Cross Document Co-Reference Resolution Applications for People in the Legal Domain

Christopher Dozier
Research and Development
Thomson Legal and Regulatory
610 Opperman Drive
Eagan, MN 55123,USA
chris.dozier@thomson.com

Thomas Zielund
Research and Development
Thomson Legal and Regulatory
610 Opperman Drive
Eagan, MN 55123,USA
tom.zielund@thomson.com

Abstract

By combining information extraction and record linkage techniques, we have created a repository of references to attorneys, judges, and expert witnesses across a broad range of text sources. These text sources include news, caselaw, law reviews, Medline abstracts, and legal briefs among others. We briefly describe our cross document co-reference resolution algorithm and discuss applications these resolved references enable. Among these applications is one that shows summaries of relationships chains between individuals based on their document co-occurrence and cross document co-references.

1 Introduction

Attorneys, judges, and expert witnesses all play important roles in legal systems. Judges decide cases. Attorneys handle the legal needs of clients. Expert witnesses testify about complex facts and play an ever-increasing part in the settlement of cases. An important part of an attorney's preparation for litigation involves researching the background of the judge deciding the case, of attorneys representing the opposing side, and of testifying experts. To help attorneys with this research need, we have created a system that automatically links across documents references to attorneys, judges, and expert witnesses. These documents include news articles, caselaw documents, law reviews, Medline abstracts, and legal briefs among others.

Our method of creating cross document co-references involves extracting from text MUC type templates for individuals and matching the templates to biographical profile records using a Bayesian based record linkage technique. We have described this method in detail elsewhere (Dozier and Haschart, 2000; Dozier et al., 2003) and briefly describe it in section 2.

The biographical records for attorneys, judges, and expert witnesses that we use for cross document co-reference resolution have been created through a combination of automatic and manual techniques. The basis of the biographical records for attorneys and judges comes from a manually created professional directory which is itself a product called the Westlaw Legal Directory. The biographical records for the expert witnesses were created through text mining. We have described the text mining application for creation of the expert witness database in (Dozier er al., 2003) and describe it briefly in sections 3 and 4.

A new application we have created from these cross document references involves creating summaries of relationships between attorneys, judges, and expert witnesses. This new application is discussed in section 5.

Since their deployment, these applications have created automatically over 7 million links between references to attorneys, judges, and expert witnesses in various document collections and their respective biographical profiles.

2 Cross Document Co-reference Resolution through Extraction and Linking

By combining information extraction techniques with record linkage techniques, we have been able to resolve cross document references for attorneys, judges, and expert witnesses. Our basic technique involves extracting a template record for an individual from a text document and matching the template record to an authority file record. Figure 1 depicts how extraction and record linkage are combined to create cross document links.

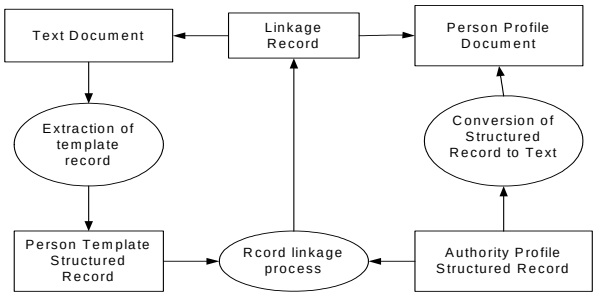

Figure 1: Cross Document Coreference Resolution Process

The extraction portion of our system is similar to template extraction systems described in the Message Understanding Conferences proceedings (MUC-6, 1995) and elsewhere (Appelt et al.,1993) (Grishman,1997). Our extraction process relies on a finite state machine that identifies paragraphs in a document containing attorney, judge, or expert names and a semantic parser that extracts from the paragraphs template information about each individual named.

The record linkage portion of our system uses a Bayesian network to match and link attorney, judge, and expert templates to biographical records. This network computes the probability that a given biographical record matches the same person specified in an extracted template. To compute this match likelihood, we treat first name, middle name, last name, firm, city, state, court, and other information as independent pieces of match evidence. We compute the prior probability of a match by calculating the probability that a randomly selected biographical record will match a template. We then compute conditional match probabilities for each piece of evidence using a manually tagged training set. For each piece of evidence, we compute the conditional probability that a biographical record matches a template when that piece of evidence matches exactly, matches in a strong fuzzy way, matches in a weak fuzzy way, is unknown, or mismatches. We define what we mean by strong fuzzy and weak fuzzy in (Dozier and Haschart, 2000) and (Dozier et al., 2003). But basically we mean that a piece of match evidence matches in a fuzzy way when it is compatible with another piece of evidence but does not match exactly.

We compute the match probability score for the records using the following form of Bayes' rule:

$$P(M \mid E) = \frac{P(M)\prod_i P(E_i \mid M)}{P(M)\prod_i P(E_i \mid M) + P(\neg M)\prod_i P(E_i \mid \neg M)}$$

$P(M|E)$ is the probability that the template and authority records refer to the same person, given match evidence E.

$P(M)$ is the prior probability that the reference records refer to the same person. $P(\neg M)$ is the prior probability that records do not refer to the same person.

$P(E_i|M)$ is the conditional probability that the match variable E_i takes on a particular value given that the template and authority records match. For example, if we let E_1 stand for middle name match evidence, the probability that the middle names in the records match exactly given that the records themselves match is $P(E_1=exact|M)$.

$P(E_i|\neg M)$ is the conditional probability that E_i takes on a particular value, given that the template and authority records do not match. For example, if we let E_1 stand for middle name match evidence, the probability that the middle names in the reference records match, given that the records themselves do not match, is $P(E_1=exact|\neg M)$.

A sample caselaw document with highlighted links to judges and attorneys is shown in figure 2. Figure 3 shows caselaw documents linked to the biographical record for attorney Gerry Spence.

3 Mining Authority Records From Text

For expert witnesses, we created an expert witness authority file of some 100,000 profiles, mined from approximately 300,000 jury verdict and settlement documents, using publicly available professional license information, an expertise taxonomy, and automatic text mining techniques. This directory can be browsed by area of expertise as well as by location and name. The profiles are automatically linked to Medline abstracts, as well as back to the relevant jury verdict and settlement documents. To the best of our knowledge, this is the largest expert witness directory of its kind and the first to be built using automatic text mining techniques.

Figure 4 shows the text mining process we used to create our authority file of expert witnesses from jury verdict and settlement documents.

Figure 2: Caselaw Document. Underscored Names are Hyperlinked to Profiles.

Figure 3: Gerry Spence Authority Record with Coreference Links. Listed in Left-hand Panel are Cases that Reference Gerry Spence.

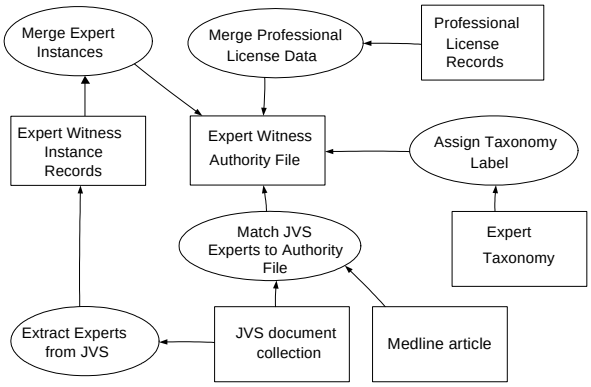

Figure 4: Text Mining of Expert Witness Profiles

First, we extract references to expert witnesses from court proceeding documents. In our initial implementation of the system, we extracted 290,000 reference records to expert witnesses from 300,000 documents. The method of extraction for these reference records is essentially the same as the method described in section 2 for template records. Reference records in fact are template records for experts.

Second, we merge the expert witness reference records together to create an expert witness profile file in which each particular expert is listed only once. Section 4 describes the merging process in more detail.

Third, we add professional license information into the expert witness profile records. These records include license information from the Drug Enforcement Agency, which licenses health care professionals to prescribe drugs, and from various other professional licensing agencies. To determine whether a license record and expert record refer to the same person, we again apply Bayesian based record linkage. The evidence we use to match license records to profiles includes first name, middle name, last name, name suffix, city-state information, area of expertise, and name rarity.

Fourth, we automatically assign each expert witness record one or more classification categories in an expertise taxonomy.

Finally, we link court documents and Medline abstracts to expert witness profile records using Bayesian based record linkage for a third and fourth time.

Figure 5 below shows a jury verdict and settlement document. Note that the reference to expert witness oncologist Arthur Ablin is highlighted. Figure 6 shows the authority file record for Arthur Ablin to which the hypertext link in the jury verdict and settlement document is linked. Figure 7 shows the cases Dr. Ablin has testified in as well as Medline articles Dr. Ablin has authored.

4 Merging Reference Records to Create Authority File

To create our directory of expert profiles (i.e., expert authority file) from expert reference records, we need to create a set of records that list a particular expert one and only one time. This means we need to merge together all the reference records that pertain to each individual.

As a first step in merging the expert reference records, we group the reference records into sets in which each record in a group shares a common last name. By doing this, we avoid the computational cost of comparing every expert witness reference record to every other reference record.

To merge expert references together within the groups, we use the following greedy algorithm:

1. Select an unmerged expert reference record from the group. Create an expert authority record from this record. Mark the expert reference record as merged.

2. Compare the new expert authority record to each unmerged reference record in the group. In each comparison, compute the probability that the expert in the authority record refers to the same individual referenced in the reference record. Use Bayesian matching to compute this match probability. If the match probability exceeds a match threshold, mark the reference record as “merged”. Note that the match threshold probability is determined empirically from training data.

3. If any unmerged records remain in the group, then return to step 1 else halt.

Note that, at this stage, it is still possible for there to be duplicate records in the authority file, if two or more reference records pertain to the same individual but have variant last name spellings. While this is not common, it can happen when a source document contains a misspelled last name. To address this situation, we make a final pass over the merged authority file and flag record pairs for manual review when the last names of two records are separated from each other by an edit distance of two or less and when the records match in all other respects.

5 Relationship Summaries

Using the links between documents and entities such as judges, attorneys, and expert witnesses, we are creating a new set of products that automatically summarize relevant relationships among these entities. For example, by following co-

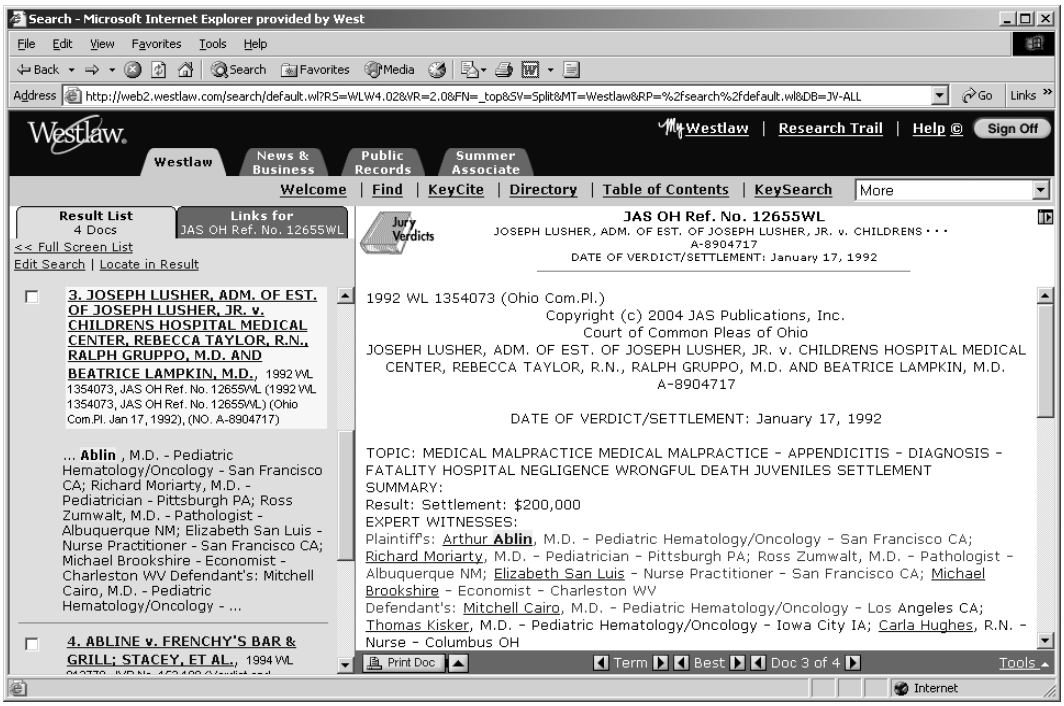

Figure 5: Jury Verdict and Settlement Document with Hyper Links to Attorneys, Judges, and Expert Witnesses including Arthur Ablin.

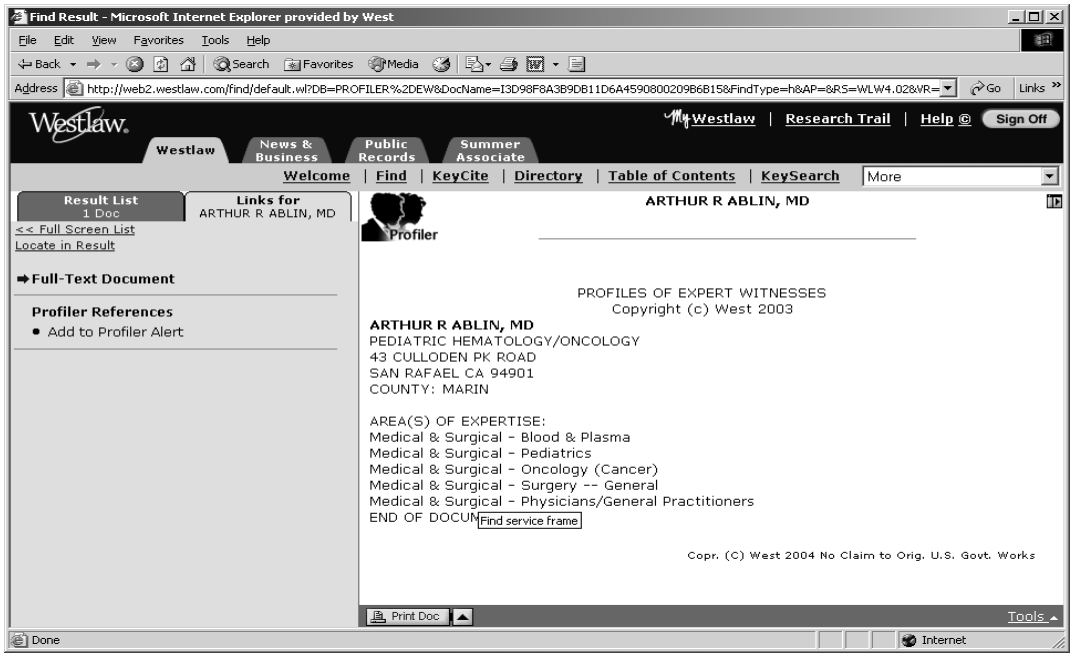

Figure 6: Profile of Oncology Expert Arthur Ablin That Was Created Through Text Mining.

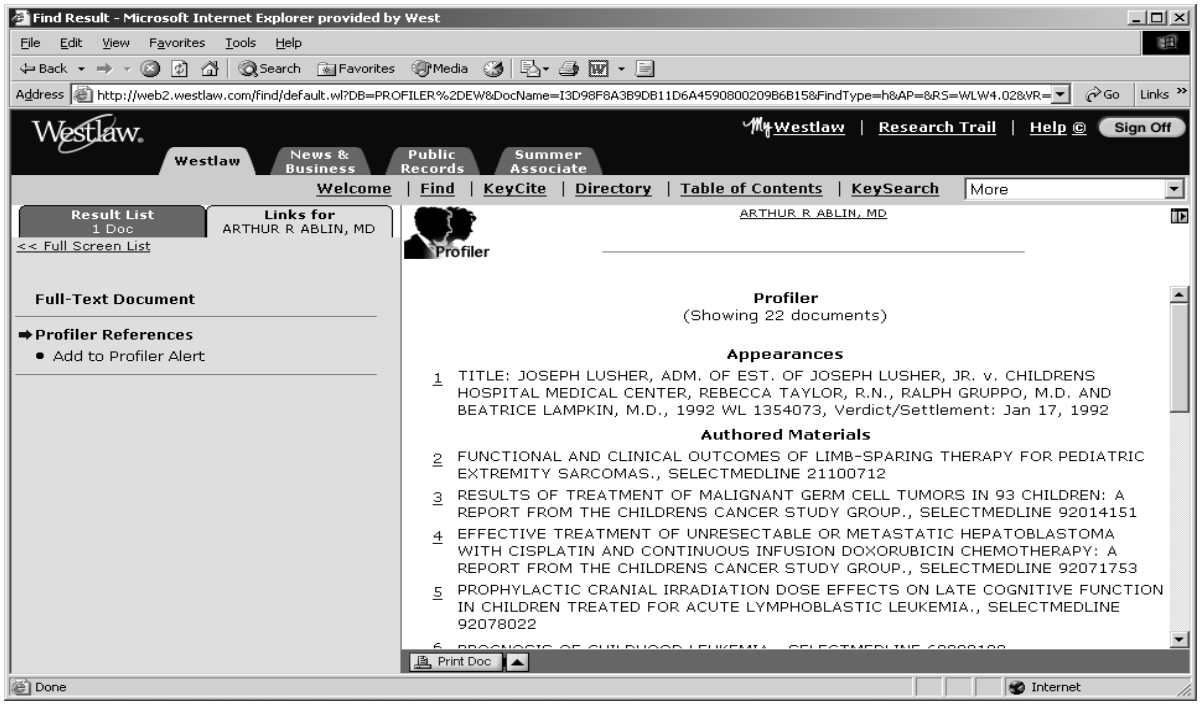

Figure 7: Documents Pertaining to Arthur Ablin.

reference and co-occurrence links, automatic reports can be generated that show which attorneys have hired which expert witnesses, which attorneys have worked together, which attorneys have appeared before which judges, which experts have co-authored papers, and so on.

Figures 8, 9, and 10 show a summary of the relationships between U.S. Chief Justice William H. Rehnquist and some other attorneys and judges who have been identified in documents with him. The summary shows some of the close working relationships one would expect to find for the chief justice. Note, however, that these relationships were discovered through cross-document co-reference, not through an editorial process.

6 Discussion

Our cross document co-reference resolution method works by attaching a person referenced in a document to a profile record that stands for a unique real world person outside the frame of the document. Person references from multiple documents that attach to the same profile are resolved by this common attachment.

In the case of attorneys and judges, we used an existing directory of attorneys and judges as the source of our profile records. In the case of experts, we created profiles by mining references from highly structured and trustworthy documents such as jury verdict and settlement documents and professional license records.

When more than one person in a single document can be attached to individual profiles and the relationship between the references in the document can be also extracted, then a relationship record can also be created outside the frame of the document and these relationships can then be chained together to generate relational inferences. This is the technology we are currently exploring within the legal domain and beyond.

There is a popular belief that between any two people, there are around six degrees of separation or less. Although (Dodds et al., 2003) and (Travers and Milgram, 1969) provide some empirical evidence in support of this theory, the popularity of the theory spawns more from a popular trivia game in which you attempt to find a path to Kevin Bacon from any Hollywood star through co-starring roles (see oracleofbacon.org for an on-line demo).

Using the database of links built from cross document co-reference resolution on legal documents, a similar technique could be used to find paths of co-occurrence between arbitrary pairs of U. S. legal professionals. Although we do not picture any popular party games developing from this, such techniques could prove useful for activities such as detecting conflicts of interest among people working on a legal matter.

Judge Summary

- **Name:** William H. Rehnquist
- **From:** United States Supreme Court
- **Location:** Washington, DC

Links from Other WestLaw Content

- **Case Law:** 847
- **Oral Arguments:** 10

Figure 8: Link Summary Statistics for William H. Rehnquist

1 Judges Appearing with William H. Rehnquist

Name	Court	Location	Co-Appearances
Antonin Scalia	United States Supreme Court	Washington, DC	782
David H. Souter	United States Supreme Court	Washington, DC	769
John Paul Stevens	United States Supreme Court	Washington, DC	690
Anthony M. Kennedy	United States Supreme Court	Washington, DC	682
Clarence Thomas	United States Supreme Court	Washington, DC	621
Sandra Day O'Connor	United States Supreme Court	Washington, DC	596
Ruth Bader Ginsburg	United States Supreme Court	Washington, DC	534
Stephen G. Breyer	United States Supreme Court	Washington, DC	471
Byron R. White	United States Supreme Court	Washington, DC	145
Thurgood Marshall	United States Supreme Court	Washington, DC	48

10 most frequently co-appearing of 27 records displayed.

Figure 9: Judges Appearing with Justice William H. Rehnquist

Attorneys Appearing with William H. Rehnquist

Name	Firm	1.1 Location	Co-Appearances
Edwin S. Kneedler	Justice Dept. Solicitor General	Washington, DC	40
Theodore B. Olson	U.S. Department Of Justice	Washington, DC	36
Michael R. Dreeben	Justice Dept. Solicitor General	Washington, DC	34
Lawrence G. Wallace	Justice Dept. Solicitor General	Washington, DC	26
Kent L. Jones	Justice Dept. Solicitor General	Washington, DC	25
James A. Feldman	Justice Dept. Solicitor General	Washington, DC	19
Warren Price, III	Price, Okamoto, Himeno & Lum	Honolulu, HI	18
Charles M. Oberly, III	Oberly & Jennings, P.A.	Wilmington, DE	18
Kenneth W. Starr	Kirkland & Ellis Llp	Washington, DC	16
Malcolm L. Stewart	Justice Dept. Solicitor General	Washington, DC	16

10 most frequently co-appearing of 1792 records displayed.

Figure 10: Attorneys Appearing with Justice William H. Rehnquist

We have described the precision and recall of the programs that create cross document co-references for attorneys and judges in (Dozier and Haschart, 2000) and that create co-references for experts in (Dozier et al., 2003). In general, for collections in which documents reference people within stereotypical syntax, attach location and job type information to the person, and usually include the person's full first name (e.g., attorney names and expert witness names in caselaw), we typically achieve better than 0.98 precision and better than 0.95 recall. For collections in which documents do not give full name or do not have highly stereotypical syntax, the precision and recall performance is worse. For example, when tagging expert witness references in Medline, where an author's first initial is given in place of the full form of the first name and the author's job type must be inferred from the topic of the scientific article, we currently achieve precision of better than 0.95 and recall of around 0.60.

7 Conclusion

By combining information extraction and record linkage techniques, we have created a repository of references to attorneys, judges, and expert witnesses across a broad range of text sources. These text sources include news, caselaw, law reviews, Medline abstracts, and legal briefs among others. We briefly describe our cross document co-reference resolution algorithm and discuss applications these resolved references enable. Among these applications is one that shows summaries of relationships chains between individuals based on their document co-occurrence and cross document co-references

References

Appelt, D. E., Hobbs, J. E., Bear, J., Israel, D., Tyson, M. (1993). FASTUS: A Finite-State Processor for Information Extraction from Real-World Text. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1172-1178.

Dodds, P., Muhammed, R. & Watts, D. An Experimental Study of Search in Global Social Networks. Science, Vol 301, Issue 5634, 8 August 2003, pages 827-829.

Dozier, C. and Haschart, R., (2000) Automatic Extraction and Linking of Person Names in Legal Text. In *Proceedings of RIAO-2000 (Recherche d'Informations Assistee par Ordinator)*, pages 1305-1321, Paris, France.

Dozier, C., Jackson. P., Guo, X., Chaudhary, M., and Arumainayagam, Y. (2003) Creation of an Expert Witness Database Through Text Mining. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*., pages 177-184, Edinburgh, Scotland, UK.

Grishman, R. (1997). Information Extraction: Techniques and Challenges. In Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology : International Summer School, Frascati, pages 13-27, Spring Verlag, Frascati, Italy.

Travers, Jeffrey and Milgram, Stanley (1969). An Experimental Study of the Small World Problem. *Sociometry*, Volume 32, Issue 4 (Dec., 1969), pages 425-443.