Sentence Boundary Detection in Legal Texts
Grading: Option 3

Stanford CS224N Custom Project

Krithika Iyer
Department of Computer Science
Stanford University
ksiyer@stanford.edu

Abstract

fThis report contains results from an effort to detect sentence boundaries in complex legal text (that do not conform to standard English syntax) using Transformer architecture based neural nets. Segmenting sentences in legal decisions are very challenging for existing algorithms as legal text do not conform to standard linguistic sentence structures. This project did not utilize any additional features for flagging punctuation and other syntactic elements in the text. Even without the use of additional features, the project demonstrated superior results over baseline studies that utilized such features. The system achieved a 97% F1-score and 96% and 97% for Precision and Recall for detecting sentence boundaries in legal text, the highest achieved so far. Visualization of attention heads are also presented to aid the understanding of the focus of attention mechanism in processing text that do not conform to standard sentence structures.

1 Introduction

The sentence is one of the basic building blocks in most NLP systems. Many definitions for what a sentence is (not) can be found in NLP and linguistic literature. Regardless of how the sentence is defined, sentence boundary detection (SBD) is a critical foundational task in many NLP applications and tasks. Incorrect SBD can propagate and generate noise and errors in NLP tasks. Thus SBD (also often referred to as sentence segmentation or sentence boundary disambiguation) plays a critical role in the development of practical NLP applications.

Despite its importance, SBD has received much less attention in recent NLP research efforts. Latest in NLP technologies such as Transformer architectures, etc. have not been used to achieve better SBD. Existing sentence boundary detection (SBD) systems are based on a number of assumptions about linguistic and sentence structure that do not hold true for legal text. Legal decisions are not a mere curiosity for NLP research. In common law system, (the norm in US), decisions made by judges are important sources of interpretation of law. The availability of large corpus of legal decisions in digital form combined with recent advances in NLP systems hold enormous potential for the creation of automated legal assistants that can help bring equity to the administration of justice. SBD is a necessary first step in such an endeavour in order to accurately extract meaning and identify precedents that build on each other.

2 Related Work

Related work from three distinct domains are reviewed in this subsection: (1) systems for detecting SBD in traditional well-formed, curated, and edited text, (2) state of the art for detecting SBD in legal domain, and (3) recent efforts aimed at visualizing and interpreting the attention mechanism in transformers.

SBD of standard English text is considered a "solved-problem" with many mature and robust solutions and has not received much attention in recent years. In their seminal paper, appropriately titled, "Sentence Boundary Detection: A Long Solved Problem?," Read et al., discuss the performance of several algorithms for detecting sentence boundaries across different corpora. The results range from 95% to 97.6% for different systems. [1] The different algorithms for SBD consist of decision tree classifier, Support Vector Machines (SVM) and Naive Bayes, and unsupervised approaches such as Punkt. [2] [3] [4] These algorithms perform well for processing text that conform to standard English perform very poorly in certain domains. These algorithms require much customization to produce reasonable results in domains such as Biomedicine, Legal, Finance and Tax texts.

Legal opinions and decisions, whether issued by a court or by an administrative tribunal, pose a difficult challenge to general purpose SBD algorithms. Sentences in legal text can be extremely long. Figure 1 illustrates a long and complex sentence typical of legal text. Two recent efforts have focused on SBD using customized Punkt tokenizer and Conditional Random Field (CRF) algorithms. CRF and customized Punkt yield results in the 75% to 90% accuracy range. [5] [6]. Bidirectional LSTM is utilized in [6], along with seven manually extracted features for annotating the input text.

Recent research efforts have demonstrated that certain attention heads correspond well to linguistic notions of syntax and coreference [7] [8]. The former [7] has demonstrated "[attention] heads that attend to the direct object of verbs, determiners of nouns, objects of prepositions, and coreferent mentions with remarkable high accuracy". -

2.1 SBD in Legal Texts

Unique challenges in the legal domain may be seen in Figure 1. Parentheticals, sentences within sentences (often with citations), are quite challenging because they contain several nested sentences without the traditional period-end-marker. Case names, abbreviations of codes, different courts, embedded page numbers (as part of a citation), citations, and foot-notes are some of the additional challenges that lead to SBD errors in legal domain. Based on recent research efforts that have demonstrated attention mechanism's ability to capture syntactic structures in standard English, we are lead to believe (and hypothesize) that the same attention mechanism can capture some structures (nested sentences, lists within nested sentences, unorthodox use of punctuation marks, and others) in long legal texts.

3 Approach

The primary obstacles faced by existing SBD algorithms are (i) lack of conformity to known sentence structures, (ii) the length of sentences, and (iii) the use of punctuation marks (especially the period) as non-sentence ending character. Will attention ("Attention is all your need") work in the legal domain and help NLP systems easily navigate around the SBD challenges posed by conventional algorithms? Can attention mechanism compensate for the lack of additional features about the legal text? The project decided to evaluate and benchmark the strengths and limitations of attention mechanism in Transformers, which can "attend to" and focus on critical elements in extremely long sentences. The project decided to investigate if attention heads can effectively distinguish among three periods (...) ellipsis, a single period (.) abbreviation, and esentence end-marker.

The PyTorch version of pre-trained BERT model for Token Classification was used in this study. It consisted of 12 layers, each layer consisting of BERT-attention, BERT-Intermediate, and BERT-output layers. More details are shown in Fig 2. The four probabilities generated (by the classifier layer) for the token labels are processed through numpy to get the predicted label for each token. BertViz was utilized for the visualization of the internal states of the model. [8]

4 Experiments

The experiments utilize a model that was pre-trained using text that conform to standard sentence structures. It is being utilized for detecting boundaries in non-conforming text. 90% of the sample data-set was used to fine tune the pre-trained model.

As used in the statute, "'act in furtherance of a person's right of petition or free speech under the United States or California Constitution in connection with a public issue' includes: (1) any written or oral statement or writing made before a legislative, executive, or judicial proceeding, or any other official proceeding authorized by law; (2) any written or oral statement or writing made in connection with an issue under consideration or review by a legislative, executive, or judicial body, or any other official proceeding authorized by law; (3) any written or oral statement or writing made in a place open to the public or a public forum in connection with an issue of public interest; (4) or any other conduct in furtherance of the exercise of the constitutional right of petition or the constitutional right of free speech in connection with a public issue or an issue of public interest." (§425.16, subd. (e), italics added; see Briggs v. Eden Council for Hope & Opportunity (1999) 19 Cal. 4th 1106, 1117-1118, 1123 [81 Cal.Rptr.2d 471, 969 P.2d 564] [discussing types of statements covered by anti-SLAPP statute].)

Figure 1: A sample passage from a legal decision. It contains a very long and complex sentence and a citation sentence. The first sentence contains a quotation which in turn contains another quotation and list. The second sentence illustrates the use of periods that are not sentence ending.

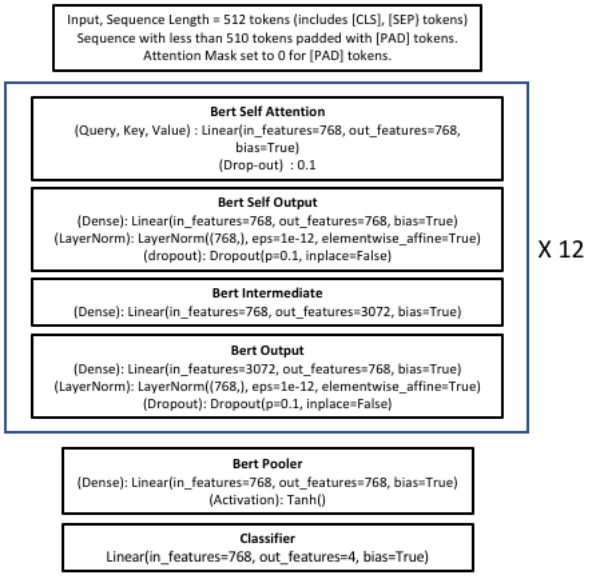

Figure 2: Configuration of the Bert Token Classification Model used in the study.

4.1 Data

Dataset consisting of 20 decisions from four different areas with manual sentence delineation was used for training the model. [9] A new dataset was curated from supreme court decisions to train tokenizers and models for use in the legal domain. [10] The first dataset consisted of 80 decisions from four different legal domains. The manual sentence delineation was based on the protocol described in [5]. Links to data sets and preprocessing procedures are described in Appendix.

Dataset	Min	Max	Mean
bva	1	134	20.73
cyber crime	1	276	19.82
ip	1	334	21.42
scotus	1	614	24.20

Table 1. Number of words per sentence (input sequence) related statistics for the four data sets used in the initial experiments.

4.2 Evaluation method

The evaluation is based on F-1, Precision and Recall scores for the evaluation dataset. 10% of the dataset was used for evaluation. Seqeval package was utilized to compute the scores. During initial experimental runs, it was determined that the system predicted correctly for sentence tokens and predicted all padded tokens incorrectly. The output of the model was modified to take only sentence based tokens into the evaluation process and ignore the padded tokens.

4.3 Experimental details

The experiments were run on Google Collaboratory notebook environment. The Pro versions were used to gain access to High-RAM (higher memory) virtual environment. Several iterations were utilized to determine a combination of max length of input tokens, batch size, etc so that the model would not run out of memory. Max Length (number of tokens) was varied from 75 tokens to 256 tokens per input sequence to the model. The batch size (due to memory constraints) was varied between 4 and 16. Future experiments may increase the size to 32. For sentences with less than 256 tokens were filled with [PAD] tokens. The attention mask set to 0 for padded tokens.

During initial experimental runs, the model created predictions for [PAD] tokens it was supposed to ignore. A large number of such predictions (for the [PAD] tokens) were incorrect. This appears to be due to a bug in HuggingFace's implementation. The sequence length for subsequent experiments were increased to 512 with 510 tokens from legal text padded with [CLS] and [SEP] tokens at the beginning and end. When multiple sentences were combined to create such sequences, a conscious decision was made not to put the [SEP] token between sentences because such tokens would clearly mark ends of sentences and may have acted as a feature.

4.4 Results

4.4.1 Comparisons to Baseline Results

The baseline research, in addition to manually created labels to delineate sentences, utilized seven additional features. These features are (for each token in the input): isUpper, isLower, isDigit, isSpace, isPunctuation, Next Word Capitalized, Previous Word Lower, and Previous Word Single Character. This research effort did not use any additional features.

The confusion matrix from this research effort and the baseline research are shown in Figure X. The performance scores are listed in Figure Y. As one can clearly see, even without the use of additional features, the Transformer based current effort performs much better than results from the baseline Bi-LSTM network with additional features.

Most of the misclassifications in this study stem from punctuation marks, and some are due to missing punctuation marks in the data itself. The misclassifications are further analyzed in detail in the following subsection. The confusion matrix shows that the "Inside" token is misclassified 95 times as "End" and 95 times as "Begin". These 95 + 95 are the same error - the system misclassifies

	Predicted		
Actual	Begin	Inside	End
Begin	6993	816	95
Inside	668	193158	870
End	113	786	7330

	Predicted		
Actual	Begin	Inside	End
Begin	2428	63	11
Inside	95	71818	95
End	4	61	2435

Figure 3: Confusion Matrix. Baseline (left), Transformer (right)

	Precision	Recall	F1-Score	Support
Beg	0.885	0.9	0.892	7774
Inside	0.992	0.992	0.992	194760
End	0.891	0.884	0.887	8295

	Precision	Recall	F1-Score	Support
Beg	0.96	0.97	0.97	2502
Inside	0.91	0.92	0.91	2556
End	0.96	0.97	0.97	2500

Figure 4: Performance Results. Baseline (left), Transformer (right)

a token in a sequence of tokens as “End” and the next token as the “Begin” of the next sentence. The performance figures (Precision, Recall and F-1 scores) are higher for the present effort than the baseline results for both "Begin" and "End" tokens (the actual determination of sentence boundaries). The baseline study shows better performance for detecting inside tokens. It may be due to the fact that the punctuations are clearly marked and provided as a feature to the neural network. The performance of the Transformer based approach can be further improved through additional fine-tuning and more samples.

4.4.2 SBD Errors

During validation of the model, 123 sequences with some mismatched predictions in them were captured and the text causing the incorrect predictions were analyzed. 18 errors were related to the use of three periods (...) in different contexts. In 11 instances, the ellipses (...) that were simply part of a sequence (with no end or beginning) were incorrectly classified as end of one sentence and beginning of the next sentence. The prediction pattern in such cases were identical (first period - part of (inside) a sentence, second period - end of the sentence, and the last period - beginning of the next sentence. When “. . . ” were actually part of an end of one sentence and beginning of another (6 instances) the system’s predictions were random and did not follow a pattern. In two instances, the data actually contained four periods and the system could not label them correctly. About 20 mis-classifications were due to text in the footnote sections of the decision text, either [9] the "[" or the "]" was mis-classified.

The email read, in part, "I told you I would not contact you by mail anymore but I am sorry, I am in agony. I’m thinking about you all the time. You really are my dream girl.... I am blinded with affection for you. I did not ask for this. Nope, it’s all your fault.... Please don’t cat dance on my emotions by failing to respond to me at all. *(excerpt from a decision that included actual quotes from harassment emails sent by a defendant in a criminal case. The defendant had used four periods in the email)*

Importantly, we said that "[Abercrombie] had suggested that the surfers had endorsed Abercrombie’s t-shirts. Accordingly, [Downing] concluded that ‘it is not the publication of the photograph itself . . . that is the basis for [plaintiffs’] claims, but rather, it is the use of the [plaintiffs’] likenesses and their names pictured in the published photographs." [8] Id. *(quoting Downing, 265 F.3d at 1003) (emphasis added). (a nested quotation inside another quotation)*

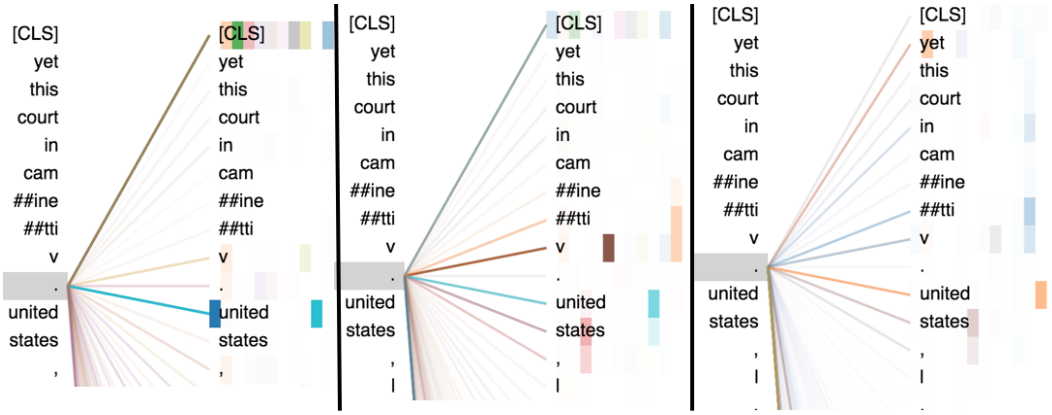

Figure 5: Attention heads in layers 3, 4, and 5. Non end of sentence period (in Caminetti v. United States

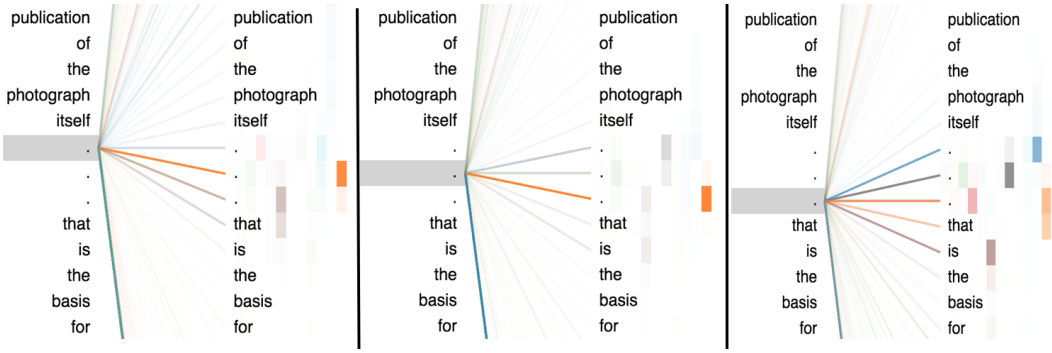

Figure 6: Attention heads in layer 4. Ellipsis (...)

" Campbell, 510 U.S. at 578-79, 114 S. Ct. 1164 (internal quotation marks, alteration and citation omitted). Courts, therefore, must examine "whether and to what extent the new work is transformative. . . . [T]he more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use." *(use of four dots, the last dot to denote end of a sentence. Only one such occurrence in the entire dataset)*

Sub-heading in decision text (often just two words) were mis-classified. The sub-headings in the dataset were marked as a separate sentence, but the system classified them as part of an existing sentence. The system was able to correctly several highly complex sentences. The following is an example of the system's ability to detect SBD (in this case had one error):

The parties differ in their interpretation of how § 114(j)(11) applies the grandfathering provision. Defendant contends that the plain text of the statute yields a two-part analysis to determine whether status as a PSS is granted under the DMCA; first, that the service provides sound recordings by ïnoninteractive audio-onlyn̈eans, and second, that it has been providing s̈uch transmissionss̈ince July 31, 1998. Def.'s Mot. Dismiss 16 (citing § 114(j)(11)). Defendant argues that this is a disjunctive test and therefore that it can meet each criteria separately to qualify for PSS rates. Def.'s Mot. Dismiss at 11-12. Defendant argues that a plain reading of the statute reveals that the language s̈uch transmissionsm̈eans the type of transmissions laid out in the beginning of that sentence, or ïnoninteractive audio-only subscription digital audio transmissions,r̈egardless of whether provided to the same customers, or a different group of 151*151 customers. Def.'s Reply Pl.'s Opp. Def.'s Mot. Dismiss 6-8 (citing 17 U.S.C. § 114(j)(11) (2015)).

Towards the end, the text **Def's Reply Pl.'s** was segmented after the Reply and before Pl.

Figure 7: Attention heads in layer 5. Ellipsis (...)

4.4.3 Visualization of Attention Heads

Attention information for three different uses of period, namely in ellipsis (...), abbreviation (.) and end of sentence (.) are illustrated in Figures 5 through 7. Significant differences in the way the different periods within the ellipse sequence and periods as abbreviation and sentence end can be seen from these images. Figs 6 and 7 clearly illustrate the token the first and second periods in the ellipse attend to is much different from the tokens the third period of the ellipse attends to. The tokens the periods attend to is also different between 4th and 5th layers. Beyond the 5th layer, (i.e., in higher layers) the periods tend to focus more on the end [SEP] token. The attention behavior of period (used in abbreviation context) exhibit different patterns than the ones observed for an ellipse.

5 Analysis

The attention heads seem to focus on both position and on context. When the experiments were run with a single sentence per sequence, the system was able to predict all sentence boundaries accurately and did not make any errors. The same system made errors while predicting labels it was supposed to ignore. The 100% success rate is most likely due to position information because it is much easier to flag start and end when there is only one sentence per sequence. When multiples sentences were combined together without the [SEP] token, the SBD accuracy dropped by a small amount.

When sentences were combined together to form sequences of 510 tokens, the system produces incorrect predictions. In all incorrect predictions were found in 123 sequences. Some sequences had multiple incorrect predictions. 443 incorrect predictions were found (among the 123 sequences). Out of the 443, 180 consisted of twin errors, i.e., a continuing sentence was predicted to end and the following token was flagged as the beginning of the next sequence. The same error (End-Begin) was counted as two. When punctuation related error (ellipses and others) were filtered out, 154 mismatches were found. Out of the 154, close to 100 incorrect predictions were found to be in foot-notes in the decisions. The system is able to decipher sentence boundaries correctly, but is not able to do the same for text in footnotes.

Access to notebook containing full analysis of the incorrect predictions is provided in Appendix.

6 Conclusion

The project has demonstrated the feasibility of using transfer learning and transformer based neural nets for robust SBD in legal domain. The system investigated by this project produces superior results than state of the art algorithms and such superior performance is accomplished without the use of additional features that are difficult to derive. The system appear to face challenges in the processing of footnotes. Some automated pre-processing of text from notes may help alleviate this shortcoming. The visualization of attention heads provide a glimpse into what tokens are attending to in a non-conforming text environment.

7 Future Work

The project has curated a new dataset consisting of 39,564 paragraphs containing complex and citation sentences extracted from 8,396 U.S Supreme Court decisions. This dataset can be used to:

- train (in unsupervised mode) modern tokenizers such as the Byte Pair Encoding Tokenizers to enable them to handle legal text.
- further investigate attention heads to evaluate how they capture non-syntactic patterns in long legal sentences.

Ultimately legal opinions may be treated as a distinct language called Legalese and separate tokenizers and models may be developed to *translate* from Legalese to standard English (parse complex and long sentences into simple senteces) so NLP tasks such as summarization, QA and others may be easily implemented using legal opinions.

8 Acknowledgement

Sample code provided at Hugging Face transformer repository and BertViz repository were utilized in this project. [11] [8]

References

[1] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, December 2012.

[2] Michael D. Riley. Some applications of tree-based modelling to speech and language. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989.

[3] Dan Gillick. 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*, NAACL-Short '09, page 241–244, USA, 2009. Association for Computational Linguistics.

[4] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4):485–525, December 2006.

[5] Jaromir Savelka, V.R. Walker, M. Grabmair, and K.D. Ashley. Sentence boundary detection in adjudicatory decisions in the united states. *TAL Traitement Automatique des Langues*, 58:21–45, 01 2017.

[6] George Sanchez. Sentence boundary detection in legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 31–38, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. *ArXiv*, abs/1906.04341, 2019.

[8] Jesse Vig. A multiscale visualization of attention in the transformer model. pages 37–42, 01 2019.

[9] https://github.com/jsavelka/sbd_adjudicatory_dec.

[10] https://chartbeat-labs.github.io/textacy/.

[11] https://huggingface.co/transformers/model_doc/bert.html.

A Appendix : Supplemental Information

A.0.1 Data Sets

Dataset with manually delineated sentence boundaries:

https://colab.research.google.com/drive/1PPsg7IMz_AMY5eCRpB3beoofG2x6WaGc

8K decisions from the US Supreme Court:

https://colab.research.google.com/drive/1Ejxg-aMCPixeOAjLamu7MIPdsnY5DN1R

A.1 Models

The following uses 512 token sequences to train the model. Sentences (of varying lengths) are combined into sequences of 510 length (+ 2 required tokens)

https://colab.research.google.com/drive/1kTVkHrAi6gXd9korZYjt6cdraw2USCZu

Following uses variable length tokens in sequences and pads the rest with [PAD] tokens. The model is supposed to ignore [PAD] tokens. The attention mask is set to 0 for the padded tokens. (Appears to be a bug in the pre-trained model.)

https://colab.research.google.com/drive/1JHEG1682-LGH_jhvnNuK3FErs_i6QuHh#scrollTo=AEUTF69oY0cv

A.2 Detailed Analysis of Incorrect Predictions

The following notebook shows analysis of incorrect predictions

https://colab.research.google.com/drive/1L7qehL8HuhbvGMrG0yqtHr1VtXNDWGT-

A.3 Visualizations of Attention Heads

Please scroll all the way to the bottom of the notebooks to view the visualization. You can view the attention data for each of the twelve layers.

Sentence Ender - single period (.)

https://colab.research.google.com/drive/1Gg1vGWQumZ2W6ZhRg-LwtTAdIsUdIDfJ

Abbreviation - single period (.)

https://colab.research.google.com/drive/1JHEG1682-LGH_jhvnNuK3FErs_i6QuHh

Ellipses - three periods (...)

https://colab.research.google.com/drive/1Wy3A6GysPxTt2-dc-qy-grQFZHhAj_cB