Enhancing Legal Discovery with Linguistic Processing

Daniel G. Bobrow, Tracy H. King, and Lawrence C. Lee
Palo Alto Research Center Inc.
www.parc.com/nlp

Introduction

The U.S. Federal Rules of Civil Procedure have greatly increased the importance of understanding the content within very large collections of electronically stored information. Traditional search methods using full-text indexing and Boolean keyword queries are often inadequate for e-discovery. They typically return too many results (low precision) or require tightly defined queries that miss critical documents (low recall.) Linguistic processing offers a solution to increase both the precision and recall of e-discovery applications. We discuss four issues in legal discovery that can be enhanced with linguistic processing: improving recall for characterization, improving precision for search, protection of sensitive information, and scalability.

Characterization: Recall

Especially in the initial stages of trial preparation, attorneys need to be able to retrieve all of the information in a collection that is relevant to some characterization of interest. These characterizations depend on the legal strategy and so need to be able to be quickly and flexibly formulated. The most natural way to describe such content is in natural language and not in heavily formalized regular expression languages. Linguistic processing on the query can help generate rules in a higher level language much closer to natural language.

Two basic linguistic tools to aid in query generation for characterization are morphological analysis and ontological information. For example, morphological analysis of the term 'buy' in a query will produce 'buy', 'buying', and 'bought'. The more abbreviated and elliptical texts found in email documents can be treated similarly. For example, common email abbreviations like 'mtg' can be run through a type of morphological analysis to match against 'mtg', 'meeting', and 'meetings'. Using a disjunction of all these forms in the search increases recall, which returns both more relevant documents and more passages with examples from which to produce novel queries. Ontologies, both domain specific and general, automatically produce synonyms ('buy'='purchase') and hypernyms (a boy is type of child is a type of human) which can be used to expand the sample query into alternatives, again allowing for greater recall at the initial stages of the characterization task. During this initial step, where recall is important and the entire information collection is being culled, linguistic processing is only being done on the queries, while the search over the information is done with more standard search techniques. This allows massive information collections to be quickly processed more rapidly and thoroughly.

Search: Precision

An important aspect of legal discovery is finding information that answer specific questions or that say specific things. By automatically processing the texts into more normalized, deep semantic structures and then indexing these structures into a large database optimized for

semantic search, queries over the information collection can be made in natural language. These linguistic structures normalize away from the vagaries of natural language sentences, encoding the underlying meaning. At the simplest level, surface forms of words are stemmed to their dictionary entry and synonyms and hypernyms are inserted. However, the linguistic processing can go much deeper, normalizing different syntactic constructions so that expressions which mean the same thing have the same linguistic structure. As a simple example, 'Mr. Smith bought 4000 shares of common stock.' and '4000 shares of common stock were bought by Mr. Smith' will be mapped to the same structure and indexed identically. Thus the creation of this semantically based index of information stores a normalized but highly detailed version of the content in the information and includes links back to the original passages in the information.

The queries against the information collection are similarly automatically processed into semantic representations at query time, and these semantic representations are used to query the database for relevant documents. Unlike more standard search techniques, using the deeper semantic structures allows for greater precision and hence fewer irrelevant documents to review. The linguistic structures encode the relations between entities and actions (e.g., who did what when) so that only documents describing entities in the desired relations are retrieved. For example, standard search techniques would retrieve both 'X hit Y' and 'Y hit X' from a search on the entities X and Y and the 'hit' relation since all of the relevant items are mentioned. However, when searching for evidence in a massive information collection, it is important to return only the text passages which refer to the intended relationship among the entities.

Redaction

E-discovery increases in complexity when issues of confidentiality are considered. Over the past several years we have been researching intelligent document security solutions, initially focusing on redaction. This line of research involves building better tools to detect sensitive material in documents, especially entities and sensitive relations between entities, determining whether inferences can be made even when sensitive passages have been redacted, and providing efficient encryption techniques to allow content-driven access control.

The detection of sensitive material works on the same underlying technology described above for enhancing recall and precision. The use of stemming, synonyms and hypernyms, and automatic alias production increase recall, allowing for a single search to retrieve entities in many surface forms. The structural normalization provided by the deep processing similarly allows for better relation and context detection. As an additional part of the content discovery for redaction, our current research examines ways to allow for collaborative work on the same document collection so that knowledge discovery workers can benefit from each other's work and so that experts can help hone the skills of novices. Another component of the project involves using the Web and other large information collections to determine whether the identity of entities can be detected even when they have been redacted. For example, removing someone's name but leaving their birthdate, sex, and zip code may uniquely identify them, thereby suggesting that further material needs to be redacted.

Once the sensitive text passages have been identified, we provide tools for encrypting document passages and assigning keys so that different users can have access to different types of redacted material. This makes it possible for the document to be viewed in different ways by different people: some may have access to the whole document, some may not be able to see anything related to entity X, and some may only be able to see publicly available material. This encryption capability can either be used actively on the electronic versions of the documents or can be used to prepare specially redacted versions for printing and shipping to different parties.

Scalability

As the average number of documents involved in each legal discovery process increases, scalability is an important issue for any technology used in the process. The linguistic processing that we advocate here is more computationally intensive than shallower methods such as keyword search or basic regular expression pattern matching over plain text. To surmount this issue, we use faster processes to go from, for example, 100 million documents to a few million documents; these faster processes may be facilitated by some linguistic processing, e.g. stemming of words so that more matches on basic keyword searches are found. Once the original information collection is reduced to a more manageable load, then the slower but more accurate linguistically-enhanced processes can be used to prune to a few hundred thousand. We have evidence that this deeper linguistic processing will scale to hundreds of thousands of documents, with processing time approaching one second per sentence. Once this initial linguistic processing is done, then the resulting indexed documents can be used repeatedly in the applications described above, thereby creating a resource to be shared across the discovery processes.

Conclusion

There are a number of benefits from using linguistic processing in e-discovery applications. Linguistic processing can provide fast and flexible characterization of large information collections in pre-trial preparation, as well as enable high precision search and confidential information access in discovery. While linguistic processing is more computationally intensive than keyword search, the technology does scale well to large information collections and can also be used in combination with standard search approaches to improve the management and discovery of electronically stored information.