Topic Similarity of Heterogeneous Legal Sources Supporting the Legislative Process

Michele Corazza[1,*], Leonardo Zilli[1] and Monica Palmirani[1,*]

[1]University of Bologna, ALMA-AI, via Galliera 3, Bologna, Italy

Abstract

The legislative process starts with a deep analysis of the existing regulations at European and national levels to avoid conflicts and fostering the into force norms. Also the Constitutional Court decisions play a fundamental role in this analysis for checking the compliance with the constitutional framework and for including the inputs coming from this relevant court in the law-making process. Finally, it is also significant to compare the forthcoming proposal with the already presented bills regarding the same topic. This comparison is crucial to avoid overlapping and to coordinate the democratic dialogue with the different parties. In this light, this paper presents an unsupervised approach for calculating similarity between heterogeneous documents annotated in Akoma Ntoso XML, with the aim to support the information retrieval of similar documents using thematic taxonomy used in legal domain. The prototype has been developed for answering to a call for manifestation of interests launched by the Chamber of Deputy of Italy in order to adopt hybrid AI in the legislation process. It uses a completely unsupervised approach based on Sentence Transformers, meaning that neither annotated data or any fine-tuning process is required.

Keywords

Unsupervised learning, Sentence Transformers, Hybrid AI, Legal NLP

1. Introduction

The legislative process inside parliaments and official assemblies includes an initial phase of preliminary discovery of the existing regulations and rules in the same domain of the proposal, in order to synchronize the new bill with the legal system and to avoid conflicting norms. Secondly, a legal preliminary study must be conducted for applying legislative drafting techniques that have the aim of creating transparent and evidence-based legislation (e.g., Better Regulation https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en).
On the other hand, the fragmentation of the legal system imposes the task of an accurate preliminary legal analysis and research at different levels of legislation to the legislative department: at the European level in order to discover the norms in Regulations and Directives; at the national level to avoid overlapping with other existing acts; at the ministerial level to synchronize the technical and operative rules. Notably, it is crucial to check the decisions of the Constitutional Court to avoid to produce norms that are unconstitutional. On the other hand, the legal sources, considering their heterogeneous nature, follow some theory of law principles: i) *lex superior derogat inferiori*, following a specific hierarchy between the legal sources (e.g., an EU regulation is directly enforceable in the Member States); ii) *lex specialis derogat legi generali* (e.g., energy regulation overrides general green deal rules); iii) *lex posterior derogat legi priori* (e.g., the norms should be applied according to the principle of point-in-time with respect to the temporal model of the provisions and facts, the normative references in the preamble are static links fixed in time when the Parliament argues on the justification reasons). Another important check is done with the existing bills already proposed in the assembly to better manage the democratic dialogue between different parties' propositions. For this reason, having a dashboard that, in a unique portal, allows the retrieval, comparison, analysis of different heterogeneous legal sources is a fundamental instrument for this preliminary legal analysis. The documents are annotated in Akoma Ntoso XML [1] for creating a common framework for their representation that is capable of capturing the legal knowledge and metadata (e.g., jurisdiction, hierarchy, temporal model).

Additionally, we provide an unsupervised approach for classifying legal documents according to their topic, which is used to retrieve the relevant legal documents concerning some main legal topics (e.g., the subject of the Chamber of Deputies Committees defined by law [1], or EUROVOC top-level thematic classes) from a user input. This work was conducted on the use-case of the Chamber of Deputy of Italy's needs and documents, answering the

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding authors.
michele.corazza@unibo.it (M. Corazza); leonardo.zilli@studio.unibo.it (L. Zilli); monica.palmirani@unibo.it (M. Palmirani)
0000-0002-7288-6635 (M. Corazza); 0009-0007-4127-4875 (L. Zilli); 0000-0002-8557-8084 (M. Palmirani)

[1] https://temi.camera.it/leg19/aree.html

call for interests launched in February 2024 concerning the use of AI in Parliament [2].

The legislative language is a peculiar language that includes qualified part of the text like the preamble, normative part, definitions, normative references, exceptions, transitional norms, etc. For this reason, the task is not trivial and should take in consideration these peculiarities.

2. Related Work

The creation of models and methods for the legal domain is a challenging endeavour, as this field is characterized by some peculiar aspects that might lead general-purpose approaches to be inaccurate. Nevertheless, a multitude of different models and strategies have been proposed in this field, including models that have been trained specifically on this domain like LEGAL-BERT[2], which was fine-tuned from BERT[3] on legislative documents from the UK, US and EU, court documents from the European Court of Justice. Another model, called custom LEGAL-BERT[4] was instead trained on a corpus comprised entirely of Case Law from the Harvard Law Library. Another prominent example of ad-hoc models for the legal domain is called Pile-of-Law (PoL), from the name of the dataset that was used to fine-tune it, which comprises data from 35 different sources in English [5]. Interestingly, in terms of natural language processing applications for the legal domain, most approaches appear to be targeted at the judiciary rather than the legislative branch. Additionally, some approaches include common-law corpora (UK/US) that for our purpose (EU) could create relevant distortions in the dataset. In particular, a common task is the prediction of a judgment for a given case. This task has been attempted using multiple methods, including using a consistency graph and a transformer model to determine which articles have been violated in a given case [6]. The research is not limited to the English language, as there are contributions for Chinese court judgments [7] and rulings from the Indian Supreme Court [8].

Another crucial aspect of research in the wider field of legal informatics is the creation of formats, ontologies and tools that support the machine-readable representation of legal documents, from both the legislative and judiciary branches. Among these, one of the founding elements of our approach is the usage of the Akoma Ntoso XML standard [1, 9], which has been adopted by many international institutions [10, 11, 12, 13, 14] to represent legal documents. This standard allows the annotation of legal definitions, references, the hierarchical structure of legal documents, as well as the temporal aspects of legal documents.

3. Datasets and resources

The documents used for the project have been collected from different sources, resulting in four distinct datasets:

- *Corte Costituzionale*: Contains the orders and judgments of the Italian constitutional court, spanning from 1956 to 2018 (10725 documents), which have been downloaded and converted to Akoma Ntoso using an ad-hoc tool [3];
- *Progetti di Legge (PDL)*: A collection of Italian legislative bills from the legislatures XVIII and XIX (March 2018 to May 2024 - 3615 documents), extracted from the official website of the Chamber of Deputies of the Italian Parliament[4] in the HTML format and converted to Akoma Ntoso using a batch python parser[5].
- *EUR-Lex*: A collection of Regulations and Directives from the European Union, spanning from 2010 to 2021, extracted from the EUR-Lex website[6] and converted from Formex to the Akoma Ntoso format using our conversion tool [7].
- *Normattiva*: A collection of Italian legislative acts extracted from the Normattiva portal[8], which contains all legislative documents from the Italian parliament in Akoma Ntoso format. The documents from 2010 to May 2024 were selected, including Primary and Secondary Law.

When not already in the Akoma Ntoso XML format, as is the case for the PDL and Eur-Lex dataset, the documents have been converted to this format. Through this conversion, it is possible for us to extract portions of the document according to its hierarchical structure (articles, commas, lists, etc). This structural information is very important for the legal domain, as it allows to chunk documents while considering their structure (e.g., legal definitions, article, list of points). Furthermore, normative references are also annotated as such, and a unique URI is used to indicate them. The Akoma Ntoso standard also follows the FRBR conceptual model, which is used to distinguish between works (i.e a specific law), expressions (the various consolidated versions of each law that have been amended over time) and manifestations (the physical embodiment of an expression or work). Through the annotation of the hierarchical structure of documents, the references and the URI naming convention based on FRBR it is possible to resolve normative references, even when they refer to a part of a document, like a single article or paragraph. Furthermore, the FRBR

[2] https://comunicazione.camera.it/archivio-prima-pagina/19-37666

[3] https://gitlab.com/CIRSFID/cortecostituzionale-py

[4] https://www.camera.it/

[5] https://gitlab.com/CIRSFID/html2aknPDL

[6] https://eur-lex.europa.eu

[7] http://u2.cirsfid.unibo.it/formexplus2akn/frontend/

[8] https://www.normattiva.it/

model allows us to retrieve the consolidated version of a document which is temporally relevant for a given reference. Akoma Ntoso also includes legal metadata (e.g., jurisdiction, temporal information, modifications, definitions, law-making process, life-cycle of the document, classification) which improves the expressiveness of legal knowledge in the XML representation.

Each dataset follows semantically descriptive naming conventions for the documents, which facilitate subsequent data handling and processing steps in the pipeline of the project. Table 1 summarizes the number of documents contained in each dataset.

Dataset	N. of Documents
Corte Costituzionale	10725
PDL	3615
EUR-Lex	14305
Normattiva	3195

Table 1
Number of documents in each dataset

In order to deal with the highly heterogeneous nature of the datasets, labels describing a number of various topics have been used for categorizing the documents. The documents concerning Italy have been classified according to the labels of the Committees of the Chamber of Deputies. These Committees are represented as a string describing them, which contains their titles (shown in Table 2), as well as their description as presented in the Circolare del Presidente della Camera (16 ottobre 1996, n. 3), the official document that regulates the matters of competence for each of them. Only regarding the dataset of the Constitutional Court, the *"Giustizia"* (Justice) and *"Affari costituzionali, della Presidenza del consiglio e interni della Camera dei deputati"* (Constitutional Affairs, Presidency of the Council and Internal Affairs of the Chamber of Deputies) commissions were excluded as they apply to the vast majority of Constitutional Court documents.

Concerning the EUR-Lex dataset, the classification leveraged the European multilingual thesaurus, EuroVoc, using the top-level terms (shown in Table 3) and their immediate subcategories separated by semicolons. As for the Constitutional Court, the term *"Unione Europea"* (European Union) has been excluded as it is too general and relevant to all documents in the dataset.

Affari esteri e comunitari
Difesa
Bilancio, tesoro e programmazione
Finanze
Cultura, scienza ed istruzione
Ambiente, territorio e lavori pubblici
Trasporti, poste e telecomunicazioni
Attività produttive, commercio e turismo
Lavoro pubblico e privato
Affari sociali
Agricoltura
Politiche dell'Unione Europea

Table 2
Italian Chamber Committees

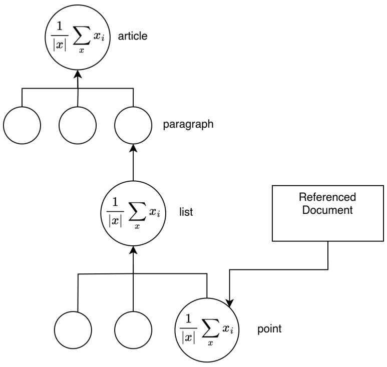
Vita politica
Relazioni internazionali
Diritto
Economia
Scambi economici e commerciali
Finanze
Questioni sociali
Istruzione e comunicazione
Scienze
Impresa e concorrenza
Occupazione e lavoro
Trasporto
Ambiente
Agricoltura, silvicoltura e pesca
Agroalimentare
Produzione, tecnologia e ricerca
Energia
Industria
Geografia
Organizzazioni internazionali

Table 3
Top level EuroVoc terms

4. Document Classification

In order to classify documents according to their content, we used an approach based on the SentenceTransformers library [15], and selected the multilingual model "paraphrase-multilingual-mpnet-base-v2"[16]. This model is made multilingual from the monolingual Sentence Transformer model "paraphrase-mpnet-base-v2", in turn based on MPNet [17], which was trained using a contrastive loss and an approach similar to siamese networks to allow the direct application of a metric (cosine similarity) to its output vectors in order to measure the semantic proximity of sentences. The monolingual model is then used as a teacher in a teacher-student configuration to train the multilingual one so that both the original and translated versions of sentences have the same vector representation in the new model. The chosen model, in particular, was trained on parallel data and supports 50+ languages, including Italian and English. Crucially, the usage of a sentence transformer allows us to operate in a completely unsupervised way, without the need to use annotated data or to fine-tune the model for the classification task, since we can directly apply cosine similarity to measure semantic relatedness.

In order to produce a classification of the documents, we selected two components of the normative documents (Eur-Lex, Normattiva, PDL), namely their titles and articles. For the Corte Costituzionale dataset, we selected



$\frac{1}{|x|}\sum_x x_i$
article
paragraph
$\frac{1}{|x|}\sum_x x_i$
list
Referenced Document
$\frac{1}{|x|}\sum_x x_i$
point

Figure 1: A graphical visualization of the aggregation strategy used to obtain a vector representation for each article. In this example, the article is composed of multiple paragraphs, one of them contains a list and one point of the list contains a normative reference. The reference is resolved and aggregated with the relevant point, then the procedure leverages the structure of the document to produce element vectors from their children, until the root of the tree (article).

the introduction as a substitute for the title, while instead of the articles we used the decision portion of the documents, in addition to all textual content between parenthesis, which contains brief descriptions of referenced documents. The text between parenthesis is fed to the model and the results are averaged to produce a single vector. In the following sections, we use "titles" and "articles" for brevity, but these correspond to introduction and decision + parenthesis for the Corte Costituzionale dataset.

These components were extracted by applying the appropriate Xpath query to the Akoma Ntoso XML tree representing each document. The first step is to compute embeddings representing each title of the document. Then, we proceed to compute the article vectors. While in the case of titles we can just apply the sentence transformer directly to the text, the length of articles might prevent the model from producing accurate result, or even exceed the maximum allowed tokens for a given model. For this reason, our approach leverages the structure of articles, represented using Akoma Ntoso, to produce one embedding for each article. In particular, we proceed traversing the XML tree in a recursive manner, until we reach the XML elements that are leaves of the tree. We exclude the elements that appear inline in the text (eg dates, references, etc) in order to maintain the textual content of each leaf node (eg paragraph, item of a list, etc) intact. A visualization of the procedure is shown in Figure 1. In addition to its own textual content, each leaf node is associated with a list of the references in its text, which are resolved as follows:

- For punctual references (eg Article 3 of Regulation xx/yyyy/EU) we obtain the specific referenced portion of the document as an XML element;
- For generic references to an entire document (eg Regulation xx/yyyy/EU) we use the title and first article of the document to represent it.

Formally, then, an article a having children and references is represented by an embedding obtained from the model M using the following recursive procedure:

$$v(a) = \frac{1}{2 + |c(a)|}\left(M(t(a)) + \sum_i v(c_i(a)) + \frac{1}{r(a)}\sum_j R(r_j(a))\right) \tag{1}$$

Where:

- $t(a)$ is the textual content of the article which is not included in any of its non inline children;
- $c(a), c_i(a)$ represent the set of all non inline children of a and the i-th child element of a, respectively;
- $r(a), r_j(a)$ represent all the references in the text of the article, and the j-th reference in the text, respectively.

In order to represent references, then, we can define a function R that works as follows:

$$R(i) = \begin{cases} v(i) & \text{if } i \text{ is a punctual reference} \\ \frac{1}{2}M(T(i)) + v(A_1(i)) & \text{otherwise} \end{cases} \tag{2}$$

Where $T(i)$ represent the title of the referenced document, while $A_1(i)$ is the first article of the document. Overall, the function $v(a)$ as defined previously computes an average vector representation for each article, which aggregates the embeddings of all its children but also considers the normative references contained in the text.

Once we obtained the vector representation of each article of each document and its titles embeddings, we can compare them with the vector representations of our topics, the EuroVoc terms for the European legislation and the Chamber commissions for the Italian documents. Then, the similarity between each document and the subjects is derived from the sum of the cosine similarity between its title and the average similarity between the topics and each article. Finally, the maximum similarity value obtained by this procedure is used to classify each document using one of the topics.

5. Searching by topic

In order to provide a topic-based search that can be used in the Italian legislative process, the final step is to provide an interface to query each of the four datasets, by providing information about the more relative topic for a given query. Our approach is based on the possibility to input an arbitrary textual input, as well as a set of keywords that are relevant to what the user is interested in. Before any further processing, the keywords are separated by a semicolon ";" and encoded as a single string. In order to obtain a vector representation of the user inputs, we can then use the model to obtain a vector representation from the arbitrary input, as well as the semicolon-separated keywords. Then, the two vectors are averaged and used in all further processing, obtaining the query vector.

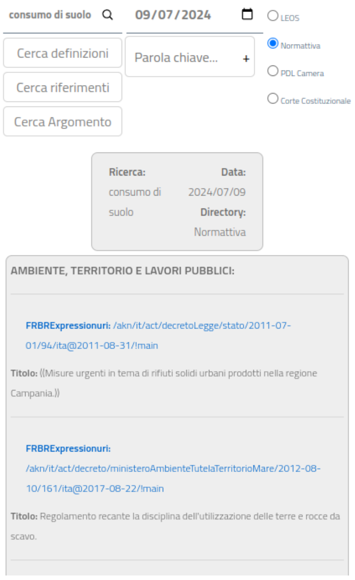
The first step of the topic-based search is the comparison between the topic list (the EuroVoc terms or the Camera commissions, according to the selected dataset) which returns the two most similar subject in terms of cosine similarity with the query vector. For these two topics, the system then computes the similarity of the query vector with each of the documents that have been classified with the specified subject.

The system described in this article is available on a website[9] which includes multiple tools for legal drafting in the context of a call from the Italian Camera dei Deputati expression of interest. The system is available under "Cerca", followed by "Ricerca Avanzata" on the panel that appears on the right, and finally by inserting the query and keyrords, followed by the "Cerca Argomento" button. An example of the layout and results of this system is shown in Figure 2. Additionally, we allow users to select a date when querying the system, meaning that only documents and consolidated versions that were in vigour at a specific time. This is a crucial feature for the legal domain, where a judge might need to know which laws were in vigour when an alleged crime was committed.

6. Evaluation and Results

In order to evaluate the performance of our subject-based classification, we asked three experts of the legal domain to annotate 100 random documents for each dataset between them, and proceeded to measure the accuracy of our classification when compared to the annotated ground truth (Table 4). The fact that experts were involved in the annotation of the results is crucial for the legal domain, since this allows the legal interpretation of the results, which can only be accomplished through an evaluation by legal experts [18].

While this is just a preliminary assessment of the classification performance of our unsupervised model, it is possible to derive that the label applied to the documents is correct in at least 39% of the cases, meaning that the approach is indeed able to link a document with its more relevant anchor with a good level of approximation.



consumo di suolo

09/07/2024

○ LEOS
● Normattiva
○ PDL Camera
○ Corte Costituzionale

Ricerca: consumo di suolo
Data: 2024/07/09
Directory: Normattiva

AMBIENTE, TERRITORIO E LAVORI PUBBLICI:

FRBRExpressionuri: /akn/it/act/decretoLegge/stato/2011-07-01/94/ita@2011-08-31/!main

Titolo: ((Misure urgenti in tema di rifiuti solidi urbani prodotti nella regione Campania.))

FRBRExpressionuri: /akn/it/act/decreto/ministeroAmbienteTutelaTerritorioMare/2012-08-10/161/ita@2017-08-22/!main

Titolo: Regolamento recante la disciplina dell'utilizzazione delle terre e rocce da scavo.

Figure 2: The first results of our search by topic system. Using the query "land consumption" in Italian on the Normattiva dataset, the system returns the appropriate Camera commission (environment, territory and public works) and the first two results are relevant (one is about waste management, the other about rocks and earth from excavation projects).

Dataset	Accuracy
Corte Costituzionale	0.45
PDL	0.39
Normattiva	0.47
EUR-Lex	0.58

Table 4
Accuracy values for all four datasets, when compared with the manually annotated documents.

When comparing the result, it is interesting to note that among the Italian datasets, which use the same categories, the Normattiva and Corte Costituzionale accuracy seems higher, while the PDL dataset shows a lower performance. This suggests that the finalized version of documents issued by the parliament and the Constitutional court might be simpler to classify in an unsupervised way, while the more draft-like qualities of the PDL dataset hinder the classification efforts.

7. Conclusions and Future Work

In this article, we present an unsupervised approach that aims to support the Italian legislative process, by providing useful insights into documents from the relevant European and Italian institutions (European Union, Constitutional Court, Italian Parliament). The system doesn't

[9] http://u2.cirsfid.unibo.it/portale-camera

only provide with a ranking of relevant documents, but it also returns the two most relevant EuroVoc terms (for EU documents) and Chamber commissions (for Italian documents). This allows the user a more thorough exploration of the relevant subjects, while also supplying suggestions in terms of specific documents.

Our approach is completely unsupervised and it does not rely on any form of annotation, meaning that scaling up the approach to more documents, or even using more performant models do not require any fine-tuning, with the procedure consisting in obtaining the article and title vectors for all documents. Furthermore, the adopted approach leverages the hierarchical nature of legislative documents, as represented in Akoma Ntoso XML in order to produce embeddings that are based on the structure of the document. Moreover, using a structured format as our input allows us to resolve normative references, without which some of the of a document will be impossible to understand for an automatic system.

The evaluation performed on the classification system showed a promising level of performance for an unsupervised model, which doesn't rely on any information about the specific task. Additionally, the multilingual model used in our method allows users to work both on English and Italian, both in terms of queries and in terms of results, with satisfying results. Nevertheless, it would be possible to improve the quality of the results by testing other models, which might yield better performance.

The validation of the search by topic task has been assessed by two senior legal researcher in the team, however it is recommendable to organize a session with relevant end-users with some concrete scenarios for returning relevant documents and categories given a user query. For this task, it would be necessary to involve the relevant stakeholders, meaning experts involved in the drafting of legislative documents in Italy. Nevertheless, the project has been evaluated by scientific experts [10] appointed by the Italian Chamber of Deputies in the context of its manifestation of interest and it was included as part of the work by of one of the two winning consortiums.

The experimental results obtained in this paper constitute a study of the application of pre-existing Sentence Transformer models in an unsupervised way to the classification and search of Italian legal documents. While we achieved satisfactory results, our approach could still be improved by improving upon the base methodology and conducting a more thorough exploration of other multilingual models. Furthermore, a formal evaluation by the stakeholders would also improve our understanding further specific parameters that arise during the legislative process.

[10] https://comunicazione.camera.it/archivio-prima-pagina/19-41329

Acknowledgments

This project is funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice PRIN 2022 (DD N. 104 del 02/02/2022), title Smart Legal Order in DigiTal Society - SLOTS, proposal code “Smart Legal Order in DigiTal Society (SLOTS)”, Proposal code 2022LRL2C2, CUP J53D23005610006. We also thank Salvatore Sapienza, Chantal Bomprezzi, Pier Francesco Bresciani for validating the results.

References

[1] M. Palmirani, R. Sperberg, G. Vergottini, F. Vitali, Akoma Ntoso Version 1.0 Part 1: XML Vocabulary, Technical Report, OASIS Standard, 2018. URL: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html.

[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[4] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, D. E. Ho, When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 159–168.

[5] P. Henderson, M. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, D. Ho, Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, Advances in Neural Information Processing Systems 35 (2022) 29217–29234.

[6] Q. Dong, S. Niu, Legal judgment prediction via relational learning, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 983–992. URL: https://doi.org/10.1145/3404835.3462931. doi:10.1145/3404835.3462931.

[7] C. Xiao, X. Hu, Z. Liu, C. Tu, M. Sun, Lawformer: A pre-trained language model for chinese legal long documents, AI Open 2 (2021) 79–84. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000176. doi:https://doi.org/10.1016/j.aiopen.2021.06.003.

[8] V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, S. K. Guha, A. Bhattacharya, A. Modi, ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4046–4062. URL: https://aclanthology.org/2021.acl-long.313. doi:10.18653/v1/2021.acl-long.313.

[9] F. Vitali, M. Palmirani, R. Sperberg, V. Parisse, Akoma Ntoso Version 1.0. Part 2: Specifications, Technical Report, OASIS Standard, 2018. URL: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html.

[10] M. Palmirani, Lexdatafication: Italian legal knowledge modelling in akoma ntoso, in: V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo, G. Sartor (Eds.), AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI JURIX 2018, AICOL-XII JURIX 2020, XAILA JURIX 2020, Revised Selected Papers, volume 13048 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 31–47. URL: https://doi.org/10.1007/978-3-030-89811-3_3. doi:10.1007/978-3-030-89811-3_3.

[11] M. Palmirani, F. Vitali, A. Bernasconi, L. Gambazzi, Swiss federal publication workflow with akoma ntoso, in: R. Hoekstra (Ed.), Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014, volume 271 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2014, pp. 179–184. URL: https://doi.org/10.3233/978-1-61499-468-8-179. doi:10.3233/978-1-61499-468-8-179.

[12] M. Palmirani, Akoma ntoso for making FAO resolutions accessible, in: G. Peruginelli, S. Faro (Eds.), Knowledge of the Law in the Big Data Age, Conference 'Law via the Internet 2018', Florence, Italy, 11-12 October 2018, volume 317 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2018, pp. 159–169. URL: https://doi.org/10.3233/FAIA190018. doi:10.3233/FAIA190018.

[13] A. Cvejić, K.-G. Grujić, A. Cvejić, M. Marković, S. Gostojić, Automatic transformation of plain-text legislation into machine-readable format, in: The 11th international conference on information society, technology and management (ICIST 2021), 2021.

[14] A. Flatt, A. Langner, O. Leps, Model-Driven Development of Akoma Ntoso Application Profiles: A Conceptual Framework for Model-Based Generation of XML Subschemas, Springer Nature, 2023.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[16] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020. URL: https://arxiv.org/abs/2004.09813.

[17] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in neural information processing systems 33 (2020) 16857–16867.

[18] M. Palmirani, S. Sapienza, K. Ashley, A hybrid artificial intelligence methodology for legal analysis, BioLaw Journal-Rivista di BioDiritto (2024) 389–409.