries is based on 855 unique queries collected during the interviews described above. We initially randomly select 100 queries and manually check how well a query expresses an information need. As an example, annotators find it unclear what information is intended to be found by the query '*felt roll laid onto insulation board*'. 21 of the randomly selected queries are replaced, as they are either vague or nearly identical to other queries in the list. Replacements are drawn from the 855 unique queries. However, during the initial annotation rounds we find that due to the format and limited scope of the interviews many of our queries still express a similar information need.

To ensure that the queries are diverse and the information need is clear, we ask two of our annotators to each select 20 of the 100 queries and provide them with a narrative. Some of the queries are rewritten to better reflect the information need, e.g.:

- Initial query: *roof covering drainage*
- Narrative: What are the requirements for pitched roof drainage in terms of adequately carrying rainwater to a suitable outlet, like gutters?
- New query: *pitched roof, drainage requirements*

The end result of this process is a set of 42 query and narrative pairs. The format and length of queries imitates that of the queries provided by interview participants. Notably, some of the queries are provided with punctuation or placed in a sentence context to aid breaking up queries with SPaR.txt.

C Annotation notes

Annotation is done by three domain experts. Initially they individually annotate all query and result pairs. Their individual annotations are compared and mismatches are discusses to find a consensus. Our general guidelines for annotation are:

- Relevance is based on the passage text; an irrelevant passage from a document with a relevant title is marked as irrelevant.
- Passages from a document index may be considered relevant based on the section titles that are part of the passage.
- A passage may be considered relevant if it describes requirements for a broader or narrower category than intended by the query, e.g.:
  - A passage on general fire performance ratings for '*roofs*' is considered relevant even when the query specifies '*flat roofs*'.
  - A passage on '*trussed rafters*' is considered relevant for a broader query on '*rafters*'.
  - However, if a specific material is described in the query, passages are usually only relevant if the document title or passage mention this material.
- A passage that indicates which document or section describes the sought-after information is considered relevant.