Event Extraction for Legal Case Building and Reasoning

Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O'Neill,

Xerox Research Centre Europe 6, chemin de Maupertuis, 38240 Meylan, France
{Nikolaos.Lagos, Frederique.Segond, Stefania.Castellani, Jacki.Oneill}@xrce.xerox.com

Abstract. We are interested in developing tools to support the activities of lawyers in corporate litigation. In current applications, information such as characters that have played a significant role in a case, events in which they have participated, people they have been in contact, etc., have to be manually identified. There is little in the way of support to help them identify the relevant information in the first place. In this paper, we describe an approach to semi-automatically extracting such information from the collection of documents the lawyers are searching. Our approach is based on Natural Language Processing techniques and it enables the use of entity related information corresponding to the relations among the key players of a case, extracted in the form of events.

Keywords: legal case building and reasoning, natural language processing, information extraction, e-discovery, event extraction, knowledge-based event analysis.

1 Introduction

We are interested in developing tools to support the activities of lawyers in corporate litigation, that is, the process of bringing and pursuing lawsuits. Typically corporate litigation involves the processing and analysis of large volumes of documents with the aim of finding evidence for or against the litigation claims. Identifying the important information is time consuming and costly and in recent years there has been a move to bring into play language processing technologies. Litigation involves a number of stages with different support requirements, from preservation and collection of all documents with possible relevance to the case; to review – a filtering process reducing the document set to those (non-confidential documents) answering specific criteria of relevance; and case construction – where arguments around facts and evidence are put together for presentation in court. The primary goal of the searching and browsing facilities offered in current litigation tools is to find relevant documents - often based on keyword/boolean based search techniques. Although this has proved to be relatively useful in the first stages of litigation, e.g. to identify responsive documents, during case construction the emphasis shifts from finding documents to finding entities and actionable information ([1], [2]) derived from these entities.

We are developing a system to help lawyers process the large document collections associated with a legal case and extract information from them to build the case. The idea is to provide some forms of semi-automatic support for the lawyers working to identify, for example, characters that have played a role in a case, events they have participated in, who they have been in contact with, etc. This kind of search is an important part of the work and tools currently on the market allow the users to store information on relevant characters and events. However, in current applications as users identify relevant information they must manually enter it in the tools database. There is little in the way of support to help them identify the relevant information in the first place. In this paper we describe how information might be semi-automatically extracted from the collection of documents the lawyers are searching. Our approach is based on Natural Language Processing (NLP) techniques and it enables the use of entity related information corresponding to the relations among the key players of a case, extracted in the form of events. Events are viewed as temporally bounded objects that have entities important within the application domain (e.g. persons and organisations) as participants. We chose a semi-automatic approach because case building requires deep semantic understanding of the events described in documents, thus people are integral to the process, but we also believe that information analysis can valuably support their work.

2 Legal Case Building and Reasoning

The e-discovery phase of litigation (preservation, collection and review) has long been a focus point for applying search ([3], [4]) and other technologies (e.g. [5]) in an attempt to help the lawyers manage the enormous document sets. The cost of this area has made it a particular focus for the implementation of technology as even small improvements can produce major savings. However, we believe that linguistic technologies could also benefit later stages of review, making it is easier to find information and construct cases [6]. One way of supporting such activities is to develop effective search mechanisms that aid in the discovery of relevant characters and events.

Even after responsiveness review large volumes of documents often remain, few of which actually contain information important to the case. Thus lawyers have to peruse many, often deadly dull, documents in search of anything that might stand as evidence. In addition, any single document often only contains part of the information which may be used to construct a fact – it is often the contents of a set of documents that constitutes evidence for any fact. Since the document set is large, it is usually divided between groups of paralegals and lawyers – either passing through iterations of relevance or being equally divided amongst the case team. In either case a potential problem is that something is not seen as relevant when first uncovered because the information which will make it relevant has not yet been found (or has been found by someone else). On finding partial information lawyers must currently choose whether to follow that line of enquiry themselves (when they may not have the relevant documents) or hope someone else turns up and notices the additional information [7]. Currently this distribution of information is managed through duplication and extra effort. Thus technologies which could help lawyers better find, explore and manage the information within the entire document set could be useful.

As an example of information searching in this context, let's consider a scenario inspired by the TREC Legal Track collection [8]. One of the issues explores the proposition that cigarettes were sold to people from defendant D, with the misinterpretation, promoted by D that they were not doing any harm while D knew that they were. A sub-issue relates to proving that D denied fraudulently that nicotine is addictive. A tool that would help the user to construct the case around the above issue should support the search for information such as:

- what tests were carried out about the addictiveness of nicotine? who conducted them? when? with what results?
- when were the test results published? who saw them?
- what meetings did key people attend after the tests? who else participated in them?
- what publicity did the company release after the production of the tests?

The following sections illustrate our approach towards the extraction of information for answering these kinds of questions.

3 Event-based Information Model for Litigation

People and organisations are typical examples of characters that may have a role in a legal case. However, depending on the litigation domain, other kinds of characters may need to be extracted. For example, in the tobacco case the following are also important: chemical substances, e.g. nicotine; products, e.g. cigarettes; and monetary expressions. The role of the characters in a case is determined, among other factors, by the events in which they participate. For instance, the role of an executive officer (EO) who publicly states something relevant to the subject of a case is more central to the case than that of other EO not involved. Naturally that is a two way relationship. The events that a key character participates in may be important for the case, but also the participants of a key event may be key characters. One of the core requirements is therefore identifying other factors (in addition to the participants) that make an event important. These include:

- the topic of an event, if any – for instance, in our example identifying that a person stated something about nicotine;
- the role of a character in the event – this enables, for example, the cases where an EO states something to be distinguished from the ones where he/she is considered in the topic of a statement;
- the relative time of an event in the chronology of the case – e.g. has the EO made a statement after contradicting results were communicated to him;

- the location that the event took place – for example did tests on tobacco take place in the company's laboratories indicating knowledge of the results from the company itself?

Events are extracted from the collection of documents associated to a legal case. They may describe situations, e.g. meetings, actions, e.g. studying, or even statuses, e.g. belong to. The events identified will depend on the matters and the domain that the legal case covers. For example, in our scenario events related to nicotine will need to be extracted (c.f. section 4.2).
Additionally, we have identified a number of classes of relations among people and organisations that we believe to be of interest to lawyers, during case construction, independently from the litigation domain. Those classes correspond to events or event abstractions and include the following:

- Role-based events such as "is employed by" (i.e. employee-employer relation).
- Interaction-based events, such as "meets", which corresponds to the act of an entity interacting with another entity (i.e. person or of type organisation).
- Reference events such as "says", correspond to the act of an entity referring to another entity through a written or spoken message.
- Cognitive events such as "knows" which indicate possible knowledge of a topic or entity. For example the publication of a study or writing of an email indicates the authors' knowledge of the contents.

4 Knowledge-based system for event extraction and analysis

In order to manipulate the information described above, a system that combines event extraction, integration, and inference is required. The architecture of such a system is illustrated in Figure 1.

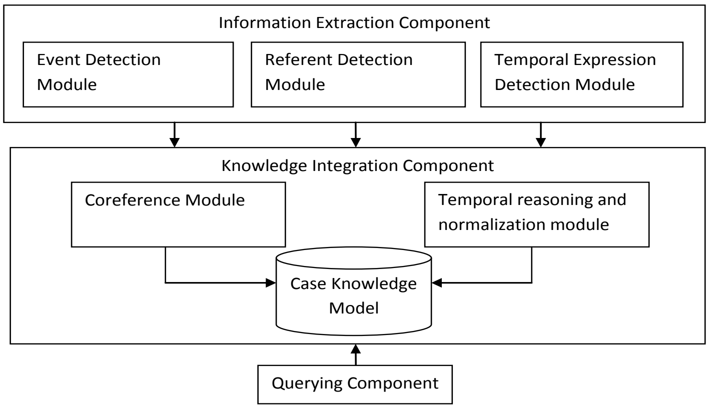

Information Extraction Component
Event Detection Module
Referent Detection Module
Temporal Expression Detection Module
Knowledge Integration Component
Coreference Module
Temporal reasoning and normalization module
Case Knowledge Model
Querying Component

Fig. 1. Overall architecture of the knowledge-based system.

The main components are: an Information Extraction Component, used to extract events, named entities along with their attributes, and temporal expressions; a Knowledge Integration Component that integrates the data extracted and infers additional information; and a Querying Component enabling user interaction. The remaining of this section details the first two components.

4.1 Information Extraction Component

The Information Extraction Component is based on the Xerox Incremental Parser (XIP) [9]. XIP combines the following five linguistic processing layers: preprocessing (tokenization, morphological analyzer and part of speech tagging); named entity extraction; chunking; extraction of dependencies between words on the basis of sub-tree patterns over chunks sequences, and a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies. XIP is altogether robust - able to process large collections of any documents' type, able to extract different types of relations between words and groups of words - and efficient (about 2000 words per second on an average PC).

XIP's event recognition module is also used, including a named entity extraction sub-module that detects and "semantically" categorizes proper nouns related to events. Event detection in our system is based on the approach described in [10] where an event description is considered as "a predicate (verb, adjective and predicative noun) related to its arguments and modifiers". For example, in our scenario it is important to find studies that talk about nicotine. Thus, we should be able to identify events referring to nicotine, as shown in the example in Figure 2.

```
"The 1983 internal study conducted by Person X allegedly concluded that nicotine
was an addictive substance."
----
   PERSON (Person X)
   OBJ-N_PRDEP_CONT-ACT_KNOW (conduct,study)
   COORDINATE (study,conclude)
   COORDINATE_ACTOR_NENTITY (conduct,Person X)
   COORDINATE_ACTOR_NENTITY (conclude,Person X)
   CORE-DRIVE_SAY (conclude,be)
   COORDINATE_ACTOR (be,nicotine)
```

Fig. 2. Named Entity and Event Recognition by XIP.

4.2 Knowledge Integration Component

The case knowledge model (or ontology), used to describe the data to be extracted for supporting case building and reasoning activities, has three different layers. The first layer supports the integration with the indexing tools focusing on low-level features (such as text zones); the second represents concepts useful in the investigation domain (such as people, organizations, and relations); and the last allows case specific information to be included (e.g. for the tobacco case, chemical elements and substances). Based on the previous analysis, the following information has to be explicitly represented:

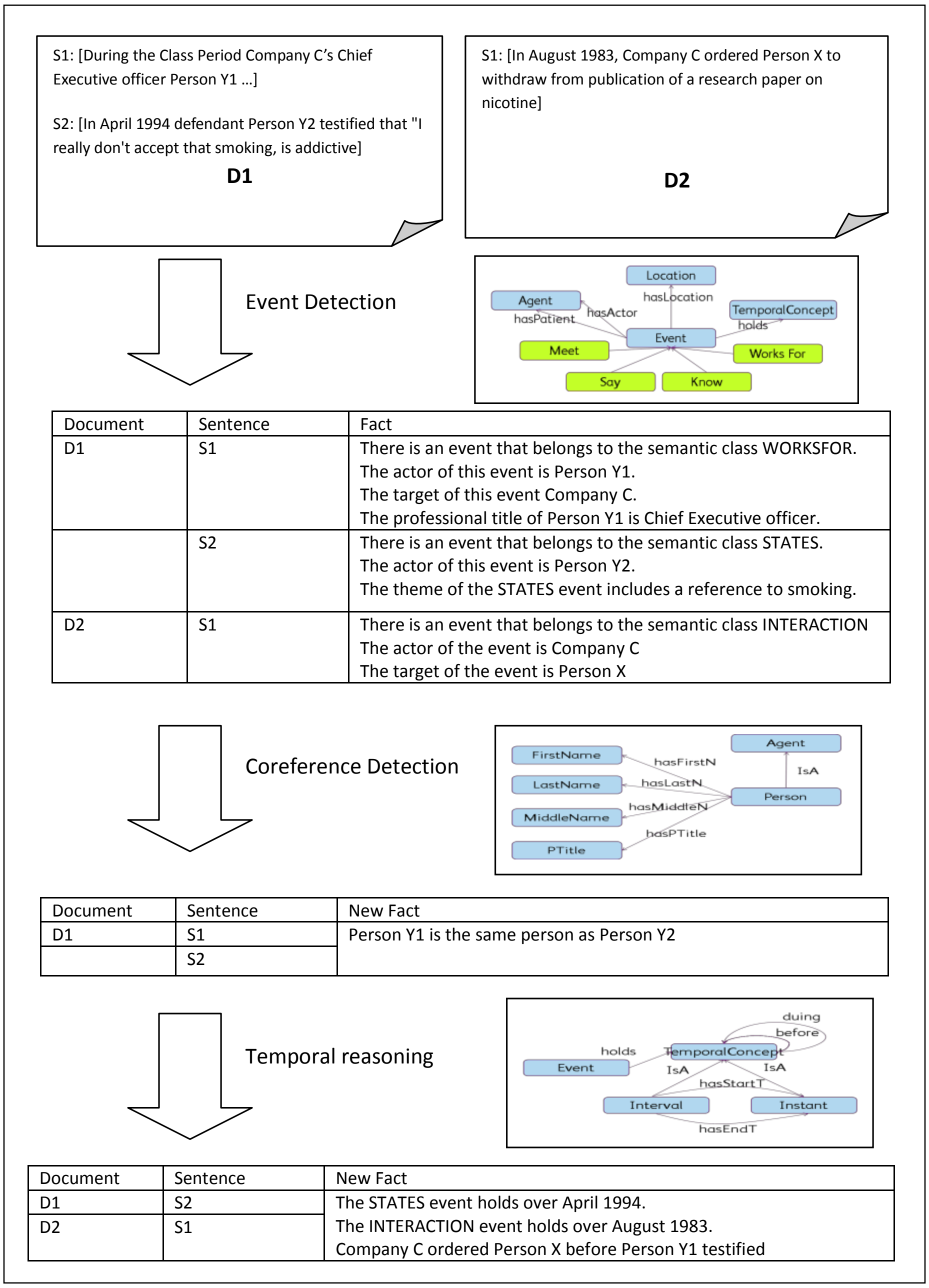
- Concepts defining organizations, people, documents, locations, dates/times and events (where an event is used to define states and transitions between states);
- Temporal and spatial features of events (when and where an event occurred);
- Event classes (e.g. X talks to Y and X emails Y can be abstracted as X CONTACT Y with CONTACT being an event class, X its agent and Y its patient). Motivated by our scenarios, key events in our context include cognitive, interaction, reference, and role-based ones.

To integrate information for actors involved in the litigation case, the coreference module is used. The module is able to identify the same entity occurring in the text several times with similar naming variations, using among others a string-edit distance algorithm, and also to track pronominal coreference (e.g. he, she). Distance between a co-reference and the nearest named entity is often used for disambiguating these cases, but this is not the only existing method. In traditional coreference, information such as birth date, birth place, etc. can be used to aid in identifying a chain of objects (i.e. referents) that refer to the same entity. In litigation though, the most important features that can be used include the name, the professional title and the social network of the referent. We have used the first two features.

In order to generate a timeline, static diary-like structures anchored to specific instants are needed, as well as temporally-based relations between events. All events are defined therefore in terms of the temporal interval attached to them. All intervals are identified based on some ordered pair of instants and a set of relations to other intervals. Conversely, any given interval or chain of intervals allows us to identify two instants, corresponding to the interval’s or chain of intervals' beginning and end time points. As in [11] we define these time points as unique and stand for corresponding points on the timeline. A temporal graph is computed within each document using Allen's temporal relations with a forward reasoning engine [12] while all instants are normalised by changing their granularity to that of a day (that granularity was selected as the most suitable according to our objectives and the data being available). Posing queries over temporally related sequences of events thus becomes possible, bearing in mind that the temporal information may be over or under-specified.

4.3 Example

In our scenario the users of such a system may want to search for chief executive officers of Company C that said something about smoking after the results of nicotine related tests were released. Figure 3 demonstrates the extraction and integration process.



S1: [During the Class Period Company C's Chief Executive officer Person Y1 …]

S2: [In April 1994 defendant Person Y2 testified that "I really don't accept that smoking, is addictive]

D1

S1: [In August 1983, Company C ordered Person X to withdraw from publication of a research paper on nicotine]

D2

Document	Sentence	Fact
D1	S1	There is an event that belongs to the semantic class WORKSFOR. The actor of this event is Person Y1. The target of this event Company C. The professional title of Person Y1 is Chief Executive officer.
	S2	There is an event that belongs to the semantic class STATES. The actor of this event is Person Y2. The theme of the STATES event includes a reference to smoking.
D2	S1	There is an event that belongs to the semantic class INTERACTION The actor of the event is Company C The target of the event is Person X

Document	Sentence	New Fact
D1	S1	Person Y1 is the same person as Person Y2
	S2	

Document	Sentence	New Fact
D1	S2	The STATES event holds over April 1994.
D2	S1	The INTERACTION event holds over August 1983. Company C ordered Person X before Person Y1 testified

Fig. 3. Example of extraction and integration process.

The first sentence with the coreference module enables to collect the facts indicating that a CEO of Company C has stated something about smoking while the temporal reasoning module enables the discovery of the fact that the statement was done after Company C ordered Person X to withdraw a relevant paper (Person X as a chief scientist involved in the nicotine addiction tests, is assumed to be one of the case's key people).

5 Related Work

Our work is closely related to ongoing research in the NLP community on extraction of events, temporal information, entities, and links among them. In particular, there are several evaluation campaigns, e.g. the TREC Entity Track [13], the Automatic Content Extraction (ACE) [14], on event extraction, the TempEval task evaluation, focusing on the discovery of temporal event chains [15], and SemEval 2010, on the identification of event participants [16]. This research forms a complementary and essential part of our work, since the accuracy of events extraction is an important factor to the acceptance of our system.

Other areas such as Knowledge Representation have mainly focused on the modeling and inference aspects. More particularly for legal data a number of models that explicitly represent events as first-class objects have been developed including the DOLCE-CLO [17], LRI-Core [18], and LKIF [19]. All those models focus on legal texts and deontic knowledge, while some of them are based on fundamental ontologies that do not recommend event subclasses specialized to the litigation domain, as in our case.

There is also work that integrates the two fields described above to achieve efficient legal document retrieval. [20] has applied event extraction for legal case retrieval. An event in that case is defined as any eventuality (event, state and attribute) related to illocutionary expressions existing in legal texts and therefore there has a different focus to ours. [21] also use an NLP system to define the rhetorical structure of legal texts and identify the parts in which the illocutionary expressions are present. [22] use Wikipedia to enrich the knowledge about entities and their relations for improving document retrieval in e-discovery.

A source of inspiration for all those works, including ours, is the TREC Legal Track [8] that, however, focuses on document retrieval rather than fine grained information extraction.

6 Conclusions and future work

In this paper we describe an NLP-based semi-automatic approach and system to support litigation case construction. We show that role, interaction, reference, and cognitive events –an event being a temporally bounded object having key entities as participants- can be used to represent relations among key characters of a case. We claim and demonstrate with an example that information integration from disparate events requires coreference resolution to identify objects describing the same entity and temporal reasoning to temporally qualify events and infer new information based on their relative chronological order. As event and participants' extraction is an integral part of our work we plan to continue improving our extraction accuracy and extending it to cross-document event extraction and tracking. In addition, we are interested in the development of user interaction components that will facilitate friendly navigation of the extracted and inferred information, as well as integration with other components of a legal case management system.

References

1. Noel, L. and Azemard, G.: From Semantic Web Data to Inform-Action: a Means to an End. In: ACM Computer Human Interaction, Florence, Italy (2008)
2. Sheth, A., Arpinar, B., Kashyap, V.: Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships, Technical Report, LSDIS Lab, Computer Science, Univ. of Georgia, Athens GA (2002)
3. Sedona Conference WG1: The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery, Vol. 8 (2007)
4. Baron, J. R., Thompson, P.: The Search Problem Posed By Large Heterogeneous Data Sets In Litigation: Possible Future Approaches To Research. In: 11th International Conference on Artificial Intelligence and Law, pp. 141--147, ACM, New York (2007)
5. O'Neill, J., Privault, C., Renders, J.-M., Ciriza, V., Bauduin, G.: DISCO: Intelligent Help for Document Review. In: Global E-Discovery/E-Disclosure Workshop – A Pre-Conference Workshop at the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain (2009)
6. Benedetti, V., Castellani, S., Grasso, A., Martin, D. and O'Neill, J.: Towards an Expanded Model of Litigation. In: DESI II, Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings, London, UK, (2008)
7. Attfield, S., Blandford, A., De Gabrielle, S.: E-discovery Viewed as Integrated Human-Computer Sensemaking: The Challenge of 'Frames'. In: DESI II, Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings, London, UK, (2008)
8. Oard, D.W., Hedin,B., Tomlinson,S., Baron, J.R.: Overview of the TREC 2008 Legal Track. 17th TREC, Gaithersburg, Maryland, USA (2008)
9. Ait-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness Beyond Shallowness: Incremental Deep Parsing. J. Nat. Lang. Eng. 8, 2-3, 121--144 (2002)
10. Capet, P., Delevallade, T., Nakamura, T., Tarsitano, C., Sandor, A., Voyatzi, S.: A Risk Assessment System with Automatic Extraction of Event Types. In: IIP2008 - 5th International Conference on Intelligent Information Processing. LNCS, vol. 288, pp. 220 – 229, Spring, Boston (2008)
11. Fikes, R., Zhou, Q.: A Reusable Time Ontology. AAAI Technical Report WS-02-11 (2002)
12. Hagege, C., Tannier, X.: XTM: A Robust Temporal Processor. In: 9th International Conference on Intelligent Text Processing and Computational Linguistics. LNCS, vol. 4919, pp. 231—240, Springer Berlin / Heidelberg, (2008)
13. Balog, K., de Vries, A.P., Serdyukov, P., Thomas, P., Westerveld, T.: Overview of the TREC 2009 entity track. In: 18th Text REtrieval Conference (2010)
14. NIST: The ACE 2005 Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/ace/ace05/doc/ace05-evalplan.v3.pdf

15. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007 Task 15: TempEval Temporal Relation Identification, ACL Workshop on SemEval (2007)
16. Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., Palmer, M.: SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. The NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09), Boulder, Colorado, USA (2009)
17. Gangemi, A., Pisanelli, D. M., Steve, G.: A Formal Ontology Framework to Represent Norm Dynamics. In: Proceedings of the 2nd International Workshop on Legal Ontologies (LEGONT) (2001)
18. Breuker, J., Hoekstra, R.: Core Concepts of Law: Taking Common-Sense Seriously. In: Proceedings of Formal Ontologies in Information Systems (FOIS-2004), pp. 210 -- 221, IOS Press (2004)
19. Hoekstra, R., Breuker, J., Bello M. D., Boer, A.: The LKIF Core Ontology of Basic Legal Concepts. In: Casanovas, P., Biasiotti, M.A., Francesconi, E., Sagri, M.T. (eds.) Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007), Stanford, CA, USA (2007)
20. Maxwell, K. T., Oberlander, J., Lavrenko, V.: Evaluation of Semantic Events for Legal Case Retrieval. In: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in information Retrieval (ESAIR), pp. 39--41, ACM, New York (2009)
21. Weber-Lee, R., Barcia, R. M., Costa, M. C., Filho, I. W., Hoeschl, H. C., Bueno, T. C., Martins, A., Pacheco, R. C.: A Large Case-Based Reasoner for Legal Cases, In: Leake, D.B., Plaza, E. (eds.) 2nd ICCBR. LNCS, vol. 1266, pp. 190—199. Springer-Verlag, London (1997)
22. Ka Kan L., Lam W.: Enhance Legal Retrieval Applications with Automatically Induced Knowledge Base. In: 11th International Conference on Artificial Intelligence and Law, California, USA (2007)