Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190312

ERST: Leveraging Topic Features for Context-Aware Legal Reference Linking

Sabine WEHNERT [1], Gabriel CAMPERO DURAND and Gunter SAAKE
University of Magdeburg, Germany

Abstract. As legal regulations evolve, companies and organizations are tasked with quickly understanding and adapting to regulation changes. Tools like legal knowledge bases can facilitate this process, by either helping users navigate legal information or become aware of potentially relevant updates. At their core, these tools require legal references from many sources to be unified, e.g., by legal entity linking. This is challenging since legal references are often implicitly expressed, or combined via a context. In this paper, we prototype a machine learning approach to link legal references and retrieve combinations for a given context, based on standard features and classifiers, as used in entity resolution. As an extension, we evaluate an enhancement of those features with topic vectors, aiming to capture the relevant context of the passage containing a reference. We experiment with a repository of authoritative sources on German law for building topic models and extracting legal references and report that topic models do indeed contribute in improving supervised entity linking and reference retrieval.

Keywords. reference linking, entity resolution, topic models, information retrieval

1. Introduction

Nowadays, institutions and businesses face the challenge of understanding the implications of legal changes, as they occur. Often multiple experts for each jurisdiction monitor a broad spectrum of legal texts, which is a challenging task. In such work, the context of a legal entity and the current situation are determining the applicability of laws. Legal knowledge bases support users in understanding such contexts, drawing out their implications. However, the development of such systems is complex, since they often rely on hand-crafted domain knowledge, thus do not scale well and are difficult to maintain. Explainable machine learning methods are a promising alternative, as they can be efficient in large data analysis. In previous work [1], we introduced a method of extracting bottom-up domain knowledge from legal literature. This approach allowed us to leverage a diverse array of authoritative resources in the field, supporting our main goal of capturing context-dependent application of laws, by using keywords, chapter and section titles in the proximity of a cited law. In our work the extracted knowledge is represented by several concept hierarchies (one per book). Hierarchies need to be aligned, allowing the complete information about entities, as spread across the diverse information sources, to

[1] Corresponding Author: Sabine Wehnert, University of Magdeburg; E-mail: sabine.wehnert@ovgu.de. The work is supported by Legal Horizon AG, Grant No.:1704/00082 and by the DFG (Grant No.: SA 465/50-1).

be connected. For this linking task, we focus in this work on legal citations and refer to these entities as *references*. Fig. 1 depicts differently complex ways in which legal references can be found. Detecting these references, and linking them across sources is one of the core challenges in developing a bottom-up legal information system. The first kind of references are *exact* matches without the need for complex identification procedures. The second level represents *approximate* matches between references to the same section, where one reference can be more specific. Given the variability of references, applying approximate string matching could be cumbersome. The third level refers to references with a specified *range*, so that all elements within that range also need to be identified, as mentioned. The most complicated level are references comprised of multiple laws forming statute chains, with their sections indicated by one of the previous three levels. These references are only relevant in certain contexts. In addition, there might be references expressed in informal language, which refer implicitly to certain laws. These references too are highly complex and need to be identified to use the information about them. In sum, the different levels of complexity in expressing legal references pose a challenge for linking references and building legal information systems. In this paper we evaluate a machine learning solution to handle references, considering the complexity levels. We focus on two tasks: First, identifying references pointing to the same legal text (e.g., a norm) and second, retrieving valid references for a given context. In supporting these tasks, we study the applicability of topic models. Retrieving characteristic keywords for a document within a corpus is often solved by topic modeling. After grouping the documents into a given number of topics, the elements within each topic share common characteristics represented by their likelihood of containing certain keywords. In this paper we evaluate whether there are benefits to the two aforementioned tasks by extending each identified reference with a topic model vector corresponding to the text window in which the reference occurred. Our contributions are methodological and summarized by:

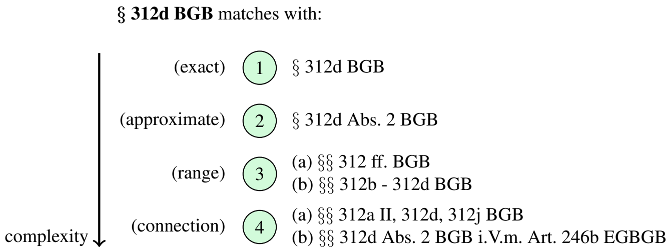

Figure 1. Complexity levels of legal reference patterns, illustrated by a reference to section 312d in the German civil code (BGB). Our linking task consists of detecting other references pointing to the same section.

- First, we identify requirements for context-dependent legal reference linking: **E**xplainality, **R**eliability, **S**tability and **T**opical Relevance **(ERST)**.
- Second, we extract and resolve legal references found in German legal literature, in a supervised setting, showing the usefulness of adding topic modelling features. We report, across several types of classifiers, that topic features assist in legal entity resolution, when combined with standard features.
- Third, we combine traditional retrieval methods with topic features for legal reference retrieval, in an unsupervised retrieval setting. We report that topic features can indeed improve the relevance of returned laws with respect to the context.

The remainder of this paper is structured as follows: In Sec. 2, we introduce our requirements for bottom-up knowledge base alignment. We then describe our method of using topic modeling features to improve entity resolution and retrieval. Sec. 3 contains results of two experiments, and their respective evaluation. In Sec. 4, we build connections to other research covering the role of rule-based legal document annotation, similarity functions for legal entity resolution, probabilistic topic modeling and legal information retrieval. In Sec. 5 we conclude our work, motivating future research directions.

2. ERST Requirements for Legal Reference Management

In our compliance checking use case, we have a high recall requirement because missing even a single regulation can lead to high costs. In fact, building on this overriding requirement, we can identify a set of requirements for bottom-up knowledge base alignment: explainability, reliability, stability and topical relevance.

Explainability: We require *explainability* (i.e., the outcome of an application can be reasonably interpreted) in two regards: ground-truth generation and in the actual application. While a common demand is the explainability of applications, the ground truth which is used to train the algorithms should also contain an explanation (e.g., for the target label). The intuition behind this requirement is to provide enough resources to understand the original thought process leading to the label. This assists in feature engineering and designing applications that can offer the same level of explanation as the ground truth. If the ground truth is generated with rules, explainability can be easily achieved by indicating the rule which generated the instance. Another aspect of explainability are the features used by the application. While feature importance is easily determined in trained models, the choice of features can also be based on explainability.

Reliability: The purpose of legal entity resolution is the matching of legal named entities, such as person, organization, location and reference to their mentions in natural language text. More precisely, we frame this as a linking task of recognizing whether two mentions refer to the same entity. We distinguish legal reference entity types from other entities because the amount of variation in the citation pattern is not only restricted to common resolution cases, such as the use of abbreviations compared to the whole word. Legal references can be very specific, occasionally pointing to a part of a sentence in an article's paragraph. Our goal is to resolve references on an article basis, despite differences in citation granularity, see Figure 1. We name the requirement from our similarity function to properly convey matches, giving high recall, as *reliability*.

Stability: Given a collection of real-world documents, it is natural to assume that they could by grouped by underlying semantic themes. Topic modeling is a broad term that covers a series of statistical methods to describe documents according to such latent semantic groups. Through such methods each document in a collection can be described as a multinomial distribution over a number of discrete topics, while topics themselves are represented as multinomial distributions over a series of keywords. As a consequence, modeled topics can be compared by their probability of including given keywords, and documents can be compared and grouped by their probability of including a given topic. Some popular methods for building topic models are Latent Semantic Analysis, Latent Dirichlet Allocation (LDA), Correlated Topic Models and Non-negative Matrix Factorization. Building a topic model multiple times on the same corpus can lead to very different results: deviations in the top keywords per topic and their rank. For example, the use of Gibbs Sampling, Expectation Maximization or Variational Bayesian Inference for approximately inferring the distribution parameters that characterize an LDA model, are expected to converge to stable & reproducible results, however this might fail to occur based on the training (e.g., its duration) and model configuration (e.g. the number of topics chosen). Such variation is not desired for legal entity linking because a variation in topic quality can affect the overall use and interpretability of topic features. We therefore require measures to ensure *stability* when topic features are used. Some starting points to assist in studying the stability of topic modeling are hyperparameter optimization strategies [2] and evaluations over repeated runs (measuring coherence, perplexity).

Topical Relevance: Statutes are written in abstract, legal jargon to be applicable to many situations. The previously described references containing statute chains are only relevant in few situations. Having those references in a knowledge base, our goal is to align them only to those references that are sharing highly related concepts, thus satisfying a requirement we call *topical relevance*. An example of topical relevance is the reference "§ 286 BGB i.V.m. § 280 BGB", which specifies the breach of a duty combined with a default of the obligation. There are many contexts, in which those regulations can apply, such as the non-issuance of a job reference or the failed transfer of an asset against the negotiated terms. Those two situations occur in different settings, so that the topical connection to other references concerning labour law is only given in the former. For our use case of context-aware legal reference linking, the *topical commonalities* between the surrounding contexts of two references determine the likelihood of a connection, regardless of reference type. Since we consider a bottom-up knowledge acquisition process of concepts related to legal references, the contexts are available in natural language.

Legal Reference Management under ERST: In the following, we explain our methodology for legal reference management enhanced with topic models. We elaborate upon the legal reference resolution task, and then show how we enable context-aware legal reference retrieval. Figure 2 illustrates the workflow. First, we preprocess the legal literature corpus *(1)* and a document, which is compared to the remaining corpus to detect matching reference pairs. This document contains besides the legal references also the context in which they are considered. In the second step, we apply topic modeling on the literature corpus *(2a)* and annotate laws *(2b)* in the query document. After the annotation, context windows *(3)* around all legal references are extracted. Then, we use those context windows to infer a feature vector *(4a)* with the topic model. The references themselves are also featurized *(4b)* regarding the capitalization of the first token *(CAP)*, the length of the whole reference *(LEN)*, the type of reference *(TYPE)* and the token set similarity *(TSS)*. Finally, we train a model to link legal references *(5a)* using the features and can retrieve *(5b)* contextually relevant laws. We satisfy the requirement for *explainability* by first, using a rule-based approach for document annotation and second, identifying matching entities with rules for generating the ground truth in legal entity resolution. The purpose of this experiment is not to replicate the ground truth which is limited to the patterns indicated by the rules, but to analyse how well the models perform for differently complex types. Considering *reliability*, the similarity between two strings shall be detected regardless of length (due to differences in granularity) and order (due to different citation styles). For this, we apply token set similarity, as done by Cohen [3], for reference string comparison. This method is comparing the intersection (t_0) and the remainders of two sorted sets of strings (t_1, t_2) concatenated with t_0 against each other.

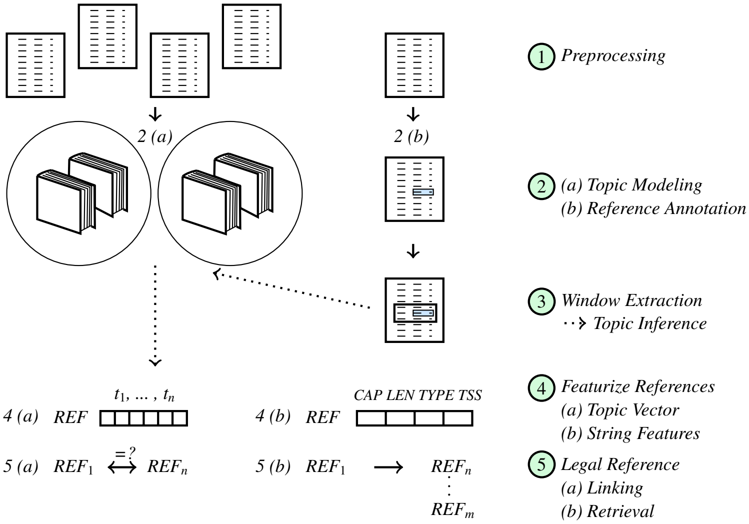

Figure 2. Overview of the featurization workflow for reference linking and retrieval.

The strings can have a different length because the comparison is allowed to end at the length of the shorter sequence. The *token set similarity (TSS)* is computed as follows:

$$\text{TSS} = \max\left(\frac{|t_i| + |t_j| - \mathscr{L}}{|t_i| + |t_j|}\right), \quad i,\ j \in \{0,1,2\}, \tag{1}$$

where $|t_i|$ and $|t_j|$ are placeholders for the strings to be compared and $\mathscr{L}$ is the Levenshtein distance [4] between the strings. We consider the length of the intersection between both strings $|t_0|$, and the lengths of the two strings, $|t_1|$ and $|t_2|$, respectively. Three combinations are compared, $(t_0,\ t_1)$. $(t_0,\ t_2)$ and $(t_1,\ t_2)$ and the maximum score is the *TSS*. When we compute the token set similarity for the example strings from Figure 1, a score of 100 is returned for all strings except for 3 (a), where a score of 78 is obtained. In this case, the character "d" is missing and thus a match between §§ *312 ff. BGB* and § *312d BGB* is implied by the abbreviation *ff.*, referring to the following articles until the end of the section. That shows that the token set similarity is well suited for partial string matching irrespective of string length. For harder cases, such as 3 (a) (i.e., with the use of *ff.*), background knowledge is needed to resolve the correct number of regulations following. It is worth noting that we do not consider token order. That assumption may not hold for references of type 4 (b) (i.e., combinations), where the connections between the law books (e.g., BGB and EGBGB) and the respective section numbers (e.g., 312d and 246b) should not be lost. For those cases, the substrings of each reference can be matched separately with the token set similarity. Aside from the token set similarity, topic features are used for entity resolution because we assume semantically overlapping content across books. Having topic models in a productive setting, they shall be optimized regarding *stability*, in order to be interpretable and maintainable. For this, we refer to the different techniques summarized in Section 4. Given those preconditions, *topical relevance* can be a helpful indicator for identifying references pointing to the same entity in similar contexts. For the specific task of pair-wise classifier-based entity resolution, where classifiers are responsible for predicting if a pair is a match or

Table 1. Distribution of reference types based on extraction rules.

Norms	Court Decisions	EU-Directives	EU-Regulations	Combinations
83,661	4,277	3,869	730	122

non-match (i.e., same entity or not), traditionally similarity/distance-based features are used. In our case, we employ as features the absolute difference between topic vectors of the paired instances. Together with features capturing the capitalization of the first token *(CAP)*, the length of the string *(LEN)* and the type of reference *(TYPE)* - as shown in Table 1, the feature vector for entity resolution is formed. We train common classifiers on the binary classification problem. Another perspective on knowledge base alignment is the retrieval task. Here, we detect references with the same surface features using *TSS* and rerank the instances based on a reference context obtained from a query. The size of the context depends on the density of entities found in the corpus. We lemmatize the tokens and infer the topic vectors using the LDA model for each reference context. We reorder all retrieved references with the topic vector distance to the querying reference, thus increasing *topical relevance*.

3. Results and Evaluation

Evaluation Setup: For our experiments, we use a corpus of 193 German books which we manually grouped into 30 categories by their title, such as IT Security Law, Labour Law and Commercial Law. To obtain a similar granularity from the topic model, we run LDA for 200 iterations with a standard parameter configuration, setting the number of topics to 30. We specifically select an approach to LDA supported by Variational Bayes optimization as proposed by Hoffman et al [5], offering a reasonable runtime to facilitate repeated studies[2]. Table 2 provides representative words for the obtained topics from LDA. Notable outlier topics are criminal activities (28), chemicals (29) and consumers (30). There are significant overlaps between many topics, such as Credits (6) and Patent law (7). Since we could give all the topics an unambiguous label, we refrain from further optimization in this study. For optimal results and in productive settings, we nevertheless recommend to optimize LDA regarding topic stability (see Section 4). We adapted the reference extraction rules from previous work [1] to the Apache UIMA Ruta annotation tool[3] and extended them to other reference types [4]. Empirical checks resulted on average at roughly 90% reference coverage. Our reference annotation patterns are based on regular expressions and constrained by part-of-speech tags (POS). Hence, we obtain a distribution of references, as shown in Table 1. The 83,661 found norms cover patterns similar to the examples 1 - 4 (a) in Figure 1. We extracted 4,277 court decisions, such as *"EuGH NJW 2006, 2465"*. Most of the 3,869 entities of type EU-Directive occur in the following shapes: *"RL 29/2005/EG"* or *"Richtlinie über den elektronischen Geschäftsverkehr"*. Among the 730 EU-Regulations, common forms are *"VO 267/2010/EU"* and *"Verordnung über die Freizügigkeit der Arbeitnehmer"*. Combinations can contain all reference types, separated by an *"i.V.m."* (meaning: "in connection to"), see type 4 (b) in Figure 1.

[2] Gensim multi-core LDA: https://github.com/RaRe-Technologies/gensim

[3] https://uima.apache.org/ruta.html

[4] Implementation: https://github.com/anybass/HONto/tree/master/reference_linking

Table 2. Topics, given names and representative words in our corpus

Topic Nr.	Given name	Representative words
1	*International business*	abs, bgb, hgb, europäisch, bag, unternehmen, arbeitnehmer, betrvg, mitgliedstaat, corporate, kosten, international,...
2	*Compliance*	unternehmen, dabei, sowei, management, neu, daten, compliance, hoch, weit, beispiel, stellen, informationen,...
3	*Employment law*	bag, arbeitnehmer, arbeitgeber, betrvg, nza, bgb, kündigung, betriebsrat, arbeitverhältnis, gelten, tarifvertrag, besehen,...
Remaining topics	*Stock Enterprises (4), Commerce (5), Credits (6), Patent law (7), European law (8), Data privacy (9), Energy (10), Trade taxes (11), Income taxes (12), Traffic/Infrastructure (13), Business ethics (14), Commercial code (15), Insurance (16), Environment (17), Vacations/Working hours (18), Cyber-security (19), Control mechanisms (20), Stock market (21), Business taxation (22), Health (23), E-mobility (24), Audits (25), Online communication (26), Corporate governance (27), Criminal activities (28), Chemicals (29), Consumers (30)*	

Experiment 1 (Legal Reference Resolution): Following the steps in Fig. 2, we identify 92,659 references in our aforementioned legal literature corpus, corresponding to a natural occurrence of references to different types of legal entities (as shown in Table 1), and of the different complexity levels described in Fig. 1. Based on domain rules and an extent of manual verification, we identify 7,459,674 pairwise matches (i.e., only 0.173% of all possible matches). We split these matched pairs into training and test data (66%, 33%), randomly sampling from the non-match classes until the same number of items as the matched class is reached per split (i.e., for having balanced examples), and checking that non-matched pairs are not repeated. This leads to a test-train split of 5,307,699 / 9,576,279 labeled items. In terms of features, we enhance each reference with a topic vector that captures the probability of topic assignations using the window of 200 characters surrounding a reference (rounded up to complete words). Next, we use string features (i.e., CAP, LEN, TYPE and TSS, as mentioned in Section 2)), topic features (the absolute difference on each dimension of the topic vectors of the paired references) and a combination of topics and standard features. Table 3, shows the features as standard, topic model and combined, respectively. We evaluated the contribution of each feature for the supervised entity linking task; so we selected 4 different classifiers. As a baseline we use a Gaussian Naive Bayes (GNB) classifier (with no priors on the class distributions), due to its simplicity and few requirements on hyperparameters. We select random forest-based methods: XGBoost (XGB, eta: 0.3, max depth: 6, alpha: 0, lambda: 1), AdaBoost (Ada, decision-tree-based, max depth: 1, 50 estimators, lr: 1) and RandomForest itself (RF, with bootstrapping, using GINI criteria, min samples for split: 2, no depth limitations), due to their computational efficiency and potential for explainability. The overall F1 score shows a consistent trend of improving with topic model features and the combination with the standards. GNB performs the worst. RF performs the best, followed by XGB. When considering the scores of the entity types, it is shown that topic features alone cannot bring improvements in several of the classifiers evaluated. The only cases where the combination of feature types brings disadvantages are for our weakest classifer (GNB), or for the combination reference types, which constitute a little-represented class. Though the RF combined model is overall the best, with a consistent performance

		F1-Score	Norms	Court Dec.	EU-Dir.	EU-Reg.	Comb.
Standard	XGB	0.81	0.84	0.79	0.85	0.91	0.91
	Ada	0.81	0.83	0.77	0.83	0.91	0.91
	RF	**0.86**	**0.87**	**0.86**	**0.90**	**0.98**	**0.98**
	GNB	0.69	0.57	0.78	0.72	0.77	0.75
Topic model	XGB	0.83	0.83	0.76	0.86	0.84	0.79
	Ada	0.81	0.82	0.76	0.87	0.84	0.77
	RF	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**
	GNB	0.77	0.75	0.54	0.75	0.69	0.68
Combined	XGB	0.89	0.90	0.93	0.91	0.94	0.89
	Ada	0.89	0.89	0.79	0.93	0.91	0.88
	RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00!**	**1.00!**
	GNB	0.78	0.76	0.78	0.74	0.76	0.78

Table 3. F1-score and entity type-based accuracy for supervised legal entity resolution on our dataset, considering different types of features and classifiers. The exclamation mark indicates zero mislabeled entities.

across all classes, we note that in spite of having a grouped F1 score of 1.00, 6,581 norm instances, 3 court decisions and 16 EU-Directives were part of mislabeled pairs (out of the 5M tested pairs). Results suggest that there is room for improvement and serving better the less represented reference type is important for our approach to contribute to the overall performance of reference linking. Common error causes are ranges, missing whitespaces, errors from extraction rules and different citation granularities.

Experiment 2 (Retrieval of Context-Dependent Reference Connections): In this experiment, we test whether topic features can help to increase the relevance of retrieved references. We frame the task with a reranking objective and compute the distance between the topic vectors via *Jensen-Shannon Divergence (JSD)* [6] and *Maxium Absolute Difference (MAD)*. The Mean Absolute Difference behaved similar to JSD in our earlier experiments, so that we employ MAD instead. We randomly draw 14 queries from 122 references of the narrow context reference group 4 (b) (see Figure 1) consisting of the topic features and the reference. For these queries, we use TSS to generate candidates from all references and compute the topic-based distances. The ground truth is created by manually assigning a binary relevance label to all references returned by TSS (ranging from 8 to 246 hits), given their natural language contexts. We use r-Precision for evaluation, which returns the precision at position r where all relevant documents have been retrieved. Results indicate that it is worthwhile to rerank the data with topic features to obtain more relevant output. The best individual score was MAD with 62.3%, followed by JSD with 60.4%. The *Term Frequency - Inverse Document Frequency (TFIDF)* baseline yielded a score of 51.6%. A combination of MAD with TFIDF achieves the best r-Precision of 63%, whereas all metrics combined achieve a lower score of 61.1%. We observe a variance of the r-Precision regardless of the amount of candidates, that we attribute to a different granularity of the relevance label that the topic model does not serve.

Table 4. r-Precision based on JSD, MAD and TFIDF and combinations on 14 queries over our dataset.

rP(TFIDF)	**rP**(JSD)	**rP**(MAD)	**rP**(TFIDF, JSD)	**rP**(TFIDF, MAD)	**rP**(TFIDF, JSD, MAD)
0.516	0.604	0.623	0.596	0.630	0.611

4. Related Work

With regards to the *ERST* requirements, we explore related research, covering topic models for entity resolution and legal information retrieval. Topic models are used for entity resolution for more than a decade, see similar work on Wikipedia by Pilz et al. [7], as well as a latent dirichlet model by Bhattacharya et al. [8]. We find that topic features are a suitable technique for context-aware legal reference linking. Considering *explainability*, Glaser et al. [9] develop a system for German legal texts which disambiguates named entities to semantic roles using templates. Similar to their work, we extracted legal reference entities by using rule-based methods in Apache UIMA Ruta. Legal named entity recognition and resolution has been studied by Dozier et al. [10] for entities of judges, attorneys, companies, jurisdictions and courts. They apply well-founded techniques for resolution, such as blocking and *reliable* string similarity metrics for each entity type and train a Support Vector Machine (SVM) classifier. Van Opijnen et al. perform legal entity linking by using national and European Law Identifiers (ELI), which we consider for follow-up work [11]. Computing entity context similarity based on word embeddings is a state-of-the-art approach, but it can hardly be interpreted. Traditional bag-of-words representations often oversimplify sensitive natural language tasks. We consider features from topic models to be a viable trade-off between both worlds. Topics capture the contextual use of words and distances at this level of abstraction are well interpretable, as shown by Yurochkin et al. [12]. They define the *Hierarchical Optimal Topic Transport (HOTT)* measure, based on the Word Mover's Distance [13] between the word distribution per topic and the optimal transport between documents as distributions of topics. The topic model LDA uses a random inference process and thus suffers from instability. Many authors have addressed *stability*, e.g., by proposing a combination with non-negative matrix factorization [2,14] or a search-based parameter optimization using differential evolution [15]. LDA performance is strongly affected by hyperparameter tuning, therefore for each corpus a different setting is recommended [15]. When the corpus is extended in the future, the topics of new documents are inferred from the existing model, or a new topic model can be computed, optionally with must-link and cannot-link constraints to preserve the original structure [16]. Considering *topical relevance*, there are similar challenges in legal information retrieval in identifying the same application context of legal references. The system by Kim et al. [17] is based on well-known retrieval methods: stopword removal, lemmatization and TFIDF. A common problem occurs when there is no lexical overlap between the query and the statute. Although word embeddings and their newer contextual variants (e.g., XLNet [18]) may be a solution to this problem, they need to be adapted to the legal terminology and trained on a sufficiently large corpus.

5. Conclusion and Future Work

In this work, we pose four requirements for bottom-up knowledge base alignment: explainability, reliability, stability and topical relevance. We describe how those requirements can be fulfilled and perform experiments on legal reference linking and contextual retrieval. We find a benefit of using topic feature vectors with standard similarity metrics for legal entity linking, which can generate further viable candidates for contextual retrieval. Hence we validate the methodological choice of leveraging topic models trained on legal literature, for creating contextual features for reference linking and retrieval. A common challenge for feature creation of domain-specific text data is the absence of word embeddings trained on a representative corpus; our topic feature vectors are a viable choice for smaller corpora. Combining topic models with word embeddings, e.g., using the *HOTT* method by Yurokchin et al. [12] can be worthwhile to investigate. Regarding supervised reference linking: Blocking and understanding better the behavior for less represented types are good avenues for continuing this research. Current approaches for knowledge base alignment use graph and word embeddings, which we want to test in follow-up work [19].

References

[1] S. Wehnert et al., Concept Hierarchy Extraction from Legal Literature, in: *Proceedings of the ACM CIKM 2018 Workshops*, CEUR-WS.org, 2018, to appear.

[2] D. Greene et al., How many topics? Stability analysis for topic models, in: *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part I*, Springer-Verlag, 2014, pp. 498–513.

[3] A. Cohen, FuzzyWuzzy: Fuzzy string matching in python, 2011. Retrievedfromhttps://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/.

[4] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: *Soviet physics doklady*, Vol. 10, 1966, pp. 707–710.

[5] M. Hoffman et al., Online learning for latent dirichlet allocation, in: *advances in neural information processing systems*, 2010, pp. 856–864.

[6] J. Lin, Divergence measures based on the Shannon entropy, *IEEE Transactions on Information theory* **37**(1) (1991), 145–151.

[7] A. Pilz et al., Named Entity Resolution Using Automatically Extracted Semantic Information., in: *LWA*, 2009, p. KDML–84.

[8] I. Bhattacharya et al., A latent dirichlet model for unsupervised entity resolution, in: *Proceedings of the 2006 SIAM International Conference on Data Mining*, SIAM, 2006, pp. 47–58.

[9] I. Glaser et al., Named entity recognition, extraction, and linking in german legal contracts, in: *Internationales Rechtsinformatik Symposium*, 2018.

[10] C. Dozier et al., Named entity recognition and resolution in legal text, in: *Semantic Processing of Legal Texts*, Springer, 2010, pp. 27–43.

[11] M. Opijnen et al., Beyond the experiment: the eXtendable legal link eXtractor, in: *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL)*, 2015.

[12] M. Yurochkin et al., Hierarchical Optimal Transport for Document Representation, *arXiv preprint arXiv:1906.10827* (2019).

[13] M. Kusner et al., From word embeddings to document distances, in: *International conference on machine learning*, 2015, pp. 957–966.

[14] M. Belford et al., Stability of topic modeling via matrix factorization, *Expert Systems with Applications* **91** (2018), 159–169.

[15] A. Agrawal et al., What is wrong with topic modeling? And how to fix it using search-based software engineering, *Information and Software Technology* **98** (2018), 74–88.

[16] Z. Zhai et al., Constrained LDA for grouping product features in opinion mining, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2011, pp. 448–459.

[17] M.-Y. Kim et al., Statute Law Information Retrieval and Entailment, in: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ACM, 2019, pp. 283–289.

[18] Z. Yang et al., XLNet: Generalized Autoregressive Pretraining for Language Understanding, *arXiv preprint arXiv:1906.08237* (2019).

[19] B.D. Trisedya et al., Entity alignment between knowledge graphs using attribute embeddings, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 297–304.