Event Extraction from Legal Documents in Spanish

Gerardo Sierra[1], Gemma Bel-Enguix[1], Guillermo López-Velarde[1], Ricardo Saucedo[2], Lucía Rivera[1]

[1] Universidad Nacional Autónoma de México, [2]Avoquate

[1]Instituto de Ingeniería, UNAM, Circuito Escolar S/N Instituto de, Ingeniería, Cd. Universitaria, 04510, Ciudad de México, México, [2]Avenida Oaxaca 31, colonia Roma Norte, delegación Cuauhtémoc, 06700, Ciudad de México, México

[1]{GSierraM, Gbele, GLopezVelardeG, LRiveraV}@iingen.unam.mx, [2]ricardos@avoquatemaker.com

Abstract

This work is part of a more general project aiming to design a tool that can help lawyers to find the information they need for litigation in a fast and efficient way. The resource is being designed for Spanish, a language that has a scarceness of Natural Language applications for legal coding, and is tested in 300 documents, mainly writs of 'amparo', a legal procedure to protect human rights, by means a judicial review of governmental action. These documents have been freely downloaded from the Mexican Instituto Federal de Telecomunicaciones. The system, implemented in Python, will include modules to perform several tasks, like automatic classification, Named Entities identification, law detection, structure summarization, and event extraction. This article is focused in one of the most complex parts of the development, event extraction. The algorithm works linking dates with events in the texts. These events are reduced to a list of verbs that have been reported as the most meaningful in this type of texts. For every verb-event, a list of pieces of information will be retrieved: 'who', 'what', 'to whom' and 'where'.

Keywords: legal documents, event extraction, natural language patterns

1. Introduction and Motivation

This paper presents a system for extracting events from legal texts in Mexican Spanish. This is part of a more general project aiming building a tool that can help lawyers to find the information they need in a fast and efficient way.

Our research has been focused in finding patterns for Spanish sentence structures that are used in legal documents. We have worked with 300 documents downloaded from the Mexican 'Instituto Federal de Telecomunicaciones' (IFT). This organization has an open webpage[1] where its resolutions can be accessed.

This article explains the methodology of the system, and its initial performance when trying to automatically detect events and its related date.

2. Previous Work

Lawyers need the processing and study of large quantities of documents as a part of their everyday life. Not having tools available for automatically obtaining the required data from texts, they perform these tasks manually, in what is an expensive and time-consuming activity. Computational linguistics can help lawyers to automatically process the documents they need. Many aspects can be taken into account when dealing with litigation documentation, from consulting laws to getting information of related trials. In what refers to laws, vlex[2] provides an extensive coverage of legislation, including Mexico. This resource offers also links to other laws the text refers to. As for documents generated by litigations, there are several attempts to build databases that can relate some documents with others.

However, it is necessary to have tools able to search in these documents for the information that a lawyer can need in the professional activity. The field of ICT applied to Law has created the area of legal technologies. Sartor et al. (2008) summarize the major types of resources related to legal technologies: legal information search, electronic data discovery, web-based communications, collaborative tools, Metadata and XML Technologies and Technologies in Courtrooms and Judicial Offices. In the last years, however, the area of legal text processing and information extraction, more closely related to Natural Language Processing, has been developed (Francesconi et al., 2010).

A key topic in automatic processing of legal texts is the identification of people and organizations related to a legal case. This is very much related to entity recognition, but must be focused in the fact that these entities need to have a given role in the legal case. In this area, there are several contributions. Dozier et al. (2010) create a hybrid system for named entities recognition and resolution in legal texts, while Quaresma and Gonçalves (2010) use machine learning techniques for solving the same problem. Kumaran & Alan (2004) design a system for NE recognition for new event detection, but it is not related to legal texts. A collection of resources that can be used to deal with legal texts can be found in the document Collection of state-of-the-art NLP tools for processing of legal text, from the project MIREL[3].

The Automatic Context Extraction (ACE)[4] evaluation defines an event as 'something that happens or leads to some change of state' (Nguyen et al., 2016). Meanwhile, Pustejovsky et al. (2002) define it as those expressions into a narrative that can be ordered temporary. This idea was the basis for the organization of TempEval shared tasks (UzZaman et al., 2013), that have helped to the development and testing of of different systems for event extraction and ordering. The area has been a trending topic in text mining. Hogenboom et al. (2011) distinguish three main approaches to the problem: a) data-driven, that try to convert data to knowledge by means to statistics, machine learning, etc.; b) knowledge-driven approaches, that are mainly pattern-based; and c) hybrid, that combine the other models.

Knowledge-driven methods are based on linguistic and lexicographic knowledge. Information is mined using

[1]http://apps.ift.org.mx/cumplimientoStp/secured/adminficum.faces

[2] https://app.vlex.com/

[3] MIREL: Mining and Reasoning with Legal Texts: http://www.mirelproject.eu

[4] https://www.ldc.upenn.edu/collaborations/past-projects/ace

semantic or syntactic patterns. Some examples are Nishihara et al. (2009) and Aone & Ramos-Santacruz (2000). The system Evita to extract events focus on verbs, nouns, nominal phrases and adjectival phrases (Saurí et al., 2005). Other works on event processing (Mani et al., 2003; Filatova y Hovy, 2001) use tools like CLAUSE-IT or CONTEX (Hermjakob y Mooney, 1997) to identify syntactic structures.

Some authors have developed methodologies to extract events from specialized domains. Yakushiji et al (2001) apply this method in the biomedical domain, while Li et al. (2002) work in the financial area and Cohen et al. (2009) focus in biology. As for legal texts event extraction, there is an interesting contribution with English documents (Lagos et al., 2010), based in a semi-automatic approach that integrates two main components: information extraction, and knowledge integration.

Our work fits in the area of knowledge-driven methods, and uses well-known common patterns from legal texts. However, the area is not enough developed in Spanish, and this work presents a small advance in the concrete space of Mexican legal system.

3. Legal Language and Patterns

In order to achieve consistency, validity, completeness and soundness, legal texts are subject to certain constraints, both with respect to content and form. They follow a rigid structural format. Legal writing uses a lot of legal terminology and scholarly words, but specially some linguistic patterns. Danet (1985) describes some legal English features, such as archaic expressions, doublets, unusual prepositional phrases, passive constructions, long sentences and syntactic complexity. Collectively, these features are often called legalese.

For example, among archaic expressions found in legal Spanish documents, there is a frequent use of the expression hereinafter (en lo sucesivo).

...se creó el Instituto Federal de Telecomunicaciones (en lo sucesivo, el "Instituto").

Knowing this type of expressions can be very important for automatic information extraction in legal texts. Being aware that 'Instituto' is a short name, or alias, to name the 'Instituto Federal de Telecomunicaciones' allows the identification of the actant intervening in the event.

Likewise, knowing the syntactic complexity of legal language is very useful to differentiate the relevant information in the description of the event. The use of large sentences and the insertion of appositions is frequent in this type of texts.

Example 1

El 13 de diciembre de 2006, de conformidad con los artículos 13 de la LFT, 16 y 21 de la LFRTV, la COFETEL otorgó a favor del Concesionario, el refrendo de la Concesión para operar y explotar el canal 7. (P_IFT_111215_577_Acc.docx)[5]

In the description of the event of Example 1, several pieces of information can be extracted. First, the date (13/12/2006). The second element is what was done (se otorgó el refrendo [the endorsement was granted]). Another item is who did this (COFETEL) and to whom (el Concesionario [the dealer]). To get all this information several steps have to be performed: a) discarding non-relevant information. In Example 1, it is the apposition (de conformidad con los artículos 13 de la LFT, 16 y 21 de la LFRTV); b) Identifying the 'Who', 'What' and 'to Whom' of the sentence, which most of times are the subject, object and indirect object, respectively.

4. Methodology

Although our goal is the application of the methodology to any type of legal text, so far we have been working with a collection of 300 texts downloaded from the IFT of Mexico, which are freely available on their website. Most of them are what is called writ of 'amparo' in mexican legislation. 'Amparo' is a legal procedure to protect human rights, by means a judicial review of governmental action.

The description of events usually follows a regular pattern involving at least two elements: the action, determined by the main verb, and the date on which the event occurred. In this sense, an analysis was made of the verbs that occur in the writings of amparo, as well as the direct objects of each verb. The description is given below.

4.1 Pre-processing

The first steps in the processing of the corpus are:

a) Change the files format from .docx to .txt.

b) Replace 17 text patterns to help FreeLing 4.0[6] make a better PoS tagging. These patterns can be divided into 3 categories: misspells, conjunctions and business entities types.

Examples of each category can be found in the Table 1:

Replace	With
Con cesiones	Concesiones
y Transportes	Y_Transportes
, S.A. de C.V.	Sadecv

Table 1: Pattern replacement in pre-processing

For example, the word 'Con cesiones' is recurrently misspelled, as it should be 'Concesiones'.

The entity 'Secretaría de Comunicaciones y Transportes' is wrongly tagged as follows: Secretaría de Comunicaciones (NP00000), y (CC), Transportes (NP00000). So, we replace 'y Transportes' with 'Y_Transportes' in order to get the whole entity tagged as NP00000.

Finally, in México there are different types of business entities that can be legally constituted, which names are always referred when a company name is mentioned, for example, the entity 'Telefonía Inalámbrica del Norte, S.A de C.V.'. In this case, the entity is not tagged as NP00000 because the type is referred after a comma. So, we replace ', S.A de C.V.' with 'Sadecv' to avoid further confusions. We also replace the instances where there is no comma

[5] This reference is the name of the document that can be downloaded from the webpage of the IFT.

[6] http://nlp.lsi.upc.edu/freeling/node/1

that separates the business entity type with the name of the company, to homogenize all business entities.
c) PoS tagging by means of FreeLing 4.0. Made with the default Spanish configuration file. In this step, FreeLing also identifies dates, assigning the tag 'W'.
d) Identification of Named Entities (NE). With what the system obtains here, a table is made that will later be modified, if necessary, during the next steps of the whole system. In the meantime, this table serves as basis to detect actants in the events.
All preprocessing is implemented in python. Step d) does not rely only on freeling to identify named entities. The table is built using another rule-based system we have developed for writs of amparo.

4.2 Verbs and dates

Within the investigations that address event detection in text, we found that most of them try to find words, phrases or indicators that establish the point in time where the events happen. That is, they seek to find what linguistic elements are used to express moments or successions, so the computers can use them in a standardized manner. One of the clearest ways in which a point in time or an interval can be represented is by identifying dates.
Every document in our corpus has the date of release, the date of submission and, sometimes, the signature date. Finding these elements is not the goal of the paper, but detecting the ones that are linked to an event in the text of the resolution.
In order to do this, the procedure starts from the idea that, in this type of document, every event is related to a date, and every event is characterized by a main verb that represents the action that is being made. So, all the dates that do not contain such verb, are not taken into account.
Additionally, every event has some actants related to the verb, which correspond to the Named Entities that have to be extracted in the pre-processing task d).
The dates are tagged by FreeLing 4.0 in the pre-processing task c).
Regarding the verbs, we found, by manually analyzing the data, that in the type of documents that we are dealing with, writs of 'amparo' of the IFT, almost every event is correlated to one of the following verbs: 'emitir' [release, issue], 'otorgar' [grant], 'presentar' [submit], 'publicar' [publish], 'solicitar' [request].
The information required for every event in the document is the one in Table 2:

	Who	What	To Whom	Where
emitir	YES	YES	NO	NO
otorgar	YES	YES	YES	NO
presentar	YES	YES	YES	NO
publicar	YES	YES	NO	YES
solicitar	YES	YES	YES	NO

Table 2: Main information items for each one of the verbs that configure the events

In the sequel, the main patterns that have been used for every one of the elements of information are discussed.

4.3 Quién [Who]

If the verb is in active voice, 'Who' is the subject, and it is at the left. This has to be a NP, present in the table of NE of the system.
If the verb is in passive voice, 'Who' is located at the right side, it must be a NP in the table of NE, and it fits into the pattern: 'por + NP'.
In legal texts, some other more elaborated models for 'Who' can be found, both at left or at right of the verb. In 'Who' patterns, the NP is always a NE.
Frequent structures are the ones in which some NPs are explained by other NPs, being both the 'Who' of the event, the second NP can be delimited by colons, or not, and it is an NE. Some common patterns for this structure are:

(1) <[NP] + (,) + 'representante legal de' + [NP](,)>
<[NP]+ (,) + legal representative of + [NP](,)>[7]
(2) <[NP] + (,) + 'mediante' + [NP](,)>
<[NP] + (,) + through+ [NP](,)>
(3) <[NP] + (,) + 'por medio de' + [NP](,)>
<[NP] + (,) + through + [NP](,)>
(4) <[NP] + (,) + 'a través de' + [NP](,)>
<[NP] + (,) + through + [NP](,)>
(5) <[NP] + (,) + [alias]>

Example 2 illustrates pattern (1), where the 'Who' is 'el representante legal de Axtel'.

Example 2
El 16 de octubre de 2009 el representante legal de Axtel presentó ante la Comisión Federal de Telecomunicaciones el escrito No. 321-2009 mediante el cual solicita la intervención de este órgano a efecto de que resuelva los términos condiciones y tarifas aplicables a partir del 10 de enero de 2010 que no ha podido convenir con Telmex y Telnor para la interconexión de sus respectivas redes públicas de telecomunicaciones.
(P_IFT_140410_191.docx)

Example 3 shows more specifically pattern (4). The 'Who' is 'Unidad de Competencia Económica', but this entity is not the one that issued the trade, but another one on its behalf, the 'Dirección General de Concentraciones y Concesiones'.

Example 3
Con fecha 14 de mayo de 2015 la Unidad de Competencia Económica a través de la Dirección General de Concentraciones y Concesiones emitió el oficio IFT mediante el cual remite la opinión correspondiente a la Solicitud de Prórroga. (P_IFT_170316_125_Acc.docx)

Finally, the 'Who' piece can follow the pattern (5), as in 'Telefonos de México, Telmex', where Telmex is an alias

[7] Translations to English are orientative. 'mediante', 'por medio de' and 'a través de' can be roughly translated to 'through'. And they mean that a person is doing something instead of another person who she/he represents.

that works usually instead of 'Teléfonos de México, S.A.B. de C.V.'.
These patterns are not all that can define 'Who' in a legal text, but the ones that can capture almost every structure in the sub-genre of writs of amparo.

4.4 Qué [What]

If the verb is in active voice, 'What' is usually immediately after the verb, at its right. If the verb is in passive voice, the 'What' is at left.
It is a NP, VMN, VMP or VMS. In the Example 2, the word 'escrito' has the form of a VMP.
In the verb 'publicar', the 'What' is usually found in quotes, as shown in Example 4, where 'Decreto por el que se expiden la Ley Federal de Telecomunicaciones y Radiodifusión, (...)' is marked as 'What'.

4.5 Dónde [Where]

After the analysis of the documents, we found that only the verb 'publicar' is expected to have this information. To find it, we use the pattern <'en' + NP>, which we observed to be the most common for this verb. In Example 4, 'Diario Oficial de la Federación' fits said pattern.

Example 4
El 14 de julio de 2014 se publicó en el Diario Oficial de la Federación el "Decreto por el que se expiden la Ley Federal de Telecomunicaciones y Radiodifusión y la Ley del Sistema Público de Radiodifusión del Estado Mexicano; y se reforman adicionan y derogan diversas disposiciones en materia de telecomunicaciones y radiodifusión" mismo que entró en vigor el 13 de agosto de 2014. (P_IFT_170216_57_Acc.docx)

4.6 A quién [To whom]

To find the phrase that stands for 'to Whom', some common patterns are:

(1) <'a' + NP>
<*to + NP*>
(2) <'ante' + NP>
<*before + NP*>
(3) <'en favor de' + NP>
<*in favor of + NP*>
(4) <'a quien' + NP>
<*to whom + NP*>

This NP must be located immediately after one of the verbs that are considered, or at least they do not have to have any other verb between both elements.
In Example 5 'C. Ricardo León Garza Limón' is marked as 'to Whom', because it fits pattern (3) and the NP comes after the main verb 'otorgó', without any other verbs in between.

Example 5
El 18 de octubre de 2005 la Secretaría de Comunicaciones Y Transportes (la "Secretaría") otorgó en favor de el C. Ricardo León Garza Limón un título de concesión.

5. Discussion and Future Work

This is a work in progress that aims at finding patterns in Legal Language in Mexican Spanish in order to extract events in writs of 'amparo'.
The application has retrieved good results so far, but the system must be improved in several ways, for: a) designing a system capable to obtain every pattern for each element of information, due to the ones that have been implemented so far do not cover every possible case, but only the more general ones; b) taking into account juridic expressions in non-Mexican Spanish; c) being extended to other specific areas of litigation, which may result in a wider variety of verbs that can define new types of events.
An important area that should be improved is evaluation. So far, the only way to do it is manually by humans.
Also, in the future we aim to implement a machine learning algorithm trained with manually annotated data, so we can compare this rule-based system to the supervised learning algorithm and figure out which approach is better in the long run, thinking that this information may someday be part of a bigger system.
From this seminal design we plan to build a system that can efficiently extract every piece of information lawyers can need from a legal text, and design friendly systems that can truly help in the court.

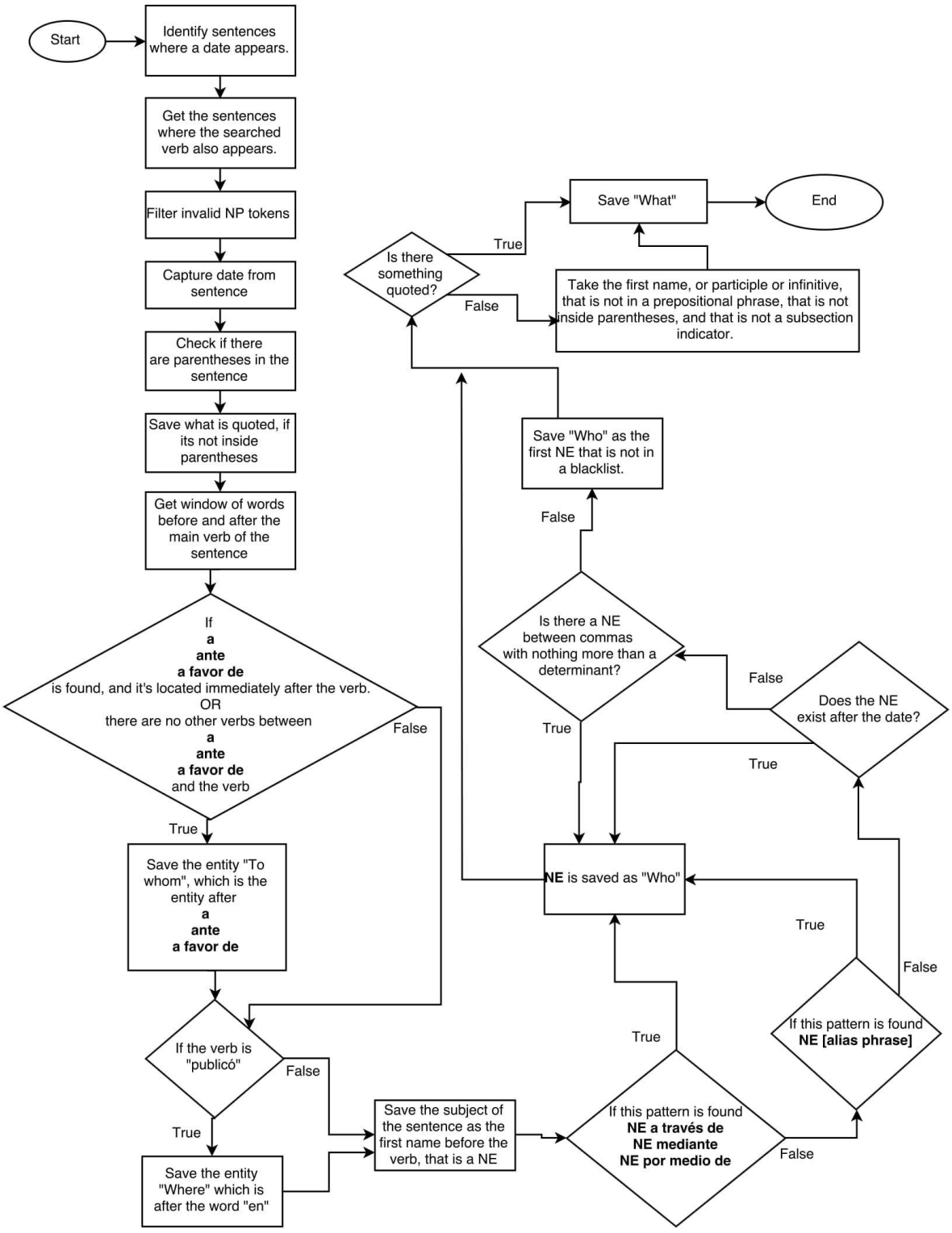

Figure 1: Scheme showing the steps used to implement the event identification system

6. Bibliographical References

Aone, C., Ramos-Santacruz, M. (2000). REES: A Large-Scale Relation and Event Extraction System. In: 6th Applied Natural Language Processing Conference (ANLP 2000): 76–83. Association for Computational Linguistics.

Cohen, K.B., Verspoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgartner, Jr., W.A., White, E., Tipney, H., Hunter, L. (2009). High-Precision Biological Event Extraction with a Concept Recognizer. In: Workshop on BioNLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. pp. 50–58. Association for Computational Linguistics (2009).

Danet, B. (1985). "Legal Discourse". In Teun A. Van Dijk (ed.) Handbook of Discourse Analysis. Vol. 1, 237 – 291 . London : Academic Press.

Dozier C., Kondadadi R., Light M., Vachher A., Veeramachaneni S. and Wudali R. (2010). Named Entity Recognition and Resolution in Legal Text. In Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds) *Semantic Processing of Legal Texts. Lecture Notes in Computer Science*, vol 6036. Springer, Berlin, Heidelberg.

Filatova, E. and Hovy, E. (2001). Assigning time-stamps to event-clauses. In Proceedings of the workshop on Temporal and spatial information processing-Volume 13, p. 13.

Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds) (2010). *Semantic Processing of Legal Texts. Lecture Notes in Computer Science*, vol 6036. Springer, Berlin, Heidelberg.

Hermjakob, U. y Mooney, R.J. (1997). Learning parse and translation decisions from examples with rich context. En Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, págs. 482–489.

Hogenboom, F., Frasincar, F., Kaymak, U, de Jong, F. (2011), An Overview of Event Extraction from Text. In Workshop on Detection, Representation and Exploitation of Events in the Semantic Web (DeRiVE 2011), vol 799, pags 48-57, CEUR Workshop Proceedings.

Kumaran, G. and Allan, J. (2004). Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '04). ACM, New York, NY, USA, 297-304. DOI=http://dx.doi.org/10.1145/1008992.100904.

Lagos, N., Segond, F., Castellani, S. and O'Neill, J. (2010). Event Extraction for Legal Case Building and Reasoning. In Zhongzhi Shi, Sunil Vadera, Agnar Aamodt, David Leake. Intelligent Information Processing V, 340, Springer, pages 92-101, 2010, IFIP Advances in Information and Communication Technology, 978-3-642-16326-5. <10.1007/978-3-642-16327-2 14>. <hal-01055067>

Li, F., Sheng, H., Zhang, D. (2002) Event Pattern Discovery from the Stock Market Bulletin. In: 5th International Conference on Discovery Science (DS 2002). Lecture Notes in Computer Science, vol. 2534, pp. 35–49. Springer-Verlag Berlin Heidelberg.

Mani, I., Schiffman, B., y Zhang, J. (2003). Inferring temporal ordering of events in news. En Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003 – short papers - Volume 2, págs. 55–57.

Morris, F. J. (2005). E-Discovery: Best practices for employment lawyers. what support do you need? how do you work with E-Discovery experts. Current Developments in Employment Law, ALI-ABA, Santa Fe, NM.

Nguyen, T.H., Cho, K. and Grishman, R. (2016). Joint Event Extraction via Recurrent Neural Networks, Proceedings of NAACL-HLT 2016, pages 300–309.

Nishihara, Y., Sato, K., Sunayama, W. (2009). Event Extraction and Visualization for Obtaining Personal Experiences from Blogs. In: Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction. Part II. Lecture Notes in Computer Science, vol. 5618: 315–324. Springer-Verlag Berlin Heidelberg.

Pustejovsky, J., Saurí, R., Setzer, A., Gaizauskas, R., and Ingria, B. (2002). TimeML annotation guidelines. TERQAS Annotation Working Group 23.

Quaresma, P., & Gonçalves, T. (2010). Using linguistic information and machine learning techniques to identify entities from juridical documents. In Semantic Processing of Legal Texts (pp. 44-59). Springer Berlin Heidelberg.

Sartor, G., Casanovas, P., Casellas, N., Rubino, R. (2008). Computable models of the law and ICT: State of the art and trends in european research. In: Casanovas, P., Sartor, G., Casellas, N., Rubino, R. (eds.) Computable Models of the Law. LNCS (LNAI), vol. 4884, pp. 1–20. Springer, Heidelberg.

Saurí, R., Knippen, R., Verhagen, M., y Pustejovsky, J. (2005). Evita: a robust event recognizer for QA systems. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, págs. 700–707.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 1-9).

Yakushiji, A., Tateisi, Y., Miyao, Y. (2001). Event Extraction from Biomedical Papers using a Full Parser. In: 6th Pacific Symposium on Biocomputing: 408–419 (2001).