Temporal Dependence in Legal Documents*

Daniel Isemann, Khurshid Ahmad, Tim Fernando, and Carl Vogel

School of Computer Science and Statistics, Trinity College Dublin, Ireland
{isemandi,kahmad,tfernand,vogel}@tcd.ie

Abstract. Tasks and difficulties inherent in the largely open problem of temporal information extraction from legal text are outlined. We demonstrate the efficacy of tools and concepts available "off-the-shelf" and suggest refinements for such applications. In particular, the frequent references between regulatory texts have to be addressed as a separate named entity recognition task that bears relevance to an analysis of the temporal ordering of legislation. A regular expression-based approach as a robust first step towards addressing this problem is tested.

Keywords: named entities, temporality, legal information extraction.

1 Introduction

Information on time and events, their ordering and causal dependencies, is critical in the legal domain and in particular in financial industry regulation. In a given regulatory framework certain requirements may have to be met by a certain deadline or particular disclosures can only be made within a certain time after an event in order to be in compliance. Such time dependence has motivated researchers in compliance information systems to experiment with and recommend the use of temporal logics [1–4].

Documents in a given regulatory domain, say capital adequacy or anti-money laundering, contain the names of key objects and events. It is not difficult for human readers to extract information about the duration of an event or the changing nature of legislation, often by analysing the tense information in verbs. A handle on dates of origination of legal documents gives a temporal sequence from earliest to latest laws which may indicate time-spans of a law's validity.

For temporal analysis we rely on in-text evidence with a view to automating the extraction of relevant information. We present an analysis of whether existing NLP systems can deal with legal and regulatory framework mnemonics which contain dates and references to jurisdictions, and following work elsewhere [5], we find regular expressions well-suited. In §2 we will make the problem statement more precise and review relevant research contributions. In §3 we report on existing off-the shelf tools and characterize their strengths and limitations for the given tasks in our domain. In §4 we motivate an alternative approach to extracting and managing the temporal interconnectedness of legislation and outline a first implementation which we have undertaken. We conclude in §5.

* We gratefully acknowledge Enterprise Ireland's *Governance Risk and Compliance Technology Centre* Initial Grant CC-2011-2601-B.

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 497–504, 2013.

2 Named Entity and Date Extraction in Legal Text

Legal draughting introduces named entities in legal texts in a variety of ways. Next to references to people, organisations, places, events and objects, conventional legal text contains references to other laws which may either be in force, have been repealed or may be forthcoming. References to other items of legislation and to judicial (court judgements) and executive documents (enforcement documents, regulations based on legislation), rely on a subtle form of abbreviation which forms mnemonics for existing laws together with dates of their origination, e.g. *UK MLR 2007* or *EU Directive 2005/29/EC*.

Named entity recognition (NER) is often taken to be the starting point of information extraction and a first step in relation detection. Similarly, event detection has been suggested as a first step in temporal analysis. Standard definitions of these tasks, however, reveal how much default assumptions are shaped by the prototypical domains of newswire text and news reportage: "Generic NER systems tend to focus on finding the names of people, places and organizations that are mentioned in ordinary news texts" [6, p. 760]. Typical subtasks of temporal analysis are "Fixing the temporal expression with respect to an anchoring date or time, typically the dateline of the story in the case of news stories; [...] Arranging the events into a complete and coherent timeline." [ibid., p. 761]. While we believe that temporal information plays an important role in legislative documents, these documents, unlike news text do not exhibit a narrative structure. In most cases there is no "story" and no "complete and coherent timeline". Instead, as we will argue in §3, we believe that there are several interacting timelines which need to be taken into account. As relationships between entities mentioned in legal documents, may involve references to time, we take named entity recognition as a precursor to temporal analysis.

Several researchers have addressed the task of named entity recognition in legal texts. Dozier and colleagues have applied a hybrid method of controlled vocabulary lookup, contextual rules and statistical models to a test set of 600 US trial documents to identify named entities such as jurisdiction, court, document title, document type and name of the judge [7]. They report high precision for identifying judges (98%) and high recall for capturing jurisdictions (87%). Quaresma and Gonçalves use linguistic information to identify named entities in a corpus of legal documents from the International Agreements/External Relations section of the EUR-Lex portal[1] and classify documents based on some of the named entities discovered [8]. Unlike Dozier et al., Quaresma and Gonçalves found only very few person names in their corpus (presumably due to nature of international agreements). For the NER part the authors focus on the categories of location and organization names, dates and references to documents and document articles. We are not aware of work which systematically addresses the identification of statutes and relationships between legal documents and specific dates.

[1] `eur-lex.europa.eu/en/index.htm` (last accessed 20/07/2013).

3 Existing Linguistic Mark-Up Tools

We present an analysis of how the extraction of time and event related information may be supported by more traditional attempts at automated text analysis (such as tagging and named entity recognition) in making sense of legal documents. We have experimented on a sample text of UK anti-money laundering (AML) regulatory guidelines. For this we hand-annotated the occurrence of named entities in a small sample of text from a regulatory authority which we then subjected to automated analysis for named entity recognition and temporal analysis in this domain with readily available domain independent natural language processing tools. We outline limits of such non-customized approaches. We use a qualitative rather than quantitative analysis to highlight where standard methods fall short and in turn motivate our own approach (§4). This is in the spirit of recall analysis of precision grammars applied to large corpora [9].

We hand-annotated named entities in two paragraphs of an HM Revenues & Customs document on anti-money laundering guidelines for money service businesses [10] and applied the Stanford University Tagger, the Stanford CoreNLP NER module[2] and the TARSQI toolkit for temporal mark-up[3] to this text.

3.1 Off-the-Shelf Named Entity Recognition in the GRC Domain

Studying this text segment we identified four different types of named entities. Three of these are well known candidates from generic named entity recognition tasks: date, location and organization. The fourth category consists of references to other legislative and regulatory documents.[4] We believe that the last category, references to other regulatory documents is particularly important in our domain and that it can be leveraged for a temporal analysis as well. The consistent identification of this category is largely an open problem. While Dozier et al. mention the need to identify such references[5] they do not attempt it in their analysis [7]. Quaresma and Gonçalves, on the other hand, emphasize the importance of this category: "we can state that these documents have a high number of references to other documents and articles [...] this kind of information [...] would allow the inference of important relations, such as, the chain of legislation references." [8, p. 55]. However, they report an error rate of 65% for their effort to automatically identify such references from an analysis of parse trees [ibid.].

The following sentence taken from our sample illustrates three of the four different types of named entities we distinguished:

> *In addition, under the Money Laundering Regulations 2007*^{REF}*, which came into force on 15 December 2007*^{DATE}*, HMRC*^{ORG} *will have powers to cancel the registration of Money Transmission Businesses*^{ORG} *where they are found to be consistently non-compliant with the Payments Regulation*^{REF}*.*

[2] Available at `nlp.stanford.edu:8080/corenlp` (last accessed 20/07/2013).

[3] Available at `www.timeml.org/site/tarsqi/toolkit/index.html` (last accessed 20/07/2013).

[4] Note that our categories are identical to the ones studied by Quaresma et al. [8].

[5] We take their example of identifying "statues [sic]" [7, p. 28] to be intended to refer to statutes.

Table 1 gives an overview of the four categories, their definition and how often they occurred in our exploratory text sample. There were 22 named entities.

Consider a set of five sentences taken form this document together with the tags attributed to the words and sometimes phrases by the Stanford Tagger and NER module (Table 2). We have only shown tags that describe location names, organization names, dates, (other) named entities and proper nouns. Phrases tagged with these tags have been underlined. Note that in some instances the tagger fails to recognize a proper noun, like 'Transfer Regulation' in Payments/Wire Transfer Regulation (sentence #1) but in other cases the NER tags a complex phrase, say Money Laundering Regulations (sentence #3), accurately as a named entity. The NER module recognizes the EC and Her Majesty's Revenue and Customs (HMRC) as an organization. The recognition of dates and organization names requires improvement insofar as dates and organisations that are part of legal or regulatory references are recognized as separate entities.

Using our hand-annotated standard as a reference we computed the Stanford system's precision (52%), recall (50%) and F1 score (51%). While the achieved F1 score still leaves room for improvement, it has to be recalled that named entity recognition is known to be a highly domain specific task. Common types of named entities, such as dates, can be identified very reliably. Domain specific categories, such as references to laws and regulations, however, (our REF category, Table 1), cannot be identified by off-the-shelf tools. Generic entity recognition therefore has to be supplemented with algorithms specific to the legal domain.

From our analysis we noted that references to regulatory documents or parts thereof, which frequently occur in legal text, are partially identified by existing parsers as proper nouns. However, there are two drawbacks of a parser-based analysis: First, fine-tuning is necessary to correctly categorize the type of such entities as references and to adequately delineate these references as they often include dates and organization names. Second, parsers and NER systems usually make no attempt to link dates and other temporal information to mentions of events or to embed these into a wider temporal context. In order to address this we have turned to TimeML and the associated Tarsqi system.

Table 1. The categories of named entities that can be found in regulatory documents and that are targeted through text analysis

Category	Definition	Number
DATE	Descriptions of dates, including days, months, years or combinations thereof in text.	3
LOC	Geographical location names.	1
ORG	Organisation names.	3
REF	References to legislative or regulatory documents including subsections or appendices of documents.	15

Table 2. Sentences from a UK HMRC document on anti-money laundering [10] showing selected entities that are recognized by the Stanford Tagger and NER application

1. The EC^{ORG} Payments/Wire$^{PROPER\ NOUN}$ Transfer Regulation came into effect on the 1st January 2007^{DATE}.
2. The UK^{LOC}'s supervision and enforcement provisions are set out in the Transfer of Funds (Information on the Payer) Regulations$^{PROPER\ NOUN}$ 2007^{DATE}.
3. In addition, under the Money Laundering Regulations$^{OTHER\ ENTITY}$ 2007^{DATE}, which came into force on 15 December 2007^{DATE}, HMRC^{ORG} will have powers to cancel the registration of Money Transmission$^{OTHER\ ENTITY}$ Businesses where they are found to be consistently non-compliant with the Payments Regulation$^{OTHER\ ENTITY}$.
4. For more information about HMRC^{ORG}'s powers to cancel registrations, please refer to MLR9 Registration notice.
5. Both the Money Laundering Regulations$^{PROPER\ NOUN}$ 2007^{DATE} and EC^{ORG} Regulation$^{PROPER\ NOUN}$ 1781/2006 on information on the payer accompanying transfers of funds (Payments Regulation/Wire Transfer Regulation$^{OTHER\ ENTITY}$) require verification of customers' identities 'on the basis of documents, data or information obtained from a reliable and independent source'.

3.2 Off-the-Shelf Temporal Mark-Up in the GRC Domain

Document mark-up has recently progressed from marking up typographical and display related information to marking up time and event related information in text. One such standard is *TimeML* [11]. There are several programs to identify such information automatically through the use of natural language processing technology [12, 13]. As a representative of such programs we have chosen the *TARSQI* toolkit, an experimental mark-up system for TimeML available under a free of charge license [12]. We have used TimeML and TARSQI for a qualitative analysis of the incidence of time related information in GRC documents.

We ran TARSQI on our sample text [10] and studied the temporal links introduced by the system. In total, TARSQI suggested 16 links between events and/or temporal anchors in the text. Twelve of these were temporal links, comprising eight BEFORE and four SIMULTANEOUS links.[6] Insofar as TARSQI identified dates correctly, it had no difficulty temporally ordering them. Some of the system guesses were inaccurate, however. In "EC Regulation 1781/2006", TARSQI identified "1781" as a year. A different picture emerged for the temporal ordering of events. Several of the links that TARSQI identified between postulated events appeared "broken" due to an incommensurability of the timelines involved.

In linguistics literature a distinction is sometimes made between *episodic* and *generic* statements or clauses [14]. We believe this distinction to be of importance in the context of temporal links. An episodic statement is one that describes one or more actual, concrete events. Generic statements on the other hand describe general truths, laws, rules or expectations. An example of an episodic statement from our text on anti-money laundering is: "The EC Payments/Wire Transfer

[6] Note that the TARSQI toolkit attempts to guess temporal as well as non-temporal TimeML links, such as modal or counter-factive subordination or aspectual links.

Regulation came into effect on the 1st January 2007." An instance of a generic statement from the same document is: "Where there is a higher risk of false identity documents or information, there may be a need to obtain additional evidence of identity." These two kinds of statements may have to be treated separately as far as temporal links (e.g. BEFORE, SIMULTANEOUS etc.) are concerned because links crossing from the episodic to the generic realm and vice versa are likely to span different, not directly comparable notions of time and timelines. In our sample paragraphs we found that one in five sentences was of episodic, the others of generic nature. Further research is needed to quantify which ratio of generic to episodic clauses is typical of regulatory documents and to establish which textual markers (e.g. mentions of concrete dates) can be used to algorithmically distinguish between the two kinds of statements.

4 Temporal Analysis of Legislative Documents

The importance of temporal information in legal texts is underlined by the advanced search facility of the EUR-Lex portal which offers no less than 14 relations between a legal document and a date for selection, including: *Date_of_publication*, *Date_of_effect*, *End_of_validity_date*, *Date_of_signature*, *Date_of_debate*, etc.[7] Relationships between documents may also have temporal implications. Most of the 17 relationships between documents listed by the EUR-Lex advanced search cover some sort of amendment or legal effect that presupposes a before and after relation between legal documents and sometimes may affect the validity of legislation. We analysed 70 EU directives from the Wolters Kluwer Compliance Resource network[8] for dates and other directives mentioned. Using regular expressions we robustly identified the occurrences of dates and references to other EU directives. On a sample of seven randomly selected EU directives, our directive pattern, which had perfect precision, achieved 86.4% recall.

With such patterns we analyzed the EU directives for cross-references and dates and made an attempt at qualifying references as legal amendments (Figure 1). We found that references to other directives though widespread throughout the documents are typically focussed on a small number of key directives per document. While on average a EU directive references 16.1 other directives (SD = 9.7), it is usually less than four directives (3.6, SD = 2.7) that account for more than half of the individual reference count in a given document.

Studying the cross-references we found that some directives receive many more incoming links than others, perhaps indicating that these are of foundational character.[9] Qualifying a reference as legally amending another one, one can begin to create timelines of laws that wholly or in part amend each other as a first step towards establishing time-spans of legislation validity.

[7] `eur-lex.europa.eu/expert/sg/sga_cnct/celexexp!dev?LANG=EN&BASE=bas-cen` (last accessed 20/07/2013). One of the 14 relations is a super-category (*All_dates*).

[8] `www.complianceresourcenetwork.com/web/crn` (last accessed 20/07/2013).

[9] See `www.scss.tcd.ie/~kahmad/IDEAL2013/` for the data set, a sample cross-reference graph and the patterns discussed here.

Directive ID	Directive Title	Publication Date	Amended Directives	Referenced Directives	Dates mentioned
Directive 2004/39/EC	Oganisational requirements and operating conditions for investment firms - Draft implementing Directive 2004/39/EC	Fri Jun 02 00:00:00 IST 2006	[]	[Directive 85/611/EEC x 8, Directive 2003/125/EC x 6, Directive 2003/6/EC x 5, Directive 2000/12/EC x 4, Directive 2003/71/EC x 4, Directive 93/22/EEC x 1, Directive 2005/29/EC x 1, Directive 2002/65/EC x 1, Directive 95/46/EC x 1, Directive 84/450/EEC x 1]	[Sat Nov 01 00:00:00 GMT 2008 x 1, Tue Nov 04 00:00:00 GMT 2003 x 1, Mon Dec 22 00:00:00 GMT 2003 x 1, Wed Apr 21 00:00:00 IST 2004 x 1, Wed Jan 31 00:00:00 GMT 2007 x 1, Thu Nov 01 00:00:00 GMT 2007 x 1, Fri Oct 31 00:00:00 GMT 2008 x 1, Fri Dec 20 00:00:00 GMT 1985 x 1]
Directive 90/211/EEC	Mutual recognition of public-offer prospectuses as stock-exchange listing particulars Directive 90/211/EEC	Mon Apr 23 00:00:00 IST 1990	[Directive 80/390/EEC x 1]	[Directive 80/390/EEC x 4, Directive 89/298/EEC x 2, Directive 87/345/EEC x 1]	[Mon Apr 23 00:00:00 IST 1990 x 2, Wed Apr 17 00:00:00 IST 1991 x 1]

Fig. 1. Pattern-based information extraction from EU Directives: Example output

5 Concluding Remarks

The analysis with tailored regular expressions for statutes and relations among dates and statutes as named entities in §4, along with cross-reference assessment involves computationally simpler technology than the more sophisticated tools considered in §3.1 and §3.2, but the methods are well suited to the purpose.

We outlined some of the tasks and difficulties inherent in the largely open problem of temporal information extraction from legal text. Two important aspects of documents in this domain were identified. First, the frequent references from one regulatory text to another have to be addressed as a separate named entity recognition task. We have motivated the relevance of this task to an analysis of the temporal ordering of legislation. A pattern-based approach provides a robust first step towards a solution. Second, in terms of the identification and interpretation of temporal links within a given regulatory text, a distinction between episodic and generic statements has to be made and care has to be taken that temporal links do not span this divide.

Several lines of future work arise from our preliminary study. First, we plan to combine our pattern-based approach for reference and link extraction with suitable machine learning techniques, hopefully enhancing and generalizing our extraction results. Second, we plan to build a classifier for telling episodic from generic statements, a useful first step for analyzing in-document temporal relations, as we have argued. Further exploring the linkages between legal documents one may create a page-rank hierarchy on a corpus of such documents, similar to those known from information retrieval. Such a hierarchy can be embellished further by analyzing the distribution of the frequency of keywords in each of the documents in the corpus and we expect semantic annotation frameworks such as RDF to be useful in this context. This approach may benefit professional providers of legal content, such as Wolters Kluwer or Thomson Reuters. Finally, the fact that in the context of European legislation, each EU member state may have a different way of interpreting a directive or regulation may necessitate a multi-lingual analysis that has to deal with local exemptions and derogations.

References

1. Mylopoulos, J., Borgida, A., Jarke, M., Koubarakis, M.: Telos: Representing knowledge about information systems. ACM Transactions on Information Systems (TOIS) 8(4), 325–362 (1990)
2. Dardenne, A., Van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. Science of Computer Programming 20(1), 3–50 (1993)
3. Fuxman, A., Liu, L., Mylopoulos, J., Pistore, M., Roveri, M., Traverso, P.: Specifying and analyzing early requirements in Tropos. Requirements Engineering 9(2), 132–150 (2004)
4. COMPAS-Project: Deliverable D2.2: Initial specification of compliance language constructs and operators, Version 2.0 (2009), `www.compas-ict.eu/compas_results/deliverables/m11/D2.2_Initial-specification-of-compliance-language-constructs-and-operators.pdf` (last accessed July 20, 2013) (last accessed July 20, 2013)
5. Breaux, T.D., Gordon, D.G.: Regulatory Requirements Traceability and Analysis Using Semi-formal Specifications. In: Doerr, J., Opdahl, A.L. (eds.) REFSQ 2013. LNCS, vol. 7830, pp. 141–157. Springer, Heidelberg (2013)
6. Jurafsky, D., Martin, J.H.: Speech and Language Processing, 2nd edn. Prentice Hall Series in Artificial Intelligence. Pearson Education International, New Jersey (2009)
7. Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., Wudali, R.: Named Entity Recognition and Resolution in Legal Text. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 27–43. Springer, Heidelberg (2010)
8. Quaresma, P., Gonçalves, T.: Using linguistic information and machine learning techniques to identify entities from juridical documents. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 44–59. Springer, Heidelberg (2010)
9. Baldwin, T., Beavers, J., Bender, E., Flickinger, D., Kim, A., Oepen, S.: Beauty and the Beast: What Running a Broad-Coverage Precision Grammar over the BNC Taught Us about the Grammar – and the Corpus. In: Kepsar, S., Reis, M. (eds.) Linguistic Evidence: Empirical, Theoretical and Computational Perspectives. Studies in Generative Grammar, vol. 85, pp. 49–69. Mouton De Gruyter (2005)
10. HM Revenue & Customs: Anti-money laundering guidance for money service businesses, We used paragraphs 10.5.6 and 10.6.1 for the analysis described in this paper (2010), `www.hmrc.gov.uk/mlr/mlr_msb.pdf` (last accessed July 20, 2013)
11. Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Fifth International Workshop on Computational Semantics (IWCS-5) (2003)
12. Verhagen, M., Mani, I., Sauri, R., Knippen, R., Jang, S.B., Littman, J., Rumshisky, A., Phillips, J., Pustejovsky, J.: Automating Temporal Annotation with TARSQI. Demo Session. In: Proceedings of the ACL (2005)
13. Llorens, H., Saquete, E., Navarro, B.: Tipsem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 284–291 (2010)
14. Carlson, G., Pelletier, F.: The Generic Book. University of Chicago Press (1995)