LawORDate: a Service for Distinguishing Legal References from Temporal Expressions

María Navas-Loro[1]

Ontology Engineering Group
Universidad Politécnica de Madrid, Madrid, España,
mnavas@fi.upm.es,
http://marianavas.oeg-upm.net/

Abstract. References to documents in the legal domain usually follow patterns containing temporal information in different forms (e.g. 'Directive 2001/29'). These references mislead algorithms detecting pure temporal references, and false positives occur in named entity recognition algorithms searching dates or intervals. This paper presents methods and techniques to identify these references, applied to two different domains. The first domain is that of news, where the temporal information plays a crucial role for their understanding and automatically building timelines can be hampered by the errors induced from these legal references. The second domain is that dataset descriptions. Dataset descriptions sometimes contain temporal information, not only in their dedicated metadata fields (e.g. dataset creation) but also within the text of their description. LawORDate, the system presented in this paper, is a web service able to detect legal references with temporal information in Spanish texts. The service identifies these references, avoiding their annotation by temporal taggers and enabling a further step of linking the references to the original sources and building co-reference graphs.

Keywords: legal references, temporal expressions, news, dataset description

1 Introduction

Temporal expressions detection, mainly focused on news, is a emerging field gaining more and more importance in NLP. Efforts such as the NewsReader project[1] and the TempEval [1, 2] initiatives in SemEval, along with subsequent more specific temporal tasks [3, 4] show the interest in processing the temporal dimension on all kind of texts. Usually processing of temporal expressions is done regarding the concrete type of text being faced, both depending on its field (such as news, clinical domain or historical texts) or extension (free texts or length-limited tweets). Due to this specialization, systems do not usually react well when they find expressions from other fields, such as is the case of legal references in news or dataset description.

[1] http://www.newsreader-project.eu/results/data/wikinews/

The boom of open data portals also present this kind of mixed information. Thousands of datasets become publicly available everyday, sometimes presenting just basic scarce metadata such as title and description. Being able to extract additional information and new search parameters from them, such as named entities or temporal references, would facilitate managing them, along with linking them resources or queries.

To this end, a system[2] was built to extract temporal coverage from both news and related datasets in Spanish, some of them in the legal domain, and be able to link them based in the temporal dimension. This system calls an existing temporal tagger, HeidelTime [5], able to detect temporal expressions in texts in Spanish and tag them following the TIMEX3 annotation standard. Nevertheless, this tagger happened to tag as temporal expressions references to Spanish laws and legal documents that led to false positives, such as shown in the example exposed in Fig. 1, extracted from a real article[3]. The result of the tagging by HeidelTime can be found in Fig. 2.

Estas actividades están reguladas por **Real Decreto 1341/2007, de 11 de octubre** *sobre la gestión de la calidad de las aguas de baño, incorporando al derecho español* **la Directiva 2006/7/CE del Parlamento Europeo y del Consejo de 15 de febrero de 2006** *relativa a la gestión de la calidad de las aguas de baño.*

Fig. 1. For English: 'These activities are regulated by ***Royal Decree*** **1341/2007, of 11th October** on the management of bathing water quality, incorporating into **Spanish law Directive 2006/7/ EC of the European Parliament and of the Council of 15th February 2006** on to the management of the quality of bathing waters.'

Estas actividades están reguladas por **Real Decreto** <TIMEX3 tid=”t2” type= ”DATE” value=”1341”>**1341**</TIMEX3>/<TIMEX3 tid=”t3” type=”DATE” value=”2007”>**2007**</TIMEX3>, <TIMEX3 tid=”t9” type=”DATE” value= ”2016-10-11”>**de 11 de octubre**</TIMEX3> sobre la gestión de la calidad de las aguas de baño, incorporando al derecho español **la Directiva** <TIMEX3 tid=”t4” type=”DATE” value=”2006”>**2006**</TIMEX3>**/7/CE del Parlamento Europeo y del Consejo** <TIMEX3 tid=”t8” type=”DATE” value=”2006-02-15”>**de 15 de febrero de 2006**</TIMEX3> relativa a la gestión de la calidad de las aguas de baño.

Fig. 2. In blue, result of HeidelTime tagging on the text in Fig. 1.

We also find this problem in the description of datasets, being specially problematic when obtaining obviously inconsistent dates such as happens in the ex-

[2] https://github.com/mnavasloro/AportaCuando

[3] http://www.castillalamancha.es/actualidad/notasdeprensa/castilla-la-mancha-cuenta-con-35-zonas-de-ba%C3%B1o-autorizadas-donde-disfrutar-de-la-naturaleza

ample in Fig.3, extracted from the description of a real dataset[4]. Here the tagged dates without a legal-focused preprocessing were *'2093'*, *'2008'* and *'2008-12-19'*. While the latest can at least be used as a lower temporal bound (since there is no additional temporal information on the coverage in the description), the year 2093 is obviously inconsistent.

Base de datos que proporciona información sobre los Centros Tecnológicos y Centros de apoyo a la Innovacin inscritos en el registro creado mediante ***el Real Decreto 2093/2008, de 19 de diciembre****. Permite la consulta por Modalidad, rea Tecnológica, Sector, Comunidad Autónoma y/o Provincia. Además, posibilita la descarga de la versión completa en PDF.*

Fig. 3. For English: 'Database that provides information on Technology Centers and Innovation Support Centers registered in the registry created by ***the Royal Decree*** **2093/2008, of December 19**. It allows consultation by Modality, Technological Area, Sector, Autonomous Community and/or Province. In addition, it allows to download the full version in PDF.'

The aim of the web service LawORDate[5] introduced in this paper is to detect common legal expressions appearing in non-legal texts that tend to mislead temporal taggers and replace them in the text, in order to obtain a clean version of it where temporal taggers are able to detect just temporal expressions. The remainder will expose a brief state-of-the-art and an analysis on usual legal expressions with patterns similar to temporal expressions in Spanish, along with examples of regular expressions able to detect most of them (tested in a case of use on descriptions of datasets from the Spanish Open Data portal). Finally, conclusions derived from this analysis and future work on this topic will be exposed.

2 State of the Art

Processing the temporal dimension of legal text has been previously tackled in literature [6–8], and the confusion between legal and temporal references has been previously exposed [9]. Nevertheless, to the best of her knowledge, the author is not aware to any previous dedicated approach to detect legal references specifically for ulterior temporal processing.

Identification of legal cross-references has been widely studied in literature [10], being targeted in different languages (such as French [11, 12], Dutch [13], Italian [14] or Japanese [15]) and with different levels of deepness. We find for instance the approach of Adedjouma et al. [11] for the Luxembourg's Legislation (later expanded to a Canadian legal corpus [12]), where a complete schema

[4] http://datos.gob.es/catalogo/e04990501-registro-de-centros-tecnologicos-y-centros-de-apoyo-a-la-innovacion-tecnologica

[5] https://github.com/mnavasloro/LawORDate (with information on how to use the web service)

identifying different parts that can be included in a reference in this context (such as *Part, Book* or *Article*), along with the different information in them (*dates, names, headers...*) are built. The authors also make a distinction between simple and complex cross reference patterns; this had been previously exposed also in [13], including also some special cases, where a grammar allowed identification of just in-collection legal references in documents from the Dutch Tax and Customs Administration. Finally, the work by Tran et al. [15] focused on references to sub-document targets, proposing machine-learning based approaches. Also more generic-aimed frameworks for managing legal documents, such as NORMA-system [16], include services for marking-up legal references, called by further works [17].

Differently from the temporal aim presented in this paper, the use of this techniques for legal references identification go from mark-up and linking [16] to normalization [14]. Most of these approaches are based in patterns; the only work in Spanish the author is aware of also follows this pattern-based approach [18].

3 Analysis of the problem

In the frame of news and dataset description processing, namely trying to locate them into a temporal instant or interval, several legal references happened to be tagged as temporal expressions by a state-of-the-art temporal tagger. Some examples are the following expressions, that refer to different official Spanish documents or laws:

- Ley Orgánica 10/1995 (*Organic Law).*
- Ley 22/2011, de 28 de julio (*Law).*
- BOE: 29/07/2011 or BOE de 22 de julio or BOE núm. 306, de 23 de diciembre (BOE: Boletín Oficial del Estado *Official State Gazette*).
- Real Decreto 1341/2007 (sometimes also expressed as RD 1463/2007, *Royal Decree*)
- Directiva 2012/27/UE.

These references are often also surrounded by a date referred to their creation (being therefore important to detect them as well). These legal expressions can also include additional words such as in 'Real Decreto Legislativo' (*Legislative Royal Decree*) or be combined such as in '*Real Decreto Legislativo 1/2004 de 5 de enero BOE de 8 de marzo*'. Also exceptions where dates near to references to legal documents can be found, such as happens when the dataset contains information about the proper legal document, such as in the example[6] depicted below, where the dates refer indeed to temporal coverage:

[6] http://datos.gob.es/catalogo/l01280148-publicaciones-en-boletin-oficial-del-estado-boe-2013-2017

Publicaciones en Boletn Oficial del Estado (BOE): 2013-2017. (for English: *'Publications in the Official State Gazette (BOE): 2013-2017.'*)

The problem of detecting these references is therefore not straightforward. In the following section some patterns for references found in a concrete application case (dataset descriptions and news) are introduced.

4 Hands on and first patterns

The corpus we have worked with is a dump of metadata from the Spanish open data portal datos.gob.es[7], consisting of almost 16k datasets. Some of them contained temporal coverage information expressed as **dcat:temporal property**[8], but most of them had their upload and creation date as only temporal information, along with information on the publisher, the title and the description.

A first analysis performed on metadata from these datasets showed that most appearances followed constrained patterns, and that texts that presented this kind of references misled the temporal tagger. Besides detecting temporal expressions that are not actually from the text timeline, another major problem derives from this misleading: temporal normalization[9] is also affected, since some dates can be wrongly normalized because of the misidentification of legal references nearby as temporal expressions.

Some of the used patterns for detecting these problematic legal references are the exposed in Fig. 4.

```
(([D|d](irectiva|IRECTIVA)) (\d*)\/(\d*)\/(\w*))(,? de (\d*) de ([E|e]nero|[F|f]ebrero|[M|m]
arzo|[A|a]bril|[M|m]ayo|[J|j]unio|[J|j]ulio|[A|a]gosto|[S|s]emptiembre|[O|o]ctubre|[N|n]ovie
mbre|[D|d]iciembre)( de (\d\d\d\d))?)?

(([R|r](eal|EAL) [D|d](ecreto|ECRETO)) (\d*)\/(\d*))(,? de (\d*) de ([E|e]nero|[F|f]ebrero|[
M|m]arzo|[A|a]bril|[M|m]ayo|[J|j]unio|[J|j]ulio|[A|a]gosto|[S|s]emptiembre|[O|o]ctubre|[N|n]
oviembre|[D|d]iciembre)( de (\d\d\d\d))?)?

((([L|l]ey [O|o]rg[|a]nica)|(LEY ORG[|A]NICA)) (\d*)\/(\d*))(,? de (\d*) de ([E|e]nero|[F|f]
ebrero|[M|m]arzo|[A|a]bril|[M|m]ayo|[J|j]unio|[J|j]ulio|[A|a]gosto|[S|s]emptiembre|[O|o]ctub
re|[N|n]oviembre|[D|d]iciembre)( de (\d\d\d\d))?)?
```

Fig. 4. Some patterns for detecting references to Spanish legal documents, as well as surrounding dates referring to them.

[7] http://datos.gob.es/

[8] https://www.w3.org/TR/vocab-dcat/#Property:dataset_temporal

[9] Temporal normalization can be described as "to assign the same value to all expressions carrying the same semantics or referring to the same point in time" [5]; this is, the temporal anchoring (often derived from context) for an incomplete temporal expression. An example can be how from the sentence *'The 4th of October of 1991 he came here. The 6th he left.'*, the date *'06/10/1991'* can be derived for *'6th'*.

Once these patterns are detected, they are replaced in the text by strings containing information of the legal references detected, but in a format that does not mislead the temporal tagger. This new version of the text maintains all the original genuine temporal expressions, being therefore the ones remaining those that must be detected by the tagger. Once the text is correctly tagged, old legal references can be recovered. Beside facilitating single-use temporal processing of isolate documents, this service also allows to generate correctly temporally tagged texts with legal references that can be used for training machine-learning based temporal taggers in order to adapt them to the legal domain.

5 Conclusions and Future Work

The work presented shows how just a basic preprocessing for detecting legal expressions to prevent temporal taggers from tagging them can improve temporal tagging on all kind of legal related texts, being for instance able to solve similar cases to the examples exposed in the introduction. Also other languages or kinds of texts could benefit from this preprocessing: the work made for Spanish and general but legal related texts (news and datasets in our case) can be adopted also for other languages and kinds of texts, such as genuine legal documents. Future work include asking experts in the field for more ways in which legal references can be written, along with increasing the amount of references detected and the languages covered by the web service.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602.

References

1. Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics, 2007.
2. Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics, 2010.
3. Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James F Allen, and James Pustejovsky. Semeval-2015 task 5: Qa tempeval-evaluating temporal information understanding with question answering. In *SemEval@ NAACL-HLT*, pages 792–800, 2015.
4. Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062, 2016.
5. Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

6. Frank Schilder. Event Extraction and Temporal Reasoning in Legal Documents. In *Annotating, Extracting and Reasoning about Time and Events, International Seminar, Dagstuhl Castle, Germany, April 10-15, 2005. Revised Papers*, pages 59–71, 2005.
7. Venkateswrlu Naik, Guda Vanitha, and Srujana Inturi. Reasoning in legal text documents with extracted event information. *International Journal of Computer Applications*, 28:8—13, 08 2011.
8. Kolikipogu Ramakrishna, Vanitha Guda, B.Padmaja Rani, and Vinaya Ch. A novel model for timed event extraction and temporal reasoning in legal text documents. *International Journal of Computer Science and Engineering Survey*, 2:39–48, 02 2011.
9. Daniel Isemann, Khurshid Ahmad, Tim Fernando, and Carl Vogel. *Temporal Dependence in Legal Documents*, pages 497–504. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
10. Monica Palmirani, Raffaella Brighi, and Matteo Massini. Automated extraction of normative references in legal texts. In *Proceedings of the 9th international conference on Artificial intelligence and law*, pages 105–106. ACM, 2003.
11. M. Adedjouma, M. Sabetzadeh, and L. C. Briand. Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, pages 63–72, Aug 2014.
12. Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, 22(2):215–237, Jun 2017.
13. Emile de Maat, Radboud Winkels, and Tom van Engers. Automated detection of reference structures in law. *Frontiers in Artificial Intelligence and Applications*, 152:41–50, 2006.
14. M. Palmirani, R. Brighi, and M. Massini. Processing normative references on the basis of natural language questions. In *Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004.*, pages 9–12, Aug 2004.
15. Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu. Automated reference resolution in legal texts. *Artificial Intelligence and Law*, 22(1):29–60, Mar 2014.
16. Monica Palmirani and Federica Benigni. Norma-system: A legal information system for managing time. In *Proceedings of the V legislative XML workshop*, pages 205–223, 2007.
17. Leonardo Lesmo, Alessandro Mazzei, Monica Palmirani, and Daniele P. Radicioni. Tulsi: an nlp system for extracting legal modificatory provisions. *Artificial Intelligence and Law*, 21(2):139–172, May 2013.
18. Mercedes Martínez-González, Pablo de la Fuente, and Dámaso-Javier Vicente. Reference extraction and resolution for legal texts. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 218–221. Springer, 2005.