Searching for legal documents at paragraph level: Automating label generation and use of an Extended Attention Mask for boosting neural models of semantic similarity

Li Tang
Institute of Computational Linguistics
University of Zurich
li.tang@uzh.ch

Simon Clematide
Institute of Computational Linguistics
University of Zurich
simon.clematide@uzh.ch

Abstract

Searching for legal documents is a specialized Information Retrieval task that is relevant for expert users (lawyers and their assistants) and for non-expert users. By searching previous court decisions (cases), a user can better prepare the legal reasoning of a new case. Being able to search using a natural language text snippet instead of a more artificial query could help to prevent query formulation issues. Also, if semantic similarity could be modeled beyond exact lexical matches, more relevant results can be found even if the query terms don't match exactly. For this domain, we formulated a task to compare different ways of modeling semantic similarity at paragraph level, using neural and non-neural systems. We compared systems that encode the query and the search collection paragraphs as vectors, enabling the use of cosine similarity for results ranking. After building a German dataset for cases and statutes from Switzerland, and extracting citations from cases to statutes, we developed an algorithm for estimating semantic similarity at paragraph level, using a link-based similarity method. When evaluating different systems in this way, we find that semantic similarity modeling by neural systems can be boosted with an extended attention mask that quenches noise in the inputs.

1 Introduction

Retrieving and understanding legal documents can take a considerable amount of time, even for legal professionals. Searching a large database of previous cases (court decisions) that apply to a particular situation is an important part of legal work, as performed by lawyers and their legal assistants (Bhattacharya et al. 2019, Bhattacharya et al. 2020, Draijer 2019, Hafner 1980, Xiao et al. 2019, Zhong et al. 2020). Such a retrieval system may not only help expert users with deep expertise in the domain but can also benefit non-expert users. The latter can, for example, try to develop a preliminary understanding of a situation, when selecting or approaching a lawyer, to better navigate the justice system (Bhattacharya et al. 2019, Boniol et al. 2020, Chen et al. 2013, Tran et al. 2020). Both expert and non-expert users may then be interested in a) identifying where the legal problem fits in regarding legal concepts, b) what legal actions are potentially relevant, and c) what were the outcomes of similar cases (Bhattacharya et al. 2019, Hafner 1980, Zhong et al. 2020).

The above problem can therefore be formulated as a specialized Information Retrieval task that involves searching through a corpus of legal documents, before ranking them according to their relevance. While the concept of relevance in this context is conceptually and practically complex, we will focus our attention on a key aspect, the semantic similarity between a query and a candidate hit (Shao et al. 2020). Early attempts at developing such systems often involved the use of Boolean expressions as queries, where a document

Natural Legal Language Processing Workshop 2021, pages 114–122

either matches or does not match a query, in accordance with the Boolean logic of a statement being either *True* or *False* (Manning et al. 2009). In the case of large document collections, the resulting number of matching documents can then easily exceed the number a human user could possibly sift through, resulting in the need to offer a rank ordering of results (Manning et al. 2009). The Boolean approach also suffers problems related to human error and cognitive load in query formulation, as the user must translate a meaningful query formulated in natural language into a Boolean expression such as *'theft AND (repeated OR sustained) AND employee'*, by guessing the exact terms used in the database in the documents of interest (Hafner 1980, Ter et al. 1996). Such guesswork about the terms likely to be used in relevant documents, through the resulting errors, can prevent the discovery of the most relevant documents in the database, for example if those documents used slightly different terms that those that were selected by the user.

As efforts to remedy such issues using linguistics and / or ontology-based approaches that explicitly model domain knowledge and language are rather expensive and difficult to keep up-to-date (Silveira et al. 2004), the research community is actively searching for new approaches to those problems, in particular the semantic similarity aspect of document relevance, as it enables improved document ranking.

Recently, data-driven neural approaches have emerged as promising components of a new generation of increasingly automatic systems that may require less human interference, curation or updating effort (Tran et al. 2020, Zhong et al. 2020). While this trend is in its early stages, its maturation could help to deal with some of the above-mentioned limitations. However, it was observed by many authors (Alberts et al. 2020, Bhattacharya et al. 2020, Chalkidis et al., 2020, Draijer 2019, Raghav et al. 2016, Shao et al. 2020, Van Opijnen & Santos 2017, Xiao et al. 2019, Wang et al. 2019, Zhong et al. 2020) that current neural systems for natural language understanding that perform very well in non-legal domains do not transfer easily to tasks in the legal domain, for a variety of reasons that make this domain especially challenging (see Table 1).

Challenge	**Description**
Length	Documents are often long and complex, with a large number of legal concepts that need to be modeled
Relevance	Domain-specific notions of relevance go beyond general concepts of document similarity
Language	A language that differs from regular language in terms of syntax, semantics, vocabulary and morphology
Accessibility	Accessibility of legal datasets is rather restricted, hindering research
Labels	Expert judgements of relevance or similarity is difficult and expensive to obtain
Models	Most current semantic similarity models (above the word level) struggle with out-of-domain data

Table 1: Domain adaptation challenges

Such research efforts can then be further complicated by the need to adjust such systems to legal documents written in non-English languages, as much of the research in the public domain has been performed on English language texts.

By performing the legal Information Retrieval task at the level of paragraphs rather than whole case documents, the modeling of semantic similarity is then focused on a more limited number of important legal concepts contained in such a paragraph, or query. For a user it can be helpful to first find the most semantically similar paragraphs with a few relevant legal concepts in a lengthy and complex case document, before reading the whole document. This way of defining the task also has another practical advantage, in the sense that many state-of-the-art (neural) models would struggle to encode the semantics of a very long text with hundreds of sentences in a useful way, while text at the level of one or a few sentences will be a more realistic input text to such models (Alberts et al. 2020, Shao et al. 2020). In the best case, such semantic similarity models at paragraph level should be invariant to differences in the input that do not matter for approximating relevance in this task, while they will be more selective to differences that do matter (Neculoiu et al. 2016).

The ability for a user to express a query in natural language rather than more artificial queries such as Boolean expressions could carry many advantages, if implemented successfully. By omitting the step

of asking the user to translate their natural language query into such a query language, human errors linked with query formulation can be avoided. The potential ability of neural systems to enable a flexible data-driven modeling of semantic similarity, beyond the ranking of exact lexical matches between a candidate hit and a query, is highly relevant for enabling a wider variety of searches using natural language for expert and non-expert users with different degrees of domain knowledge. For the non-expert users, this aspect is likely to be even more important, due to their lack of familiarity with legal language and topics. In other words, progress in this area could help to further increase the accessibility of the legal system itself.

We have therefore identified the following three fundamental challenges for enabling German natural language queries in a collection of Swiss federal court decisions, namely a) how to best encode paragraph text as vectors that enable a good ranking of search hits (semantic modeling), so that the most relevant paragraphs would be scored highly, b) how to assess the semantic similarity between a query and a candidate hit in the absence of expensive expert judgements of relevance, using an automated mechanism that approximates such human judgement, c) the generation of relevant datasets for enabling such work, in particular Swiss cases and statutes that will be relevant for different tasks, including the extraction of relationships between cases and statutes. We also investigate the possibility of combining neural systems with some well-established components of non-neural systems in a way that would enhance the ability of the combined system to focus on the most important parts of the input, while modeling semantic similarity.

2 Related Work

Understanding the central concept of relevance in Information Retrieval tasks can depend on the particular user, task and other contextual factors (van Opijnen & Santos 2017), hindering the emergence of a clear consensus on this topic. Also, the construction of datasets that capture relevance as determined by human experts is challenging and expensive, leading to a paucity of such datasets in the public domain (Shao et al. 2020). Therefore, algorithms that could help to estimate important aspects of relevance could therefore help to enable more research on the challenges outlined above. With semantic similarity as a key aspect of relevance, and the paucity of human expert annotations in the field, the question of the automated generation of labels that can approximate such human judgement poses itself.

On a conceptual level, the computation of similarity between objects can not only be based on their context, but also on the link structure of a graph that describes relationships between those objects (Lu et al. 2006). Such link-based similarity can often complement content-based similarity measures. The application of this principle can be found in an area that is related to our task, namely a field called *context-aware citation recommendation*, which builds on link-based similarities between scientific publications using citations between them (He et al. 2020, Jeong et al. 2019). In the legal domain, citation networks can be described as linking case documents with relevant statutes, thereby generating a directed graph with two types of nodes (cases and statutes). The number of common out-citations can then be defined as *bibliographic coupling* (Kessler 1963), indicating link-based similarity between two case documents (Bhattacharya et al. 2020, Kumar et al. 2013, Raghav et al. 2015, Raghav et al. 2016). Bhattachyarya et al. (2020) have noted that the notion of semantic similarity is domain-specific, and not completely defined, as they developed a framework for the comparison of content-based with link-based similarity methods. Again, bibliographic coupling is used by calculating the number of common out-citations as a way of estimating semantic similarity between case documents.

As semantic similarity modeling of text at paragraph level (beyond a ranking of hits that is solely based on exact lexical matches with terms in the query), is a key challenge in this domain, lessons learned from the comparison of systems that model semantic similarity at sentence level (in various domains) may be relevant. For example, SentEval is an effort to evaluate sentence-to-vector encoders on a variety of tasks, including sentence similarity (Conneau et al. 2018). Bhattacharya et al. (2020) describe a dataset of 47 pairs of Indian Supreme Court case documents where similarity

between each pair of documents is annotated on a scale from 0-10 by law experts, which the authors use to compare methods for modeling legal document similarity. However, it is important to keep in mind the many known domain-specific challenges (see Table 1), in addition to language-specific and legal system-specific challenges. In other words, we cannot assume that systems that did well in those benchmarking efforts will also do well in our task, even if such systems have shown great potential across languages, in other domains.

The most widely used classic non-neural model is often referred to as a *bag-of-words* model, with every word in the modeled vocabulary represented as a separate dimension in a vector space. Note that while such vector space models can enable a user to use natural language as a query, it will be unable to model an important aspect of semantic similarity, namely contributions based on word order (Mitchell & Lapata 2010). As semantically similar words are represented as distinct dimensions in that vector space, such similarities are not captured in bag-of-word models. These non-neural systems can provide useful baseline systems for comparisons with (neural) systems that aim to model these aspects of semantic similarity.

Neural systems for Natural Language Understanding published in recent years have demonstrated promising abilities in this regard, in terms of their ability to model both word order aspects as well as other aspects of semantic similarities in natural language text (Zhang et al. 2020, Bhattacharya et al. 2019).

3 Materials and Methods

3.1 Datasets

German language case documents published after the year 2000 that contain citations of Swiss law articles (statutes) were obtained from the website of the Swiss Federal Court ('Bundesgericht') [1]. Using the Python library *urllib* we developed a dedicated Web crawler to download the HTML of 2168 case documents for paragraph and citation extraction. Paragraph extraction was then performed on those HTML files, filtering for case paragraphs that contain citations, resulting in a set of 7562 case paragraphs for further processing. Swiss federal statutes cited by those case paragraphs were obtained from the official Swiss government website ('Systematische Rechtssammlung') [2], also in HTML format. Again, paragraphs were extracted, creating a hierarchy of documents containing many articles, and those articles containing one or more paragraphs ('Absatz'). In total, 38994 statute paragraphs were extracted from 109 statute documents.

3.2 Triples

Here, triples are directed relations that link a single case paragraph (the source) to specific statute articles (the target). To qualify, a citation needs to match one of two citation styles that unambiguously identify at least a specific article of that law: style 1 (e.g. 'Art. 1 Abs. 3 StGB') or style 2 (e.g. 'Art. 3 StGB'). Note that style 1 is more specific, as it points to a single paragraph ('Abs.' = 'Absatz') in that article, which will then consist of a single sentence or multiple sentences. Citations expressed in the case paragraphs were identified using regular expressions, in an automated way. Each of the resulting 14299 triples then contained a citation in the form of either style 1 (10998 triples) or style 2 (3301 triples), the full text of the statute paragraph the citation refers to, the case paragraph in which the citation was found, and the relevant case document ID ('caseName').

3.3 Baseline models

A random baseline was implemented as a minimal performance reference, to estimate the performance achieved by a random ranking of paragraphs in the search collection. This system omits the actual similarity model by generating random numbers for use in ranking the search results. To estimate the distribution of results from this random baseline, ten independent runs were performed.

Idf scores (Manning et al. 2009), that represent the number of documents in which a particular token occurs in the search collection, were calculated for all tokens in the vocabulary. Here, a document corresponds to a concatenation of all texts from all case paragraphs in the triples collection that originated from the same HTML-formatted case document.

[1] https://www.bger.ch

[2] https://www.fedlex.admin.ch

Using the above idf values, a non-neural baseline (NNB) that performs exact lexical matching at paragraph level was implemented. As in other bag-of-words models, the vectorization of an individual paragraph creates a separate dimension for each token in the vocabulary, with the final tf-idf score calculated by multiplying the frequency of that token in the paragraph with its idf value (Manning et al. 2009). A boosted version of that non-neural baseline (Boosted NNB) uses an additional token filter that reduces the final vocabulary, by accepting only tokens that start with a sequence of three alphabetical (non-numeric) characters and have a minimal length of 5 characters. This filter was motivated by the observation that many semantically non-informative tokens started with various non-alphabetical characters. The alternative of using each paragraph in the search collection as a document for tf-idf calculation using the standard Python module '*sklearn'* was also tested. In that case, idf values were not based on all case paragraphs in the triples collection, but on a smaller search collection of 69 case paragraphs used in the model evaluation.

3.4 Neural models

We tested different neural approaches for the semantic modeling of paragraphs. If not mentioned otherwise, default parameters were applied.

Two different implementations of the Universal Sentence Encoder (USE, Cer et al. 2018) were tested, a) the model 'xx use md' from the 'spacy' Python library, and b) the huggingface (*huFace*) implementation 'distilUSE'. Note that the huFace variant is a knowledge-distilled, smaller version of the original 15-language USE model.

The doc2vec model (Le & Mikolov 2014) was applied after training the model on our case paragraphs using a vector size of 100, a window size of 5 tokens, 12 epochs for training, and a maximal vocabulary size of 500, after an effort to optimize its hyper-parameters for maximal performance in the task outlined below.

Pre-trained BERT-based transformers that were tested include DistilBERT (German model) and GermanBERT (trained on German text including 20% legal documents), including standard *WordPiece* word segmentation on the input. As a vector that aggregates information from all input tokens, the hidden state of the 'CLS' token of the final BERT layer was used, without further fine-tuning of the pre-trained model. Such transformers offer the use of an *attention mask* for padding, to

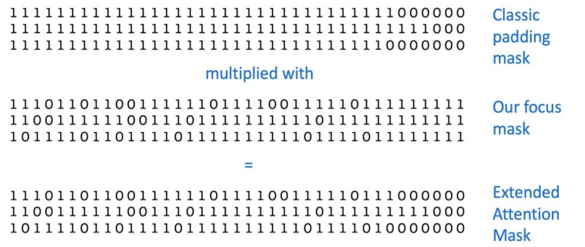


Fig. 1: Generation of the Extended Attention Mask, by multiplying a classic BERT attention mask for padding batch inputs with a focus mask we developed.

mask empty token positions after the last token in the input, when a batch of input sequences has varying length of sequences (Fig. 1). This attention mask represents a vector of binary 'hard gates' that cancel positions that should not get any attention calculated on them.

To combine the idf score aspect of the non-neural baseline systems with the neural models, tokens with low idf values were also assigned a zero value in that attention mask, resulting in what we call an *Extended Attention Mask.* This mask can suppress less informative tokens (with low idf values) in the input, changing the encoding of the paragraph as a result. At a more granular level, this Extended Attention Mask is generated as follows: before tokenization by *WordPiece*, tokens are screened against the idf threshold for masking. Tokens that do not pass are replaced by an underscore character, which in a second step, after using WordPiece for tokenization and subword segmentation, is replaced by a zero value instead of the original value of 1, in the focus mask. This means that transformer self-attention cannot be calculated on that token, even if the surrounding tokens receive self-attention. With the use of the underscore, subword segmentation is also avoided for the masked token.

3.5 Automated labels and model evaluation

An evaluation approach was designed using a set of queries that represent information needs and a set of case paragraphs as a search collection to match with those queries. To approximate the judgement of a human expert on the relevance of a

search hit based on semantic similarity, an automatically generated, link-based similarity definition was developed. This link-based similarity definition uses *case pairs* that are deemed to be sufficiently similar as proxies for true hits, when calculating precision values for each

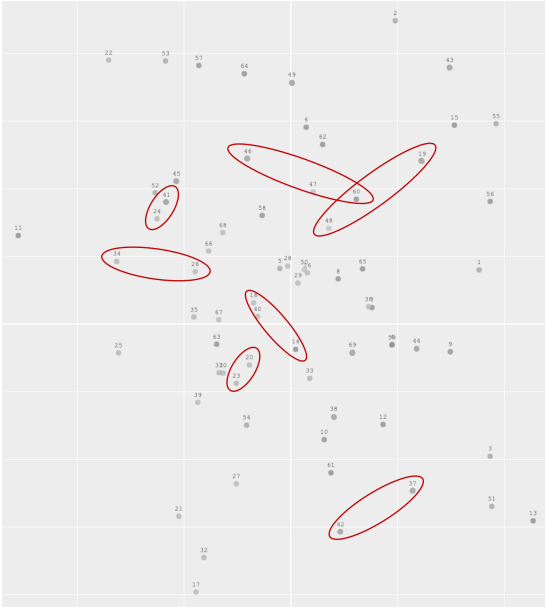


Fig. 2: A projection of the 69 query vectors used for model evaluation, using t-SNE to reduce 100-dimensional doc2vec-generated vectors to a position on this 2D map. Each vector is identified by an integer that represents its position in this list of query texts. Some case pairs are illustrated using red ellipses.

query. Such case pairs are defined as a pair of case paragraphs that share at least two different outgoing citations to the same statute articles in the triple collection. With the employed citations filter used in the generation of those triples, a sufficient level of specificity in terms of the statute target is ensured, thereby selecting a limited number of legal concepts as the semantic scope. To qualify, these two citations need to point to two different laws, or to two different law articles in the same law. A single shared citations was found insufficient in our exploratory work, to guarantee that most case pairs would indeed share a similar meaning that enables an approximation of relevance in this task, for non-experts. With those at least two shared citations as outlined above, we observed that the desired level of semantic similarity between case paragraphs was well approximated. Note that this definition of semantic modeling includes semantics encoded by word order, which cannot be modeled by (the baseline) bag-of-words models.

As a diversity of information needs is represented by a set of 69 queries (Fig. 2) with a large number of case pairs we selected for model evaluation, the

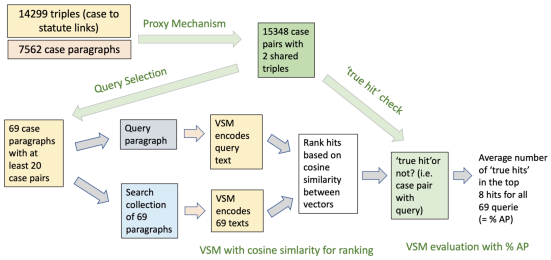


Fig. 3: Schematic overview of our method for model evaluation. From 7562 case paragraphs and 14299 triples, 15348 case pairs were identified (automated label generation). Those case pairs are then used to a) estimate 'true hits' for each query, as a basis for calculating an average precision over all 69 queries, b) as a filter for defining queries based on the triples.

average of the precision calculated for each query is then informative of a model's performance across this set of diverse queries. For calculating precision, the number of case pairs for a query is then used as an approximation of the number of true hits. As case paragraphs that are semantically similar may not always fulfil the selection criteria we defined for generating case pairs, the obtained average precision values may constitute an underestimation of the number of true hits.

When comparing different models, each of the 69 queries selected from the 7562 case paragraphs was then encoded as a single vector by the tested model and compared with the vectors of the other 68 queries using cosine similarity. In other words, for each query, the other 68 queries were used as the search collection, representing paragraph texts in a database of legal documents.

4 Results and Discussion

For ten independent runs of the random baseline, the mean average precision (AP) was 13.15%, with a standard deviation of 1.06. The non-neural baselines performed considerably better, with a performance of 24.09% achieved by the simple sklearn baseline, and 41.30% by the basic non-neural baseline (NNB v1). The performance of the

basic NNB can be further enhanced by applying a token filter using idf values calculated at case document level, up to a maximal level of 42.75% at an idf threshold of 2.12 (NNB v2). The curve in Fig. 4 shows a fine-grained analysis of the optimal idf threshold values, while Table 2 offers an overview of results obtained for different neural and non-neural models tested. The optimal threshold value of 2.12 for the idf score fits a human inspection of the relative importance of different words in case pairs (Fig. 5).

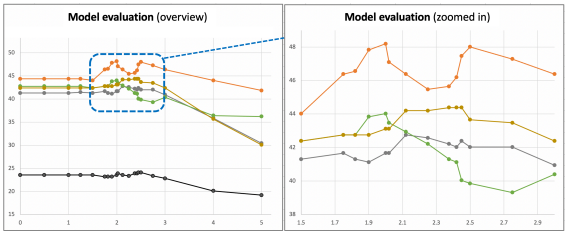


Fig. 4: Determination of the optimal idf threshold for non-neural and neural models. The performance at the optimal idf value corresponds to the v2 variant of the model shown in Table 2. idf threshold values are shown on the X axis, while the achieved %AP is shown on the Y axis. In dark grey, the *sklearn* model (left panel only). Non-neural baseline in light grey, and its boosted version (the best-performing non-neural model) in gold. DistilBERT in green, and GermanBERT (the best-performing neural model) in orange. The right panel shows a zoomed-in view, for the models that achieved a %AP of at least 38.

The ability of the neural models to outperform the best non-neural baseline (Boosted NNB) was then tested. The USE implementation *huFace* reached a performance of only 39.13%, so did not show a clear benefit over the Boosted NNB baseline. The *spacy* implementation, however, achieved 42.41 %, a gain of more than 3 percentage points. This difference between the two USE implementations may be explained by the fact that the *huFace* variant is not the full USE model, but a smaller, distilled version.

A clearly improved performance over the Boosted NNB was observed with GermanBERT, the model that was trained on 20% legal documents (cases and statutes in German language). This BERT-based model achieved a performance of 44.38%. With the use of the Extended Attention Mask, a maximal performance of 48.19% was reached at an optimal idf threshold of 2.0. For all tested BERT-based transformers, the use of the Extended Attention Mask resulted in clear improvements, at idf threshold values around 2 (compare Fig. 5), while distilled, smaller models (of USE and DistilBERT) did not perform well. In comparison, doc2vec achieved a performance of 45.47%, above NNB, but nowhere near the performance we saw

Model	%AP (v1)	%AP (v2)	Opt. idf
sklearn	23.55	24.09	2.5
NNB	41.30	42.75	2.12
Boosted NNB	42.39	**44.38**	2.45
doc2vec	45.47	-	-
USE (huFace)	39.13	-	-
USE (spacy)	42.21	-	-
DistilBERT	42.75	44.02	2
GermanBERT	44.38	**48.19**	2

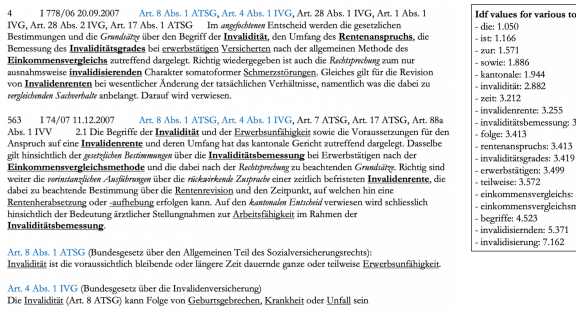
Table 2: Summary of model evaluation results. Some models are listed as two variants, one (v1) using a threshold of zero for idf scores (resulting in no filtering), and another (v2) that uses the optimal idf threshold. The performance of the best neural and non-neural models is shown in bold. A double line separates the non-neural (top) from the neural models (bottom). For BERT-like transformers that allowed the use of our Extended Attention Mask, the optimal idf value (see Fig. 4) is shown in the column 'Opt. idf' for the v2 variant of the model.

with the Extended Attention Mask-boosted GermanBERT. While both DistilBERT and GermanBERT were trained solely using German data, only GermanBERT had legal documents in the training data. Note that this performance of GermanBERT was observed on a pre-trained model without any finetuning to the task.

5 Conclusions

In the described semantic similarity modeling task, we found that the best neural model that was trained not only in the relevant language, but also in-domain data, offered a clear added value over the best non-neural model. This top performance was obtained by extending the use of the transformer attention mask to quench signals from less informative tokens (using idf scores), so they would not be considered in the model. The generated datasets of paragraphs extracted from Swiss cases and statutes, as well as citations in cases that point to statutes, when published can facilitate further research in this domain and task, which currently lacks such datasets. With the link-based automated labeling method we developed,

the challenge of obtaining costly human expert annotations in this domain becomes less of an obstacle, aiding research into improved models (and their finetuning) that could help to reveal semantically similar paragraphs that may have been missed with the models we tested. Therefore, with this new framework, a new generation of


4 I 778/06 20.09.2007 Art. 8 Abs. 1 ATSG, Art. 4 Abs. 1 IVG, Art. 28 Abs. 1 IVG, Art. 1 Abs. 1 IVG, Art. 28 Abs. 2 IVG, Art. 17 Abs. 1 ATSG Im *angefochtenen* Entscheid werden die gesetzlichen Bestimmungen und die *Grundsätze* über den Begriff der **Invalidität**, den Umfang des **Rentenanspruchs**, die Bemessung des **Invaliditätsgrades** bei erwerbstätigen Versicherten nach der allgemeinen Methode des **Einkommensvergleichs** zutreffend dargelegt. Richtig wiedergegeben ist auch die *Rechtsprechung* zum nur ausnahmsweise **invalidisierenden** Charakter somatoformer Schmerzstörungen. Gleiches gilt für die Revision von **Invalidenrenten** bei wesentlicher Änderung der tatsächlichen Verhältnisse, namentlich was die dabei zu *vergleichenden Sachverhalte* anbelangt. Darauf wird verwiesen.

563 I 74/07 11.12.2007 Art. 8 Abs. 1 ATSG, Art. 4 Abs. 1 IVG, Art. 7 ATSG, Art. 17 ATSG, Art. 88a Abs. 1 IVV 2.1 Die Begriffe der **Invalidität** und der Erwerbsunfähigkeit sowie die Voraussetzungen für den Anspruch auf eine **Invalidenrente** und deren Umfang hat das kantonale Gericht zutreffend dargelegt. Dasselbe gilt hinsichtlich der *gesetzlichen Bestimmungen* über die **Invaliditätsbemessung** bei Erwerbstätigen nach der **Einkommensvergleichsmethode** und die dabei nach der *Rechtsprechung* zu beachtenden *Grundsätze*. Richtig sind weiter die *vorinstanzlichen Ausführungen* über die *rückwirkende Zusprache* einer zeitlich befristeten **Invalidenrente**, die dabei zu beachtende Bestimmung über die Rentenrevision und den Zeitpunkt, auf welchen hin eine Rentenherabsetzung oder -aufhebung erfolgen kann. Auf den *kantonalen Entscheid* verwiesen wird schliesslich hinsichtlich der Bedeutung ärztlicher Stellungnahmen zur Arbeitsfähigkeit im Rahmen der **Invaliditätsbemessung**.

Art. 8 Abs. 1 ATSG (Bundesgesetz über den Allgemeinen Teil des Sozialversicherungsrechts): Invalidität ist die voraussichtlich bleibende oder längere Zeit dauernde ganze oder teilweise Erwerbsunfähigkeit.

Art. 4 Abs. 1 IVG (Bundesgesetz über die Invalidenversicherung) Die Invalidität (Art. 8 ATSG) kann Folge von Geburtsgebrechen, Krankheit oder Unfall sein

Idf values for various to
- die: 1.050
- ist: 1.166
- zur: 1.571
- sowie: 1.886
- kantonale: 1.944
- invalidität: 2.882
- zeit: 3.212
- invalidenrente: 3.255
- invaliditätsbemessung: 3
- folge: 3.413
- rentenanspruchs: 3.413
- invaliditätsgrades: 3.419
- erwerbstätigen: 3.499
- teilweise: 3.572
- einkommensvergleichs:
- einkommensvergleichsn
- begriffe: 4.523
- invalidisierenden: 5.371
- invalidisierung: 7.162

Fig. 5: Example for a case pair (case paragraphs No. 4 and No. 563, with their caseNames and full list of citations) with two shared citations (in blue) pointing to different statutes (ATSG and IVG). Note that semantic similarity of both case paragraphs, as defined using this link-based similarity method using shared citations, goes beyond exact lexical matching for many of the key terms, except the term denoting the main topic, 'Invalidität'. For comparison, the idf values of terms denoting key legal concepts in this case pair that contribute substantially to semantic similarity are shown below. Additional terms and expressions that are specific to legal language and not likely to occur in other domains are shown in italics. Also note how the statute texts are very concise and broadly understandable, which the case paragraph language is more about legal reasoning.

improved models can now be developed, for the task of searching Swiss court decisions using a wide range of natural language queries that represent information needs from expert and even non-expert users. The ability to model semantic similarity beyond exact lexical matching could then help a larger audience of non-expert users who may struggle with query formulation, to better access, understand and navigate the legal system.

Supplementary Material

The created datasets and Python code will be published on GitHub at https://github.com/lilytang2017

Acknowledgments

Martin Volk for his guidance on natural language processing and artificial intelligence in the legal domain.

References

Houda Alberts, Akin Ipek, Roderick Lucas, and Phillip Wozny. 2020. COLIEE 2020: Legal information retrieval & entailment with legal embeddings and boosting. In: Okazaki N., Yada K., Satoh K., Mineshima K. (eds) New Frontiers in Artificial Intelligence. JSAI-isAI 2020. Lecture Notes in Computer Science, vol 12758. Springer, Cham.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumber. 2019. FIRE 2019 AILA Track: artificial intelligence for legal assistance. Proc 11th Forum for Information Retrieval Evaluation.

Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. Methods for computing legal document similarity: a comparative study. arXiv:2004.12307v1.

Paul Boniol, George Panagopoulos, Christos Xypolopoulos, Rajaa El Hamdani, David Restrepo Amariles, Michalis Vazirgiannis. 2020. Performance in the courtroom: automated processing and visualization of appeal court decisions in France. arXiv:2006.06251v2.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, et al. 2018. Universal Sentence Encoder. arXiv:1803.11175

Ilias Chalkidis, Manos Fergadiotis, Prodomos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. EMNLP 2020 Findings, ACL: 2898-2904.

Yen-Liang Chen, Yi-Hung Liu, and Wu-Liang Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. Journal of ASIS&T 64(2):280–290.

Alexis Conneau, and Douwe Kiela. 2018. SentEval: an evaluation toolkit for universal sentence representations. arXiv:1803.05449

Wilco Draijer. 2019. The creation and integration of a legal information retrieval system without manual query construction. Master thesis, Leiden University. Supervisor: Suzan Verbene.

Carole D. Hafner. 1980. Representation of knowledge in a legal information retrieval system. SIGIR '80: Proc 3rd Annual ACM Conf Res Dev Inf Retrieval. P 139-153

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. Proc 19th Int Conf WWW, pages 421-30. ACM.

Arthur Ter, Henderik Alex Proper, and Theo P. van der Weide. 1996. Query formulation as an information retrieval problem. The Computer Journal 39.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, Sungchul Choi. A context-aware citation recommendation model with BERT and Graph Convolutional Networks. arXiv:1903.06464v1.

M.M. Kessler. 1963. Bibliographic coupling between scientific papers. American Documentation, 14: 10-25.

Sushanta Kumar, Polepalli Krishna Reddy, V. Balakista Reddy, and Malti Suri. 2013. Finding similar legal judgements under common law system. In DNIS, pages 103-116. Springer.

Quoc V. Le, and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1188–1196

Wangzhong Lu, Jeannette Janssen, Evangelos Milios, and Nathalie Japkowicz. 2006. Node similarity in the citation graph. Knowl Inf Syst 11:105-29.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. An introduction to Information Retrieval. Online edition. Cambridge University Press, Cambridge, UK.

Jeff Mitchell, and Mirella Lapata. 2010. Composition in distributional models of semantics. Cognitive Science 34: 1388-1429.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with Siamese recurrent networks. Proc 1st workshop representation learning for NLP, pages 148-157.

K. Raghav, Pailla Balakrishna Reddy, V. Balakista Reddy, and Polepalli Krishna Reddy. 2015. Text and citations-based cluster analysis of legal judgments', in MIKE, pp. 449–459. Springer.

K. Raghav, Pailla Balakrishna Reddy, and V. Balakista Reddy. 2016) Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. AI4JCArtificial Intelligence for Justice, page 30.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. Proc 29th Int Joint Conf AI (IJCAI-20).

Da Silveira, and Berthier A. Ribeiro-Neto. 2004. Concept-based ranking: a case study in the juridical domain. Inf. Process. Manage. 40 pages 791–805.

Vu Tran, Minh Le Nguyen, and Ken Satoh. 2020. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. ICAIL '19: Proc 17th Int Conf AI Law. p275-82.

Marc van Opijnen, and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. AI & Law 25:65-87.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: a multi-task benchmark and analysis platform for natural language understanding. ICLR 2019.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu et al. 2019. CAIL2019-SCM: a dataset of similar case matching in legal domain. arXiv:1911.08962v3

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. arXiv:2004.12158v5