Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-193

Concept Recognition in European and National Law

Rohan NANDA [a,b,1], and Giovanni SIRAGUSA [a] and Luigi DI CARO [a] and Martin THEOBALD [b] and Guido BOELLA [a] and Livio ROBALDO [b] and Francesco COSTAMAGNA [a]

[a] *University of Turin, Italy*
[b] *University of Luxembourg*

Abstract. This paper presents a concept recognition system for European and national legislation. Current named entity recognition (NER) systems do not focus on identifying concepts which are essential for interpretation and harmonization of European and national law. We utilized the IATE (Inter-Active Terminology for Europe) vocabulary, a state-of-the-art named entity recognition system and Wikipedia to generate an annotated corpus for concept recognition. We applied conditional random fields (CRF) to identify concepts on a corpus of European directives and Statutory Instruments (SIs) of the United Kingdom. The CRF-based concept recognition system achieved an F1 score of 0.71 over the combined corpus of directives and SIs. Our results indicate the usability of a CRF-based learning system over dictionary tagging and state-of-the-art methods.

Keywords. Concept Recognition, European Law, Information Retrieval

1. Introduction

With the increasing volume of European and national legislation available online, the identification of domain concepts in legal texts is very important for the development of legal information retrieval systems. The identification of domain concepts provides a deeper insight into the interpretation and understanding of texts. The recognition of concepts in legal texts would also be useful for the harmonization and integration of European and national law. Research in this domain has mainly focused on identification of named entities like person, organization and location names. However, European and national legislation contains very few instances of named entities. They primarily comprise legal and domain-specific jargon which can be represented by concepts.

In this paper, we develop a system for concept recognition in European directives and national law (statutory instruments of the United Kingdom). The concept recognition system was used for automatically identifying concepts in a corpus of 2884 directives and 2884 SIs. We generated an annotated corpus using a semi-supervised approach to

[1] Corresponding Author

save human effort and time for evaluation of our system. Further, we also generated a mapping to link similar terms in directives and SIs under the same concept.

The rest of the paper is organized as follows. In the next section, we discuss the related work. Section 3 describes the concept recognition system. Section 4 discusses the results and analysis. The paper concludes in Section 5.

2. Related Work

Related work is mainly focused in the domain of named entity recognition (NER) systems. In [1], the authors developed a legal named entity recognizer and linker by aligning YAGO[2] (WordNet-and Wikipedia-based ontology) and the LKIF ontology. The alignment was carried out manually by mapping a concept node in LKIF to its equivalent in YAGO. They utilized different models like support vector machines (SVM), Stanford Named Entity Recognizer (NER) [4], and neural networks and evaluated the system on a small sample of judgements from the European Court of Human Rights (ECHR). Their results indicate that the LKIF level of generalization is not suitable for named entity recognition and classification as their system was unable to distinguish between the classes defined in LKIF. However, their NER system achieved a better performance while distinguishing YAGO classes. The authors in [3] developed a named entity recognition and classification system to recognize entities like judges, attorneys, companies, courts and jurisdictions in US case law, depositions, pleadings and other trial documents. They utilized dictionary lookup, contextual pattern rules and statistical models for identifying named entities. The NER system was trained using a SVM classifier and evaluated on manually and automatically acquired training datasets of case law. The authors in [2] developed a NER system using AdaBoost. The system uses a window, along with a set of features (part-of-speech tags and dictionary of words) to capture the local context of a word. Current NER systems are based on conditional random fields (CRF) [5], which allow to train a unique model for the classification and recognition of named entities. In [4], the authors developed a CRF which used Gibbs sampling instead of the standard Viterbi algorithm. They demonstrated that the use of Gibbs sampling allowed the system to distinguish between mentions of organization or person on the basis of context, thus enforcing label consistency.

3. Concept Recognition System

In this section, we describe the concept recognition system for European and national law. In the legal domain, concepts are generally represented using ontologies or vocabularies. Previous NER systems (based on the concepts represented in the LKIF ontology) demonstrated that the LKIF level of generalization was not suitable [1]. This is because NER systems could not clearly distinguish between the classes defined in LKIF. Therefore, in this paper we investigate the use of vocabularies for developing our concept recognition system. We utilize Inter-Active Terminology for Europe [3] (IATE), which is the EU's inter-institutional terminology database. IATE consists of 1.3 million entries in

[2]http://www.yago-knowledge.org/

[3]http://iate.europa.eu

English. Every entry (concept) in IATE is mapped to a subject domain. We filtered out some irrelevant entries in IATE (stopwords and concepts mapped to 'NO DOMAIN').

We utilized a corpus of 2884 directives and 2884 statutory instruments for our experiments. Since training data was not available, we utilized a semi-supervised approach to generate an annotated corpus. The development of NER or concept recognition systems require a large amount of manually annotated datasets, which are expensive to obtain. We manually annotated a few documents with IATE subject domains. Then we developed a dictionary lookup program to tag terms (both words and phrases) in the text with IATE subject domains. Each term in the text was compared to entries in the IATE vocabulary and matching terms were tagged with the relevant subject domains. Table5 shows some examples of these terms and subject domains. We also used spaCy[4], a state-of-the-art NER system to annotate time, date and monetary units. We filtered out irrelevant candidate entities by using Dexter [7], a Wikipedia entity linker.[5] Then, we annotated all the documents in the corpus. After generating the annotated corpus for both directives and SIs we divided each dataset into an 80% training (2307 documents) and a 20% test set (577 documents) to build the concept recognition system. The combined corpus comprised 80% training set (2307 directives + 2307 SIs) and 20% test set (577 directives + 577 SIs). Table 1 shows the number of documents, tokens and vocabulary size for both the directive and SI datasets, respectively. We observe that SIs have a much larger vocabulary than directives. We utilized conditional random fields (CRFs) to build our concept

Table 1. Number of documents, number of tokens and the vocabulary size (|V|) for directives (left) and SIs (right). We computed $|V_{total}|$ as $|V_{train}| + |V_{test}| - |V_{train} \cap V_{test}|$

Dataset	# docs	# tokens	|V|
Train	2,307	4,646,286	24,522
Test	577	1,226,338	14,127
Total	2,884	5,872,624	38,649

Dataset	# docs	# tokens	|V|
Train	2,307	4,189,157	83,172
Test	577	1,096,246	33,757
Total	2,884	5,285,403	116,929

recognition system as they have been known to work well in tasks which require labeling sequence data (especially natural language text). They are discriminative probabilistic models where each observation is a token from a sentence and the corresponding label (tag of subject domain or entity) represents the state sequence. We utilize the following features for our CRF model: word suffix, word identity (whether a word represents a subject domain/named-entity or not), word shape (captialized, lowercase or numeric) and part-of-speech (POS) tags. We used the limited-memory BFGS training algorithm with L1+L2 regularization.

4. Results and Analysis

In this section, we present the results of our system. Table 2 reports the F1 score of our CRF-based concept recognition model for each subject domain and entity class. We observe that all IATE subject domains are clearly distinguished due to the achievement of a reasonable F1 score for each domain for each corpus. The lower F1 score of domain 'INTERNATIONAL ORGANISATIONS' and named entities like 'QUANTITY', 'MONEY' and 'ORDINAL' is explained by a smaller number of tagged tokens, resulting

[4]https://spacy.io/

[5]Wikipedia Entity Linkers find named entities in the text that can be linked to a Wikipedia page.

in only a few training instances The other subject domains had sufficient training data and therefore were classified with a higher F1 score. These results also indicate that European and national legislation consist of very few named entities and are therefore more suited for concept recognition.

Table 2. Results (F1 score) for concept recognition for each class by the CRF-based concept recognition system

Tag name	Directives	SIs	Directives + SIs
IATE Subject Domains			
FINANCE	0.68	0.62	0.62
POLITICS	0.70	0.74	0.71
ENVIRONMENT	0.68	0.41	0.66
EDUCATION AND COMMUNICATIONS	0.68	0.72	0.71
LAW	0.92	0.81	0.89
INTERNATIONAL ORGANISATIONS	0.52	0.14	0.32
EMPLOYMENT AND WORKING CONDITIONS	0.70	0.68	0.70
AGRI-FOODSTUFFS	0.75	0.73	0.68
INDUSTRY	0.67	0.45	0.60
PRODUCTION TECHNOLOGY AND RESEARCH	0.69	0.67	0.69
BUSINESS AND COMPETITION	0.78	0.77	0.77
ENERGY	0.81	0.50	0.74
TRANSPORT	0.59	0.60	0.58
EUROPEAN UNION	0.79	0.77	0.76
AGRICULTURE FORESTRY AND FISHERIES	0.70	0.58	0.64
SOCIAL QUESTIONS	0.68	0.65	0.66
ECONOMICS	0.66	0.57	0.68
GEOGRAPHY	0.52	0.76	0.75
INTERNATIONAL RELATIONS	0.70	0.59	0.59
SCIENCE	0.60	0.48	0.59
TRADE	0.77	0.66	0.76
spaCy Named Entities			
QUANTITY	0.00	0.00	0.00
MONEY	0.00	0.00	0.00
ORDINAL	0.00	0.00	0.00
TIME	0.62	0.00	0.60
DATE	0.00	0.19	0.47

Table 3. Results of concept recognition with CRF model and comparison with a baseline ("Most frequent class") and the Stanford NER model

Corpus	System	Precision	Recall	F1 score
Directive Corpus	Most frequent class	0.74	0.53	0.61
	CRF	0.80	0.71	0.75
	Stanford NER	0.80	0.71	0.75
SIs Corpus	Most frequent class	0.61	0.40	0.48
	CRF	0.73	0.61	0.66
	Stanford NER	0.68	0.53	0.59
Combined Corpus (Directives + SIs)	Most frequent class	0.66	0.47	0.54
	CRF	0.76	0.68	0.71
	Stanford NER	** (did not finish training)	**	**

The average F1 scores of our CRF-based concept recognition system for directive, SI and combined corpus were 0.75, 0.66 and 0.71 respectively (Table 3). We also compare the performance of the CRF with a baseline method (the "Most frequent class" model). We observe that the CRF outperforms the baseline model. This is because the baseline model does not take into account the context information for a particular token while assigning it to a class. We also compared the CRF with Stanford NER for both the directive corpus and the SIs corpus. The CRF model had similar performance to the Stanford NER

in the directive corpus. However, it outperformed the Stanford NER in the SIs corpus by achieving a higher F1 score. For the combined corpus,the Stanford NER was still in training and we could not record the results in time (Stanford NER takes several days for training, perhaps due to use of long n-gram sequences). These runs are indicated by ** in Table 3. Our CRF system did not include n-gram features.

One drawback of using dictionary tagging to annotate a corpus is that some terms are missed and not tagged due to inconsistent rules to accommodate different phrases and tokenization errors. In the IATE dictionary, an entry, e.g., 'integrated energy performance', is linked to a subject domain, e.g., 'INDUSTRY'. Table 4 presents an example sentence with tagged labels of the IATE dictionary and predicted CRF labels. The CRF classifies both 'energy' and 'performance' to the 'INDUSTRY' subject domain, whereas the dictionary missed them. This is because the dictionary lookup utilizes state-of-the-art tokenizers which may not be 100% accurate and may lead to an incorrect tokenization, thus resulting in a mismatch. The CRF on the other hand, had some training instances from which it learns that the terms 'energy' and 'performance' are related to 'INDUSTRY'. Thus it was able to correctly classify them. Therefore, training a CRF model is advantageous also on automatically annotated corpora because it can improve the tagging of the dictionary by learning these semantic relations between terms and subject domains. Thus, it can be used to improve the quality of annotations and develop a better gold standard for further work.

Table 4. Comparison of CRF output with the dictionary tagging

	CRF predicted labels	Dictionary
calculation	O	O
of	O	O
the	O	O
integrated	O	O
energy	**INDUSTRY**	**O**
performance	**INDUSTRY**	**O**
of	O	O
buildings	O	O

In order to utilize the concept recognition system, it is important to align similar terms across European and national law. This semantic alignment of terms is highly useful for legal professionals to understand the differences in terminologies at the European and national level. The concept recognition system generates a large collection of terms under each subject domain from both directives and statutory instruments. We divided the terms under each subject domain into two lists: directive terms and SI terms. We computed the set difference of these two lists to obtain a list of terms present in the directives but not in the SIs. Similarly, we also obtained a list of terms present in SIs but not in the directives. We then computed text similarity (using Levenshtein distance) to find the most semantically similar term in the SIs (but not present in the directives) for a particular term in the directive. Table 5 shows a few examples of such terms. In future work, we intend to use the mapping of such terms to extend our text similarity system of detecting also transposing provisions for EU directives [6].

5. Conclusion

In this paper, we developed and evaluated a CRF-based concept recognition system for European and national law. We generated a labeled corpus of directives and statutory

Table 5. Aligned terms from European and national law

Subject Domain	Aligned terms (*Directive→SIs*)
EMPLOYMENT AND WORKING CONDITIONS	*professional qualification→vocational qualification* *seniority→job security* *occupational disease→industrial disease*
FINANCE	*life assurance→endowment assurance* *financial institution→financial administration* *dividend→tax on dividends*

instruments with subject domains of the IATE vocabulary, Wikipedia and a state-of-the-art named entity recognition system. We evaluated the system on both a European and national law corpus and analyzed its performance with respect to a baseline model and the Stanford NER tagger. Our results indicate that the concept recognition system is able to identify concepts in both directives and UK statutory instruments with a F1 score of 0.71 over the combined corpus. It can also be used to iteratively improve the dictionary-lookup based tagging from IATE. We also demonstrated that concept recognition systems are useful to align legal terminology at European and national level to assist legal practitioners and domain experts.

Acknowledgements

Research presented in this paper is conducted as a PhD research at the University of Turin and the University of Luxembourg within the Erasmus Mundus Joint International Doctoral (Ph.D.) programme in Law, Science and Technology. This work has been partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts". The authors would like to thank Ba Dat Nguyen from the Max-Planck Institute for his comments and suggestions.

References

[1] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. *16th International Conference on Artificial Intelligence and Law (ICAIL), 2017.*

[2] Xavier Carreras, Lluis Marquez, and Lluıs Padro. Named entity extraction using AdaBoost. *6th Conference on Natural language learning*, volume:20, pages:1-4, 2002.

[3] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in Legal text. *Semantic Processing of Legal Texts*, pages:27-43 Springer, 2010.

[4] Jenny Rose Finkel, Trond Grenager, and Christopher Manning Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages:363-370. 2005.

[5] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages:282-289, 2001.

[6] Rohan Nanda, Luigi Di Caro and Guido Boella. A Text Similarity Approach for Automated Transposition Detection of European Union Directives. *Proceedings of the 29th International Conference on Legal Knowledge and Information Systems (JURIX2016)*, pages:143-148, 2016.

[7] Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Dexter 2.0: An Open Source Tool for Semantically Enriching Data. *ISWC-PD'14 Proceedings of the 2014 International Conference on Posters and Demonstrations Track*, volume:1272, pages:417-420. 2014.