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

THOMSON REUTERS

Pan Du, Hawre Hosseini*, George Sanchez, Filippo Pompili

ThomsonReuters 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.

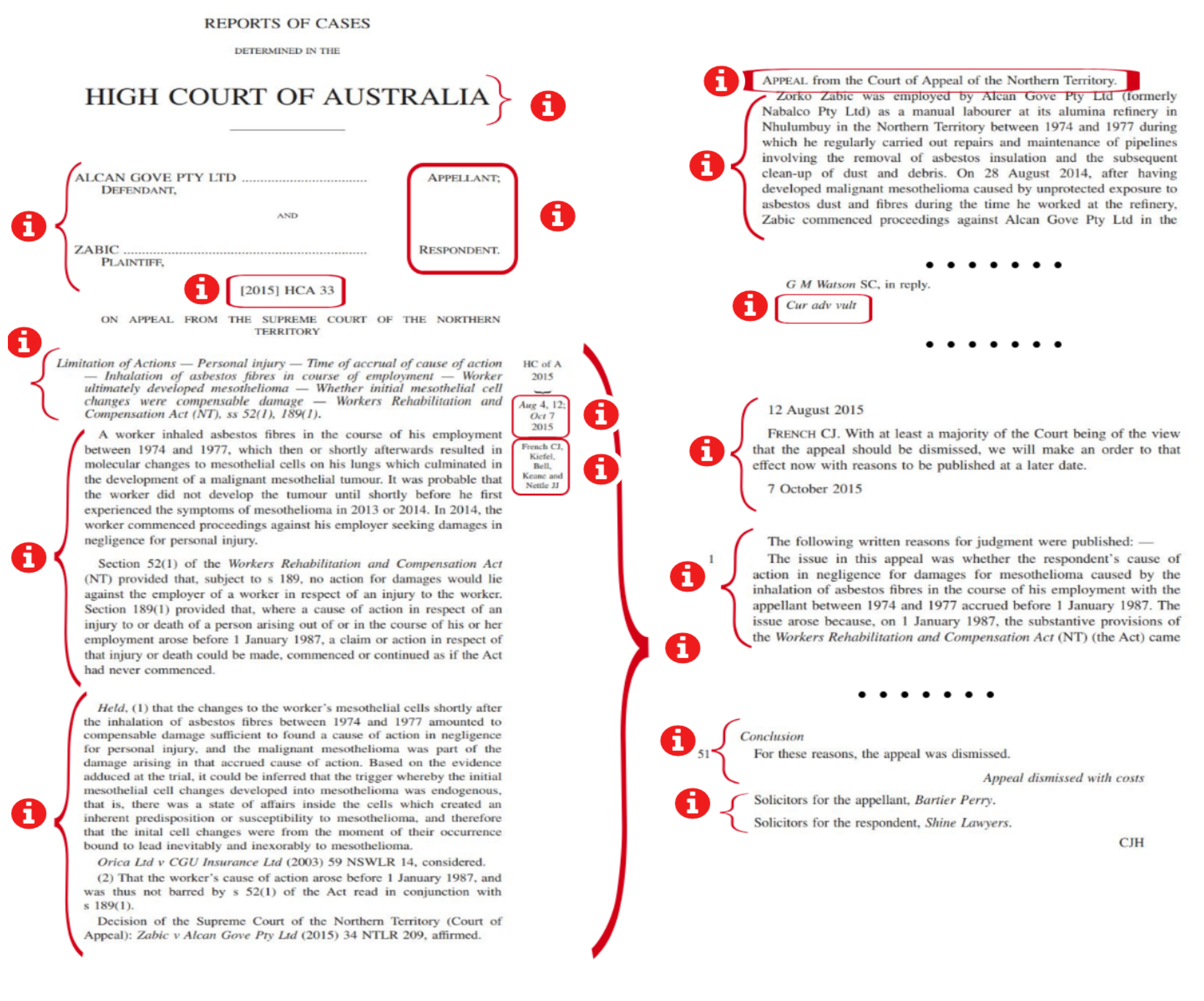
Case Law Retrieval Characteristics. Considering the legal research process and its expected outcome, the nuances and performance evaluation of case law search system can be different from a general-purpose search engine:

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.



REPORTS OF CASES

DETERMINED IN THE

HIGH COURT OF AUSTRALIA

ALCAN GOVE PTY LTD .... APPELLANT;
DEFENDANT,

AND

ZABIC .... RESPONDENT.
PLAINTIFF,

[2015] HCA 33

ON APPEAL FROM THE SUPREME COURT OF THE NORTHERN TERRITORY

Limitation of Actions — Personal injury — Time of accrual of cause of action — Inhalation of asbestos fibres in course of employment — Worker ultimately developed mesothelioma — Whether initial mesothelial cell changes were compensable damage — Workers Rehabilitation and Compensation Act (NT), ss 52(1), 189(1).

HC of A
2015

Aug 4, 12; Oct 7
2015

French CJ, Kiefel, Bell, Keane and Nettle JJ

A worker inhaled asbestos fibres in the course of his employment between 1974 and 1977, which then or shortly afterwards resulted in molecular changes to mesothelial cells on his lungs which culminated in the development of a malignant mesothelial tumour. It was probable that the worker did not develop the tumour until shortly before he first experienced the symptoms of mesothelioma in 2013 or 2014. In 2014, the worker commenced proceedings against his employer seeking damages in negligence for personal injury.

Section 52(1) of the *Workers Rehabilitation and Compensation Act* (NT) provided that, subject to s 189, no action for damages would lie against the employer of a worker in respect of an injury to the worker. Section 189(1) provided that, where a cause of action in respect of an injury to or death of a person arising out of or in the course of his or her employment arose before 1 January 1987, a claim or action in respect of that injury or death could be made, commenced or continued as if the Act had never commenced.

Held, (1) that the changes to the worker's mesothelial cells shortly after the inhalation of asbestos fibres between 1974 and 1977 amounted to compensable damage sufficient to found a cause of action in negligence for personal injury, and the malignant mesothelioma was part of the damage arising in that accrued cause of action. Based on the evidence adduced at the trial, it could be inferred that the trigger whereby the initial mesothelial cell changes developed into mesothelioma was endogenous, that is, there was a state of affairs inside the cells which created an inherent predisposition or susceptibility to mesothelioma, and therefore that the inital cell changes were from the moment of their occurrence bound to lead inevitably and inexorably to mesothelioma.

Orica Ltd v CGU Insurance Ltd (2003) 59 NSWLR 14, considered.

(2) That the worker's cause of action arose before 1 January 1987, and was thus not barred by s 52(1) of the Act read in conjunction with s 189(1).

Decision of the Supreme Court of the Northern Territory (Court of Appeal): *Zabic v Alcan Gove Pty Ltd* (2015) 34 NTLR 209, affirmed.

APPEAL from the Court of Appeal of the Northern Territory.

Zorko Zabic was employed by Alcan Gove Pty Ltd (formerly Nabalco Pty Ltd) as a manual labourer at its alumina refinery in Nhulumbuy in the Northern Territory between 1974 and 1977 during which he regularly carried out repairs and maintenance of pipelines involving the removal of asbestos insulation and the subsequent clean-up of dust and debris. On 28 August 2014, after having developed malignant mesothelioma caused by unprotected exposure to asbestos dust and fibres during the time he worked at the refinery, Zabic commenced proceedings against Alcan Gove Pty Ltd in the

....

G M Watson SC, in reply.

Cur adv vult

....

12 August 2015

French CJ. With at least a majority of the Court being of the view that the appeal should be dismissed, we will make an order to that effect now with reasons to be published at a later date.

7 October 2015

The following written reasons for judgment were published: —

1 The issue in this appeal was whether the respondent's cause of action in negligence for damages for mesothelioma caused by the inhalation of asbestos fibres in the course of his employment with the appellant between 1974 and 1977 accrued before 1 January 1987. The issue arose because, on 1 January 1987, the substantive provisions of the *Workers Rehabilitation and Compensation Act* (NT) (the Act) came

....

Conclusion

51 For these reasons, the appeal was dismissed.

Appeal dismissed with costs

Solicitors for the appellant, *Bartier Perry*.

Solicitors for the respondent, *Shine Lawyers*.

CJH

Illustration from https://libguides.csu.edu.au/

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
Toronto, Zug, London, Gdańsk, Bangalore, and Minneapolis-St. Paul.