Finding Implicit Links Between Norms Using HONto

Sabine Wehnert^{1,2}, Ernesto William De Luca^{1,2}

¹Georg Eckert Institute Leibniz Institute for International Textbook Research, Germany ²Otto von Guericke University Magdeburg, Germany

Abstract

Linking legislation is an important task in knowledge modeling for the legal domain. This also involves implicit links which are not apparent from mere content analysis of the respective norms. Current systems mostly rely on ontologies or classifiers to enrich the norm content and thereby detect implicit links. The drawback of this approach is that the required concepts and their relations may not already be present in publicly available ontologies, requiring manual effort in conceptualizing and integrating new knowledge. In contrast, we employ a general approach called HONto to automatically extract knowledge about explicitly and implicitly related norms from textbooks. Having a textbook source, we can refer to the individual text passage where we extracted the relation from and thus offer evidence for each implicit link. Therefore, the HONto approach can assist knowledge engineers in finding implicit links and automate a part of the engineering process. We intend to implement HONto in an information retrieval scenario to detect relevant changes in law. With its in-built mechanism to incorporate implicit relationships between norms during similarity scoring, HONto aims for providing high-recall results.

Keywords

implicit norm links, information extraction, bottom-up knowledge base

1. Introduction

Most systems in the field of Legal AI base their knowledge on the content of legal texts by extracting relevant information regarding entities or concepts and integrating it into the respective knowledge representation. In general, regulatory texts are not seen as stand-alone documents, they have many relationships and dependencies among each other. Some of them are mentioned explicitly in the text by referring to another article or document with a citation. Other relationships - which we refer to by *implicit links* - are not directly encoded in the content of the regulatory document. Those implicit links may not apply in all contexts, thus they are only valid if a certain situation is given. In earlier work [1], we presented requirements for linking legal documents, and one of them is particularly relevant for implicit links: *topical relevance*. Hence, the conditions under which such a link is formed need to be also considered during the knowledge modeling process. In rule-based systems, this may require a dedicated classifier for the condition types and an intelligent mechanism to decide when topical relevance is given.

In our work called *HONto* we focus on information extraction for a retrieval system to determine norm changes which are relevant to a user. The user has to explore the knowledge

RELATED - Relations in the Legal Domain Workshop, in conjunction with ICAIL 2021, June 25, 2021, São Paulo, Brazil Sabine.wehnert@gei.de (S. Wehnert); deluca@gei.de (E. W. De Luca)

^{© 0 2021} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

base and then mark relevant concepts or norms, for the system to determine the similarity to upcoming (changes in) regulatory documents [2]. For all of this to happen though, the knowledge base needs to be populated with legal documents, their relationships and the concepts related to them. In this aspect, HONto can be distinguished from other related work: We extract this domain knowledge automatically from textbooks [3]. This also includes implicit relations between legal documents, which are the main subject of this work.

Hence, our contributions are:

- We show how our system forms links between legal texts.
- We discuss under which circumstances an implicit link shall be made.

The remainder of this work is organized as follows: Section 2 contains our concept for linking legal documents based on topical relevance. Section 3 is about a preliminary experiment we made on implicit link detection on the German Civil Code. In Section 4, we collect related work for automated knowledge extraction in regulatory monitoring and for linking implicit references to legal documents. With the final section we conclude this research and mention future work.

2. Topical Relevance in HONto for Linking Laws

The HONto system is based on a knowledge base containing many concept hierarchies, where each hierarchy is extracted from one textbook in the corpus. Those hierarchies need to be connected, however we do not intend to perform any extensive manual effort in terms of concept integration or merging. Instead, each hierarchy stays intact, having its own logical structure from the author's point of view. This means that HONto is based on the concept of a lightweight ontology, a shallow meaning representation [4] from section titles, keywords and legal document entity relations in the way that they are expressed in the text. The only modification we make on those concept hierarchy structures is to add links between them. Those links can be established between concepts or between legal text references made in the textbooks which share the same context. Adding more links between those concepts can help bring similar contexts closer in retrieval settings for query expansion. For instance, we can take into account the newly formed graph structure from the linked concept hierarchies and use its node encoding [5] for or in addition to other similarity computation to perform similarity scoring on HONto.

We already introduced the concept of entity linking in a previous work [1], where we resolved references to different types of legal texts: Norms, Court Decisions, EU-Directives, EU-Regulations, and combinations of the aforementioned types. However, we did not make the distinction between implicit and explicit references yet because until now, we did not use the original content of the referenced legal texts. In this research, we showcase the linking mechanism for explicit and implicit relationships between norms based on the knowledge we can extract from textbooks.

Figure 1 illustrates our approach for explicit and implicit linking of the norms. In this example, § 823 BGB is the article of interest from the German Civil Code (BGB). We see three concept hierarchies which reference this article. The references §§ 823, 253 II BGB and § 823 BGB are

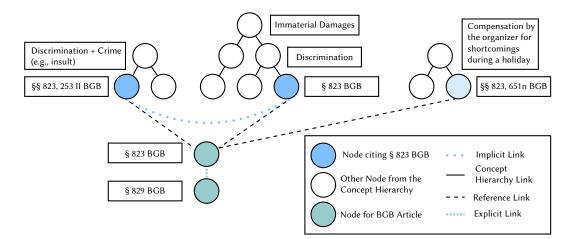


Figure 1: Linking explicitly and implicitly related norms.

both made in the same context of damages for discrimination, whereas the third reference is about compensation for shortcomings during a holiday, thus not topically relevant in the discrimination context. Therefore, we establish an implicit link between the first two references, but not with the third one. There is a reference link between all three reference nodes and the node for § 823 BGB. Considering the actual article text of § 823 BGB, there is no other article referenced. Also the articles § 253 II BGB and § 651n BGB do not contain any citations of other norms. Therefore, those article texts do not generate any explicit links and we can only form a link between them via implicit links generated from the extracted textbook knowledge. While searching in the remaining articles of the German Civil Code, we find another article (§ 829 BGB) that references § 823 BGB. Between those two articles, we can therefore establish an explicit link.

To summarize, we only form implicit links among references if they share the same context and only if no explicit link is present. Explicit links are made among the nodes of the legal texts directly. In this way, explicit links are intentionally forming a shorter path between related norms than than implicit links.

Our whole process of node linking is set up as a knowledge graph, where node distance is determined by path length and node features. The node features are obtained by vectorizing the strings we extracted from the textbook. The vectorization can be done with different methods, such as a bag-of-words approach with TF-IDF scoring or embedding the sequence with a BERT model [6]. The choice of the encoding model depends on the user's individual preference for explainability. Nowadays, there is often a trade-off between performance and explainability (e.g., between TF-IDF scoring, topic models and BERT for the Statute Retrieval Task of the Competition on Legal Information Extraction/Entailment [7]). Comparing those different encoding methods in the context of legal norm linking is subject to future research. In the following, we present a small experiment on our knowledge base to show the linking process and the amount of implicit references in the German Civil Code that we can extract from our textbook corpus.

3. Experiment

We choose the German Civil Code for this experiment because the citation pattern for its articles is the most uniform one (compared to the other legal document types we extracted) and therefore, we expect the most accurate results in terms of reference extraction and the subsequent reporting of established links. Furthermore, we found earlier $[1]^1$ that norms are the most prevalent reference type in the type of textbooks we use, with 83,661 norm references over our corpus of 193 textbooks. The similarity computation of the previous work among those candidates has been performed with the Latent Dirichlet Allocation (LDA) topic model. We note that the choice of LDA was based on its relatively good performance and explainability on the textbook corpus, but this choice can be made differently depending on user preferences. We analyze the same corpus also in this work. The implicit link candidates we consider in this work are based on jointly cited legal texts, of which we have 122 instances. For the sake of the following experiment we avoid any uncertainty stemming from links between multiple references that are formed under similarity computation, instead we use norms that are already linked together via a reference link and see how meaningful the link between the cited references is. Of course, in a real-world setting there will be an error from the similarity computation for implicit link creation.

We consider the knowledge gained from those jointly cited implicit references as meaningful, if there is no ingoing or outgoing explicit citation from the affected norms towards each other. Jointly cited norms usually apply to very specific contexts, so that in this case, we can treat the involved articles as instances to be linked. The question we want to answer in this experiment is: How many instances may receive explicit or implicit links?

Out of the 122 instances, there are 39 instances referring to the German Civil Code (BGB), as shown in Table 1. Since this number of references is manageable, we manually checked how many of the 39 combinations are obtainable from the original article texts (i.e., explicit links), compared to how many of those relationships we could only get from the textbook knowledge (i.e., implicit links). We consider two different search spaces: one called "BGB-only" for combinations only within the BGB, and "External" with relations to articles in other laws than the German Civil Code. For "BGB-only", we have 13 references, for the "External" category 26 references. For the search direction, we can search for an explicit reference in either both articles we compare (="Bidirectional"), or in case of the "External" category, we also distinguish between "Outbound" for explicit references from the BGB article to other laws and "Inbound" for explicit references from other laws to the corresponding article in the BGB. The "Outbound" and "Inbound" sets in the "External" category overlap in several parts, however we only find 11 explicit links if we search in both directions for each reference, otherwise we obtain fewer links. In total, from 39 references, we are able to detect 15 of those references with the bidirectional search method, such that those cases become explicit links in our system. The remaining 24 references are candidates for implicit links. We call those instances "candidates" because it depends on the similarity scoring we employ to detect topical relevance whether an implicit link between two references can be formed. Before forming implicit links, we need to perform a user study and obtain relevance feedback for different scoring methods in the final HONto

¹Code for reference extraction: https://github.com/anybass/HONto/tree/master/reference_linking

Search Space	Search Direction	References	Explicit Links	Implicit Link Candidate
BGB-only	Bidirectional	13	4	9
External	Outbound	26	9	17
External	Inbound	26	7	19
External	Bidirectional	26	11	15
Total	Bidirectional	39	15	24

 Table 1

 Amount of explicit and implicit relationships based on BGB combinations identified by HONto.

recommender system. In this work, we focus more on the nature of implicit links once they have been formed, regardless of the scoring method. Characterizing the types of implicit links may serve as a basis to justify future research on several scoring methods for detecting those links.

Upon closer inspection, we find several, mutually not exclusive properties of the implicit link candidates, which we refer to as follows:

- default link
- situational relationship
- implicit mention

The most obvious category of implicit links in our dataset are those with default links, which are norms that are always cited together as a default setting. For instance, we find a reference to § 195 BGB in connection to § 199 I BGB, where the former denotes the duration of a limitation period and the latter its beginning at the end of the year as a default setting. Exceptions in which the start of the limitation period is determined differently, belong to situational relationships. Default links are rather simple to find, either by spatial norm article proximity in the same part of the law book, or they may be modeled in domain ontologies where they could constitute a rule. Also in textbook sources, we can identify the standard combinations, since they occur comparatively often.

Some situational relationships are particularly hard to obtain for a non-domain expert, especially if there is no relevant term overlap or even paraphrases among the texts to be linked. Such an example is the entitlement to continued employment as a claim of an employee who is in dispute with his employer about the employment relationship, until the existence or non-existence of the employment relationship has been clarified. The involved norms for this case do not contain any mention about the entitlement itself; instead we find § 611a BGB (employment contract), § 613 BGB (non-transferable duties of service) and § 242 BGB (performance in good faith) cited. Our dataset also suggests a relationship between § 134 BGB and § 203 Abs. 1 StGB in specific situations of a void practice purchase agreement due to a violation of private secrets (§ 203 Abs. 1 StGB) for transferring patient data to the buyer. Consequently, the legal transaction that violated a statutory prohibition is void (§ 134 BGB). Another example for a situational relationship found in the textbooks is tort liability of a money courier § 823 Abs. 2 BGB (liability in damages) in connection to § 261 StGB (money laundering). Those situational relationships have in common that they only apply to a very narrow sense, hence those combinations may

be irrelevant, if retrieved without determining prior semantic similarity to a given user context. On the other hand, the knowledge about those relationships is valuable when the user faces the exact same situation, so that a retrieval system from the textbooks shall be able to return those relationships in exactly these situations and offer proof of the relationship by pointing to the textbook source.

Implicit mentions can consists of keyphrases or references to other norms. Keyphrases in that regard are known by experts to be defined by a norm that is not always explicitly cited along (such as the above example of "performance in good faith"). Implicit mentions may also be a reference to another norm, which is not following the standard format or is just an approximate pointer. For instance, the combination of § 75c I 2 HGB and § 343 I 1 BGB is intended for a reduction of a disproportionately high penalty when acting against the non-competition clause. § 75c I 2 HGB contains an implicit mention of the other involved law, but instead of citing § 343 I 1 BGB explicitly, we find a subsumption of multiple articles under the section: "Where the commercial employee has promised to be subject to a penalty in the event of failing to perform the obligation undertaken in the agreement, the principal may assert claims only in accordance with the provisions of section 340 of the Civil Code. The provisions of the Civil Code concerning reduction of a disproportionately high contractual penalty shall remain unaffected.". Such cases require reference resolution (to detect "provisions of section 340") and relevance scoring thereof for the given situation to obtain the most relevant section in the given context (§ 343 I 1 BGB).

In view of the aforementioned properties we find in examples of implicit references, we see potential in further research in this regard. Overall, we find in this experiment that our textbook corpus offers domain knowledge about implicit links, which we cannot obtain from the law texts directly. This finding exposes a general necessity of methods to find those implicit links to improve performance in Legal AI systems, especially in our legal norm retrieval use case. By using the HONto system, this knowledge is automatically extracted and inserted into a knowledge base. Depending on the similarity threshold that is set, the HONto system connects laws not only based on their explicit links, but also based on their context-dependent relationships via implicit links between references to those laws.

4. Related Work

In this section, we first describe related research regarding automatic knowledge extraction for finding compliance-relevant norm changes and second, we collect other methods for implicit relationship extraction. The work by Schönhof et al. [8] presents an approach for automatically extracting entity information from natural language text for automatic ontology creation. They also apply their method on the German Civil Code, however, they do not use textbooks as a knowledge source but other ontologies and wikidumps. Our HONto system differs from their approach because our knowledge base explicitly models implicit links between norms. HONto's main use case is information retrieval and regulatory change recommendation based on concept similarity from many not-integrated knowledge sources. Thereby, HONto circumvents issues related to integration conflicts between individual concept hierarchies at the cost of not providing any reasoning capabilities. Some examples for systems which offer reasoning capabilities for compliance checking are CARiSMA [9], the works by Amantea [10], as well as

Palmirani and Governatori [11].

In their case study on EU legislation, Sulis et al. [12] extract implicit relationships between recitals and articles using stemming and use word overlaps to assign a weight to the relationships between both nodes in their graph. They also further categorize the implicit relationships, but experiments with two human annotators show that there may be no easy agreement regarding the category assignment. The advantage of the way we evaluated the implicit links is that we base our analysis on evidence of such a relationship in terms of jointly cited norms in published textbooks. On the other hand, our evaluation method may only apply for a subset of possible implicit relationships. Our definition of topical relevance for implicit relationships fits most to their category "Conceptually Similar", although we do not only consider word or stem overlap, but also semantic similarity (e.g., from contextual word embeddings). While deep learning approaches and pre-trained models become increasingly available for the legal domain, Nanda et al. [13] report from a similar use case that nowadays still combining TF-IDF scoring and cosine similarity can achieve the best performance. Devyatkin et al. [14] employ deep learning techniques for detecting implicit relations between fragments of legal documents and achieve good results, but it is not clear if a BM25 or TF-IDF-based model with cosine similarity may perform better on that dataset, as well. A further category for implicit links between norms are also hierarchical dependencies, as mentioned by Opiła and Pełech-Pilichowski [15]. For future experiments, we may check if this category also is covered at least partially by our textbook knowledge, since this certainly depends on the books or commentary included in the corpus.

Therefore, we conclude the related work section by noting that there are systems which are similar to HONto in their use case, while HONto's extraction of domain knowledge from textbooks and the knowledge modeling suited to implicit links has not been done before for retrieving compliance-relevant changes in law.

5. Conclusion

This work focuses on the formation of links between implicitly related legal documents in the HONto knowledge base. We condition the link formation on topical relevance which is determined by the contextual similarity between a reference to legislation in textbooks. We find that textbooks offer valuable knowledge about implicit relationships between norms, which cannot be extracted from the norm content itself. In future research, we intend to enhance our textbook corpus by further open knowledge from online commentary, Wikipedia and ontologies. This work describes design decisions which are still part of an ongoing research process for the HONto system. Therefore, a final evaluation of alternative methods to determine topical relevance, as well as a user study to obtain multiple relevance assessments for HONto's recommendations is subject to future work.

Acknowledgments

This work is supported by Legal Horizon AG.

References

- [1] S. Wehnert, G. C. Durand, G. Saake, ERST: leveraging topic features for context-aware legal reference linking, in: M. Araszkiewicz, V. Rodríguez-Doncel (Eds.), Legal Knowledge and Information Systems - JURIX 2019: The Thirty-second Annual Conference, Madrid, Spain, December 11-13, 2019, volume 322 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2019, pp. 113–122. URL: https://doi.org/10.3233/FAIA190312. doi:10.3233/FAIA190312.
- [2] S. Wehnert, W. Fenske, G. Saake, Context selection in a heterogeneous legal ontology, in: H. Meyer, N. Ritter, A. Thor, D. Nicklas, A. Heuer, M. Klettke (Eds.), Datenbanksysteme für Business, Technologie und Web (BTW 2019), 18. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 4.-8. März 2019, Rostock, Germany, Workshopband, volume P-290 of *LNI*, Gesellschaft für Informatik, Bonn, 2019, pp. 129–134. URL: https://doi.org/10.18420/btw2019-ws-13. doi:10.18420/btw2019-ws-13.
- [3] S. Wehnert, D. Broneske, S. Langer, G. Saake, Concept hierarchy extraction from legal literature, in: A. Cuzzocrea, F. Bonchi, D. Gunopulos (Eds.), Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018, volume 2482 of CEUR Workshop Proceedings, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2482/paper33.pdf.
- [4] G. Ferraro, H. Lam, S. C. Tosatto, F. Olivieri, M. B. Islam, N. van Beest, G. Governatori, Automatic extraction of legal norms: Evaluation of natural language processing tools, in: M. Sakamoto, N. Okazaki, K. Mineshima, K. Satoh (Eds.), New Frontiers in Artificial Intelligence JSAI-isAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI, Yokohama, Japan, November 10-12, 2019, Revised Selected Papers, volume 12331 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 64–81. URL: https://doi.org/10.1007/978-3-030-58790-1_5. doi:10.1007/978-3-030-58790-1_5.
- [5] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 855–864. URL: https://doi.org/10.1145/2939672.2939754. doi:10.1145/2939672.2939754.
- [6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423.
- [7] J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, K. Satoh, Coliee 2020: Methods for legal document retrieval and entailment, https://sites.ualberta.ca/~rabelo/COLIEE2021/ COLIEE2020_summary.pdf, 2020. Accessed: 2021-05-09.
- [8] R. Schönhof, A. Tenschert, A. Cheptsov, Practical aspects of ontology-based analysis and reasoning for law information represented in textual form, International Journal On Advances in Intelligent Systems 8 (2015) 45–55.
- [9] T. Humberg, C. Wessel, D. Poggenpohl, S. Wenzel, T. Ruhroth, J. Jürjens, Using ontologies to analyze compliance requirements of cloud-based processes, in: M. Helfert,

F. Desprez, D. Ferguson, F. Leymann (Eds.), Cloud Computing and Services Science -Third International Conference, CLOSER 2013, Aachen, Germany, May 8-10, 2013, Revised Selected Papers, volume 453 of *Communications in Computer and Information Science*, Springer, 2013, pp. 36–51. URL: https://doi.org/10.1007/978-3-319-11561-0_3. doi:10.1007/978-3-319-11561-0_3.

- [10] M. Araszkiewicz, I. A. Amantea, S. Chakravarty, R. van Doesburg, M. Dymitruk, M. Garin, L. Gilpin, D. Odekerken, S. S. Salehi, ICAIL doctoral consortium, montreal 2019, Artif. Intell. Law 28 (2020) 267–280. URL: https://doi.org/10.1007/s10506-020-09267-z. doi:10. 1007/s10506-020-09267-z.
- [11] M. Palmirani, G. Governatori, Modelling legal knowledge for GDPR compliance checking, in: M. Palmirani (Ed.), Legal Knowledge and Information Systems - JURIX 2018: The Thirtyfirst Annual Conference, Groningen, The Netherlands, 12-14 December 2018, volume 313 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2018, pp. 101–110. URL: https://doi.org/10.3233/978-1-61499-935-5-101. doi:10.3233/978-1-61499-935-5-101.
- [12] E. Sulis, L. Humphreys, F. Vernero, I. A. Amantea, L. D. Caro, D. Audrito, S. Montaldo, Exploring network analysis in a corpus-based approach to legal texts: A case study, in: A. Tagarelli, E. Zumpano, A. K. Latific, A. Calì (Eds.), Proceedings of the First International Workshop "CAiSE for Legal Documents" (COUrT 2020) co-located with the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, June 9, 2020, volume 2690 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 27–38. URL: http://ceur-ws.org/Vol-2690/COUrT-paper3.pdf.
- [13] R. Nanda, L. Humphreys, L. Grossio, K. J. Adebayo, Multilingual legal information retrieval system for mapping recitals and normative provisions, in: V. Serena, J. Harasta, P. Kremen (Eds.), Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, volume 334 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 123–132. URL: https://doi.org/10.3233/ FAIA200856. doi:10.3233/FAIA200856.
- [14] D. Devyatkin, A. Sofronova, V. Yadrintsev, Revealing implicit relations in russian legal texts, in: S. O. Kuznetsov, A. I. Panov, K. S. Yakovlev (Eds.), Artificial Intelligence -18th Russian Conference, RCAI 2020, Moscow, Russia, October 10-16, 2020, Proceedings, volume 12412 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 228–239. URL: https://doi.org/10.1007/978-3-030-59535-7_16. doi:10.1007/978-3-030-59535-7_16.
- [15] J. Opiła, T. Pełech-Pilichowski, Visual analysis of similarity and relationships between legal texts, in: M. Koricic, K. Skala, Z. Car, M. Cicin-Sain, V. Sruk, D. Skvorc, S. Ribaric, B. Jerbic, S. Gros, B. Vrdoljak, M. Mauher, E. Tijan, T. Katulic, P. Pale, T. G. Grbac, N. F. Fijan, A. Boukalov, D. Cisic, V. Gradisnik (Eds.), 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020, Opatija, Croatia, September 28 -October 2, 2020, IEEE, 2020, pp. 1482–1487. URL: https://doi.org/10.23919/MIPRO48935. 2020.9245296. doi:10.23919/MIPRO48935.2020.9245296.