Plan TL
IBERLEGAL

Ontology Engineering Group

Terminology Extraction in the Legal Domain for Spanish Administration

Pablo Calleja Ibáñez
Estudiante de doctorado
Universidad Politécnica de Madrid

11 de diciembre de 2019

Agradecimientos

- Lote 4 Exp. 017/18 - DOC20180307171649PCT SEAD and INCIBE
- Horizon 2020 No 780602 - Lynx
- Plan de Tecnologías del Lenguaje
- OEG

Índice

1. Problems in information extraction

Plan TL IBERLEGAL

- Legal documentation is usually **difficult to understand**
  - long, intricate sentences, complex expressions and the particular legal terminology

- The **identification of accurate terms** is a key task to improve its comprehension

- Such legal terms are understood as words **and multi-word expressions** and **verbs or adjectives** used in a particular context

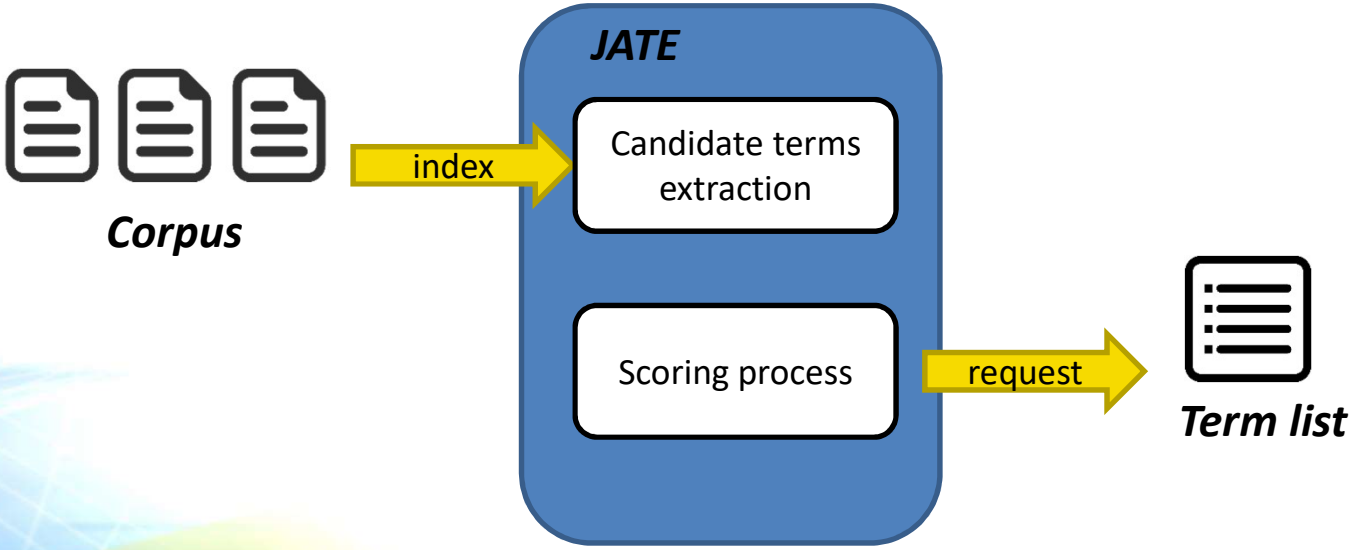
1. Problems in information extraction

- Current Automatic Terminology Extractions tools are oriented for **general terms**
- **Data privacy problems**. Cannot rely on online or third party tools
- Problems with a **massive payload** of documents.

2. Developed solution

SELECTED TOOL: **JATE 2**

- Free to use – (alpha version)
- Built over Apache Solr
- Configurable for different use cases
- 10 state of the art Terminology Extraction algorithms (TTF-IDF, Cvalue, ...)



2. Developed solution

- Extension for Spanish language
- NLP tasks (tokenization, POS tagging, etc.) rely on **OpenNLP**
- **POS tags** are coded with CAST3LB format
- There is **no Chunker** for Spanish in OpenNLP. JATE allows the creation of patterns to identify chunks based on the POS tags
- General patterns have been translated and adapted to the legal domain

Patterns coded for the identification of Spanish legal chunks

```
default (\bAQ\b)
default (\bVM\b)
default (\bNC\b)
default (\bAQ\b) (\bNC\b)
default (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bAQ\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b)
default (\bNC\b) (\bSP\b) (\bDA\b) (\bNC\b) (\bAQ\b)
default (\bNC\b) (\bSP\b) (\bNC\b) (\bSP\b) (\bNC\b)
default (\bNC\b) (\bAQ\b) (\bSP\b) (\bNC\b) (\bAQ\b)
default (\bAQ\b) (\bNC\b) (\bSP\b) (\bAQ\b) (\bNC\b)
default (\bAQ\b) (\bNC\b) (\bSP\b) (\bNC\b) (\bAQ\b
default (\bNC\b) (\bAQ\b) (\bSP\b) (\bDA\b) (\bNC\b
```

3. Obtained results

USE CASE 1: *Lote 4 del proyecto: Servicios de obtención y clasificación de información para la caracterización del sector de la ciberseguridad*

3 Spanish Corpus

- Tenders Electronic Daily (**TED**) in Spanish which is comprised of public procurement notices from the EU and beyond. 448 documents
- **CODICE**. Spanish platform for the public procurement. 200 documents
- A proprietary **INCIBE corpus** of involving 1.297 documents of cybersecurity alerts

MAIN CHALLENGE: Terminology extraction with high number of private data

3. Obtained results

Algorithms: Cvalue and TTF-IDF

Examples of Cvalue terms

TED Corpus	Codice Corpus	Incibe Corpus
plataforma de logística sanitaria (sanitary logistics platform)	pliego de cláusulas administrativas (specifications of administrative clauses)	protección de datos (data protection)
suministro de energía eléctrica (power supply)	colaboradora con la seguridad social (social security partner)	política de privacidad (política de privacidad)
pliego de prescripciones técnicas (Technical specification sheet)	clausulas administrativas particulares (private administrative clauses)	seguridad de la información (security of the information)
medidas de contratación publica (public procurement measures)	sección de asuntos económicos (economic affairs section)	reglamento general de protección (general protection regulation)

3. Obtained results

USE CASE 2: *Lynx - European Project*

Corpus : Spanish collective agreements

Algorithms: C-Value and TTF-IDF

Results: 6% of new terms applying legal patterns with C-value and 3.5% with TTF-IDF in the first 200 relevant terms

MAIN CHALLENGE:

Legal Spanish terminology extraction

New terms C-value	New terms TTF-IDF
ley de prevención de riesgos (risk prevention law)	anónima (anonymous)
representación legal de los trabajadores (legal representation of workers)	flexible (flexible)
miembros del comité de empresa (members of the company committee)	discontinuo (discontinuous)

Examples of new terms

4. Conclusions

- The results have shown the importance of the complex nominal chunks, adjectives and verbs in the legal domain
- Poor **POS tagging** process. Results could be better by improving this task
- Other processes should be considered such as **Named Entity Recognition**
- The developed solution works for high amount of documents without compromising the privacy policies.

4. Future lines

- Integrate other **libraries** for the NLP tasks
  - e.g., **IXA Pipes** and the **EAGLE** tags
- Test and implement other Terminology Extraction algorithms
- Improve the distributed request in JATE
- Improve the working interface

Gracias

www.PlanTL.es

PlanTecnologiasLenguaje@mineco.es