A Free Format Legal Question Answering System

Soha Khazaeli
Janardhana Punuru
Chad Morris
Sanjay Sharma
Bert Staub
Michael Cole
LexisNexis | Legal & Professional, NC, USA
firstname.lastname@lexisnexis.com

Sunny Chiu-Webster
Facebook / WA, USA
sunnycw1@gmail.com

Dhruv Sakalley
Sensibill / Ontario, CANADA
dhruv.sakalley@gmail.com

Abstract

We present an information retrieval-based question answer system to answer legal questions. The system is not limited to a predefined set of questions or patterns and uses both sparse vector search and embeddings for input to a BERT-based answer re-ranking system. A combination of general domain and legal domain data is used for training. This natural question answering system is in production and is used commercially.

1 Introduction

Question answering (QA) applications range from simple yes/no systems to complex questions where answers might be synthesized from several sources (Voorhees and Tice, 2000). This work concerns a QA system for legal research. The system is designed to answer both factoid (Agichtein et al., 2005) and non-factoid questions. A short answer can satisfy factoid questions, e.g., "What is the burden of proof for breach of contract?". In contrast, non-factoid questions are open-ended and an adequate answer needs opinions or explanations (Hashemi et al., 2020), e.g., "Why does child support increase with income?".

A typical system user is a litigator seeking answers to case-specific legal questions. Those answers inform the creation of litigation documents, such as pleadings, briefs, and motions. A system should provide complete multiple-sentence answers with context that can be cited. A legal QA system must also handle questions where no single answer exists. For example, the useful answer may be jurisdiction specific, or be time frame dependent because the law evolves. Importantly, the best answers can depend on the lawyer's perspective because application of the law can make fine distinctions per the case facts, or recognizes competing principles or mitigating factors. Legal practice areas can have distinctive concerns. For example, the scope of legal principles and practice affecting family law legal matters is distinguishable from those applied in bankruptcy law. Developing an effective and useful question answering system in this setting faces state of the art challenges, including creation of training collections and performance metrics.

We present a retrieval-based legal domain QA system designed to provide useful answers for all legal practice areas. It is designed for customers with real-world tasks while meeting cost, response time, and scalability constraints. This system is in production and serving customers. We also describe an experiment methodology found to have pragmatic value in system development. The methodology can be useful in domains with similar context-laden QA characteristics.

2 Related Work

Recent deep learning open domain QA research successfully applied a retrieve and read paradigm. The retrieval step selects candidate documents, then a reading component finds answers (Chen et al., 2017; Das et al., 2019; Yang et al., 2019). We adopted a similar approach.

Often, the systems employ standard retrieval methods based on sparse vector space approaches like TF-IDF (Jones, 1972) and BM25 (Robertson and Spärck Jones, 1994). Dense vector representations with distributional semantic properties based on LSA (Landauer et al., 1998), GLoVe (Pennington et al., 2014), and sentence embedding (Reimers and Gurevych, 2019) are also used. Custom domain embedding methods such as Legal-bert have been studied (Chalkidis et al., 2020). The QA retrieval target is usually at the passage level rather than complete documents (Luan et al., 2021). Paragraph-based legal domain search has been studied (Zhang and Steiner, 2018). Answer extraction techniques have used Machine Reading Comprehension (MRC) (Seo et al., 2016) and DrQA (Chen et al., 2017). Several ranking models have been ap-

Natural Legal Language Processing Workshop 2021, pages 107–113

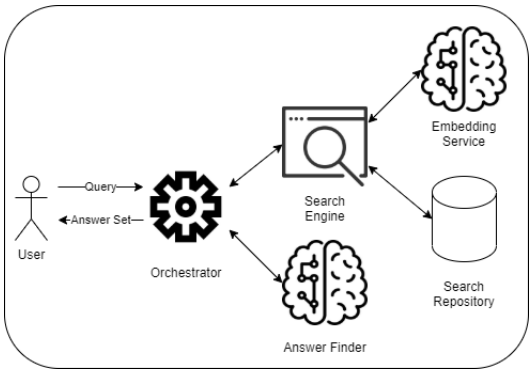

Figure 1: Two stage architecture for QA systems

plied in QA systems (Yang et al., 2019; Nogueira and Cho, 2019). One COLIEE[1] task is to answer yes/no legal questions and retrieve germane legal documents. Competitors have used TF-IDF and BM25, and also contextual embedding vectors such as BERT and ELMo, to score passages against the question (Rabelo et al., 2019). LexisNexis legacy answer cards address knowledge-based QA by detecting the user query intent and then serving a previously mined answer if appropriate (Kumar and Politi, 2019; Shankar and Buddarapu, 2018). The WestSearch QA system categorizes non-factoid questions to different frames. To answer a frame-specific question, a trained frame-specific question-answer pair classifier is used to recognize a retrieved passage as an answer (McElvain et al., 2019). Both of these commercial legal QA system can only handle a limited range of questions.

In contrast to the previous work, the system presented here is designed to answer almost all legal content questions without legal practice area restrictions. In particular, the question coverage is not limited by pattern or frame.

3 Methodology

The system selects answers by re-ranking search results obtained using both sparse vector techniques (BM25) and a dense vector approach (semantic embedding). Figure 1 shows a simplified system architecture. The *search repository* contains the passage text and passage embeddings. The *search engine* retrieves passages by text and embedding similarity. The *answer finder* re-ranks the retrieved passages.

[1]Competition on Legal Information Extraction/Entailment

3.1 Retrieving Passages

The search collection is a pool of passages that highlight key aspects of each legal document. It consists of case-law Headnotes [2] and RFCs [3]. The collection contains over 100 million passages. Our experiments were conducted with a 10% sample of the collection. The search engine has the usual goal of retrieving relevant passages. Beyond topical germaneness, the retrieval engine must also detect sufficient context in the passage. To that end, the retrieval answer set coverage is enriched using both sparse vector and semantic embedding passage representations. The system used a Query By Document method (QBD) (Yang et al., 2018), *more like this*, implemented with BM25 for sparse vector passage representation retrieval. The dense embedding enrichment used Legal GloVe and Legal Siamese BERT embeddings. The GloVe embeddings are built using 64GB of legal text with a 300K word vocabulary and 200 dimensions. Passages and questions are encoded using the average of the word embeddings.

The Siamese Legal BERT system is trained to retrieve similar passages in a contextual vector space (Reimers and Gurevych, 2019). The training data is a sample of 100,000 headnotes. Given a headnote, the most similar headnote using BM25 is identified as a positive similar passage. Five random headnotes are added as negative instances. We ensured discriminative challenge amongst the negative instances using a procedure described below. The system was trained using a regression objective function with cosine loss. The input sentence embedding uses the Legal BERT base model with mean pooling of the tokens embedding. The Legal BERT model was trained in-house from scratch with a custom legal vocabulary on the last 20 years of US case-law documents. The model is trained with $train\_batch\_size = 16$. We used Spearman and Pearson correlation upward trends as convergence indicators.

3.2 Answer Finder

The *answer finder* accepts a question passage pair and computes the probability the passage answers

[2]A LexisNexis headnote is a point of law expressed in a case written by a judge which is picked and edited by editors as a general point of law.

[3]An RFC (Reason For Citing) is an automatically extracted passage of a case which contains sentences near a document citation, such as a court case citation, that suggest the Reason(s) For Citing (RFC) (Humphrey et al., 2005).

the question. A BERT sequence binary classifier (Devlin et al., 2019) is trained on question-answer pairs. The *answer finder* input is the concatenation of question(Q) and passage(P) as "[CLS]<Q>[SEP]<P>[SEP]". *answer finder* is trained by fine tuning Legal BERT. The BERT classifier uses [CLS] representation with two fully connected layers with a final softmax layer.

The nature of legal language and the high-stakes nature of lawyer tasks require subject matter experts (SMEs) to judge legal QA system performance. The legal training data was created in-house by certified lawyers. The training dataset had over 10,000 annotated legal question answer pairs covering legal practice areas. Questions with long paragraph answers (107,089) are selected from Natural Questions (NQ) (Kwiatkowski et al., 2019) and added to training data to improve generalization.

The system was fine-tuned in two stages. First, question-answer pairs were created by selecting a random negative passage for each question. The tuned *answer finder* then predicts the probability on all negative samples. In the second stage, for each question the negative answer with the highest probability of being a good answer is selected. The goal is to increase the challenge of good answer discrimination. The validation set is real-world questions extracted from user-logs. The training hyper-parameters are: $learning\_rate : 2e-5$, $max\_seq\_length : 512$, $num\_train\_epochs : 3$, $do\_lower\_case : True$, $batch\_size = 8$. We found that examination of the first 128 word-pieces ($max\_seq\_length : 128$) doesn't significantly lower validation set accuracy, so that was used in large batch experiments. Reduced max_seq_length also improves latency performance in production.

4 Results

Table 1 shows the answer evaluation scale used by the subject matter experts (SMEs). Simple "yes/no" judgments are clearly inadequate for complex domain QA systems.

Two test sets were used: 100 and 1000 questions. The small set proved valuable for rapid development cycles. Both sets consist of 50% actual user questions with additional questions from SMEs to ensure coverage of legal research goals including content, entity, and analytic questions. The QA system presented here focuses on answering content questions and may not provide good answers for entity and analytic questions. This limitation affects the performance for the smaller test set reported in Table 3. The performance of the system using the large test set are reported in Appendix A. The search corpus was bulk embedded and queried using BM25, Siamese Legal BERT, and Legal GloVe. Similarity calculations use the Euclidean metric. The input to *answer finder* is the top 30 retrieved passages for each retrieval method. *answer finder* re-ranks the list by answer probability and selects the top 3 passages. The top 3 passages for each method were evaluated. SMEs evaluated these answer sets in a random order.

Score	Criteria
-1	so unrelated that a user will lose patience with the system ('silly')
0	off point and is not reasonably related
1	right topic but not an answer
2	partial answer
3	good answer

Table 1: Answer evaluation scale

Table 2 shows retrieved passages, system answers, and the SME evaluation for "Is an airline liable for its pilot's negligence?" BM25_MLT picked a long passage with multiple occurrences of 'airline', 'pilot', 'liable' and 'negligence'. It was judged as off point, despite some topical overlap. Legal GloVe and Legal Siamese BERT picked a semantically-similar short passage even though 'pilot' does not appear. It was judged as topical but not an answer. *answer finder* on BM25_MLT picked a good answer passage with all the question elements and actors, and also discusses the conditions in which the carrier is liable for a pilot's negligence. This passage was promoted from rank 27 of the 30 BM25_MLT passages.

The answer evaluation scale has five levels and the SMEs agreed that an acceptable answer will score a '2' or '3'. SME evaluator interrater agreement (IRR) is excellent, given the complexity of legal question answering task, with an agreement level of 84% for acceptable answers and 91% for unacceptable answers. Setting a threshold of '2' in the test set, the system performance was: F1 0.766, precision 0.827, recall: 0.713, and accuracy 0.759.

F1 and accuracy provide classifier performance metrics. The production legal research system presents the top three answers to users, so we also use ranked search results evaluation metrics

Table 2: Example question and passages (BM25_MLT: BM25 more-like-this, SL_BERT: Siamese Legal BERT, L_GloVe: Legal GloVe, AF: Answer Finder)

Is an airline liable for its pilot's negligence?	**SME label**	**Method**
A carrier would not be liable for an error of judgment of the pilot, not constituting positive negligence on his part in exercising such judgment; but liability is incurred if the pilot, by his negligent and careless conduct, has created a situation requiring the formation of a judgment and then errs in the exercise thereof.	3	AF on BM25_MLT
An airline corporation is not an insurer of the safety of its passengers. The liability of an airline corporation must be based on negligence.	1	L_GloVe, SL_BERT
Airline pilot who was accused of raping flight attendant has no tort claim against airline based upon its alleged negligent investigation of accusation, even if airline's policy of investigating sexual harassment complaints creates duty to use due care in conducting investigation,...	0	BM25_MLT

to assess the system performance with DCG (Discounted Cumulative Gain) and MRR (Mean Reciprocal Rank). In this system *answer finder* is employed as a re-ranker, but it could also be used as a threshold filter on the answers. Table 3 compares alternative combinations of system components. Promising system combinations are bolded. Production systems have practical constraints and cannot simply optimize performance without considering factors like user experience, cost, and scalability. Re-ranking retrieved passages by each retrieval method increased the average DCG, confirming the value of adding *answer finder* to the system. Another advantage of the re-ranker is the ability to combine multiple retrieval methods to improve DCG performance. Combining dense and sparse retrieved passages both increases system cost and creates scaling challenges. However, the richer representation can improve user experience because users are able to ask their question using a greater variety of words. Another advantage of using a re-ranker is the capacity to apply a threshold to help filter unrelated passages.

5 Error Analysis

Case-specific error-analysis helps to identify shortcomings in the training data coverage. Categorized DCG analysis indicates the QA system provides good answers for well-defined legal questions such as *standard of review*[4] question (e.g. What is standard of review for marital property allocation decision in Florida?) . *Single topic questions* are free format questions about specific legal issue, e.g "Can an executor compromise a claim without beneficiary consent?". The QA system performs significantly better than previous systems on single topic questions, although there remains room for improvement. Optimizing the real world performance of a legal QA system needs to consider the frequency of each question category, the space for improvement, and the question category importance judged by SMEs. We identified 16 categories to prioritize for high return on development investment. Single topic question is at the top of the list. Training data for this category will be created to improve system accuracy. Table 4 reports the average DCG@3 for seven high priority question categories.

6 Discussion and Future Work

This paper presents a new legal domain QA system that is deployed and serving customers. Legal QA systems must address a specialized domain language and provide contextualized answers in a high-stakes setting. We were able to develop a performant product using a 10% content sample and 100 questions drawn from previously seen real-world queries supplemented with expert-generated questions. This proved effective to evaluate alternative embedding methods and performance trade-offs for combination of fast retrieval system on all corpus and relatively slow re-ranking system on limited retrieved passages. The sampled collection and smaller question set enabled rapid design and test cycles. Later evaluation using the full collection and a much larger question set confirmed the

[4]The deference an appeals court will apply to a decision of a lower court http://cdn.ca9.uscourts.gov/datastore/uploads/guides/stand_of_review/I_Definitions.html

Table 3: QA metrics for methods (100 questions)(BM25_MLT: BM25 more-like-this, SL_BERT: Siamese Legal BERT, L_GloVe: Legal GloVe, AF: Answer Finder, AF 0.2: Answer Finder as an answer filter with threshold 0.2)

Method	DCG@3[a]	95% C.I.[b]	N silly[c]	Answered[d]	MRR@3[a]
BM25_MLT	4.052	-	7	100	0.411
SL_BERT	3.386	1.26	2	100	0.326
L_GloVe	2.855	1.25	7	100	0.285
AF BM25	**5.464**	1.43	7	100	**0.493**
AF SL_BERT	4.862	1.43	0	100	0.416
AF L_GloVe	4.281	1.40	7	100	0.397
AF (BM25, SL_BERT)	**5.605**	1.47	5	100	**0.483**
AF (BM25, L_GloVe)	5.502	1.47	8	100	0.481
AF (BM25, SL_BERT, L_GloVe)	5.533	1.45	6	100	0.492
AF 0.2 (BM25, SL_BERT)	**6.269**	1.52	**2**	89	**0.543**

[a] Avg for answered questions.
[b] confidence interval.
[c] Number of answers that are bad enough to impact user trust in the system.
[d] Number of questions answered.

Table 4: Seven high priority question categories on 1000 questions

Category	DCG@3[a]	Priority
Single topic questions	6.99	1
Questions about rules or statutes	5.46	2
Relationship questions[5]	7.58	3
Definitions	7.11	4
Statute of limitations[6]	9.23	5
Standard of review	11.52	6
Elements [7]	9.76	7

[a] Avg for answered questions in the category.

findings. The system is oriented to answering content questions and performs lower on entity and analytic questions. We intend to address this limitation by developing alternative approaches to handle other question types that can be combined with this content-oriented system.

Various development directions are being explored. Direct extensions of this work include use of a smaller Siamese BERT, vocabulary improvements for the models, training set enhancements targeting low performing legal areas, and training the *answer finder* component on augmented training data and smaller Legal-BERT models.

Acknowledgements

The complete production QA system includes important components not described here. It was developed by the authors and Shyjee Mathai, Aaron Pohl, Kishore Ethiraj, Randall Fee, Urvi Luhana, Jonathan Mitchall, Harris Joseph, David Borchers, Megan Bramhall, Serena Wellen, Rick McFarland, Eric Weiss, Sachin Kumar, Shashi Penumarthy, and Nagaraju Buddarapu. The evaluation team of in-house lawyers included Megan Bramhall, Elizabeth Hickey, Donna McMurry, and Brian Rouse.

References

Eugene Agichtein, Silviu Cucerzan, and Eric Brill. 2005. Analysis of factoid questions for effective relation extraction. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 567–568.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1870–1879.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-

[5] Asks about applicability of a legal concept on another one e.g. Is the continuous treatment doctrine applicable to negligence claims in Ohio?

[6] The time limit for plaintiff to file a complaint https://www.nolo.com/dictionary/statute-of-limitations-term.html

[7] Essential requirements to make a claim https://www.law.cornell.edu/wex/element e.g. What are the elements of constructive eviction?

reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

Timothy L Humphrey, Xin Allan Lu, Afsar Parhizgar, Salahuddin Ahmed, James S Wiltshire Jr, John T Morelock, Joseph P Harmon, Spiro G Collias, and Paul Zhang. 2005. Automated system and method for generating reasons that a court case is cited. US Patent 6,856,988.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Sachin Kumar and Regina Politi. 2019. Understanding user query intent and target terms in legal domain. In *International Conference on Applications of Natural Language to Information Systems*, pages 41–53. Springer.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the ACL*, 7:453–466.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the ACL*, 9:329–345.

Gayle McElvain, George Sanchez, Sean Matthews, Don Teo, Filippo Pompili, and Tonya Custis. 2019. Westsearch plus: A non-factoid question-answering system for the legal domain. In *Proceedings of the 42nd International ACM SIGIR Conference*, pages 1361–1364.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. A summary of the coliee 2019 competition. In *JSAI International Symposium on Artificial Intelligence*, pages 34–49. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Stephen E Robertson and Karen Spärck Jones. 1994. Simple, proven approaches to text retrieval. Technical report, University of Cambridge, Computer Laboratory.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Arunprasath Shankar and Venkata Nagaraju Buddarapu. 2018. Deep ensemble learning for legal query understanding. In *Proceedings of CIKM 2018 Workshop on Legal Data Analytics and Mining (LeDAM 2018), CEUR-WS. org*.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference*, pages 200–207.

Eugene Yang, David D Lewis, Ophir Frieder, David A Grossman, and Roman Yurchak. 2018. Retrieval and richness when querying by document. In *DESIRES*, pages 68–75.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the NAACL (Demonstrations)*, pages 72–77.

Paul Zhang and David Steiner. 2018. Systems and methods for paragraph-based document searching. US Patent 10,002,196.

A Some Analysis on the Production System

The production QA system is evaluated periodically on the larger 1000 question set. These question collections were created by legal domain experts to be representative of various practice areas and different question types. Each experiment requires evaluation of tens of thousands of question-answer pairs. Table 5 shows higher DCG compared to the internal evaluation because the production system is based on the complete content set. Some

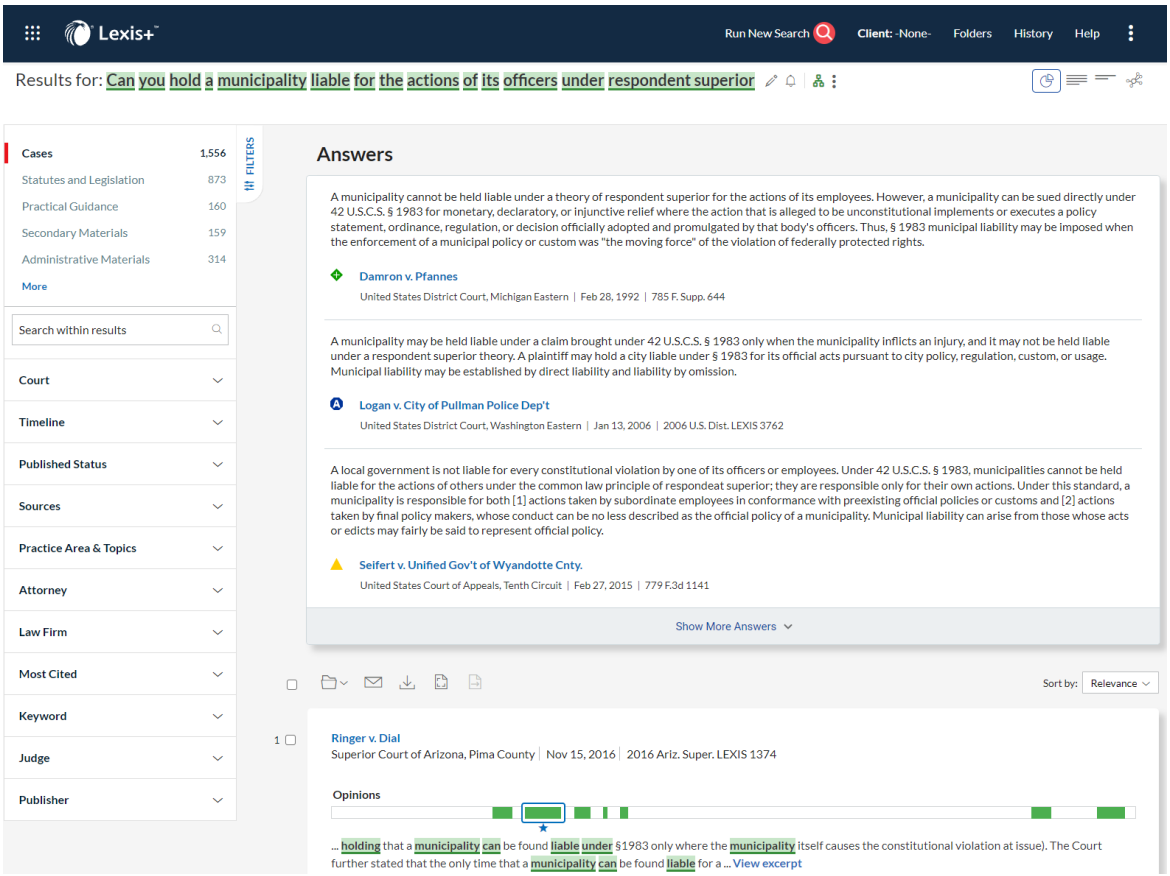

Figure 2: Customer facing application using the discussed components

additional filters are also implemented. Nonetheless, the method performance ranking is unchanged as compared to that using the smaller test collection and question sets. This confirms the usefulness of work with the smaller sets for development and tuning.

We also investigated answer variations for retrieval methods. Table 6 shows the distribution of surfaced answers, i.e. presented in the product UI, and customer-engaged answers by retrieval query method. Both retrieval methods provide passages recognized by the system as worthy answers. If an answer was retrieved by both retrieval methods there is a higher probability of user engagement.

B QA System in Production

Figure 2 shows a screen-shot of a product UI for the production QA system.

Table 5: Production system evaluation (1000 questions) (April 2020)

Method	**Avg DCG@3**
BM25_MLT	4.82
SL_BERT	3.62
AF on BM25	6.27
AF on SL_BERT	5.18
AF on BM25 + SL_BERT	6.28
AF on BM25 + L_GloVe	6.18
AF 0.3 on BM25 + SL_BERT	7.75

Table 6: Answer distribution by retrieval methods (May 2021)

Answers	**BM25_MLT**	**Si_L_Bert**	**Both**
surfaced	58.85%	31.53%	9.62%
engaged	56.48%	29.02%	14.50%