An Intelligent Hybrid Approach for Improving Recall in Electronic Discovery

Eniafe Festus Ayetiran

CIRSFID, University of Bologna, Bologna-Italy

eniafe.ayetiran2@unibo.it

Abstract. In this work, we propose a hybrid method for improving recall in electronic discovery proceedings. This approach takes ideas from Natural Language Processing (Word sense disambiguation) and Information Retrieval in enhancing retrieval of responsive documents using the semantics of query terms instead of direct text matching. Preliminary results from disambiguation of user queries show that this approach is promising to improve recall at the same time maintaining high degree of precision in the retrieval of relevant documents to help lawyers and their clients during litigations.

Keywords: eDiscovery, artificial intelligence, information retrieval, natural language processing

1 Introduction and General Background of the Study

There have been studies as early as the 1950s comparing automated methods for classification of documents [3]. eDiscovery is an emerging problem domain that calls for solutions provided from two separate disciplines: Law and Information Systems [3]. The term eDiscovery refers to electronically stored information (ESI) sought by an opposing party during litigation [2], is an important area that poses difficulties for lawyers, litigants and the entire court all alike. Discovering and producing required document(s) among huge volume of data created and stored electronically in various formats in repositories is a big challenge which needs to be addressed. It can be viewed as a form of legal research, which is the process of identifying and retrieving information necessary to support legal decision-making. For many years, lawyers and their clients have relied upon manual and physical methods for retrieving and providing requested documentation during litigations.

At present, the process is commonly carried out mostly through the use traditional technologies such as keyword searching to speed up the process due to advent and subsequent ubiquitous use information systems.

Recently, there have been a lot of research efforts in Machine Learning to improve the present situation. Machine learning is a branch of artificial intelligence which concerns the construction and study of systems that can learn from data and using the knowledge learned on some other new data.

For instance, the 3 emerging Artificial Intelligence techniques for eDiscovery proposed by [5] all of which fell in the line of Machine Learning. These techniques include: (1) Machine learning to extend and apply theories of relevance (2) Generalizing relevance theories with a hypothesis ontology (3) Social network analysis to apply relevance theories

Here we present a proposal which attempts produce a novel approach to eDiscovery by combining techniques from Natural Language Processing and traditional Information Retrieval in overcoming the problems in the existing methods. Natural Language Processing (NLP), a field classified under Artificial Intelligence and Linguistics. NLP enables computers to derive meaning from human or natural language. The idea is to learn from the user queries to improve recall and high degree of precision yet economical.

2 Electronic Discovery

Electronic Data Discovery or eDiscovery is any process (or series of processes) in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case [10]. eDiscovery, born on April 12, 2006 as a result of the approved amendments to the Federal Rules of Civil Procedure by The United States Supreme Court governing the discovery of electronically stored information (ESI). These amendments took effect on December 1, 2006. It has been the major decisive factor in many cases. According to a report by Socha and Gelbmann [11], the consensus among legal consumers is that 60% of today's cases warrant some form of eDiscovery activity. This percentage will continue to grow over the next several years. Regarding EDD content, according to Corporate Counsel, at least 50% of eDiscovery documents will be in the form of e-mail, with another large chunk coming in the form of office documents (e.g Word, spreadsheets, etc.), together with small databases (e.g MS Access) or larger databases (e.g Oracle), as well as less conventional forms of digitized data (e.g., software code) or other forms (e.g voice mail or video clips) [1]. Today eDiscovery has spread to different parts of the world including Australia, United Kingdom (eDisclosure) and parts of Asia.

2.1 Electronic Discovery and Information Retrieval

eDiscovery is a form of information retrieval. In any Information Retrieval system there is always a trade-off between precision and recall. eDiscovery is a recall-centred task because under production and over production of responsive may have effects on the litigation process as there have been several cases where these situations have been penalized. Although the legal community is familiar with key word search, which historically has been the foundation of case law and statutes searching, standard key word search alone is inadequate for obtaining complete, high recall solutions. There is a wide spectrum of eDiscovery software and service providers today, many that rely on conventional IR techniques, while others harness alternative technologies such as machine learning or concept search along with more standard techniques

2.2 Critical Problems in Existing/ State-of-the-art Approaches and Motivation for Research.

Keyword search, which uses direct text matching between query terms and terms in the document collection, does not provide an intelligent search approach that can cater for the requirements of eDiscovery as the search results includes too many false hits in terms of irrelevant documents. This is because the two foundational issues which arise when searching in an unstructured information domain has not been addressed. The first is the synonym problem – words having the same meaning. The second problem is known as "polysemy," - many words having more than one meaning [9]. Synonyms and polysemies are two factors that reduce the power and accuracy of information retrieval systems. Hence the present generic tools cannot be effectively used to discover relevant documents electronically. Hence, there is need for more intelligent approach.

Machine Learning, an intelligent approach provides a good search that can cater for the requirements of the present day eDiscovery by training a system on a set of data and applying it to new set of data to predict an outcome. One major concern about using Machine Learning is how to get a wide coverage of data enough to cover reasonable level for a problem like eDiscovery may be an almost impossible task knowing the fact testing a Machine Learning system on an entirely different data domain for which it has not been trained may lead to poor results. The big issue is what can be done about this since discovery documents cut across all areas of human endeavour and not limited to a particular domain.

Therefore, we see this as more of a human language problem and propose an intelligent system which learns from the user queries may be a better approach. Computing the actual meaning of each query terms used in context can greatly help improve the overall retrieval process.

2.3 Research Questions

The research questions to be addressed are as follow:

1. How can we conduct an intelligent search and improve recall with only the user query?
2. How do we produce a scalable system to handle large volume of documents usually involved in eDiscovery?
3. How we handle the heterogeneous nature of document formats within the document collection indexing and retrieval?

3 Research Methodology

We present in Figure 3.1 below the proposed architecture of the eDiscovery system.

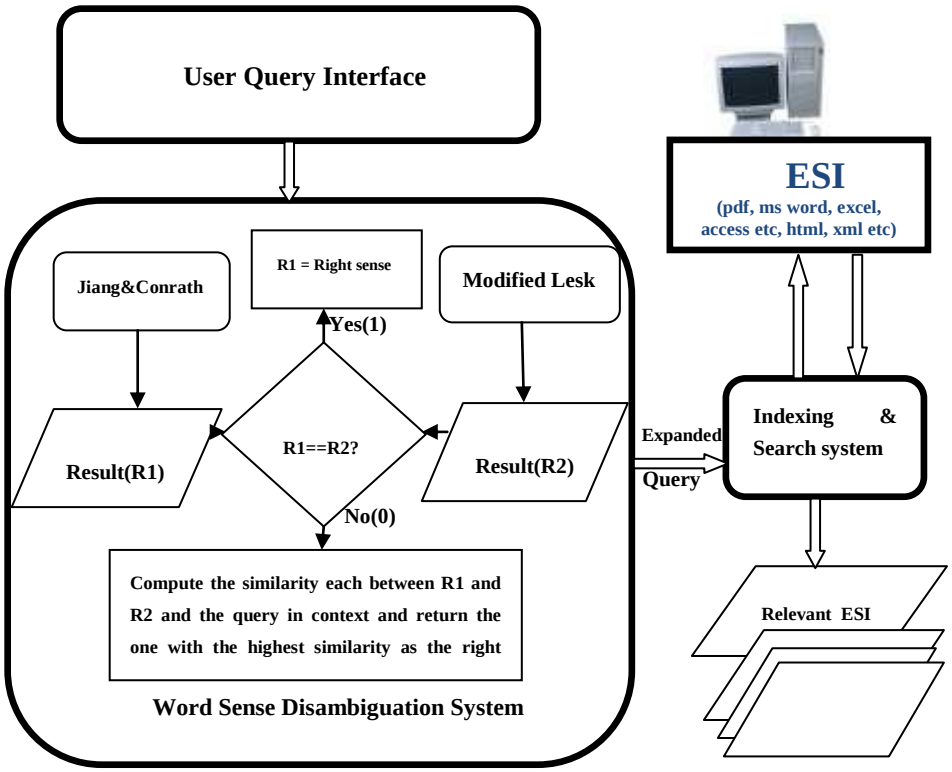

Fig. 1. General Architecture of the Proposed eDiscovery System

Below is an outline of the methodology as depicted in Figure 3.1 above:

a. Sense disambiguation of user query.
b. Expansion of query with semantically related terms to the query terms
c. Development of format-independent indexing and search system using vector space classification
d. Classification and retrieval of responsive documents by the indexing and search system using the expanded query

The whole idea is to compute the meaning of each query terms using word sense disambiguation techniques. The disambiguation will lead to the production of other semantically related words to each of the query terms. The query terms and their semantically related words will serve as input to the indexing and search system which will then classify the documents and subsequently retrieve the responsive documents. The indexing and search system is an accumulation of various technologies that can handle documents of several formats as each format have their own characteristics and tools to index them.

3.1 Word Sense Disambiguation

Ambiguity is a fundamental characteristic of every language of which the English language is not an exception. A considerable number of English words have more than one meaning. The meaning of word intended by a particular user can be inferred considering the context of usage.

For example: (a) *I have a permit to stay in the lodge* (b) *A permit was brought from for dinner preparation*. Based on the context of the usage of the word, *permit,* in the two sentences above, we can infer that the first instance (sentence (a)) is referring to a legal document or an authority to do something and the second instance (sentence (b)) is referring to a large game fish found in the waters of the west Indies. However, human identification of the right word sense is relatively simple compared to machines which need to process large unstructured textual information, carrying out complex computations in order to determine the sense of a word used in a particular context.

The computational identification of meaning of words in context is called Word Sense Disambiguation (WSD) also known as Lexical Disambiguation. Considering the instances in the examples above, the sentences can be sense-tagged as follows: (a) I have the permit/authority/license to stay in the lodge (b) A permit/fish was caught in the Indian Ocean. Basically, the output of any word sense disambiguation system with the right synonymous word (if any). Word Sense Disambiguation relies on knowledge. This means, it uses a knowledge source or knowledge sources to associate the most appropriate senses with words in context. Ideally, Word Sense Disambiguation is a means to an end but not usually the end itself, enhancing others tasks in different fields and application development such as parsing, semantic interpretation,

machine translation, information retrieval and extraction, text mining, and lexical knowledge acquisition. “Polysemy” means to have multiple meanings. It is an intrinsic property of words (in isolation from text), whereas “ambiguity” is a property of text. Whenever there is uncertainty as to the meaning that a speaker or writer intends, there is ambiguity. So, polysemy indicates only potential ambiguity, and context works to remove ambiguity.

In our approach, we have employed a method of inter-technical cross validation of two widely used techniques in the field leveraging on their strengths. These algorithms are the Modified Lesk algorithm – a modified version of the original Lesk algorithm and the Jian & Conrath algorithm. Both algorithms are forms of knowledge-based approach based to WSD.

3.1.1 The original Lesk Algorithm.

A basic knowledge-based approach relies on the calculation of the word overlap between the sense definitions of two or more target words. This approach is named *gloss overlap* or the *Lesk* algorithm after its author [6]. It is one of the first algorithms developed for the semantic disambiguation of all words in unrestricted text. The only resource required by the algorithm is a set of dictionary entries, one for each possible word sense, and knowledge about the immediate context where the sense disambiguation is performed. The idea behind the Lesk algorithm represents the starting seed for today’s corpus-based algorithms. Almost every supervised WSD system relies one way or another on some form of contextual overlap, with the overlap being typically measured between the context of an ambiguous word and contexts specific to various meanings of that word, as learned from previously annotated data.

The main idea behind the original definition of the algorithm is to disambiguate words by finding the overlap among their sense definitions. Namely, given two words, W1 and W2, each with NW1 and NW2 senses defined in a dictionary, for each possible sense pair W1i and W2 j, i = 1…...NW1, j = 1…..NW2, we first determine the overlap of the corresponding definitions by counting the number of words they have in common. Next, the sense pair with the highest overlap is selected, and therefore a sense is assigned to each word in the initial word pair. The Algorithm is summarized in Listing 2.1 below:

1. *for each sense i of* W_1
2. *for each sense j of* W_2
3. *compute Overlap(i,j), the number of words in common between the definitions of sense i and sense j*
4. *find i and j for which Overlap(i,j) is maximized*
5. *assign sense i to W1 and sense j to W2*

Listing 3.1: The Original Lesk Algorithm

3.1.2 Jiang & Conrath Algorithm

Jiang & Conrath propose a combined model that is derived from the edge-based notion by adding the information content as a decision factor. The model is based on the lexical taxonomy of the lexicon and statistics in the information content. In particular, attention is given to the determination of the link strength of an edge that links a parent node to a child node. Jiang and Conrath [4] (Equation 3.1) uses the difference in the information content of the two concepts to indicate their similarity. He used the information content defined by Resnik[8] and augmented it with notion of path length between concepts. This approach includes the information content of the concepts themselves along with the information content of their lowest subsumer.

$$\text{Similarity} = 2 \times \text{IC}(\text{LCS}(C_1, C_2)) - \text{IC}(C_1) + \text{IC}(C_2) \quad (3.1)$$

Where IC is the information content, LCS is the lowest common subsume, C_1 and C_2 are the concepts under consideration

3.1.3 Inter-technical Cross Validation Algorithm

Our technique has been derived from the two algorithms discussed above using WordNet [7] as the knowledge resource. We have modified the original Lesk algorithm adopting WordNet lexical and semantic taxonomy and direct implementation of the Jiang & Conrath algorithm using all the words in context as the window size. In the Modified Lesk implementation, we have not considered the glosses of only the target word and that of their surrounding neighbours, but also that of their semantically related ones in the WordNet taxonomy and these include the hypernyms, hyponyms, meronyms, antonyms etc. We then cross validate the results produced by both Modified Lesk and the Jiang and Conrath algorithms with query terms in context. The main idea is that the glosses of the right sense and that of their semantically related ones in the WordNet hierarchy should be similar as much as possible with the query. The process starts by tokenizing the query with each term in the query as a token and tagging the terms into their part of speech based on the usage in the query. That is for a set of terms, Ti ϵ Qi , where Qi is the query, tag T ϵ Ti into their part of speech based on the usage in the query. For monosemous terms, return the sense accordingly. For polysemous tokens, obtain the synsets from the WordNet with the sense definitions, the lemma names, semantic relations i.e hypernyms, hyponyms, meronyms, etc and examples. We consider the sense definitions of each synset with their associated lemma names, their glosses, glosses of their hypernyms, hyponyms, meronyms etc. We compute the initial score based on the overlap of terms in the gloss of the target word, its hypernyms, hyponyms etc, and that of each of the surrounding words. The overall score for each senses of a term is obtained by summing the all the initial scores with other words in the window size (in this case, all the terms in the sentence). We chose the sense with the highest score as the appropriate sense for the Modified Lesk algorithm.

In the same manner, we compute initial semantic similarity scores for the target word in the query with each of the terms in the query using Jiang & Conrath method. Compute final semantic similarity scores for the target word from the addition of all initial semantic similarity scores. Again, we chose the sense with the highest final semantic similarity score as the appropriate sense for the Jiang & Conrath method.

Finally, we then compare the senses returned by Modified Lesk and Jiang & Conrath algorithms for agreement. We chose the sense for which they agreed as the right sense, otherwise where they disagree, we compute score based on the overlap of their glosses, that of their hypernyms, hyponyms etc with the original query in consideration. The sense with highest score between the two senses is selected as the right sense.

3.2 Vector Space Model

Our aim is to classify documents in the collection or repository into relevant (responsive) and irrelevant (non-responsive) and retrieve the relevant once based on a determined threshold in the weighting and scoring of terms in the expanded query terms and terms in the document collection.

The representation of a set of documents as vectors in a common vector space is known as the vector space model and is fundamental to a host of information retrieval operations ranging from scoring documents on a query, document classification and document clustering. In a typical setting we have a collection of documents each represented by a vector, a free text query represented by a vector, and a positive integer K. We seek the K documents of the collection with the highest vector space scores on the given query.

3.3 Innovation of Research Methodology

Why *Disambiguation, Expanded Query, Indexing and Retrieval Instead of Directly Using Latent Semantic Indexing?*

Latent Semantic Indexing [9] is a method for automatic indexing and retrieval taking into account the issues of synonyms and polysemies. The approach is to take advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents on the basis of terms found in queries. The particular technique used is singular-value decomposition (SVD), in which a large term by document matrix is decomposed into a set of ca. 100 orthogonal factors from which the original matrix can be approximated by linear combination. However, the computational cost of the SVD is significant; LSI works best in applications where there is little overlap between queries and documents. Also, it is most suitable where small number of documents are involved.

Hence, it is not suitable for eDiscovery where we have to deal with large volume of data.

Furthermore, the original LSI works with clustering but not with statistical/probabilistic techniques (classification) used for scoring and ranking in information retrieval. eDiscovery is purely a classification rather clustering hence direct implementation of LSI for this type of problem may not be a suitable idea.

Finally, implementing the solution through a method of disambiguation, query expansion, indexing and scoring documents for retrieval brings about the solution to scalability problem while also taking into account the problems of polysemy and synonyms.

4 Preliminary Results and Discussion

We implemented the inter-technical cross validation algorithm and evaluated with the Semeval 2007 coarse-grained English All-words dataset. The result produced 76.516% accuracy (F1 score). The results from this will be used to expand the query which will serve as input to the indexing and retrieval system.

5 Conclusion

With this high performance result of semantic determination of query terms, we believe is a good performance result that will positively enhance the entire retrieval system. In the preceding phase of this research, we hope to effectively adopt the results as an expended query to the indexing and retrieval system using the techniques we discussed previously.

References

1. Counsel C: The American Bar Association (ABA), section of litigation, committee on Corporate Counsel. http://www.abanet.org/litigation/committees/corporate/((2006)
2. D. Oard, B. Hedin, S. Tomlinson, J. R. Baron, "Overview of the TREC 2008 Legal Track." TREC Conference 2008, Proceedings
3. H.S. Hyman, "Designing a Text Mining Artifact for eDiscovery Solutions." Working Paper (2010)
4. Jiang, Jian & David Conrath.: Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics*, Taipei, Taiwan.(1997)
5. Kelvin. D. Ashley & Will Bridewell: Emerging AI & Law Approaches to Automating Analysis and Retrieval of Electronically Stored Information in Discovery Proceedings. In: ICAIL 2009 Global eDiscovery/eDisclosure Workshop (2009)

6. Lesk, Michael.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: *Proceedings of the ACM-SIGDOC Conference*, Toronto, Canada, 24–26 (1986)
7. Miller, George.: Wordnet: A lexical database. *Communications of the ACM*, 38(11): 39–41.(1995)
8. Resnik, Philip. 1995. Using information content to evaluate semantic similarity. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, 448
9. S. Deerwester, S. T. Dumais, G.W. Furnas, T. K. Landauer, R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*. (Sep. 1990)
10. SearchFinancialSecurity.com: Definitions: Electronic Discovery. http://searchfinancialsecurity.techtarget.com (2009)
11. Socha G, Gelbmann T: The 2006 Socha-Gelbmann electronic discovery survey report. Socha Consulting LLC and Gelbmann & Associates, MN ((2006)