Title
CLAS: Civil Lawsuit Analysis System

Author
Shapiro, Max

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

CLAS: Civil Lawsuit Analysis System

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Computer Science

by

Max Shapiro

2015

ABSTRACT OF THE THESIS

CLAS: Civil Lawsuit Analysis System

by

Max Shapiro

Master of Science in Computer Science

University of California, Los Angeles, 2015

Professor Michael Dyer, Chair

Antiquated technology plagues the legal industry. While the legal market generates billions of dollars of revenue annually, the technology that supports it is decentralized, inconsistent and outdated. With the purpose of promoting modernization and illustrating how technology could improve existing inefficiencies in legal research, this thesis evaluates the creation and execution of an automated lawsuit classification system called the Civil Lawsuit Analysis System ("CLAS"). CLAS classifies the nature of a lawsuit based on information contained in a subset of attorney-authored court documents called "Memorandum of Points and Authorities" that are filed with trial courts during the process of the lawsuit. In light of the fact that no similar automated classification system exists today, developing the process by which CLAS operates proved to be a collaborative and time-intensive process that ultimately required specialized knowledge in both the computer science and the legal fields.

The thesis of Max Shapiro is approved.

Todd Millstein

Carlo Zaniolo

Michael Dyer, Committee Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Antiquated technology plagues the legal industry. While the legal market generates billions of dollars of revenue annually, the technology that supports it is decentralized, inconsistent and outdated. Unlike many other industries that have been revolutionized through the integration of transformative, intelligent technologies, the legal sector has not. As a result, significant inefficiencies exist when trying to locate legal information or run meaningful analytics on legal data. Additionally, outdated technologies enable a general lack of transparency. This lack of advancement creates opportunity, particularly with regard to the vast abundance of legal data that currently sits stagnant, segregated, and unanalyzed in public court documents.

With the purpose of promoting modernization and illustrating how technology could improve existing inefficiencies in legal research, this thesis evaluates the creation and execution of an automated lawsuit classification system called the Civil Lawsuit Analysis System ("CLAS"). CLAS classifies the nature of a lawsuit based on information contained in a subset of attorney-authored court documents called "Memorandum of Points and Authorities"[1] that are filed with trial courts during the process of the lawsuit. After being processed by CLAS each Memorandum of Points and Authorities is assigned a "nature-of-suit" classification that categorizes the nature of the underlying dispute that gave rise to the lawsuit.

CLAS's automated classification of a "nature-of-suit" offers a mechanism to improve current legal search engines and streamline the retrieval of relevant legal information by allowing users to narrow their search results to only those documents that relate to a specific type of dispute. For example CLAS could be used limit document retrieval to only those documents filed in a

[1] CLAS focuses on extracting, analyzing, and interpreting data from "Memorandum of Points and Authorities" because, unlike many other case documents, these specific documents detail the underlying legal research and legal support used to bolster the specific legal arguments presented in a lawsuit.

medical malpractice lawsuit. Or, CLAS's nature-of-suit classification could be coupled with existing document title search mechanisms and thereby allow users to limit search results to a show only a specific type of document filed in a specific type of lawsuit (*e.g.*, retrieve only those "Memoranda of Points and Authorities in support of a Motion to Dismiss" (the documents' title) that were filed in a medical malpractice lawsuit).

A. Components of CLAS

CLAS operates through three components: a document preprocessing component, a "Key Phrase Selection" component, and a "Nature-of-Suit Classifier."

1. The document preprocessing component standardizes and extracts the information contained in the Memorandum of Points and Authorities;
2. The "Key Phrase Selection" component identifies and extracts sentences within the Memorandum of Points and Authorities that provide information about the nature of the lawsuit.
3. The "Nature-of-Suit Classifier" component implements a machine learning algorithm to interpret the Key Phrases and, based on its Key Phrase interpretations, assign a nature of suit to the Memorandum of Points and Authorities.

In light of the fact that no similar automated classification system exists today,[2] developing the process by which CLAS operates proved to be a collaborative and time-intensive process that ultimately required specialized knowledge in both the computer science and the legal fields.

B. Challenges Faced in Constructing CLAS

There were three major challenges in developing CLAS: (1) acquiring an adequate number of electronic court documents that contain machine readable text, (2) understanding what characteristics of a lawsuit contain information relevant to the nature of the lawsuit, and (3) automatically identifying and extracting Key Phrases that contain these characteristics.

[2] Still today lawsuits are manually classified and annotated, usually by either the filing attorneys or by attorneys employed by for-profit companies that offer commercial libraries of legal documents.

Because the Nature-of-Suit Classifier (CLAS's third component) relies on a machine learning algorithm, before CLAS could be evaluated on a set of test documents, the algorithm first needed to be trained. This is an expected prerequisite in any system that relies on a learning algorithm to interpret data. However, accessing an adequate number of electronic court documents that contain machine readable text was an unexpected hurdle. Although court documents are public information, they are not readily accessible. Very few court documents are stored in electronic formats and even when they are, most of the time it is not in a format that contains machine readable text. This is why CLAS's first component (document preprocessing) was developed and necessary. In this vein, the legal landscape itself complicated the development of CLAS by making the acquisition of enough raw data on which to train the algorithm difficult.

Another key challenge in developing CLAS centered on crafting the Key Phrase Selection component (CLAS's second component). This required numerous trial and errors and significant attorney collaboration to even understand what characteristics a Key Phrase must have in order to contain information relevant to the nature of the lawsuit. Once these characteristics were defined, CLAS was designed to automatically identify and extract these characteristics. The Key Phrase Selection component identifies Key Phrases by analyzing the grammatical relationships in every sentence and extracting only those sentences that:

1. Reference Persons, Organizations, and Other Proper Nouns (collectively called "Entities" throughout this paper)
2. Reference at least one party in the Lawsuit; and
3. Describe an interaction between two Entities that occurred in the past.

Thus, CLAS requires its Key Phrase Selection component to progress through a multi-step analysis focused on ensuring that each requirement is met before a sentence is identified as a Key Phrase.

One of the three characteristics requires Key Phrases to reference at least one party in the suit (*e.g.*, a plaintiff or a defendant). The Key Phrase Selection component requires a reference to

a party because the underlying relationship between the plaintiff(s) and defendant(s) drives the nature of the legal claims. Thus, references to a plaintiff or defendant helps locate sentences that may provide data about the nature of the lawsuit. In the legal world this is an "obvious" legal relationship; however, it is not so apparent to someone without specialized legal knowledge. Likewise, the technical value of the plaintiff/defendant relationship (i.e. its ability to direct CLAS to valuable data) is not readily apparent to attorneys who lack specialized knowledge about how to build information retrieval systems.

The challenges faced in creating CLAS highlight the industry-specific barriers to the successful implementation of disruptive legal technologies. The legal sector is inherently (and purposefully) decentralized and specialized, causing legal data to be systematically inconsistent and housed in a market that is very poorly understood by outsiders. Ultimately, this thesis was only possible as a result of significant collaboration with (and continued testing and feedback from) practicing attorneys. This collaboration not only exposed the value in automating the classification of a lawsuit, it was essential to assist in locating and understanding the documents and the data.

Overall, while significant opportunities exists for technology to improve many inefficiencies in the legal sector, it is an area riddled with industry-specific challenges that cannot go overlooked if such improvements are to be successfully introduced to the legal market. The complexities and inconsistencies of the legal data demand application of intelligent and complex technologies grounded in a firm understanding of the legal market. Thus, a valuable and viable solution, requires a team with knowledge of both law and computer science.

II. THE LEGAL LANDSCAPE: THE IMPORTANCE OF PRIOR LAWSUITS IN LEGAL REASONING

Broadly speaking, the legal industry offers services to evaluate whether certain behaviors are compliant with relevant laws. Two central components are required to do this: (1) identify relevant laws and (2) formulate predictive arguments or advisory opinions about the application of those laws. While the first component requires a general baseline knowledge about the existence of certain laws, the second component requires the more specialized ability to analyze how the law may be applied (or not applied). In order to reach reliable predictions or arguments about how a law may apply, lawyers reason by analogy, researching and analyzing prior similar applications or analyses of the law.[3] Given the importance of reviewing similar prior precedents to draft predictive assessments and persuasive arguments, the legal industry invests significant time and money creating systems and methods to facilitate lawyer's ability to locate and retrieve prior applications of law.

C. Lawyers use Prior Judicial Rulings to form Reliable Predictive Arguments

Within the legal system, courts function to clarify the application of a law through the issuing of judicial opinions. Lawyers who represent clients in formal lawsuits (called litigators) attempt to persuade courts to apply the law in certain ways. Through a series of written and verbal arguments, litigators summarize relevant prior cases for the court and draw similarities or distinguish the facts of their case from the facts of the prior case. Judges review these arguments and then reach a final conclusion (called a judicial opinion or a ruling) about how to apply (or not apply) the law to the facts of the case before them.

[3] While there is no exhaustive list of what may be relevant or useful, common sources of legal information include: judicial rulings from prior lawsuits, attorney-authored documents filed with a court in a prior lawsuit, transcripts of prior legislative discussions or negotiations that occurred during the enactment or revisions of the law, analyses contained in academic law journals.

If a judge sits in a high enough court then his or her ruling will bind judges in lower courts in that jurisdiction to apply the law in the same manner in future cases.[4] In other words, lower courts with "inferior jurisdiction" must accept the application of the law declared by courts of "superior jurisdiction." A judicial opinion that must be followed or applied is known as "binding precedent", "case law", or "binding case law". These prior judicial opinions establish the law going forward.

However, even if not binding case law, all prior judicial rulings are considered persuasive or advisory and are valuable to provide insight as to how the law may apply. For example, because trial courts are the lowest level court, judicial opinions generated out of trial courts are not binding precedent on any other courts, but lawyers and other judges will still look to those rulings for guidance.

D. Lawyers use Prior Attorney-Authored Arguments to form Reliable Predictive Arguments

In addition to the value of researching and analyzing judge-authored opinions, lawyers also research and analyze the prior attorney-authored arguments that influenced prior rulings. These court documents are valuable for numerous reasons. First, lawyers often do not recreate the wheel.

---

[4] In a vertical court system, like the United States, there are multiple levels of courts. Typically three tiers exist: (1) trial courts, (2) intermediate appellate courts and (3) a supreme court. The trial or inferior level courts conduct almost all trial proceedings (*e.g.*, examine evidence, hear witness testimony, etc...) and reach a ruling based on those proceedings. An intermediate appellate court will review the trial-level ruling, if it is appealed. The appellate review does not re-examine all the evidence, but rather determines whether the trial court ruling was proper given the evidence. Intermediate appellate court rulings may be appealed to and reviewed by the highest level court in that jurisdiction.

However, inferior courts are only bound to obey the rulings established by the higher courts in their jurisdiction. For example, the state courts in California have authority over issues of California state law. Thus, the highest court in California (the Supreme Court of California) only binds lower courts seeking to interpret California law. Thus, a trial-level court in Arizona seeking to interpret an Arizona law is not bound by a prior ruling of the Supreme Court of California; even if the Arizona law at issue is identical to a California law that has already been interpreted by the Supreme Court of California in a factually similar case. This is because the Arizona trial court tasked with interpreting Arizona law is not under the jurisdiction of the Supreme Court of California.

Instead of writing arguments from scratch, lawyers often recycle prior legal arguments and adapt them for the facts of the current case. The ability to efficiently locate and evaluate similar prior arguments saves attorneys (and therefore clients) valuable time and resources in developing and drafting arguments.

Second, reviewing prior legal arguments allows attorneys to identify the most persuasive citations and can further bolster the attorney's argument in the case. Given that judicial rulings tend to be wordy and complex, lawyers derive value in reviewing other attorneys' descriptions and citations from prior rulings. Moreover, reviewing citations on which others have previously relied may reveal potential authorities that may otherwise go overlooked.

Third, reviewing of prior arguments may shed light on the expected arguments of opposing counsel and assists in preparing counter-arguments.

Finally, at times when relevant published case law does not exist (*e.g.*, if the case settled prior to the judge issuing to a final ruling) there is great value in reviewing how other attorneys handled the issue.

Ultimately, the ability to craft persuasive predictive arguments or reliable advisory opinions requires lawyers to preform in-depth review and analyses of prior case law. In this vein, the most valuable legal knowledge is not identifying the existence of the law, but rather understanding how the law applies in certain factual situations. Thus, significant time and money is spent in the legal industry locating and analyzing prior case documents and prior rulings.

E. The Process of a Civil Lawsuit & Civil Case Documents

The process of a lawsuit and its related case documents are largely determined by certain characteristics of the case. For example, a lawsuit may be civil or criminal and could involve alleged violations of federal laws or state laws or both. These characteristics identify what court

has authority to preside over the case and dictate what formal procedural rules will govern the process. Because CLAS captures and analyzes data contained in civil case documents, this section centers on the litigation process for civil cases.

A civil case is a noncriminal lawsuit that seeks to redress an alleged private wrong that usually arises out of disputes between individuals, organizations, businesses, or other private entities.[5] Common examples of the types of civil lawsuits (without limitation) include property damage, personal injury, breach of contract, wrongful termination, medical malpractice, and fraud. The lawsuit type is classified based on the nature of the underlying dispute that gave rise to the litigation. Table 1, on page 45, provides a list of nature-of-suit categories as defined by the California Superior Court, County of Sacramento ("Sacramento Superior Court").

[5] Some jurisdictions further divide civil cases. For example, California courts separate out those civil cases that assert lower damage amounts, allowing lawsuits below a certain damage threshold to proceed through a shorter, less expensive, trial process. These "smaller" claims are frequently called "limited civil cases" and the damage threshold is usually $10,000-$25,000.

Ultimately, the process of a civil lawsuit can be separated into four stages: (1) PLEADING, (2) DISCOVERY, (3) TRIAL and (4) POST TRIAL. Figure 1 below, summarizes these four stages:

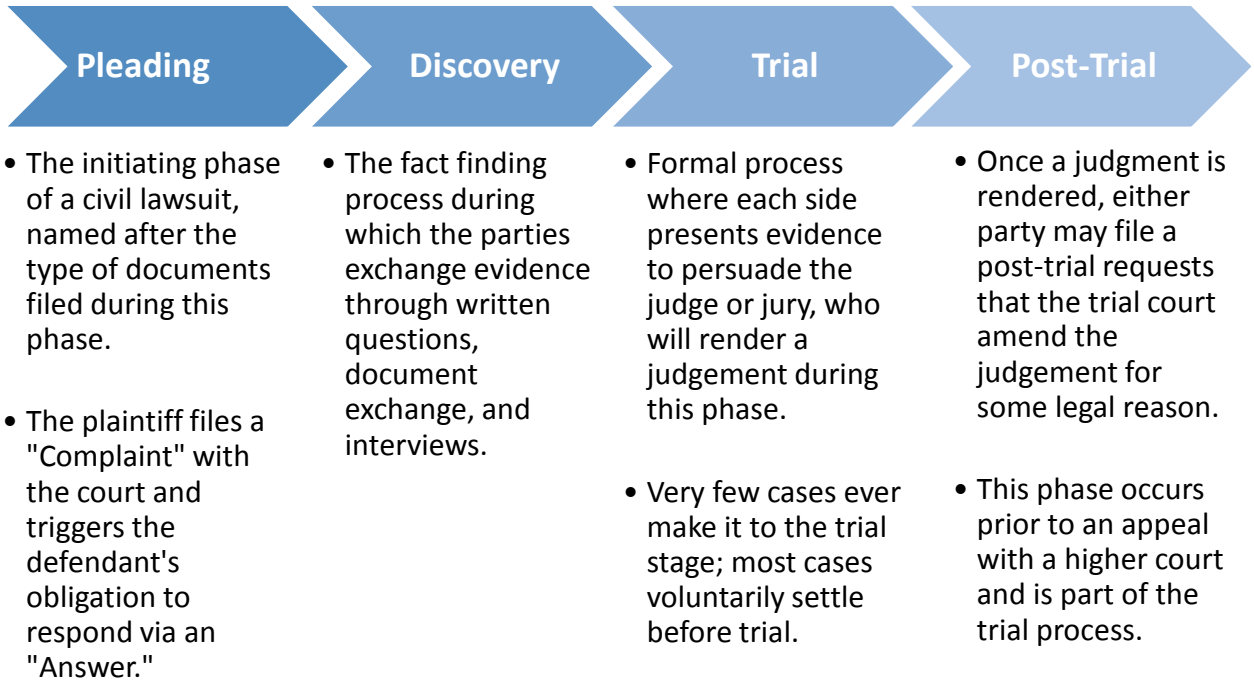


Figure 1 - The Four Stages of Litigation

Throughout every stage of litigation, case documents accumulate to create an official record of each lawsuit.[6] Some case documents may be generated by the court itself, such as transcripts of in-court proceedings or the specific judicial orders issued throughout the litigation process (*e.g.*, the judge granting a party an extension of time). However, many case documents are prepared and filed by the parties (typically, lawyers prepare and file these documents on behalf of their clients).[7] These case documents are broadly referred to as "Filings," because they are filed

[6] The documents that make up the court record are public records, unless otherwise sealed by the court certain laws or court rules may limit public availability of trial-court documents by permitting the "sealing" of selected documents or even entire case records under certain circumstances.

[7] Traditionally trial-court documents are filed manually with a designated court clerk. Some courts today accept electronic filing of uploaded PDF documents over a secure website.

with and maintained by the court clerk or record custodian. (*See* Black's Law Dictionary, 10th ed. 2014). Procedural rules of the presiding court detail the requisite formats of these various Filings and the timeframes during which certain documents may be or must be filed.

Filings occur throughout the litigation process and serve various purposes. There is no exhaustive list of Filing types, however, broadly speaking, Filings usually serve one of the following purposes: (1) setting forth or responding to allegations or claims in the case, (2) requesting the exchange of information between the parties or (3) requesting an order from a judge on a specific issue.[8] Some Filings are mandatory and are required to be filed during a particular litigation stage. For example, as discussed in more detail below, pleadings are specific to the PLEADING stage. However, another broad category of Filings called MOTIONS are not stage-specific and instead appear throughout the entire litigation process. MOTIONS (and their related documents) are one of the most common civil case documents. As a civil lawsuit progresses each party will likely file multiple types of MOTIONS throughout the litigation. The following sections identify some of the more common Filings that may be submitted throughout a litigation.

(i) *Motions & Memorandum of Points and Authorities*

Broadly speaking, a MOTION is a verbal or written application requesting a judge to make a specific ruling or order. Importantly, MOTIONS request judicial orders on legal issues only (*e.g.*, whether the claims in the lawsuit are legally appropriate or whether a party must legally turn over certain information to the other party during the DISCOVERY-stage or whether evidence may legally be submitted to a jury for consideration). Factual disputes of the case are not decided during MOTION practice and are reserved for trial. For example, in a "personal injury – auto"

[8] As a lawsuit progresses, the courts maintain an itemized table of all the court documents and Filings, called a "docket" or "docket sheet." One civil court case has no limit on the number of total documents it may contain; many cases contain hundreds of trial-court documents.

lawsuit, the Defendant may claim he was out of town and was not the driver of the car. However, there may be a red-light photo of the defendant running a red light earlier that day in another part of town. Whether that evidence actually proves that the defendant was driving the car at the time of the accident is a factual issue to be determined at trial. Whether that photography must be shared with the Plaintiff during discovery is a legal issue that could be decided through a Motion to Compel.

There are many different types of MOTIONS, some are substantive to the issues of the lawsuit and others raise incidental issues. A substantive MOTION relates to the legal legitimacy of the actual claim at issue in a lawsuit. For example a substantive MOTION may request that the judge dismiss the lawsuit prior to trial because, even if all the facts alleged by the plaintiff are assumed to be true, the plaintiff would still not be entitled to the requested relief under the law. An incidental MOTION relates to a procedural matter, such as a request for an extension on existing page limitations.

As a Filing, MOTIONS are unique because they trigger a document-heavy hearing process that will occur within the litigation process. Notably, a MOTION itself is really no more than a cover page that provides a concise statement about the specific request. To persuade the judge to grant the MOTION, the filing party also files "supplemental documents" that provide the legal and factual support for the motion. These supplemental documents are also Filings and are entered into the court record. One type of supplement document that is *always* filed alongside a MOTION is a "Memorandum of Points and Authorities" that sets out the issues to be decided, the filing party's position, and the arguments and legal authorities in support of the request. Sometimes other documents may also be filed in support of the MOTION, for example if courts in other jurisdictions

have ruled favorably on similar MOTIONS, then the filing party may provide a supplemental index of persuasive authority for the judge to consider.

The filing of a motion entitles the opposing party to file a written OPPOSITION to that motion, detailing the legal justifications why the court should not grant the motion. The originating party is then permitted to file a REPLY to the OPPOSITION. Usually an in-court hearing is held to provide the judge a final opportunity to seek clarity on the attorneys written arguments. The court will then issue an ORDER on the motion, in which the court may deny the motion, grant the motion, or grant part of the motion. The ORDER also becomes part of the case record.

In terms of legal research, Memoranda of Points and Authorities are among the most valuable attorney-authored documents because they contain valuable information, like citations to prior similar cases and summaries and analysis regarding prior applications of the law. For this reason these documents serve as the input in CLAS. CLAS then outputs aa nature-of-suit classification for each inputted document, allowing lawyers to target their research to only those Memoranda of Points and Authorities that pertain to a specific type of lawsuit.

(ii) *PLEADING Stage Documents*

A civil lawsuit is initiated when the person (or persons or organization(s)) bringing the suit (the "plaintiff(s)") files a document known as a COMPLAINT; copies of which are also served on the other party or parties (the "defendant(s)"), thereby providing notice that a legal action has been filed and enabling the defendant(s) to respond to the COMPLAINT with denials, available defenses, or counterclaims. This begins the PLEADING stage of the litigation.[9] Pleadings are a

[9] The pleading stage derives its name from the fact that during this phase the parties file formal, written pre-trial documents, called "pleadings," with the court. Common pre-trial pleadings include "Complaints," "Answers," "Counterclaims," and "Replies."

broad category of Filings in which a party to the lawsuit sets forth or responds to allegations, claims, denials, or defenses. (*See* Black's Law Dictionary, 10th ed. 2014).

The COMPLAINT is one type of pleading. It sets forth the plaintiff's version of the facts, identifies the law(s) under which the case is brought (*e.g.,* assault, breach of contract, trespass) and specifies the damages sought. Ultimately, a Complaint must identify what facts and law(s) entitle the plaintiff to damages. However, a Complaint does not provide legal analysis; it does not present legal arguments as to why the law should apply or why the court should grant the requested damages. Instead it serves more as an outline of the allegations.

Once filed, a defendant must respond to a COMPLAINT within a specified timeframe and may do so in several ways. The most basic response is for the defendant to simply serve an ANSWER. An ANSWER is another type of pleading in which the defendant responds to each allegation in the COMPLAINT, either by admitting, denying or setting forth an appropriate legal defense. The filing of a responsive pleading by all parties against whom allegations have been made closes the PLEADING stage of litigation.

However, it is common practice for parties to assert pre-answer MOTIONS that substantially prolong the PLEADING stage beyond that of a simple COMPLAINT and ANSWER. Pre-answer MOTIONS include (but are not limited to) motions to dismiss, motions to strike, and a motion for a more definite statement. The types of MOTIONS attack the sufficiency of the opposing party's pleading, claiming that the judge should legal dismiss or strike portions of the pleading or otherwise declare the pleading legally insufficient and thereby excuse the defendant from having to respond. For example a MOTION to dismiss or a MOTION to strike requests the judge dismiss the COMPLAINT or at least strike portions thereof because even if the facts alleged may be true, they are insufficient for the plaintiff to legally state a claim for relief. For example,

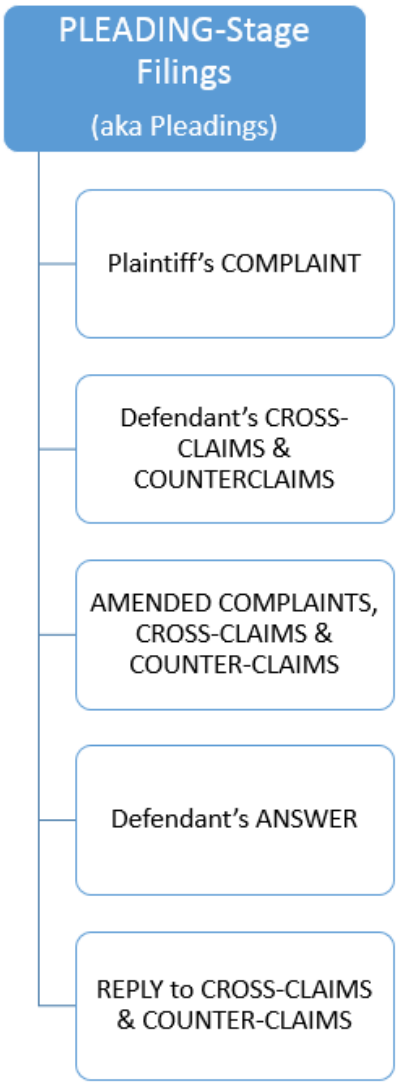
if the plaintiff waited to file the lawsuit for too long and was now outside of the legal time limits, a defendant may file a MOTION to dismiss the case.

These motions are usually unsuccessful in dismissing the case entirely. Even when these types of dispositive motions[10] are granted, the judge will often grant the plaintiff(s) leave to amend the Complaint so it asserts legally sufficient claims. Other times, a judge may grant the MOTION in part, dismissing only a single claim from a multiple claim COMPLAINT. Often, unless a MOTION is denied, a plaintiff may have to amend its original complaint and file an AMENDED COMPLAINT, thereby restarting the entire process. Often, following an AMENDED COMPLAINT, opposing parties will again file another round of MOTIONS. This cycle will continue until the court fully grants or denies all such MOTIONS, thereby either dismissing the case or causing the opposing party to formally respond to the COMPLAINT. Some litigations continue in this MOTION practice through a FOURTH or FIFTH AMENDED COMPLAINT; as long as the court provides the plaintiffs leave to amended, then the lawsuit will continue.

In addition to responding to a COMPLAINT via a MOTION or an ANSWER, the defendant may also file a CROSS-COMPLAINT or a COUNTERCLAIM in which the defendant alleges claims against other parties (*e.g.*, the plaintiffs or other co-defendants). These pleadings may be filed separately or as part of the ANSWER and provide the accused parties with the same opportunity to respond by filing a REPLY or by filing a pre-reply MOTION.

Figure 2, below, provides a list of common PLEADING-stage Filings.

[10] In law, a dispositive motion is a motion seeking a trial court order entirely disposing of all or part of the claims in favor of the moving party without need for further trial court proceedings. This is in contrast to an incidental motion, which deals with procedural issues or other ancillary issues of the case, that do not attempt to dismiss the claims of the lawsuit.


PLEADING-Stage Filings
(aka Pleadings)
Plaintiff's COMPLAINT
Defendant's CROSS-CLAIMS & COUNTERCLAIMS
AMENDED COMPLAINTS, CROSS-CLAIMS & COUNTER-CLAIMS
Defendant's ANSWER
REPLY to CROSS-CLAIMS & COUNTER-CLAIMS

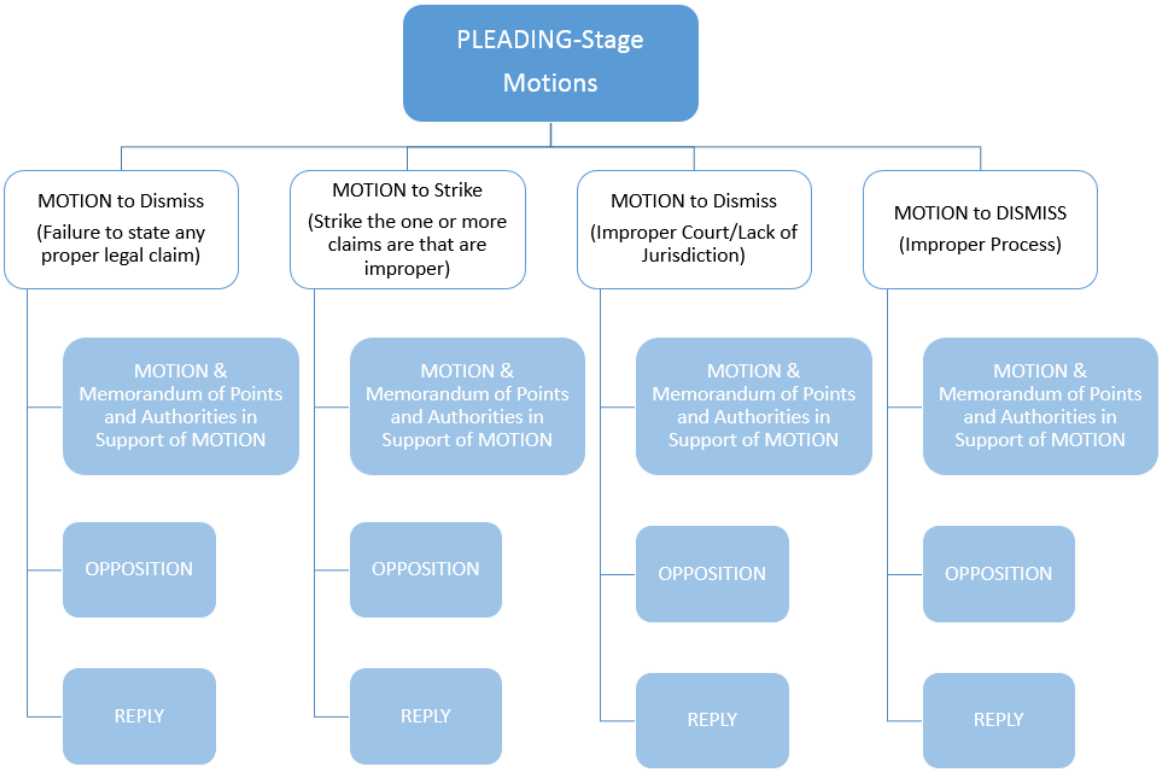


Figure 2 - Common PLEADING-Stage Filings

(iii) *DISCOVERY Stage Documents*

During the DISCOVERY stage the parties acquire, evaluate, and exchange relevant evidence (*e.g.*, admissions to facts, relevant documents, photographs, expert reports, interviews, etc.). Depending on the presiding court's procedural rules, the DISCOVERY-stage usually begins before the close of the pleading stage. Often the initiation of the DISCOVERY-stage is permitted either after the Defendant(s) file an ANSWER or after some specified time period after the filing of the COMPLAINT (*e.g.* 20 days), whichever is first. This makes sense considering the fact that the PLEADING-stage can be a very long and drawn out process (in some litigations it could last years) and the DISCOVERY-stage allows parties to evaluate the strengths and weakness of their arguments and could help facilitate settlement.

Although it may start early in the lawsuit process, the DISCOVERY-stage is often a long and expensive process. Both parties often spend significant time, money and resources producing and reviewing evidence. Any evidence a party intends to submit at trial, must be collected and shared during this stage. Except in unusual circumstances, "surprise" evidence is not admissible during trial.

During this stage the evidence is not submitted to the court for consideration. Discovery is the pre-trial evaluation of the evidence by the parties, so each side can determine what to submit at trial to support their case. Although the evidence is not presented to the court, the exchange of evidence occurs through a formal series of written notices and requests for certain types of evidence and each of these requests is filed with the court. Frequently, parties disagree about what information must be shared. When disagreements occur, the parties can file MOTIONS requesting orders to either compel the other side to release the requested information or protective orders to limit what has to be exchanged. It is common for each party to file multiple discovery-related MOTIONS.

(iv) *TRIAL & POST-TRIAL Stage Documents*

Unless otherwise extended by the court, Discovery has a limited time frame. During the TRIAL stage of litigation, the attorneys (or the parties, if they are not represented) appear in court and present evidence and arguments to support their respective claims and defenses. After evaluating only the evidence actually presented at trial, the judge or jury makes "finding of fact" to resolve any factual disputes where the parties disagree as to what actually happened. The judge will then accept those factual findings and apply them to the relevant laws to reach a final ruling on the matter and will order that judgment be entered for the party who wins. Before during and after trial the parties will file various motions requesting the judge to issue orders to resolve

ongoing disagreements. For example, common motions include requests that the court prohibit one side from presenting evidence that may be inappropriately prejudicial or irrelevant at trial. Or following trial, one side may disagree with the findings of fact and may submit a motion for reconsideration, requesting the judge reverse the jury's findings. Even after a final judgment is entered by the judge, it is common that the parties may file requests that the judge reconsider and overturn the ruling.

III. FORMALIZING LEGAL RELATIONSHIPS WITH FIRST-ORDER LOGIC

First-order logic can be used to formalize many legal relationships. These formalizations can help computer scientists who lack any prior legal background to gain the basic understandings of legal relationships that are necessary to develop reliable and successful legal technologies. Specifically, formalizing legal relationships helps expose gaps in knowledge representations, allows automated systems to make inferences, and can assist in automated classification.

Formalizing legal relationships requires an exhaustive review of potential legal interactions. Formalization helps expose all the potential pieces of relevant information. For example, the process of initiating a lawsuit requires one party to file an initiating Complaint. However, filing an initiating Complaint also requires additional knowledge, such as identifying the defendant(s), the court, the attorney(s), and the judge that will all participate in the lawsuit. The process of formalizing this knowledge reveals this information.

This thesis includes an initial attempt at formalizing legal knowledge as it pertains to civil cases and nature-of-suit classification. While CLAS does not use this formalized first-order logic, future versions of CLAS will be extended to utilize first-order logic, as detailed in the FUTURE WORK section.

Formalizing legal relationships with first-order logic allows automated systems to infer information about legal proceedings, events, and Entities. These types of inferences are similar to the way an attorney infers information in a lawsuit. For example, using first-order logic to formalize the stages of a civil case provides a list of all stages as well as an enforcement of order. Figure 3, below, provides the stages of a civil lawsuit expressed in first-order logic.

Stages in a Civil Lawsuit

BEFORE(STAGE(pleading), STAGE(discovery))
BEFORE(STAGE(discovery), STAGE(trail))
BEFORE(STAGE(trial), STAGE(post-trial))

$\forall$ x, y, z BEFORE(x, y) $\wedge$ BEFORE(y, z) $\Rightarrow$
BEFORE(x, z)

Figure 3 - The Four Stages of Litigation in First-Order Logic

The first-order logic in Figure 3 allows an automated system to infer the order of the stages. Understanding the order of the stages can provide additional context to formalize further information, such as the types of documents that can be filed in each stage. As previously discussed, when one party files a Complaint, this allows the opposing party to file an Answer. Similarly, when a party files a Motion, they may attach a supporting document such as a Memorandum of Points and Authorities. The filing of the Motion allows the opposing party to file an Opposition to the Motion. Thus, the process of filing one document permits (and sometimes mandates) the filing of another document. Figure 4, below, provides first-order logic for filing a document, permitting the filing of a document, and attaching one document to another document.

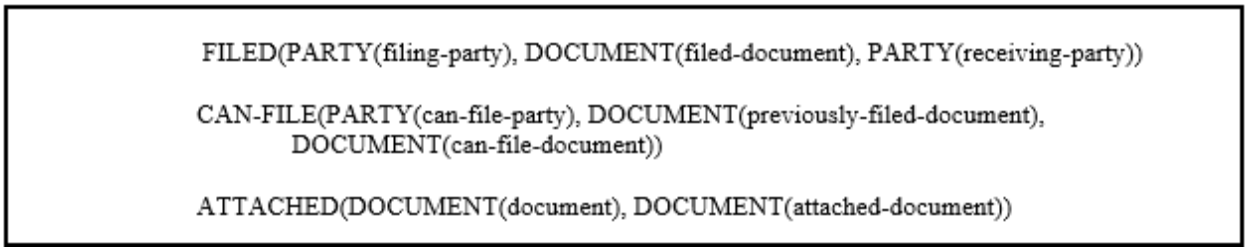


Figure 4 - Filing and Enabling Filing in First-Order Logic

In Figure 4, filing a document ("FILED") requires knowledge of the party filing the document, the document type being filed (*e.g.* Complaint, Answer, Motion, etc…), and the receiving party. Allowing a document to be filed ("CAN-FILE") requires knowledge of the party that is allowed to file the document, the document that was previously filed permitting the filing of a new document, and finally the new document that can be filed. Attaching a document ("ATTACHED") requires the original document and the document being attached. These relationships are important because they create a formalized structure for filed documents, allowing an automated system to interpret relationships between filed documents.

With a formalized structure for filing documents, the process of a Motion can be formalized. Motions can be filed at any stage in a civil case, thus their formalization does not require knowledge of the stage. Formalizing the process of filing Motions allows an automated system to understand the pre-conditions and post-conditions for each step in the Motion process. Figure 5, below, builds on the knowledge in Figure 4 to provide the process of a motion expressed in first-order logic.

Process of a Motion

$\forall$ x DOCUMENT(x) $\Rightarrow$
MOTION(x) $\vee$ OPPOSITION(x) $\vee$ REPLY(x) $\vee$
MEMORANDUM-OF-POINTS-AND-AUTHORITIES(x)

$\forall$ x SUPPORTABLE(x) $\Rightarrow$ MOTION(x) $\vee$ OPPOSITION(x) $\vee$ REPLY(x)

$\forall$ p1, p2, x, y FILED(p1, DOCUMENT(x), p2) $\wedge$ SUPPORTABLE(x) $\Rightarrow$
ATTACHED(x, DOCUMENT(y)) $\wedge$ FILED(p1,
MEMORANDUM-OF-POINTS-AND-AUTHORITIES(y), p2)

$\forall$ p1, p2, m, o PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, MOTION(m), p2) $\Rightarrow$
CAN-FILE(p2, m, OPPOSITION(o))

$\forall$ p1, p2, m, o PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, MOTION(m), p2) $\wedge$ FILED(p2, OPPOSITION(o), p1) $\Rightarrow$
CAN-FILE(p1, o, REPLY(r))

$\forall$ p1, p2, m, o PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, MOTION(m), p2) $\wedge$ FILED(p2, OPPOSITION(o), p1) $\wedge$ FILED(p1, REPLY(r), p2) $\Rightarrow$
ORDER(MOTION(m))

$\forall$ x ORDER(MOTION(x)) $\Rightarrow$ DENY(x) $\vee$ GRANT(x) $\vee$ GRANT (PART-OF(x))

Figure 5 - Process of a Motion in First-Order Logic

In Figure 5, the process of a motion is described. In order to represent the process of a motion, the types of documents in the motion process must be defined (Motion, Opposition, Reply), along with the types of documents that can be attached in support (Memorandum of Points and Authorities). Since the process of a motion concludes in an order by the judge, the resulting order must also be represented. Additional finer grained information about Motions can also be formalized. For example, as previously discussed, Motions are substantive or incidental. A substantive Motion relates to the legal legitimacy of the actual claim at issue in a lawsuit and an incidental Motion relates to a procedural matter. Figure 6, below, provides an example of formalizing these types of Motions using first-order logic.

$$\text{TYPE(MOTION)} \Rightarrow [\text{SUBSTANTIVE, INCIDENTAL}]$$
$$\text{ISSUE-TYPE(MOTION)} \Rightarrow [\text{LEGAL-ISSUE, PROCEDURAL-ISSUE}]$$

$$\forall x\ \text{SUBSTANTIVE(MOTION(x))} \Rightarrow \text{LEGAL-ISSUE(MOTION(x))}$$

$$\forall x\ \text{INCIDENTAL(MOTION(x))} \Rightarrow \text{PROCEDURAL-ISSUE(MOTION(x))}$$

Figure 6 - Motion Type and Issue Type in First-Order Logic

Figure 6 provides a formalization for the fact that substantive Motions infer legal issues and incidental Motions infer procedural issues.

While motions can be filed in any stage of a civil case, knowledge of the current stage can allow an automated system to reason about what documents can be filed and what events trigger a transition between stages. Figure 7, below, provides first-order logic for (1) initiating a lawsuit and entering the PLEADING-stage, (2) the documents filed during the PLEADING-stage, and (3) the events that trigger a transition from the PLEADING-stage to the DISCOVERY-stage or allow the PLEADING-stage and DISCOVERY-stage to occur simultaneously.

Pleading Stage

$\forall$ x DOCUMENT(x) $\Rightarrow$
COMPLAINT(x) $\vee$ ANSWER(x) $\vee$ COUNTERCLAIM(x) $\vee$ CROSSCLAIM(x) $\vee$
REPLY(x)

$\forall$ p1, p2, c, pleading, a, m, cc PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, COMPLAINT(c), p2) $\Rightarrow$
STAGE(pleading) $\wedge$ $\neg$TIME-EXPIRED(STAGE(pleading)) $\wedge$
(CAN-FILE(p2, c, ANSWER(a)) $\vee$ CAN-FILE(p2, c, MOTION(m)) $\vee$
CAN-FILE(p2, c, COUNTERCLAIM(cc)) $\vee$ CAN-FILE(p2, c, CROSSCLAIM(cc)))

$\forall$ p1, p2, c, r, a, m, cc PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, COMPLAINT(c), p2) $\wedge$
(FILED(p2, COUNTERCLAIM(cc), p1) $\vee$ FILED(p2, CROSSCLAIM(cc), p1)) $\Rightarrow$
STAGE(pleading) $\wedge$ $\neg$TIME-EXPIRED(STAGE(pleading)) $\wedge$
(CAN-FILE(p1, cc, REPLY(r)) $\vee$ CAN-FILE(p2, c, ANSWER(a)) $\vee$
CAN-FILE(p1, cc, MOTION(m)) $\vee$ CAN-FILE(p2, c, MOTION(m))))

Pleading Stage to Discovery Stage

$\forall$ p1, p2, c, a, x, discovery PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, COMPLAINT(c), p2) $\wedge$
FILED(p2, ANSWER(a), p1) $\Rightarrow$
STAGE(discovery) $\wedge$ $\neg$END-DATE(STAGE(discovery))

$\forall$ p1, p2, c, cc, r, a, discovery PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, COMPLAINT(c), p2) $\wedge$
(FILED(p2, COUNTERCLAIM(cc), p1) $\vee$ FILED(p2, CROSSCLAIM(cc), p1)) $\wedge$
FILED(p2, ANSWER(a), p1) $\wedge$ FILED(p1, REPLY(r), p2) $\Rightarrow$
STAGE(discovery) $\wedge$ $\neg$END-DATE(STAGE(discovery))

$\forall$ pleading, discovery STAGE(pleading) $\wedge$ TIME-ELASPED(STAGE(pleading)) $\Rightarrow$
STAGE(pleading) $\wedge$ STAGE(discovery) $\wedge$ $\neg$END-DATE(STAGE(discovery))

Figure 7 - Process of the Pleading Stage in First-Order Logic

The first-order logic in Figure 7 allows an automated system to determine when the PLEADING-stage is initiated and the types of documents that can be filed during the PLEADING-stage. This formalization requires knowledge of each type of document that can be filed during the PLEADING-stage and the order in which they are filed. The first-order logic also displays the transitions between the PLEADING-stage and DISCOVERY-stage. For example, if a plaintiff files a Complaint and the defendant files an Answer, the PLEADING-stage ends and the DISCOVERY-stage begins. In some circumstances, the PLEADING-stage can overlap with the DISCOVERY-stage. This occurs when the amount of time allowed for the PLEAIDNG-stage

expires, thus starting the DISCOVERY-stage, however, the PLEADING-stage still continues until an Answer is filed.

While the PLEADING-stage is fairly complicated, the DISOVERY-stage and TRIAL-stage can be represented using simpler first-order logic. Figure 8, below, provides the process of the DISCOVERY-stage and TRIAL-stage using first-order logic.

Discovery Stage

$\forall$ x DOCUMENT(x) $\Rightarrow$
MOTION(x) $\vee$ OPPOSITION(x) $\vee$ REPLY(x) $\vee$
MEMORANDUM-OF-POINTS-AND-AUTHORITIES(x) $\vee$
EVIDENCE-REQUEST(x) $\vee$ EVIDENCE-RESPONSE(x)

$\forall$ p1, p2, er, erp PARTY(p1) $\wedge$ PARTY(p2) $\wedge$ FILED(p1, EVIDENCE-REQUEST(er), p2) $\Rightarrow$
CAN-FILE(p2, er, EVIDENCE-RESPONSE(erp))

* Same as Process of a Motion

Discovery Stage to Trial Stage

$\forall$ discovery, pleading, trial STAGE(discovery) $\wedge$ END-DATE(STAGE(discovery)) $\wedge$
$\neg$STAGE(pleading) $\Rightarrow$
STAGE(trail)

Trial Stage

EVALUATE(evidence, RULING-BODY(body))
APPLY-LAWS(FACTS(facts), JUDGE(judge))

$\forall$ x RULING-BODY(x) $\Rightarrow$ JUDGE(x) $\vee$ JURY(x)

$\forall$ x, evidence, facts EVALUATE(evidence, RULING-BODY(x)) $\Rightarrow$ FACTS(facts)

$\forall$ x, facts FACTS(facts) $\wedge$ JUDGE(x)) $\Rightarrow$ APPLY-LAWS(facts, x)

$\forall$ x, facts, p APPLY-LAWS(facts, JUDGE(x)) $\Rightarrow$ RULING(PARTY(p))

Trial Stage to Post-Trail Stage

$\forall$ trail, p, post-trial STAGE(trail) $\wedge$ RULING(PARTY(p)) $\Rightarrow$ STAGE(post-trail)

Figure 8 - Process of the Discovery Stage and Trial Stage in First-Order Logic

Figure 8 completes the first-order logic for the stages of a civil case. The DISCOVERY-stage includes evidence requests, evidence responses, and Motions. The DISCOVERY-stage ends

after a predetermined date has been reached, thus starting the TRIAL-stage. The TRIAL-stage requires a ruling body (judge or jury) to determine the facts (which may be under dispute) in the lawsuit. Once these facts have been determined, a judge applies the law to these facts in order to decide a ruling. After a ruling has been made, the TRIAL-stage ends and the POST-TRIAL-stage begins.

Although these formalizations require further work to more accurately represent the details of the litigation in each stage, these initial formalizations allow an automated system to infer the current stage of a civil case based on previously filed documents as well as infer the next step in the case.

While formalizing the process of a civil case using first-order logic provides context for an automated system to infer procedural understanding, additional formalization can improve nature-of-suit classification. Because the underlying relationship between the plaintiff(s) and defendant(s) drives the nature-of-suit classification, formalizing these relationships using first-order logic provides an automated method for classifying the nature-of-suit category. Figure 9, below, provides an example of using first-order logic to formalize how the underlying relationship between parties in a lawsuit can infer certain types of nature-of-suit categories.

PARTICIPANT(PARTY(party), LAWSUIT(lawsuit))
LAWSUIT-IN(LAWSUIT(lawsuit), NATURE-OF-SUIT(nature-of-suit))
CONTRACT-BETWEEN(CONTRACT(c), PARTY(p), PARTY(d))
FULFIL(CONTRACT(c), PARTY(p))
DRIVER(PARTY(p), AUTOMOBILE(a))
COLLISION(AUTOMOBILE(a1), AUTOMOBILE(a2))

$\forall$ x PARTY(x) $\Rightarrow$ PLAINTIFF(x) $\vee$ DEFENDANT(x)

$\forall$ x NATURE-OF-SUIT(x) $\Rightarrow$
BREACH-OF-CONTRACT(x) $\vee$ PERSONAL-INJURY-AUTO(x)

$\forall$ l, n, p, d, c LAWSUIT(l) $\wedge$ NATURE-OF-SUIT(n) $\wedge$ LAWSUIT-IN(l, n) $\wedge$ PLAINTIFFS(p) $\wedge$
PARTICIPANT(p, l) $\wedge$ DEFENDANTS(d) $\wedge$ PARTICIPANT(d, l) $\wedge$ CONTRACT(c) $\wedge$
CONTRACT-BETWEEN(c, p, d) $\wedge$ $\neg$FULFIL(c, d) $\Rightarrow$
BREACH-OF-CONTRACT(n)

$\forall$ l, n, p, d, pa, da LAWSUIT(l) $\wedge$ NATURE-OF-SUIT(n) $\wedge$ LAWSUIT-IN(l, n) $\wedge$
PLAINTIFFS(p) $\wedge$ PARTICIPANT(p, l) $\wedge$ DEFENDANTS(d) $\wedge$ PARTICIPANT(d, l) $\wedge$
AUTOMOBILE(da) $\wedge$ DRIVER(d, da) $\wedge$ AUTOMOBILE(pa) $\wedge$ DRIVER(p, pa) $\wedge$
COLLISION(da, pa) $\Rightarrow$
PERSONAL-INJURY-AUTO(n)

Figure 9 - Underlying Relationship between Parties in a Lawsuit Infers the Nature-of-Suit Category

In Figure 9, first-order logic is used to describe the underlying relationship between parties in a lawsuit that can infer a nature-of-suit category. In a breach of contract lawsuit, the parties must enter into a contract and the defendant must fail to fulfil the contract. In a personal injury – auto lawsuit, the parties must be involved in an automobile accident. Although these are simple implementations, they provide an example of how a classification system can use this formalized knowledge to infer the nature-of-suit category based on the relationship between the parties in the lawsuit.

However, due to the inherent complexity in law, it can be very difficult to represent certain relationships using first-order logic. For example, throughout the stages of a civil case, Motions may include requests that the court prohibit one side from presenting evidence that may be inappropriately prejudicial or irrelevant at trial. Determining if evidence is inappropriately prejudicial or irrelevant requires specific context, reasoning, and an understanding of both law and

the world. For example, in a hypothetical case, John Smith rear-ended Jane Doe on the freeway causing Jane Doe to lose control of her car and collide with the median. Her car was totaled and she suffered a broken collarbone. Jane Doe sued John Smith for the damages to her vehicle and medical expenses. John Smith had rear-ended three other drivers over the past year and this evidence was brought before the judge (due to disagreements between the parties) to determine if it is inappropriately prejudicial or irrelevant. Thus, formalizing the determination of inappropriately prejudicial evidence requires developing reasoning that is currently only allowed by a judge. The judge is deciding if the evidence will influence the jury in such a way that it leads the jury to reason improperly about the case. Figure 10, below, provides a simple example of first-order logic that describes the assumption that past behaviors provide evidence for accused behaviors.

DO-WRONG(PARTY(p), WRONG(w), TIME(t))
ACCUSED-OF(PARTY(p), WRONG(w), TIME(t))
IRRELEVANT(WRONG(w1), WRONG(w2))
BEFORE(TIME(t1), TIME(t2))

$\forall$ x PARTY(x) $\Rightarrow$ PLAINTIFF(x) $\vee$ DEFENDANT(x)

$\forall$ p, w1, t2, w2, t2 PARTY(p) $\wedge$ DO-WRONG(p, w1, t1) $\wedge$ WRONG(w1) $\wedge$ TIME(t1) $\wedge$
ACCUSED-OF(p, w2, t2) $\wedge$ WRONG(w2) $\wedge$ TIME(t2) $\wedge$ BEFORE(t1, t2) $\Rightarrow$
DO-WRONG(p, w2, t2) $\vee$ $\neg$IRRELEVANT(w1, w2)

Figure 10 - Prejudicial or Irrelevant in First-Order Logic

The example in Figure 10, can be used to describe a scenario where a previous conviction would result in a jury inferring that the party committed the action of which they are accused. This example could be applied to the hypothetical case because the fact that John Smith had previously rear-ended three other drivers may convince a jury he was at fault in the current lawsuit. While this example provides a simple first-order logic implementation of prejudicial reasoning, it is extremely difficult to handle complex cases because the first-order logic must reason about incorrect reasoning (*e.g.* a jury might reason incorrectly based on a previous conviction).

IV. THE TECHNOLOGY LANDSCAPE: USING TECHNOLOGIES TO ASSIST IN LEGAL REASONING

A. Current Landscape of Legal Research, a Review of Commercial Systems

Given the importance of reviewing prior cases and case documents in the practice of law, there are a number of specialized legal research tools that attempt to consolidate and index valuable legal documents into digital databases. However, the large majority of these tools explicitly target attorney users and require specialized training and extensive legal knowledge in order to find desired documents. For example, the two market leaders in legal research systems, WestLaw (http://www.westlaw.com) and LexisNexis (http://www.lexisnexis.com), both employ full-time representatives at law schools across the United States to train law students on how to use their respective search engines successfully. While there are some free commercial offerings, the most effective legal research tools are the expensive, subscription-based Westlaw and LexisNexis.

WestLaw and LexisNexis both offer an array of legal research tools. Both companies offer natural language search options which can be further filtered by court, state, jurisdiction, time period, judge, attorneys, law firms, and more. Both companies also offer a service in which users can call the company and describe the type of documents they are looking for and a trained employee conducts the search for the user. Once a user has located a case which is relevant to their search, they can review each document in the case as well as case metadata. The metadata includes plaintiffs, defendants, judges, attorneys, law firms, outcome of the case, the citations used in the case, a list of topics in the case, and additional information. The list of topics are analogous to the nature-of-suit categories discussed in this thesis, however, WestLaw and LexisNexis include many additional categories. In addition, these companies have each developed high-level summaries or annotations of certain key aspects of the rulings that allow users to quickly ascertain whether the document would be useful or relevant. However, very little of this process utilizes

technology that would help automate the extraction or summary of relevant data from these documents. Instead today, any summary data that exists is heavily reliant on manual extraction, review, and interpretation. When manual interpretation is used, individual opinions are unavoidable, introducing subjectivity. Figure 11, below, provides an example of a WestLaw annotated judicial opinion. The first main paragraph is a WestLaw summary of the lawsuit. The "West Headnotes" box provides a summary of judicial opinion relating to a certain area of law.

11 Cal.4th 85
Supreme Court of California.

FREEMAN & MILLS, INCORPORATED, Plaintiff and Appellant,
v.
BELCHER OIL COMPANY, Defendant and Appellant.

No. S042831. Aug. 31, 1995.

Accounting firm brought action against oil company to recover fees for services alleging, inter alia, claim for bad faith denial of contract. The Superior Court, Los Angeles County, No. C740915, Edward Y. Kakita, J., entered judgment for accounting firm, and oil company appealed. The Court of Appeal, Miriam A. Vogel, J., 33 Cal.App.4th 837, 33 Cal.Rptr.2d 585, reversed, and accounting firm sought review. The Supreme Court, Lucas, C.J., held that oil company's denial of contract did not create remedy in tort.

Affirmed.

Kennard, J., filed concurring opinion in which Arabian, J., joined.

Mosk, J., filed opinion concurring in part, and dissenting in part.

West Headnotes (1)

Change View

1 **Action** Nature of Action
Torts Duty, breach, or wrong independent of contract
Tort recovery does not lie in action for noninsurance contract breach, in absence of violation of independent duty arising from principles of tort law other than bad faith denial of existence of, or liability under, breached contract; overruling *Seaman's Direct Buying Service, Inc. v. Standard Oil Co.*, 36 Cal.3d 752, 206 Cal.Rptr. 354, 686 P.2d 1158.

173 Cases that cite this headnote

Figure 11 - Example Westlaw Annotation

Moreover most nearly all of the commercial offerings in the legal research arena focus predominantly on aggregating and annotating binding case law. As discussed in the previous

section, binding case law is composed of published judicial rulings from appellate courts and higher. Therefore, the majority of existing legal research tools focus on a different population of legal documents than CLAS. This is because CLAS is focused on a classifying the nature of a lawsuit using a single type of trial-court document called a Memorandum of Points and Authorities. Trial-court documents cannot create binding case law because trial courts are the lowest level courts. In WestLaw and LexisNexus, the ability to research and locate trial-court documents or prior attorney authored documents is still very limited. WestLaw and LexisNexus offer limited libraries of these documents, but because their current systems are so heavily reliant on manual review and annotations, their libraries of court documents are not easily searchable. In this first implementation of CLAS, the system is focused on extracting valuable data points (ie. parties in the lawsuit, non-party Entities, and Key Phrases) from Memorandum of Point and Authorities and classifying the nature of suit.

Beyond these limited databases, researching and locating court documents is limited to law firm internal databases of their own prior Filings and individual court's databases, which generally require one to know the case number in order to retrieve any of the documents. Free online services are offered by Justia (https://www.justia.com/), FindLaw (http://www.findlaw.com/), PlainSite (www.plainsite.org/), and Google Scholar (https://scholar.google.com/). However, these services also focus on allowing the user to search for binding case law and do not include high-level summaries or annotations.

Ultimately, the true value in legal research tools comes from the work done by manual extraction and interpretation. Writing high-level summaries or annotations requires understanding circumstantial facts within an individual case and how the law applies based on those facts. Due to these extremely complex requirements, fully automated systems have not been introduced into

the commercial market. If automated systems could be developed to extract case information and write annotations, these systems would significantly impact the market by lowering costs and providing information faster than the current human standard. CLAS takes a small step towards this ultimate goal of automated annotations.

B. Case-Based Reasoning Research and Development in the Field of Artificial Intelligence and Law

Together with the commercial tools, the research shows the challenges of successfully implementing useful technologies in the legal field. The intricacies of legal relationships and meanings of legal terminology make it difficult to develop systems that model this information.

Case-based reasoning[11] is the process of using previous experiences to understand and solve new problems. Evaluating the use of case-based reasoning systems to model legal data and emulate the decision making abilities of legal experts are prevalent research topics in the artificial intelligence in law community. This makes sense because, as discussed in previous sections, the legal practice centers on one's ability to reason by analogy in order to draw conclusions and guide decision making. Legal practitioners use analogies between target problems and past precedents to draft predictive assessments and persuasive arguments that the target problem should be decided in accordance with prior precedent. A major challenge when implementing case-based reasoning systems in law is abstracting away enough details to be able to draw similarities between prior lawsuits. For example, in Lawsuit A, a driver of a car is at a stop sign and a tree limb falls on their car, totaling it. In Lawsuit B, a motorcyclist runs a red light, collides with a car, and causes damage to their motorcycle and the car they hit. Determining if Lawsuit A and Lawsuit B are related

[11] Case-based reasoning systems have been used in a variety of tasks. One example is a helpdesk system which attempts to diagnose problems on computers [Kolodner, 1992]. A user provides a description of the problem they are experiencing with their computer and the case-based reasoning system finds the closest related previous case.

requires abstracting away certain details, such as treating a car as a vehicle and a motorcycle as a vehicle. With this simple abstraction, the two lawsuits are related because they both contain vehicles and both contain damage to vehicles. However, the law applies differently in these two scenarios because Lawsuit B is the fault of the motorcyclist while Lawsuit A contains no fault. Because of the complexities in the application of law, CLAS does not attempt to draw similarities to prior cases using abstractions.

Overall the legal practice involves a wide variety of case-based tasks and methods, making it ripe for disruption with case-based reasoning systems. For example, case-based reasoning systems can contribute to the design of intelligent legal data retrieval systems and improve legal document assembly programs, thereby reducing the time spent in repetitive, labor-intensive legal tasks. Or case-based reasoning systems can be used to apply conventional rule-based systems to legal problem-solving, increasing the speed and reliability of legal advice.

However, a review of the research illustrates that the inherent complexity and decentralization of the legal industry causes numerous challenges for successful implementation of case-based reasoning systems. At a high-level, case-based reasoning systems are composed of two parts, a storage component and a retrieval component. The storage component of any case-based reasoning system functions to hold a collection of extracted information obtained from prior cases. The retrieval stage then employs various techniques to locate and retrieve relevant information from the storage tool. Research in the artificial intelligence in law community investigates challenges in both of these components.

(i) *Storing Legal Information*

There are four methods that researchers have used to store legal information: (1) identifying, extracting, and storing legal concepts, (2) legal ontologies, (3) record aggregation, and (4) vector space models.

a. Identifying, Extracting, and Storing Legal Concepts

A number of researchers have worked on identifying, extracting, and storing legal concepts. For example, in Gelbart and Smith [Gelbart and Smith, 1993], a system was developed which could extract legal concepts from legal documents. Legal concepts were extracted by matching sections of text with terms in a "Legal Phrase Dictionary", a manually constructed collection of legal words and phrases. Every word or phrase in the dictionary also contained one or more links to legal concepts contained in a "Concept dictionary", a manually constructed collection of legal concepts. Therefore, each matching section of text would result in one or more concepts found in the "Concept dictionary."

In Gelbart's system, the sentence "They allege that he was negligent in carrying out his duties as a solicitor, and that his negligence caused them to suffer economic loss", would have two matches in the "Legal Phrase Dictionary", "negligent" and "negligence". Both of the matches in the "Legal Phrase Dictionary" would have links to the legal concept of "negligence" in the "Concept dictionary" [Gelbart and Smith, 1993]. Extensions to the Gelbart system were made by Smith in 1997 [Smith, 1997]. The extensions included a legal conceptual hierarchy which was used to relate legal concepts to one another. The legal concept hierarchy was manually created and maintained.

Similar to Gelbart's approach, CLAS uses a dictionary of words that have links to nature-of-suit categories. The dictionary of words is built during training of the CLAS system, thus

avoiding the manual work required to construct the "Legal Phrase Dictionary" and "Concept dictionary" in Gelbart's system.

Alternatively, other research efforts test the extraction of legal concepts from legal documents by generalizing the text in legal documents [Bruninghaus and Ashley, 2001]. For example, in a hypothetical case where the plaintiff is Mason and the defendant is Randle, the sentence "Mason disclosed part of the recipe to Randle" is generalized into "Plaintiff disclosed the information to defendant." In this example, Mason and Randle are the generalized into their roles of plaintiff and defendant, respectively. The event in the sentence is generalized into a legal concept of "disclosed information."

The process of generalizing sentences was possible due to the Case Argument TutOrial system ("CATO") [Aleven and Ashley, 1997]. CATO was designed as an intelligent tutoring system to teach law students basic skills of case-based legal argument, such as how to distinguish current facts from prior cases. The CATO system was able to classify the plaintiffs, defendants, and a limited set of legal concepts in legal arguments which conformed to a specific format. CATO employs a didactic approach, presenting law students with good and bad examples of distinguishing, based on legal information stored from prior lawsuits. In this way CATO teaches a process of case-based reasoning, literally comparing and contrasting problems to past cases in order to draw and justify inferences about the problems [Aleven and Ashley, 1997].

b. Legal Ontologies

Research has also investigated the use of legal ontologies as the storage component in legal case-based reasoning systems. In the artificial intelligence domain, an ontology is a collection (a description or a formal specification) of the terms, concepts, and relationships that are used to describe and represent an area of knowledge. A legal ontology can play an important role in

storage by promoting consistency among accumulated information [Gruber, 1993]. In this manner, a legal ontology can support the storage structure of legal knowledge modelling.

While CLAS does not implement a legal ontology, it does extract and index parties in a lawsuit, roles of the parties (plaintiff or defendant), and non-party Entities. Additionally, CLAS indexes events that occurred in the past between two identified parties, two non-party Entities or between an identified party and another Entity. An example of a legal ontology with a broad scope is the Legal Knowledge Interchange Format (the "LKIF Core Ontology" or "LKIF").

As it is described by its developing consortium, the Estrella Consortium[12], LKIF caters to the "continuing need for a standard vocabulary of basic legal terms when exchanging knowledge between different knowledge based systems." (*see* Estrella, "LKIF Core Ontology," available at http://www.estrellaproject.org/?page_id=3). Initially, the LKIF Core Ontology intended to build on an already existing core-ontology, the LRI-Core [Breuker and Hoekstra, 2004], however development proved more complicated once researchers realized that "in law, and in particular in legal reasoning, complex patterns of concepts are used, which do not easily fit into an ontology" ["Deliverable 1.4: OWL Ontology of Basic Legal Concepts (LKIF-Core)," 2007, p. 3]. LKIF employs concepts from both legal and commonsense domains, acting more as a library of ontologies relevant for the legal domain than a monolithic body of definitions. For example LKIF models basic legal concepts such as legal actions, court judgments, legal codes defined by European courts, and legal entities (*e.g.*, plaintiffs, defendants, attorneys, and judges). Unlike LKIF, CLAS is limited to Memorandum of Points and Authorities from the Sacramento Superior Court. The LKIF ontology also stores common sense vocabulary and concepts. This includes

[12] "ESTREALLA" stands for the European project for Standardised Transparent Representations in order to Extend Legal Accessibility.

physical concepts (like moving a chair), physical objects (like a book), mental concepts (like remembering an event), and social concepts (like roles in organizations such as a CEO in a company). Although the LKIF project was fully funded and developed over a period of 30 months, the project faced a number of implementation hurdles due to the complexities of legal relationships and has not been continued or extended. The project has not been updated since the original derivable documents in 2007 (https://github.com/RinkeHoekstra/lkif-core).

Other, more specific legal ontologies also exist. One such ontology catalogs the "Code of Federal Regulations Parallel Table of Authorities and Rules" [Richards and Bruce, 2011]. The "Code of Federal Regulations Parallel Table of Authorities and Rules" links federal regulations to the federal statute(s) that authorized the governmental agency to enact the regulation. Richards and Bruce developed an ontology to automate the process of connecting these regulations to the statute that authorizes them. While the Richards and Bruce ontology modelled the authorization of laws, CLAS does not attempt to model any actual application of law. The research from Richards and Bruce also emphasized the issues created by the complex relationship within the legal domain. For example, Richards and Bruce found that multiple types of relationships could exist between a regulation and how a statute may authorize it. For example, an authorization may be considered an "express authority for the regulation" or an "implied authority for the regulation" or it could be "applied by the regulation" or "interpreted by the regulation" [Richards and Bruce, 2011]. Each of these relationships between the regulation and the authorizing statute hold different legal consequences. Thus a useful ontology must capture all of this legal information.

c. Record Aggregation

Record aggregation systems can also serve as a storage component in legal case-based reasoning systems. A record aggregation system indexes a corpus and combines existing records

or in the case of a new record, determines if the new record should be aggregated with existing ones. For example, CLAS indexes parties in a lawsuit, roles of the parties (plaintiff or defendant), and non-party Entities. When indexing parties in the lawsuit, records that represent the same party (plaintiffs or defendants) are aggregated together.

Another example of record aggregation is a system called PeopleMap [Conrad et al., 2011]. PeopleMap stores public records relating to the legal relationships between people and assets. The records stored in this system include ownerships, contracts, bankruptcies, seizures, and more. The system leverages entity resolution[13] techniques [Benjelloun et al., 2009; Cohen and Richman, 2002] in order to assign multiple records to a single entity. The system also uses a semi-supervised learning algorithm [Veeramachaneni and Kondadadi, 2009] to match records. More recently, PeopleMap was acquired and adapted to a commercial system owned by WestLaw [31].

d. Vector Space Models

While case-based reasoning systems commonly use separate approaches for storage and retrieval, some approaches serve as both the storage component and retrieval component. For example, vector space models can serve as the architecture for both storage and retrieval. Vector space models are representations of text in the form of a vector of values. These values are assigned using a transformation. For example, the term frequency-inverse document frequency transformation (tf-idf). tf-idf is used to determine the how important a word is within a corpus. tf-idf is calculated by taking the term frequency (Number of times a term n appears in a document divided by the total number of terms in the document.) multiplied by the inverse document frequency (The logarithm of the total number of documents divided by the number of documents

[13] Entity resolution means identifying and linking or grouping different references of the same real world entity. In the PeopleMap system, the references are real world entities are people, thus the goal is to link different references to a single person.

with term n.). Figure 12, below, provides a simple example of calculating tf-idf in a corpus of two documents.

Document 1	
Term	**Term Count**
vehicle	1
motorcycle	1
car	10
the	2
a	3

Document 2	
Term	**Term Count**
motorcycle	1
the	2
a	3

Term	Term Frequency (tf)	Inverse Document Frequency (idf)	tf-idf
car	10 / 17 = 0.5882	log(2 / 1) = 0.3010	0.1770
the	2 / 17 = 0.1176	log(2 / 2) = 0	0.0000
vehicle	1 / 17 = 0.0588	log(2 / 1) = 0.3010	0.0177

Figure 12 - Example of tf-idf Calculation

CLAS uses vector space models to represent word occurrences in Key Phrases. Specifically, each vector represents the number of times a word occurs in a single Key Phrase. These vectors are aggregated together and indexed in CLAS for use during the classification of the nature of the suit. This process is discussed in detail in section "CLAS COMPONENTS."

Gelbart and Smith [Gelbart and Smith, 1993] adapted existing vector space models [Salton and McGill, 1983] to the legal domain. The vector space models represent a legal document as one vector that contains the number of occurrences for each word in the document. The collection of all document vectors are the storage component. Queries are also modeled a vector of words. Comparison against the existing documents uses two approaches. First, the query vector is used to locate and retrieve the most similar documents using cosine similarity[14] to measure the distance

[14] In cosine similarity, words are converted into vector space and the cosine angle between the two vectors is measured.

between the vectors. Second, the queries natural language text is converted into weighted terms and used to find matches in the documents. The similar documents are aggregated together to determine a final relevancy score and the most similar documents are returned first.

(ii) *Retrieving Legal Information*

Researchers have explored three different approaches to legal information retrieval: (1) natural language and structured queries, (2) goal oriented, and (3) automated summarization.

a. Natural Language and Structured Queries

Das-Gupta [Das-Gupta, 1987] used a statistical approach and developed an algorithm which infers logical AND and OR operators based on natural language text queries. Logical AND and OR operators are used to construct a subset of boolean queries[15] to retrieve relevant information. Smith [Smith, 1990] furthered this idea by developing an algorithm that transforms natural language text queries into full boolean queries. For example a natural language text query of "John Smith and Robert Thompson but not Jane Smith" results in a search for documents containing "John Smith" and "Robert Thompson" and not containing "Jane Smith". Different boolean queries have also been developed and evaluated for retrieving legal information. Salton, Fox, and Wu [Salton, Fox and Wu, 1983] created extended boolean queries which provide weighted interpretations of each term. Thus allowing a query to specify which term is the most important in finding a match. For example, the search criteria of "(0.8)attorney OR (0.1)john" values a match of "attorney" higher than a match of "john" due to the weight value of 0.8 for

[15] Boolean queries are used as a means of finding relevant data. In boolean queries, operators are specified to construct a query. The operators are AND, OR, and NOT. An example query of "attorney AND john" would return data which contained the term "attorney" and "john." Alternatively, a query of "attorney OR NOT john" would return data which contained the term "attorney," or did not contain the term "john," or meet both requirements.

"attorney" and 0.1 for "john". An implementation of weighted boolean queries is currently used in the commercial legal research system from WestLaw. Unlike WestLaw, CLAS does not include boolean models to enable users to search for relevant documents.

b. Goal Oriented

Legal assessment is a goal-oriented method of determining the consequences from cases and defining the conditions required to enact those consequences. For example, in order to enact the death penalty in California, the defendant must be convicted of treason, first-degree murder with special circumstances (financial gain, multiple murders, explosives, etc…), train-wrecking causing the death of an innocent person, or perjury causing the death of an innocent person. In Valente and Breuker [Valente and Breuker, 1995], a system was developed that retrieves information relevant to legal assessment in two different modes. In the first mode, the goal is to find all the consequences which can be enacted based on a user-provided set of conditions. In the second mode, the goal is to find all conditions which are required to enact a user-provided consequence. The system is built on top of an existing legal reasoning system[16] [Breuker, 1993] which is used to identify consequences and conditions in cases. Improvements in legal reasoning systems [Valente and Breuker, 1994; Valente, 1995] led to improvements in legal assessment systems. In Valente [Valente, 1995], legal knowledge was modeled in three different forms, rule-based, case-based, and logic-based. These forms are used together to model legal assessment. The case-based form is used to store legal documents, the rule-based form is used to enact consequences when conditions are met, and the logic-based form is used to determine if conditions are met.

[16] Legal reasoning examines how attorneys analyze legal issues and develop arguments.

c. Automated Summarization

Automated summarization is a popular research topic in its own right. Extensive research has been conducted in summarizing news articles [Conroy and O'Leary, 2001; Nenkova 2005; McKeown and Radey, 1995; Radev et al., 2004]. News summarization focuses on capturing specific events in news articles. By design, news articles list the most important information at the beginning of an article and provide further details in the remainder of the article. Legal documents do not conform to this design. Legal documents contain information regarding events that have occurred in the past and how the framework of law can be applied to these events in order to enact a consequence. Summarizing legal documents requires a summarization of events, arguments, and laws being applied. Developing an effective legal summarization tool has proven to be very difficult. Hachey and Grover [Hachey and Grover, 2005] developed a system which determines the relevance of each sentence in a legal document. The relevance is judged based on whether the sentence should be included in an automatically generated summary. The sentences are evaluated using two different probabilistic classifiers, a naïve Bayes classifier[17], and a maximum entropy classifier[18]. The system was developed based on previous work in classifying sentences for use in an automatically generated summary [Hachey and Grover, 2004] and older legal summarization systems [Grover, Hachey, and Hughson, 2003].

(iii) *Evaluating Systems*

Researchers have also explored different methods of evaluating legal case-based reasoning systems. Evaluation frameworks have been developed using linear regression with neural

[17] Naïve Bayes Classifier is a simple probabilistic classification technique which applies Bayes' theorem (Bayesian statistics) with strong (naive) independence assumptions. Naïve Bayes assumes that all features are independent even if the features are actually dependent on each other.
[18] Maximum Entropy Classifier is similar to the Naïve Bayes Classifier except that the features are not treated independently.

networks, user acceptance surveys, system predictions compared against past case results, and system outputs evaluated by a panel of lawyers [Stranieri and Ballarat, 1999]. The goal of these frameworks is to provide enough variation in evaluation approaches to be able to evaluate any legal case-based reasoning system in a meaningful way.

While existing evaluation frameworks offer a wide variety of evaluation options, CLAS is evaluated using a set of annotated data. Because CLAS classifies the nature of the suit for Memorandum of Points and Authorities, it is evaluated using a collection of Memorandum of Points and Authorities that have been annotated with their nature-of-suit category.

(iv) *Hybrid Systems*

Legal case-based reasoning systems have also been integrated with other information retrieval research to form hybrid systems. The DataLex Legal Workstation (DataLex) [Greenleaf, Mowbray, and Tyree, 1991] combined a case-based reasoning system and free text retrieval system. DataLex allows the user to search for relevant legal documents using natural language text or a structured query language. If natural language text is used, the query is transformed into a structured query for use by the case-based reasoning system. The queries are also processed by the free text retrieval component to search for exact text matches in the documents. A different hybrid system combined a case-based reasoning system and information retrieval system [Rissland and Daniels, 1995]. This system accepts structured queries which are used by the case-based reasoning system to find relevant documents. The relevancy of documents is determined using a Bayesian inference[19] probabilistic model [Turtle and Croft, 1991]. Documents deemed relevant are returned to the information retrieval system which uses a set of internal relevance weights to

[19] Bayesian inference is a method of creating a probability of an event occurring using Bayes' Rule. The probability of an event occurring will change as more information about the event's causes are collected.

alter the relevancy of the documents. The documents are then displayed to the user. The user provides feedback indicating which documents they believe are relevant. Feedback from the user alters the internal relevancy weights in the information retrieval system which impacts later queries.

V. SELECTING & ACQUIRING MEMORANDA OF POINTS AND AUTHORITIES

In California, the trial courts, which serve as the first level courts, are called "superior courts."[20] The California trial court system is organized by county, thus consists of 58 county superior courts. Each county operates fairly autonomously, managing its own court house locations, employees, procedural rules, websites, and document storage systems.

The Sacramento Superior Court was selected as the document source because, unlike the majority of trial courts in California, the Sacramento Superior Court (1) has an electronic "Public Case Access System" that offer's users the ability to download (as PDF files) certain court documents filed as part of an unlimited civil case[21] and (2) requires the initiating attorney in each unlimited civil case to simultaneously file a case coversheet that identifies the nature of suit with the initiating papers. In identifying the nature of the suit, the filing attorney has the option to select from 39 pre-established categories. A complete list of the 39 pre-established nature-of-suit categories, with a brief description of each is provided in Table 1, below.

[20] The secondary and tertiary levels of judicial review in California are called the Court of Appeals and the California Supreme Court, respectively.

[21] "Unlimited civil case" broadly describes the case category and identifies what court department will hear the case. For example, a civil case is heard in the general civil department of the court and is a noncriminal lawsuit brought to redress an alleged private wrong (*e.g.* breach of contract claims, wrongful termination claims, and property damage claims) and usually involves disputes between persons or organizations. Other common judicial departments include criminal, family law, juvenile, and small claims. The term "unlimited" means that the amount of damages potentially at issue in the civil case exceeds some pre-determined dollar value, which in Sacramento is $25,000 (this pre-determined value may vary by county).

Nature of Suit Category	Description
Antitrust/Trade Regulation	Protect trade and commerce from unfair restraints, monopolies, and price fixing.
Asbestos	Damages from asbestos used as insulation or as a fire retardant.
Asset forfeiture	The relinquishment of money or property as a consequence of a legal action.
Breach of Contract/Warranty	Failing to meet requirements in a contract or warranty.
Business Tort	Wrong doing to a corporation causing it to lose intangible assets such as reputation or intellectual property.
Civil Rights	Individual rights or privileges bestowed on an individual.
Construction Defect	Construction defect leading to damages in design, materials, or other.
Contract - Other	Contract issues which do not fall into Breach of Contract/Warranty.
Defamation	Communication that causes damage to a person's reputation which may lead to other damages.
Eminent domain/Inverse condemnation	Legal action available to a property owner who is not fairly compensated for property taken from him/her by a government for a public use.
Enforcement	Legal action available to provide enforcement of a previous action which was not followed.
Fraud	Intentional misrepresentation or nondisclosure leading to damages.
Harassment	Actions which threaten, intimidate, or create fear for a person's safety.
Insurance Coverage	Financial coverage for events listed in the insurance coverage policy.
Insurance Coverage Claims	Claims in order to obtain financial remedy for insurance coverage.
Intellectual Property	Property right for intangible property such as ideas, discoveries, and inventions.
Judicial Review - Other	Legal action to review the three wings of the government, legislative, executive, and administrative.
Mass Tort	Collective injury to a group of people, such as an airplane crash or damaging food product.
Medical Malpractice	Failure of a medical professional to follow the standards of practice resulting in damage to the patient.
Misc Complaints - Other	Legal action for complaint which does not fall into any other nature-of-suit category.
Non-PI/PD/WD tort - Other	Wrong doing to a person which does not involve personal injury, property damage, or wrongful death.
Other Collections	Collections of debt from a person or organization which are greater than or equal to $25,000.
Other employment	Employment related legal action which does not fall into Wrongful Termination.
Other Real Property	Rights to ownership and future ownership of property.
Petition re: Arbitration Award	Determination of arbitration award from arbitration tribunal, analogous to judgment of a court.
Petitions - Other	Request to the count to take specific action defined in the petition.
PI/PD/WD - Auto	Personal injury, property damage, or wrongful death related to automobile collision or accident.

Nature of Suit Category	Description
PI/PD/WD - Other	Personal injury, property damage, or wrongful death which does not fall into PI/PD/WD - Auto or PI/Property Damage/Wrongful Death.
PI/Property Damage/Wrongful Death	Legal action pertaining to personal injury, property damage, or wrongful death.
Product Liability	Liability of manufacturer for damages caused by defective merchandise.
Professional Negligence	Failure to follow the standards of practice resulting in damages.
Rule 3.740 Collections	Collections of debt from a person or organization which is under $25,000.
Toxic Tort/Environmental	Damages from toxic environmental substance.
Uninsured Motorist	Legal action due to accident or collision with uninsured motorist.
Unlawful Detainer - Commercial	Retaining possession of property without legal right.
Writ of Mandate	Court order to a government agency to follow the law by correcting prior actions or ceasing illegal acts.
Wrongful Eviction	Removal of tenant from property without following proper legal procedures.
Wrongful Termination	Termination of an employment contract by employer which breaches employment contract.
N/A	Case does not fall into any other category.

Table 1 - Available Categories for the "Nature of the Suit" as established by the Sacramento Superior Court

CLAS is designed to classify the nature of suit for Memorandum of Points and Authorities into one of the 39 categories provided in Table 1.

As mentioned previously, Memorandum of Points and Authorities are supplemental documents filed simultaneously alongside a motion. They were selected as the target documents for this thesis because they detail the underlying legal research and legal support used to bolster the specific legal arguments in favor of the motion, and thereby hold valuable legal information regarding how the law(s) have been applied or should be applied. Additionally, because Memorandum of Points and Authorities are filed to persuade a judge to apply a law in the requested manner, they also will summarize the underlying facts of the lawsuit, and thus supply enough information to gain insight as to the nature of the suit. Finally, because motions are filed throughout the each stage of a lawsuit (PLEADING-stage, DISCOVERY-stage, TRIAL-stage, and

POST-TRAIL-stage), Memorandum of Points and Authorities are also prominent in all four stages. Thus, these documents contain valuable data that CLAS uses to determine the nature of the suit; such as parties in the lawsuit, non-party Entities involved in the lawsuit, and past events that have caused the lawsuit.

In order to obtain an adequate number of electronic Memorandum of Points and Authorities (to both train and test the system), a web scraping system called the Sacramento Civil Lawsuit Scraper (SCLS) was developed to automatically collect Memorandum of Points and Authorities from the Sacramento Superior Court's "Public Case Access System." The creation of SCLS was necessary because manual collection of the documents did not offer a realistic option due to a number of search limitations imposed by the Sacramento Superior Court's "Public Case Access System."

First, the Public Case Access System only allows users to search for cases by name, case number, or filing date. (Sacramento Superior Court, "Public Case Access System," https://services.saccourt.ca.gov/PublicCaseAccess/Civil). There is no keyword search. A user cannot limit searches to pull only certain types of case documents or only certain nature of suit types. Nor can a user pull documents from multiple cases at one time. From a legal research perspective, these search limitations significantly limit the value of the Public Case Access System.

For example, when searching previously filed court documents lawyers search for a particular type of document in a particular nature of suit. Identifying the type of document is valuable because it is likely that the many of the overarching legal principals will remain the same across document types. For example, all Memorandum of Points and Authorities in support of Motions to Dismiss a Complaint for Failure to State a Claim, discuss the legal standard regarding when a Complaint must be dismissed – regardless of the type of case, this procedural standard will

remain consistent. For the most part it is easy to identity document type by the title of each document. For example, a Complaint would include "COMPLAINT" in the title, a Motion to Compel would include "MOTION TO COMPEL" in the title, and an Opposition to a Motion to Compel would include "OPPOSITION TO MOTION TO COMPEL" in the title.

However, locating information about the nature of the suit is not as readily apparent. Filtering searches by nature of suit allows a user to pool documents with similar facts to their current legal dispute. This filtering is likely to increase the value of the search results by providing lawyers with applications of the law in similar factual situations.

However, neither document title searches nor nature of suit filters are possible in the Public Case Access System. Instead, a user must either know of a specific lawsuit name or lawsuit number (a "case number") or can wade through a plethora of cases filed within a specified time frame. If a lawyer does have an exact lawsuit name or case number, they can locate that particular lawsuit and view court documents filed within that lawsuit.

While the "search by filing date" option does allow for a user to pull more than one lawsuit at a time (search results are limited to 1,000 results), the search provides a list of lawsuits initiated during the specified timeframe. Thus, the Public Case Access System does not return a list of all court documents filed during the specified time frame. Figure 13, below, provides an example screen shot of the search results page.

Public Case Access
Sacramento Superior Court

Log On | Create Account | Forgot Password? | Find Order | Document Cart (0)

Home | Civil | Criminal | Family | Probate | Small Claims | Traffic | Unlawful Detainer

Civil Cases
Search by Name
Search by Case
Search by Filing Date

Tentative Rulings
Search By Case
Search By Department

General Info
How This Site Works
Civil Home Page on Saccourt
Arbitrator Selection Process
Trial Readiness Notification
Trial Setting Process
Complex Case Calendar

Search Results - Civil
Filing Date Range: 1/7/2013 - 1/7/2013

Back to Search By Filing Date

	Case Number	Case Title	Filing Date	Case Type
View	2013-00137987	Helene Karcher vs. Sutter Health	01/07/2013	Medical Malpractice
View	2013-00137998	Portfolio Recovery Associates LLC vs. Bernice Phillips	01/07/2013	Rule 3.740 Collections
View	2013-00138001	Midland Funding LLC vs. Jeffrey Price	01/07/2013	Rule 3.740 Collections
View	2013-00138039	Portfolio Recovery Associates LLC vs. Stephanie M Kamanu	01/07/2013	Rule 3.740 Collections
View	2013-00138040	Christopher Garcia vs. Daniel Reid Thomas	01/07/2013	Breach of Contract/Warranty
View	2013-00138064	Caroline Isaac vs. Aaron Claude Fink	01/07/2013	PI/PD/WD - Auto
View	2013-00138067	Genevieve Lawson vs. California Highway Patrol	01/07/2013	Wrongful Termination
View	2013-00138081	Petition of Devon Chase Zanter	01/07/2013	Petitions - Other
View	2013-00138105	Capital One Bank USA N A vs. Robert J Gibber	01/07/2013	Rule 3.740 Collections
View	2013-00138122	Investment Retrievers vs. Saddoris	01/07/2013	Enforcement
View	2013-00138136	Midland Funding LLC vs. Ramon Fernandez	01/07/2013	Rule 3.740 Collections
View	2013-00138139	Midland Funding LLC vs. Powell Kennedy	01/07/2013	Rule 3.740 Collections
View	2013-00137969	Response Indemnity Company of California as Subrogee vs. Max Rasumussen	01/07/2013	PI/PD/WD - Auto
View	2013-00137973	Ann Oluwadare vs. Martin Huang	01/07/2013	PI/PD/WD - Auto
View	2013-00137992	Ford Motor Credit Company LLC vs. David Collenberg	01/07/2013	Rule 3.740 Collections
View	2013-00138005	Portfolio Recovery Associates LLC vs. Ernesto Castillo	01/07/2013	Rule 3.740 Collections
View	2013-00138006	Sierra Central Credit Union vs. Vera Williams	01/07/2013	Breach of Contract/Warranty
View	2013-00138017	Union Bank NA vs. Janak K Mehtani Trustee	01/07/2013	Breach of Contract/Warranty
View	2013-00138032	David P Stone vs. Devry Inc	01/07/2013	Other employment

Figure 13 - Excerpt from the Sacramento Superior Courts Public Case Access System Search Results Page for Filing Date January 7, 2013

The lawsuit search results page in Figure 13 shows a list of lawsuits that were initiated on the date searched (1/7/2013). While these results include a nature-of-suit category for each lawsuit (the "Case Type" column in the Figure 13), the results do not provide any information on the types of documents filed in the lawsuits. Furthermore, the user has no method of filtering the search by nature of suit or document type. In order to access each respective lawsuit's documents, a user must click each individual "view" button in order to see the individual case details and to access a

list of the lawsuit's court documents ("docket"). There is no method to simultaneously view documents filed in multiple cases. There is also no method to pull only a certain type of document.

Figure 14, below, is a screen shot of the lawsuit summary details provided by the Sacramento Superior Court in a medical malpractice lawsuit titled "*Helene Karcher vs. Sutter Health*." Figure 15, below, provides an excerpted screen shot of the Sacramento Superior Court's docket for that same medical malpractice lawsuit; the red rectangle indicates a Memorandum of Points and Authorities filed in the lawsuit.

Public Case Access
Sacramento Superior Court

Log On | Create Account | Forgot Password? | Find Order | Document Cart (0)

Home | Civil | Criminal | Family | Probate | Small Claims | Traffic | Unlawful Detainer

Civil Cases
Search by Name
Search by Case
Search by Filing Date

Tentative Rulings
Search By Case
Search By Department

General Info
How This Site Works
Civil Home Page on Saccourt
Arbitrator Selection Process
Trial Readiness Notification
Trial Setting Process
Complex Case Calendar

Civil Case Details

Case Information

Case Title	Helene Karcher vs. Sutter Health		
Case Number	34-2013-00137987-CU-MM-GDS	**Case Type**	Medical Malpractice
Filed Date	01/07/2013	**Case Category**	Civil - Unlimited

Participants

Participant Name	Role	Represented By
Amaral PA, Steven	Defendant	Thomas J Doyle
Amaral PA, Steven	Defendant	Thomas J Doyle
Does 1-100	Defendant	
Karcher, Edmund	Plaintiff	Khaldoun A Baghdadi
Karcher, Helene	Plaintiff	Khaldoun A Baghdadi
Rabii MD, Cyrus	Defendant	Thomas J Doyle
Rabii MD, Cyrus	Defendant	Thomas J Doyle
Schneidewind DO, Barry	Defendant	Thomas J Doyle
Schneidewind DO, Barry	Defendant	Thomas J Doyle
Sutter Health	Defendant	
Sutter Medical Foundation	Defendant	Thomas J Doyle
Sutter Medical Foundation	Defendant	Thomas J Doyle
Sutter Roseville Medical Center	Defendant	Ronald R Lamb

Figure 14 - Excerpted Screen Shot of Sacramento Superior Court's Lawsuit Summary Data case number 34-2013-00137987-CU-MM-GDS (Helene Karcher vs. Sutter Health)

					Document Cart (0)	
ROA#	ROA Entry	Filed Date	Filed By	Pages		Cart
52	Civil Settlement Conference - Civil Settlement Conference scheduled for 09/17/2014 at 01:30:00 PM in Department 59 at Gordon D Schaber Courthouse was vacated .	12/17/2014				
51	Case disposed with disposition of Request for Dismissal .	12/17/2014				
50	Case dismissed with disposition of Request for Dismissal .	12/10/2014				
49	Amended Complaint disposed with disposition of Request for Dismissal .	12/17/2014				
48	Request for Dismissal with Prejudice - Entire Action filed.	12/10/2014	Karcher, Edmund(Plaintiff); Karcher, Helene(Plaintiff)	3	Preview	
47	Notice of Conditional Settlement filed.	10/15/2014	Karcher, Edmund(Plaintiff); Karcher, Helene(Plaintiff)	3	Preview	
46	Civil Trial Assignment - Long Cause - Civil Trial Assignment scheduled for 10/28/2014 at 08:30:00 AM in Department 47 at Gordon D Schaber Courthouse was vacated .	10/15/2014				
45	Minute Order Following Mandatory Settlement Conference filed.	09/17/2014		1	Preview	
44	Notice of Entry of Dismissal filed.	09/09/2014	Karcher, Edmund(Plaintiff); Karcher, Helene(Plaintiff)	4	Preview	
43	Case disposed with disposition of .	07/24/2014				
42	Case dismissed with disposition of Request for Dismissal .	07/23/2014				
41	Request for Dismissal with Prejudice - Party (Sutter Roseville Medical Center ONLY) filed.	07/23/2014	Karcher, Edmund(Plaintiff); Karcher, Helene(Plaintiff)	2	Preview	
40	Minutes finalized for Motion to Compel - Other - Civil Law and Motion heard on 05/15/2014 02:00:00 PM .	05/15/2014		1	Preview	
39	Correspondence (Confirming Letter - 5/15/14 hearing dropped from calendar) filed.	05/02/2014	Schneidewind DO, Barry(Defendant)	2	Preview	
38	Motion to Compel - Other - Civil Law and Motion scheduled for 05/15/2014 at 02:00:00 PM in Department 53 at Gordon D Schaber Courthouse .	04/15/2014				
37	Declaration - Other (of Thomas J. Doyle) filed.	04/14/2014	Amaral PA, Steven(Defendant); Rabii MD, Cyrus(Defendant); Schneidewind DO, Barry(Defendant); Sutter Medical Foundation(Defendant)	31	Preview	
36	Statement - Other filed.	04/14/2014	Amaral PA, Steven(Defendant); Rabii MD, Cyrus(Defendant); Schneidewind DO, Barry(Defendant); Sutter Medical Foundation(Defendant)			
35	Memorandum of Points and Authorities filed.	04/14/2014	Amaral PA, Steven(Defendant); Rabii MD, Cyrus(Defendant); Schneidewind DO, Barry(Defendant); Sutter Medical Foundation(Defendant)	10	Preview	
34	Motion to Compel - Other filed.	04/14/2014	Amaral PA, Steven(Defendant); Rabii MD, Cyrus(Defendant); Schneidewind DO, Barry(Defendant); Sutter Medical Foundation(Defendant)	3	Preview	

Figure 15 - Excerpted Screen Shot of the Sacramento Superior Court's Docket Sheet for case number 34-2013-00137987-CU-MM-GDS (Helene Karcher vs. Sutter Health)

To download the available court documents in each case, the user then must select the box in the far right column on the docket sheet (Figure 15) and proceed through a "checkout" page, which requires inputting of personal information and subsequent email verification.

The SCLS system automated the otherwise time-intensive process of navigating through the court website to locate lawsuits, navigate to the lawsuit summary pages, and review each document in the lawsuits to locate Memorandum of Points and Authorities. At the same time, SCLS also recorded any lawsuit summary data provided on the "Civil Case Details" page (Figure 14, above). This summary data included a court-assigned case number, plaintiffs, defendants, and the nature of the suit, which could then be related back to each of the downloaded documents using a document identification value and court-assigned case number. Table 2, below, illustrates the lawsuit summary information scraped from the Sacramento Superior Court's website.

Court Assigned Case Number	Nature of Suit	Plaintiff(s)	Defendant(s)
2012-00116470	Fraud	Robert M. Espinoza; Wilma Maria Espinoza	U.S. Bank National Association; DSL Service Company; FCI Lender Services, Inc.
2010-80000413	Writ of Mandate	David Wojan	Therese Delgadillo; California Department of Developmental Services
2012-00119118	PI/PD/WD - Other	Mary Cooper; Larry Cooper	Delta Skilled Nursing Center Inc.; Delta Rehabilitation and Care Center
2010-00067784	Breach of Contract/Warranty	Stephen Furman	Hayes Family Enterprise Inc.; Brett W. Hayes
2010-00068759	PI/PD/WD - Auto	Thomas Harris II; Ingersoll-Rand Company	Joseph Urbon; Complete Equipment Repair

Table 2 - Court Provided Lawsuit Summary Data (Multiple plaintiffs and defendants are delimited by a semicolon.)

In total, SCLS downloaded 10,000 Memorandum of Points and Authorities, from 5,659 distinct lawsuits, and scraped the lawsuit summary data for each lawsuit (Table 2). SCLS located Memorandum of Points and Authorities by scraping the documents in each lawsuit and locating the titles which contained the text "Memorandum of Points and Authorities filed" in the "ROA

Entry" column seen in Figure 15, above. In order to obtain the 10,000 Memorandum of Points and Authorities, SCLS scraped 826,110 documents across 21,391 different lawsuits. The reason such a large number of documents and lawsuits were scraped prior to obtaining 10,000 Memorandum of Points and Authorities was because many lawsuits were settled early in the PLEADING-stage process and did not contain any motions or supporting documents. Another issue was that attorneys included Memorandum of Points and Authorities as part of their motions and submitted the two documents as a single motion document. This led the SCLS system to identify the single document as a motion because it was labeled as such on the Sacramento Superior Court website.

Based on the lawsuit summary data scraped from the Sacramento Superior Court's website, there was at least one Memorandum of Points and Authorities for each of the 39 case type categories. However, over half the documents were filed as a part of a lawsuit classified into one of the following five categories:

- "Breach of Contract/Warranty,"
- "PI/PD/WD – Other,"
- "Other Real Property,"
- "PI/PD/WD – Auto," and
- "Medical Malpractice."

Together these five categories account for 5,319 or 53.19% of the 10,000 Memorandum of Points and Authorities used in this thesis. These same five categories account for 2,914 or 51.49% of the 5,659 distinct lawsuits represented by the documents used in this thesis.

According to the Sacramento Superior Court's classifications, the category of Breach of Contract/Warranty alone accounts for over fifteen percent (15.43%) of all of the documents with

PI/PD/WD – Other in close second, accounting for 13.80% of the scraped documents. Table 3, below, provides the actual document count and distinct lawsuit count for each category, as classified by the Sacramento Superior Court. The table also includes the average number of documents per case, highest number of documents in a single case, and lowest number of documents in a single case for each category.

Nature of Suit Category	Number of Documents	Number of Lawsuits	Average Documents per Lawsuit	Highest Number of Documents in a Lawsuit	Lowest Number of Documents in a Lawsuit
Breach of Contract/Warranty	1543	844	1.8294	23	1
PI/PD/WD - Other	1380	685	2.0146	21	1
Other Real Property	956	537	1.7803	17	1
PI/PD/WD - Auto	883	606	1.4571	30	1
Medical Malpractice	556	242	2.2975	35	1
Fraud	446	215	2.0744	22	1
Wrongful Termination	427	214	1.9953	12	1
Misc. Complaints - Other	426	248	1.7177	11	1
Rule 3.740 Collections	393	390	1.0077	2	1
N/A	351	176	1.9943	12	1
Other Collections	335	249	1.3454	12	1
Petitions - Other	265	213	1.2441	7	1
Business Tort	264	98	2.6939	23	1
Professional Negligence	251	100	2.5100	19	1
Contract - Other	189	107	1.7664	20	1
Writ of Mandate	151	115	1.3130	18	1
Civil Rights	144	74	1.9459	9	1
Non-PI/PD/WD tort - Other	144	73	1.9726	9	1
Construction Defect	137	76	1.8026	10	1
Defamation	108	44	2.4545	11	1
Insurance Coverage	106	53	2.0000	12	1
Other employment	101	48	2.1042	15	1
Product Liability	94	47	2.0000	11	1
Eminent domain/Inverse condemnation	90	57	1.5789	12	1
Asbestos	58	14	4.1429	27	1
Unlawful Detainer – Commercial	43	28	1.5357	4	1
Wrongful Eviction	36	15	2.4000	11	1
Petition re: Arbitration Award	30	22	1.3636	6	1
Enforcement	25	18	1.3889	4	1
Uninsured Motorist	22	16	1.3750	3	1
Asset forfeiture	19	19	1.0000	1	1

Judicial Review - Other	9	6	1.5000	3	1
PI/Property Damage/Wrongful Death	6	2	3.0000	3	3
Harassment	3	1	3.0000	3	3
Insurance Coverage Claims	3	1	3.0000	3	3
Intellectual Property	2	2	1.0000	1	1
Toxic Tort/Environmental	2	2	1.0000	1	1
Antitrust/Trade Regulation	1	1	1.0000	1	1
Mass Tort	1	1	1.0000	1	1

Table 3 - Summary of Nature of Suit Classification, as Identified by the Sacramento Superior Court, in descending order by document count

All court documents are on what is called "pleading paper,"[22] which contains numbers down the left hand margin. These numbers serve as line references and allow for easy citing of these documents. Additionally, every court document has a cover page called a "caption page" that contains formatting and other images such as stamps, handwriting, headers, and footers that are unique to that individual document. Figure 16 - Sample Court Document Caption Page and Figure 17 - Sample Court Document Body Page, below, provide examples of a court document caption page and a body page.

[22] All Filings, whether a pleading or not, are on what is commonly referred to as "pleading paper". Attorneys often use the terms pleading and filing synonymously in casual conversation.

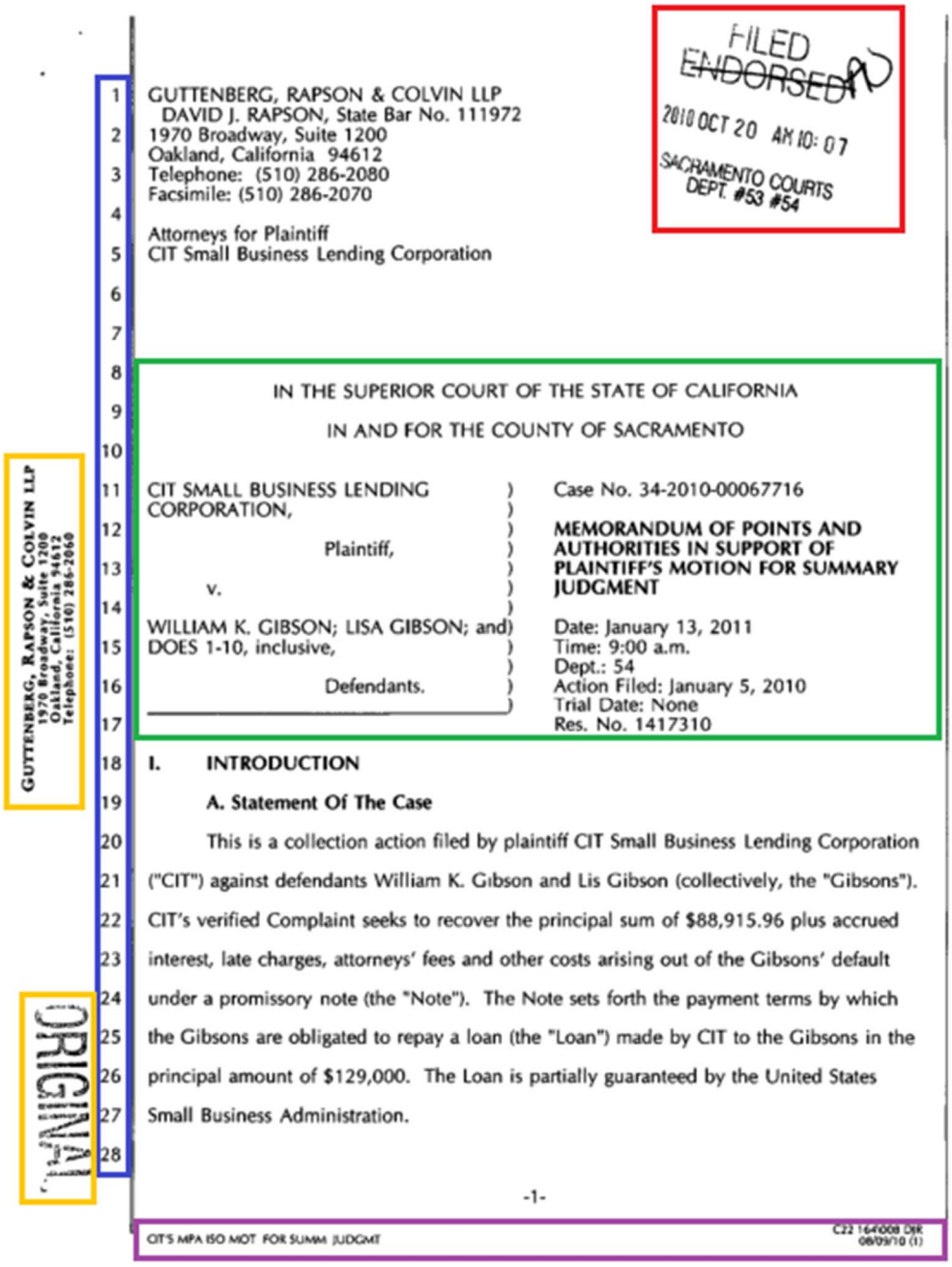


Figure 16 - Sample Court Document Caption Page

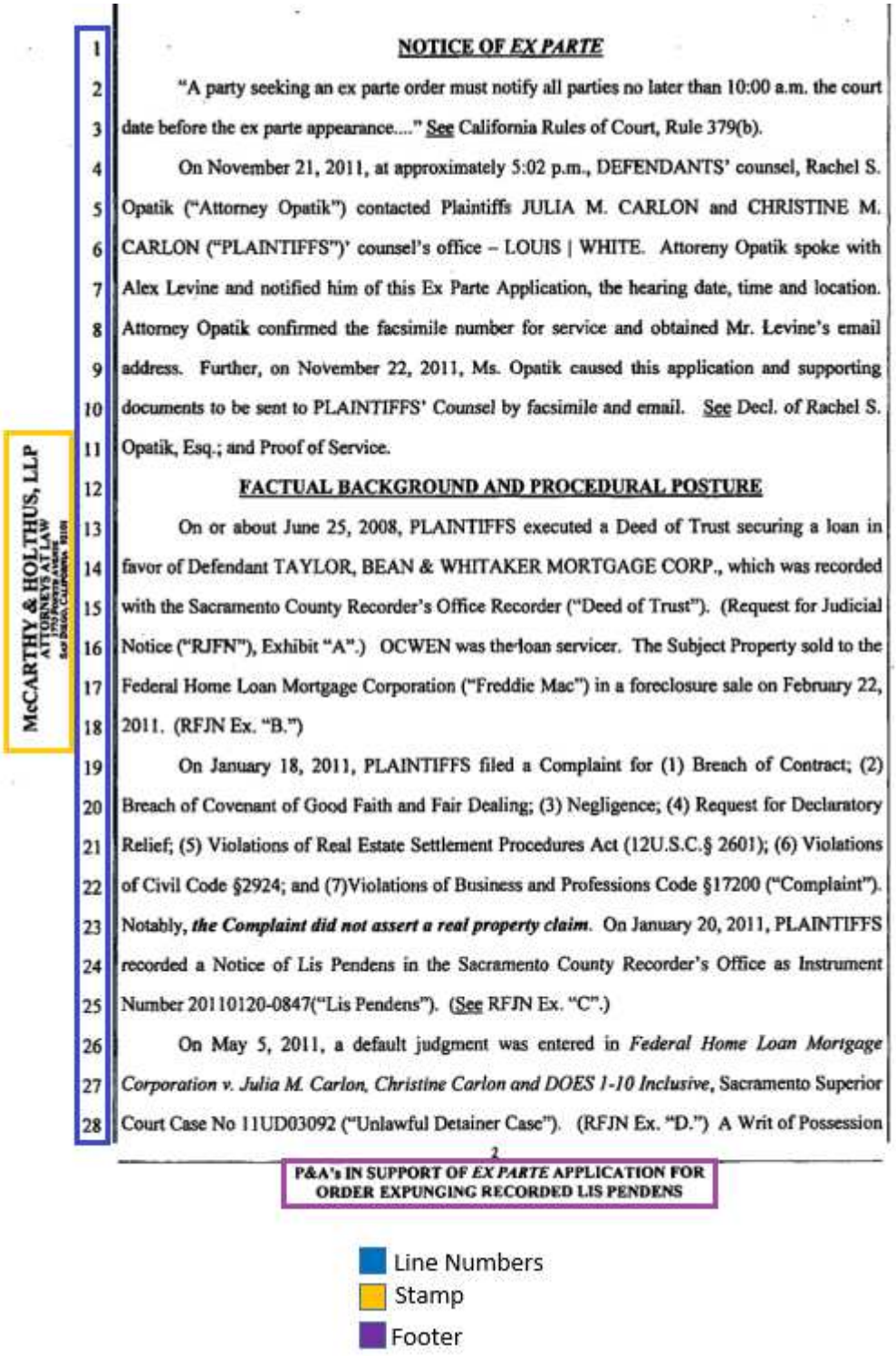


Figure 17 - Sample Court Document Body Page

VI. CLAS COMPONENTS

CLAS classifies the nature of a lawsuit based on information contained in a subset of attorney-authored court documents called Memorandum of Points and Authorities that are filed with trial courts during the process of the lawsuit. CLAS is limited to Memorandum of Points and Authorities because unlike many of the other documents filed during a lawsuit, these documents detail the legal research and support used to strengthen the arguments in the lawsuit. Memorandum of Points and Authorities also summarize the events that caused the lawsuit. When attorneys are attempting to locate documents that contain similar issues and arguments to their current lawsuit, they attempt to locate lawsuits with the same nature of suit. Once they have located a relevant lawsuit, they review the documents in the lawsuit to locate the particular document type they are looking for. While locating a particular type of document within a lawsuit can be done by reviewing the titles of each document, no simple solution exists to determine the nature of a lawsuit without manual review.

CLAS is divided into three components. These components are (1) document preprocessing, (2) Key Phrase Selection, and (3) Nature-of-Suit Classifier. These three components are executed in order because each component depends on the prior component.

The document preprocessing component extracts machine readable text from Memorandum of Points and Authorities. The Key Phrase Selection component extracts Key Phrases from the machine readable text. The Key Phrase Selection component extracts only those sentences that (1) reference persons, organizations, and other proper nouns (collectively called “Entities”), (2) reference at least one party in the lawsuit, and (3) describe an interaction between two Entities that occurred in the past. By capturing these sentences, CLAS deliberately focuses on information about the nature of the suit Thereby, excluding sentences that describe legal arguments or legal

analysis because they do not characterize the nature of the suit. The extracted Key Phrases are used to train the last component, the Nature-of-Suit Classifier. The following sections detail the technical implementation of each component.

A. Document Preprocessing

(i) *Extracting Machine Readable Text from the Documents*

After acquiring the court documents, it is necessary to extract machine readable text from each of them. Unfortunately, two challenges complicated this otherwise fairly straightforward step. First, of the 10,000 Memorandum of Points and Authorities scraped from the Sacramento Superior Court website, only 852 contain machine readable text, causing the remaining 9,148 to require additional processing in order to render their text machine readable and extractable. Second, even for the 852 documents that contain machine readable text, the layout and formatting of the documents limits the degree to which the text is extractable.

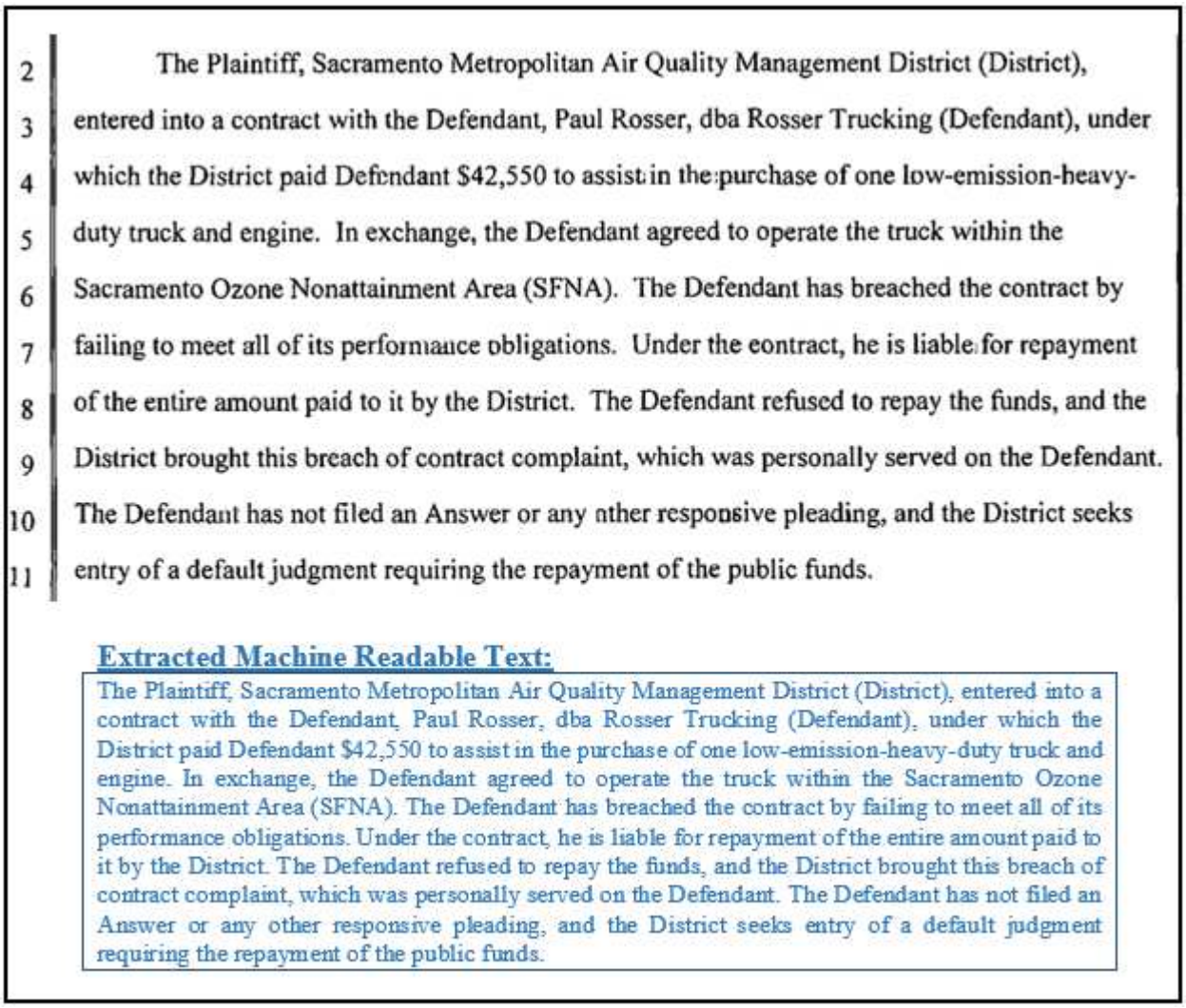
For the 9,148 documents without machine readable text an optical character recognition engine is used to render machine readable and extractable text.[23] The first step in this process requires conversion of the PDF documents into a digital image file called Portable Network Graphics (PNG). Next, the PNG files are run through the optical character recognition engine, which is capable of identifying and extracting machine readable text from a PNG file.

However, the extraction of machine readable text is complicated by the layout and formatting of the documents. Figure 16 and Figure 17, above, provide examples of a court document caption page and a body page. The court documents are neither uniform nor simple

[23] The optical character recognition engine was called Tesseract-ocr. Tessaract-ocr was originally developed at HP labs from 1985 to 1995. Few improvements were made until 2006, when Google continued its development and released it as open source. ("Tesseract-ocr," is available at https://code.google.com/p/tesseract-ocr/).

text, and thus the quality of extracted machine readable text varies. Below is a comparison of the extractions from three different court documents.

The first example, in Figure 18, below, is an excerpt from a Motion of Points and Authorities filed in 2011. This document is one of the 852 documents that contained machine readable text without needing any additional processing. The quality of the extracted machine readable text in this example is very high.



The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine. In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA). The Defendant has breached the contract by failing to meet all of its performance obligations. Under the contract, he is liable for repayment of the entire amount paid to it by the District. The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant. The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

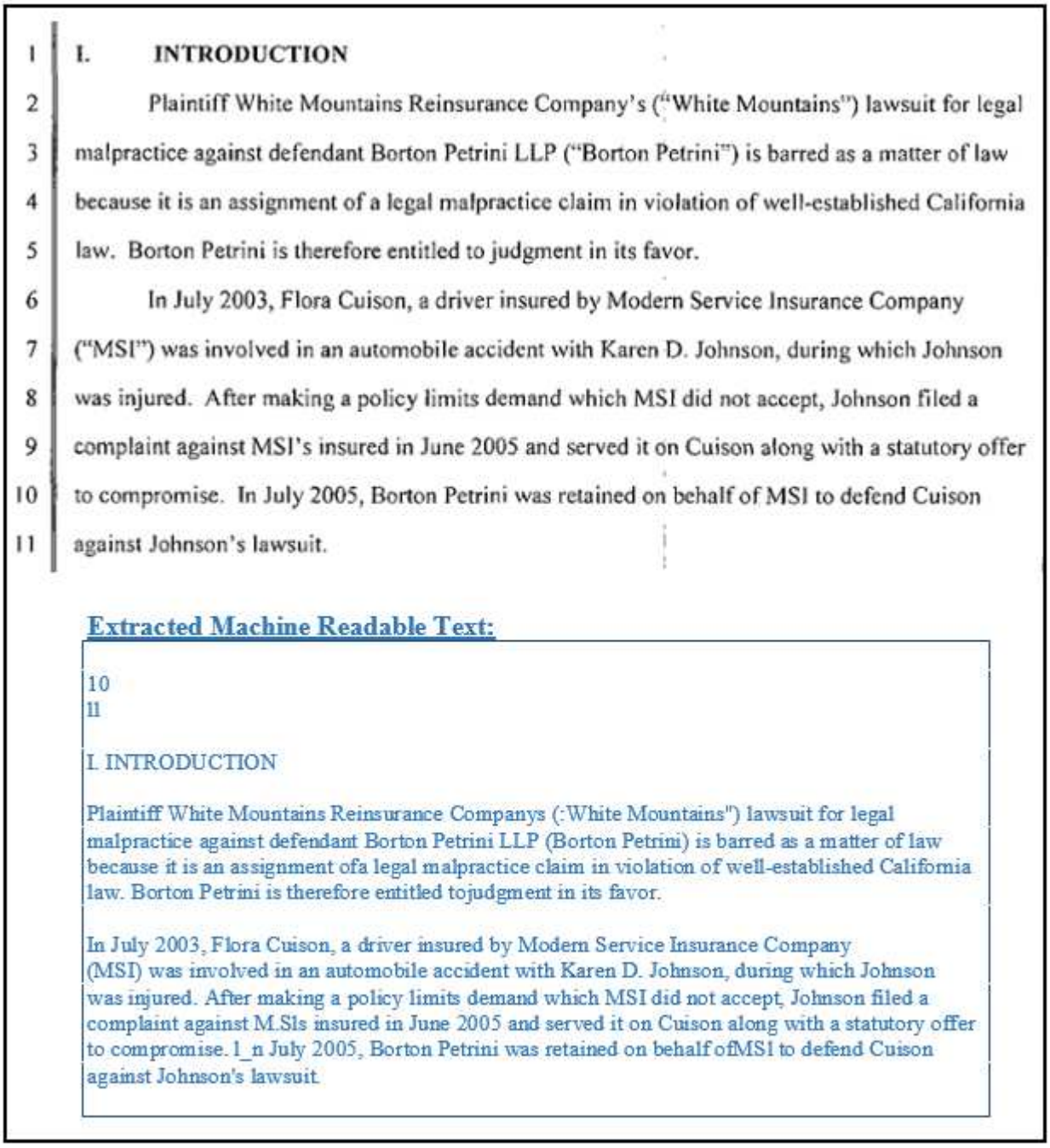
Extracted Machine Readable Text:

The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine. In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA). The Defendant has breached the contract by failing to meet all of its performance obligations. Under the contract, he is liable for repayment of the entire amount paid to it by the District. The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant. The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

Figure 18 - Extracting Machine Readable Text, Excerpt from Memorandum of Points and Authorities filed in 2011

Figure 19, below, provides an example of extracted text from a document that does not originally contain machine readable text. The optical character recognition engine is able to extract

all of the text from the excerpt, however, the extraction lacks the fluidity of the previous example and is riddled with formatting and grammatical errors.



I. INTRODUCTION

Plaintiff White Mountains Reinsurance Company's ("White Mountains") lawsuit for legal malpractice against defendant Borton Petrini LLP ("Borton Petrini") is barred as a matter of law because it is an assignment of a legal malpractice claim in violation of well-established California law. Borton Petrini is therefore entitled to judgment in its favor.

In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company ("MSI") was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured. After making a policy limits demand which MSI did not accept, Johnson filed a complaint against MSI's insured in June 2005 and served it on Cuison along with a statutory offer to compromise. In July 2005, Borton Petrini was retained on behalf of MSI to defend Cuison against Johnson's lawsuit.

Extracted Machine Readable Text:

10
ll

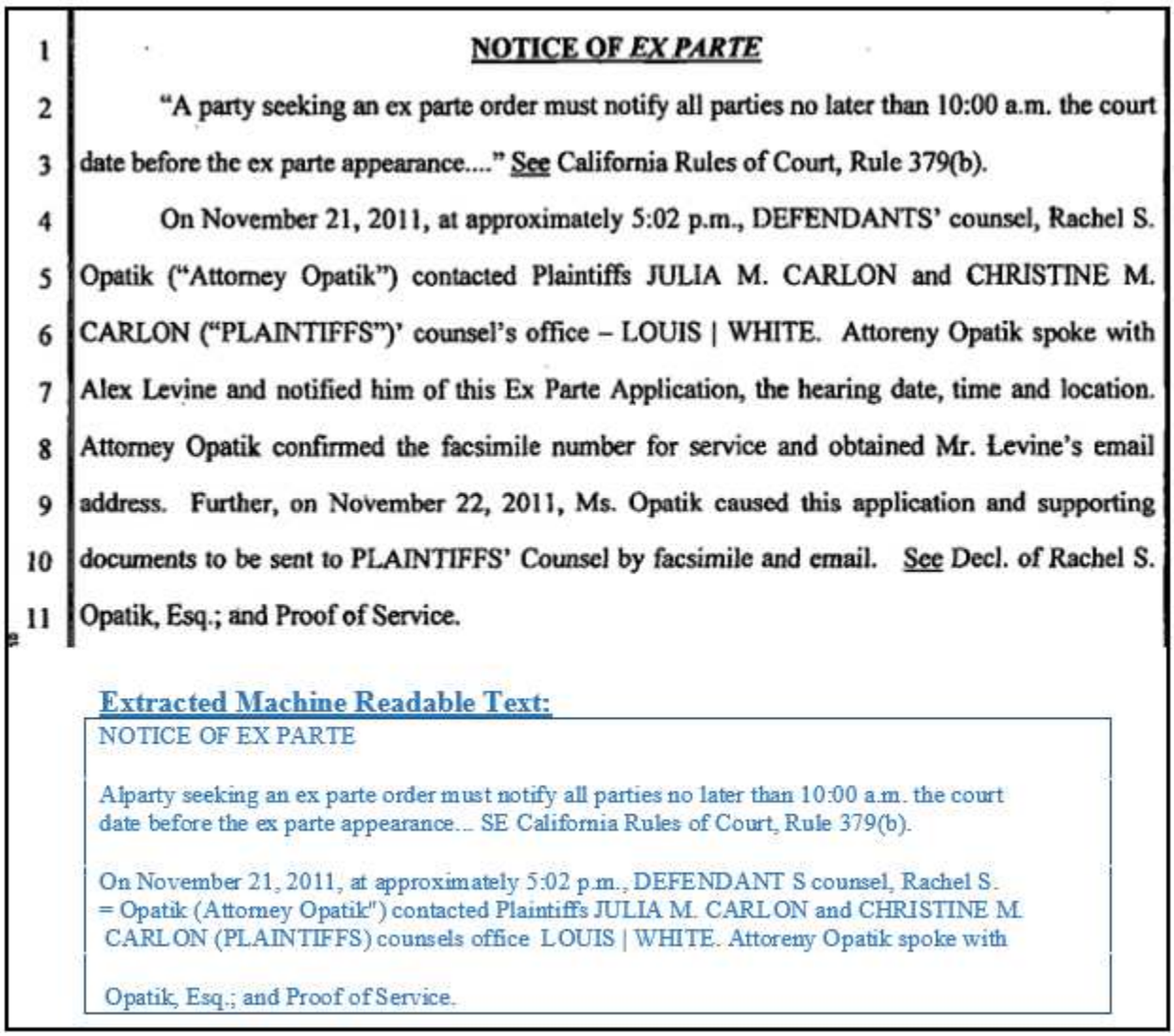
I. INTRODUCTION

Plaintiff White Mountains Reinsurance Companys (:White Mountains") lawsuit for legal malpractice against defendant Borton Petrini LLP (Borton Petrini) is barred as a matter of law because it is an assignment ofa legal malpractice claim in violation of well-established California law. Borton Petrini is therefore entitled tojudgment in its favor.

In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company (MSI) was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured. After making a policy limits demand which MSI did not accept, Johnson filed a complaint against M.SIs insured in June 2005 and served it on Cuison along with a statutory offer to compromise. l_n July 2005, Borton Petrini was retained on behalf ofMSI to defend Cuison against Johnson's lawsuit.

Figure 19 - Extracting Machine Readable Text, Excerpt from Memorandum of Points and Authorities filed in 2012

Finally, a low quality machine readable text is exemplified in Figure 20, below. In this extraction, lines 7 through 10 altogether failed to produce any machine readable text. The gaps in extracted text are a result of the typeface and bold font used in the excerpt.



NOTICE OF *EX PARTE*

"A party seeking an ex parte order must notify all parties no later than 10:00 a.m. the court date before the ex parte appearance...." See California Rules of Court, Rule 379(b).

On November 21, 2011, at approximately 5:02 p.m., DEFENDANTS' counsel, Rachel S. Opatik ("Attorney Opatik") contacted Plaintiffs JULIA M. CARLON and CHRISTINE M. CARLON ("PLAINTIFFS")' counsel's office – LOUIS | WHITE. Attoreny Opatik spoke with Alex Levine and notified him of this Ex Parte Application, the hearing date, time and location. Attorney Opatik confirmed the facsimile number for service and obtained Mr. Levine's email address. Further, on November 22, 2011, Ms. Opatik caused this application and supporting documents to be sent to PLAINTIFFS' Counsel by facsimile and email. See Decl. of Rachel S. Opatik, Esq.; and Proof of Service.

Extracted Machine Readable Text:

NOTICE OF EX PARTE

Alparty seeking an ex parte order must notify all parties no later than 10:00 a.m. the court date before the ex parte appearance... SE California Rules of Court, Rule 379(b).

On November 21, 2011, at approximately 5:02 p.m., DEFENDANT S counsel, Rachel S. = Opatik (Attorney Opatik") contacted Plaintiffs JULIA M. CARLON and CHRISTINE M. CARLON (PLAINTIFFS) counsels office LOUIS | WHITE. Attoreny Opatik spoke with

Opatik, Esq.; and Proof of Service.

Figure 20 - Extracting Machine Readable Text, Excerpt from Memorandum of Points and Authorities filed in 2011

(ii) *Separating the Machine Readable Text into Sentences*

After extracting machine readable text, a sentence split annotator segregates the text into individual sentences. This is a necessary component because CLAS analyzes each sentence individually when selecting Key Phrases.

Sentences are detected using the sentence split annotator included in the Stanford CoreNLP toolkit [Manning et al., 2014]. The sentence split annotator breaks text into individual tokens (Sequences of characters grouped together as a word, or in some cases, as punctuation.) and then reconstructs the tokens into sentences.

Figure 21 and Figure 22, below, provide examples of the sentence split annotator converting machine readable text into individual sentences.

Extracted Machine Readable Text:

10
11

I. INTRODUCTION

Plaintiff White Mountains Reinsurance Companys (:White Mountains") lawsuit for legal malpractice against defendant Borton Petrini LLP (Borton Petrini) is barred as a matter of law because it is an assignment ofa legal malpractice claim in violation of well-established California law. Borton Petrini is therefore entitled tojudgment in its favor.

In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company (MSI) was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured. After making a policy limits demand which MSI did not accept, Johnson filed a complaint against M.Sls insured in June 2005 and served it on Cuison along with a statutory offer to compromise. 1_n July 2005, Borton Petrini was retained on behalf ofMSl to defend Cuison against Johnson's lawsuit.

Sentences:

10 I1 I.

INTRODUCTION Plaintiff White Mountains Reinsurance Companys (:White Mountains") lawsuit for legal malpractice against defendant Borton Petrini LLP (Borton Petrini) is barred as a matter of law because it is an assignment ofa legal malpractice claim in violation of well-established California law.

Borton Petrini is therefore entitled tojudgment in its favor.

In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company (MSI) was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured.

After making a policy limits demand which MSI did not accept, Johnson filed a complaint against M.Sls insured in June 2005 and served it on Cuison along with a statutory offer to compromise.

1_n July 2005, Borton Petrini was retained on behalf ofMSl to defend Cuison against Johnson's lawsuit.

Figure 21 - Sentence Splitting, Excerpt from Memorandum of Points and Authorities filed in 2012

Extracted Machine Readable Text:

The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine. In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA). The Defendant has breached the contract by failing to meet all of its performance obligations. Under the contract, he is liable for repayment of the entire amount paid to it by the District. The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant. The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

Sentences:

The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine.
In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA).
The Defendant has breached the contract by failing to meet all of its performance obligations.
Under the contract, he is liable for repayment of the entire amount paid to it by the District.
The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant.
The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

Figure 22 - Sentence Splitting, Excerpt from Memorandum of Points and Authorities filed in 2011

B. Key Phrase Selection

CLAS identifies and selects Key Phrases using three stages. The stages are executed in order because each stage relies on the prior stage. The first stage extracts people, organizations, and other proper nouns (collectively referred to as "Entities"). The second stage identifies the parties in the lawsuit using the Entities extracted in the first stage. The third stage limits the selection of Key Phrases to sentences that reference a party in the lawsuit and describe a past

interaction between two Entities. After Key Phrases are identified and selected, they are indexed for use by the third component, the Nature-of-Suit Classifier.

(i) *Stage 1: Extracting Persons, Organizations, and Other Proper Nouns*

Each document is searched for persons, organizations, and other miscellaneous proper nouns. Locating and extracting these Entities is a necessary stage in Key Phrase selection because this is the first step in identifying the parties in the lawsuit.

Entities are extracted using the named entity recognizer ("NER") in the Stanford CoreNLP toolkit [Finkel, Grenager, and Manning, 2005]. The NER in the Stanford CoreNLP toolkit is trained on the CoNLL-2003 English training dataset (http://www.cnts.ua.ac.be/conll2003/ner/). The CoNLL-2003 English training dataset is a collection of Reuter's newswire articles annotated with four entity types: "person", "location", "organization", and "miscellaneous." The NER is not trained on data from the legal domain because it is being used to identify a limited subset of Entities, namely "person", "organization", and "miscellaneous." The training data from the CoNLL-2003 English training dataset accomplishes this task. The NER is not used to directly identify any parties in a lawsuit, it simply identifies all persons, organizations, or miscellaneous. The identification of parties in a lawsuit is handled by CLAS and uses the Entities identified by the NER.

Figure 23, below, shows an example of extracting Entities. From top to bottom, the example contains a sentence extracted from a document, output from the NER after processing the sentence, and the Entities that are extracted.

Sentence:

In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company (MSI) was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured.

Named Entity Recognizer Output:

In July 2003, <PERSON>Flora Cuison</PERSON>, a driver insured by <ORGANIZATION>Modern Service Insurance Company</ORGANIZATION> (<ORGANIZATION>MSI</ORGANIZATION>) was involved in an automobile accident with <PERSON>Karen D. Johnson</PERSON>, during which <PERSON>Johnson</PERSON> was injured.

Extracted Entities:

<PERSON>Flora Cuison</PERSON>
<ORGANIZATION>Modern Service Insurance Company</ORGANIZATION>
<ORGANIZATION>MSI</ORGANIZATION>
<PERSON>Karen D. Johnson</PERSON>
<PERSON>Johnson</PERSON>

Figure 23 - Entity Extraction, Excerpt from Memorandum of Points and Authorities filed in 2012

Figure 23 uses tags to indicate the type of the identified Entity. An opening tag is <PERSON>, <ORGANIZATION>, or <MISCELLANEOUS> and the closing tag is </PERSON>, </ORGANIZATION>, or </MISCELLANEOUS>, respectively. The text between the opening and closing tag is the identified Entity. The result of processing the example sentence reveals three person Entities and two organization Entities. The PERSON Entities are "Flora Cuison", "Karen D. Johnson", and "Johnson". The ORGANIZATION Entities are "Modern Service Insurance Company" and "MSI".

Figure 24, below, provides another example of Entity extraction, this time analyzing six sentences from a single document. The extracted Entities are "Sacramento Metropolitan Air Quality Management District", "Paul Rosser", "Rosser Trucking", and "Sacramento Ozone Nonattainment Area".

Sentences:

The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine.
In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA).
The Defendant has breached the contract by failing to meet all of its performance obligations.
Under the contract, he is liable for repayment of the entire amount paid to it by the District.
The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant.
The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

Named Entity Recognizer Output:

The Plaintiff, <ORGANIZATION>Sacramento Metropolitan Air Quality Management District</ORGANIZATION> (District), entered into a contract with the Defendant, <PERSON>Paul Rosser</PERSON>, dba <ORGANIZATION>Rosser Trucking</ORGANIZATION> (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine.
In exchange, the Defendant agreed to operate the truck within the <ORGANIZATION>Sacramento Ozone Nonattainment Area</ORGANIZATION> (SFNA).
The Defendant has breached the contract by failing to meet all of its performance obligations.
Under the contract, he is liable for repayment of the entire amount paid to it by the District.
The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant.
The Defendant has not filed an Answer or any other responsive pleading, and the District seeks entry of a default judgment requiring the repayment of the public funds.

Extracted Entities:

<ORGANIZATION>Sacramento Metropolitan Air Quality Management District</ORGANIZATION> <PERSON>Paul Rosser</PERSON> <ORGANIZATION>Rosser Trucking</ORGANIZATION> <ORGANIZATION>Sacramento Ozone Nonattainment Area</ORGANIZATION>

Figure 24 - Entity Extraction, Excerpt from Memorandum of Points and Authorities filed in 2011

(ii) *Stage 2: Identifying the Parties in a Lawsuit*

After locating any referenced Entities within a Memorandum of Points and Authorities, CLAS proceeds to use that information to specifically identify all potential references to the parties in the suit (*e.g.* plaintiff(s) or defendant(s)). This is more complex than a simple name identification because parties in the court documents can be referred to by multiple names and/or abbreviations. Identifying the parties in a lawsuit is essential because Key Phrases must include at least one reference to a party.

d. Obstacles in Identifying the Parties in a Lawsuit

Each document has a caption page which contains metadata about the case and the document. The caption page contains various summary information about the lawsuit, such as plaintiffs, defendants, judge, case number, filing date, document title, law firms, attorneys, and additional information. The plaintiff and defendant names listed on the caption page are the formal legal names of the parties in the case. However, throughout the body of the document, the parties in the lawsuit are often referred to as some abbreviation of their formal name or simply assigned an altogether different name. For example, the author of the document may combine of the formal names of multiple defendants or multiple plaintiffs into an acronym, which is then used throughout the rest of the document to reference those parties. It is necessary for CLAS to be able to identify the parties based on these abbreviated names because these are the most common references to the parties throughout a document. For this reason, CLAS cannot use the caption page to identify plaintiffs and defendants.

Figure 25, below, shows the formal names of the parties as they appear on the caption page. It then provides an excerpt of the body of the document where the author assigned abbreviated names to the parties.

TAYLOR WHITNEY RANCH, L.P. a California limited partnership and SHEHADEH/TAYLOR LLC, a California limited liability company,

Plaintiff

v.

SIERRA PROPERTY ASSOCIATES - ROCKLIN LLC, a California limited liability company, and Does 1 through 10, inclusive,

Defendant(s).

"On June 22, 2012, the Court entered judgment in favor of SHEHADEH/TAYLOR, LLC and TAYLOR WHITNEY RANCH, L.P. ("TAYLOR WHITNEY"), and against Defendant SIERRA PROPERTY ASSOCIATES-ROCKLIN, LLC ("SIERRA PROPERTY")."

Figure 25 - Plaintiff and Defendant Renaming to Abbreviations (Top: Cover page excerpt with plaintiffs outlined in blue, defendants outlined in red. Bottom: Text excerpt from the body of the document.)

As seen in the caption page excerpt in Figure 25, the formal names of the plaintiffs are "TAYLOR WHITNEY RANCH, L.P. a California limited partnership" and "SHEHADEH/TAYLOR, LLC a California limited liability company". However, renaming occurs in the body of the document and the two plaintiffs are renamed to "TAYLOR WHITNEY". The formal names of the defendants are "SIERRA PROPERTY ASSOCIATES-ROCKLIN, LLC, a California limited liability company" and "Does 1 through 10, inclusive."[24] After renaming, the defendant is referenced as "SIERRA PROPERTY".

Another example of renaming is displayed in Figure 26, below.

[24] The defendant "Does 1 through 10, inclusive" is a placeholder used for potential unidentified defendants which may be added to the lawsuit at a later date.

ELLIS LAW GROUP, LLP,

Plaintiff,

v.

MARISON M. MULL; GEORGE W.M. MULL; NEVADA CITY SUGAR LOAF PROPERTIES, LLC and DOES 1 through 20 inclusive,

Defendants.

"By this Motion, Plaintiff Ellis Law Group LLP "ELG" hereby requests the Court compel responses from Defendant Nevada City Sugar Loaf Properties, Inc. "NCSLP" to ELG's Request for Admissions…"

Figure 26 - Plaintiff and Defendant Renaming to Abbreviations (Top: Cover page excerpt with plaintiffs outlined in blue, defendants outlined in red. Bottom: Text excerpt from the body of the document.)

In this example, the plaintiff "ELLIS LAW GROUP LLP" is renamed to the acronym "ELG" which is used for the remainder of the document. The defendants "MARISON M. MULL", "GEORGE W.M. MULL", "NEVADA CITY SUGAR LOAF PROPERTIES LLC", and "Does 1 through 20, inclusive" undergo a partial renaming. One defendant, "NEVADA CITY SUGAR LOAF PROPERTIES LLC", is renamed to the acronym "NCSLP", while the remaining defendants are not renamed. Abbreviations can also simply be a shortening of the formal name. For example, the defendant "MARISON M. MULL" is referred to as "MULL" later in the document.

e. Implementation of Identifying the Parties in a Lawsuit

CLAS identifies both the formal names and abbreviations for all Entities and uses a disambiguation process to relate analogous references to a single Entity. For example, CLAS

identifies “Microsoft”, “Microsoft Corp”, and “Microsoft Corporation” as a single Entity. Disambiguation is necessary not only because the document abbreviates formal names but also because frequently documents may use numerous names to refer to a single Entity. In Figure 26, above, defendant “MARISON M. MULL” may be referred to as “MARISON” or “Mr. MULL” or “Defendant MULL.”

Disambiguation is executed in three phases. The first phase applies pattern matching to find formal names that are renamed to abbreviations. For example, in Figure 26, the first phase detects the renaming of “ELLIS LAW GROUP LLP” to “ELG” and “NEVADA CITY SUGAR LOAF PROPERTIES LLC” to “NCSLP.” Pattern matching is implemented using regular expressions designed to match the 11 different patterns seen in the documents. Table 4, below, provides each pattern and the total number of times the pattern occurred in the documents.

Type	Formal Name Abbreviation Pattern	Total Number of Occurrences in Documents
1	NAME ("ABBREVIATION")	11954
2	NAME "ABBREVIATION"	10568
3	NAME (ABBREVIATION)	3284
4	NAME-1, NAME-2, ..., and NAME-N ("ABBREVIATION")	2018
5	NAME referred to as ABBREVIATION	985
6	NAME-1 and NAME-2 ("ABBREVIATION")	824
7	NAME (the "ABBREVIATION")	328
8	NAME (the ABBREVIATION)	296
9	NAME abbreviated to ABBREVIATION	194
10	NAME renamed to ABBREVIATION	164
11	NAME-1, NAME-2, ..., NAME-N collectively referred to as ABBREVIATION	33

Table 4 - Formal Name Abbreviation Patterns

Regular expressions are implemented to match each abbreviation pattern in Table 4. The regular expressions are used to detect any words made up of all capital letters in the location of ABBREVIATION in the abbreviation pattern. For example, the regular expression for the

abbreviation pattern type 1 matches any word that contains all capital letters and is surrounded by quotes and parenthesis. When a match is found, the preceding words are checked to see if they match any extracted Entities. If a single Entity is found, each letter in the abbreviation is matched against the first letter in each word in the Entity. If each letter in the abbreviation matches a word in the Entity, in order, then the abbreviation is identified. For example, in Figure 26 "ELLIS LAW GROUP LLP" is abbreviated to "ELG" using abbreviation pattern type 2 in Table 4. The regular expression designed to detect abbreviation pattern type 2 patterns detects the word "ELG" and the preceding Entity "ELLIS LAW GROUP LLP." The "E" in "ELG" is matched against "ELLIS", the "L" is matched against "LAW", and the "G" is matched against "GROUP." Since all letters are identified to have a match, the abbreviation is identified. However, if multiple Entities are found in the words preceding the matched abbreviation, the abbreviation is identified for all of the preceding Entities in the sentence.

The second phase applies pattern matching to search for all possible permutations of a formal name. For example, "MARISON M. MULL" is converted to "MARISON", "MARISON M.", "M. MARISON", "MARISON MULL", "MULL MARISON", "M. MARISON MULL", "MARISON MULL M.," "MULL M. MARISON," "M. MULL," "MULL M.," and "MULL." Thus, in Figure 26, when the document refers to "MULL" CLAS relates that reference to defendant "MARISON M. MULL."

The third phase measures the similarity among the Entities referenced in the document and relates those references that are above a specified similarity threshold to a single Entity. To do this, CLAS first encodes each Entity into a vector. Entities are encoded into vectors by counting the occurances of each character and storing them as a vector. Figure 27, below, provides an example of the conversion of the terms "Microsoft" and "Microsoft Corp" into vectors.

Term	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Microsoft	0	0	1	0	0	0	0	0	1	0	0	0	1	0	2	0	0	1	1	1	0	0	0	0	0	0
Microsoft Corp	0	0	2	0	0	0	0	0	1	0	0	0	1	0	3	1	0	2	1	1	0	0	0	0	0	0

In order to compare the vectors, they must contain the same number of entries, however, common zeroes can be removed to form the final vectors:

Term	c	i	m	o	p	r	s	t
Microsoft	1	1	1	2	0	1	1	1
Microsoft Corp	2	1	1	3	1	2	1	1

Or using a different syntax:

- "Microsoft": (1, 1, 1, 2, 0, 1, 1, 1)
- "Microsoft Corp": (2, 1, 1, 3, 1, 2, 1, 1)

Figure 27 - Term Vector Encoding

After Entities have been encoded, CLAS uses cosine similarity to compare the selected Entity to every other Entity found during entity extraction. If the cosine similarity value is greater than or equal to 0.5, then the two discrete references are related to a single Entity. For example, the words "Microsoft" and "Microsoft Corp" have a cosine similarity of 0.7071. The words "Microsoft Corp" and "Microsoft Corporation" have a cosine similarity of 0.5000. Thus, "Microsoft", "Microsoft Corp", and "Microsoft Corporation" are all treated as a single Entity.

The third phase also controls for unintentional author typos or misspellings and inaccurate text extractions that may occur during the optical character recognition process. (As exemplified above in Figure 19 and Figure 20, the extraction of machine readable text does not always result in the exact text of the document).

After the disambiguation process, CLAS analyzes the identified Entities to determine what Entities are parties in the lawsuit. Pattern matching is again applied to every sentence that contains an identified Entity. In order for an identified Entity to be classified as a party in the lawsuit, the

immediately preceding word must be a party title or another pre-established party-identification term (*e.*g, "plaintiff," "defendant," "complainant," or "suspect"). CLAS implements a stemming algorithm[25] to include the derivatives of these terms in this party identification process. Table 5, below, offers a list of the predecessors that CLAS considers when identifying an Entity as a party in the lawsuit. The table also provides each term's association to a role in the lawsuit.

Stemmed Word	Party in Lawsuit
Plaintiff	Plaintiff
Complainant	Plaintiff
Sue	Plaintiff
Defendant	Defendant
Suspect	Defendant
Offend	Defendant

Table 5 - Stemmed Entity Classification Words

If a match is found, the Entity is assigned to the corresponding role listed in the table. The plaintiffs and defendants cannot share Entities, thus if an Entity is in the role of a plaintiff, this prevents a later assignment to the role of a defendant and vice versa. In this sense CLAS prioritizes the first association identified.

By default, the terms "Plaintiff" and "Plaintiffs" are assigned the role of plaintiffs, and the terms "Defendant" and "Defendants" are assigned the role of defendants. If no matches are found when identifying the parties in the lawsuit, these are the only plaintiffs and defendants. If a formal name is renamed to the term "Plaintiff(s)" (or "Defendant(s)") the Entity is assigned to the role of plaintiffs (or defendants).

[25] Stemming is the process in which any affix or suffix is removed from a word to reduce it to its "stem." For example, the word "Plaintiffs" would be stemmed to "Plaintiff," the word "Offender" would be stemmed to "Offend."

Figure 28, below, provides an example of a match for plaintiff and defendant in a single sentence. The match for plaintiff is highlighted in yellow and the match for defendant is highlighted in green. The example also includes the identified plaintiffs and defendants.

The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine.

Plaintiffs	Defendants
Plaintiff	Defendant
Plaintiffs	Defendants
Sacramento Metropolitan Air Quality Management District	Paul Rosser
District	Rosser Trucking

Figure 28 - Identifying Parties in Lawsuit with Pattern Matching (Plaintiff match highlighted in yellow, defendant match highlighted in green.)

The match for the plaintiff occurs for the Entity "Sacramento Metropolitan Air Quality Management District." The word preceding "Sacramento Metropolitan Air Quality Management District" matches the word "Plaintiff" in the Stemmed Word column of Table 5. Therefore, "Sacramento Metropolitan Air Quality Management District" and all other terms used to reference the Entity ("District") are assigned the role of plaintiffs. The match for the defendant occurs for the Entity "Paul Rosser." The word preceding "Paul Rosser" matches the word "Defendant" in the Stemmed Word column of Table 5. Therefore, "Paul Rosser" is assigned to the role of defendant. "Rosser Trucking" is also assigned to the role of defendant. This is because "Rosser Trucking" is renamed to an abbreviation, "Defendant." This results in "Paul Rosser" and "Rosser Trucking" being collectively identified as the defendants. The identified parties in the lawsuit are shown in Figure 28, above.

(iii) *Stage 3: Limiting Selection to Sentences that Reference an Identified Party and Describe a Past Interaction between two Entities*

Limiting Key Phrase Selection to only those sentences that reference parties in the lawsuit is not sufficient to reliably select sentences that provide information regarding the nature of the suit. For example, the sentence "Plaintiff Shepard Johnson seeks leave to file a memorandum of points and authorities not exceeding approximately 35 pages in length," provides no information relevant to the nature of the suit and does not contribute to understanding the underlying relationship that gave rise to the lawsuit. Because CLAS classifies the nature of suit using Key Phrases, it is important for the Key Phrase Selection component to only extract sentences that contain information relevant to the nature of suit.

Therefore, the Key Phrase Selection component limits the definition of Key Phrases further to include only (1) those sentences that reference a party and (2) describe a past interaction between two Entities. The first step in the process is to determine if the sentence contains a reference to a party in the suit. The second step is analyzing the grammatical relationships in the sentence to determine whether the phrase describes any interaction (past or present or future) between two Entities. The third step is to select only those sentences that reference an interaction that occurred in the past based on the tense of the verb(s) that relates the two Entities. Thus, in the third step, the Key Phrase Selection component validates that the verb used to establish a relationship between two Entities is in the past tense.

Thus, selected Key Phrases describe a past interaction between two identified parties, two non-party Entities or between an identified party and another Entity. The Key Phrase Selection component was developed in this fashion because often disputes giving rise to a lawsuit involve interactions with other third parties that may not be named as a party in the suit but whose role in the prior dispute may still shed valuable information relating to the nature of the suit. If the Key

Phrase Selection component only selected sentences that referenced identified parties, essential information could be missed and lead to unreliable classifications. Moreover, this broader selection scope helps protect against not selecting Key Phrases that in fact do reference both a plaintiff and a defendant but that reference the parities by a name that was not identified previously as a reference to plaintiff or defendant. While the Key Sentence Selection component can select sentences that do not reference both a plaintiff and a defendant, 88% of the selected Key Phrases, or 65,358 out of 74,270, did actually reference both the plaintiff and defendant.

CLAS employs the Stanford CoreNLP toolkit to analyze the grammatical structure of each sentence that references an identified party. The Stanford CoreNLP toolkit contains a statistical parser[26] that analyzes the context of a sentence by segregating the words into grammatical categories (*e.g.*, noun, adjective, or verb) and then relates the words back to each other to gain context about those words. [Chen, and Manning, 2014; Toutanovam, Klein, Manning, and Singer, 2003]. The grammatical parsing of the sentence allows CLAS to automatically identify events that occurred between two previously identified Entities (or parties) and determine if the event took place in the past

Table 6, below, provides examples of different grammatical categories along with a tag (a shorthand representation) for the grammatical categories, and an example word. (A complete list of the Stanford CoreNLP toolkit's grammatical categories is provided in the Appendix, section Penn Treebank).

[26] The statistical parser is trained on a dataset of the entire English Wikipedia (http://download.wikimedia.org) which was transformed into suitable training data [Collobert et al., 2011]. The training dataset also consisted of 221 million words extracted from the Reuters RCV1 [Lewis et al., 2004].

Grammatical Categories	Tag	Example Word
Noun, singular	NN	Book
Noun, plural	NNS	Books
Verb, base form	VB	Watch
Verb, past tense	VBD	Watched
Verb, past participal	VBN	Been
Adjective	JJ	Carefully

Table 6 - Example Subset of Parts of Speech

Table 7, below, provides examples of the grammatical parsing of sentences.

Grammatical Relation	Sentence	Shorthand
Direct Object Relationship	She gave me a raise.	dobj(gave, raise)
Indirect Object Relationship	She gave me a raise.	iobj(gave, me)
Negation Modifier	Bill does not drive.	neg(drive, not)
Nominal Subject	Clinton defeated Dole.	nsubj(defeated, Dole)
Noun Compound Modifier	Oil price futures.	nn(futures, oil)

Table 7 - Example Subset of Grammatical Relations

The first column in Table 7 identifies the grammatical relation between two words, the second column provides the full sentence, and the third column provides a shorthand representation for the grammatical relation as well as the two words that are related. (A complete list of the Stanford CoreNLP toolkit's 50 different grammatical relations are provided in the Appendix, section Stanford Grammatical Relations).

Figure 29, below, diagrams the results of the grammatical parser.

Figure 29 – Diagrams of Grammatical Relations (Adapted from Stanford CoreNLP. Retrieved January 23, 2015, from http://nlp.stanford.edu:8080/corenlp/. Copyright Stanford University 2011.)

These diagrams identify the (1) different grammatical category of each word and (2) the relationships between the words. The grammatical category is displayed directly above each word. The contextual relationship(s) between the words is displayed as an arrow from one word to another word. The arrows between the words diagram the directional relationship of how the words in the sentence relate to each other.

For example, the diagram for the sentence "She gave me a raise," in Figure 29 relates the following words:

- The terms "gave" and "She" are related as a nominal subjects or "nsubj." In other words the term "She" is the subject of the clause and "gave" is the action.
- The terms "me" and "gave" are related as an indirect object or "iobj." In other words the term "me" is the indirect object or the receiver of the action "gave."
- The terms "gave" and "raise" are related as a direct object or "dobj." In other words the term "raise" is the object on which the action, "gave," is directed.
- The terms "raise" and "a" are related as a determiner or "det." In other words "a" is the determiner or a description of "raise".

Understanding the contextual relations between words in a sentence is an integral part of Key Phrase Selection component. The process of extracting Key Phrases relies on the grammatical parsing of sentences to gain understanding about the information contained in the sentences. Figure 30, below, provides a diagram of a more complex sentence extracted from a Memorandum of Points and Authorities processed by Key Phrase Selection component.

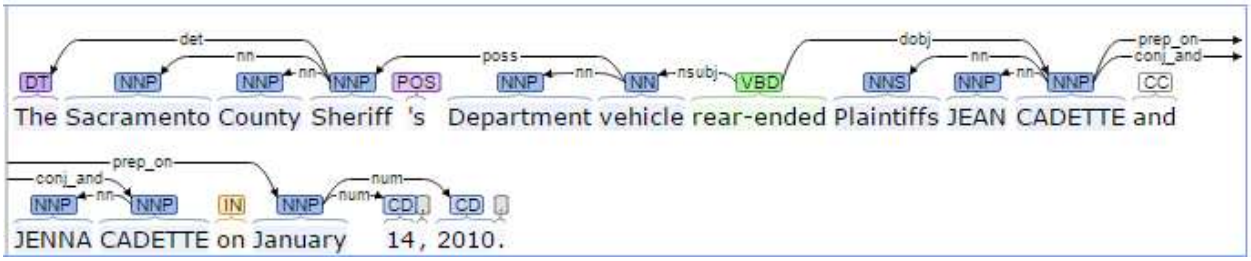


Figure 30 - Diagram of Grammatical Relations (Adapted from Stanford CoreNLP. Retrieved February 23, 2015, from http://nlp.stanford.edu:8080/corenlp/. Copyright Stanford University 2011.)

The sentence diagramed in Figure 30 references previously identified parties to the lawsuit, which is why the Key Phrase Selection component selects it for grammatical parsing. The grammatical parsing of the sentence identifies that the sentence describes an interaction between two Entities. In this case, a plaintiff ("JEAN CADETTE") and the defendant ("Sacramento County Sheriff's Department"). The Key Phrase Selection component identifies this interaction because both Entities (the plaintiff and defendant) relate to the verb "rear-ended." As the diagram above illustrates,

- The plaintiff "JEAN CADETTE" and "rear-ended" are related because "JEAN CADETTE" is the direct object of the action "rear-ended" (referred to as "dobj" in the diagram).
- The defendant "Sacramento County Sheriff's Department" and "rear-ended" are related because "vehicle" is the subject of the action "rear-ended" (referred to as "nsubj" in the diagram) and "vehicle" is a possession of the "Sacramento County Sheriff's Department" (referred to as "poss" in the diagram).

While grammatical parsing occurs on the entire sentence, the Key Phrase Selection component's determination of whether to select it as a possible Key Phrase only truly depends on the contextual relationships described above, relating both Entities (the plaintiff and defendant) to the verb "rear-ended," to establish that this sentence does describe an interaction between two Entities. The remaining grammatical relations are irrelevant to the Key Phrase selection process.

The final step limits the selection of Key Phrases to only those sentences that describe a past interaction between two Entities. The Key Phrase Selection component filters out present and future tense sentences because too often those sentences describe information that does not relate to the nature of the suit. This is the case because lawsuits can only arise as a result of events or interactions that have already occurred. The two sentences below, illustrate a present tense and

future tense sentence that are not Key Phases and do not contribute information about the nature of the suit[27]:

> **Present Tense**
> "Plaintiff Shepard Johnson seeks leave to file a memorandum of points and authorities not exceeding approximately 35 pages in length."
>
> **Future Tense**
> "Demurrer for uncertainty will be sustained only where the complaint is so bad that the defendant cannot reasonably determine what issues must be admitted or denied, or what counts or claims are directed against him or her."

While these sentences both reference a party in the lawsuit neither provide any information relevant to the nature of the suit; neither contributed to understanding the underlying relationship between the plaintiffs and defendants in the lawsuit.

The Key Phrase Selection component evaluates the tense of the verb that describes an interaction. If the verb is in past tense, then the sentence is identified as a Key Phrase. This can be a past interaction between two identified parties, two non-party Entities or between an identified party and another Entity.

The statistical parser from the Stanford CoreNLP toolkit provides grammatical categories that identify past tense verbs as "past tense" or "past participle." "Past tense" verbs are labeled by the shorthand value of "VBD" and "past participle" verbs are "VBN" (Table 6). In Figure 30, the two Entities are related through the past tense verb "rear-ended," thus the sentence passes the final step and is selected as a Key Phrase.

---

[27] The present tense sentence is from a Memorandum of Points and Authorities that was part of the 2013 lawsuit Shepard Johnson vs. David Miner, et al., and was filed in support of the plaintiff's request that the court allow it to exceed the page limitation in another document that the plaintiff intended to file in the future. The Sacramento Superior Court's case details classified this lawsuit as PI/PD/WD – Other.

The future tense sentence is from a Memorandum of Points and Authorities that was part of the 2011 lawsuit Mr Kenneth Smith vs. UCD Davis Medical Center Hospital, and was filed in support of the defendant's special demurrer to the plaintiff's complaint. The Sacramento Superior Court's case details classified this lawsuit as Professional Negligence.

While the sentence in Figure 30 is a Key Phrase and describes a clear relationship between two Entities, it is also important to see the challenges in identifying the relationships between Entities.

Figure 31, below, provides another sentence selected as a Key Phrase.

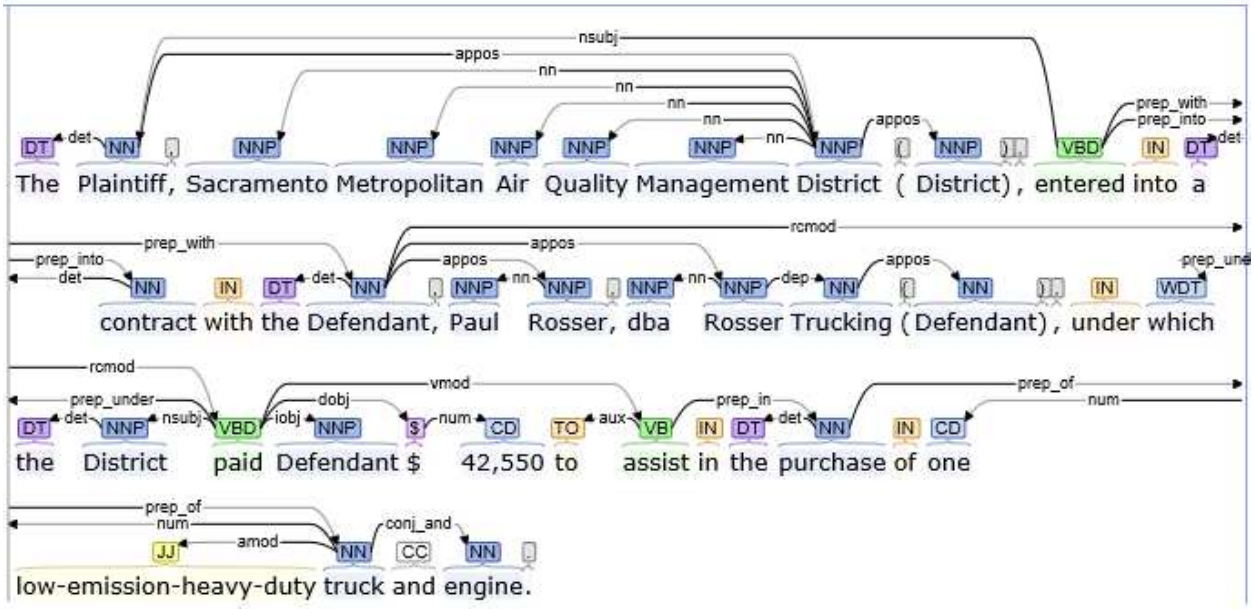


Figure 31 - Diagram of Grammatical Relations (Adapted from Stanford CoreNLP. Retrieved January 23, 2015, from http://nlp.stanford.edu:8080/corenlp/. Copyright Stanford University 2011.)

The plaintiff in the case is Sacramento Metropolitan Air Quality Management District (referred to as "Plaintiff" and "District"), and the defendants are Paul Rosser and Rosser Trucking (referred to as "Defendant"). The sentence in Figure 31 meets all three Key Phrase requisites:

1. It contains a reference to a party.
2. It describes an interaction between the referenced party ("Plaintiff") and another Entity ("Defendant").
3. It describes a past interaction between two Entities.

The verb "entered" establishes the relationship between the two Entities; the plaintiff entered into a contract with the defendant. The verb "entered" is in the past tense ("VBD"), and therefore describes the past.

However, this thesis calls attention to this specific Key Phrase not because of its selection as a Key Phrase, but to illustrate a potential limitation of CLAS's Key Phrase Selection component, which at the time of this thesis treats all selected Key Phrases equally and is not capable of ranking the potential value or likelihood that the Key Phrase will provide reliable information about the nature of the suit. This limitation is caused by the fact that the Key Phrase Selection component only considers a single past tense verb that connects two referenced entities. No other words or relations are considered by the Key Phrase Selection component. (To be clear, CLAS's subsequent component, the Nature-of-Suit Classifier, analyzes and interprets every word in selected Key Phrases. Thus, the final classification of the Key Phrase will rely on all words contained therein). The fact that the Key Phrase Selection component only considers past tense verbs and their related Entities causes the Key Phrase Selection component to miss other, potentially valuable information that could provide insight into whether the sentence should be considered a high quality or low quality (or inconclusive quality) Key Phrase.

In the example provided in Figure 31 the Key Phrase Selection component correctly selects the sentence as a Key Phrase (and the subsequent Nature-of-Suit Classifier component correctly classified it as "Breach of Contract/Warranty"). However, during the Key Phrase selection process the Key Phrase Selection component did not consider other valuable information contained in the sentence. For example, the verb "entered" is followed by the prepositional phrase "into a contract." These types of descriptors, whether adverbs or prepositional phrases or something else, may provide valuable information about the quality of information contained in the Key Phrase. Whether an Entity entered "into a tunnel" or "into a contract" or just "entered the wrong way" causes similar sentences to have completely different meanings. Some sentences, on the other hand, may contain verbs with no additional descriptors at all. Consideration of the verb by itself

provides no insight into the quality of the information contained in the Key Phrase. Allowing for the evaluation of the information contained in the Key Phrases during the selection component could allow for better results during the classification component. As discussed in the FUTURE WORK section, extensions to CLAS will examine the quality of the information contained in a Key Phrase prior to classification.

While the sentences in Figure 30 and Figure 31 are selected as Key Phrases, it is important to see an example of a sentence that is not a Key Phrase and understand why that sentence is not selected.

Figure 32, below, provides a sentence that is not a Key Phrase. This is an important example because the sentence contains a direct request to the court that does not describe past events between two Entities and therefore does not contribute to determining the nature of suit. This example also illustrates that a sentence can have a reference to a party, references to at least two Entities, and a past tense verb, and still not meet the threshold to be identified as a key phrase by the Key Phrase Selection component. Simply having a past tense verb is not enough, the past tense verb must establish a relationship between two Entities. In Figure 32, below, the plaintiffs in the case are White Mountains Reinsurance Company of America ("White Mountains") and Modern Service Insurance Company ("MSI"), and the defendant is Borton Petrini LLP ("Borton Petrini"). Other Entities in the document are Flora Cuison and Karen D. Johnson.

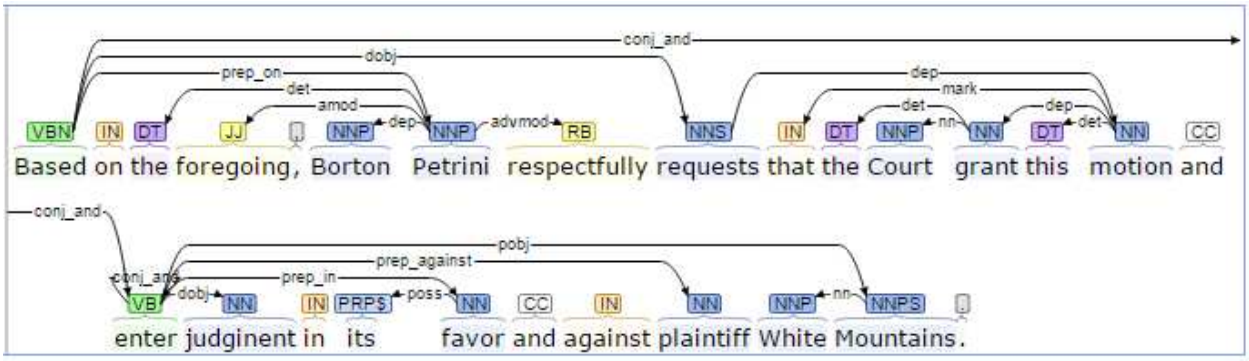


Figure 32 - Key Phrase Selection (Adapted from Stanford CoreNLP. Retrieved January 23, 2015, from http://nlp.stanford.edu:8080/corenlp/. Copyright Stanford University 2011.)

The sentence in Figure 32 only meets two of the three requisites:

1. It contains a reference to a defendant, Borton Petrini.
2. It describes an interaction between the referenced party and another Entity, White Mountains.
3. But, it does not describe a past interaction between two Entities.

The Key Phrase Selection component identifies that the relationship between the defendant, Borton Petrini, and the plaintiff, White Mountains, does not contain a verb in the past tense or past participle. Importantly, the sentence does contain a past participle verb, "based", however, this verb does not function to relate two Entities and does not function to qualify the sentence as a Key Phrase.

Figure 32 also highlights different challenges faced by CLAS. For example, the quality of the machine readable text can impair CLAS's ability to analyze Key Phrases. Had this example been selected as a Key Phrase, the word "judgment" was misspelled as "judginent" and could have affected the classification. While this is not a particular issue during Key Phrase selection of this sentence, it is possible that misspellings exist for verbs that would have otherwise allowed the sentence to be selected as a Key Phrase and instead caused the Key Phrase Selection component to incorrectly not select the sentence.Nature-of-Suit Classifier

(i) *Training the Nature-of-Suit Classifier*

The Nature-of-Suit Classifier must be trained in order to build relationships between the text contained in Key Phrases and the nature-of-suit categories. The 10,000 Memorandum of Points and Authorities collected from the Sacramento Superior Court are randomly divided into 8,500 training documents and 1,500 evaluation documents.

To conduct training, the 8,500 training documents are processed by the first two components in CLAS, (1) document preprocessing, and (2) Key Phrase Selection. The document preprocessing component extracts machine readable text. The Key Phrase Selection component produces Key Phrases from the machine readable text. A total of 62,447 Key Phrases are selected from the training documents, an average of 7.3467 Key Phrases per document. The Nature-of-Suit Classifier is trained using the Key Phrases produced for each Memorandum of Points and Authorities together with the nature of suit assigned by the Sacramento Superior Court. Table 8, below, provides examples of Key Phrases used for training. The nature of suit column contains the assigned nature of suit and the Key Phrase column contains Key Phrases that are selected by the Key Phrase Selection component.

Nature of Suit	**Key Phrase**
Breach of Contract/Warranty	This matter involved an action brought by ELG against NCSLP for attorneys fees and costs incurred by ELG during its representation of NCSLP, which remain unpaid.
	In late 2006, George Mull, NCSLPs manager executed a legal services agreement, retaining ELG to represent NCSLP.
	On or about September 12, 2011, NCSLP led and served an answer, asserting 38 affirmative defenses, and a cross - complaint alleging a single cause of action against ELG for breach of fiduciary duty.
Other Real Property	Plaintiff had received the offer on August 3, 2011, the offer was accepted on or about September 6, 2011 and then sent to defendant Bank of America for its approval.
	After uploading the offer, defendant Bank of America, who had initiated foreclosure proceedings on the property, agreed to the Short Sale.
	The copy of the most recent Short Sale perusals which was submitted and approved by the Bank of America, and are attached as Exhibit N to the First Amended Complaint.

	Without justification, Bank of America refused to extend the close of escrow and the date of trustees sale.
	In the meantime the defendant Bank of America also served the Plaintiff with Notice of Trustees Sale, and , recorded a Notice of Default and the said Notice of Trustees Sale with the County of Sacramento on or about September 26 , 2011 , while the short sale was in process.
Medical Malpractice	Plaintiff has failed to present clear and convincing evidence that Daemeron acted with the requisite level of recklessness required to show elder abuse.
	In a medical malpractice action, both plaintiff and defendant must present expert opinion testimony to prove or disprove, to a reasonable degree of medical certainty, that the act or omission of the alleged negligent actor breached the standard of care.

Table 8 – Example Training Key Phrases

During training, the Key Phrase Selection component selects Key Phrases from Memorandum of Points and Authorities that have been assigned as training documents. Each Key Phrase is converted into vectors and indexed. The vectors are created by splitting the Key Phrase into individual words and removing any stop words (Words such as “and”, “a”, “the”, and others. A complete list of Stop Words is provided in the Appendix, section Stop Words). One vector is created for each word. Each vector represents the number of times the word appears in the Key Phrase. Figure 33, below, provides an example of converting a Key Phrase into vectors.

Nature of Suit
Medical Malpractice

Key Phrase
"In a medical malpractice action, both plaintiff and defendant must present expert opinion testimony to prove or disprove, to a reasonable degree of medical certainty, that the act or omission of the alleged negligent actor breached the standard of care."

Vectors from Key Phrase			
Term	**Term Count**	**Term**	**Term Count**
medical	2	reasonable	1
malpractice	1	degree	1
action	1	certainty	1
plaintiff	1	act	1
defendant	1	omission	1
present	1	alleged	1
expert	1	negligent	1
opinion	1	actor	1
testimony	1	breached	1
prove	1	standard	1
disprove	1	care	1

Figure 33 - Key Phrase Converted into Vectors

All of the vectors from all of the Key Phrases in each training document are aggregated together and categorized by the nature-of-suit category that was assigned by the Sacramento Superior Court to the document's respective lawsuit. Figure 34, below, provides an excerpt of the aggregated training document vectors organized by nature-of-suit category. For example, for nature of suit PI/PD/WD – Auto (Table 1 on page 45 provides a listing of all nature-of-suit categories with descriptions), the term "vehicle" occurred a total of 7,113 times in all of the Key Phrases in all of the training documents. In Figure 34, the term count directly corresponds to the strength of the association to the nature-of-suit category; a higher term count means the term is more strongly associated to the nature-of-suit category.

Aggregated Vectors Nature of Suit = PI/PD/WD - Auto	
Term	**Term Count**
vehicle	7113
accident	3440
damages	1094
injured	748
collision	655
driver	590
automobile	561
motor	383
street	219
insured	203

Aggregated Vectors Nature of Suit = Medical Malpractice	
Term	**Term Count**
medical	9921
hospital	5204
emergency	2136
healthcare	918
physician	809
malpractice	766
surgery	430
nurses	387
treatment	209
rehabilitation	193

Figure 34 - Excerpt of Aggregated Vectors

It is interesting to note that many of the terms that distinguish nature of suit are non-legal terms. For example, in PI/PD/WD – Auto the word "vehicle", "automobile", "driver", and "motor" are non-legal terms that associate with this nature of suit. Many of the terms that relate to Medical Malpractice are also non-legal terms. In this vein, when categorizing nature of suit, CLAS relies on non-legal terms because non-legal terms provide more information about the underlying nature of the dispute. This makes sense because all lawsuits are likely to contain many legal terms, such as "liability", "breach", "negligence", and "damages". Thus, such common legal terms will not provide insight to the underlying dispute that gave rise to the lawsuit.

It is also important to note that CLAS uses a naïve Bayes classifier to independently calculate the probability that a word is classified in a nature-of-suit category. Because the naïve Bayes classifier maintains the independent correlation between the terms and the nature of the suit, CLAS can automatically build a lexicon during training. The automated building of a lexicon provides allows for a more comprehensive collection of potentially relevant terms than a manually constructed counterpart.

(ii) *Classifying Memoranda of Points and Authorities*

After training is complete, CLAS is equipped to categorize documents. CLAS contains a set of aggregated vectors for each nature-of-suit category. To classify a new Memorandum of Points and Authorities, CLAS selects Key Phrases and classifies the nature-of-suit category for each Key Phrase. CLAS then considers all of the Key Phrase classifications to make a final determination.

In order to classify each Key Phrase, the Key Phrase is converted into vectors (As seen in Figure 33). CLAS's naïve Bayes classifier is used to calculate the probability that the vectors are assigned to a particular nature-of-suit category. These calculations are repeated for each nature-of-suit category. In order to calculate these probabilities, the naïve Bayes classifier first calculates the probability that each individual vector is in a nature-of-suit category. Each vector's probability contributes independently to the calculation of how likely it is that the Key Phrase is a particular nature-of-suit category. These independent probabilities are then multiplied together to reach the overall probability that the Key Phrase is in the nature-of-suit category. The Key Phrase is classified as the category with the highest probability.

Figure 35, below, provides an example of the Key Phrase classification process.

Key Phrase

"In July 2003, Flora Cuison, a driver insured by Modern Service Insurance Company (MSI) was involved in an automobile accident with Karen D. Johnson, during which Johnson was injured."

Vectors from Key Phrase			
Term	Term Count	Term	Term Count
july	1	msi	1
2003	1	involved	1
flora	1	automobile	1
cuison	1	accident	1
driver	1	karen	1
insured	1	johnson	1
modern	1	during	1
service	1	johnson	1
company	1	injured	1

Calculating the Probability the Key Phrase is PI/PD/WD – Auto

kp = Key Phrase

w_i = ith word in Key Phrase

ns = PI/PD/WD – Auto

$$P(kp \mid ns)$$
$$= P(w_1 \mid ns) * P(w_2 \mid ns) * \dots * P(w_n \mid ns)$$
$$= P("july" \mid ns) * P("2003" \mid ns) * \dots * P("injured" \mid ns)$$

Calculating the Probability a Single Word is PI/PD/WD – Auto

ε = Laplace estimator

$$P(w_i \mid ns) = \frac{P(ns \mid w_i) * P(w_i) + \varepsilon}{P(ns)}$$

Figure 35 – Example of the Key Phrase Classification Process

The Key Phrase in Figure 35 had a 3.93% probability that is was filed in a PI/PD/WD – Auto lawsuit and a 3.51% probability that it was filed in a PI/PD/WD – Other lawsuit. Figure 36, below, provides the top five probabilities for the Key Phrase in Figure 35.

Nature of Suit	Probability
PI/PD/WD - Auto	3.93%
PI/PD/WD - Other	3.51%
Fraud	0.99%
Medical Malpractice	0.95%
Professional Negligence	0.68%

Figure 36 - Key Phrase Classification Results

In Figure 36, the first two probabilities are above 3% while the remaining are below 1%. Probability values below 1% indicate a small amount of confidence in the nature-of-suit category. The highest probability is associated with the PI/PD/WD – Auto category because the Key Phrase contains terms that are strongly associated with this nature-of-suit category, as shown in the excerpt of the aggregated vectors in Figure 34, above. The second highest probability is associated with PI/PD/WD – Other due to the terms "injured" and "accident". While "accident" is most strongly associated with PI/PD/WD – Auto, the second highest association is with PI/PD/WD – Other. Because PI/PD/WD – Auto had the highest probability, this Key Phrase is classified as PI/PD/WD – Auto.

After classifying each Key Phrase, an overall classification of the Memorandum of Points and Authorities must be determined. CLAS applies three different methods to make this overall classification; (1) maximum occurrence, (2) maximum probability, and (3) weighted probability.

The maximum occurrence method counts the number of nature-of-suit categories assigned to the Key Phrases in that document. The nature-of-suit category with the highest number is used as the classification of the document.

The maximum probability method uses the probability values returned from the classification of each Key Phrase and selects the classification with the highest probability value.

The weighted probability method also uses the probability values returned from the classification of each Key Phrase. These probability values are summed for each nature-of-suit category and the highest sum is selected.

Figure 37, below, provides an example of the classification methods using six key phrases that have already been classified.

Key Phrase Number	Classification	Probability
1	Breach of Contract/Warranty	3.00%
2	Fraud	8.00%
3	Construction Defect	3.00%
4	Breach of Contract/Warranty	0.10%
5	Breach of Contract/Warranty	0.30%
6	Construction Defect	7.00%

Classification Method	Document Classification
Maximum Occurrence	Breach of Contract/Warranty
Maximum Probability	Fraud
Weighted Probability	Construction Defect

Figure 37 - Example Classification Method Results

In Figure 37, six example Key Phrases from a Memorandum of Points and Authorities are classified. The classifications and associated probabilities are displayed for each Key Phrase. The Key Phrase classifications are used to determine the classification for the Memorandum of Points and Authorities from which they derived. In this example, each classification method results in a different classification for the Memorandum of Points and Authorities. The maximum occurrence method selects Breach of Contract/Warranty because it occurs more frequently than any other classification with a count of three. The maximum probability method selects Fraud because it has the highest probability of all the classifications with a value of 8.00%. The weighted probability method selects Construction Defect because the sum of the probabilities is the highest

(Construction Defect sum is 10.00%, Fraud sum is 8.00%, Breach of Contract/Warranty sum is 3.40%).

Figure 38, below, provides an example of the classification methods using three Key Phrases extracted from a Memorandum of Points and Authorities. The correct classification for the Memorandum of Points and Authorities is Breach of Contract/Warranty.

Key Phrases:

1. The Plaintiff, Sacramento Metropolitan Air Quality Management District (District), entered into a contract with the Defendant, Paul Rosser, dba Rosser Trucking (Defendant), under which the District paid Defendant $42,550 to assist in the purchase of one low-emission-heavy-duty truck and engine.

2. In exchange, the Defendant agreed to operate the truck within the Sacramento Ozone Nonattainment Area (SFNA).

3. The Defendant refused to repay the funds, and the District brought this breach of contract complaint, which was personally served on the Defendant.

Key Phrase Number	Classification	Probability
1	Civil Rights	2.35%
2	Breach of Contract/Warranty	3.45%
3	Breach of Contract/Warranty	5.17%

Classification Method	Document Classification
Maximum Occurrence	Breach of Contract/Warranty
Maximum Probability	Breach of Contract/Warranty
Weighted Probability	Breach of Contract/Warranty

Figure 38 - Example Document Classification

Although the first Key Phrase in Figure 38 is classified as Civil Rights, the remaining Key Phrases are correctly classified as Breach of Contract/Warranty with higher probabilities than the misclassified Key Phrase. These Key Phrases result in all three classification methods correctly classifying the document as Breach of Contract/Warranty.

VII. OUTCOMES

For the purpose of this thesis, the remaining 1,500 Memorandum of Points and Authorities (From the original 10,000 Memorandum of Points and Authorities downloaded from the Sacramento Superior Court) are processed through CLAS and categorized by nature of suit. A total of 11,823 Key Phrases are extracted from the 1,500 evaluation documents, an average of 7.8820 per document.

The overall nature-of-suit classification accuracy for two out of the three methods is higher than 70%, with the third method scoring above 69%. These results indicate a degree of success in CLAS's classification of nature of suit. Table 9, below, displays the overall accuracy of each classification method. The overall accuracy is calculated by counting the number of correctly classified evaluation documents and dividing it by the total number of evaluation documents.

Classification Method	Accuracy
Maximum Occurrence	69.13%
Maximum Probability	72.87%
Weighted Probability	76.93%

Table 9 - Nature of Suit Classification Accuracies

Table 10, below, provides the precision[28] and recall[29] scores for each classification method in each nature-of-suit category. Precision and recall are calculated against a "Golden Standard." The Golden Standard contains the nature-of-suit category assigned to each document by the Sacramento Superior Court. Therefore, the Golden Standard contains the correct nature-of-suit

[28] Precision is calculated for each nature-of-suit category using the following formula:

$$\frac{\text{the number of correctly classified documents}}{\text{the number of incorrectly classified documents} + \text{the number of correctly classified documents}}$$

[29] Recall is calculated for each nature-of-suit category using the following formula:

$$\frac{\text{the number of correctly classified documents}}{\text{the total number of documents}}$$

classification for every document. The symbol "*" indicates that CLAS classified no documents with that nature-of-suit category and no documents with this nature-of-suit category are contained in the evaluation documents.

	Classification Method					
	Maximum Occurrence		Maximum Probability		Weighted Probability	
Nature of Suit Category	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Antitrust/Trade Regulation	*	*	*	*	*	*
Asbestos	100.00%	60.00%	100.00%	60.00%	100.00%	60.00%
Asset forfeiture	*	*	*	*	*	*
Breach of Contract/Warranty	41.40%	92.22%	63.89%	63.89%	68.90%	80.00%
Business Tort	72.22%	65.00%	85.71%	90.00%	100.00%	90.00%
Civil Rights	*	*	*	*	*	*
Construction Defect	100.00%	58.00%	86.00%	86.00%	100.00%	86.00%
Contract – Other	100.00%	100.00%	50.00%	100.00%	100.00%	100.00%
Defamation	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Eminent domain/Inverse condemnation	100.00%	100.00%	48.28%	63.64%	100.00%	100.00%
Enforcement	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fraud	69.62%	62.50%	76.99%	98.86%	82.86%	98.86%
Harassment	*	*	*	*	*	*
Insurance Coverage	100.00%	61.90%	0.00%	0.00%	0.00%	0.00%
Insurance Coverage Claims	*	*	*	*	*	*
Intellectual Property	*	*	*	*	*	*
Judicial Review - Other	*	*	*	*	*	*
Mass Tort	*	*	*	*	*	*
Medical Malpractice	90.28%	100.00%	90.28%	100.00%	90.28%	100.00%
Misc Complaints - Other	83.72%	65.45%	74.14%	78.18%	74.14%	78.18%
Non-PI/PD/WD tort - Other	100.00%	80.56%	100.00%	80.56%	100.00%	100.00%
Other Collections	100.00%	24.14%	100.00%	48.28%	100.00%	48.28%
Other employment	83.72%	41.38%	76.92%	57.47%	74.14%	49.43%
Other Real Property	90.28%	52.85%	64.36%	52.85%	82.76%	58.54%
Petition re: Arbitration Award	100.00%	25.00%	0.00%	0.00%	0.00%	0.00%
Petitions - Other	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PI/PD/WD - Auto	93.07%	68.61%	70.80%	70.80%	76.64%	76.64%
PI/PD/WD - Other	57.92%	80.21%	76.33%	100.00%	68.18%	96.26%
PI/Property Damage/Wrongful Death	*	*	*	*	*	*
Product Liability	100.00%	52.63%	80.00%	42.11%	100.00%	52.63%

Professional Negligence	69.15%	82.28%	52.55%	91.14%	52.55%	91.14%
Rule 3.740 Collections	90.28%	84.42%	91.14%	93.51%	91.14%	93.51%
Toxic Tort/Environmental	*	*	*	*	*	*
Uninsured Motorist	*	*	*	*	*	*
Unlawful Detainer - Commercial	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Writ of Mandate	100.00%	100.00%	63.64%	100.00%	63.64%	100.00%
Wrongful Eviction	100.00%	100.00%	50.00%	100.00%	100.00%	100.00%
Wrongful Termination	75.86%	44.00%	67.44%	58.00%	100.00%	58.00%
N/A	80.56%	58.00%	80.56%	58.00%	74.14%	86.00%

Table 10 - Nature of Suit Classification Results (indicates there are no documents with this nature-of-suit category in the evaluation documents)*

Since the CLAS classification process relies on individual words in Key Phrases for training and evaluation, nature-of-suit categories that are correlated to unique terms are more likely to be correctly classified. It is also interesting to note that the unique terms are most commonly non-legal terms. For example, all three methods score highly in precision and recall when classifying Medical Malpractice, Non-PI/PD/WD tort – Other, Rule 3.740 Collections, and Unlawful Detainer – Commercial. This is because these nature-of-suit categories contained unique terminology which was rarely found elsewhere. A subset of the unique terminology used in each of these nature-of-suit categories can be seen in Table 11, below.

Medical Malpractice	**Non-PI/PD/WD tort – Other**	**Rule 3.740 Collections**	**Unlawful Detainer – Commercial**
malpractice	distress	credit	detainer
hospital	emotional	financial	unlawful
emergency	reckless	payment	unsigned
healthcare	infliction	consolidated	barred

Table 11 - Unique Terminology for Subset of Nature of Suits

One interesting case of unique terminology is in Rule 3.740 Collections. Rule 3.740 Collections and Other Collections both involve collections. Intuition would dictate that these two nature-of-suit categories would share terminology. However, Other Collections is specifically for amounts over $25,000. Terminology in Other Collections documents is concerned with collecting

or seizing real estate properties, vehicles, and other high valued assets. These terms include "sanctions", "seize", "instruments", and "assets". This results in Other Collections sharing terminology with other nature-of-suit categories such as Other Real Property ("assets" and "seize") and Antitrust/Trade Regulation ("sanctions" and "instruments").

While unique terminology leads to high scores in precision and recall, the lack of unique terminology leads to incorrect classifications. For example, all three methods failed to correctly classify any Memorandum of Points and Authorities as Enforcement or Defamation. These nature-of-suit categories share terminology with other nature-of-suit categories because they are used to describe general concepts. For example, Enforcement is a legal action available to provide enforcement of a previous action that was not followed. This is an abstract type of lawsuit when compared to PI/PD/WD – Auto which directly relates to an automobile accident. Enforcement lawsuits could be caused by any previous action that was not followed and thus do not contain consistent terminology. Similar to Enforcement, Defamation is also an abstract type of lawsuit. Defamation is communication that causes damage to a person's reputation and may lead to other damages. Lawsuits with a nature-of-suit category of Defamation could involve any form of communication and any number of damages, thus they do not contain unique terminology. For example, a plaintiff may bring a Defamation lawsuit alleging that the defendant inaccurately told others that the plaintiff was a bad driver who sped excessively and has caused multiple car accidents. This example lawsuit would most likely be classified as PI/PD/WD – Auto.

Shared terminology also leads to common misclassifications into particular nature-of-suit categories. For example, all three methods commonly misclassify documents as Breach of Contract/Warranty, PI/PD/WD – Other, and Professional Negligence. Each of these nature-of-suit categories contains low precision scores and high recall scores (Table 10), indicating that many

documents are incorrectly classified into these categories. Misclassifications as Breach of Contract/Warranty occur because other nature-of-suit categories commonly include contracts and therefore have Key Phrases which discuss entering contracts, pre-existing contracts, breaching contracts, or exiting contracts. For example, construction contracts are found in Construction Defect lawsuits, insurance contracts are found in Insurance Coverage lawsuits, and employment contracts are found in Wrongful Termination lawsuits. Although these are marked as misclassifications in this evaluation, in many cases they are the correct classification. This is because these documents contain multiple nature-of-suit categories. For example, a Construction Defect document could involve a contract which is breached due to defects in the construction of a building. The attorney who filed the initial document in the case may have decided it was a Construction Defect case, however, it still clearly involves a breach of contract as well.

Misclassifications as PI/PD/WD – Other and Professional Negligence are also due to shared terminology. PI/PD/WD – Other is commonly misclassified in place of Medical Malpractice or PI/PD/WD – Auto because both of these nature-of-suit categories involve personal injury, property damage, or wrongful death. Professional Negligence is often misclassified as Fraud due to shared terminology. Professional negligence cases commonly include non-disclosures of information and failures to meet a standard of accepted expectations. These concepts are also seen in Fraud. For example, a lawsuit in which a lawyer delayed a trial causing a statute of limitations to expire may result in professional negligence. The lawyer may have knowingly misled their client and withheld information involving their reasons for delaying the trial. Although this case may be classified as Fraud, it also involves professional negligence.

While the three classification methods respond similarly to unique terminology and shared terminology, each individual classification method has different strengths and weaknesses. The

maximum occurrence method is the only method which correctly classifies Insurance Coverage. The other methods fail to do so because the Key Phrases which are identified as Insurance Coverage contain consistently low probabilities. Of the three methods, the maximum occurrence method misclassifies the largest number of Memorandum of Points and Authorities as Breach of Contract/Warranty. The improvement in the other two methods reveals that although many Key Phrases are classified as Breach of Contract/Warranty, these classifications do not have consistently high probabilities.

The maximum probability method shows improvement over the maximum occurrence method in many categories, such as Business Tort, Construction Defect, Fraud, and PI/PD/WD – Other. These improvements are attributed to individual Key Phrases that are classified with a high probability. Probabilities above 20% represent strong associations to the nature of suit.

Table 12, below, provides examples of Key Phrases that had strong associations to these nature-of-suit categories.

Key Phrase	Classification	Probability
In addition, MGG now alleges that ICL contracted to supply the State with the AquaGel K product that ICL knew did not meet the States specifications “as a ruse” to induce the State to purchase other non-conforming fire suppressant from ICL .	Business Tort	29.23%
27 Cross-Defendant A & M FENCE contracted with MOURIER to serve as the fencing 28 subcontractor for the homes in the Northpointe developments that are the Subject of this MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT OF CROSS-DEFENDANT A 8 .	Construction Defect	27.12%
Plaintiff was a shareholder and Defendant Laurence had a fiduciary duty to the Plaintiff-shareholder to disclose such a transfer as an officer of Defendant Luppen , and Defendant Luppen had a duty to effect such notication by and through the actions of the 1995 Board -RRB- -LRB- L.O. Depo .	Fraud	31.21%

25 26 To properly plead and prove a cause of action for Intentional Iniction of Severe Emotional 27 Distress , plaintiff must assert facts which demonstrate extreme and outrageous conduct by the LEWIS 28 defendant with the intention of causing , or reckless disregard of the probability of causing emotional distress .	Non-PI/PD/WD tort – Other	26.11%

Table 12 - Maximum Probability Method High Probability Key Phrases (Probabilities above 20% represent strong associations to the nature of suit.)

In Table 12, each Key Phrase is classified correctly and contains a strong association to the nature-of-suit category. In the first Key Phrase, the terms "ruse" and "non-conforming" contain strong associations to the nature-of-suit category Business Tort. In the second Key Phrase, the terms "contractor", "subcontractor", and "homes" are all associated to the nature-of-suit category Construction Defect. In the third Key Phrase, the terms "shareholder", "fiduciary", and "disclose" are all associated to the nature-of-suit category Fraud. In the fourth Key Phrase, the terms "emotional", "distress", and "reckless" are all associated to the nature-of-suit category Non-PI/PD/WD tort – Other.

Although high probability Key Phrases resulted in correct classifications, they also resulted in misclassifications by the maximum probability method. The maximum probability method frequently misclassifies documents as Writ of Mandate and Wrongful Eviction. These misclassifications are due to individual Key Phrases containing terminology which is strongly associated to these categories.

The weighted probability method is designed to account for the limitations in the maximum occurrence method and maximum probability method. While the maximum occurrence method ignores the probabilities for each Key Phrase and the maximum probability method ignores the classifications for low probability Key Phrases, the weighted probability method leverages all the information to determine the nature of suit. The weighted probability method displays an accuracy improvement of 7% over the maximum occurrence method and 4% over the maximum probability

method (Table 9). Improvements are also seen in precision and recall, most notably in Breach of Contract/Warranty and Fraud. These improvements are due to the summation of probabilities. Documents which are misclassified as Breach of Contract/Warranty or Fraud by the maximum probability method due to a single high probability Key Phrase, are often reclassified into the correct nature of suit. Figure 39, below, provides an example of a correct reclassification.

Key Phrases	Probability	Classification
Plaintiff identified LYON as the listing agent.	0.80%	Breach of Contract/Warranty
PLAINTIFF is informed and believes and thereon alleged that CHAS hired LYON to list 8445 Barton Road .	1.97%	Breach of Contract/Warranty
In addition , PLAINTIFF has incurred and will continue incurring additional damages , cost and egipenses , including attomeys ' fees , as a result of LYON 'S and ACKERMAN breach of their statutory duty .	28.22%	Breach of Contract/Warranty
DEFENDANT CHERYL ACKERMAN GRIFFIN was a salesperson licensed by the California Department of Real Estate , and working as an agent for DEFENDAN WILLIAM L. LYONS & ASSOCIATES .	22.16%	Other Real Property
DEFENDANT WILLIAM L. LYON & ASSOCIATES , INC. was the real estate broker for the sale of 8445 Barton Road .	9.10%	Other Real Property

Classification Mode	Document Classification
Maximum Occurrence	Breach of Contract/Warranty
Maximum Probability	Breach of Contract/Warranty
Weighted Probability	Other Real Property

Figure 39 - Weighted Probability Reclassification

In Figure 39, five Key Phrases are displayed from a Memorandum of Points and Authorities with a nature-of-suit category of Other Real Property. Three of the five Key Phrases are classified as Breach of Contract/Warranty, one of which has a high probability of 28.22%. These classifications cause the maximum occurrence method and maximum probability method to

misclassify the document as Breach of Contract/Warranty. However, the weighted probability method correctly classifies the document as Other Real Property due to summing the probabilities (Other Real Property sum is 31.26%, Breach of Contract/Warranty sum is 30.99%).

The strengths and weaknesses of the classification methods are showcased when attempting to differentiate between PI/PD/WD – Other and PI/PD/WD – Auto. Differentiating between these two nature-of-suit categories proved to be a difficult task. Both categories involve personal injury, property damage, and wrongful death, but PI/PD/WD – Auto also includes an automobile accident or collision. Table 13, below, provides the average precision and recall scores for these categories.

	Classification Method					
	Maximum Occurrence		**Maximum Probability**		**Weighted Probability**	
Nature of Suit	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
PI/PD/WD – Auto	93.07%	68.61%	70.80%	70.80%	76.64%	76.64%
PI/PD/WD – Other	57.92%	80.21%	76.33%	100.00%	68.18%	96.26%
Average	**75.50%**	**74.41%**	**73.57%**	**85.40%**	**72.41%**	**86.45%**

Table 13 - Average Precision and Recall for PI/PD/WD - Other and PI/PD/WD – Auto

The maximum occurrence method performs the worst when differentiating between the two. The maximum probability method and weighted probability method display significant improvements. While the precision scores decline slightly, the recall score improves over 10%. The improvements are due to the strong associations of automobile related terminology. Terms such as “automobile”, “accident”, and “vehicle” yield high probabilities in Key Phrases.

VIII. FUTURE WORK

Future work involves two categories of goals, short term and long term. The short term goals are made up of small improvements that will have major impacts on the existing version of CLAS. Short term improvements include adapting CLAS to (1) identify and extract legal citations,

(2) assess the quality of the information contained in Key Phrases prior to classification, (3) implement a classification process that will support vectors representing up to three words, and (4) support the classifications of multiple categories for a single Memorandum of Points and Authorities. For the long term, CLAS will be extended into a hybrid case-based reasoning system that can evaluate a Memorandum of Points and Authorities by comparison with other similar Memoranda of Points and Authorities, and provide suggestions for improvements.

Identifying and extracting legal citations introduces a new data point and mitigates a current limitation created by the fact that legal documents often contain summary explanations of the cited prior lawsuits and those summary explanations usually reference the parties in those cited lawsuits. CLAS currently has the possibility of misidentifying these cited case summary explanations as Key Phrases. Beyond mitigating a current limitation of CLAS, the identification of the legal citations in a suit can provide an understanding of how these citations apply within the law. For example, if the legal citation to *John Smith v. Jane Doe* appears in 80% of the cases that are classified as Civil Rights then this citation could assist in the classification of legal arguments contained in that document. Thus, the ability to define relevance for a citation can provide a method for evaluating the quality of the legal arguments. Further, a document could be evaluated by determining how relevant the citations are to the classified nature of suit. However, identifying citations provides additional challenges too, like overcoming the inconsistencies in the citation formatting across types of references. Figure 40, below, provides examples of the different citation formats for a lawsuit, providing both lawsuit citations (*e.g.*, *Brock v. Superior Court*) versus a statute citation (Cal Civ. Proc. § 2017.010).

[20] Declaration of Amy Roberts, ¶8
[21] Cal.Civ. Code, § 3287(a); Children's Hosp. & Med. Ctr. V. Bonta (2002) 97 Cal.App.4th 740, 774 (test for determining whether damages are certain and prejudgment interest is owed is whether defendant actually knows the amount owed or could reasonable compute it based on reasonably available information).
[22] Cal. Civ. Code, § 3289.

obtain discovery regarding any matter, *not privileged*, that is relevant to the subject matter involved in the pending action, if the matter either is itself admissible in evidence or appears reasonably calculated to lead to the discovery of admissible evidence (Code Civ. Proc. § 2017.010.)

Attorneys may make binding stipulations on a great many matters of procedure, and, where no public policy is opposed, may waive procedural rights. (See *Brock v. Superior Court* (1947) 29 Cal.2d 629, 634). Since the parties here previously stipulated that Napa County was the proper

Figure 40 - Inconsistent Citation Formats

Being able to assess the quality of the information contained in Key Phrases prior to classification will also improve the accuracy of CLAS's nature-of-suit classification. High quality information provides detailed facts about the past interaction that gave rise to the current lawsuit. Whereas, low quality information may only provide high-level summaries of the past interaction. Identifying high quality relationships will be implemented by extending the first-order logic described in this thesis and using a set of known objects and events specifically developed for nature-of-suit classification. These extensions will be incorporated into the next version of CLAS. If a high quality relationship is found, the Key Phrase will have a higher weight during the classification of the document, thus having more influence over the document classification. For example, first-order logic will be used to determine if a Key Phrase contains a relationship between two Entities that entered or exited a contract, instead of just considering whether the entities "entered" something. So when selecting the Key Phrase, "Plaintiff John Smith entered into a

contract with Defendant Jane Doe," the Key Phrase Selection component will not exclusively rely on the verb "entered," but will instead consider the verbs descriptors as well to include the phrase "entered into a contract" to assess the quality of the information contained in the sentence. The Key Phrase must contain a specific event in order to qualify. Figure 41, below, provides an example of the first-order logic used to identify a contract relationship between two Entities.

$$\forall x\ \text{PARTY}(x) \Rightarrow \text{PLAINTIFF}(x) \vee \text{DEFENDANT}(x)$$

$$\forall x\ \text{ENTITY}(x) \Rightarrow \text{PARTY}(x) \vee \text{PERSON}(x) \vee \text{ORGANIZATION}(x) \vee \text{PROPER-NOUN}(x)$$

$$\text{STARTED}(\text{CONTRACT}) = [\text{ENTERED-INTO}, \text{SIGNED}, \text{AGREED-ON}, \text{INITIATED}]$$
$$\text{FINISHED}(\text{CONTRACT}) = [\text{BREACHED}, \text{COMPLETED}, \text{VOIDED}]$$

$$\forall c, x, y\ \text{CONTRACT}(\text{STARTED}(c), \text{ENTITY}(x), \text{ENTITY}(y)) \Rightarrow \text{CONTRACT}(\text{ENTITY}(x), \text{ENTITY}(y))$$

$$\forall c, x, y\ \text{CONTRACT}(\text{FINISHED}(c), \text{ENTITY}(x), \text{ENTITY}(y)) \Rightarrow \neg\text{CONTRACT}(\text{ENTITY}(x), \text{ENTITY}(y))$$

Figure 41 - Identifying a Contract Relationship between two Entities

In Figure 41, the first-order logic representation uses a set of known events for starting a contract or finishing a contract. These known events are used to identify a contract relationship between two Entities. The events for starting a contract are ENTERED-INTO, SIGNED, AGREED-ON, and INITIATED. The events for finishing a contract are BREACHED, COMPLETED, and VOIDED. Further first-order logic extensions will be implemented for identifying relationships that involve automobile accidents, medical procedures, non-disclosures of information, and monetary collections. Any Key Phrase that contains a high quality relationship will have an increased weight when determining the nature-of-suit classification for the document.

Adapting CLAS to support vectors that represent multiple words will also directly improve its value, this time by improving the classification accuracy. Although the current version of CLAS focuses on non-legal terms, some multi-word legal terms have direct relationships with nature-of-suit categories. These legal terms cannot be captured using the single word vectors that

CLAS currently indexes, instead, CLAS will be extended to support indexing multi-word vectors. An example of multi-word legal terms and their associated nature-of-suit category can be seen in Table 14, below.

Term	Related Nature of Suit
fraudulent transfer	Fraud
breached contract	Breach of Contract/Warranty
actual loss	Fraud
transactional malpractice	Professional Negligence
reaffirmation agreement	Rule 3.740 Collections / Other Collections
wage garnishment	Rule 3.740 Collections / Other Collections

Table 14 - Multi-Word Legal Terminology

CLAS's current process of using individual words as vectors does not maintain the association between a multi-word term and the related nature of suit. The inclusion of multi-word vectors would increase the accuracy of the classifier because the direct relationships to nature-of-suit categories would be maintained.

Similarly, supporting multiple nature-of-suit categories for a single Memorandum of Points and Authorities will improve the ability of CLAS to accurately represent the information contained in the document. The current limitation of one nature of suit per document fails to represent the complexity of the documents. A single Memorandum of Points and Authorities (and indeed a lawsuit) can have many different nature-of-suit categories. Supporting the classification of multiple nature-of-suit categories creates a more robust system which more accurately reflects the complexity of legal information. Supporting multiple nature-of-suit categories also provides additional insight into the relationships between the categories. For example, if 90% of the documents which contain a classification of Professional Negligence also contain a classification of Fraud, these two nature-of-suit categories may be related. Multiple nature-of-suit categories could also provide insight into the relationship between citations and the nature of suit. For

example, if a specific citation is frequently used in Memoranda of Points and Authorities which are classified as both Professional Negligence and Fraud, the citation may be related to both nature-of-suit categories. Evaluating multiple nature-of-suit categories would require an extension to the current evaluation data. The evaluation data would need to include all of the nature-of-suit categories for each Memorandum of Points and Authorities. Obtaining this information would require a person with legal expertise to manually identify the nature-of-suit categories in each document.

While these short terms goals will have immediate impact, the long term goal is geared towards directly assisting attorneys by developing CLAS as a hybrid system. CLAS will include an evaluation component that provides feedback on the expected effectiveness of a user-provided Memorandum of Points and Authorities and suggestions for improvements.

CLAS will accept a Memorandum of Points and Authorities (such as a lawyer's draft of a future filing) and then compare that input to locate similar Memorandum of Points and Authorities already indexed in CLAS. CLAS will include an improved record system that indexes multiple nature-of-suit categories for each document and extracted citations. CLAS will compare the current document with prior documents using the classified nature-of-suit category and extracted citations. Documents that share the nature-of-suit categories and citations will be returned to the user. CLAS will also prioritize similar documents with positive outcomes (positive rulings). Then CLAS will evaluate the current document based on the outcomes of the similar prior documents, and provide feedback to the user of how to improve by highlighting missing citations or incorrectly used citations.

In order to perform the evaluation, it would be necessary to collect and correlate court-ordered outcomes for each prior document. This would present a difficult challenge. First, due to

current access inefficiencies in the legal market, locating the corresponding court orders is not an easy task. Moreover, court orders either (1) grant the filing parties request, (2) deny the request or (3) grant the request in part. Thus, it is not always a straight-forward result. Evaluating and classifying the outcome of each would likely require a person with legal expertise to review each document and find the associated outcome(s). It may be useful to first identify each request contained in the underlying document and then provide the corresponding outcome for each request, as opposed to relating each document to a single outcome.

These extensions to CLAS would have real world applications for practicing attorneys and law students (collectively referred to as "users"). CLAS could be used as a research tool to find relevant Memorandum of Points and Authorities using a variety of input sources. If a user has an existing Memorandum of Points and Authorities that is relevant to their research, they can use the document as input to CLAS to find additional relevant documents. Users may also use CLAS to find documents which relate to their own case or their opponents case in order to be prepared for future arguments or to validate they have exhausted all relevant citations and are correctly using their current citations. Newly suggested citations would allow users to further strengthen their arguments.

IX. APPENDIX

Penn Treebank

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Stanford Grammatical Relations

Function	Grammatical Relation
dep	dependent
aux	auxiliary
auxpass	passive auxiliary
cop	copula
arg	argument
agent	agent
comp	complement
acomp	adjectival complement
ccomp	clausal complement with internal subject
xcomp	clausal complement with external subject
obj	object
dobj	direct object
iobj	indirect object
pobj	object of preposition
subj	subject
nsubj	nominal subject
nsubjpass	passive nominal subject
csubj	clausal subject
csubjpass	passive clausal subject
cc	coordination
conj	conjunct
expl	expletive (expletive “there”)
mod	modifier
amod	adjectival modifier
appos	appositional modifier
advcl	adverbial clause modifier
det	determiner
predet	predeterminer
preconj	preconjunct
vmod	reduced, non-finite verbal modifier
mwe	multi-word expression modifier
mark	marker (word introducing an advcl or ccomp)
advmod	adverbial modifier
neg	negation modifier
rcmod	relative clause modifier
quantmod	quantifier modifier
nn	noun compound modifier
npadvmod	noun phrase adverbial modifier

tmod	temporal modifier
num	numeric modifier
number	element of compound number
prep	prepositional modifier
poss	possession modifier
possessive	possessive modifier (’s)
prt	phrasal verb particle
parataxis	parataxis
punct	punctuation
ref	referent
sdep	semantic dependent
xsubj	controlling subject

Stop Words

a	can't	he	it's	over	they	when
about	cannot	he'd	its	own	they'd	when's
above	could	he'll	itself	same	they'll	where
after	couldn't	he's	let's	shan't	they're	where's
again	did	her	me	she	they've	which
against	didn't	here	more	she'd	this	while
all	do	here's	most	she'll	those	who
am	does	hers	mustn't	she's	through	who's
an	doesn't	herself	my	should	to	whom
and	doing	him	myself	shouldn't	too	why
any	don't	himself	no	so	under	why's
are	down	his	nor	some	until	with
aren't	during	how	not	such	up	won't
as	each	how's	of	than	very	would
at	few	i	off	that	was	wouldn't
be	for	i'd	on	that's	wasn't	you
because	from	i'll	once	the	we	you'd
been	further	i'm	only	their	we'd	you'll
before	had	i've	or	theirs	we'll	you're
being	hadn't	if	other	them	we're	you've
below	has	in	ought	themselves	we've	your
between	hasn't	into	our	then	were	yours
both	have	is	ours	there	weren't	yourself
but	haven't	isn't	ourselves	there's	what	yourselves
by	having	it	out	these	what's	

X. REFERENCES

1. Aleven, V., & Ashley, K. D. (1997, August). Teaching case-based argumentation through a model and examples empirical evaluation of an intelligent learning environment. In *Artificial intelligence in education* (Vol. 39, pp. 87-94).

2. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. *The VLDB Journal—The International Journal on Very Large Data Bases*, *18*(1), 255-276.

3. Breuker, J. (1993). Modelling artificial legal reasoning. In *Knowledge Acquisition for Knowledge-Based Systems* (pp. 66-78). Springer Berlin Heidelberg.

4. Breukers, J. A. P. J., & Hoekstra, R. J. (2004). Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law.

5. Brüninghaus, S., & Ashley, K. D. (2001, May). Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th international conference on Artificial intelligence and law* (pp. 42-51). ACM.

6. Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740-750).

7. Cohen, W. W., & Richman, J. (2002, July). Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475-480). ACM.

8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493-2537.

9. Conrad, J. G., Dozier, C., Molina-Salgado, H., Thomas, M., & Veeramachaneni, S. (2011, June). Public record aggregation using semi-supervised entity resolution. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law* (pp. 239-248). ACM.

10. Conroy, J. M., & O'leary, D. P. (2001, September). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 406-407). ACM.

11. Creenleaf, G., Mowbray, A., & Tyree, A. (1991, May). The DataLex legal workstation: integrating tools for lawyers. In *Proceedings of the 3rd international conference on Artificial intelligence and law* (pp. 215-224). ACM.

12. Das-Gupta, P. (1987). Boolean interpretation of conjunctions for document retrieval. *Journal of the American Society for Information Science*, *38*(4), 245-254.

13. Deliverable 1.4: OWL Ontology of Basic Legal Concepts (LKIF-Core). (2007). Retrieved from http://www.estrellaproject.org/doc/D1.4-OWL-Ontology-of-Basic-Legal-Concepts.pdf

14. Gelbart, D., & Smith, J. C. (1993, August). FLEXICON: an evaluation of a statistical ranking model adapted to intelligent legal text management. In *Proceedings of the 4th international conference on Artificial intelligence and law* (pp. 142-151). ACM.

15. Grover, C., Hachey, B., Hughson, I., & Korycinski, C. (2003, June). Automatic summarisation of legal documents. In *Proceedings of the 9th international conference on Artificial intelligence and law* (pp. 243-251). ACM.

16. Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

17. Hachey, B., & Grover, C. (2004). A rhetorical status classifier for legal text summarisation. In *Proceedings of the ACL-2004 Text Summarization Branches Out Workshop*.

18. Hachey, B., & Grover, C. (2005, June). Automatic legal text summarisation: experiments with summary structuring. In *Proceedings of the 10th international conference on Artificial intelligence and law* (pp. 75-84). ACM.

19. Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, *6*(1), 3-34.

20. Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, *5*, 361-397.

21. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

22. McKeown, K., & Radev, D. R. (1995, July). Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 74-82). ACM.

23. Nenkova, A. (2005, July). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI* (Vol. 5, pp. 1436-1441).

24. Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

25. Quaresma, P., & Rodrigues, I. P. (1999, June). A collaborative legal information retrieval system using dynamic logic programming. In *Proceedings of the 7th international conference on Artificial intelligence and law* (pp. 190-191). ACM.

26. Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*,*40*(6), 919-938.

27. Richards Jr, R. C., & Bruce, T. R. (2011, June). Adapting specialized legal metadata to the digital environment: The code of federal regulations parallel table of authorities and rules. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law* (pp. 126-130). ACM.

28. Rissland, E. L., & Daniels, J. J. (1995, May). A hybrid cbr-ir approach to legal information retrieval. In *Proceedings of the 5th international conference on Artificial intelligence and law* (pp. 52-61). ACM.

29. Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of documentation*, *33*(4), 294-304.

30. Rose, J. F., Grenager, T., & Manning, C. D. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (pp. 363-370). Association for Computational Linguistics.

31. Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, *26*(11), 1022-1036.

32. Smith, J. C. (1997, June). The use of lexicons in information retrieval in legal databases. In *Proceedings of the 6th international conference on Artificial intelligence and law* (pp. 29-38). ACM.

33. Smith, M. (1990). *Aspects of the P-Norm model of information retrieval: syntactic query generation, efficiency, and theoretical properties*. Cornell University.

34. Stranieri, A., & Zeleznikow, J. (1999, June). The evaluation of legal knowledge based systems. In *Proceedings of the 7th international conference on Artificial intelligence and law* (pp. 18-24). ACM.

35. Thomson Reuters. (2015). PeopleMap. Retrieved from http://legalsolutions.thomsonreuters.com/law-products/westlaw-legal-research/public-records/peoplemap.

36. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.

37. Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS)*, *9*(3), 187-222.

38. Valente, A. (1993). Preliminary steps in causal legal reasoning. In Cohn A., Lee M., Preist C., & Price C., editors. *Proceedings of the AISB Workshop on Qualitative and Causal Reasoning*.

39. Valente, A. (1995). Legal Knowledge Engineering. *A Modelling Approach, IOS Press, Amsterdam, Dissertation*.

40. Valente, A., & Breuker, J. (1994). Making ends meet: conceptual models and ontologies in legal problem solving. In *Proceedings of the XI Brazilian AI Symposium (SBIA'94)* (pp. 1-15).

41. Valente, A., & Breuker, J. (1995, May). ON-LINE: An architecture for modelling legal information. In *Proceedings of the 5th international conference on Artificial intelligence and law* (pp. 307-315). ACM.

42. Veeramachaneni, S., & Kondadadi, R. K. (2009, June). Surrogate learning: from feature independence to semi-supervised classification. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 10-18). Association for Computational Linguistics.

43. Verma, M., & Varma, V. (2011, June). Applying Key Phrase extraction to aid invalidity search. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law* (pp. 249-255). ACM.