IIT Kanpur

SURGE 2023
End Term Report

SENTENCE BOUNDARY DETECTION IN INDIAN LEGAL TEXT

Mentor : Dr. Ashutosh Modi

Submitted by:
Aniket Suhas Borkar: 2330298
Divyansh: 2330137
Shourya Trikha: 2330125
Vikas Yadav: 2330595

Abstract

Sentence boundary detection is a natural language processing task that aims to identify the boundaries between sentences in a given text or document. It involves determining where one sentence ends and another begins, often using punctuation marks, such as periods, question marks, and exclamation marks, as indicators. Accurate sentence boundary detection is crucial for various language processing tasks, including machine translation, text summarization, and sentiment analysis. Models trained on usual text perform poorly on legal text. Accurate sentence boundary detection in legal text is important for various applications, such as legal information retrieval, contract analysis, and legal document summarization. There are several previous works done on this task but on legal text the performances are not at par. We have started with the paper, "Efficient Deep Learning-based Sentence Boundary Detection in Legal Text" which performs sentence boundary detection on US legal text documents, however they restrict to a limited set of delimiters for sentence boundary detection. Our aim is to develop an annotated legal text corpus and subsequently training the models to perform better on Indian legal text documents.

1 Introduction

Sentence boundary detection is a fundamental task in natural language processing (NLP) that plays a crucial role in various language-related applications, including machine translation, text summarization, information retrieval, and sentiment analysis. Accurate sentence boundary detection is particularly significant in legal texts, where precise segmentation of sentences is essential for effective comprehension, analysis, and processing. Despite the importance of accurate sentence boundary detection in Indian legal texts, existing NLP tools and models primarily cater to standard English text, thereby overlooking the intricacies and nuances specific to Indian legal language. Consequently, the performance of such tools in this domain is suboptimal, leading to potential errors and inaccuracies in downstream NLP tasks. This report aims to address the aforementioned gap by focusing on sentence boundary detection in Indian legal text. We aim to build and train models on dataset specific to Indian legal documents.

2 Methodology

The methodology for the Sentence Boundary detection project involves the development of a model to effectively and accurately predict sentence boundaries in legal text, specifically in the Indian context. To achieve this goal, the project followed a three-step approach. The first step involved learning various deep learning architectures as well as literature review of the state-of-the-art methods for sentence boundary detection. The second step involved experimentation with the already existing US legal text data with different architectures and preexisting models and comparing their performance. The architectures used in this step have been detailed in the following section. The third step involved preparing a dataset of Indian Legal documents and using them to train various deep learning architectures.

3 Architectures Used

In order to predict whether a given sentence delimiter, such as "." is a sentence boundary or not, we must provide the model with a context on both sides of the delimiter. Thus the left and right context is fed as input to the embedding layer. Taking inspiration from Sheik et al.[1] we decided to go with the following architectures to perform sentence boundary detection. We finally proceeded with the CNN model.

3.1 CNN

Convolutional Neural Networks are a type of artificial neural network which consist of at least one convolutional layer, possibly followed by pooling and fully connected layers. The convolutional layer performs a convolution operation on the inputs. A pooling layer is used to reduce the dimensionality of the data.

In our use-case, the CNN model used consists of a 1D convlutional layer with a kernel size of 5, and a total of 6 filters. The outputs of all the filters is concatenated

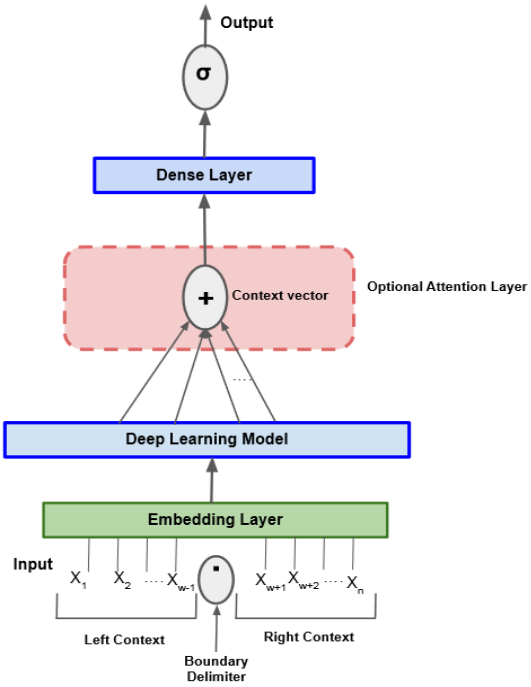

Figure 1: Block diagram of the architecture used, proposed in [1]

and fed through a global max pooling layer, and then fed through a 250-dimensional hidden layer with ReLU activation before the final prediction layer. For each delimiter, the model predicts whether it is a sentence boundary or not.

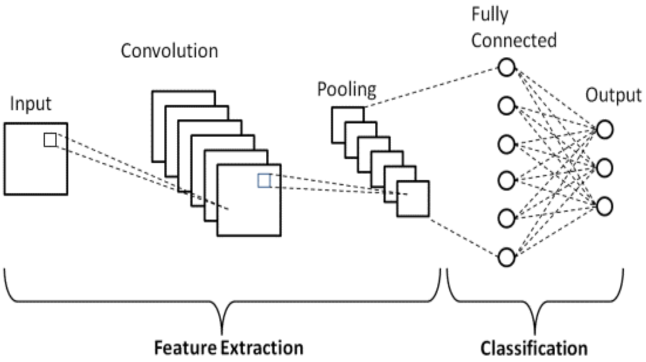

Figure 2: Block diagram of the CNN architecture

3.2 LSTM

An LSTM (Long Short Term Memory) is a type of recurrent neural network, which has the ability to learn long term dependencies in data. Traditional RNNs suffer from the "vanishing/exploding gradient" problem, which makes it difficult for them to capture information over long sequences. LSTMs address this issue by introducing a more sophisticated memory cell that can selectively remember or forget information over time.

An LSTM consists of a memory cell, which has an input gate, a forget gate and an output gate. These gates each having a sigmoid activation, together control how information propagates in an LSTM. This makes LSTMs well-suited for various tasks involving sequential data, such as natural language processing.
In our case, we utilised a 128 embedding size LSTM. The number of features in the hidden state is taken to be 256, and only a single layer of LSTMs was used.

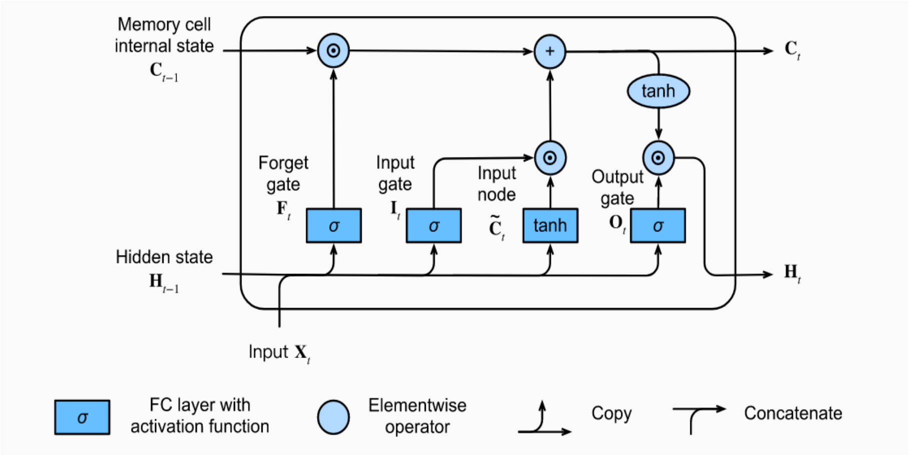
Memory cell internal state $\mathbf{C}_{t-1}$
$\mathbf{C}_t$
tanh
Forget gate $\mathbf{F}_t$
Input gate $\mathbf{I}_t$
Input node $\tilde{\mathbf{C}}_t$
Output gate $\mathbf{O}_t$
σ
σ
tanh
σ
Hidden state $\mathbf{H}_{t-1}$
$\mathbf{H}_t$
Input $\mathbf{X}_t$
σ
FC layer with activation function
Elementwise operator
Copy
Concatenate

Figure 3: Block diagram of the LSTM architecture [2]

3.3 BiLSTM

A BiLSTM (Bidirectional Long Short-Term Memory) is a type of recurrent neural network architecture that consists of two LSTM networks. One of it processes the input sequence in a forward direction, and another processes the input sequence in a backward direction. The forward and backward LSTMs capture the dependencies in the input sequence from both past and future contexts. The outputs of the forward and backward LSTMs are usually combined in some way, such as concatenation or addition, to obtain the final representation of each time step in the input sequence. This allows BiLSTMs to learn from long input sequences while capturing patterns from both the past and future. This is especially useful in applications such as Natural language processing where both the left and right context is important.

3.4 Pre-trained Model : SpaCy

SpaCy is a Natural Language Processing Library, which provides a general purpose sentence boundary detection model as a part of its library. The model is trained on general English text, and does not specifically cater to legal text. The default model used is a simple linear layer with softmax activations which tags whether a given token is a sentence boundary or not.

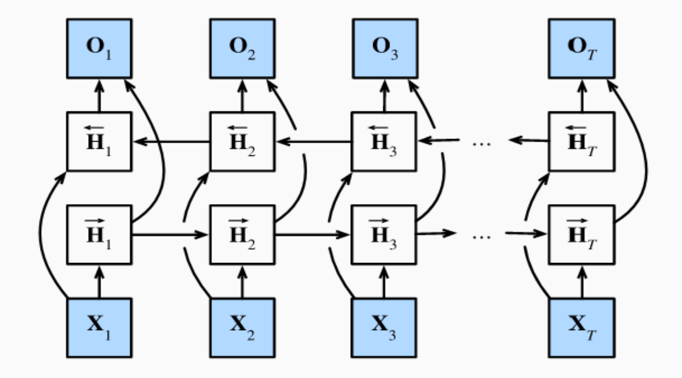
$\mathbf{O}_1$
$\mathbf{O}_2$
$\mathbf{O}_3$
$\mathbf{O}_T$
$\overleftarrow{\mathbf{H}}_1$
$\overleftarrow{\mathbf{H}}_2$
$\overleftarrow{\mathbf{H}}_3$
...
$\overleftarrow{\mathbf{H}}_T$
$\overrightarrow{\mathbf{H}}_1$
$\overrightarrow{\mathbf{H}}_2$
$\overrightarrow{\mathbf{H}}_3$
...
$\overrightarrow{\mathbf{H}}_T$
$\mathbf{X}_1$
$\mathbf{X}_2$
$\mathbf{X}_3$
$\mathbf{X}_T$

Figure 4: Block diagram of the BiLSTM architecture. [2]

4 Dataset

The algorithm was trained using a dataset of 60 Supreme Court of India decisions which were available in the .txt format. These decisions were then put in JSON files along with the list of offsets that basically denotes the sentence boundaries i.e the START and END of a sentence.

4.1 Data Preprocessing

The protocol used for the preprocessing of the data was the manual demarcation of the sentence boundaries. For every decision, the sentences were manually labelled and converted into JSON using Label Studio. These labels are important for comparing the results of the models with baseline frameworks. Then these json files were divided into 3 sets: training set (80%), testing set (10%) and validation set (10%).

Input format of data:

```
{
    "DOCUMENT IDENTIFIER": {
        "Text": "legal text here",
        "Annotation": [
            {
                "start": 0,
                "end": 497
            },
            {
                "start": 1081,
                "end": 1128
            },
            {
                "start": 1129,
                "end": 1301
            },
            {
                "start": 1302,
                "end": 1406
            }
        ]
    }
}
```

5 Future Work

While this report addresses the gap in sentence boundary detection in Indian legal text and presents a CNN-based model trained on a dataset of Supreme Court of India decisions, there are several avenues for future research and improvement in this area. The following are potential directions for future work:

- **Expansion of the dataset**: The current dataset used for training the model consists of 60 Supreme Court of India decisions. To further enhance the performance and generalization of the model, it would be beneficial to expand the dataset by including a larger and more diverse collection of Indian legal documents. This could involve incorporating judgments from lower courts, legal articles, legislative texts, and other relevant sources.

- **Annotation quality and consistency**: The accuracy and consistency of sentence boundary annotations play a crucial role in training robust models. In future work, efforts should be made to ensure high-quality and consistent annotations of sentence boundaries in Indian legal text. This could involve refining the annotation guidelines, conducting multiple rounds of annotation, and employing expert annotators with domain expertise to improve the reliability of the annotated dataset.

- **Fine-tuning and transfer learning**: Transfer learning techniques can be explored to leverage pre-trained models on large-scale language tasks. Fine-tuning the models on the specific task of sentence boundary detection in Indian legal text could potentially improve the performance and efficiency of the model. Pre-training the models on a large corpus of general Indian legal text or related domains could capture more nuanced linguistic patterns specific to Indian legal language.

- **Investigation of domain-specific features**: Indian legal text exhibits specific linguistic characteristics, including unique sentence structures, legal jargon, and cultural references. Future work could explore the incorporation of domain-specific features into the model architecture. This could involve analyzing syntactic and semantic patterns in Indian legal language and integrating them into the model to enhance its understanding and prediction capabilities.

6 Conclusion

Sentence boundary detection is a crucial task in natural language processing, with significant implications for various language-related applications in the legal domain. This report addressed the gap in existing NLP tools and models such as SpaCy that primarily cater to standard English text, overlooking the intricacies and nuances specific to Indian legal language. Our focus was on sentence boundary detection in Indian legal text, aiming to develop an effective and accurate model trained on a dataset of Supreme Court of India decisions.

Through the utilization of a Convolutional Neural Network (CNN) model, we trained it for predicting sentence boundaries in Indian legal text. The model incorporated left and right context inputs to provide the necessary information for determining whether a given delimiter marks a sentence boundary. The experimental evaluation on the dataset demonstrated the model's effectiveness in identifying sentence boundaries, thereby contributing to more accurate comprehension, analysis, and processing of Indian legal documents. The outlined future work provides a roadmap for further research, aiming to enhance the model's performance, expand its applicability, and advance the field of Indian legal language processing.

References

[1] R. Sheik, G. T, and S. Nirmala, "Efficient deep learning-based sentence boundary detection in legal text," in *Proceedings of the Natural Legal Language Processing Workshop 2022*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 208–217. [Online]. Available: https://aclanthology.org/2022.nllp-1.18.

[2] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *arXiv preprint arXiv:2106.11342*, 2021.