Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-1

Named Entity Recognition, Linking and Generation for Greek Legislation

Iosif ANGELIDIS [a], Ilias CHALKIDIS [b] and Manolis KOUBARAKIS [a]
[a] *National and Kapodistrian University of Athens, Greece*
[b] *Athens University of Economics and Business, Greece*

Abstract. We investigate named entity recognition in Greek legislation using state-of-the-art deep neural network architectures. The recognized entities are used to enrich the Greek legislation knowledge graph with more detailed information about persons, organizations, geopolitical entities, legislation references, geographical landmarks and public document references. We also interlink the textual references of the recognized entities to the corresponding entities represented in other open public datasets and, in this way, we enable new sophisticated ways of querying Greek legislation. Relying on the results of the aforementioned methods we generate and publish a new dataset of geographical landmarks mentioned in Greek legislation. We make available publicly all datasets and other resources used in our study. Our work is the first of its kind for the Greek language in such an extended form and one of the few that examines legal text in a full spectrum, for both entity recognition and linking.

Keywords. Named Entity Recognition and Linking, Dataset Generation, Entity Reference Representation, Deep Learning

1. Introduction - Related Work

Recently, there has been an increased interest in the adaptation of Artificial Intelligence technologies to the legal domain including text processing, knowledge representation and reasoning. Legal text processing [1] is a growing research area, comprising of tasks such as legal question answering [2] and legal entity extraction [3,4]. The same applies to the area of legal knowledge representation, where new standards have been developed and started to be adopted based on semantic web technologies. Relevant contributions here are the European Legislation Identifier (ELI) [5] for legislation, the European Case Law Identifier (ECLI) [6] for case laws, as well as LKIF [7] and LegalRule ML [8] for the codification of advanced legal concepts, such as rules and norms. The research community aims to develop tools and applications to help legal professionals as well as ordinary citizens. Based on these principles, Chalkidis et al. [9] developed Nomothesia (`http://legislation.di.uoa.gr`), a platform which makes Greek legislation available on the Web as linked open data to aid its sophisticated querying using SPARQL and the development of relevant applications.

Deepening this effort in order to build a bridge, as a point of reference, between those relative research fields of artificial intelligence (natural language processing and semantic web), we developed a Named Entity Recognizer (NER) and Linker (NEL) for Greek legislation.

Our main contributions are listed below:

- We study the task of named entity recognition in Greek legislation (Section 2) by evaluating state-of-the-art neural architectures that have been applied in legal text for other tasks (contract elements extraction [3], recognition of requisite-effectuation parts [4]). In these experiments, we compare two alternative token shape encodings, which signify the importance of an expressive feature representation.
- We introduce a novel RDF vocabulary for the representation and linking of textual references to entities (Subsection 3.1). As Chalkidis et al. [9], we consider RDF as a single data model for representing both metadata of a legislative document and knowledge that is encoded in the text.
- We deploy Nomothesia NER, based on the best model BILSTM-BILSTM-LR (Subsection 2.2) with a macro-averaged F_1 of 0.88, in the Greek legislation dataset [9] and produce new data for entity references, that we describe using the new RDF vocabulary.
- We link the references with open public datasets (Greek administrative units and Greek politicians) using rule-based techniques and the Silk framework [10] (See Section 3).
- We make publicly available new benchmark datasets (Subsection 2.1.1) of 254 annotated pieces of legislation related to named entity recognition and linking. Pre-trained word embeddings specialized in Greek legal text, demonstration code in python and other supplementary material are also provided.
- We generate a new RDF dataset of Greek geographical landmarks based on the results of Nomothesia NER by applying heuristic rules (Subsection 3.5).
- Based on the above procedures, we augment the knowledge base and the querying capabilities of the Nomothesia platform in two significant ways: tracing legislation citation networks and searching using entity-based criteria (Subsection 3.6).

All methods and practices that are described throughout this work can be applied in any given language given the appropriate pre-trained word embedding and datasets.

2. Entity Recognition

In this paper, we focus on extracting 6 entity types, when present:

Person Any formal name of a person mentioned in the text (e.g., Greek government members, public administration officials, etc.).

Organization Any reference to a public or private organization, such as: international organizations (e.g., European Union, United Nations, etc.), Greek public organizations (e.g., Social Insurance Institution) or private ones (e.g., companies, NGOs, etc.).

Geopolitical Entity Any reference to a geopolitical entity (e.g., country, city, Greek administrative unit, etc.)

Geographical Landmark References to geographical entities such as local districts, roads, farms, beaches, which are mainly included in pieces of legislation related to topographical procedures and urban planning.

Legislation Reference Any reference to Greek or European legislation (e.g., Presidential Decrees, Laws, Decisions, EU Regulations and Directives, etc.)

Public Document Reference Any reference to documents or decisions that have been published by a public institution (organization) that are not considered a primary source of legislation (e.g., local decisions, announcements, memorandums, directives).

2.1. Datasets

2.1.1. Benchmark Datasets

The benchmark datasets[1] contain 254 daily issues for classes A and D of the Greek Government Gazette over the period 2000-2017. Every issue contains multiple legal acts. Class A issues concern primary legislation published by the Greek government (e.g., laws, presidential decrees, ministerial decisions, regulations, etc.). Class D issues concern decisions related to urban, rural and environmental planning (e.g., reforestations, declassifications, expropriations, etc.). We uniformly splitted the issues across training (162), validation (45), and test (47) in terms of publication year and class. Thus the possibility of overfitting due to specific linguistic idiosyncrasies in the language of a government or due to specific entities and policies has been minimized. Our group annotated all of the above documents for the 6 entity types that we examine. We also created datasets that contain pairs of entity references and the respective matching Universal Resource Identifiers (URIs) in other open public datasets.

2.1.2. Word Embeddings

The last few years, feature engineering in NLP, which results in sparse feature representations, has gradually been replaced by the use of dense word vectors, most notably word embeddings. Word embeddings are pre-trained using unsupervized algorithms [11,12] over large corpora based on the linguistic observation that similar words tend to co-occur in similar contexts (phrases). Thus, word embeddings capture both semantic and syntactic information as well as correlations between words. In our work, we applied WOR2VEC (skip-gram model) [11] to an unlabelled corpus, which contains: 150,000 issues of the Greek Government Gazette in the period of 1990-2017; all publicly available pieces of legislation from the foundation of the Greek Nation in 1821 until 1990, which sum up to 50,000; 1,500 case laws published online by Greek Courts; all EU Treaties, Regulations and Decisions that have been translated in Greek and published in EUR-Lex; and the Greek part of the European Parliament Proceedings Parallel Corpus.

We produced 100-dimensional word embeddings for a vocabulary of 428,963 words (types), based on 615 millions of tokens (words), included in the unlabelled corpus. We used Gensim's implementation of WORD2VEC (http://radimrehurek.com/gensim/). Out of vocabulary words were mapped to a single 'UNK' embedding. To generalize across numbers with similar patterns and tokens that differ in letter-case formats, the unlabeled corpus was pre-processed to be upper-cased, de-accented and underwent replacement of all digits by 'D' for all tokens. We opted to do such transformations, in contrast to the usual lower-casing, based on the fact that the modern Greek alphabet consists of 24 upper-case letter cases and 25 lower-case ones, which also in many circumstances can be accented with 2 different accents in their lower-case formats. So, we normalize every word in an upper-case non-accented form (i.e.,'Νόμος', '1ή' encoded as 'ΝΟΜΟΣ', '1Η'). English words were also mapped to a single word named 'ENGLISH_WORD'.

Moreover, we experimented with generic pre-trained 200-dimensional word embeddings (publicly available), trained with FASTTEXT [12] (https://fasttext.cc) and

[1] Datasets and Supplementary Material are published in https:/http://legislation.di.uoa.gr/publications under a non-commercial license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

based on a much larger corpus with Greek Wikipedia articles. The experimental results were worse, possibly because legal expressions are under-represented (or do not exist) in generic corpora (e.g., wikipedia or news articles) that were used, while also the preprocessing seems really poor based on our observation.

We also experimented with two different formats of token shape embeddings [13]. Chalkidis et al. [3] proposed 5-dimensional token shape embeddings that represent the following seven possible shapes of tokens: token consisting of alphabetic upper-case characters, possibly including periods and hyphens (e.g., 'ΠΡΟΕΔΡΟΣ', 'Π.Δ.', 'ΠΔ/ΤΟΣ'); token consisting of alphabetic lower-case characters, possibly including periods and hyphens (e.g., 'νόμος', 'ν.', 'υπερ-φόρτωση'); token with at least two characters, consisting of an alphabetic upper-case first character, followed by alphabetic lower-case characters, possibly including periods and hyphens (e.g., 'Δήμος', 'Αναπλ.'); token consisting of digits, possibly including periods and commas (e.g., '2009', '12,000', '1.1'); line break; any token containing only non-alphanumeric characters (e.g., '.', '€'); and any other token (e.g., '1o', 'OIK/88/4522', 'EU').

In addition and for comparison, we used 25-dimensional token shape embeddings by generating 1578 shapes for tokens by replacing each alphabetic lower-case and upper-case character with 'c' and 'C' respectively, each digit with 'd' and punctuation characters kept from the original token (i.e. 'Π.δ./τος', 'Αναπλ.', '1963' encoded as 'C.c./ccc', 'Ccccc.' 'dddd'). If the same character is encountered more than 4 times in a token in the row, it is limited to 4 times (e.g., '123456' is mapped into 'dddd'). Intuitively, having such representations for tokens offers more information regarding the context. In general, the shape (form) of its token relies on the existence and relative position of alphabetic characters, digits and punctuations.

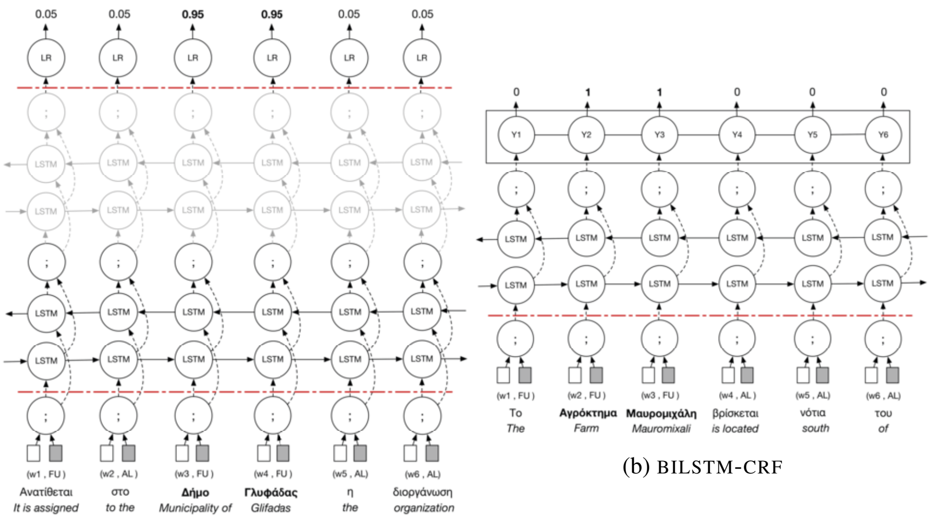
We were unable to embed the part-of-speech tag of each token due to the fact that so far there is no reliable POS tagger for the Greek language. We verified this by experimenting with the NLTK (`http://www.nltk.org`) POS tagger and, as also the one provided by CLTK (`http://cltk.org`), but both of them had a vast amount of wrong predictions, a fact that is even more profound in legal text.

2.2. LSTM-based methods

Until the recent advances in deep learning [14,15], NLP techniques were dominated by machine-learning methods that used linear models such as support vector machines or logistic regression. Currently, the focus of the community has switched from such linear models to multi-layer neural-network models. In this work, we experiment with Recurrent Neural Networks (RNNs), more specifically LSTM-based models that produce state-of-the-art results for language modeling [16,17], as well as for named entity recognition and part-of-speech tagging [18], all of which are sequence tagging tasks.

In this section, we describe the three LSTM-based methods we experimented with[2]. The LSTM (Long Short-Term Memory) [19] units are a more sophisticated form of the Simple Recurrent (RNN) units, tailored to resolve memory issues on sequential data. Each unit contains a self-connected memory cell and three multiplicative units: the input, output and forget gates, which provide continuous analogues of write, read and reset operations for the cells in each timestep (word). Further on, the LSTM layers are bidirectional in the sense that the input sequence is being processed from left to right and from right

[2]The methods were implemented using KERAS (`https://keras.io/`).



(b) BILSTM-CRF

(a) BILSTM(-BILSTM)-LR

Figure 1. LSTM-based methods

to left, conditioning information from the previous and next timesteps (tokens), respectively.

The first LSTM-based method that we have used, called BILSTM-LR (Figure 1a), uses a bidirectional LSTM (BILSTM) layer to convert the concatenated word and token shape embeddings of each token (lower ; nodes of Figure 1a) in each sentence to context-aware token embeddings (upper ; nodes), which better describe the semantics of each token given the specific task. Each context-aware token embedding is then passed on to the logistic regression (LR) layer (LR nodes of Figure 1a including the SOFTMAX activation) to estimate the probability that the corresponding token belongs to each of the examined categories (e.g., person, organization, etc.).

Based on experimentation, the pre-trained word embeddings are not updated during training on the labeled dataset, while in contrast the token shape embeddings are not pre-trained. The corresponding shape vectors are being learned during the actual training. We used Glorot initialization [20], binary cross-entropy loss, and the Adam optimizer [21] to train the BILSTM-LR recognizer with early stopping by examining the validation loss. Hyper-parameters were tuned by grid-searching the following sets, and selecting the values with the best validation loss: LSTM hidden units {100, 150}, batch size {16, 24, 32}, DROPOUT rate {0.4, 0.5}.

The second LSTM-based method, called BILSTM-BILSTM-LR, has an additional BILSTM layer (transparent upper LSTM nodes of Figure 1a) between the context-aware token embeddings (lower ; nodes) of the lower BILSTM layer, and the logistic regression (LR) layer (LR nodes). Stacking LSTM (or BILSTM) layers has been reported to improve efficacy in several natural language processing tasks [22] at the expense of increased computational cost. Our experimental results (presented in Section 2.3 below) show that the additional BILSTM chain of BILSTM-BILSTM-LR also leads to significant improvements.

In the third LSTM-based method, called BILSTM-CRF, we replace the logistic regression layer of the BILSTM-LR method with a linear-chain of Conditional Random Fields

(CRFs), as illustrated in the Figure 1b. CRFs [23] have been widely used in NLP sequence labeling tasks (e.g., POS tagging, named entity recognition). They have also shown exceptional results on top of BILSTMs in sequence labeling [18]. Hyper-parameter tuning and training are performed as in the previous LSTM-based methods.

2.3. Experiments and Evaluation

For each of the three methods we measured the performance on *precision*, *recall*, and F_1 scores based on the MUC guidelines [24][3]. Table 1 lists the results of this group of experiments based on the average of five individual runs for each method.

Entity Type	BILSTM-LR				BILSTM-CRF				BILSTM-BILSTM-LR			
	P	R	F1	F1'%	P	R	F1	F1'%	P	R	F1	F1'%
Person	0.95	0.86	0.90	-1%	0.94	0.91	0.92	-2%	0.94	0.92	**0.93**	-3%
Organization	0.90	0.71	0.79	-4%	0.87	0.72	0.79	-6%	0.92	0.76	**0.83**	-5%
GPE	0.90	0.80	0.85	-	0.90	0.76	0.83	+2%	0.92	0.85	**0.88**	-1%
GeoLandmark	0.88	0.77	0.83	-11%	0.83	0.80	0.81	-11%	0.93	0.86	**0.89**	-5%
Legislation Ref.	0.95	0.79	0.86	-3%	0.94	0.79	0.86	-7%	0.96	0.84	**0.90**	-5%
Public Document	0.82	0.69	0.75	+3%	0.76	0.69	0.73	-1%	0.88	0.76	**0.82**	-
Macro AVG	0.89	0.79	0.84	-1%	0.87	0.80	0.83	-2%	0.91	0.85	**0.88**	-2%

Table 1. Precision (P), Recall (R), and F_1 score. Best F_1 per entity type shown in bold font. The F_1 score show the performance using the 25-dimensional shapes, while $F_1'\%$ show the perfomance reduction using the 5-dimensional shapes for the seven predefined categories.

The results are highly competitive for all the examined methods. The best results, based on the macro-averaged F_1, are coming from BILSTM-BILSTM-LR (0.88), which indicates that the extra BILSTM layer, which deepens the model, expands its capacity by a significant margin, compared to BILSTM-LR (0.84) and BILSTM-CRF (0.83). Considering the generic FASTTEXT pre-trained embeddings, instead of our domain-specific ones, leads to a macro-averaged F_1 of 0.81 for the best reported method BILSTM-BILSTM-LR, especially in the latter four categories, in which domain knowledge matters the most (e.g., geographical aspects and codification of documents). Considering more expressive token shapes also seems to improve the performance of the examined model (by a factor of 2% in the case of the BILSTM-BILSTM-LR, 0.88 vs 0.86). Further on, we are going to rely on the BILSTM-BILSTM-LR classifier based on the fact that it outperforms the other models in every entity-type.

3. Entity Linking

As we already mentioned, the complexity of legal text and the particularities of the Greek language itself provide an additional challenge in our goal to link the identified references. Based on heuristic rules, we were able to segment and normalize entity references and proceed to the task of entity linking[4].

[3]MUC guidelines consider partial token overlaps between the gold annotations and the predicted entities (sequences of consecutive tokens that belong in the same class), given the correct (gold) class.

[4]Technical details have been stated in the assisting Supplementary Material document of the datasets.

3.1. A vocabulary for textual entity references

The first step towards linking entity references extracted by Nomothesia NER with the entities described in public open datasets is to represent those references using the RDF specification. The *legal text* of a document contains subdivisions (passagges of individual laws) that are defined as *LegalResourceSubdivisions* based on the Greek legislation ontology. Since some of those contain text, it is also possible to contain (*has_reference* to) a *Reference* to an entity (e.g., a law passage referring to a specific law that it modifies, or an organization). This reference is realized in an interval of characters. In other words, it *begins* and *ends* on specific sequential characters inside the text of the subdivision. This *Reference* most likely refers to (or, in another sense, is *relevant_for*) an *Entity*, which is probably described in open public datasets. Therefore, in our case, a *LegalResource-Subdivision* contains references to persons, administrative units and legal resources (e.g., laws, decisions etc.). The former description is depicted in Figure 2.

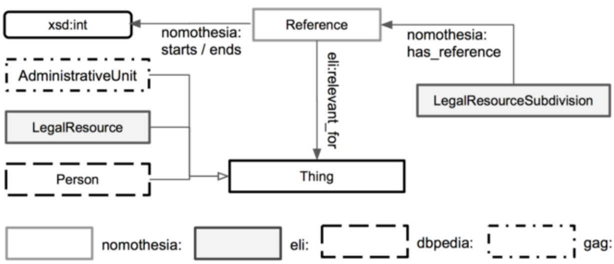


Figure 2. Textual Reference RDF Vocabulary

3.2. Linking entity references

We linked legal references with legal documents provided by the Greek legislation dataset[5]. We based on heuristic rules to directly interpret the relevant URI by capturing the *type*, *year* of publication and the *serial number*. We provide performance evaluation in Table 2.

3.3. Linking Greek Politicians and Greek Administrative Units using Silk

We linked person references with Greek politicians retrieved from the Greek DBpedia (`http://el.dbpedia.org/`) dataset and geopolitical entity references with the Greek administrative units, as they are described in the Greek Administrative Geography (GAG) dataset (published in `http://linkedopendata.gr/dataset`).

For both entity types, we proceed in interlinking the corresponding datasets using the Silk framework [10]. We experimented with two different textual linking operators: Levenshtein and substring distance [25] over the `rdfs:label` values provided by each dataset. For the case of the Greek Administrative Units, we also provided the *type* of the administrative units (e.g., local community, municipality, region, etc.) based on the naming conventions that we identified in the validation part of the labeled dataset.

[5]Published in `http://legislation.di.uoa.gr/legislation.n3`

3.4. Evaluation

For each interlinking method that we tried, we examine the performance of the interlinking in terms of *precision*, *recall*, and F_1 score *measured per entity pair* on the test part of our labeled dataset. Here, true positives (*TP*) are references correctly paired with an entity of each set, *false positives* (*FP*) are references incorrectly paired with entities, and *false negatives* (*FN*) are references incorrectly not paired with the relative entities of the examined sets. The acceptance threshold for both linking operators was tuned on the validation part of our datasets, while the entity pairs provided are those presented in the test part. Table 2 lists the results for this group of experiments.

METRICS	LINKING TECHNIQUE								
ENTITY	RULES			LEVENSHTEIN			SUBSTRING		
TYPE	P	R	F1	P	R	F1	P	R	F1
Person	-	-	-	0.99	0.55	0.71	0.90	0.68	**0.77**
GPE	-	-	-	0.99	0.79	0.88	0.95	0.92	**0.94**
Legislation Ref	0.99	0.97	**0.98**	-	-	-	-	-	-

Table 2. Precision (P), Recall (R), and F_1 score, *measured per entity pair*

Linking persons to Greek politicians was a great challenge, mainly because legislators tend to refer to a person's first name by its initials (e.g., 'A. Tsipras'), thus a fair amount of person references have been misclassified (precision: 0.71) for persons with the same surname. We successfully linked the geopolitical entities with the Greek administrative units (F_1: 0.92). Minor issues are related to the segmentation of compound references of multiple administrative units. The results for legislation references are robust (F_1: 0.98), while a short margin of documents are mis-linked due to the fact that ministerial decisions do not have a standard codification (nor a standard reference pattern), which varies from one ministry to another.

3.5. Greek geographical landmarks dataset generation

Greek geographical landmarks are a major asset for our legal recognizer since they are related to planning and architectural interests. However, there is no such public dataset to interlink between the references and the actual entities. We proceed in generating a new dataset by applying linguistic heuristics in order to form the entities and classify their type in 4 different abstract categories (classes):

Local District such as villages and small local communities.

Area sub-classified into *agricultural*, *forest*, *coastal* and *marine* areas.

Road sub-classified into *highway*, *local*, *bypass* roads.

Point of Interest such as a *farm*, an *islet*, or a *peninsula* which are commonly referred to urban planning legislation.

Further on, we interlink the new dataset with the Greek administrative units when there is a connection between them (*belongs_to*), indicated in terms of text (e.g., 'Beach Kavouri at Municipality of Varis-Voulas-Vouliagmenis').

3.6. *Querying the augmented Greek legislation datasets*

In this section we demonstrate new forms of querying the augmented Greek legislation dataset. The linking process expanded the Greek legislation dataset from approximately 2,9M triples to 4,4M triples in order to describe knowledge for 194,102 references of the supported entity types. Based on the above, we have the ability to pose queries against the resulting RDF graph using SPARQL (Table 3). Nomothesia platform also provides a SPARQL endpoint (http://legislation.di.uoa.gr/endpoint), as also a designated page for searching named entities (http://legislation.di.uoa.gr/entities).

<table>
<tr><td colspan="3">Q1: Retrieve any legal act that refer to municipalities, which belong to the regional unit of Florinas.</td></tr>
<tr><td>SELECT DISTINCT ?municipality_name (group_concat(?act;separator=", ") as ?act_ids)
WHERE {
?act eli:id_local ?act_id. ?act eli:has_part+ ?part.
?part lego:has_reference ?reference. ?reference eli:relevant_for ?gpe.
?gpe owl:sameAs ?municipality. ?municipality rdfs:label ?local_district_name.
?municipality a gag:Municipality. ?municipality lego:belongs_to ?reg_unit.
?reg_unit rdfs:label "REGIONAL UNIT OF FLORINA"@en.
} GROUP BY ?municipality_name LIMIT 5</td><td><u>Municipality</u>

“DIMOS PRESPON”
“DIMOS AMINTAIOU”
“DIMOS FLORINAS”</td><td><u>Act ID</u>

“leg:pd/1998/310, ...”
“leg:law/1997/2539, ...”
“leg:law/2013/4109, ...”</td></tr>
<tr><td colspan="3">Q2: Retrieve any legal acts that contain references to persons that have been born in Athens.</td></tr>
<tr><td>SELECT DISTINCT ?person_name
(group_concat(?act_id;separator=", ") as ?act_ids)
WHERE {
?act eli:id_local ?act_id. ?act lego:published_by ?signer.
?signer lego:relevant_for ?person. ?person a lego:Person.
?person owl:sameAs ?per. ?per rdfs:label ?person_name.
?per dbpedia-owl:birthPlace <http://el.dbpedia.org/resource/>.
} GROUP BY ?person_name LIMIT 5</td><td><u>Signer</u>

“ST. LABRINIDIS”
“G. GENNIMATAS”
“P. PIKRAMENOS”
“G. PAPAKONSTANTINOY”
“D. AVRAMOPOULOS”</td><td><u>Act ID</u>

“leg:dec/.3440_48844, ...”
“leg:law/1990/1874, ...”
“leg:la/1_14.06.2012, ...”
“leg:la/1_31.12.2009, ...”
“leg:dec/0546_.6551_55, ...”</td></tr>
<tr><td colspan="3">Q3: Retrieve any legal act and its written code that is being referred to in Law 2014/4261.</td></tr>
<tr><td>SELECT DISTINCT ?ref_act ?ref_code
WHERE {
<http://legislation.di.uoa.gr/eli/law/2014/4261> eli:has_part+ ?part.
?part lego:has_reference ?ref. ?ref lego:relevant_for ?ref_act.
{?ref_act a lego:PresidentialDecree} UNION {?ref_act a lego:Law}.
?ref lego:has_original_label ?ref_label
} LIMIT 5</td><td><u>Act URI</u>

leg:law/2017/3356
leg:law/2007/3606
leg:law/2008/3691
leg:law/1920/2190
leg:law/2006/3455</td><td><u>Act Code</u>

“ν. 3556/2007”
“ν. 3606/2007”
“ν. 3691/2008”
“κ.ν. 2190/1920”
“ν. 3455/2006”</td></tr>
</table>

Table 3. Entity-based queries

Conclusion and Future Work

We evaluated LSTM-based methods for named entity recognition in Greek legislation, demonstrating the effectiveness of such techniques in language-diverse environments. We introduced and applied a novel vocabulary for the representation of textual references in RDF. Further on, we evaluated entity-linking between textual references and entities from third-party datasets, while we generated a new dataset for Greek geo-landmarks.

Our future plans include the investigation of deeper and/or CNN-based architectures and also the use of ELMo embeddings [17], replacing the current feature representation. In relation to the entity linking, we will conduct experiments with character-level neural approaches as alternative textual linking operators. We also endeavour to extract more geospatial information to augment the newly-published Greek geo-landmarks dataset.

Acknowledgement

This research has been cofinanced by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-01000, project name: GeoREgC).

References

[1] K. Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017.

[2] M. Kim and R. Goebel, "Two-step cascaded textual entailment for legal bar exam question answering," in *Competition on Legal Inf. Extraction/Entailment*, (London, UK), 2017.

[3] I. Chalkidis and I. Androutsopoulos, "A deep learning approach to contract element extraction," in *Int. Conf. on Legal Knowledge and Inf. Systems*, (Luxembourg), pp. 155–164, 2017.

[4] S. N. Truong, N. L. Minh, K. Satoh, T. Satoshi, and A. Shimazu, "Single and multiple layer BI-LSTM-CRF for recognizing requisite and effectuation parts in legal texts," in *Int. Conf. on Artificial Intelligence and Law*, (London, UK), pp. 159–168, 2017.

[5] J. Dann, "European Legislation Identifier "ELI"," tech. rep., European Commission, 2014.

[6] M. V. Opijnen, "European Case Law Identifier: Indispensable Asset for Legal Inf. Retrieval," in *From Inf. to Knowledge* (M. A. Biasiotti and S. Faro, eds.), vol. 236 of *Frontiers in Artificial Intelligence and Applications*, pp. 91–103, IOS Press, 2011.

[7] R. Hoekstra, J. Breuker, M. Di Bello, and A. Boer, "Lkif core: Principled ontology development for the legal domain," in *Conf. on Law, Ontologies and the Semantic Web: Channelling the Legal Inf. Flood*, (Amsterdam, The Netherlands, The Netherlands), pp. 21–52, IOS Press, 2009.

[8] T. Athan, G. Governatori, M. Palmirani, A. Paschke, and A. Z. Wyner, "Legalruleml: Design principles and foundations," in *Reasoning Web. Web Logic Rules*, 2015.

[9] I. Chalkidis, C. Nikolaou, P. Soursos, and M. Koubarakis, "Modeling and querying greek legislation using semantic web technologies," in *European Semantic Web Conf.*, (Portorož, Slovenia), pp. 591–606, 2017.

[10] C. B. Anja Jentzsch, Robert Isele, "Silk - generating rdf links while publishing or consuming linked data," in *Int. Semantic Web Conf.*, (Shanghai, China), 2010.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Int. Conf. on Neural Inf. Proc. Systems*, (Stateline, NV), 2013.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword inf.," *arXiv preprint arXiv:1607.04606*, 2016.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Proc. :An Introduction to Natural Language Proc., Computational Linguistics, and Speech Recognition (3rd ed. draft)*. 2018.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] Y. Goldberg, *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers, 2017.

[16] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," *CoRR*, vol. abs/1708.02182, 2017.

[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Conf. of NAn Chapter of the Association for Computational Linguistics*, (New Orleans, Louisiana, USA), 2018.

[18] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-cNNs-CRF," in *Annual Meeting of the Association for Computational Linguistics*, (Berlin, Germany), pp. 1064–1074, 2016.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Int. Conf. on Artificial Intelligence and Statistics*, (Sardinia, Italy), pp. 249–256, 2010.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations*, (San Diego, CA), 2015.

[22] O. Irsoy and C. Cardie, "Deep recursive neural networks for compositionality in language," in *Int. Conf. on Neural Inf. Proc. Systems*, (Montreal, Canada), pp. 2096–2104, 2014.

[23] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Int. Conf. on Machine Learning*, (Williamstown, MA), pp. 282–289, 2001.

[24] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[25] G. Stoilos, G. Stamou, and S. Kollias, "A string metric for ontology alignment," in *Int. Semantic Web Conf.*, (Galway, Ireland), pp. 624–637, 2005.