Challenges in the development of effective systems for Professional Legal Search

Piyush Arora, Murhaf Hossari, Alfredo Maldonado, Clare Conran, Gareth J. F. Jones
ADAPT Centre, Dublin City University
Dublin, Ireland
firstname.lastname@adaptcentre.ie
Alexander Paulus, Johannes Klostermann, Christian Dirschl
Wolters Kluwer, Germany
firstname.lastname@wolterskluwer.com

Abstract

The key objective of an information retrieval (IR) system is to identify and return to the user content relevant or useful in addressing the information need which required them to use the system. The development and evaluation of IR systems relies on the availability of suitable datasets or test collections. These typically consist of a target document collection, example search queries representative of those that users of the system to be developed, are expected to enter, and relevance data indicating which documents in the collection are relevant to the information needed as expressed in each query. Public research in IR has focused on popular content, e.g. news corpora or web content, for which average users can pose queries expressing information needs and judge the relevance of retrieved documents. This is not the case for professional search applications, for example *legal, medical, financial* search where domain experts are required for these tasks. We describe our experiences from the development of a professional legal IR application employing semantic search technologies. Our activities indicate the vital need for close interaction between the professionals for which the application is being developed and the IR researchers throughout the development life cycle of the search system. Such engagement is vital in order for the IR researchers to understand the working practices of the professional searchers, the specifications of their information needs and the domain in which they are searching, and to study how they engage and interact with information. A key reason to seek to understand these topics so deeply is to facilitate meaningful evaluation of the effectiveness of IR technologies as components in the system being developed.

1 Introduction

Information retrieval (IR) systems seek to address users' information needs by identifying and returning information relevant or useful in addressing their information needs. A key element of the development of effective IR systems for specific tasks or application areas is the evaluation of their effectiveness in retrieving relevant

In: Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop on Data Search (DATA:SEARCH18). Co-located with SIGIR 2018, Ann Arbor, Michigan, USA – 12 July 2018, published at http://ceur-ws.org

content. Doing this relies on the availability of suitable datasets or test collections representative of the retrieval task for which the system is being developed.

Published research in IR has focused almost exclusively on popular or publicly available content, e.g. news corpora or web content. For this content average users are able to compose queries expressing realistic information needs and to judge the relevance of retrieved documents. While test collections of this sort are sufficient to enable research into new approaches to IR in general, they do not support research for professional search applications, e.g. *legal, medical, financial*. In these settings, the realistic search queries need to be posed by task experts, and relevance assessment of retrieved content can similarly only be carried out by experts who understand the context of the search and the relevance of information to it. Even if an IR researcher believes that they do understand the content sufficiently well to compose queries and judge relevance, it is often the case that in professional terms they do not, and that they are not able to interpret retrieved documents to understand the reasons for them being deemed relevant or not-relevant by a domain professional assessor. Thus, they are not able to use informal inspection of the behaviour of an IR system in development to guide the development of IR methods suitable for the task in hand. For example, examination of the interaction between queries and the relevant and non-relevant content which they retrieve, often an important element of IR system development, is not possible. This issue is further compounded if the language of the content and queries is not one with which the IR system's developer is familiar.

IR research generally takes place in laboratory settings where researchers are able to specify the details of the search task to be explored and often design tasks and datasets to investigate specific research questions of their choosing. Search in professional settings does not enjoy this form of flexibility. If an IR system in a professional setting is to be successful, it must work with the document collection to be searched, to address the actual information needs of the professional users within their work tasks, and to successfully retrieve content which is useful to them in fulfilling their work objectives.

In the remainder of this paper, we present our experience in a project developing an IR application for professional legal search in the *German* language. We describe the development challenges we encountered where the IR developers were neither legal experts nor native German language speakers. This includes details of test set development for the evaluation of the performance of retrieval methods and a preliminary experimental investigation using this dataset. It is our hope that in sharing our experiences in the development of a professional search application, we can contribute to discussions on establishing methods for development of effective IR solutions in *legal* and other professional search domains.

This paper is structured as follows: Section 2 overviews the details of the requirements of our project, Section 3 describes our preliminary investigation, and Section 4 presents the conclusion of this work with directions for future work.

2 Overview of Legal IR Application Development

Our project focuses on the development of a fact based search application for use by legal professionals. The goal of the system is to enable a legal professional working on a case to retrieve similar cases from a large archive available to them.

2.1 Understanding user information needs

Prior work on legal search has examined the nature of typical information needs and queries used in legal search [2]. It was found in this work that the majority of queries aimed at framing an issue or learning about a particular case which is currently being worked on, for which a searcher wants to investigate particular features relevant to the case. Similarly we began our investigation by exploring the nature of queries for our legal search task. An English translation of two German input queries are shown below.[1]

Case 1: "The client works as a **personal protector**. In the course of his activity, he **crashed a photographer spectacles** with a stroke, and **added an injury to the eye**, through which the **victim was nearly blind**. The client was sentenced to a probation penalty and received the **termination without notice**. He would like to **sue for continuation of the employment situation**, because he believes to have only done his job."

Case 2: "The client has been employed by his employer since 1 April 1998. Since mid-2012 he is **suffering from depression**. As a result, there were **repeated miscarriages of different duration**. The **client leads**

[1]Translation is performed using Google Translate (`https://translate.google.com/`)

his illness back to the high psychological stress at the workplace. On 19.11.2016 he received a **illness-related termination**. The client would no longer want to work with his employer, But would still like to **stand up against the dismissal** in order to be able to **win a higher compensation**. The **client has a legal protection insurance.**"

From examination of these and other queries, we found that legal search queries often have specific phrases, concepts discussing different entities, and actions as shown in boldface in *Case 1* and *Case 2*. Generally queries exhibit rich relationships between the concepts being described.

This poses an IR challenge since the bag-of-words assumption commonly used in IR, where words are treated as independent terms disregarding the grammar and structure of the information, have their limitations, cannot capture these rich concepts and relationships between these concepts, and are thus unlikely to be able to effectively satisfy the user's information need in many cases. This led us to consider alternatives to traditional IR models which do not attempt to capture such relationships.

2.2 Development of professional legal test collection

For initial investigation in our project we used a collection of approximately 50K documents relating to legal cases provided by Wolters Kluwer. They also provided 15 user queries relating to search for similar cases. Similar to the development of test collections for public IR tasks, judging the relevance of all the documents in a collection for each query is impractical. We thus adopted a standard *pooling* strategy. The top k documents returned by a number of different IR systems were combined to form a document subset for manual relevance assessment.[2]

2.3 Relevance Judgement

Relevance in legal search is more focused and broader as compared to definition typically used in information seeking behaviour studies, which range from known-items to exploratory and investigative topics [6]. Within this work, the task of the IR application is to locate legal cases that have *legal facts* similar to those expressed in the searcher's query. Relevance is determined by the factual legal information contained in the query. Recognising that IR developers and general users were not able to make assessments of relevance, in our project legal experts from Wolters Kluwer were asked to classify the relevance of each retrieved document in the assessment pool on a scale of [0 - 2] where:
2 – indicates Relevant document, indicating that the facts in the document are topically similar to the query.
1 – indicates Partial relevant document, indicating that the facts in the document are in part/somewhat topically similar to the query.
0 – indicates Non-relevant document, indicating that the facts in the document are unrelated to the query.

The procedure followed thus far will be familiar to IR researchers. However, it was at this point in the process that we encountered new challenges.

Working with the legal experts carrying out the relevance assessments at Wolter's Kluwer, it became clear that depending on variations in the context in which the entities appear, and the relationship between them, can lead to completely different interpretation and meaning with respect to relevance and usefulness from a legal expert perspective. Not only were legal experts required to make the assessments of relevance and content usefulness, but it quickly became apparent that the retrieval challenges associated with the task are much greater than those encountered with public IR tasks. Effective IR methods for our task require a degree of nuanced semantic interpretation of the needs expressed in the queries and the contents of the documents in order to be able to make a meaningful comparison to compute am effective retrieval score.

Returning to the examples in section 2.1. For *case 1*, the query is focused on the facts relating to a "bodyguard hitting a photographer to protect his client", which has specific entities and actions. Thus a case referring to a "policeman hitting a photographer" might appear to be partially relevant from the perspective of an IR system developer, but in fact is completely non-relevant from the point of view of a legal expert. A policemen might use force and power to maintain law and order, and not be in violation of law when doing so, but a private personal protector using force and involving in a physical attack will come in strict violation of law. We would also anticipate a concept of the form "X hitting Y" will have more occurrences in the collection and might be considered as a partially relevant match from an IR developer perspective whereas "a personal protector hitting photographer" is a very particular concept that a legal expert would expect to match with cases to be deemed as relevant.

[2]Technical details and parameter settings are provided in Section 3 describing our preliminary experiments.

Similarly for example case 2, the main focus is on *illness-related termination*. A case discussing an employee who was dismissed because of a *disease related termination*, where a disabled employee who cannot work as an assembly worker anymore seemed partially relevant from an IR developer point of view, but would be deemed non-relevant by a legal expert. A general document focusing on disability related termination would seem similar to illness related termination, but would not help a legal expert o address their information need.

Based on ongoing interactions between an IR developer and a legal expert over relevance judgements, we found that without understanding what constitutes useful and relevant document from the perspective of a professional legal expert inaccurate judgements were made, which often impacted on the reliability of prediction of the quality of the performance of a retrieval model for this task.

3 Preliminary Experimental Investigation

In this section we report on our initial experimental investigation for this IR task using our test collection with standard IR models and a simple semantic retrieval model.

3.1 Retrieval Models

For the standard IR models, we compute retrieval results for three models: *TF-IDF* [9], *BM25* [8] and *Language Modelling* [7]. These models have been shown to exhibit high retrieval effectiveness over a wide range of search applications [1]. Our semantic models used distributed representation of words and documents commonly referred to as *embeddings* which aim to capture semantic similarity between content items. This has emerged as a popular area of investigation with good results for document retrieval tasks [5]. We learn embeddings for each of the documents in our collection [3]. Each of the document is embedded as a vector of 100 dimensions. The input query embedding vector is searched over the document embeddings collection to retrieve documents based on their cosine similarity. The embeddings are learnt in an unsupervised fashion using neural networks trained over the document collection [3].

3.2 Retrieval Tools

Our legal documents were supplied in a raw XML format. We preprocessed them using XML parsing to extract only the "facts" section from the full documents which were then used for these initial experiments. We used the *Lucene* toolkit[3] to perform document retrieval. Extracted content was preprocessed using stemming and stopword removal using the Lucene German Analyzer while indexing the collection, as well as while searching queries over the collection. We used the TF-IDF, Language Model and BM25 model implementations of Lucene with default parameters.

We used gensim[4] implementation of *paragraph vectors* to learn and incorporate document embeddings in our experiments for relevance prediction. We use the distributed bag of words (DBOW) algorithm with an embedding size of 100 [3].

3.3 Evaluation Measures

Our goal is to help a legal expert to identify relevant cases easily. Due to the limited amount of relevance assessment and early stage exploratory nature of our IR investigation, we focus on measuring precision of retrieved results. We evaluate Precision at rank 10 (P@10) and 20 (P@20), and count the overall number of relevant documents retrieved at rank 50.

3.4 Creation of Evaluation Set

To create the evaluation set we created a pool of 50 documents for each query for relevance assessment as described in Section 2.2 by merging the top 100 retrieval results for the TF-IDF, BM25, Language Modelling, and semantic model. The pooled documents were shared with the legal experts with guidelines for assigning relevance labels to the documents, as discussed in Section 2.3. Each document was annotated by 3 annotators on a scale of [0-2]. Table 1 presents the per query relevance judgements for the pooling exercise. The number of relevant documents found for the 15 queries varies considerably across the three annotators. To make the analysis easier, we combined partially relevant and relevant documents (referred as *Rel docs*). Based on the

[3] `https://lucene.apache.org/core/4_4_0/`
[4] `https://radimrehurek.com/gensim/`

Query-Id	Rel docs Annotator-1	Rel docs Annotator-2	Rel docs Annotator-3	Rel docs, annotators agreement	
				All 3 annotators agree	2 annotators agree
Query-1	8	14	19	6	6
Query-2	17	29	23	17	6
Query-3	20	38	36	18	18
Query-4	4	8	3	3	1
Query-5	14	19	15	8	9
Query-6	13	21	21	13	3
Query-7	1	18	12	1	10
Query-8	2	10	16	2	7
Query-9	3	17	35	3	12
Query-10	1	6	3	1	1
Query-11	1	4	12	1	2
Query-12	2	3	15	2	1
Query-13	8	33	46	8	25
Query-14	0	0	50	0	0
Query-15	11	37	39	11	25
Overall Count	105	257	345	94	126

Table 1: Pooling results

Model	Rel Docs	P@10	P@20
BM25	82	0.167	0.160
TF-IDF	84	0.147	0.153
LM	78	0.133	0.120

Table 2: Retrieval Model results

number of *Rel docs*, Query-1, 2, 3, 5, 6, 13 and 15 appear to be easier to satisfy while Query-4, 10 appear to be more difficult queries. Results for the remaining queries differ significantly across the annotators, we speculate that the reason for this is the subjective nature of assessments and different interpretation of results by the annotators.

Due to the low agreement between the annotators, we selected the judgements by annotator 1, for our preliminary evaluation of retrieval performance for the different IR models.

3.5 Experimental Results

Table 2 show results for the traditional IR retrieval models. The BM25 model showed the best results for this initial study. In terms of finding relevant documents, the system performs well for 5 queries: 1, 2, 3, 5, 6, and not so well for remaining 10 queries. We speculate that there are two possible reasons for the low performance of traditional models: i) relevance judgements are challenging in this legal IR task, ii) traditional retrieval models seem to be not so effective for this legal search task.

Using semantic models where we learn a whole document embedding and match it with the query embeddings performed poorly, with P@10 and P@20 results for all queries using the test set actually being zero. We speculate two reasons for this: i) matching query and document embedding is not an effective way of using this information, and ii) due to the shallow nature of the pooling, documents returned at top using semantic based models may not have been judged and are being treated as non-relevant, resulting in precision scores of zero.

4 Conclusions and Future work

Our experiences working on the development of an IR application for legal professionals revealed the importance of having a close, ongoing work relationship with legal professionals who are able to reliably judge the relevance of retrieved content, with whom the behaviour of the retrieval system can be discussed more generally.

An important feature of working with legal professionals or other domain experts in the evaluation of IR applications is to communicate effectively the role of evaluation in the development of effective IR technologies for specific tasks, and beyond this how relevance assessment labels should be assigned. From our experiences, it is

clear from the inconsistency of the relevance assessments that we obtained from three different legal professionals familiar with the search task at hand, that communicating this is not straightforward to domain experts not familiar with the research concepts of IR. Improving this aspect of the work is a key element of the next stage of our project which we are currently working on.

Professional search tasks can require more complex analysis of content and queries and more advanced matching algorithms than the ones usually considered in traditional IR tools. The availability of query and click logs in web applications can mean that such detailed analysis is not required in order to satisfy complex information needs in web search tasks, which essentially rely on user action based recommendations. Such user based signal information is not available for professional tasks, and researchers need to focus on the development of content analysis and matching techniques. There is a need to capture and extract concepts and their relationships in our legal queries and documents, and to incorporate this rich information in the search model.

As discussed in Section 3.5, the shallow pool based test set collection, and the initial methods being used to create sub-collection can limit the evaluation of diverse and combination models being explored at the development side. At times we need to carry out separate relevance judgement exercises at early stages of the research. We plan to look at alternative approaches for conducting pooling exercise in our future work [4], to be able to compare systems effectively.

The work described in this paper is part of an ongoing project. Our ongoing work focuses on alternative and extended retrieval models, capturing different aspects and combining information from semantic representation, bag-of-words representations, and also the incorporation of other methods including for example query expansion. Instead of learning complete document embeddings, we plan to explore sentence embeddings for each sentence in a document and use these sentence-based embeddings to retrieve relevant documents for a given query following the earlier work on passage retrieval [10].

Acknowledgements

We would like to thank Ahmed Abdelkader for helping with running some of the experiments. This research is supported by Science Foundation Ireland (SFI) and Wolters kluwer as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

References

[1] Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Volume 283. Addison-Wesley Reading (2010)

[2] Ferrer, A.S., Hernández, C.F., Boulat, P.: Legal search: foundations, evolution and next challenges. the wolters kluwer experience. Revista Democracia Digital e Governo Eletrônico **1**(10) (2014) 120–132

[3] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). (2014) 1188–1196

[4] Lipani, A., Palotti, J., Lupu, M., Piroi, F., Zuccon, G., Hanbury, A.: Fixed-cost pooling strategies based on ir evaluation measures. In: European Conference on Information Retrieval, Springer (2017) 357–368

[5] Mitra, B., Craswell, N.: Neural models for information retrieval. arXiv preprint arXiv:1705.01509 (2017)

[6] Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.: Evaluation of information retrieval for e-discovery. Artificial Intelligence and Law **18**(4) (2010) 347–386

[7] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of SIGIR 1998, ACM (1998) 275–281

[8] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. NIST special publication (500225) (1995) 109–123

[9] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24** (1988) 513–523

[10] Trotman, A., Geva, S.: Passage retrieval and other xml-retrieval tasks. In: Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology, Department of Computer Science, University of Otago (2006) 43–50