Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230982

Re-Framing Case Law Citation Prediction from a Paragraph Perspective

Henrik PALMER OLSEN [a,1], Nicolas GARNEAU [b], Yannis PANAGIS [a], Johan LINDHOLM [c] and Anders SØGAARD [b]
[a] *Faculty of Law, University of Copenhagen; Institut D'Études Avancées de Paris*
[b] *Computer Science Department, University of Copenhagen*
[c] *Faculty of Law, Umeå University*

Abstract.
Case law citation prediction, i.e., predicting what historical cases are relevant for your current case, can assist legal discovery and decision-making, but legal documents are long, and often only parts of them are relevant for a particular use case. We therefore reframe case law citation prediction as a paragraph-to-paragraph citation task, introduce a new dataset, and train and evaluate new models. We also evaluate our models qualitatively. Our resources provide a first step toward discovering citation patterns and modeling legal rules in EU law from precedent documents.

Keywords. legal dataset, case law citation, link prediction, legal rules

1. Introduction

A key aspect of both legal research and legal practice is the task of retrieving relevant case law. In this paper, we focus on the retrieval of relevant case law from the Court of Justice of the European Union (CJEU). The CJEU is a prolific court: It decides around 800 new cases every year and has, since its establishment in 1952, issued more than 14,000 judgments, each of which involves several pages of complex legal text. Extracting the relevant information from this large and ever-increasing amount of data constitutes a significant challenge. Existing Legal Information Systems (LIRs) provide entire judgments (as well as other legal documents) as the output to users with access to CJEU case law. This includes, inter alia, the EU's own LIRs: EUR-LEX[2] and the CJEU's own database, Curia[3]. These LIRs seek to assist the legal research process by providing a list of the judgments given a specific search input, but the results are presented at the case level, thereby leaving it to users to manually read through those documents in their entirety to find the relevant piece of information. We propose a complementary approach for LIRs which returns individual paragraphs that restate legal rules instead of entire judgments. We believe that retrieving individual paragraphs is closer to what lawyers seek and need when crafting legal arguments. If a lawyer has found a relevant rule restated in one judg-

[1] Corresponding Author: Henrik Palmer, hpo@jur.ku.dk.
[2] https://eur-lex.europa.eu/collection/eu-law/eu-case-law.html
[3] https://curia.europa.eu/juris/recherche.jsf?cid=43048

ment, the value of finding an alternative authority for an identical restatement in another judgment is limited. However there is value for the user in being able to see alternative formulations of the same rule, as this may provide legal nuance. In this contribution, we show how semantic similarity analysis can identify paragraphs across several judgments that restate the same rule, but at the same time differ in a legally relevant way.

Our research thereby places itself in the context of what has broadly been called Computational Legal Studies [1] Previous research in this field has used these methods to show how the Court's jurisprudence develops [2,3,4,5,6,7] and to challenge traditional textbook accounts of the judicial salience of individual cases [8,9]. Computational legal studies have also been applied to other areas of law and using other approaches. In the field of European Human Rights Law, for example, there has been research aimed at predicting judgment outcomes [10,11,12]. In the field of International Trade Law, there has been research on the use of pathos-related arguments in World Trade Organization decisions [13]. [4] In this article, we expand on previous CJEU-related research by analyzing a whole new paragraph-to-paragraph dataset which is, to the best of our knowledge, a first attempt in this direction. This paper thus makes the following contributions; the release of a new paragraph-to-paragraph citation dataset along with several baseline models trained for the paragraph-to-paragraph link prediction task. We also conduct a manual analysis of the clustered rules according to the best-performing model. We release all related code and data.[5].

2. Methodology

The paragraph-to-paragraph dataset that we introduce consists of *paragraph-to-paragraph citations*, ranging from case-law paragraphs dating back to February 1978, until cases decided in October 2021. Both the paragraphs and their associated metadata are collected from EUR-LEX. In order to collect the court cases' paragraphs and metadata, we use the *XML Web-Service* provided by EUR-LEX. Given a specific case identifier called `CELEX`, we are able to retrieve all data available for a specific court case. Once the data is downloaded, we extract the metadata and paragraphs using XPath expressions. We also extract citations to case law using the metadata, however, for cases before 2015 we used GATE[6] to parse text and extract the relevant citations to paragraphs since the citations are not made available from EUR-LEX in the previous format. The paragraph-to-paragraph citation dataset contains 110,609 rows with information about the citing and the cited paragraphs (e.g. CELEX number, decision date, paragraph number & text).

We then introduce the task of predicting citations, i.e. if a link exists (or should exist) between two paragraphs, and further detail our experiments. Formally, we define the set of paragraphs in the paragraph-to-paragraph dataset as $\mathbf{p}$, a *citing* paragraph as $\overrightarrow{p}$ and a *cited* paragraph as $\overleftarrow{p}$. The goal of the task is to, given the ith citing paragraph $\overrightarrow{p}_i$, find the set of cited paragraphs $\overleftarrow{\mathbf{p}}_i$ such that $\overleftarrow{\mathbf{p}}_i \subseteq \mathbf{p}_i$ where $\mathbf{p}_i$ contains only paragraph published *before* $\overrightarrow{p}_i$. We thus frame this task from an information retrieval perspective so that a ranking model takes as input $\overrightarrow{p}_i$ and $\mathbf{p}_i$ and produces a ranked list of paragraphs $\tilde{\mathbf{p}}_i$. We

[4] For a broader overview of the "judgment prediction" literature and a critique, see [14].

[5] https://github.com/coastalcph/paragraph_network

[6] https://gate.ac.uk/

then compute the performance of the ranker using the Average Precision of $\tilde{\mathbf{p}}_i$ with respect to $\overleftarrow{\mathbf{p}}_i$. To assess the generalization performance of the ranker, we split the original dataset into training and testing sets with 46,637 and 11,891 examples respectively. The training set contains citing paragraphs that appeared before 2018 and the test set contains citing paragraphs that appeared in 2018 and onwards. We consider only a paragraph's text to retrieve citations among the set of possible candidates. We use several baselines as well as state-of-the-art models for the paragraph link prediction task. We split the set of models into unsupervised or supervised models. The unsupervised setting uses models that are not explicitly trained to predict a link between two paragraphs. We use their pretrained vectorial representations to perform citation prediction by computing the cosine similarity between two representations in order to rank the set of candidates $\mathbf{p}_i$ with respect to the citing paragraph $\overrightarrow{p}_i$. We consider two unsupervised models: Term Frequency - Inverse Document Frequency (TF-IDF) and SentenceBERT (SBERT) [15]. In the supervised setting we use our dataset to train the models to explicitly predict if there is a link between two paragraphs. However, The paragraph-to-paragraph citation dataset contains only positive examples; there are only pairs of citing and cited paragraphs. Hence, a link prediction model needs to learn to predict when there is a link between a citing and a cited paragraph, and when *there is not*. Thus, we need to create negative examples. To this end, we use a TF-IDF model to retrieve negative examples in two different ways. The first method takes the farthest paragraph, p_i^*, w.r.t the cited paragraph $\overleftarrow{p}_i$ according to the distributional semantics of the TF-IDF model i.e. $1 - \text{cosine}(\overleftarrow{p}_i, p_i^*)$. The other method samples a negative paragraph among the set of possible paragraphs $\mathbf{p}_i$ according to the distance provided by the TF-IDF model. That is, a paragraph highly dissimilar to the cited paragraph has more chance of being selected as a negative example. We denote each method as *hard negatives* (HN) and *sampled negatives* (SN) respectively from now on. Both methods will create, for each positive example $(\overrightarrow{p}_i, \overleftarrow{p}_i, 1)$, a negative example w.r.t $\overleftarrow{p}_i$ $(\overrightarrow{p}_i, p_i^*, 0)$ so that we have a perfectly balanced dataset derived from the original paragraph-to-paragraph training set. This dataset thus contains a total of 181,114 examples, split into a pseudo-train and pseudo-test set of 144,891 and 36,223 examples respectively. In the supervised setting, we consider two models; SBERT, trained on our pseudo-train dataset, and SimCSE [16]. Both models use Contrastive Loss where they try to bring the citing and cited paragraphs closer in the vector space while pulling apart the citing and negative paragraphs.

We evaluate each model on the link prediction task using the Mean Average Precision of the test set. We can see from **Table 1** that the TF-IDF model is a strong baseline, outperforming SentenceBERT in both untrained and trained settings. The SimCSE model yields the best performance either using *hard negatives* or *sampled negatives* with Mean Average Precision of 0.441 and 0.489 respectively. After manually analyzing a dozen rankings, we found out that the actual nature of the data (sparsity of the citation graph) leads us to think that AP underestimates the real performance of the rankers. For a given example, according to the best ranker (SimCSE-SN), the cited paragraph has been ranked in the third position, yielding an average precision of 0.33. However, the two other paragraphs at rank 1 and 2 would have been equally valid citations for the citing paragraph, but they were not explicitly cited. This is mainly caused by the fact that judges do not aim for completeness but rather prefer efficiency where a few relevant citations are enough instead of providing an exhaustive list of all possible citations. There are two interesting facts resulting from this discovery; i) the model is indeed able to cluster para-

Table 1. Results on the Link Prediction Task measured using the Mean Average Precision. TF-IDF and SentenceBERT (SBERT) are untrained models. Trained version of SentenceBERT and SimCSE are either using Hard Negatives (HN) or Sampled Negatives (SN).

	TF-IDF	SBERT	SBERT-HN	SimCSE-HN	SBERT-SN	SimCSE-SN
Mean AP	0.393	0.351	0.384	0.441	0.334	**0.489**

graphs about the same legal matter and ii) this suggests that we might be able to deduce legal rules from the paragraph-to-paragraph citation dataset using the SimCSE-SN.

Hence, we conjecture that a good enough ranking model is able to cluster legally related paragraphs together in order to implicitly exhibit legal rules. We thus verify this hypothesis by looking at the rankings provided by the SimCSE-SN model with respect to a given citing paragraph. To limit the scope of our analysis, we selected citing paragraphs that are associated with only one subject matter. We considered 9 subject matters having several paragraphs (>100): Social Policy (332), Free Movement of Capital (201), Right of Establishment (165), Provisions Governing the Institutions (164), Approximation of Laws (145), Social Security of Migrant Workers (131), Transport (126), Area of Freedom, Security and Justice (104), and Free Movement of Workers (102). From each of these 9 subject matters, we selected 10 citing paragraphs that have various Average Precision with respect to SimCSE-SN, one for each tenth. For each of these citing paragraphs, we selected the top 10 ranked paragraphs according to SimCSE-SN along with their respective subject matter. Two law professors with a thorough background and understanding of EU Law in general have been selected as evaluators to conduct this manual analysis. The evaluators were provided the 10 citing paragraphs along with their associated citations and the list of candidates with their associated similarity scores and subject matter. Given the citing paragraph, the citations, and the top 10 rankings, the evaluators were asked the following question: *Does the citing paragraph contain a statement of an EU law rule or principle?* The evaluator answers Yes or No. If Yes, then the evaluators were asked if the cited paragraph contained a verbatim version of the rule in the *citing* paragraph (Yes/No) and if the cited paragraph contain a different or more expanded version of the rule in the *citing* paragraph (Yes/No). Hence, for every example where the evaluator has answered Yes to the first question, there are four possible outcomes for each of the 10 ranked paragraphs: 1) Yes and Yes; 2) Yes and No; 3) No and Yes and 4) No and No. We consider that both 1), 2), and 3) count as positive findings - ie. the ranked paragraph contains useful information supplementing the citing paragraph. Even if it is a verbatim repetition of the rule (outcome 2), it will be stated in a different case and hence in a different context. This different context provides additional information to the citing paragraph. We consider that 4) is an indication that the ranked paragraph is not related to the rule in the citing paragraph, and therefore does not provide additional information about the rule in question (although it may provide related information). 1) and 3) are the most interesting outcomes as they provide relevant rule information that is not a verbatim repetition of the rule in the citing paragraph.

The results of the manual evaluation are displayed in **Table 2**. The first column displays the ratio of citing paragraphs containing EU Law or principle for which both evaluators agreed. From these relevant citing paragraphs are displayed in columns two and three the ratios of candidate paragraphs being a verbatim version of the cited paragraph or containing a different or more expanded version. Lastly, column four contains the ratio of positive candidates, which is defined by containing either a verbatim and/or a differ-

Table 2. Manual analysis of the 9 subject matters regarding the citing and candidate paragraphs. The evaluators verified if the citing paragraph actually contained an EU Law or principle, if the candidate paragraphs, proposed by the SimCSE-SN model, were verbatim or different/expanded versions of the citing paragraph, and if the candidate paragraph was a positive one. The values are the *agreed-upon* ratios with their respective inter-annotator agreement in parentheses.

	Citing Paragraph	**Candidate Paragraphs**		
Subject Matter	**Contains EU Law?**	**Verbatim Version?**	**Expanded?**	**Positive?**
Approximation of Law	0.88 (0.80)	0.27 (0.41)	0.11 (0.51)	0.73 (0.86)
Area of Freedom	1.00 (0.80)	0.05 (0.30)	0.15 (0.37)	0.72 (0.82)
Free Movement of Capital	1.00 (0.80)	0.14 (0.46)	0.25 (0.48)	0.64 (0.83)
Free Movement of Workers	0.71 (0.78)	0.22 (0.62)	0.24 (0.47)	0.64 (0.80)
Provisions Gov. Institutions	1.00 (0.80)	0.11 (0.64)	0.43 (0.65)	0.71 (0.88)
Right of Establishment	0.66 (0.90)	0.23 (0.42)	0.17 (0.43)	0.97 (1.00)
Social Policy	1.00 (0.80)	0.24 (0.48)	0.27 (0.51)	0.79 (0.88)
Social Security for Migrant	1.00 (1.00)	0.1 (0.57)	0.11 (0.37)	0.59 (0.74)
Transport	1.00 (1.00)	0.12 (0.66)	0.34 (0.55)	0.69 (0.73)
Average	0.93 (0.85)	0.16 (0.51)	0.23 (0.48)	0.72 (0.84)

ent/expanded version of the citing paragraph. Almost every citing paragraph contained a statement of an EU law rule or principle (93%) for which the evaluators agreed 85% of the time. Regarding the verification of whether a candidate paragraph is a verbatim or different/expanded version of the citing paragraph, there is a discrepancy between evaluators (51% and 48% agreement on average respectively). However, both evaluators agreed (84%) that most of the candidates (74%) proposed by the SimCSE-SN model, across different subject matters, were actually relevant. This result suggests that the model is able to cluster relevant paragraphs in the same vector space according to the subject matter. Here are the key findings from our manual analysis;

1. The model can identify many different formulations of the same rule, similar to the one in the citing paragraph, for a multitude of subject matters and languages.
2. The ranked paragraphs may contain statements of the same rule as the citing paragraphs, however, the similarity scores sometimes capture subtle but nevertheless legally important differences.
3. The model can identify paragraphs containing the same legal rule as the citing paragraph, even when for rules that are complex or quite different language.
4. The model is capable of identifying paragraphs containing rules that are related to, but distinct from the rule in the citing paragraph.
5. There are challenges when it comes to determining where distinct rules start and end as they can be closely related.

From this qualitative analysis, we can see that the SimCSE-SN model is an interesting contender for the task of predicting paragraph citations.

3. Conclusion

This paper introduces two new resources which are the Paragraph-to-Paragraph citation dataset and a strong Paragraph-to-Paragraph link prediction model. We provided evi-

dence that these resources will be useful for the research community by conducting a manual analysis of the model's output. Although inter-coder reliability is low overall, it is remarkable that only very few extracted paragraphs were coded as No and No (outcome 4) by any of the coders, meaning that few paragraphs were unrelated to the rule in the citing paragraph. This indicates that in the context of our paragraph-to-paragraph dataset, semantic similarity is generally a good proxy for legal similarity. However, the low inter-coder reliability also indicates that the concept of "same legal rule" is ambiguous and relative to subjective perception.

References

[1] Whalen R. Computational legal studies: the promise and challenge of data-driven research. Edward Elgar Publishing; 2020.
[2] Mones E, Sapieżyński P, Thordal S, Olsen HP, Lehmann S. Emergence of network effects and predictability in the judicial system. Scientific reports. 2021;11(1):2740.
[3] Sadl U, Tarissan F. The relevance of the network approach to European (case) law: reflection and evidence. Oxford University Press; 2020.
[4] Kilpatrick C, Scott J, et al. New Legal Approaches to Studying the Court of Justice: Revisiting Law in Context. Oxford University Press, USA; 2021.
[5] Šadl U, Olsen HP. Can quantitative methods complement doctrinal legal studies? Using citation network and corpus linguistic analysis to understand international courts. Leiden Journal of International Law. 2017;30(2):327-49.
[6] Derlén M, Lindholm J. Characteristics of precedent: The case law of the European court of justice in three dimensions. German Law Journal. 2015;16(5):1073-98.
[7] Frankenreiter J. Network Analysis and the Use of Precedent in the Case Law of the CJEU–A Reply to Derlén and Lindholm. German Law Journal. 2017;18(3):687-94.
[8] Derlén M, Lindholm J. Goodbye van g end en l oos, hello b osman? Using network analysis to measure the importance of individual CJEU judgments. European Law Journal. 2014;20(5):667-87.
[9] Frese A, Olsen HP. Citing Case Law: A Comparative Study of Legal Textbooks on European Human Rights Law. Eur J Legal Stud. 2018;11:91.
[10] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ computer science. 2016;2:e93.
[11] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317-23. Available from: https://aclanthology.org/P19-1424.
[12] Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. p. 226-41. Available from: https://aclanthology.org/2021.naacl-main.22.
[13] Busch ML, Pelc KJ. Words Matter: How WTO Rulings Handle Controversy. International Studies Quarterly. 2019.
[14] Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. Artificial Intelligence and Law. 2023;31(1):195-212.
[15] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3982-92. Available from: https://aclanthology.org/D19-1410.
[16] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 6894-910. Available from: https://aclanthology.org/2021.emnlp-main.552.