GerDaLIR: A German Dataset for Legal Information Retrieval

Marco Wrzalik and **Dirk Krechel**
RheinMain University of Applied Sciences, Germany
{firstname.lastname}@hs-rm.de

Abstract

We present GerDaLIR, a **Ger**man **Da**taset for **L**egal **I**nformation **R**etrieval based on case documents from the open legal information platform *Open Legal Data*. The dataset consists of 123K queries, each labelled with at least one relevant document in a collection of 131K case documents. We conduct several baseline experiments including BM25 and a state-of-the-art neural re-ranker. With our dataset, we aim to provide a standardized benchmark for German LIR and promote open research in this area. Beyond that, our dataset comprises sufficient training data to be used as a downstream task for German or multilingual language models.

1 Introduction

There are few non-English datasets dedicated to Natural Legal Language Processing (NLLP) or Legal Information Retrieval (LIR). To our knowledge, not a single dataset exists that provides a standardized benchmark for LIR models on the German language. To this end we contribute GerDaLIR, a legal document retrieval dataset comprising a large document collection and corresponding queries forming a document ranking task. We provide a large amount of training data such that both unsupervised and supervised methods can be benchmarked. This also enables GerDaLIR to be used as a downstream task for German or multilingual language models. The task provided is a precedent retrieval task. As illustrated in Figure 1, we build GerDaLIR by extracting passages that reference other cases. For that we utilize 201,825 cases from the open legal information platform *Open Legal Data* (Ostendorff et al., 2020). We present baseline experiments on classic term-based retrieval methods, a semantic search approach based on word embeddings and a transformer-based re-ranker giving an orientation to other researchers using our dataset. In contrast to other LIR datasets based on precedent case retrieval, GerDaLIR offers the following unique characteristics:

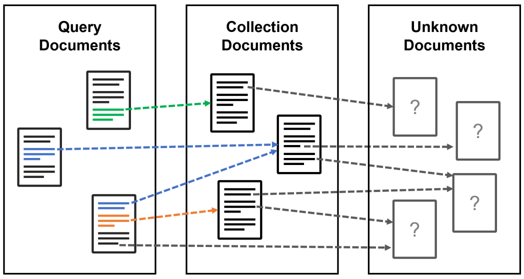

Figure 1: GerDaLIR comprises query-document pairs from passages that cite known collection documents.

Large Corpus Size. With a total of 144K relevance labels for 123K query passages and a collection of 131K documents comprising over 3M passages, GerDaLIR is – to our knowledge – bigger than any other LIR dataset. Its size enables GerDaLIR to be used for full-ranking evaluation.
German Language. The German language is GerDaLIR's most prominent feature and fills a gap in the community. Furthermore, in combination with the large training set, GerDaLIR can be used as a downstream task for German or multilingual language models, of which there are quite few.
Query Passages. Most other LIR datasets based on precedent retrieval provide entire case documents to be used as queries. It may be unclear to which part of a given case relevant cases should be retrieved. As GerDaLIR provides query *passages* rather than whole documents, it better reflects a practical use case.

The download links and descriptions to the format of GerDaLIR can be accessed via GitHub[1].

2 Other Datasets

There are several sources of LIR datasets and tasks. In the following, we outline those based on precedent retrieval.

[1] https://github.com/lavis-nlp/GerDaLIR

Natural Legal Language Processing Workshop 2021, pages 123–128
November 10, 2021. ©2021 Association for Computational Linguistics

COLIEE 2020 Task 1. The *Competition on Legal Information Extraction/Entailment* (COLIEE) is a workshop that annually provides a number of tasks in the areas of legal document retrieval, question answering and entailment. COLIEE 2020 Task 1 (Rabelo et al., 2020) is a re-ranking task that provides for training: "520 base cases, each with 200 candidate cases from which the participants must identify those that should be noticed with respect to the base case". From the total of 104,000 candidates, 2,680 are labelled as positive. For testing they provide 130 base cases, a total of 26,000 candidates and 646 positive labels. The case documents are written in English and originate from the Federal Court of Canada. The workshop also provides tasks derived from the Japanese jurisdiction including original Japanese texts as well as English translations.

SigmaLaw. This dataset originates from the paper *Legal Document Retrieval using Document Vector Embeddings and Deep Learning* (Sugathadasa et al., 2018). It comprises 2,500 case documents and a citation graph indicating for each case which of the other cases are considered relevant.

FIRE 2017 IRLED. The *FIRE 2017 IRLED Track* presents a precedence retrieval task comprising 200 query cases, 2000 collection cases and 1000 positive relevance labels. The provided case documents originate from the Indian Supreme Court, which uses the English language in their proceedings.

3 Dataset Generation

GerDaLIR is based on parsed references in German case documents taken from *Open Legal Data*. Although precedent cases have no binding effect in the German law system, references to prior cases are very common and arguably play a big role in supporting a line of argument. By definition, such referenced cases are relevant to the case at hand. With this in mind, the idea behind GerDaLIR's task is simple: Passages containing one or more references to known cases become queries while the referenced cases are labelled as relevant. However, if a passage is used as a query, the document the passage originates from should not be used as an retrievable collection document. To achieve that, we classify case documents – as depicted in Figure 1 – into the following classes: Unknown documents belong to cases that we have seen references to, but are not part of *Open Legal Data*. Collection documents comprise the cases that will be indexed for the retrieval task, which mainly consist of those that only refer to unknown documents. Query documents are the documents from which queries are sampled. Assigned to these are cases that contain references to collection documents. If a case document refers to other query cases, but not to collection cases, it is also classified as a collection document. The case documents are divided into passages along margin numbers. It regularly happens that the references to a passage follow with a margin number. Those passages typically start with *Vgl.* ("compare") or *Siehe* ("see"). We use these and more indicator words in the beginning of a passage to detect such referential passages and assign their references to the previous passage, which is assumed to contain the corresponding statement or line of argument. The text describing the references, however, is not added to the passage, since we want models to rely on natural language rather than exploiting references or parts of them. For this reason, we attempt to replace any reference including those to statutes with a `[REF]` token. However, a small portion of references that we were unable to parse remain in the text. From the final text, we also remove any braced content, since they mostly contain comprehensively described references that are difficult to sanitize otherwise. After that we collect all passages from the query documents that are marked with references to one or more collection documents. These form a set of multi-sentence queries, each with at least one label to a relevant collection document. Finally, we perform a 0.8/0.1/0.1 split on the queries for training, development and testing respectively. The resulting size of GerDaLIR's collection, the queries and the labels are summarized in Table 1.

Table 1: GerDaLIR's dataset size

	Documents	Passages
Collection	131,446	3,095,383

	Train	Dev	Test
Queries	98,380	12,297	12,298
Pos. Labels	115,360	14,570	14,394

4 Baseline Methods

We conduct a series of baseline experiments demonstrating that GerDaLIR can be used to benchmark retrieval methods or to evaluate the language models used by them. The resulting measures also serve as an orientation to other researchers using the dataset. The methods considered are described below.

4.1 BM25 and TF-IDF

TF-IDF and BM25 are known as term-based or sparse retrieval methods. They are efficiently realized using inverted indexing. With that, they rely on exact term matches resulting in the tendency of missing relevant items. This tendency is often mitigated by employing a *stemmer* or *lemmatizer* normalizing each word to its base form. More detailed information can be found in *Introduction to Information Retrieval* by Manning et al. (2008).

4.2 Word Centroid Similarity

We introduce the *Word Centroid Similarity* (WCS), an unsupervised semantic textual similarity measure based on word embeddings (Mikolov et al., 2013). Retrieval with WCS can be described as a dense retrieval method, since a dense representation is assigned to each query, document or passage. This vector is the *centroid* or mean vector of the embeddings of the words that occur in the given text. Based on the centroids, we calculate the relevance score using the cosine similarity measure. Aggregating word embeddings for the measurement of textual similarity has been studied in the past with various aggregation methods, word embedding models and vector similarity or distance measures (Kusner et al., 2015; Glasgow et al., 2016; Rücklé et al., 2018; Landthaler et al., 2018). We include WCS as a semantic search counterpart to the term-based retrieval methods. With that we also demonstrate how GerDaLIR can be used to evaluate word embeddings in terms of their utility to information retrieval.

4.3 Neural Re-ranking

We conduct neural re-ranking experiments with a simple binary relevance classifier based on *Transformer Encoders* (Vaswani et al., 2017) that follows BERT's cross-encoding design for *sentence pair classification* (Devlin et al., 2019). Rankings result from the order of confidence with which given query-passage pairs are classified as relevant. Nogueira and Cho (2019) provide a more detailed description on the implementation of this model.

5 Experiments

In this section, we outline the experimental setup, describe the implementation of the methods described above, and briefly discuss the results.

5.1 Metrics

We measure standard information retrieval metrics for the evaluation of our baseline models. With the *mean reciprocal rank* cut off at the tenth position (MRR@10) and the *normalized discounted cumulative gain* cut off at the twentieth position (nDCG@20), we measure the ranking quality on the top positions. MRR@10 only considers the first hit and penalizes strongly for each rank below rank one while nDCG@20 takes all positively labeled documents into account and penalizes softer. Complementary to the ranking quality measures, we measure recall cut off at positions 100 and 1000 to coarsely illustrate the distribution of positive documents and the portion of documents that were missed completely.

5.2 With Passages to Document Rankings

There are various good reasons to perform passage retrieval although the actual targets are documents. In our work the reason behind this is two-fold: First, depending on the model, it could result in better rankings. Second, the model at hand might not be able to process whole documents. The neural re-ranker we use is limited to input sequences of 512 tokens. To cast passage rankings to document rankings, we map passages back to the documents they originate from and perform max-pooling on the scores along documents. However, many documents are represented by multiple passages and after pooling the lengths of the resulting document rankings are smaller than the initial passage ranking. For this reason we retrieve 2000 passages although we only analyze top-1000 document rankings. For the re-ranking, however, we utilize only the first 1000 passages as candidates (including multiple document occurrences) and cast to document ranking afterwards.

5.3 TF-IDF and BM25

We use *Elasticsearch*[2] to perform TF-IDF and BM25 retrieval. Its German analyzer includes a pre-processing pipeline that removes stop-words and performs stemming in accordance to the German language. BM25's parameters k1 and b are

[2] https://www.elastic.co/

Table 2: Baseline measures. Mode P and D denote passage-wise and document-wise retrieval (Section 5.2).

Method	Mode	MRR@10	nDCG@20	Recall@100	Recall@1000
TF-IDF	P	0.333	0.375	0.651	0.768
	D	0.336	0.386	0.701	0.809
BM25 ($k1 = 1.20, b = 0.75$)	P	0.365	0.409	0.693	0.800
	D	0.386	0.434	0.734	0.827
BM25$_{tuned}$ ($k1 = 0.51, b = 0.72$)	P	0.372	0.417	0.703	0.803
BM25$_{tuned}$ ($k1 = 0.90, b = 0.98$)	D	0.391	0.439	0.737	**0.829**
WCS – GloVe	P	0.242	0.278	0.539	0.695
	D	0.134	0.166	0.420	0.625
WCS – fastText	P	0.257	0.295	0.582	0.726
	D	0.153	0.188	0.468	0.668
Neural Re-ranking – BERT	P	0.416	0.465	**0.745**	0.789
Neural Re-ranking – ELECTRA	P	**0.436**	**0.481**	**0.743**	0.789

tuned based on the development set in the ranges of $[0.5, 2.0]$ and $[0.3, 1]$ respectively. For that we employ the Bayesian optimization algorithm provided by *Optuna*[3], with 100 trials and nDCG@20 as the metric being optimized. The default parameters and those resulted from the tuning are listed in Table 2. It is worth noting that for term-based retrieval methods, the document-wise retrieval (D) outperforms the passage-wise retrieval (P) as shown in Table 2. We hypothesize that in relevant documents, the important keywords occur more frequently throughout the entire document while in other, non-relevant documents, they occur only marginally in a few passages. This can be exploited through the term frequency in document-wise retrieval.

5.4 Word Centroid Similarity

We employ *GloVe* (Pennington et al., 2014) and *fastText* (Bojanowski et al., 2017) word embeddings in our WCS retrieval experiments. For that we train the embeddings based on the entire lowercased text from the cases in *Open Legal Data*, which comprises more than 465 million words. The training is performed using the original implementations for *GloVe*[4] and *fastText*[5]. For inference, we filter stopwords and normalize each word centroid to L2-norm before indexing in an *faiss*[6] inner product index with which a cosine similarity search index is realized. As shown in Table 2, fastText slightly outperforms GloVe. We hypothesize fastText is favorable for the German language, since it virtually realizes a compound splitter: FastText expands each word by character n-grams and calculates an aggregated representation using the n-grams and a representation for the entire word. Furthermore, if a word is out of vocabulary, there is a good chance of generating a meaningful representation using the character n-grams. The minimum and maximum size of those character n-gram are hyperparameters. We found that 5 for both minimum and maximum n-gram size perform best among various tested settings.

5.5 Neural Re-ranking

In recent years many modifications to the BERT model and its training procedure have been proposed such as ALBERT (Lan et al., 2020) or RoBERTa (Liu et al., 2019). The effectiveness gains of those modifications are often demonstrated on downstream tasks. In our neural re-ranking experiment, we compare a pretrained BERT model with a pretrained ELECTRA discriminator (Clark et al., 2020). We use the "base" variants of BERT and ELECTRA trained by Chan et al. (2020), which can be accessed via *Hugging Face*[7] with the identifiers `deepset/gbert-base` and `deepset/gelectra-base`. During fine-tuning we use top-100 BM25 rankings from which we randomly sample one negative candidate for each positive example. The pre-trained models are trained for 100 epochs on GerDaLIR's training

[3] https://optuna.org/

[4] https://github.com/stanfordnlp/GloVe

[5] https://fasttext.cc/

[6] https://faiss.ai/

[7] https://huggingface.co/deepset/

data with a learning rate of 1e-4 and an effective batch size of 768 samples (e.g. batches of 16 samples and 48 gradient accumulation steps for each update step). The final models are tested based on top-1000 passage rankings from BM25 as candidates. To those, the score max-pooling is not applied, but on the final re-ranked rankings. Therefore, many documents are represented by multiple passages and the final rankings are much shorter than 1000 documents, which negatively affects recall@1000. Due to the sequence length limitation, document-wise retrieval can not directly be performed with BERT or ELECTRA. Passages that exceed this limitation are divided along sentence boundaries, and the maximum score is applied to the passage. As shown in Table 1, the use of the ELECTRA model results in higher re-ranking quality in terms of MRR@10 and nDCG@20 compared to the BERT model, which is consistent with external experiments on other downstream tasks (Clark et al., 2020; Chan et al., 2020).

6 Conclusion & Future Work

We present GerDaLIR, a dataset filling the gap of a standardized IR benchmark for the German legal domain. We provide several baselines with which other researchers can compare their results. Our experiments demonstrate the use of GerDaLIR as a downstream task for German or multilingual language models. In future work, we plan to investigate the importance of in-domain pre-training to neural LIR models. We also intend to explore unsupervised methods that effectively leverage language models for domain-specific information retrieval, as well as approaches combining these with traditional term-based retrieval methods.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6788–6796. International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kimberly Glasgow, Matthew J. Roos, Amy J. Haufler, Mark A. Chevillet, and Michael Wolmetz. 2016. Evaluating semantic models with word-sentence relatedness. *CoRR*, abs/1603.07253.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jörg Landthaler, I Glaser, E Scepankova, and F Matthes. 2018. Semantic text matching of contract clauses and legal comments in tenancy law. In *Tagunsband IRIS: Internationales Rechtsinformatik Symposium*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 385–388, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. COLIEE 2020: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15-17, 2020, Revised Selected Papers*, volume 12758 of *Lecture Notes in Computer Science*, pages 196–210. Springer.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *CoRR*, abs/1803.01400.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2018. Legal document retrieval using document vector embeddings and deep learning. In *Science and information conference*, pages 160–175. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.