Open Challenges in the Application of Dense Retrieval for Case Law Search

Pan Du, Hawre Hosseini*, George Sanchez, Filippo Pompili
Thomson Reuters Labs
hawre.hosseini@tr.com

1. Introduction

Dense retrieval (DR) has gained remarkable success in general-purpose search over the past few years and a considerable body of literature has been produced. We are exploring the feasibility of a DR based search engine for case law search on a large scale. In this presentation, we draw upon some open questions that have also been studied within the broader literature; however, they still need to be further researched within the context of case law search. We hope this will draw attention of interested researchers to explore those questions.

2. Case law retrieval and legal research

Case law. It is the body of judicial decisions that establish precedent within a given jurisdiction. In a common law system, case law is of primary importance in determining what the law is and how it applies to any particular set of facts. Legal Research. Lawyers spend significant amounts of time searching in this body of law, find relevant cases to an issue or situation, and analyze its legal bearing based on precedents — this process is referred to as Legal Research.

Case Law Retrieval Characteristics. Considering the legal research process and its expected outcome, the nuances and performance evaluation of case law search system can be different from a general-purpose search engine:

1. It is a cost-constrained high-recall task.
2. Complex document structure. There is a web of documents with various levels of importance and each document with multiple segments and associated meta-data to be searched over.
3. Complex relevance aspects. There are several relevance aspects beyond textual similarity including precedence, facts of the query, the points of law, whether the case discusses law that is not reversed or under challenge, etc.

3. Anatomy of case documents

Facts about case documents:

• Fact 1. Contain multiple segments and meta-data;
• Fact 2. Exhibit semantic cross-dependencies due to argument-like narrative;
• Fact 3. Have unique, complex, language;
• Fact 4. Greater average document length with extreme cases; an average of 2000 words.

4. Document segmentation and label propagation for case documents

Long documents in DR frameworks. Dealing with long documents in DR frameworks has been one of its major drawbacks due to the length constraints of its underlying pretrained language models. Literature has approached this through different methodologies including:

1. as a whole and with one label;
2. through chunking into passages and transferring the document labels to its passages.

For case law documents, it sounds reasonable to segment documents to prepare for DR framework due to the excessive length of these documents and the logical sections that they contain.

Label propagation issues:

1. The gap between document and passage labels is wider for case law document due to the heterogeneity in their topical composition;
2. Label propagation is difficult due to contextual information.

Segmentation issues:

1. Case documents have logical sections, hence requiring special segmentation strategies.
2. There are long-range semantic dependencies due to argument-like structure of case documents that span longer than typical passage lengths.

5. Learning relevance from case law search user interactions logs

Extracting training signals from user logs is particularly challenging for case law because of severe click noise and domain-specific biases in legal search.

Severe click noise:

1. Cases frequently are first collected exhaustively and then narrowed down;
2. Definition of relevance for a case is complex and multi-faceted, and cannot be easily captured by short snippets from ranking lists;
3. There are multiple interaction types beyond just clicks.

Expert-knowledge bias:

1. Users willingly skip top-ranked relevant cases due to prior knowledge of them: either from personal experience, or from other searches on the same topic;
2. The induced bias is different than position, trust, popularity, or selection biases.

6. Takeaways

There are several takeaways that can introduce interesting research questions to be explored:

1. Topic-aware segmentation might be beneficial as it helps in preserving topical coherence within segments;
2. Better solutions for label propagation need to be aware of the topical composition of passages in a case law document;
3. Section- and content-aware score aggregation can be very beneficial for the retrieval process;
4. Utility of approaches for logs interactions de-noising needs to be examined;
5. Correlation of interaction types with relevance is not straightforward;
6. Expert knowledge bias; it's specific to the legal domain and it requires special modeling, different than other studied biases such as position, selection, popularity, and trust bias.

Contact us

Come work with us!
Thomson Reuters Labs