Legal Definition Annotation in EU Legislation using Symbolic AI

Muhammad Asif [1, 2][0000-0001-9580-5626] and Monica Palmirani[1][0000-0002-8557-8084]

[1] CIRSFID, University of Bologna, Bologna, Italy
[2] Department of Computer Science, University of Luxembourg, Luxembourg
{muhammad.asif19, monica.palmirani}@unibo.it

Abstract. The definition is an integral component of the legislation. It is essential to the adequacy and legitimacy of legislation. In legislation, legal definitions are used for clarity, consistency, and legitimate certainty, but also for creating new legal concepts (e.g., personal data) or crimes (e.g. stalking, mobbing). With the advancement in society with technological innovation, the significance of accurate and precise definitions in legislation is indeed more articulated. In order to avoid ambiguity and to ensure, as far as possible, a strict interpretation of Law, Legal Texts (LT) usually define the specific lexical terms used within their discourse by means of normative rules. Due to the continuous increase of LT and a large number of domain-specific rules, extracting these definitions from the LT would be costly and time-consuming if it's done humanely. Definition extraction is widely used in Legal domains to perform legal and compliance analysis. In this paper, we detect and annotate the legal definitions using Symbolic Artificial Intelligence (AI) based on Natural Language Processing (NLP) and fostering LegalXML annotation. The goal is to qualify a very valuable part of legislation for supporting further AI applications also in the judiciary domain. The detection and annotation of definitions are performed on the delimiting type of definitions. The dataset consists of EU Legislation in the span of time from 2010 to 2021 in Akoma Ntoso[1] (AKN) file format. The resultant 15082 AKN files are annotated.

Keywords: Definition Extraction, Definition Annotation, Symbolic AI.

1 Introduction

In recent years, the amount of data has been growing exponentially. Traditional tools and techniques are not good enough to process, refine and extract necessary information from it. To process this huge amount of data, the concept of Bigdata is introduced [1]. This data is either in structured (credit card numbers, dates, geolocation, address, stock information,) or unstructured (spreadsheets, emails, text files, surveillance footage, survey reports, and machine-generated formats) form. Text data is the best example of unstructured data. Text data is generally analyzed to extract patterns to take action [2].

---

1 Akoma Ntoso OASIS LegalDocML XML Standard. http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html

Due to the huge amount of unstructured data in text form, makes it practically impossible to extract structured content and patterns manually. Several attempts are witnessed for the extraction of the required contents using automated channels [2]. Most notable works were done in the domain of Data/Text Mining [3][4] and Computational Linguistics [5]. In text mining, Information Retrieval techniques are used to extract and preprocess information [6]. Information Extraction (IE) is one of the effective techniques to extract meaningful information from text. However, it is considered one of the challenging tasks in the domain of NLP [7]. In IE, some efforts target domain-specific problems to deal appropriately with their inherent qualities. One such domain-specific technique is Definition Extraction (DE) [8].

DE is a process of identifying and extracting definitions from text [9]. Different automated and semi-automated techniques exist to extract definitions from the text [10]. These automated and semi-automated techniques of DE significantly reduced the efforts and time required in manual extraction. Over the last few years, researchers [11-13] have worked significantly aiming to achieve DE with varying degrees of success. DE is useful in many domains like question answering systems, ontology engineering, developing e-learning applications, and lexicon or glossary construction. DE is becoming a hot topic in Law. In law, DE is performed on from LT [14-15]. Most of the work that is witnessed in DE is based on statistics. A very few research works are carried out on DE on legal aspects. So, there is a need to carry out research on the extraction of definitions, based on the legal aspects.

2 Literature Review

Research studies are witnessed on IE by using different techniques of NLP, Deep Learning (DL) and Knowledge-Based Extraction (NBE). It is used in various applications, i.e., machine translation, question-answering systems, automatic summarization, ontology engineering, text classification and dictionary construction. Assorted studies have been witnessed on data extraction and DE from legal documents.

DE [16] was performed on the BCTL (Brazilian Collection of Telecommunication Laws) dataset which consists of 1940 reference documents. Punkt algorithm was used for the segmentation of paragraphs. The 6120832 tokens have been created that consists of words and punctuations. Brill tagger was used for Part of speech tagging and it shows 90.44% accuracy. For DE the dataset was split into training (70%) and testing (30%) and during evaluation, it shows 73.50% extraction was performed correctly.

In [17] detection of definitions from Federal administrative documents of the Swiss Confederation was performed. After extraction, the documents were converted into XML from HTML (HyperText Markup Language) format. At the auto-detection of Enumerated and Bracketed definitions model performed 99% and 94% precision respectively. In [18] a model was designed to help and assist legal researchers and lawyers in accessing legal data from most applied cases.

A Machine Learning based approach was presented [13] for DE. World-Class Lattices (WCL) dataset was used with 4718 manually annotated sentences. Long Short-Term Memory was implemented using Part of Speech tagging it shows 90.70% F-score in validation. The Trained model has a good understanding of the relation between the words in the sentence. However, when the results were compared with the state-of-the-art, the results were not good. Whereas the BERT outperforms with a 97.4% F-score at DE. WCL was also used in [11], the BERT performs 85.30% in terms of F-score.

3 Methodology

For the annotation, this study used the dataset that consists of EU Legislation for Agri-Food in the span of time from 2010 to 2021 in AKN format that has 15082 documents. A NLP technique is used for auto-detection and annotation of legal. The annotation is performed on a delimiting type of definition, which is a very famous type of definition to define norms in legislation. For this purpose, in Python programming language, the Element Tree (ET) library is used to handle the AKN files, and a tree is created using Element Tree to handle them easily. In legal documents (AKN format), first, we must find the heading tag (τ_H) having targeted keywords, “definition” or “definitions” ($k_{Definition}$ or $k_{Definitions}$). After finding the keyword enter τ_H and find the paragraph tag (τ_ρ). After finding τ_ρ enter body of τ_ρ and find the keyword “means” (k_{means}) being followed by some quoted text. Store this quoted text ($Q_{Definiendum}$). Annotate $Q_{Definiendum}$ as per rules defined in equation 1, and then replace it with new quoted text ($Q'_{Definiendum}$).

$$Q'_{Definiendum} = < \mathrm{def} \ \ \mathrm{eId} = \mathrm{def} \ \ \_\mathrm{i} > Q_{Definiendum} </\mathrm{def}> \quad (1)$$

Next, store the rest of the text in the body of τ_ρ i.e., T_{Next}. Annotate T_{Next} as per the rules defined in equation 2 and then replace T_{Next} with T'_{Next}.

$$T'_{Next} = < \mathrm{defBody} \ \mathrm{eId} = {}''\mathrm{defBody}_{\_\mathrm{i}} > T_{Next} </\mathrm{defBody}> \quad (2)$$

Where i is the unique numeric identifier that is assigned in this annotation, and it is the same in both equations.

After detecting the equations patterns the annotation is done as explained above in equations 1 and 2. The found quoted text which is the definiendum is tagged as the definition heading. The definiendum is marked with a start tag i.e. “<def …>” and an end tag i.e. </def>. Rest i.e. the text that follows “means” and ends at either “,” or “;” is marked as the body of the definition that is annotated with start tag <defBody …> and end tag </defBody>. All siblings and children are annotated in the same way. After the annotation of the definiendum (def) and Rest (defBody), a unique (eId`s) is assigned to all the annotated definiendum and rest. All the annotated material (outputs) are exported using the Pretty XML library. The camelCase [19] is used to avoid inconsistencies in format. The indentation method is used for the validation of AKN files.

4 Results Discussions

In legislation, annotation is a process to provide explanatory remarks, comments, notes, or interpretations regarding legal texts. The annotation of legislation aims to explain the legislative clauses and clarify the meaning of legislation to assist the readers to understand legal implications. The annotation of legislation is used in various forms i.e., definition annotation, legislative intent, case & cross references etc. This study performs annotation of definitions using Symbolic AI as per rules explained above. Manual annotation is also performed to compare the results with auto-annotation.

For the manual annotation, first the guidelines are prepared. After the guideline's preparation, the services of three volunteers were acquired, V1, V2 and V3. The volunteers were well-educated and were familiar with the definition in legislation and the concepts of definition annotation and had a good grip on the law. Initially they were introduced with the annotation process. After the training, the dataset (files having definitions to be annotated) was given to the first two annotators—annotator V1 and V2. Once they completed the annotation, a short meeting was arranged to resolve conflicts by involving the third annotator too. After training and conflict resolution, 5 same files were given to all three annotators for annotation. After manual annotation, the inter-annotator agreement is computed between them, using Kappa statics, formula is shown in Equation 3. The p_o is the observed agreement between the annotators and p_e is the expected agreement of random judgements.

$$k = ((p_o - p_e))/((1 - p_e)) \tag{3}$$

According to Cohen`s Kappa, if the agreement between the annotators is '1' it shows perfect agreement is found between the annotators. If the score is between 0.81-0.99 shows a nearly perfect agreement [20]. The total number of definitions to be annotated in all 5 files that were given to the annotators for manual annotation was 70. Annotator V1 annotated 69 files and missed one definition from the 2nd file. The annotator V2 annotated 69 definitions and missed one definition from the 4th file and the annotator V3 also missed one definition from the 3rd file, and he annotated 69 definitions successfully. The inter-annotator agreement between all three annotators is 0.96%, which means performed annotation is nearly Perfect. Shown in equation 4.

$$k = ((70 * 67 - 70))/((70 * 70 - 70)) = 0.96 \tag{4}$$

To compute the results of annotation, the auto-annotation was compared with the best manual annotated (BMA) files. The count of best manual annotation was 70 from 5 selected files, whereas, in the results of auto annotation, the number of annotated definitions was also 70. The Calculated value of the Kappa coefficient is 1 between the best manual annotation and auto-annotation, which means performed auto-annotation is perfect. Comparison among the performance of different annotators and auto-annotation can be seen in Table 1.

Table 1. Results of Manual Annotation

file #	V1	V2	V3	BMA	Auto-Annotation
32010L0024	5	5	5	5	5
32010L0030	10	11	11	11	11
32010L0031	19	19	18	19	19
32010L0035	25	24	25	25	25
32010L0043	10	10	10	10	10

As the result of annotation, the 899 files have found the annotated material, the remaining 14183 AKN files, were not annotated. The reason behind this is that the remaining files, in spite of being in the standard format, were not having the required tags. Hence these files are skipped for further processing.

5 Conclusions

The EU and FAO are working in collaboration to preserve food to overcome hunger issues worldwide. EU makes policies regarding food safety. Some inconsistent norms exist in EU Legislation, to detect the inconsistencies in norms and to inline the policies, there is a need to extract and annotate definitions. This research is carried out on annotation of definitions using Symbolic AI. Every annotation carries a definition heading and definition-body. For the purpose, the dataset consists of files of EU legislation of the years 2010 to 2021. To perform the annotation of the definitions and their subparts, the algorithm is designed and implemented in Python programming language with the library named Element Tree (which is used to handle the XML files). First, the targeted pattern is found using rule-based mining. After the detection of the targeted pattern, the heading and the body of the definitions are annotated with different tags like "def" and "defBody" etc. The annotated information is listed and finally, the AKN is validated using the indentation technique. A total of 899 files have found the annotated material and successfully annotated it. To verify the results of annotation, 5 files are manually annotated by the three volunteer annotators. After performing the manual annotation, the inter-annotator agreement is counted by applying Cohen's kappa and a 0.96 score is found, which shows a "Nearly Prefect Agreement" between the three annotators. After performing the manual annotation, the best-annotated files are separated to compare with the auto-annotated files. When these files are compared, the perfect agreement was found between them which shows the performed annotation is perfect.

Acknowledgments. This study is funded by PON grants of the Italian Government and also the ERC HyperModeLex.

References

[1] M. A. Qureshi *et al.*, "A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis," *CMC-COMPUTERS MATERIALS \& CONTINUA*, vol. 70, no. 3, pp. 4987–5004, 2022.

[2] B. Oral and G. Eryi\ugit, "Fusion of visual representations for multimodal information extraction from unstructured transactional documents," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–19, 2022.

[3] E. Senave, M. J. Jans, and R. P. Srivastava, "The application of text mining in accounting," *International Journal of Accounting Information Systems*, vol. 50, p. 100624, Sep. 2023, doi: 10.1016/j.accinf.2023.100624.

[4] N. Farzana, P. Mohammed1, F. R. Shaikh2, T. Priya3, K. Khan4, and S. Jamadar5, "Obtaining and Analyzing Data from Texts," vol. 12, p. 2023.

[5] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," *NPJ Comput Mater*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41524-022-00784-w.

[6] A. Kumar, V. Dabas, and P. Hooda, "Text classification algorithms for mining unstructured data: a SWOT analysis," *International Journal of Information Technology*, vol. 12, no. 4, pp. 1159–1169, 2020.

[7] S. Singh, "Natural language processing for information extraction," *arXiv preprint arXiv:1807.02383*, 2018.

[8] D. Toluhi, R. Schmidt, and B. Parsia, "Concept Description and Definition Extraction for the ANEMONE System," in *Engineering Multi-Agent Systems: 9th International Workshop, EMAS 2021, Virtual Event, May 3--4, 2021, Revised Selected Papers*, 2022, pp. 352–372.

[9] N. Gardner, H. Khan, and C.-C. Hung, "Definition modeling: literature review and dataset analysis," *Applied Computing and Intelligence*, vol. 2, no. 1, pp. 83–98, 2022, doi: 10.3934/aci.2022005.

[10] A. Zaki-Ismail, M. Osama, M. Abdelrazek, J. Grundy, and A. Ibrahim, "RCM-extractor: an automated NLP-based approach for extracting a semi formal representation model from natural language requirements," *Automated Software Engineering*, vol. 29, no. 1, pp. 1–33, 2022.

[11] A. P. Ben Veyseh, F. Dernoncourt, D. Dou, and T. H. Nguyen, "A joint model for definition extraction with syntactic connection and semantic consistency," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 9098–9105, 2020, doi: 10.1609/aaai.v34i05.6444.

[12] L. Claassen *et al.*, "Cold brew coffee—Pilot studies on definition, extraction, consumer preference, chemical characterization and microbiological hazards," *Foods*, vol. 10, no. 4, p. 865, 2021.

[13] P. Kumar, A. Singh, P. Kumar, and C. Kumar, "An explainable machine learning approach for definition extraction," in *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, 2020, pp. 145–155.

[14] A. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: a survey," *Artif Intell Rev*, vol. 51, no. 3, pp. 371–402, 2019, doi: 10.1007/s10462-017-9566-2.

[15] C. Niculi\ct\ua and L. Dumitriu, "The Relational Parts of Speech in Text Analysis for Definition Detection, for Romanian Language," in *2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 2019, pp. 1–6.

[16] E. Ferneda, H. A. do Prado, A. H. Batista, and M. S. Pinheiro, "Extracting definitions from brazilian legal texts," in *International Conference on Computational Science and Its Applications*, 2012, pp. 631–646.

[17] S. Höfler, A. Bünzli, and K. Sugisaki, "Detecting legal definitions for automated style checking in draft laws," *Technical Reports in Computational Linguistics*, no. CL-2011.01, 2011.

[18] T. Padayachy, B. Scholtz, and J. Wesson, "An information extraction model using a graph database to recommend the most applied case," in *2018 International Conference on Computing, Electronics \& Communications Engineering (iCCECE)*, 2018, pp. 89–94.

[19] L. Weissweiler, V. Hofmann, M. J. Sabet, and H. Schütze, "CaMEL: Case Marker Extraction without Labels," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.10010

[20] F. Dai, M. Leach, A. M. Macrae, M. Minero, and E. D. Costa, "Does thirty-minute standardised training improve the inter-observer reliability of the horse grimace scale (HGS)? A case study," *Animals*, vol. 10, no. 5, May 2020, doi: 10.3390/ani10050781.