Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220477

Multi-Granularity Argument Mining in Legal Texts

Huihui Xu [a,b,1], Kevin Ashley [a,b,c]

[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *Learning Research and Development Center, University of Pittsburgh*
[c] *School of Law, University of Pittsburgh*

Abstract. In this paper, we explore legal argument mining using multiple levels of granularity. Argument mining has usually been conceptualized as a sentence classification problem. In this work, we conceptualize argument mining as a token-level (i.e., word-level) classification problem. We use a Longformer model to classify the tokens. Results show that token-level text classification identifies certain legal argument elements more accurately than sentence-level text classification. Token-level classification also provides greater flexibility to analyze legal texts and to gain more insight into what the model focuses on when processing a large amount of input data.

Keywords. Argument mining, Information retrieval, Natural language processing, Deep learning

1. Introduction

Argument mining is "the automatic discovery of an argumentative text portion, and the identification of the relevant components of the argument presented there." [1]. The goal is to identify and extract the structure of inference and reasoning expressed as arguments presented in natural language [2]. Legal argument mining identifies and extracts arguments in legal texts.

In previous work, we applied and demonstrated that supervised machine learning (ML) and deep learning methods can classify sentences of legal cases in terms of the roles they play in a legal argument to some extent.

In this paper, we take legal argument mining to a finer-grained level – token-level argument mining where the tokens are words. That is, we treat it as a word classification task. Token-level argument mining has several potential advantages. First, it is more robust against errors in sentence segmentation [3]. Secondly, it can efficiently handle single sentences that exhibit multiple argumentative elements. For example, as shown in Figure 1, different parts of a single sentence have been labeled as conclusion and reason. If we apply sentence-level classification methods for each label to the same sentence, we confuse the classifier and lose ordering information as compared with training on those sub-sentences. Finally, token-level argument mining can provide insights about the contributions of particular words to sentence-level classification.

[1]Corresponding Author: huihui.xu@pitt.edu

> Allowing the appeal, that s. 23(1) of the Social Assistance Act places a mandatory responsibility upon social service committees to provide assistance for all persons in need as defined in s. 19(e) of the Act.

Figure 1. An example of a legal summary sentence whose parts are labeled with two argumentative elements. Green-colored text represents conclusion, and blue-colored text represents reasons.

Our contributions in this work are, first, to apply token-level argument mining to legal texts. Secondly, we show that this token-level approach improves the accuracy of classifying sentences in terms of legal argument elements. Finally, our error analysis shows new ways to understand the significance of certain tokens/words in classifying sentences by legal argumentative roles.

2. Related Work

Argument mining in the legal domain includes training classifiers on different types of extracted features to classify premises and conclusions [4,5], investigating discursive and argumentative characteristics of legal documents [6], identifying argument schemes [7] or rhetorical roles that sentences play in legal cases [8], summarizing legal cases in terms of argument elements [9,10], and accounting for sentence position and embedding in legal argument classification [11].

Some recent argument mining research has focused on a more granular level, the token-level, which means assigning labels to every word. For example, [3] showed that the token-level argument mining employed in Argument Unit Recognition and Classification (AURC) retrieves a larger number of arguments than sentence-level mining. [12] also treated the Kaggle competition "Feedback Prize - Evaluating Student Writing" as a token-level argument mining task.

Some sequential labeling techniques have been applied in this context, like BERT, Conditional Random Fields, and Bi-LSTM [3,13]. This conceptualized token-level classification resembles the classic sequence labeling tasks in NLP like Named Entity Recognition (NER). Hidden Markov Models (HMM), Maximum entropy Markov models (MEMMs) [14], and Conditional Random Fields (CRF) are the most commonly used sequential labeling techniques in the pre-neural model era. Recently, researchers have applied neural models to tackle sequence labeling problems such as convolution networks [15], bidirectional LSTM-CRF models [16], and BERT-CRF [17].

As far as we know, token-level argument mining has not yet been applied in legal argument mining. We have applied it to a corpus of expert-annotated legal cases and summaries as described below.

3. Dataset

Our dataset comprises 28,733 legal cases and summaries prepared by attorneys, members of legal societies, or law students and provided by the Canadian Legal Institute

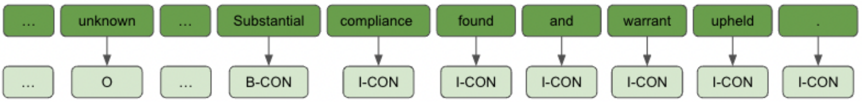

Figure 2. An example of using BIO format to tag every token in a conclusion sentence.

(CanLII).[2] As noted our IRC type system for labeling sentences in legal cases and case summaries includes: **Issue** – Legal question which a court addressed in the case; **Conclusion** – Court's decision for the corresponding issue; **Reason** – Sentences that elaborate on why the court reached the Conclusion. All un-annotated sentences are treated as non-IRC sentences.

We employed two third-year law school students to annotate sentences from the human-prepared summaries in terms of issues, reasons, and conclusions. They annotated 1049 randomly selected case/summary pairs. Cohen's κ [18] metric is used to assess the degree of agreement between two annotators. The mean of Cohen's κ coefficients across all types for summaries is 0.734; the mean for full texts of cases is 0.602. Both scores indicate substantial agreement between two annotators according to [19]. The full texts annotation agreement is lower than that of summaries since the sentences of full texts and summaries are not in a one-to-one mapping.

The BIO or IBO tagging scheme was first proposed in [20]. We adapt the BIO tagging format to our annotated summary/full text pairs. One advantage of this tagging format is it allows tokens to carry both the sentence structure and sentence type information. As shown in the Figure 2, the B-prefix of a tag indicates the beginning of an annotated conclusion sentence, the I-prefix of a tag indicates the token is inside a conclusion sentence, while the O tag indicates the token does not belong to any typed sentence.

4. Experiment

We pre-processed our dataset using the BIO format: we first tokenized all the sentences in summaries and full texts, then assigned the corresponding BIO tags to every token. Those BIO-tagged tokens were then put into the pretrained Longformer [21] model for token classification. We chose Longformer over the traditional BERT [22] model because of its ability to process longer documents. The maximum input length is 1024 tokens due to the GPU limitation.[3] We decided to segment the full text documents into multiple chunks of length 1024 to avoid information loss. We experimented with two types of Longformer: Lonformer-base-4096 and Longformer-large-4096.[4] We split our datasets of summaries and full texts into 80% training, 10% validation and 10% test sets.

[2] https://www.canlii.org/en/
[3] We use a single NVIDIA Titan X GPU with 12 GB memory.
[4] https://github.com/allenai/longformer

Table 1. Results of BIO token-level classification on summaries and full texts. All the results are reported in terms of precision, recall and F_1 scores. The scores inside parentheses are produced by Longformer-base-4096, while the scores outside of parentheses are produced by Longformer-large-4096.

	Summary						
	B-Issue	I-Issue	B-Reason	I-Reason	B-Conclusion	I-Conclusion	O
Precision	0.83 (0.79)	0.83 (0.80)	0.72 (0.67)	0.75 (0.70)	0.83 (0.77)	0.80 (0.73)	0.78 (0.77)
Recall	0.79 (0.78)	0.78 (0.81)	0.80 (0.75)	0.80 (0.76)	0.84 (0.80)	0.82 (0.72)	0.75 (0.72)
F1-score	0.81 (0.78)	0.81 (0.81)	0.75 (0.71)	0.77 (0.73)	0.83 (0.78)	0.81 (0.72)	0.77 (0.74)
	Full-texts						
	B-Issue	I-Issue	B-Reason	I-Reason	B-Conclusion	I-Conclusion	O
Precision	0.66 (0.62)	0.80 (0.75)	0.54 (0.44)	0.69 (0.64)	0.53 (0.46)	0.65 (0.61)	0.98 (0.98)
Recall	0.55 (0.52)	0.70 (0.69)	0.36 (0.36)	0.63 (0.62)	0.43 (0.44)	0.63 (0.61)	0.98 (0.98)
F1-score	0.60 (0.56)	0.74 (0.72)	0.43 (0.40)	0.66 (0.63)	0.47 (0.45)	0.64 (0.61)	0.98 (0.98)

Table 2. Results of classification on summaries and full texts. All the results are reported as F_1 scores.

	Summary				Full text			
	Issue	Reason	Conclusion	Non-IRC	Issue	Reason	Conclusion	Non-IRC
Longformer(large)-BIO	0.81	0.77	0.87	0.79	0.66	0.68	0.67	0.98
Longformer(base)-BIO	0.82	0.72	0.81	0.77	0.63	0.67	0.64	0.98
Longformer(base)-no BIO	0.75	0.73	0.80	0.75	0.49	0.30	0.49	0.95
Longformer(large)-no BIO	–	–	–	0.58	–	–	–	–
Legal-BERT	0.76	0.73	0.81	0.76	0.52	0.47	0.56	0.98
BERT	0.73	0.70	0.79	0.69	0.50	0.49	0.52	0.98

5. Results

Table 1 shows the results of token-level classification in summaries and full texts. As seen in the table, the classification results on the summaries are better than on the full texts in terms of precision, recall, and F1 score. The better results on summaries are expected because the summaries are shorter than full texts and more clearly organize the sentences. To determine the sentence type from the resulting token labels, we used the token type that appears most frequently in the sentence. Table 2 reports the results of assigning sentence type utilizing the token labels.

For purposes of comparison, we trained three techniques on sentence-level annotation: Legal-BERT [23], BERT [22] and Longformer. None of these baseline techniques employ token-level annotation. In order to compare the results across different models, we tested them on the same test set. As shown in the Table 2, Longformer(large)-BIO achieved better F_1 scores in sentence labeling across all sentence types (e.g., issues, reasons, and conclusions).

6. Discussion

We trained Longformer on annotated summary and full text sentences, respectively. It confirms that the BIO approach classifies the sentences more effectively. For token-level classification, as shown in Table 1, we can see that the F_1 scores of I-prefixed token types (i.e., inside) are higher than B-prefixed token types (i.e., beginning). Our intuition is that

I-prefixed token types benefit from more training data, because each annotated sentence has only one beginning token while I-prefixed tokens dominate the rest of the annotated sentence.

After investigating the results of token-level classification, we find that the model is more likely to assign I-Reason to a Non-IRC (O) type in both summaries and full texts. A large portion of those misclassified tokens are stop words, like 'the', 'to', 'of' etc., which are commonly used within issue sentences. Those stop words, of course, appear everywhere in a document; their type depends more on their context than their semantic meaning. The token-based classification indicates some contexts where even a stop word like 'the' appears to have an effect.

We also found that 'HELD' appears most frequently in the correctly classified B-Conclusion tokens in the summaries; 'the' appears most frequently in the correctly classified I-Issue tokens in the full texts. The human summarizers tend to make conclusions more noticeable to readers by using indicators such as 'HELD'. Those indicators are captured by the model.

For sentence-level classification, the sentence type is determined by the token type that appears most often in the sentence. We observed that conclusions in summaries are prone to be misclassified as reasons. We investigated those misclassified conclusion sentences and find most sentences were completely misclassifed on a token-level. That is, the model identified no conclusion tokens. Only one sentence had several correctly identified conclusion tokens including 'support' and 'allowed'. We have been unable to explain the token-level misclassification.

7. Conclusion

In this work, we experimented with multi-granular argument mining from legal texts. We employed two label classification tasks: token-level (i.e., word-level) classification and sentence-level classification. The sentence-level classification is based on the results of the token-level classification. Results showed that token-level classification achieved more accurate sentence classification than state-of-the-art sentence-classification models. The token-level classification not only improved the sentence classification performance but also gave insights into how the model behaves with respect to certain tokens.

In future work, we plan to use the token-based approach to more accurately classify issues, conclusions, and reasons and to use these IRC argument elements to improve automatic case summarization. We will explore using these finer-grained indicators to identify other legal argumentative units, such as factors, and to better evaluate the quality of legal summaries in terms of coverage of argument elements.

Acknowledgement

This work has been supported by grants from the Autonomy through Cyberjustice Technologies Research Partnership at the University of Montreal Cyberjustice Laboratory and the National Science Foundation, grant no. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice. The Canadian Legal Information Institute provided the corpus of paired legal cases and summaries. This work was supported in part by the

University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

References

[1] Peldszus A, Stede M. From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI). 2013;7(1):1-31.

[2] Lawrence J, Reed C. Argument mining: A survey. Computational Linguistics. 2020;45(4):765-818.

[3] Trautmann D, Daxenberger J, Stab C, Schütze H, Gurevych I. Fine-grained argument unit recognition and classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34; 2020. p. 9048-56.

[4] Moens MF, Boiy E, Palau RM, Reed C. Automatic detection of arguments in legal texts. In: Proceedings of the 11th international conference on Artificial intelligence and law; 2007. p. 225-30.

[5] Mochales-Palau R, Moens M. Study on sentence relations in the automatic detection of argumentation in legal cases. Frontiers in Artificial Intelligence and Applications. 2007;165:89.

[6] Mochales R, Moens MF. Study on the structure of argumentation in case law. In: Proceedings of the 2008 conference on legal knowledge and information systems; 2008. p. 11-20.

[7] Feng VW, Hirst G. Classifying arguments by scheme. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies; 2011. p. 987-96.

[8] Saravanan M, Ravindran B. Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artificial Intelligence and Law. 2010;18(1):45-76.

[9] Xu H, Savelka J, Ashley KD. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In: Proceedings of the eighteenth international conference on artificial intelligence and law; 2021. p. 250-4.

[10] Elaraby M, Litman D. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. arXiv preprint arXiv:220901650. 2022.

[11] Xu H, Savelka J, Ashley KD. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In: Legal Knowledge and Information Systems. IOS Press; 2021. p. 33-42.

[12] Ding Y, Bexte M, Horbach A. Don't Drop the Topic-The Role of the Prompt in Argument Identification in Student Writing. In: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022); 2022. p. 124-33.

[13] Ajjour Y, Chen WF, Kiesel J, Wachsmuth H, Stein B. Unit segmentation of argumentative texts. In: Proceedings of the 4th Workshop on Argument Mining; 2017. p. 118-28.

[14] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization. AAAI/IAAI. 2000;2000:584-9.

[15] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of machine learning research. 2011;12(ARTICLE):2493-537.

[16] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015.

[17] Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:190910649. 2019.

[18] Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.

[19] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.

[20] Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Natural language processing using very large corpora. Springer; 1999. p. 157-76.

[21] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. arXiv:200405150. 2020.

[22] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[23] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. arXiv preprint arXiv:210408671. 2021.