MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset

Tobias Brugger*
University of Bern
Switzerland
tobias.brugger@students.unibe.ch

Matthias Stürmer
University of Bern
Switzerland
Bern University of Applied Sciences
Switzerland
matthias.stuermer@unibe.ch

Joel Niklaus*
University of Bern
Switzerland
Bern University of Applied Sciences
Switzerland
Stanford University
United States
joel.niklaus@unibe.ch

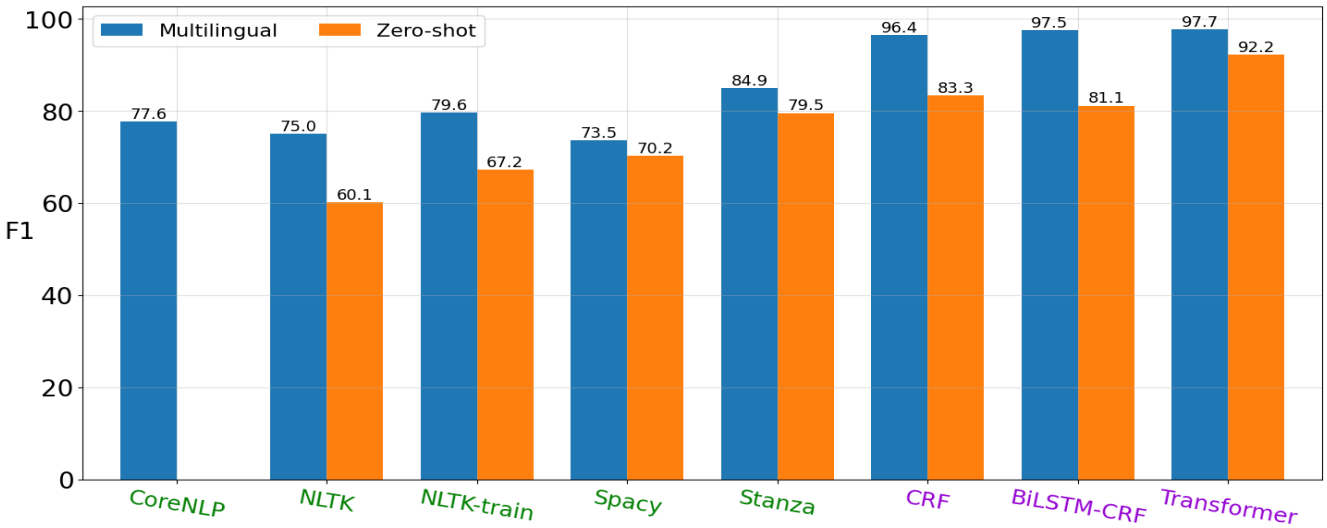

Figure 1: Mean performance comparison of baseline (green) and our multilingual models (violet) on multilingual legal text (French, Italian, Spanish, English, and German) and a zero-shot experiment on Portuguese data

ABSTRACT

Sentence Boundary Detection (SBD) is one of the foundational building blocks of Natural Language Processing (NLP), with incorrectly split sentences heavily influencing the output quality of downstream tasks. It is a challenging task for algorithms, especially in the legal domain, considering the complex and different sentence structures used. In this work, we curated a diverse multilingual legal dataset consisting of over 130'000 annotated sentences in 6 languages. Our experimental results indicate that the performance of existing SBD models is subpar on multilingual legal data. We trained and tested monolingual and multilingual models based on CRF, BiLSTM-CRF, and transformers, demonstrating state-of-the art performance. We also show that our multilingual models outperform all baselines in the zero-shot setting on a Portuguese test set. To encourage further research and development by the community, we have made our dataset, models, and code publicly available.

CCS CONCEPTS

• **Applied computing** → *Law*; Annotation; • **Computing methodologies** → **Natural language processing**; *Supervised learning*; **Machine learning**.

KEYWORDS

Sentence Boundary Detection, Natural Language Processing, Legal Document Analysis, Text Annotation, Multilingual

ACM Reference Format:
Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. In *Proceedings of 19th International Conference on Artificial Intelligence and Law (ICAIL '23)*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

ICAIL '23, June 19–23, 2023, Braga, Portugal

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent methodological advances, e.g., transformers [34], have lead to substantial progress in quality and performance of language models as well as growth in the general field of Natural Language Processing (NLP). This trend is also evident in legal NLP, with research papers increasing drastically in recent years [14].

Not as much attention and resources have been directed to the Sentence Boundary Detection (SBD) task, being viewed as solved by some, as high baseline performances can be achieved by utilizing simple lookup methods capturing frequent sentence-terminating characters such as periods, exclamations marks and question marks combined with hand-crafted rules [26]. This approach is feasible when applied to well-formed and curated text such as news articles. Noisier domain-specific data containing differently structured text combined with the ambiguity of many sentence-terminating characters [8, 15] – e.g., the period occurring in abbreviations, ellipses, initials etc. as a non-terminating character – often overwhelm the aforementioned methods and also more complicated off-the-shelf SBD systems. This has been illustrated in a number of specific SBD applications such as user-generated content [9, 26] as well as in the clinical [20] and financial domain [7, 19].

In legal documents, the aforementioned difficulties are increased with legal text consisting of smaller parts such as paragraphs, clauses etc., making it quite different from standard text. Furthermore, sentences are long and may contain complex structures such as citations, parentheses, and lists. These structures are often utilized to convey additional information to the reader (e.g., citations referencing another text) or formatting the text in a specific way (e.g., lists emphasizing ideas or increasing the readability of long paragraphs). However, these structures or special sentences do not follow a standard sentence structure, thus posing an additional challenge to SBD systems, illustrated in several works on English [27, 29] and German [10] legal documents.

1.1 Motivation

Having a reliable SBD system is crucial for accurate NLP analysis of text. Poor SBD can result in errors propagating into higher-level text processing tasks, which hinders overall performance. For instance, the curation of the multilingual EUROPARL corpus required proper SBD to align sentences in both languages for statistical machine translation. Koehn [16] noted the difficulty of SBD as it requires specialized tools for each language, which are not readily available for all languages. Inadequate SBD weakens the performance of sentence alignment algorithms and reduces the quality of the corpus. Therefore, a high-quality SBD system, especially one customized for the legal domain, can significantly improve performance.

Another example is Negation Scope Resolution (NSR), focusing on finding negation words (e.g., "not") in sentences and their impact on surrounding words' meaning. Negations are vital in text's semantic representation, reversing proposition values. This is particularly useful in the legal domain, enabling models extracting information from documents to better understand input text meaning, such as recognizing court decisions' outcomes based on exact wording. NSR models often require data split into sentences for labeling training data and application input, making a reliable SBD system crucial. Incorrect sentence predictions by the SBD system may significantly lower input data quality and model performance. Proper SBD is also crucial in other NLP tasks such as Text Summarization, Part-of-Speech-Tagging, and Named Entity Recognition, all relevant in the legal domain.

1.2 Main Research Questions

In this work, we pose and examine three main research questions:

RQ1: What is the performance of existing SBD systems on legal data in French, Spanish, Italian, English, and German?

RQ2: To what extent can we improve upon this performance by training mono- and multilingual models based on CRF, BiLSTM-CRF, and transformers?

RQ3: What is the performance of the multilingual models on unseen Portuguese legal text, i.e., a zero-shot experiment?

1.3 Contributions

The contributions of this paper are twofold:

(1) We curate and publicly release a large, diverse, high-quality, multilingual legal dataset (see Section 3) containing over 130'000 annotated sentence spans for further research in the community.

(2) Using this dataset, we showcase that existing SBD systems exhibit suboptimal performance on legal text in French, Italian, Spanish, English, and German. We train and evaluate state-of-the-art monolingual SBD models based on Conditional Random Fields (CRF), BiLSTM-CRF and transformers, achieving F1-scores up to 99.6%. We showcase the performance and feasibility of multilingual SBD models, i.e., trained on all languages, achieving F1-scores in the higher nineties, comparable or better than our monolingual models on each aforementioned language. In a zero-shot experiment, we demonstrate that it is possible to achieve good cross-lingual transfer by testing the multilingual models on unseen Portuguese legal text. We publicly release the datasets[1], all of our monolingual and multilingual models[2] (see Section 5) as well as our code[3] for further use in the community.

2 RELATED WORK

In this section, we discuss the literature at our disposal. First, we look at works showcasing the need for more research in regard to SBD. Second, we take a look at works tackling the problem of SBD in legal text in several languages. Lastly, we investigate SBD research in other domains and present multilingual datasets in the legal domain for thoroughness.

Read et al. [26] questioned the status quo of SBD being "solved", especially in more informal language and special domains, by reviewing the current state-of-the-art SBD systems on English news articles and user-generated content. The systems were able to reach F1-scores in the higher nineties for the former, however the performance on user-generated content weakened perceptibly with scores down to the lower nineties, showcasing the need for "a renewed research interest in this foundational first step in NLP." [26]

[1] https://huggingface.co/datasets/rcds/MultiLegalSBD
[2] https://huggingface.co/models?search=rcds/distilbert-sbd and https://github.com/tobiasbrugger/MultiLegalSBD/tree/master/models
[3] https://github.com/tobiasbrugger/MultiLegalSBD

2.1 SBD in the Legal Domain

Savelka et al. [29] continued this research in the English language by curating a legal dataset, consisting of adjudicatory decisions from the United States. When testing existing systems on the dataset, they report F1-scores between 75% and 78%. Training or adapting these systems to the dataset improved their F1 score to the mid-eighties, which is still lower than their respective performance in more standard domains [26], showcasing the subpar performance of state-of-the-art SBD in the English legal domain. To improve this issue, they trained a number of CRF models as well as a model based on hand-crafted rules, reporting F1-scores of 79% for the hand-crafted model and up to 96% for the CRFs. Additionally, they developed a publicly available, comprehensive set of annotation-guidelines for sentence boundaries in legal texts which we used as a foundation for our guidelines.

Sanchez [27] experimented on the same dataset reporting an F1-score of 74% using the Punkt Model [15]; adapting it to the dataset slightly improved performance. They also trained and evaluated CRF and Neural Network (NN) models, reporting F1-scores up to 98.5% and 98.4% respectively. Our multilingual models achieve F1-scores between 95.1% and 97% on the same dataset.

Similarly, Glaser et al. [10] curated a German legal dataset, split into laws and judgements; a similar distribution is used in our work. They established a baseline performance of existing SBD systems and compared it to CRF and NN models trained on the aforementioned dataset. Their findings outline F1-scores between 70% to 78% for off-the-shelf systems, supporting the view that the performance of existing SBD system is subpar on legal data. The CRFs and NNs models achieve F1-scores up to 98.5%. However, a significant decrease in performance was reported, when applying them to previously unseen German legal texts with scores down to 81.1%. Our multilingual models showcase F1-scores between 91.6% to 97.6% on the German dataset.

2.2 SBD in Other Domains

In the financial domain, Du et al. [7] experimented with Bidirectional Long Short-Term Memory (BiLSTM) models combined with a CRF layer as well as the transformer-based model BERT [6] and compared their performance, approaching SBD as a sequence labelling task to extract useful sentences from noisy financial texts. They demonstrate that BERT significantly outperforms BiLSTM-CRFs across all evaluation metrics, including F1-scores. In their work they also underline the fact that "SBD has received much less attention in the last few decades than some of the more popular subtasks and topics in NLP."

Schweter and Ahmed [31] compared the performance of Long Short-Term Memorys (LSTMs), BiLSTMs and Convolutional Neural Networks (CNNs) to OpenNLP [4] in an SBD task on the Europarl [16], SETimes [33] and Leipzig Corpora [11] containing around 10 different languages, showcasing the use of their models as robust, language-independent SBD systems.

2.3 Multilingual Datasets in the Legal Domain

Niklaus et al. [23] present LEXTREME, a novel multilingual benchmark dataset containing 11 datasets in 24 languages, designed to evaluate natural language processing models on legal tasks. The authors assess five prevalent multilingual language models, providing a benchmark for researchers to use as a basis for comparison. Savelka et al. [30] investigate the application of multilingual sentence embeddings in sequence labeling models to facilitate transfer across languages, jurisdictions, and other legal domains. They demonstrate encouraging outcomes in allowing the reuse of annotated data across various contexts, which leads to the development of more resilient and generalizable models. Additionally, they create a vast dataset of newly annotated legal texts using these models. Chalkidis et al. [3] introduce MultiEURLEX, a multilingual and multilabel legal document classification dataset containing 65000 EU Laws. Aumiller et al. [1] present a EurLexSum, a multilingual summarization dataset curated from Eur-Lex data. Niklaus et al. [21, 24] introduce Swiss-Judgment-Prediction, a multilingual judgment prediction dataset from the Federal Supreme Court of Switzerland.

3 DATASET

We annotated sentence spans for three diverse multilingual legal datasets in French, Italian, and Spanish, each containing approximately 20,000 sentences evenly split between judgments and laws. We chose a variety of legal areas to capture a broad selection. The laws included the Constitution, part of the Civil Code, and part of the Criminal Code, with the Constitution used only for evaluation. The judgments comprised court decisions from various legal areas and sources. We also annotated a smaller Portuguese dataset with approximately 1800 sentences, divided into the same subsets as the other datasets. This dataset was used for zero-shot experiments.

Additionally, we standardized and integrated two publicly available datasets, an English collection of legal texts [29], consisting of Adjudicatory Decision from the United States as well as a German dataset [10], comprising laws and judgments, into our dataset to further increase its diversity.

Figure 2 illustrates the sentence length distribution of our dataset, showing the relative frequency of sentence length in tokens for laws and judgments, with a bin size of 5. We used an aggressive tokenizer, resulting in a larger number of tokens per sentence than usual. For clarity, we did not include sentences longer than 101 tokens, which comprised only ~2% (2634) of the sentences. Only 26 sentences were longer than 512 tokens.

For each language, we used random sampling to split the dataset into three parts: train, test and validation. The test and validation splits each contain 20% of the dataset. Every model is trained on the train split, and we report their performance on the test split. Selected statistics and information about the dataset are in Table 1.

3.1 Annotation

The human annotator was tasked with correcting the sentence-spans predicted by an automatic SBD system[5] [29] based on CRF, which was trained on data annotated using annotation guidelines by Savelka et al. [29]. This helped improve the quality and consistency of our annotations. Furthermore, a practical rule set, heavily influenced by the aforementioned guidelines, was utilized to aid the annotator in the annotation process, reducing the complexity of the task and helped provide dependable and well-founded data. The

[4] https://opennlp.apache.org/

[5] https://github.com/jsavelka/luima_sbd

Table 1: Statistics on datasets per language and subset

Language	Subset	Sentences	Tokens	# of Documents	Source
French	Judgments	9971	342469	315	Niklaus et al. [21]
	Laws	11055	334453	3	Wipolex
Italian	Judgments	10129	340041	183 + 60	Niklaus et al. [21] + Multi-Legal-Pile (MLP)
	Laws	10849	301466	3	Jus.unitn.it
Spanish	Judgments	10656	356681	20 + 84	Wipolex + MLP
	Laws	11501	229240	3	Wipolex
Portuguese	Judgments	759	20590	6	Wipolex
	Laws	1010	25947	3	Wipolex
German	Judgments	21409	506009	131	Glaser et al. [10]
	Laws	20330	484816	13	Glaser et al. [10]
English	Judgments	25899	712433	80	Savelka et al. [29]
Total	Laws & Judgments	133568	3654145	906	

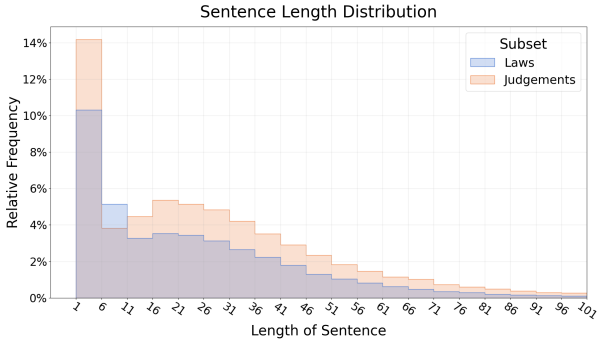

Figure 2: Sentence length distribution in tokens

rule set is outlined in Section 3.1.1, containing the most important sentence structures followed by an example.

The documents were annotated using Prodigy (https://prodi.gy/). Because Prodigy requires pre-tokenized text, a customized tokenizer was applied to the input text, further described in Section 3.2. The decision to annotate the full sentence-span, in lieu of just the first and last token in the sequence, was made to incentivize the annotator to read the text instead of skimming it for sentence-terminating characters. To make the annotation easier, laws were split into smaller chunks with one to three articles per chunk, while judgments were only split, if they surpassed ~15000 characters since Prodigy was unable to handle longer documents.

3.1.1 Legal Sentence Structures. In this section, we briefly describe the most important sentence structures in legal text, heavily influenced by Savelka et al. [29], followed by an example in French.

Standard Sentence have subject, object and verb in the correct order and the last token in the sequence is a sentence-terminating character.
- *Il s'est établi comme ingénieur indépendant.*

Linguistically Transformed Sentence are similar to a standard sentence, but slight transformations such as changes to the word order are applied.
- *Tout porte à croire, en réalité, qu'elle est condamnée au surendettement, puis à la faillite.*

Headlines determine the structure of the text and show relatedness between parts of the document and therefore convey important information about the overall structure of the text.
- *Considérant en fait et en droit*
- *PAR CES MOTIFS*
- *DÉCLARATION*

Data fields provide the name and data of a field. This is annotated as a sentence, as for example in English "Civil Chamber: Madrid" has a similar meaning to "The civil chambers are located in Madrid".
- *Numéro d'appel: 1231/2015*

Parentheses appear frequently in legal text, often combined with citations. We annotate parentheses with the sentence they belong to. Sequences inside the parentheses are not annotated separately, as seen in the following example, containing a single sentence:
- *Ce dernier étant domicilié à l'étranger, il ne peut en effet prétendre à des mesures de réadaptation (art. 8a. 1er paragraphe. Convention de sécurité sociale entre la Suisse et la Yougoslavie du 8 juin 1962).*

Colons should not be annotated as a sentence-terminating character, unless the colon is immediately followed by a newline. The reasoning here is that a sequence ending in a colon followed by a line break usually introduce a list or block quote, which should be annotated separately to the introductory sentence.

Lists are annotated differently depending on its type. For lists with incomplete sentences as list items, often ended with a semi-colon, the whole list is annotated as a sentence. The following example consists of 2 sentences, the introductory sentence to the colon and 1° to the period.
- *Au cours du délai fixé par la juridiction pour accomplir un travail d'intérêt général, le condamné doit satisfaire aux mesures de contrôle suivantes:*
  1° Répondre aux convocations du juge de l'application des peines;
  2° (...) une affection dangereuse pour les autres travailleurs.

However, if the list items themselves are sentences, the list number (or letter) and items are both annotated as one sentence each, the reason being that they express separate thoughts. In the example below we have 3 sentences (introductory, list number, list item).

- *Considérant en droit:*
  1.- En instance fédérale, peut seul être examiné le point de savoir si la commission de recours a exigé à bon droit de la recourante une avance de frais de 500 fr. pour la procédure de recours de première instance.

Ellipses are used to indicate when part of a sentence or part of the document are left out. The following example shows the use cases for ellipses. The first ellipsis is annotated separately, as it indicates sentences that are missing. The second ellipses indicates, that part of that single sentence was left out and is therefore not annotated separately.

- *(...) La faute de X. est d'une exceptionnelle gravité tant les faits qui lui sont reprochés (...), commis avec une certaine froideur sont insoutenables et comportent un caractère insupportable pour les victimes.*

Footnotes / Endnotes convey additional information to the reader. Indicators for end- and footnotes such as numbers or letters should always be annotated as being inside the sentence span, even if they occur after the sentence-terminating character. As an example, the sequence below is just one sentence, with "(2)" as the indicator:

- La loi ne dispose que pour l'avenir; elle n'a point d'effet rétroactif. (2)

Furthermore, endnotes appearing as numbered lists, should be annotated as following the guidelines for lists. In the example below, (2) is one sentence, followed by a normal sentence:

- (2) Le remplacement des membres du Parlement a lieu conformément aux dispositions de l'article 25.

3.2 Tokenizer

We implemented an aggressive tokenizer based on Regex to segment text into tokens, also employed in other research [10, 29]. This tokenizer was utilized for all languages. Words, numbers and special characters such as newlines and whitespace are separated into individual sequences. This was done to ensure no information (e.g., a line break indicating a sentence boundary), vital to the SBD process, was lost. An example is showcased below; tokenized whitespace is left out for clarity's sake:

- *D._ est entré à l'école le 16 juillet 1979.*
- *D | . | _ | est | entré | à | l | ' | école | le | 16 | juillet | 1979 | .*

4 EXPERIMENTAL SETUP

We conducted a series of experiments to answer our research questions posed in Section 1.2. Firstly, we compared selected existing models to establish a baseline performance. Secondly, we trained and evaluated various monolingual and multilingual models based on CRF, BiLSTM-CRF, and transformers, comparing them to baselines. Lastly, we evaluated the multilingual models' performance on unseen data in a zero-shot experiment.

4.1 Baseline Systems

We conducted a thorough evaluation of several widely used systems utilizing various technologies, including CoreNLP, NLTK, Stanza, and Spacy, which served as our baselines. In the following section, we will provide a detailed description of each system.

4.1.1 NLTK. A fully unsupervised SBD system created by Kiss and Strunk [15]. The main thought behind the system is that most falsely predicted sentence boundaries stem from periods after abbreviations. The system therefore discovers abbreviations by looking at the length, the collocational bond, internal periods and occurrences of abbreviations without an ending period of each token in the text. We test a pre-trained model as well as a model trained on our data.

4.1.2 CoreNLP. A rule-based system from the Stanford CoreNLP toolkit [18], which predicts sentence boundaries based on events like periods, question marks, or exclamation marks.

4.1.3 Stanza. A multilingual system based on a BiLSTM model [25]. We only use the first part of its NLP pipeline, the tokenizer. It addresses tokenization and sentence splitting jointly, treating it as a character sequence tagging problem, predicting if a character is the end of a token or sentence.

4.1.4 Spacy. A multilingual system [12] with pre-trained models using technologies like CNN and transformers. For our purposes, only the tokenizer and sentence splitter were used.

4.2 Our Models

Following the works presented in Section 2, we chose to test models based on CRFs, BiLSTM-CRFs and transformers. We further describe these models in the following subsections. For testing, we trained[6] and evaluated monolingual models for each language as well as multilingual models using all languages except Portuguese, once for laws, once for judgments and both types together.

4.2.1 Conditional Random Fields. The tokenizer in Section 3.2 tokenized input text, including whitespaces. Each token was translated into a list of simple features representing the token, and the features of tokens within a pre-defined window around the token were added. Window sizes for each feature varied, inspired by Glaser et al. [10] and Savelka et al. [29], as shown in Table 2. We labeled input data using the "BILOU" system following Lin et al. [17].

For training our CRF models, we used the python-crfsuite[7] implementation. We trained each model for 100 iterations, with regularization parameters 1 and $1e^{-3}$ for C1 and C2, L-BFGS as the algorithm, and including all possible feature transitions.

4.2.2 Bidirectional LSTM - CRF. A BiLSTM connects two LSTMs with opposite directions to the same output, allowing it to capture information from past and future states at the same time. The outputs of each LSTM are concatenated into a representation of each input token. For a BiLSTM-CRF model, a CRF layer is connected to the output of the BiLSTM network, using the aforementioned representation as features to predict the final label.

[6] GPU: NVIDIA GeForce RTX 3060 TI, CPU: Intel Core i5-8600K CPU @ 3.60GHz
[7] https://pypi.org/project/python-crfsuite/

Table 2: Description of CRF-Features

Feature	Description	Window
Special	Each token is categorized using the following translation: Sentence-terminating tokens as "End", opening and closing parentheses as "Open" and "Close" respectively, newline characters as "Newline", abbreviation characters as "Abbr" and the rest as "No".	10
Lowercase	The token in lowercase.	7
Length	The length of the token.	7
Signature	Each character is represented using the following translation: Lower case and upper case character are rewritten as "c" and "C" respectively, digits are written as "N" and special characters as "S".	5
Lower	Whether the first character is lower case.	3
Upper	Whether the first character is upper case.	3
Digit	Whether the token is a digit.	3

We utilized the Bi-LSTM-CRF[8] library to train our models. We used a word embedding dimension of 128, hidden dimension of 256 and a maximum sequence length of 512. The batch-size was 16 with a learning rate of 0.01 and a weight decay of 0.0001. We trained each model for 8 epochs and saved the model with the smallest validation loss. We extracted word embeddings for training from our documents. To label the training data, we utilized the "BILOU" labeling system described in Section 4.2.1. For training, gold sentences were put together into batches with a token-limit of 512 to simulate longer paragraphs.

4.2.3 Transformer. Transformers are a type of NN that utilizes self-attention mechanisms to weigh the importance of difference parts of the input when making predictions. Transformer models such as BERT use a multi-layer encoder [34] to pre-train deep bidirectional representations by jointly conditioning on both left and right context across all layers [6]. Thus, we can fine-tune transformer models to the SBD task by adding an additional output layer. In our case we used a pre-trained model[9] based on DistilBERT [28], a smaller, more lightweight version of BERT, for all languages on our SBD task.[10] We trained the models using PyTorch[11] and Accelerate[12] with the Adam optimizer for 5 epochs with a batch-size of 8 and learning rate of $2e^{-5}$.

A limitation of DistilBERT is the input length limit of 512 tokens because the runtime of the self-attention mechanism scales quadratically with the sequence length. This issue is exacerbated, since DistilBERT relies on a WordPiece Tokenizer [32], splitting the text into subwords resulting in a higher token count per sequence. Thus, to get around the 512 token-limit, each document was split into sentences using the gold annotation. Each consecutive sentence was added to a collection until the total length was as close to the token-limit as possible. Next, the model predicted the sentence boundaries for each collection. Sentences longer than 512 tokens were truncated.[13] An obvious downside to this solution is that the input text already has to be split into sentences or short sections, making it difficult to apply BERT models to unknown text.

For future work, it would be interesting to see, whether it is feasible to chain SBD models (i.e., first, apply a CRF model on the input text to split the text into sections smaller than 512 tokens and second apply a transformer based model). Another solution might be using pre-trained transformer models that support longer input text utilizing an attention mechanism scaling linearly with sequence length, such as Longformers [2].[14]

4.3 Evaluation

A characteristic of the SBD task is the inherent imbalance towards non-sentence boundary labels, as each sentence can at most have two sentence boundaries. Thus, to more accurately score our models, we used commonly utilized measures to evaluate our models - Precision (P), Recall (R) and F1-Score (F1). Although the SBD task is not yet solved in specialized domains, it is comparatively easier than other NLP tasks such as Questions Answering or Summarization. Because SBD is a pre-processing task, it is necessary to achieve higher scores to prohibit the propagation of errors into downstream tasks. Thus, we expect that state-of-the-art SBD models exhibit F1-scores in the high nineties to be useful in practice.

For the evaluation process, we let models predict the sentence spans of every document. These annotated spans are tokenized by our tokenizer (Section 3.2). Each token is then assigned a binary value, depending on whether it was a sentence boundary or not. This decouples the predicted sentence spans or boundaries from the tokenizer used, as the tokenizer of some models might designate a slightly different token as the first or last in a sentence, further described in the following example in French: *"C'est en outre ..."*. While our tokenizer would designate "C" as the first token in the sequence, a different tokenizer might designate "C'" or even "C'est". This would lead to a wrongly predicted sentence boundary when compared to the gold annotations, although the prediction was actually correct.

True and predicted labels for each document type are compared using Scikit-Learn to calculate binary F1-Scores. Scores are averaged for subsets: "Laws" encompass Criminal Code, Civil Code, and Constitution; "Judgments" include various court decisions.

We trained each CRF model once and the BiLSTM-CRF and transformer models 5 times with random seeds, reporting the mean performance including standard deviation. If not specified differently, reported values are binary F1-scores.

[8] https://github.com/jidasheng/bi-lstm-crf
[9] https://huggingface.co/distilbert-base-multilingual-cased
[10] For efficiency, we used a smaller model; a bigger model is advisable for future work.
[11] https://pytorch.org/
[12] https://huggingface.co/docs/accelerate/index

[13] This led to some wrongly predicted sentence boundaries, however this only occurred a few times and is therefore insignificant to the overall score.
[14] Unfortunately, to the best of our knowledge, so far there do not exist multilingually pretrained efficient transformer models.

Table 3: Mean (±std) F1 Score of baseline and multilingual models on all languages and the Portuguese zero-shot experiment. Best scores are in bold.

Language	French		Spanish		Italian		English	German		Portuguese (Zero-shot)	
Type / Model	Judg.	Laws	Judg.	Laws	Judg.	Laws	Judg.	Judg.	Laws	Judg.	Laws
CoreNLP	74.7	76.7	71.4	89.0	79.8	75.6	81.7	69.0	64.0	-	-
NLTK	72.5	75.8	70.2	89.2	72.3	66.3	77.2	72.3	73.8	64.9	57.0
NLTK-train	82.9	75.8	72.1	81.6	84.8	77.5	84.9	74.2	73.5	71.7	64.3
Spacy	86.6	67.2	60.0	70.3	73.9	73.7	79.7	87.5	67.0	59.0	77.7
Stanza	81.9	81.0	83.2	90.2	85.7	87.4	92.3	72.6	64.7	88.6	73.4
CRF	97.8	98.1	94.8	98.9	97.3	97.7	95.1	95.2	91.6	90.2	78.6
BiLSTM-CRF	97.6±0.3	**98.5**±0.2	97.3±0.1	**99.3**±0.2	97.8±0.1	**99.2**±0.1	95.4±0.3	**97.2**±0.2	97.5±0.5	93.0±0.6	73.2±3.3
Transformer	**98.3**±0.1	98.1±0.2	**97.8**±0.1	99.0±0.0	**98.3**±0.1	99.1±0.1	**97.0**±0.1	92.9±0.2	**97.6**±0.1	**93.6**±0.3	**91.3**±1.1

5 RESULTS

5.1 Baseline Models

The performance of baseline models in Section 4 on each language in our dataset is summarized in the upper section of Table 3.

The results for the baseline models are clearly lower than the reported scores for user-generated content by Read et al. [26], supporting the hypothesis that the performance of out-of-the-box models is subpar on legal data for all tested languages. The difference in performance could be explained in one part by the special sentence structures presented in Section 3.1, while the challenging nature of legal text accounts for another part.

Of interest is the gap between NLTK and NLTK-train in most languages, as training NLTK improves its ability to recognize and correctly predict abbreviations. This showcases that abbreviations are one part of the challenging nature of legal texts. To note here is that Spacy uses a slightly different notion of a sentence compared to the other models: Usually, when two sentences are separated by a newline character, the newline character would not be part of any sentence span, however Spacy would include it in the span of the second sentence. This leads to a false prediction, even though Spacy correctly recognized that there are two sentences. Therefore, the scores Spacy achieves are lower than expected.

5.2 Monolingual Models

We report the performance of our trained monolingual models in Table 4. Each model was trained and tested on the same language.

We observe that each model's performance, when applied to their training subset, reaches high nineties for almost all languages, significantly improving over the baseline models from Section 5.1 and comparable to reported SBD system performance on English news articles [26]. Our models also perform similarly to the reported performance of CRFs and CNNs on English [27, 29], as well as CRFs and NNs on German datasets [10].

Comparing the performance of the models when trained on one subset and evaluated on the other unseen set, i.e. a zero-shot experiment, the transformer model outperforms CRF and BiLSTM-CRF on most languages, dropping down to 81.8% on the Italian dataset, comparable to the best baseline models, when trained on judgements and evaluated on laws. Unsurprisingly, the models' performance in the zero-shot experiment is almost always lower than the performance on the subset they were trained on. This gap can be explained by the large difference of writing and formatting styles between judgements and laws, with the transformer model being the best at generalizing knowledge between the two subsets. We further hypothesize that it was easier for the models to generalize their knowledge to different domains, when being trained on judgements, than when being trained on laws, resulting in higher scores on unseen data. One factor here might be that legal text in judgements contain a higher variety of different sentence structures, while laws usually reuse the same structures.

The CRF and BiLSTM-CRF model showcase especially poor performance on the Spanish dataset when trained on laws and evaluated on judgements, with scores down to 43.4% and 54.3%. We hypothesize that both models possess a worse ability to generalize to different domains compared to transformer models.

To conclude, while training on both laws and judgments together not always produces the absolute best performance, it is most robust and does not result in performance degradation.

5.3 Multilingual Models

The performance of our multilingual models trained on laws and judgements is reported in the lower section of Table 3. Each multilingual model was trained on all languages except Portuguese.

The multilingual models clearly outperform the baseline models by a large margin, with F1-scores up to 99.2%. Both the BiLSTM-CRF and transformer models perform very well, with transformers performing slightly better on judgements and BiLSTM-CRFs on laws. The CRF model is close behind the other two, mostly reaching scores in the higher nineties. Comparing the performance of the multilingual models to the monolingual models, showcases that there is no loss of performance when training on a much larger dataset, with multilingual models performing comparably or in case of the transformer and BiLSTM-CRF model even better than the monolingual models on each respective language.

5.4 Zero-shot Experiment on Portuguese Data

We conducted a more challenging experiment, evaluating multilingual models on Portuguese data, comparing them to the baseline. Figure 1 provides an overview, while Table 3 details the differences in judgements and laws against the baseline.

Table 4: Mean (±std) F1 Score of monolingual models on their respective language. Best scores are in bold.

	Language	French		Spanish		Italian		English	German	
	Type	Judg.	Laws	Judg.	Laws	Judg.	Laws	Judg.	Judg.	Laws
Model	Trained on									
CRF	Judg.	97.9	73.2	97.0	98.3	**98.5**	95.6	96.8	97.8	76.5
	Laws	78.5	**98.8**	54.3	**99.6**	88.6	**99.6**	-	75.8	**97.7**
	Laws + Judg.	97.8	**98.8**	97.0	99.5	98.3	99.5	-	97.2	97.2
BiLSTM-CRF	Judg.	97.3±0.3	56.7±3.0	94.7±0.5	92.1±0.9	95.9±0.3	71.3±2.2	**97.3**±0.4	97.0±0.3	76.9±0.4
	Laws	66.1±4.2	97.9±0.2	43.4±6.8	98.7±0.2	74.1±1.2	98.4±0.2	-	71.9±2.5	97.3±0.3
	Laws + Judg.	97.0±0.4	98.1±0.1	95.6±0.5	98.9±0.4	96.2±0.2	98.2±0.1	-	97.2±0.2	97.6±0.2
Transformer	Judg.	98.2±0.1	84.7±1.2	96.9±0.2	96.9±0.4	97.8±0.2	81.8±0.9	96.5±0.1	98.0±0.2	87.2±0.4
	Laws	92.4±0.5	97.6±0.4	89.5±0.6	97.1±3.7	89.4±0.7	98.8±0.5	-	89.4±0.5	97.4±0.1
	Laws + Judg.	**98.4**±0.1	98.2±0.2	**97.3**±0.1	99.0±0.1	97.1±0.3	99.1±0.1	-	**98.3**±0.1	97.5±0.2

Table 5: Mean F1 Score of monolingual and multilingual models on unseen Portuguese data

Model	CRF		BiLSTM-CRF		Transformer	
Type	Judg.	Laws	Judg.	Laws	Judg.	Laws
Model Language						
French	79.3	75.4	25.5	51.7	82.5	87.1
Spanish	**91.5**	79.4	80.3	**73.5**	88.0	**94.0**
Italian	81.8	**83.3**	12.6	64.8	70.0	73.7
English	90.6	72.1	80.6	62.4	87.6	89.9
German	59.0	25.2	43.6	30.3	79.9	71.1
Multilingual	90.2	78.6	**93.0**	73.2	**93.6**	91.3

For judgements performance is adequate with F1-scores between 90.2% and 93.6%, comparable to user-generated content [26], and outperforming most baselines. However, for laws, only the transformer model scores in the lower nineties, while CRF and BiLSTM-CRF drop to 78.6% and 73.2%, respectively, similar to our usual baseline values. The transformer model's large-scale multilingual pretraining likely makes it more robust to distribution shifts, leading to better cross-lingual transfer to unseen languages than CRFs or BiLSTM-CRFs.

The difficulty of the writing and formatting style in Portuguese law texts could explain the difference between laws and judgements, indicated by lower than usual Portuguese baseline performance. BiLSTM-CRF's reduced performance could also result from the lack of Portuguese word embeddings used in training, as we only extracted embeddings from our training data. To improve BiLSTM-CRF models, future research could explore adding Portuguese word embeddings or using larger, multilingual embedding vocabularies during training. To improve transformer models, fine-tuning larger pre-trained models like XLM-RoBERTa [5] on the SBD task could be a potential avenue as they improve significantly in cross-lingual transfer compared to mBERT [6] or DistilBERT [28] models.

When evaluating the effectiveness of monolingual and multilingual models, trained on the entire monolingual dataset, on previously unseen Portuguese data (Table 5), we observe that the multilingual models outperform corresponding monolingual models in most languages, with Spanish being a notable exception. We hypothesize that the disparity in performance is due to close linguistic ties between Spanish and Portuguese, which enabled the Spanish monolingual models to excel in cross-lingual transfer. However, on other languages linguistically less close to Spanish, the multilingual model is expected to perform better than the monolingual ones.

5.5 Inference Time

Table 6 reports the inference times of our multilingual models trained on laws and judgments. We measured inference time three times on both a GPU (NVIDIA GeForce RTX 3060 TI) and a CPU (Intel Core i5-8600K CPU @ 3.60GHz), and show the average. We did not report standard deviation since there were no significant outliers. Notably, the transformer model saw significant improvements in inference time on a GPU. However, CRF does not benefit from GPU evaluation as it uses sequential operations.

Table 6: Mean inference time in minutes (min), seconds (s), milliseconds (ms) for each multilingual model to predict the entire dataset of ~130000 sentences and one sentence, measured on a GPU and CPU

	full dataset (~130000 sentences)		One sentence	
Model	CPU	GPU	CPU	GPU
CRF	11 min 57 sec	-	~5.37 ms	-
BiLSTM-CRF	10 min 6 sec	9 min 23 sec	~4.54 ms	~4.21 ms
Transformer	34 min 26 sec	9 min 18 sec	~15.47 ms	~4.18 ms

Considering the results presented in Sections 5.2, 5.3 and 5.4, inference times and ease of use, a recommendation for the multilingual transformer model can be made for most cases, as long as a GPU is available for inference. For language specific tasks or tasks requiring longer input texts, we recommend the CRF models for the respective language, although they have a longer setup time compared to the BiLSTM-CRF and transformer model.

5.6 Error Analysis

We inspected random samples – two thirds of the Portuguese dataset (8 judgements, 20 laws) – predicted by the multilingual

transformer model for the zero-shot experiment on Portuguese texts. We selected the multilingual transformer following our recommendation in Section 5.5, and the Portuguese dataset because the model already performed very well on the other datasets.

Standard sentence boundaries are rarely missed and the model performs adequately in that regard; yet, we identified a few sources of common mistakes. We discuss examples with |T| and |P| indicating true and predicted sentence boundaries, respectively. Many errors stem from citations and parentheses as shown in the example below:

- (Bittar, Carlos Alberto. *|P|* Direito de autor. *|P|* Rio de Janeiro: Forense Universitária, 2001, p. 143) *|T| |P|*

In this example, we have a citation sentence with periods being wrongly predicted as sentence boundaries inside the citation.

Another source of errors are datafields and headlines, since there is often little indication e.g., a sentence-terminating character, for the model to recognize it as such:

(1) RELATOR: MINISTRO SIDNEI BENETI *|T|*
(2) ACÓRDÃO *|T|*

The model failed to predict a sentence boundary at the end of both sequences. The errors showcased in the examples above mainly stem from our particularly defined sentence structures (Section 3.1.1) as well as the challenging nature of the legal SBD task.

Another set of errors were caused by the different formatting styles and words used in the Portuguese language, unknown to the model, such as:

(1) A Turma, por unanimidade, deu provimento ao recurso especial, nos termos do voto do(a) Sr(a). *|P|* Ministro(a) Relator(a). *|T| |P|*
(2) Exmos. *|P|* Desembargadores MAURÍCIO PESSOA (Presidente), CLAUDIO GODOY E GRAVA BRAZIL. *|T| |P|*

In (1), we have the abbreviation "Sr(a)", which the model did not recognise as such, thus marking the period as a sentence boundary. A similar mistake is shown in (2), with the abbreviation "Exmos".

5.7 Limitations

Due to the language skills of our annotator, we only annotated data from two language groups (Germanic and Italic). Therefore, our languages have high lexical overlap, making cross-lingual transfer comparatively easy. Future work may investigate legal text from additional diverse language groups to build systems even more robust towards language distribution shifts.

The annotator is a native German speaker, with intermediate French language skills. Due to the similarity of Italian, Spanish, and Portuguese to French, and because the SBD task is largely structural, the annotations were possible. However, having the annotations performed by a native speaker in the respective languages may further increase annotation quality. On the other hand, having one annotator (as done in our case) annotate the entire dataset, enables more consistency across languages.

Because of financial limitations, we performed the annotations using only one annotator. Having a second annotator validate the annotations may further increase annotation quality.

Augmenting the qualitative error analysis from Section 5.6 quantitatively may provide more concrete and actionable evidence for improving the systems further. To achieve this, a more detailed annotation of the sentence type would be helpful, so statistics over the sentences can be computed to get quantitative results of the sentence types performing worst.

6 CONCLUSION AND FUTURE WORK

6.1 Answers to the Research Questions

RQ1: *What is the performance of existing SBD systems on legal data in French, Spanish, Italian, English, and German?*
Existing SBD systems are subpar in all tested languages, lower than reported scores by Read et al. [26] on user-generated content, indicating that SBD is not solved in the legal domain.

RQ2: *To what extent can we improve upon this performance by training mono- and multilingual models based on CRF, BiLSTM-CRF and transformers?*
The monolingual models achieved state-of-the-art F1-scores in the high nineties for all tested languages, comparable to reported scores on news articles [26]. The multilingual models performed similarly to monolingual models, demonstrating the potential of training with larger datasets. The transformer model exhibited superior cross-domain transfer compared to CRF and BiLSTM-CRF models.

RQ3: *What is the performance of the multilingual models on unseen Portuguese legal text, i.e., a zero-shot experiment?*
The transformer models performs adequately on the judgements and laws subsets, reaching F1-scores in the lower nineties, demonstrating the best cross-lingual transfer, while the CRF and BiLSTM-CRF models perform decently around 90% on judgements, but drop down to baseline values on the laws, most likely requiring additional optimization.

6.2 Conclusion

In this work, we curated and publicly released a diverse legal dataset with over 130'000 annotated sentences in 6 languages, enabling further research in the legal domain. Using this dataset, we showed that existing SBD methods perform poorly on multilingual legal data, at most reaching F1-scores in the low nineties. We trained and evaluated mono- and multilingual CRF, BiLSTM-CRF and transformer models, achieving binary F1-scores in the higher nineties on our dataset, demonstrating state-of-the art performance. For a more challenging task, we tested our multilingual models in a zero-shot experiment on unseen Portuguese data, with the transformer model reaching scores in the lower nineties, outperforming the baseline trained on Portuguese texts as well as the CRF and BiLSTM-CRF models by a large margin. We publicly release these models and the code for further use and research in the community.

6.3 Future Work

Further improvement for all models might be achieved by preprocessing the input text more, e.g., replacing newlines with spaces, special characters with more widely used equivalent characters e.g., double quotes (“) with single quotes ("). Furthermore, thorough hyperparameter optimization tailored to the specific dataset could improve multilingual CRF and BiLSTM-CRF models. Finally, transformer models may benefit from legal-oriented models [4, 13, 22], larger pre-trained models like BERT [6], or models designed for cross-lingual transfer tasks, like XLM-RoBERTa [5].

Augmenting the dataset with legal texts from multiple languages and documents from various sources like privacy policies and terms of service may improve multilingual models' performance, particularly in the zero-shot scenario. An interesting impact on the model performance could be observed if the sentence spans were labeled with their sentence structure type such as "Citation" (Section 3.1.1) during training instead of being assigned a single label.

An investigation into whether the positive cross-lingual transfer observed in their study also applies to languages from a different family, such as Hungarian. This assumption is based on the common origin of the languages studied, as mentioned in Section 5.

REFERENCES

[1] Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. http://arxiv.org/abs/2210.13448 arXiv:2210.13448 [cs].

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs]

[3] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.00904 arXiv: 2109.00904.

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]* (Oct. 2020). http://arxiv.org/abs/2010.02559 arXiv: 2010.02559.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.48550/arXiv.1911.02116 arXiv:1911.02116 [cs]

[6] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Jinhua Du, Yan Huang, and Karo Moilanen. 2019. AIG Investments.AI at the FinSBD Task: Sentence Boundary Detection through Sequence Labelling and BERT Fine-tuning. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. -, Macao, China, 81–87.

[8] Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S.. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Boulder, Colorado, 241–244.

[9] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 42–47.

[10] Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021. Sentence Boundary Detection in German Legal Documents:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, Online Streaming, — Select a Country —, 812–821. https://doi.org/10.5220/0010246308120821

[11] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* -, - (2012), –.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. - -, - (2020), –. https://doi.org/10.5281

[13] Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. https://doi.org/10.48550/arXiv.2212.08204 arXiv:2212.08204 [cs].

[14] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain.

[15] Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32, 4 (Dec. 2006), 485–525. https://doi.org/10.1162/coli.2006.32.4.485

[16] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, Vol. -. -, Phuket, Thailand, 79–86.

[17] Chun-Wei Lin, Yinan Shao, Ji Zhang, and Unil Yun. 2020. Enhanced Sequence Labeling Based on Latent Variable Conditional Random Fields. *Neurocomputing* 403, - (May 2020), –. https://doi.org/10.1016/j.neucom.2020.04.102

[18] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/P14-5010

[19] Ditty Mathew and Chinnappa Guggilla. 2019. AI_Blues at FinSBD Shared Task: CRF-based Sentence Boundary Detection in PDF Noisy Text in the Financial Domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. -, Macao, China, 130–136.

[20] Denis Newman-Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert Lai. 2016. A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. *AMIA Joint Summits on Translational Science proceedings. AMIA Summit on Translational Science* 2016 (July 2016), 88–97.

[21] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 19–35. https://aclanthology.org/2021.nllp-1.3

[22] Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? https://doi.org/10.48550/arXiv.2211.17135 arXiv:2211.17135 [cs].

[23] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. https://doi.org/10.48550/arXiv.2301.13126 arXiv:2301.13126 [cs].

[24] Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online only, 32–46. https://aclanthology.org/2022.aacl-main.3

[25] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

[26] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jrgen Solberg. 2012. Sentence Boundary Detection: A Long Solved Problem?. In *Proceedings of COLING 2012: Posters*. -, Mumbai, India, 985–994.

[27] George Sanchez. 2019. Sentence Boundary Detection in Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota, 31–38. https://doi.org/10.18653/v1/W19-2204

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *ArXiv* abs/1910.01108, - (2019), –.

[29] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States. *TAL* 58, 21 (2017), –.

[30] Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S. Alexander, Jayla C. Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, Kevin D. Ashley, Alexandra Ashley, Karl Branting, Mattia Falduti, Matthias Grabmair, Jakub Harašta, Tereza Novotná, Elizabeth Tippett, and Shiwanni Johnson. 2021. Lex Rosetta: Transfer of Predictive Models across Languages, Jurisdictions, and Legal Domains. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ACM, São Paulo Brazil, 129–138. https://doi.org/10.1145/3462757.3466149

[31] Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, Vol. -. -, -, 5.

[32] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast WordPiece Tokenization. - -, - (2020), –. https://doi.org/10.48550/ARXIV.2012.15524

[33] Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* -, - (2012), –.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs]

Received 16 February 2023; accepted 12 April 2023; revised 2 May 2023