Extracting facts from case rulings through paragraph segmentation of judicial decisions*

Andrés Lou[1], Olivier Salaün[2], Hannes Westermann[3], and Leila Kosseim[1]

[1] CLaC Lab, Dept of Computer Science and Software Engineering, Concordia University, Montréal QC H3G 1M8
{andres.lou, leila.kosseim@}.concordia.ca
[2] RALI, Université de Montréal, Montréal QC H3T 1J4
olivier.salaun@umontreal.ca
[3] CyberJustice Lab, Université de Montréal, Montréal QC H3T 1J4
hannes.westermann@umontreal.ca

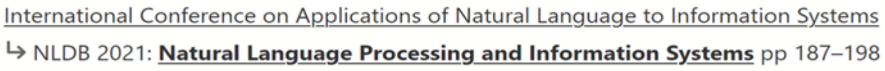
Abstract. In order to justify rulings, legal documents need to present facts as well as an analysis built thereon. In this paper, we present two methods to automatically extract case-relevant facts from French-language legal documents pertaining to tenant-landlord disputes. Our models consist of an ensemble that classifies a given sentence as either Fact or non-Fact, regardless of its context, and a recurrent architecture that contextually determines the class of each sentence in a given document. Both models are combined with a heuristic-based segmentation system that identifies the optimal point in the legal text where the presentation of facts ends and the analysis begins. When tested on a dataset of rulings from the *Régie du Logement* of the city of ANONYMOUS, the recurrent architecture achieves a better performance than the sentence ensemble classifier. The fact segmentation task produces a splitting index which can be weighted in order to favour shorter segments with few instances of non-facts or longer segments that favour the recall of facts. Our best configuration successfully segments 40% of the dataset within a single sentence of offset with respect to the gold standard. An analysis of the results leads us to believe that the commonly accepted assumption that, in legal documents, facts should precede the analysis is often not followed.

Keywords: legal document · text classification · text segmentation.

1 Introduction

Understanding the rationale behind a particular ruling made by a judge is a complex task that requires formal legal training in the relevant case law. Nevertheless, the rulings are still made using traditional methods of human discourse and reasoning, including default logic (Walker, Lopez, et al. 2015), deontic logic (Dragoni et al. 2015), and rhetoric (Wald 1995). In particular, knowing all the

* Supported by the CLaC Lab and the CyberJustice Lab at the University of Montréal.



International Conference on Applications of Natural Language to Information Systems
↳ NLDB 2021: **Natural Language Processing and Information Systems** pp 187–198

relevant facts surrounding a case is of the utmost importance to understand the outcome of a ruling, as they are necessary to arrive at the best-possible decision, since facts are what give way to what is usually called the *Best Evidence* (Stein 2005; Nance 1987).

Within a ruling, however, determining what constitutes a fact, as opposed to other types of content that motivate and illustrate a judge's decision, is also a matter that requires formal training. Figure 1 shows a sample from our dataset. As the figure shows, a variety of linguistic factors, such as specialised terminology, textual structure, linguistic register, as well as domain knowledge, make the ruling stray from more general-domain texts. As such, fact extraction from legal texts is time consuming, expensive, and requires legal expertise. Additionally, as Westermann et al. (2019) have shown, even amongst trained experts, inter-annotator agreement tends to be low. For example, in the task of labelling a corpus of rulings with a pre-established set of labels on the ruling's subject matter, Westermann et al. (2019) reported low inter-annotator agreement and suggested that its cause might be the general vagueness and lack of explicit reasoning in the texts.

This paper proposes an automatic method to identify and segment facts in texts of rulings. The extraction of facts from a ruling is performed by classifying each sentence in the text as either belonging to the facts or not; this is followed by the identification of the boundary between the segment that holds the facts and the segment that holds everything else. To this end, we use two different approaches based on Deep Learning (DL) to perform the sentence classification task: a **sentence ensemble classifier**, which individually takes each sentence in the corpus and classifies it either as fact or non-fact, regardless of its context, and a **recurrent architecture**, which encodes each document in the corpus as a binary string, where each sentence is classified as fact or non-fact as a function of its context.[4].

2 Related work

The use of Natural Language Processing (NLP) to mine judicial corpora is not new; however, very little work has used neural methods, as most of the cited literature uses rules or hand-crafted features to perform their stated tasks. Maat, Winkels, and Van Engers (2006) developed a parser that automatically extracts reference structures in and between legal sources. A few years later, Maat and Winkels (2009) developed a model based on syntactic parsing that automatically classifies norms in legislation by recognising typical sentence structures. Dell'Orletta et al. (2012) proposed a shared task on dependency parsing of legal texts and domain adaptation where participants compared different parsing strategies utilising the DeSR parser (Attardi 2006) and a weighted directed graph-based model (Sagae and Tsujii 2010). Grabmair et al. (2015) demonstrated the feasibility of extracting argument-related semantic information and

[4] The source code is publicly available at `https://gitlab.com/Feasinde/fact-extraction-from-legal-documents`

Original sentence	**English translation**	**Tag**
Comme le mandat fourni à l'audience par Madame <NAME> émane de <ORG>. qui n'est pas le véritable locateur, ce mandat n'est pas conforme à l'article 72 de la Loi sur la Régie du logement.	*As the mandate provided at the hearing by Ms. <NAME> emanates from <ORG>, who is not the real landlord, the mandate does not comply with section 72 of the Act respecting the Régie du logement.*	Fact
Rien ne prouve par ailleurs que <NAME> est employée de <ORG> et <ORG> puisque ces compagnies ne lui ont pas donné de mandat.	*There is also no evidence that <NAME> is an employee of <ORG> and <ORG> since these companies did not give him a mandate.*	Fact
Ceci étant dit, revenons à l'argumentation de Monsieur <NAME> et de Madame <NAME> voulant que toutes ces compagnies soient liées entre elles et puissent représenter l'autre sans plus de formalité.	*That said, let us return to the argument of <NAME> and <NAME> that all of these companies are linked and can represent the other without further formality.*	non-Fact
Cet argument ne tient pas; en effet, même si les compagnies sont dirigées par les mêmes personnes et que l'une soit l'actionnaire majoritaire de l'autre, il n'en demeure pas moins qu'il s'agit d'entités juridiques distinctes.	*This argument does not hold; even if the companies are managed by the same people and one is the majority shareholder of the other, the fact remains that they are separate legal entities.*	non-Fact

Fig. 1: Example of sentence classification as either stating a case Fact or non-Fact. (English translations provided by the authors. Proper names redacted.)

used it to improve document retrieval systems. Walker, Han, et al. (2017) introduced an annotated dataset based on propositional connectives and sentence roles in veterans' claims, wherein each document is annotated using a typology of propositional connectives and the frequency of the sentence types that led to adjudicatory decisions. Jaromír Savelka and Ashley (2017) used Conditional Random Fields (CRF) to extract sentential and non-sentential content from decisions of United States courts, and also extract functional and issue-specific segments (Jaromír Savelka and Ashley 2018), achieving near-human performance. Finally, Dragoni et al. (2015) performed rule-extraction from legal documents using a combination of a syntax-based system using linguistic features provided by WordNet (Miller 1995), and a logic-based system that extracts dependencies between the chunks of their dataset. Their work is closely related to our task; however, whereas Dragoni et al. (2015) base their model on syntactic and logical rulesets, we base our model on Recurrent and Convolutional Neural models, introduce our own segmentation heuristic, and use independent and contextual word embeddings (Mikolov et al. 2013; Martin et al. 2019).

Outside the frame of legal texts, semantic sentence classification has recently achieved new benchmarks thanks to the application of neural methods. The use Recurrent Neural Networks (RNN), and their derivative recurrent architectures (in particular encoder-decoder models (Cho et al. 2014)), has steadily produced results that outperform traditional models in many NLP tasks such as Machine Translation, Question Answering and Text Summarisation (eg the Attention mechanism Bahdanau, Cho, and Bengio 2014 and the Sequence-to-Sequence model (Sutskever, Vinyals, and Le 2014), two of the most important architectural developments in RNNs). Sentence classification using Deep Learning (DL) was first proposed by Kim (2014), who used Convolutional Neural Networks (CNN) and showed that their models outperformed classical models on many standard datasets. Their work was quickly followed by the application of RNN architectures for similar tasks, including those of Lai et al. (2015), Zhang, Zhao, and LeCun (2015), and Zhou et al. (2016). A major breakthrough was achieved with the Transformer (Vaswani et al. 2017), whose non-recurrent, Multi-Head Attention architecture allowed for richer language model representations that capture many more linguistic features than the original attention mechanism. Subsequently, Google's BERT (Devlin et al. 2018) has given way to a whole new family of language models able to produce state-of-the-art contextual embeddings for both individual tokens in a sentence and for the sentence itself. The following paragraphs will explain these architectures in detail.

3 Methodology

3.1 The dataset

The current work was developed as part of the JusticeBot project. JusticeBot aims to provide a gateway to law and jurisprudence for lay people (Westermann et al. 2019) through a chatbot where users can seek remedies to terminate their lease because of landlord-tenant disputes. The chatbot was developed using a corpus of 1 million written decisions in French, provided by the *Régie du Logement* of the city of Montréal. One of the numerous tasks related to the development and training of the chatbot is the extraction of case-related facts from a given document in the corpora.

The dataset used for fact extraction consists of a subset of the *Régie du Logement*'s corpus and includes 5,605 annotated rulings; these were selected from the original dataset because they include an explicit separation between two distinct sections: Facts and Analysis, as determined by the original author of the ruling (the judge), and delimited by appropriate headings. These two sections have been used as gold standard annotations to train and test our model. The **Facts** section should consist of all the case-relevant facts on which the ruling is supported, while the **non-Facts** should contain the analysis and discussion of the facts that ultimately lead to the resolution presented in the document. Table 1 shows statistics of the dataset.

Number of documents	5,605
Total number of sentences	454,210
Total number of sentences in Facts segments	239,077
Av number of sentences in Facts segments	36.25
Total number of sentences in non-Facts segments	215,133
Av number of sentences in non-Facts segments	32.62

Table 1: Statistics of the dataset

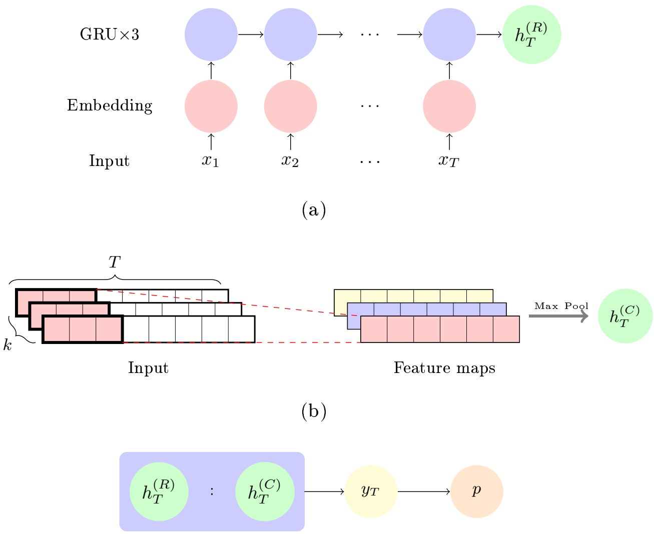
As Table 1 shows, Fact and non-Fact segments are similar both in terms of average number of words per sentence and average number of sentences per segment, which is why the process of detecting either cannot rely on simple word or sentence statistics.

3.2 Sentence Classification

Given a document, the first step in our approach is to represent its contents as a binary sequence, where sentences that include case-related facts are represented by continuous sub-sequences of 1's and the sentences containing everything else are represented by continuous sub-sequences of 0's. In order to produce this binary encoding of the document, we examine two methods: a **sentence ensemble classifier** method, and a **recurrent architecture** method.

The sentence ensemble classifier The sentence ensemble classifier method processes a given document as a collection of sentences whose classes are independent of one another. Each sentence in the corpus, regardless of the document in which it is found, is classified as either *fact* or *non-fact* using an ensemble model consisting of the combination of a Gated Recurrent Unit network (GRU) (Cho et al. 2014)) and a Convolutional Neural Network (CNN). The process is illustrated in Figure 2. In the recurrent part of the classifier (Figure 2a), a tokenised sentence is passed through an Word2Vec embedding layer (Mikolov et al. 2013) whose outputs are passed into a stack of GRU layers, producing a context vector $h_T^{(R)}$; In the CNN part of the classifier (Figure 2b), the input is a tensor of size $1 \times T \times k$, where T is the sequence length, and k is the size of the word vector. The output feature maps are passed through a 1-D Max Pool layer that produces an output vector $h_T^{(C)}$. The concatenation of $h_T^{(R)}$ and $h_T^{(C)}$ (Figure 2c) is passed through an affine layer with a softmax activation function to produce the probability of the sentence being Facts.

The recurrent architecture The recurrent architecture approach tries to determine whether a sentence is classified as *fact* or *non-fact* by using the sentences around it. We experimented with two different models to process a given document: a bidirectional GRU and an Encoder-Decoder model using an Attention



(a)

(b)

(c)

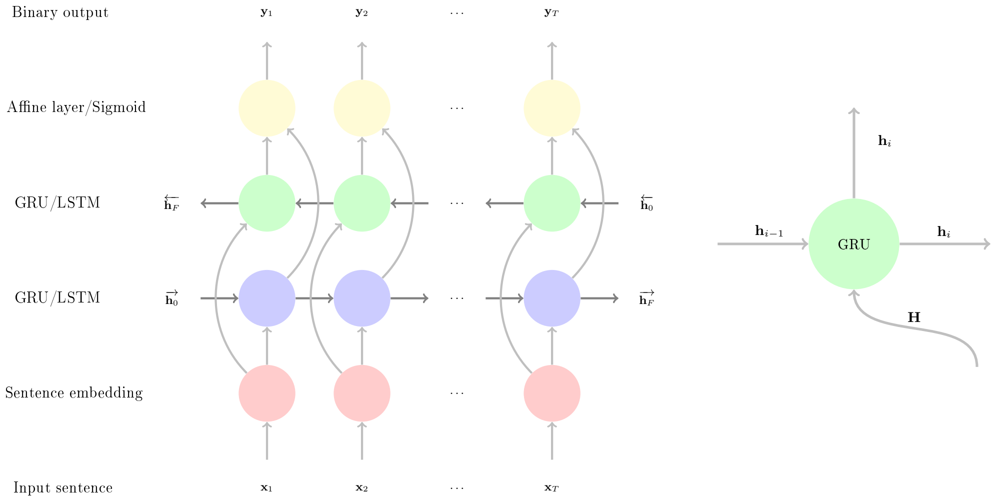
Fig. 2: Architecture of the sentence ensemble classifier.

mechanism (Bahdanau, Cho, and Bengio 2014). The process is illustrated in Figure 3. The input document is split into sentences, and each sentence is vectorised using the CamemBERT language model (Martin et al. 2019). The bidirectional GRU model (3a) produces an output ($\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_j}$) at each time step as a function of the previous and following steps, and the sentence input; after this, the output is passed through a affine layer with a binary output. The encoder of the attention mechanism (3b) is similar to this architecture, except the output of the bidirectional GRU is gathered as an input matrix $\mathbf{H}$ and passed to the decoder such that each time step output becomes a weighted function of every other token in the input sequence. The networks are trained so that the outputs correspond to the binary representation of the document.

3.3 Text segmentation

Once each sentence is classified as fact (1) or non-fact (0), the next step is to optimally divide the sequence into two substrings, each representing the *Facts* and *non-Facts* segments of the ruling. Figure 4 illustrates this. To perform the segmentation, we establish the following propositions:

- Let L represent the number of sentences document.
- Let L_f represent the number of sentences in its Facts segment.
- Let n_f represent the number of Facts sentences found in L_f (such that $n_p \leq L_f$).
- Define $p_f = \frac{n_f}{L_f}$ as the *purity* of L_f.



(a) (b)

Fig. 3: Architecture of the LSTM/GRU and Attention Encoder-Decoder method

$$\underbrace{\overbrace{1111011111111}^{L_f}00000100001}_{L}$$

Fig. 4: Visual representation of the segmentation of the binary string where 1 refers to a sentence classified as *fact* and 0 as *non-fact*. In this example, $L_f = 13$, $L = 24$ and $n_f = 12$

.

Maximising L_f is equivalent to maximising the recall of facts in the segmentation, and maximising p_f ensures the segmentation corresponds as closely as possible to the gold standard. Hence, our approach aims at maximising both L_f and p_f. This can be described as the following optimisation problem:

$$\max J(L_f) = \max\left(\alpha L_f + \beta p_f\right) \tag{1}$$

where J is a loss function of L_f, and $\alpha, \beta \in \mathbb{R}$ are arbitrary weights representing the importance of each term. We can rewrite and differentiate Equation 1 to find an expression that optimises L_f:

$$L_f = \frac{\beta}{\alpha} p_f \tag{2}$$

Equation 2 indicates that there is a linear relation between the purity (p_f) of a substring and its length (L_f). Therefore, for all possible substrings of length L_{f_i}

in the original string, we select the one that maximises p_f. Since any substring comprised exclusively of 1's will have a trivial purity $p_{f_i} = 1$, we select L_{f_i} that maximises p_f such that $p_f \neq 1$.

By considering the problem of finding the optimal segmentation point as extracting the substring with the highest purity from a binary representation of the document, our model can compute a splitting index, l, which can be empirically weighted by a factor, $\gamma \propto \frac{\beta}{\alpha}$ in order to favour either shorter or longer substrings. Shorter substrings will be purer, favour precision, and will contain few instances of non-Facts, while longer substrings will favour the recall of sentences containing facts. Hence, we can compute the weighted splitting index: $L_\gamma = \gamma l$

4 Results and discussion

4.1 Results of the sentence classification task

Recall from Section 3.1 that we evaluated the approach with a dataset of 5,605 rulings from the *Régie de Logement* of the city of Montréal. The dataset was randomly split into fractions of 90% and 10% for training and testing respectively. Using the standard classification metrics, we obtained the results shown in Table 2. As shown in Table 2, the ensemble model reached an F_1 of 77%, where as the GRU and the Attention models reached 99% and 90% respectively. Given the low performance of the ensemble, we experimented with a data augmentation technique. We used part-of-speech lexical substitution for data augmentation (PLSDA) (Xiang et al. 2020), generating new sentences by randomly replacing POS-annotated tokens in a given sentence with syntactically identical synonyms. We used the spaCy tokeniser and annotator (Honnibal and Montani 2017) and WordNet (Miller 1995). We observed no considerable improvement in our performance when doubling the number of training instances.

As Table 2, the recurrent architecture's improved performance suggests that contextually determining whether a sentence is *fact* or *non-fact* is a much better approach than assuming individual sentences are independently distributed from one another.

Model	**Accuracy**	**Precision**	**Recall**	$\mathbf{F_1}$
Ensemble	0.79	0.78	0.76	0.77
Ensemble (PLSDA)	0.79	0.78	0.76	0.77
GRU	0.98	0.98	0.99	0.99
Attn	0.92	0.91	0.92	0.90

Table 2: Intrinsic performance of the sentence classification task

4.2 Results of the text segmentation task

We evaluated the text segmentation on the test set of documents (660) using the headings separating Facts and non-Facts as gold standard and using both the augmented sentence ensemble classifier and the recurrent architecture methods. Given a value of γ, for each document, we compute its corresponding splitting index l_{pred} and split the text according to the weighted splitting index L_γ. We then compute the percentage of sentences by which the resulting text is off compared to the gold standard of number of sentences in the Facts section. Results are shown in Table 3.

		Offset										
		< -4	-4	-3	-2	-1	0	1	2	3	4	> 4
	0.6	602	25	16	8	3	3	2	0	1	0	0
	0.8	570	25	30	7	13	6	1	3	3	0	2
γ	1	529	19	26	21	15	**11**	13	8	3	5	10
	1.2	525	21	24	22	16	9	11	8	6	8	10
	1.4	520	17	26	21	16	11	13	6	9	7	14

(a) Sentence Ensemble Classifier

		Offset										
		< -4	-4	-3	-2	-1	0	1	2	3	4	> 4
	0.6	517	44	60	23	10	3	1	0	0	0	2
	0.8	319	74	75	89	77	16	1	3	2	1	3
γ	1	176	8	2	6	30	**408**	16	1	4	0	9
	1.2	137	2	7	7	30	445	17	0	4	1	10
	1.4	112	2	2	9	33	468	16	3	3	2	10

(b) Recurrent Architecture: GRU

		Offset										
		< -4	-4	-3	-2	-1	0	1	2	3	4	> 4
	0.6	518	42	57	21	11	5	2	0	0	1	3
	0.8	326	70	73	90	74	15	1	5	0	0	6
γ	1	179	8	4	5	46	**385**	15	4	4	0	10
	1.2	140	2	12	7	50	406	14	8	9	2	10
	1.4	112	2	6	7	66	413	17	10	11	1	15

(c) Recurrent Architecture: Attention

Table 3: Different values of γ and the number of sentence by which the predicted segmentation is off with respect to the gold standard. Bold indicates the number of documents obtained at the expected splitting index for $\gamma = 1$.

Table 3 shows the number of documents that fall within a distance (in sentences) from the expected index, given a weighting value of γ. For example, as shown in Table 3b, for the GRU recurrent architecture at $\gamma = 1$, 408 of the documents (61%) are segmented exactly where the gold standard indicates, while 454 documents (30+408+16, 68% of the test dataset) fall within a single sentence of difference with respect to the gold standard; nevertheless, 176 documents (27%) have their index underestimated and fall short of the target, having more than 4 sentences fewer than the gold standard. Increasing the value of γ favours the recall of sentences annotated as Facts, but the percentage of documents whose segmentation falls short by more than 4 sentences does not decrease as quickly as the percentage of overestimated segmentations; for $\gamma = 1.4$, the number of underestimated segmentations by a margin greater than 4 is still 112 (17%).

For the different values of γ, the distribution of offsets presents a large number of underestimated splitting indices, which suggests that the distribution of fact sentences and non-fact sentences does not actually follow our base assumption, namely, that facts should always give way to analyses. The gold standard expects us to find many more facts after the predicted splitting index, weighted or otherwise, which suggests that some cases either contain an imbalance of facts and analyses or contain facts and analysis interspersed with each other on a larger scale than expected.

5 Conclusion

This paper presented a method to automatically extract case-relevant facts in French-language legal documents pertaining to tenant-landlord disputes using text segmentation. We used two approaches based on classifying the sentences of a given document as either *facts* or *non-facts*: considering each sentence as independent from all others, and using the context in which the sentence is found to predict its class. Subsequently, we used a metric based on the purity of the facts substring to find an optimal splitting index and perform the segmentation

Experiments with French-language rulings of the *Régie du Logement* of the city of Montréal produced a significant number of underestimations (up to 27%); this seems to indicate that the standard assumption that the discourse structure should be such that all facts will precede the analysis is not always followed. Indeed, our text segmentation approach, based on the heuristic of maximising the density of facts on the purported facts segment of the ruling, has shown that the distribution of facts is not usually concentrated in the first segment of the text.

Our work has considered sentences as the unit of classification; a sentence that contains facts is considered fact-bearing even if it might also contain analysis. Future work might explore a more fine-grained intra-sentence analysis in order to find smaller fact-bearing units than sentences. Additionally, future work should also involve the classification and rearrangement of sentences, perhaps by means of standard automatic summarisation techniques (Nenkova and McKeown 2012), in order to produce coherent paragraphs that both maximise the purity of a substring and the recall of facts. Finally, rather than segmenting legal documents as a single fact-analysis block, it might be worth considering breaking them down into smaller fact-analysis constituents.

6 Acknowledgments

The authors would like to thank the anonymous reviewers for their comments on an earlier version of this paper.

References

Attardi, Giuseppe (June 2006). “Experiments with a multilanguage non-projective dependency parser”. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City, pp. 166–170.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (May 2014). “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473*.

Cho, Kyunghyun et al. (Sept. 2014). “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078*.

Dell’Orletta, Felice et al. (May 2012). “The SPLeT-2012 shared task on dependency parsing of legal texts”. In: *Proceedings of the 4th Workshop on Semantic Processing of Legal Texts*. Istambul, pp. 42–51.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Dragoni, Mauro et al. (2015). “Combining natural language processing approaches for rule extraction from legal documents”. In: *AI Approaches to the Complexity of Legal Systems*. Braga, Portugal: Springer, pp. 287–300.

Grabmair, Matthias et al. (June 2015). “Introducing LUIMA: An experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA-type system and tools”. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM. San Diego, CA, pp. 69–78.

Honnibal, Matthew and Ines Montani (2017). “spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing”. In: *To appear* 7.1.

Kim, Yoon (Sept. 2014). “Convolutional neural networks for sentence classification”. In: *arXiv preprint arXiv:1408.5882*.

Lai, Siwei et al. (Feb. 2015). “Recurrent convolutional neural networks for text classification”. In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2267–2273.

Maat, Emile de and Radboud Winkels (June 2009). “A next step towards automated modelling of sources of law”. In: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. ACM, pp. 31–39.

Maat, Emile de, Radboud Winkels, and Tom Van Engers (Dec. 2006). “Automated detection of reference structures in law”. In: *Frontiers in Artificial Intelligence and Applications*, p. 41.

Martin, Louis et al. (Nov. 2019). “CamemBERT: a Tasty French Language Model”. In: *arXiv e-prints*, arXiv:1911.03894, arXiv:1911.03894. arXiv: `1911.03894 [cs.CL]`.

Mikolov, Tomas et al. (2013). “Efficient Estimation of Word Representations in Vector Space”. In: arXiv: `1301.3781 [cs.CL]`.

Miller, George A (1995). “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11, pp. 39–41.

Nance, Dale A (1987). “The Best Evidence Principle”. In: *Iowa Law Review* 73, p. 227.

Nenkova, Ani and Kathleen McKeown (Jan. 2012). “A survey of text summarization techniques”. In: *Mining Text Data*. Springer, pp. 43–76.

Sagae, Kenji and Jun-Ichi Tsujii (2010). “Dependency parsing and domain adaptation with data-driven LR models and parser ensembles”. In: *Trends in Parsing Technology*. Springer, pp. 57–68.

Savelka, Jaromír and Kevin D Ashley (2017). “Using conditional random fields to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection”. In: *Workshop on Automated Semantic Analysis of Information in Legal Texts*, p. 10.

— (2018). “Segmenting US Court Decisions into Functional and Issue Specific Parts.” In: *JURIX: The 31st International Conference on Legal Knowledge and Information Systems*, pp. 111–120.

Stein, Alex (2005). “Foundations of evidence law”. In: Oxford University Press. Chap. Epistemological Corollary.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (Dec. 2014). “Sequence to sequence learning with neural networks”. In: *Advances in Neural Information Processing Systems*. Montréal, QC, pp. 3104–3112.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*. Long Beach, CA, pp. 5998–6008.

Wald, Patricia M (Oct. 1995). “The rhetoric of results and the results of rhetoric: Judicial writings”. In: *The University of Chicago Law Review* 62.4, pp. 1371–1419.

Walker, Vern R, Ji Hae Han, et al. (2017). “Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans’ claims dataset”. In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. London, UK, pp. 217–226.

Walker, Vern R, Bernadette C Lopez, et al. (2015). “Representing the Logic of Statutory Rules in the United States”. In: *Logic in the Theory and Practice of Lawmaking*. Springer, pp. 357–381.

Westermann, Hannes et al. (June 2019). “Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice”. In: *Proceedings of the 17th International Conference on Artificial Intelligence and Law*. ACM. Montréal, QC, pp. 133–142.

Xiang, Rong et al. (2020). “Lexical Data Augmentation for Text Classification in Deep Learning”. In: *Canadian Conference on Artificial Intelligence*. Springer, pp. 521–527.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (Dec. 2015). “Character-level convolutional networks for text classification”. In: *Advances in Neural Information Processing Systems*. Montréal, QC, pp. 649–657.

Zhou, Peng et al. (Apr. 2016). “Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling”. In: *arXiv preprint arXiv:1611.06639*.