Temporal information extraction from legal documents

Frank Schilder[1] and Andrew McCulloh[1]

R&D, Thomson Legal & Regulatory
610 Opperman Drive, Eagan 55123, U.S.A.
{Frank.Schilder|Andrew.McCulloh}@Thomson.com

Abstract. The aim of this paper is to analyze what kinds of temporal information can be found in different types of legal documents. In particular, it provides a comparison of different legal document types (case law, statute or transactional document) and how one can do further reasoning with the extracted temporal information.

Keywords. extraction of temporal information, temporal reasoning, legal documents

1 Introduction

In the recent past, only few research has been carried out in legal reasoning looking at formalizing temporal information. This should come in particular as a surprise since laws, regulations and legal documents in general are normally filled with temporal information:

(1) Celltech owns a family of patents called the "Adair" patents and sought to claim royalties from Medimmune under a patent licence dated 19 January 1998.

Although temporal information is actually ubiquitous in legal text, systems for legal reasoning deal normally only on an 'ad-hoc-basis' with this important phenomenon [1]. With the exception of the special issue of Information & Communications Technology Law in 1998 [1,2,3], there is hardly any research on temporal information in legal text carried out. A couple of recent attempts focused on the specification of legal text in XML including temporal information [4,5,6]. Apart from these few research projects the extraction of temporal information has not been looked at in the literature. Traditionally, legal reasoning has been the focus of AI-related research, where the content of laws and regulations may, for example, become formalized in the event calculus [7]. Time may play a role within such a formalization, but it has not been the main focus of the formalization apart from a few exceptions.

However, it is important to note that legal reasoning is not the main focus of the current paper. Instead of looking at temporal information in legal reasoning, we are interested in temporal information in legal text and doing reasoning

Dagstuhl Seminar Proceedings 05151
Annotating, Extracting and Reasoning about Time and Events
http://drops.dagstuhl.de/opus/volltexte/2005/313

with the temporal information. We want to look at different types of legal text and investigate what kind of temporal information they can contain and after discussing how this information could be automatically extracted, how one could do reasoning with the temporal information in order to add more value to the document.

Section 2 contains an overview of different kinds of legal documents and provides a brief introduction on how temporal information and constraints can be important for researching these legal documents. Section 3 focuses on three types of legal documents and discusses in more detail how temporal information can be extracted from them. Section 4 concludes and discusses possible avenues of future research.

2 Legal documents and temporal information

Legal documents can be categorized in different ways. For this paper, we make the following distinction for different U.S. legal documents:

- Statutes (issued by the federal government)
- Proclamations, code of Federal Regulations, administrative decisions (issued by the President, Executive Departments and administrative departments (e.g. National Labor Relations Board (NLRB))
- Case law (authorized by trial courts, appellate courts or the supreme courts)
- Transactional documents (written by lawyers)
- Documents used as evidence for a case
- News documents that mention parties or people relevant to a case

There are different ways of how to look at temporal information and legal documents. For one thing we can look at the documents and their creation date or the date when the law described by them takes effect. Legal documents can be ordered along a time line according to these dates. This ordering of documents could be called extrinsic temporal ordering.

Another ordering would be an intrinsic temporal ordering of the events described within the document and placing them onto a time line. This type of temporal extraction is clearly more sophisticated and requires deep NLP processing techniques.

Another way of processing temporal information derived from legal documents is the mining of information about the participating parties mentioned in the document. Based on the creation date, one can derive that a lawyer works for a particular company at that time. A different case may show the same lawyer working for a different company at a latter point in time. Other text types such as news messages about companies, law firms or lawyers may also give information about the current affiliation of the people mentioned in the text. This information could be used to update databases on companies, law firms or lawyers.

All these three dimensions of temporal extraction and reasoning can be found if we look at the normal life circle of a case. Traditionally, the search for precedent

cases is the centerpiece for the American legal system and most often the starting point for the legal researcher. Hence, it is absolutely essential to find precedent cases relevant to the current case that are also not superseded by decisions of a higher court made at a later date. Services such as Keycite^{TM} offer a legal researcher the tool to search the *history* and status of U.S. and state court cases and statutes. In order to ensure accuracy this information is annotated by editors a couple of hours after the decisions have become public.

Apart from this classic case of ordering legal cases according to a time line, there are other applications where the automatic temporal ordering of documents can become crucial for a legal researcher. In the following, we will look at three different kinds of legal text: legal narratives, statutes and transactional documents. We will discuss the different kind of temporal expressions these documents can contain and how a standard off-the-shelf temporal tagger performs on these different kinds of data.

3 Types of temporal information in legal documents

This section discusses three different types of legal documents in more detail. First, we discuss fact-based narratives in case law which are most similar to news messages, because they mention mainly actual events that are linked to temporal expressions. Second, we investigate what kind of temporal expressions can be found in statutes. They are concerned with normative legal concepts rather than with concrete events. Consequently, event types are described that are linked to a temporal expressions. We found a higher number of durations than that is normally the case in news messages. Third, we looked at transactional documents that are similar to the normative laws presented in statutes but also contain more concrete dates and events (e.g. of a purchase event).

3.1 Legal narratives in case law

Narrative language describing the facts of the case most often contains temporal expressions. At the beginning of a case the judge normally describes the facts and the reasoning that follows should be based on the relevant laws, statutes or regulations relevant to these facts.

(2) On November 12, 1998, Illinois State Police Trooper Daniel Gillette stopped defendant on Interstate Route 80 in La Salle County for driving 71 miles per hour in a zone with a posted speed limit of 65 miles per hour. Trooper Gillette radioed the police dispatcher that he was making the traffic stop.

Such narratives are very similar to news messages and off-the-shelf temporal tagger could extract temporal expressions reasonably well from this type of text. In addition research focusing on temporal information derived from narratives [8] could be leverages for deriving a formal representation of the chain of events. Having derived the temporal constraints on the event described in the case,

searches could be carried out that contain temporal constraints. A query such as "*Banana /s slip /before fall*" would return only cases where a *slipping* event occurred before an *falling* event. Note that this is a (temporal) relation between events and not sentences.

3.2 Temporal restrictions in statutes or regulations

Statutes and regulations contain several different types of temporal expressions. In contrast to the fact-based narratives one finds in case law, they often contain periods of time (e.g. *30 days*) or sets of times (e.g. *every year*). These two types of temporal expressions are used to add time constraints to event types rather than to an actual event, as this is the case in news messages or the facts sections of a case.

(3) ATTORNEY GENERAL OPTION TO ELECT TO APPLY NEW PROCEDURES.- In a case described in paragraph (1) in which an evidentiary hearing under section 236 or 242 and 242B of the Immigration and Nationality Act has not commenced as of the title III-A effective date, the Attorney General may elect to proceed under chapter 4 of title II of such Act (as amended by this subtitle). The Attorney General shall provide notice of such election to the alien involved **not later than 30 days** before the date any evidentiary hearing is commenced. If the Attorney General makes such election, the notice of hearing provided to the alien under section 235 or 242(a) of such Act shall be valid as if provided under section 239 of such Act (as amended by this subtitle) to confer jurisdiction on the immigration judge.

The anchor for the duration in (3) is found in the date an evidentiary hearing is commenced. It is important to note that the link between the temporal expression and this event is conditional. Only if such an evidentiary hearing exists does the 30-days restriction apply.

Statutes may also contain date expression. These can be linked to an actual event, as for an effective date (or termination date) in (e.g. (4)). But mostly, even these date expressions are linked to an event type as a temporal constraint, as in (5).

(4) Amendment by Pub. L. 99177 effective **Dec. 12, 1985**, and applicable with respect to fiscal years beginning after Sept. 30, 1985, but with subsec. (c) to expire **Sept. 30, 2002**, see section 275(a)(1), (b) of Pub. L. 99177, as amended, set out as an Effective and Termination Dates note under section 900 of Title 2, The Congress.

(5)) (. . .) is an alien who entered the United States on or before **December 31, 1990**, who filed an application for asylum on or before **December 31, 1991**, and who, a**t the time of filing such application**, was a national of the Soviet Union, Russia, any republic of the former Soviet Union, Latvia, Estonia, Lithuania, Poland, Czechoslovakia, Romania, Hungary, Bulgaria, Albania, East Germany, Yugoslavia, or any state of the former Yugoslavia;

In a preliminary study of the United State Code we investigated the performance of an off-the-shelf temporal tagger (i.e. TempEx by [9]) on a small test set drawn from the United States Code by hand-annotating this test set with respect to the links between temporal expressions and events or event types (i.e. `TLINK`).

First we ran the TempEx tagger and computed precision and recall for a randomly selected set of 26 statute sections extracted from the 8th United States Code on Aliens and Nationality. Of the 64 temporal expressions in the sampled sections, the temporal tagger identified 24. Of these four contained incorrect date attributions. Results on this test are shown in table 1 (a). Take into consideration that the Tempex tagger was written for news messages and that such a test can only be seen as a baseline for temporal taggers that are more fine-tuned for legal language in statutes or regulations.

	correct	occurrences	percent
Precision	20	24	83.33%
Recall	20	64	31.25%

count	DT	PT	FT	AE
raw	22	26	11	5
	59			5
total	64			

Table 1. (a) Tagging accuracy (b) Distribution of temporal expression types

Then we hand-annotated all temporal expressions in these 26 sections according to the subordinated link and temporal link between the temporal expression and the event (type). We defined the following categories:[1]

PT Period linked to event type
FT Set of times linked to event type
DT Date links to event type
AE Date linked to actual event

The results of our preliminary study of the distribution of different types of links between temporal expressions and event (types) can be found in table 1 (b). From the distribution of these different link types one can conclude that temporal expressions in statutes serve a different function than in news messages or in the facts sections of cases. Statutes define event types that can be restricted by temporal constraints. A set of people may be defined by their actions within a certain time frame in addition to other conditions that have to hold (e.g. (5)). Such conditional definitions do not occur that often in factive text.

Nevertheless, the TimeML specification allows for such a link via an `SLINK` [10]:

[1] We did not find any periods or sets of times linked to actual events (e.g. *John wrote the note within 2 minutes.*)

(6) On Dec. 2 Marcos promised to return to the negotiating table if the conflict zone was demilitarized.

```
<SLINK eventInstanceID="ei1" subordinatedEventInstance="ei2"
signalID="s1" relType="CONDITIONAL"/>
```

Important signals for conditional `SLINKs` are conjunctions *when* or *if*, as described in the TimeML annotation guide. Those signals, however, are not found in statutes. Instead these temporal expressions are often used within a modal context (cf. *The Attorney General shall provide notice of such election to the alien involved* ***not later than 30 days***).

Extracting these links can be useful for the shallow processing of statutes where conditions including temporal ones are extracted and a matching algorithm could filter those statutes or regulations relevant to a given case (e.g. *former citizen of East Germany entered the United States om November 11th, 1990 and filled an application for asylum 20 days after he entered the country* fulfills all conditions stated in (5)).

Another important temporal dimension one encounters with this type of document is the history of the statute. Arnold-Moore describes a system that keeps track of the amendments that were added to a statutes of regulation. This system is currently being used for legislations in Tasmania.[2]

3.3 Dates in transactional documents

Some of the most common documents handled by lawyers in their daily work include transactional documents. These include contracts, purchase or sales agreements, and others which represent some kind of legal transaction. These documents almost always contain time expressions important for the legal stature of the document. The most important of these is the execution date, or the date when the transaction takes effect. In addition transactional documents may also contain duration clauses. These, for example, may establish a time frame for one party to establish or meet some condition necessary to satisfy the contact. In practice, an attorney may want to search their document management system to find all contracts signed after a particular date. We developed a system to recognize these dates in transactional documents.

Dates in legal documents are typically expressed in a form containing the day, month, and year. Contractual documents are overly specific, often using complicated language to rule out any possible future cause a party may have to contest the document. Dates need to be fully defined in this sense and so a reader is never required to infer the year or month based on other evidence in the document. In addition to the many, often wordy, date expressions, transactional documents typically have a particular date format not found in most document collections. Because these agreements need to take effect the same day they are signed, the author often leaves the actual day of the month blank, to be filled in at the time of signing.

[2] `http://www.thelaw.tas.gov.au/index.w3p`

Common off the shelf date recognition systems tended to over-match areas of the document that might otherwise appear to be date information. Simple rule or regular expression based systems often misconstrued other information as dates. Common errors included plot numbers and acreage sizes in real-estate transactions or citations to civil code that are found in many of these documents. In addition, these systems were unable to cope with the cases where the day was left to be filled in at the time of signing. These are especially important as they usually represent the execution date and the paragraph containing them may also have other information pertaining to the timeliness of the contract.

We undertook a study to see if it was possible to build a system which could recognize dates in transactional documents. We received approximately 1000 documents of various types from a local law firm. We manually identified several date types that we wished to recognize. In addition to the standard American date form of MM/DD/YY there were many more verbose examples. Many involving ordinal values, which were often spelled out. Some of the examples can be seen here

```
January 1, 2001
15<sup>TH</sup> DAY OF JANUARY, A.D. 2002
January 15, 2002
15th day of January, 2002
January 31, 2000
the 24<sup>th</sup> day of January 2002
January ____, 2002
this _______ day of January, 2002
first (1st) day of June, 2002
this 25th day of August, 2002
```

Given the small number of date types and the relatively few variations, we decided to write a recursive descent parser to identify dates. We used the Antlr compiler toolkit [11] for the implementation. We first constructed a tokenizer, which only recognized a limited number of token types. The most important being numbers both cardinal and ordinal, months, and underlines. The grammar for the parser could then be very specific. It was written to recognize either fully specified dates (containing day, month, and year) or a partially specified date, which contained a blank to be filled in at the time of signing. In practice, the program would tokenize the document, and then scan through the token lists until it located a token that could begin a date production. At this point the recursive descent parsing mechanism would take over and attempt to recognize a date. If successful, a date object would be created and stored with the document as searchable meta-data in the law firms document management system.

We compared the output of our parser to the output of the same off-the-shelf temporal tagger as before [9]. The test collection consisted of 6 documents and contained 20 date references. The time tagging system was able to identify 13 dates spread across the bodies of the documents but could not correctly identify the 6 partial dates (i.e. those with a blank for the date to be filled in when

the document was signed.) In addition there were three false positives where addresses were considered as years by the off-the-shelf tagger. The complete results are in table 2. Because our parser only looks for fully specified dates it does not confuse other numbers as parts of dates. In addition the off-the-shelf tagger requires a separate pass to tag parts of speech before processing input. To its credit the system did recognize date ranges and other indicators which could be useful in the analysis of transactional documents. Our tagger could not do this.

	correct	occurrences	percent
Precision	13	16	81.25%
Recall	13	20	65.00%

Table 2. Off the shelf tagger on transactional data

4 Conclusions

This paper reports on work-in-progress on temporal information extraction techniques to legal documents. More specifically, we focused on three types of legal documents and discussed the applicability of temporal taggers to these different types of documents.

- Legal narratives in case law are similar to news messages and off-the-shelf temporal taggers should provide a good coverages with respect to extracting temporal expressions. In addition, the narrative structure should give additional clues for ordering the events of the current case. Applications that could benefit from a temporal extraction techniques are more detailed searches with temporal connectors or temporal reasoning of witness accounts in order to detect inconsistencies among the witnesses' statements.
- Statutes or regulations have a different languages and differ in many respect from other legal texts by providing legal rules that should match the facts of the current case. This is also reflected in the temporal information encoded into these rules. In a preliminary study, we found a large amount of temporal expressions that are linked to event types rather than actual event. A temporal and event tagger has to take this into account when applied to this kind of data. Consequently, an off-the-shelf temporal tagger we used had a very low recall. Future applications could use the temporal constraints mentioned in the statutes and match them against the actual case and suggest relevant passages.
- Transactional documents describe legal rules as well as actual dates. In addition, many numbers mentioned in the document could be confused by dates. We also found underspecified temporal expressions with the day information

left open in these documents. A temporal tagger tuned to this kind of data was able to deal with these special requirements sufficiently.

References

1. Vila, L., Yoshino, H.: Time in automated legal reasoning. Information and Communications Technology Law **7** (1998) 173–197
2. Brian Knight, J.M., Nissan, E.: Representing temporal knowledge in legal discourse. Law, Computers, and Artificial Intelligence / Information and Communications Technology Law **7** (1998) 199–211
3. Farook, D.Y., Nissan, E.: Temporal structure and enablement representation for mutual wills:. Law, Computers, and Artificial Intelligence / Information and Communications Technology Law **7** (1998) 243–268
4. Arnold-Moore, T.: About time: legislation's forgotten dimension. In: Proceedings of the 3rd AustLII Law via the Internet Conference 2001, Sydney, Australia (2001)
5. Arnold-Moore, T.: Point in time publication for legislation (xml and legislation). In: Proceedings ot the 6th Conference on Computerisation of Law via the Internet, Paris, France (2004)
6. Grandi, F., Mandreoli, F., Tiberio, P., Bergonzini, M.: A temporal data model and system architecture for the management of normative texts (extended abstract). In: Proceedings of SEBD 2003 - Natl' Conf. on Advanced Database Systems, Cetraro, Italy (2003) 169–178
7. Kowalski, R., Sergot, M.: A logic-based calculus of events. New Gen. Comput. **4** (1986) 67–95
8. Mani, I., Pustejovsky, J.: Temporal discourse models for narrative structure. In Webber, B., Byron, D.K., eds.: Proceedings of the ACL 2004 Workshop on Discourse Annotation, Barcelona, Spain, Association for Computational Linguistics (2004) 57–64
9. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000), Hong Kong (2000) 69–76
10. Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., Mani, I.: The specification language TimeML. In Mani, I., Pustejovsky, J., Gaizauskas, R., eds.: The Language of Time: A Reader. Oxford University Press, Oxford (2005)
11. Parr, T.J., Quong, R.W.: Antlr: A predicated- *ll(k)* parser generator. Softw., Pract. Exper. **25** (1995) 789–810