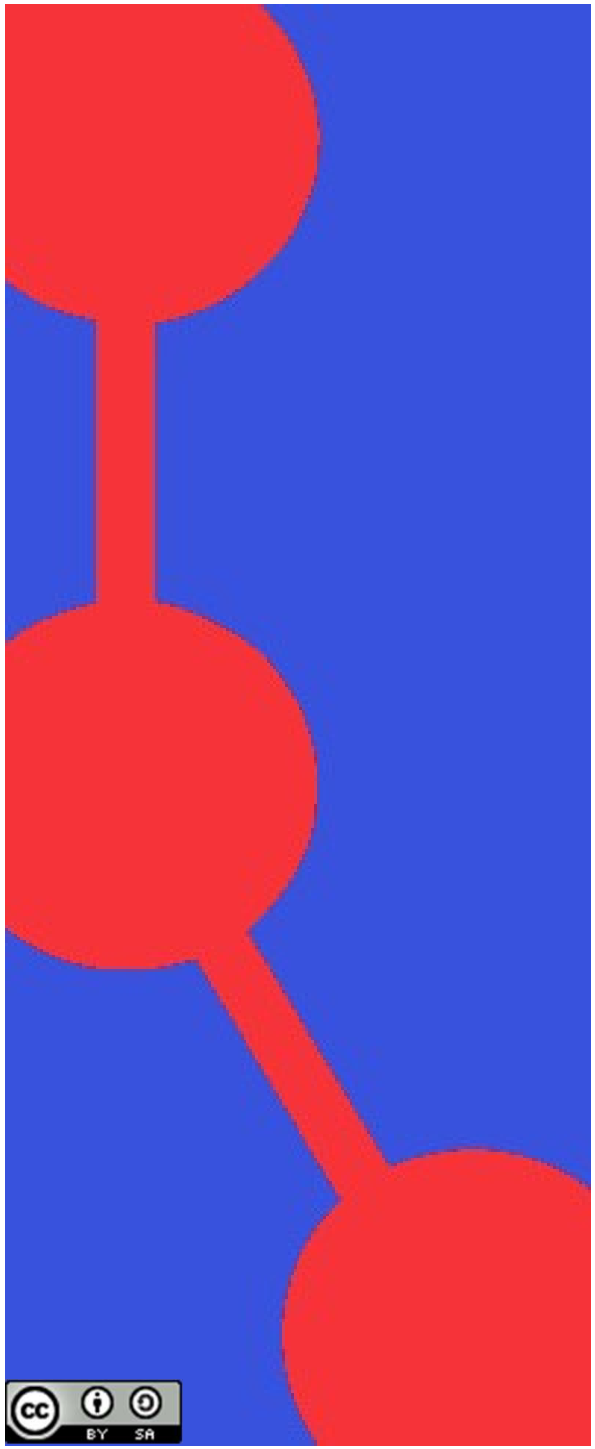


TimeLex:
Tools for Processing Temporal Information in Legal Texts

María Navas-Loro
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

Víctor Rodríguez-Doncel
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

mnavas@fi.upm.es
https://marianavas.oeg-upm.net

11/12/2019
IberLegal 2019

Outline

- Brief Introduction
- Tools
  - lawORdate
  - Añotador
  - Event extractor
- Conclusions

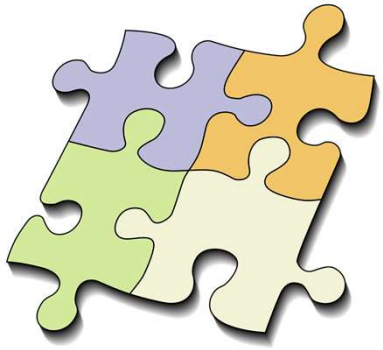
Temporal Annotation

BRIEF INTRODUCTION

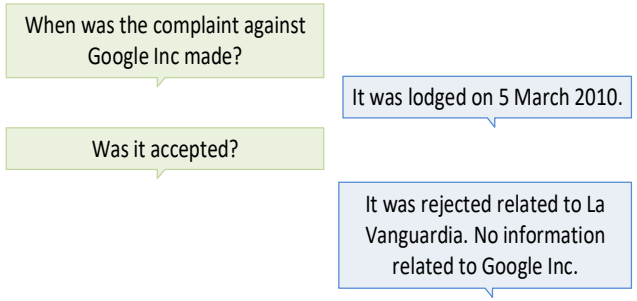
Temporal information is everywhere:

news, medical texts, legal documents…

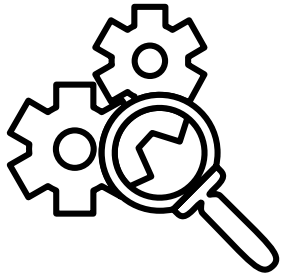
… and we can use it for a lot of things!



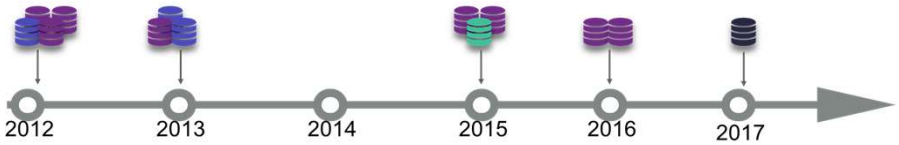
PATTERN DETECTION



QA



SEARCH ENGINE



TIMELINE GENERATION

SUMMARIZATION

… but there are not a lot of tools to extract temporal information in Spanish and for the legal domain

Example

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.

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.

Most common misleading expressions

Directiva 2012/27/UE

RD 1463/2007

Real Decreto 1341/2007

Ley Orgánica 10/1995

Ley 22/2011, de 28 de julio

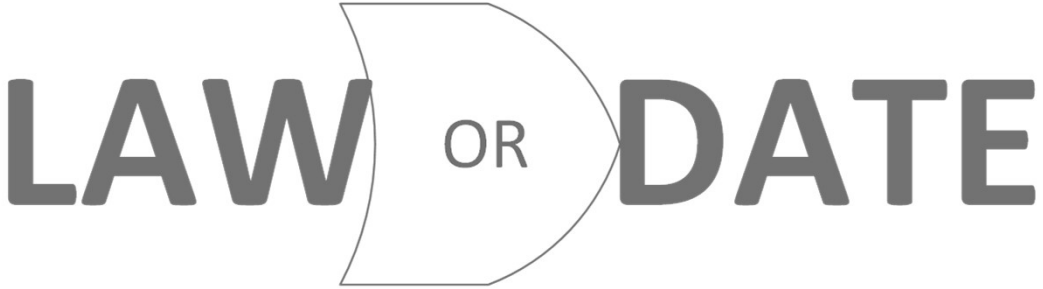
BOE: 29/07/2011

BOE de 22 de julio

BOE núm. 306, de diciembre

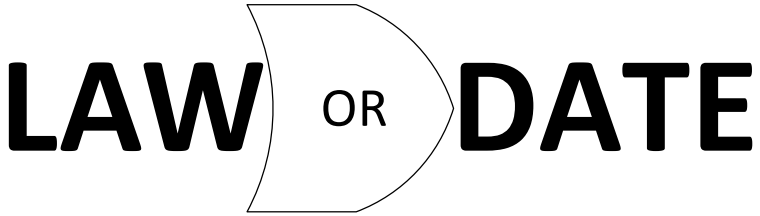
So the first step will be preprocessing the text to delete all these misleading expressions…

…and then we could start looking for temporal information.



Step 1

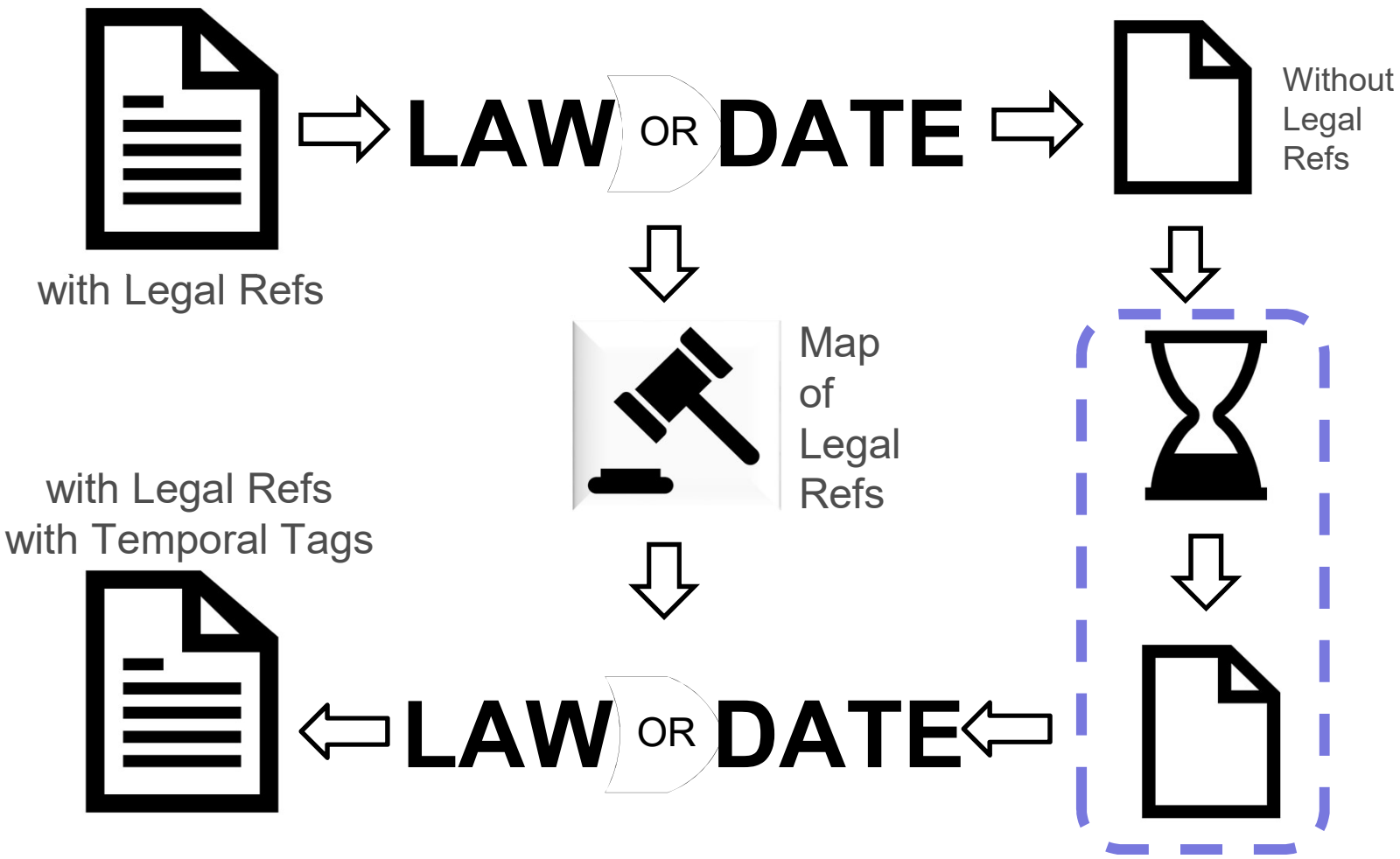
LAWORDATE



http://legalwhen.appspot.com/ http://ceur-ws.org/Vol-2049/04paper.pdf

lawORdate [1] is a webapp that takes a text with legal references including dates and returns it without them. Once the temporal tagging is done, the references are restored in the text.

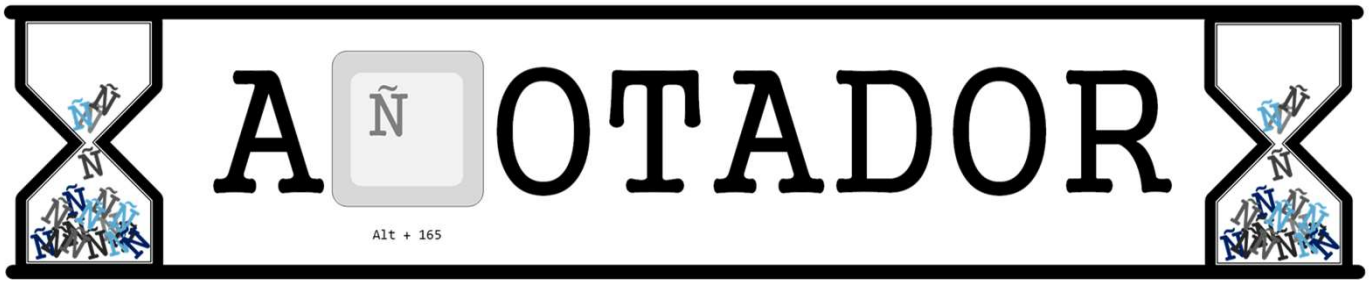
An example: "En la presente base de datos se recogen los elementos inscritos en el registro creado via el **Real Decreto 2093/2008, de 19 de diciembre**. Ha sido actualizado por ultima vez el 13 de agosto de 2017."


with Legal Refs
LAW
OR
DATE
Without
Legal
Refs
Map
of
Legal
Refs
with Legal Refs
with Temporal Tags
LAW
OR
DATE

Once we have the text free of misleading expressions, we can start looking for temporal information.

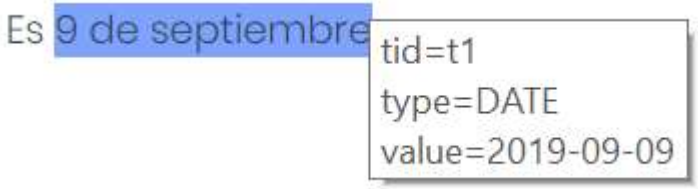
Step 2

AÑOTADOR



http://annotador.oeg-upm.net/

Añotador (aka Annotador) [2] is a temporal tagger for Spanish and English. It is able to find and normalize dates, times, durations and sets (temporal expressions that repeat periodically, such as “monthly”, “twice a year” or “every Tuesday”).

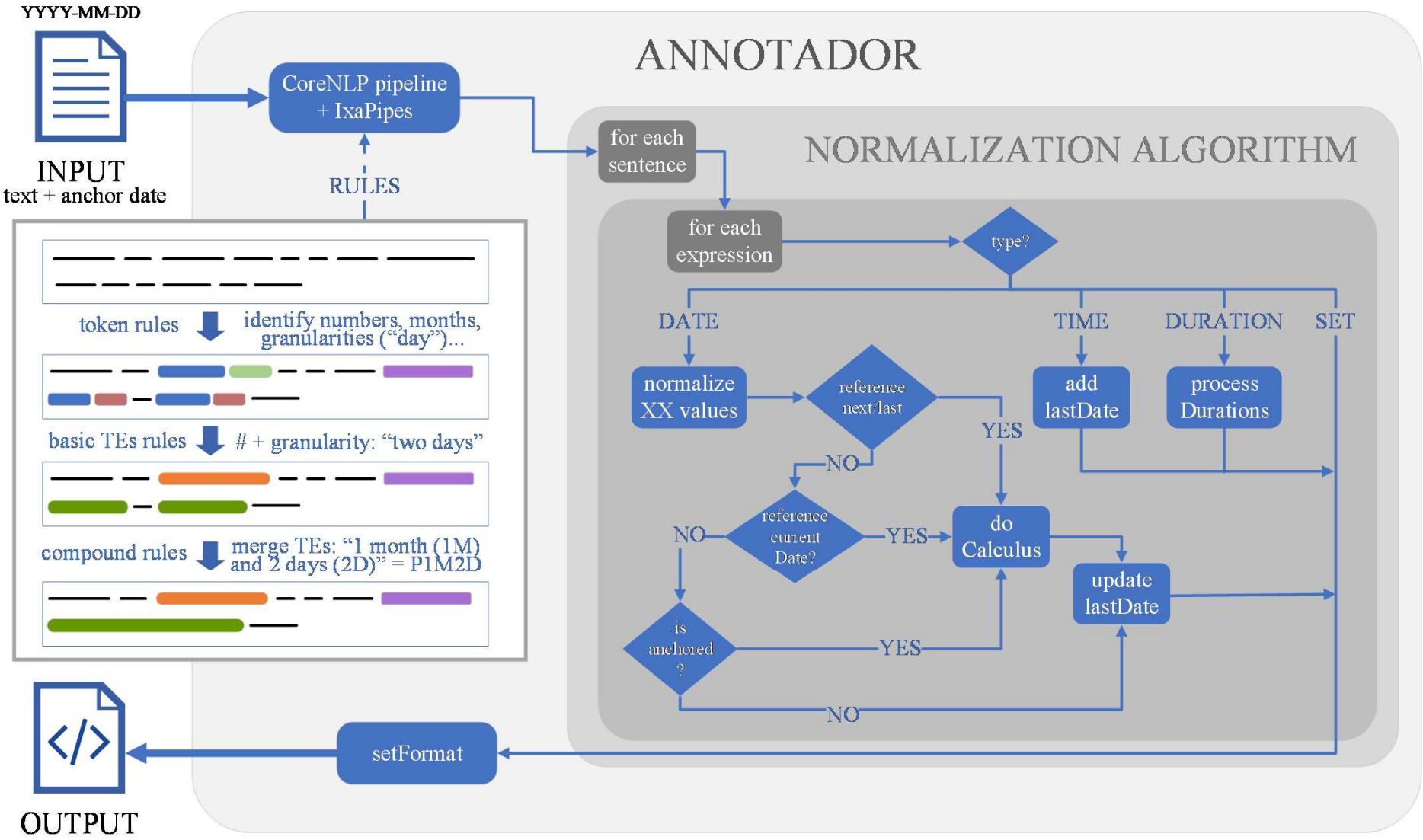


Añotador

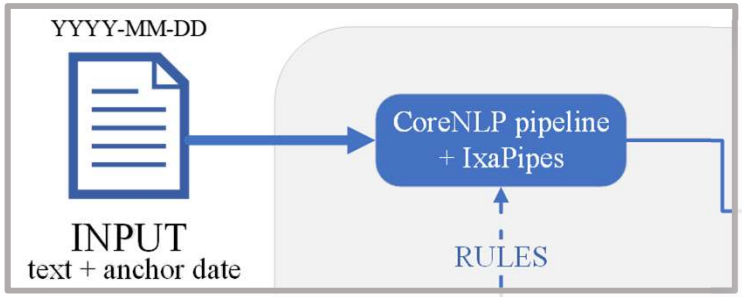
- There is a demo available in its webpage, and can be called as webservice or downloaded from GitHub (where all the code and a docker image are available).

- Results:
  - Outperforms other Spanish taggers in general texts, including colloquial and Hispanic American expressions [2].
  - Outperforms other English taggers in legal European English texts (from the European Court of Human Rights and the European Court of Justice).

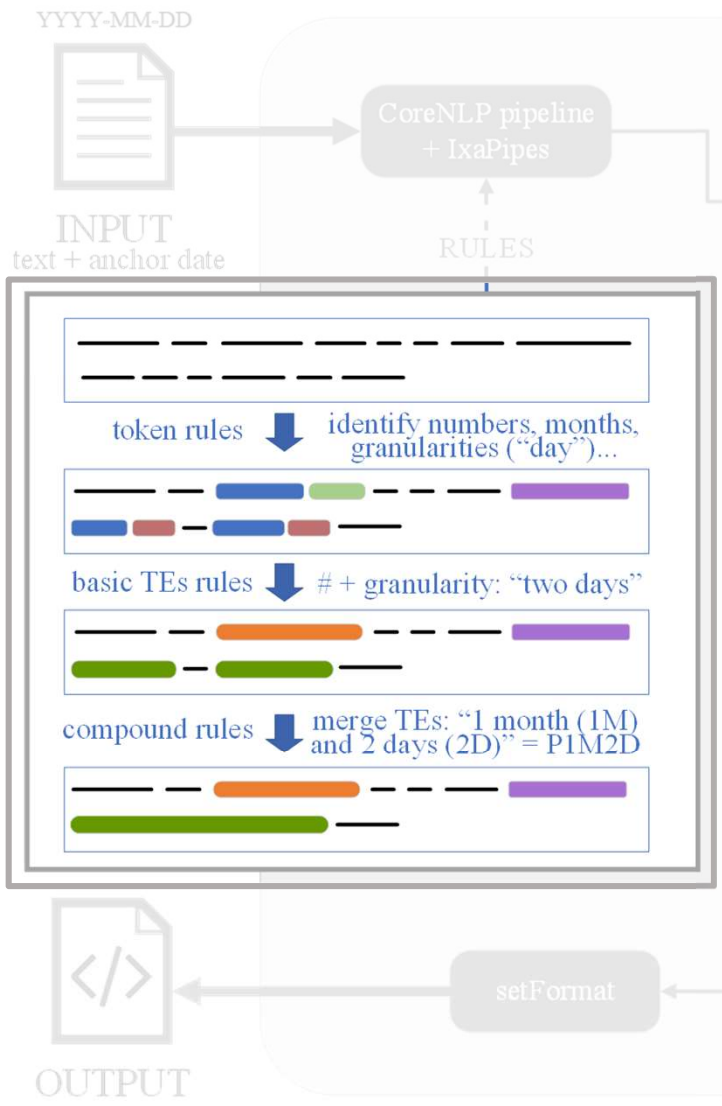
Pipeline of Añotador



Pipeline of Añotador



1. **Preprocessing:**
   - We get as input the text and the anchor date (if none, we assume the current day)
   - We use CoreNLP for lemmatizing, sentence splitting…
   - We added IxaPipes models for Spanish to improve the quality of the output.



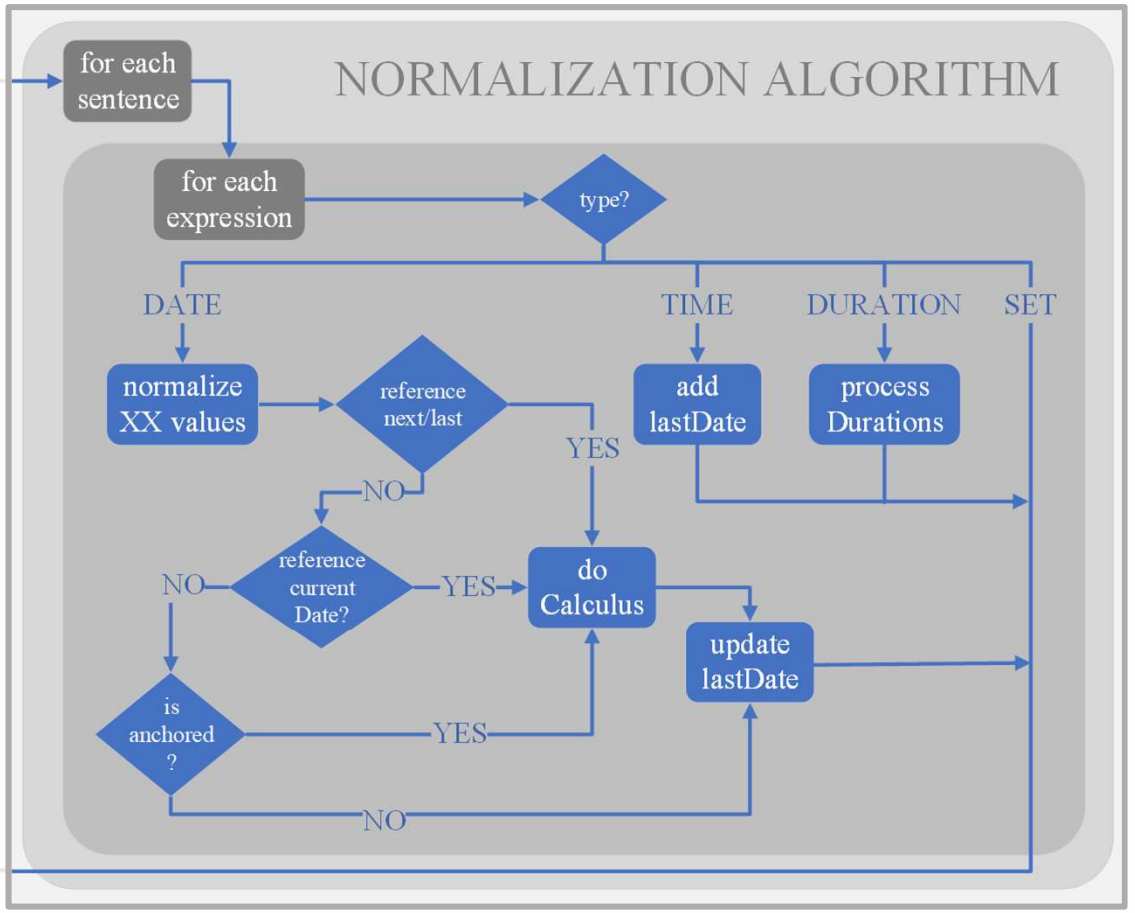
2. Rules:

More than 100 rules written in CoreNLP TokensRegex format.

1. **Token-based rules** for expressions such as numerals, granularities…
2. **Basic temporal expression rules**, working on previously found basic expressions
3. **Compound expression rules**, for inheritance values or composition.
4. **Literal expression rules**, for specific expressions such as "previously", polysemic expressions (Spanish "mañana") or expressions including tokens that can appear in other temporal expressions, requiring a dedicated approach.

3. Normalization algorithm:

- Works for each sentence sepparately.
- Different approaches for each type of expressions.
- Can take into account different reference dates.



Pipeline of Añotador

4. Output format:

- TimeML
- JSON

OK, now we know “when”… but we are also interested in “what”, “who”, “where”…

We need *events*!

Step 3

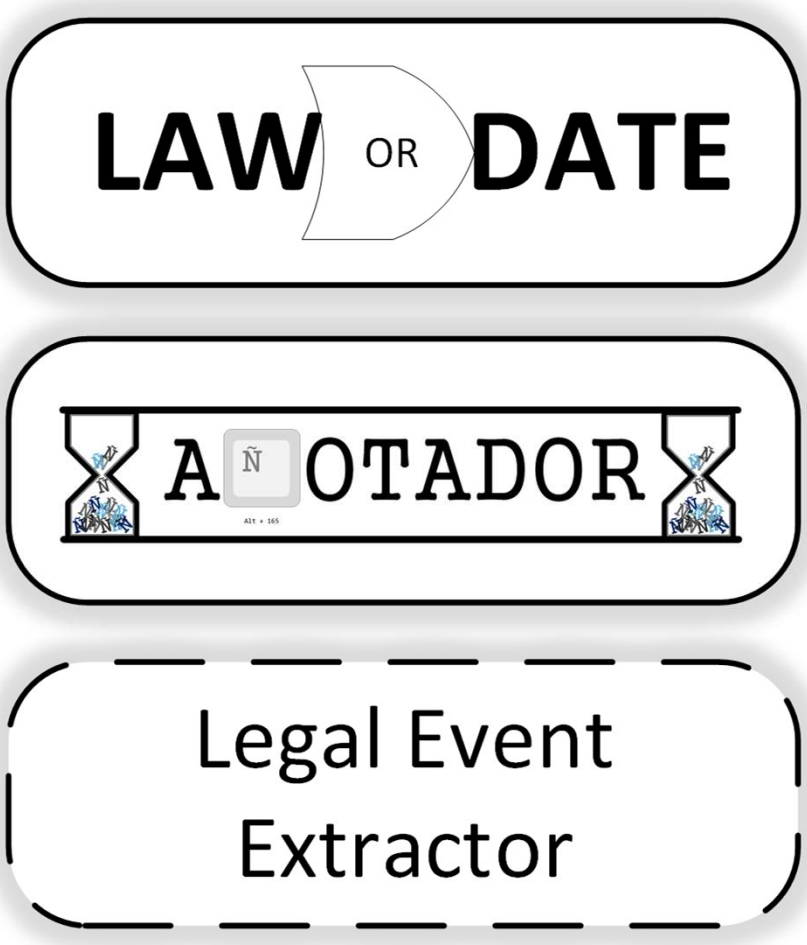
EVENTS

We are currently working on the detection of events, specifically trying to define what are legal *relevant* events in judgments.

We have already some results on:

- Theorical analysis [3, 4].
- Detection within contracts (ContractFrames [5]).

Pipeline



lawORdate prepares the legal document in order to make it is easier to be processed by the other tools

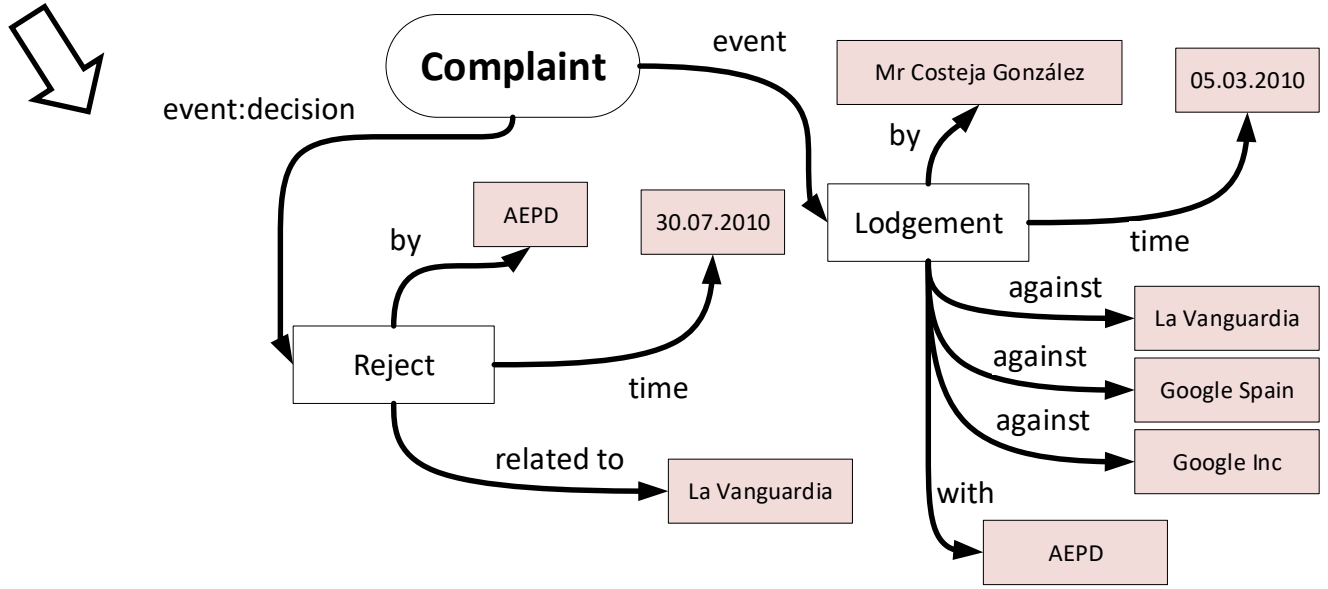
Añotador finds and normalizes all the temporal expressions.

With the tool we are currently developing, we find the relevant legal events in a text.

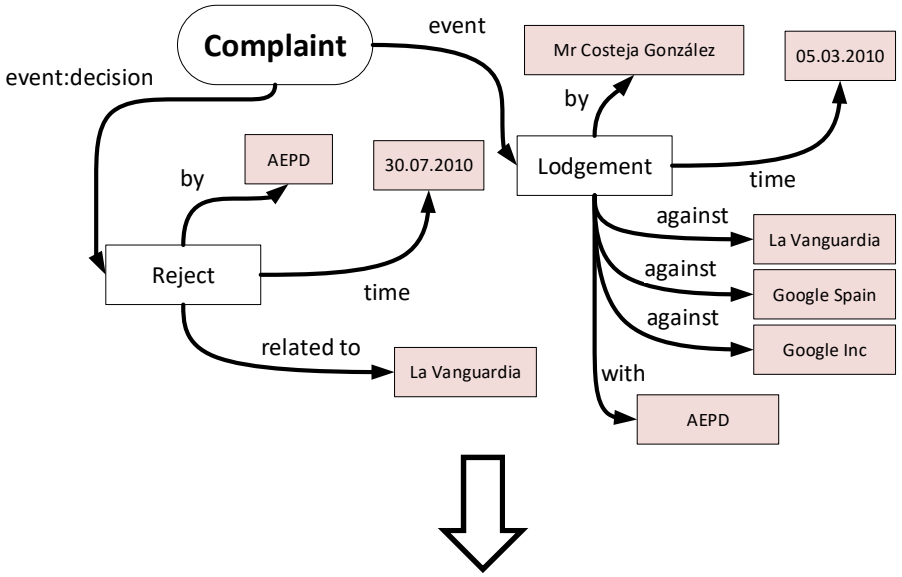
From text to event representation

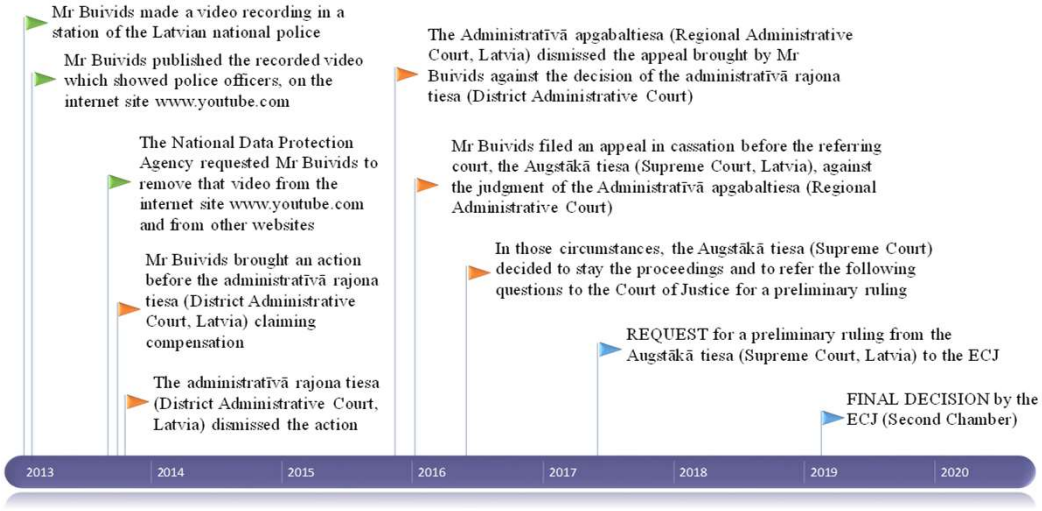
14
On **5 March 2010**, Mr Costeja González, a Spanish national resident in Spain, lodged with the AEPD a complaint against La Vanguardia Ediciones SL, which publishes a daily newspaper with a large circulation, in particular in Catalonia (Spain) ('La Vanguardia'), and against Google Spain and Google Inc. The complaint was based on the fact (…)

16
By decision of 30 July 2010, the AEPD rejected the complaint in so far as it related to La Vanguardia (…)



From event representation to applications





An event-based representation of judgments would facilitate the understanding of legal documents to both layman and legal practitioners by allowing, among others, the generation of:

- Timelines
- Event-based summaries
- Pattern-based analysis of judgments
- Question Answer systems
- … and more!

Future work

CONCLUSIONS

Conclusions

Results:

- **lawORdate**: Preprocessing of legal documents to facilitate the task of temporal taggers (Añotador).
- **Añotador**: Detection of temporal expressions to facilitate event detection and anchoring in Spanish and English texts.

Current work:

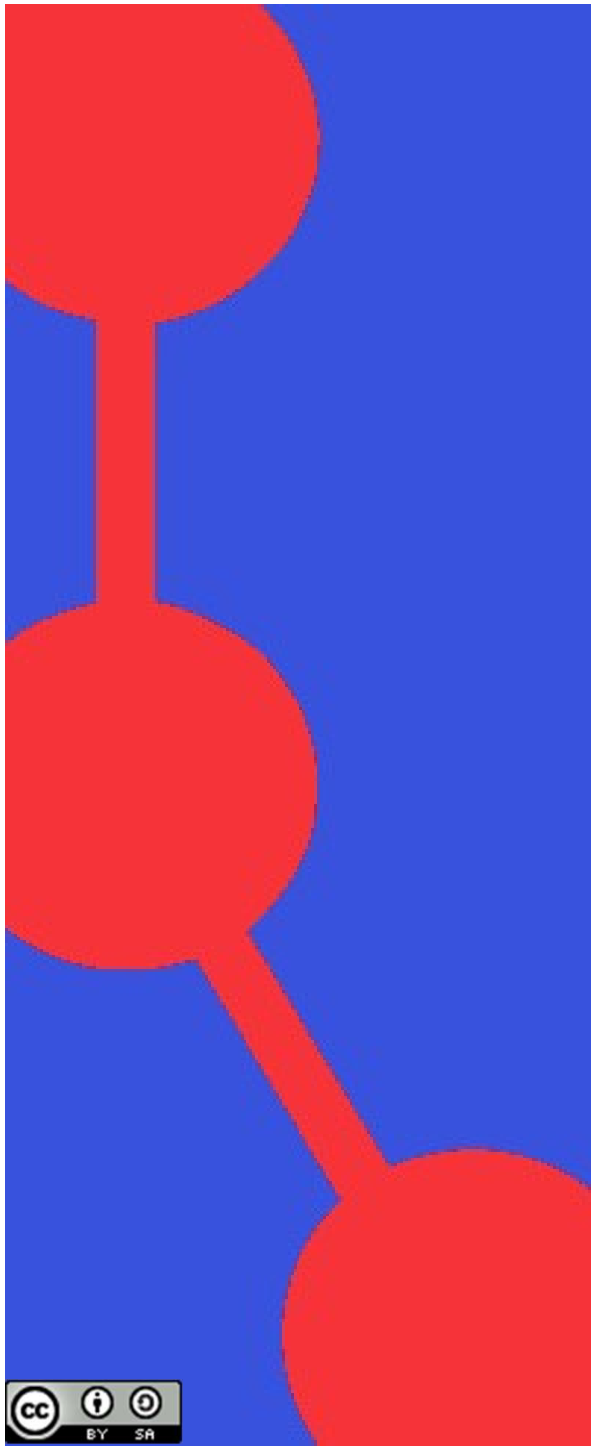
- **Event extractor in the legal domain**: Detection of relevant events in the legal domain.

Application:

TimeLex suite offers tools that allow to represent legal documents as sets of events to facilitate tasks such as summarization, question answering, timeline generation, pattern detection...

References

1. Navas-Loro, M. (2017). *LawORDate: a Service for Distinguishing Legal References from Temporal Expressions*. Proceedings of TeReCom 2017: Workshop on Technologies for Regulatory Compliance at JURIX (TeReCom 2017).
2. Navas-Loro, M., Rodríguez-Doncel, V. (2019). "Annotador: a Temporal Tagger for Spanish". To be published in the Journal of Intelligent and Fuzzy Systems.
3. Navas-Loro, M., Filtz, E., Rodríguez-Doncel, V., Polleres, A., Kirrane, S. (2019). "TempCourt: Evaluation of Temporal Taggers on a new Corpus of Court Decisions". To be published in the Knowledge Engineering Review journal.
4. Navas-Loro, M., Santos, C. (2018). *Events in the legal domain: first impressions*. In: Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance co-located with the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018), Groningen, The Netherlands, December 12, 2018. Pp. 45–57.
5. Navas-Loro, M., Rodríguez-Doncel, V., Satoh, K., *ContractFrames: Bridging the Gap Between Natural Language and Logics in Contract Law* – In: Kojima K., Sakamoto M., Mineshima K., Satoh K. (eds) New Frontiers in Artificial Intelligence. JSAI-isAI 2018. Lecture Notes in Computer Science, vol 11717. Springer, Cham. doi: 10.1007/978-3-030-31605-1_9.



TimeLex:
a Suite of Tools for Processing Temporal Information in Legal Texts

María Navas-Loro
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

Víctor Rodríguez-Doncel
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

mnavas@fi.upm.es
https://marianavas.oeg-upm.net

11/12/2019
IberLegal 2019

Results against the Hourglass corpus

In the HourGlass corpus, Añotador always shows the best performance.

		strict			lenient			average		
Tagger	Attribute	P	R	F1	P	R	F1	P	R	F1
Annotador	value	0.7231	0.7068	0.7148	0.7949	0.7769	0.7858	0.7590	0.7419	0.7503
	type	0.7923	0.7744	0.7833	0.8846	0.8647	0.8745	0.8385	0.8195	0.8289
	extent	0.8333	0.8145	0.8238	0.9462	0.9248	0.9354	0.8897	0.8697	0.8796
Heidel	value	0.5672	0.4762	0.5177	0.6358	0.5338	0.5804	0.6015	0.5050	0.5490
	type	0.6060	0.5088	0.5531	0.8239	0.6917	0.7520	0.7149	0.6003	0.6526
	extent	0.6239	0.5238	0.5695	0.8716	0.7318	0.7956	0.7478	0.6278	0.6826
SUTime	value	0.3019	0.0802	0.1267	0.4528	0.1203	0.1901	0.3774	0.1003	0.1584
	type	0.4717	0.1253	0.1980	0.8019	0.2130	0.3366	0.6368	0.1692	0.2673
	extent	0.4717	0.1253	0.1980	0.8868	0.2356	0.3723	0.6792	0.1805	0.2851

Value means the normalization is correct

Type means that the temporal expression was correctly classified

Extent means that the temporal expression was correctly marked in the text

Results against the TempEval2 corpus

Results of the temporal taggers in the TempEval 2 corpus. HeidelTime is slightly better at finding the normalized value (0.0027 on average), but Añotador is better in the rest of metrics. SUTime shows good Precission.

		strict			lenient			average		
Tagger	Attribute	P	R	F1	P	R	F1	P	R	F1
Annotador	value	0.8021	**0.7778**	0.7897	0.8281	**0.8030**	0.8154	0.8151	**0.7904**	0.8026
	type	0.8438	**0.8182**	**0.8308**	0.9063	**0.8788**	**0.8923**	**0.8750**	**0.8485**	**0.8615**
	extent	0.8646	**0.8384**	0.8513	0.9323	**0.9040**	**0.9179**	0.8984	**0.8712**	**0.8846**
Heidel	value	**0.8418**	0.7525	**0.7947**	**0.8644**	0.7727	**0.8160**	**0.8531**	0.7626	**0.8053**
	type	**0.8531**	0.7626	0.8053	0.8870	0.7929	0.8373	0.8701	0.7778	0.8213
	extent	**0.9040**	0.8081	**0.8533**	0.9435	0.8434	0.8907	**0.9237**	0.8258	0.8720
SUTime	value	0.6377	0.2222	0.3296	0.8261	0.2879	0.4270	0.7319	0.2551	0.3783
	type	0.6522	0.2273	0.3371	**0.9275**	0.3232	0.4794	0.7899	0.2753	0.4082
	extent	0.6667	0.2323	0.3446	**0.9565**	0.3333	0.4944	0.8116	0.2828	0.4195

References - Corpora

1. J. Pustejovsky et al., “The timebank corpus,” in Corpus linguistics, vol. 2003, p. 40, Lancaster, UK, 2003.
2. http://semeval2.fbk.eu/semeval2.php?location=tasks&taskid=5
3. L. Minard et al., “Meantime, the newsreader multilingual event and time corpus,” in Proc. LREC 2016, 2016.
4. W. Styler IV et al., “Temporal annotation in the clinical domain,” Transactions of ACL, vol. 2, pp. 143–154, 2014.
5. P. Mazur et al., “Wikiwars: A new corpus for research on temporal expressions,” in Proc. EMNLP 2010, pp. 913–922, 2010.
6. J. Strötgen et al., “Temporal tagging on different domains: Challenges, strategies, and gold standards,” in LREC, vol. 12, pp. 3746–3753, 2012.
7. J. Tabassum et al., “Tweetime: A minimally supervised method for recognizing and normalizing time expressions in twitter,” arXiv preprint arXiv:1608.02904, 2016.
8. https://catalog.ldc.upenn.edu/LDC2012T12
9. https://catalog.ldc.upenn.edu/LDC2012T01

References - Taggers

10. J. Strötgen and M. Gertz, “Multilingual and cross-domain temporal tagging,” LREv, vol. 47, no. 2, pp. 269–298, 2013.
11. A. X. Chang and C. D. Manning, “Sutime: A library for recognizing and normalizing time expressions,” in LREC, vol. 2012, pp. 3735–3740, 2012.
12. X. Zhong et al., “Time expression analysis and recognition using syntactic token types and general heuristic rules,” in Proc. the 55th Annual Meeting of the ACL, vol. 1, pp. 420–429, 2017.
13. S. Bethard, “ClearTK-TimeML: A minimalist approach to TempEval 2013,” in Proc. the Workshop SemEval 2013, pp. 10–14, ACL, 2013.
14. L. Derczynski et al., “Usfd2: Annotating temporal expresions and tlinks for tempeval-2,” in Proc. the Workshop SemEval, pp. 337–340, ACL, 2010.
15. K. Lee et al., “Context-dependent semantic parsing for time expressions,” in Proc. the 52nd Annual Meeting of the ACL, vol. 1, pp. 1437–1447, 2014.
16. M. Verhagen et al., “Automating Temporal Annotation with TARSQI,” in Proc. the ACL 2005 on Interactive Poster and Demonstration Sessions, pp. 81–84, ACL, 2005.
17. N. Chambers et al., “Dense event ordering with a multi-pass architecture,” Transactions of ACL, vol. 2, pp. 273–284, 2014.
18. H. Llorens et al., “Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2,” in Proc. the Workshop SemEval, pp. 284–291, ACL, 2010.

Temporal taggers:

1) Detect temporal expressions:

- DATE
- DURATION
- SET
- TIME

2) Normalize them.

The 3rd of April 2016 is OK, but the 20th May would be better.

```
<?xml version="1.0"?>
<!DOCTYPE TimeML SYSTEM "TimeML.dtd">
<TimeML>
<TIMEX3 tid="t4" type="DATE" value="2016-04-03">The 3rd of April 2016</TIMEX3>
is OK, but <TIMEX3 tid="t6" type="DATE" value="2016-05-20">the 20th May</TIMEX3>
would be better.
</TimeML>
```

3) Additionally, event and relation extraction