Legal_try at SemEval-2023 Task 6: Voting Heterogeneous Models for Entities identification in Legal Documents

Junzhe Zhao and **Yingxi Wang** and **Nicolay Rusnachenko** and **Huizhi Liang**
Newcastle University, Newcastle Upon Tyne, England
{j.zhao33,Y.wang318,nicolay.rusnachenko,huizhi.liang}@ncl.ac.uk

Abstract

Named Entity Recognition (NER) is a crucial subtask of Natural Language Processing (NLP) that involves identifying and categorizing named entities in text. The resulting annotation enables unstructured natural language texts for various NLP tasks, such as information retrieval, question answering, and machine translation. NER is essential in the legal domain as an initial stage for extracting relevant entities. However, legal texts contain domain-specific named entities, including applicants, defendants, courts, statutes, and articles, rendering standard named entity recognizers incompatible with legal documents. This paper proposes an approach that combines multiple model results through a voting mechanism to identify unique entities in legal texts. This study's primary focus is extracting named entities from legal texts in the context of SemEval-2023 Task 6, Sub-task B: Legal Named Entities Extraction (L-NER). The goal is to create a legal NER system for unique entity annotation in legal documents. Our experiments' results and our system's implementation are published and accessible online[1].

1 Introduction and Related Work

The significant growth of legal texts, particularly in highly populated countries, leads to a considerable burden on the workload of competent judges. Automated text structurization is the first and essential step for other downstream tasks that may support judges' work. The distinct nature of the Indian legal process and the terminology employed in legal texts are primary reasons why the SemEval-2023 competition (Modi et al., 2023) encourages research in this area (Kalamkar et al., 2022).

Initially, researchers focused on rule-based methods, such as Ralph Grishman's 1995 NER system (Grishman and Sundheim, 1996), which depended on manually created rules and patterns to identify named entities in text. Machine learning approaches were subsequently employed, with the *Hidden Markov Model* (HMM) (Zhou and Su, 2002) as one of the earliest methods modeling the probability distribution of named entity labels given a sequence of words. It laid the foundation for further research, culminating in the development of neural networks for NER, including the application of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997; Hammerton, 2003), Bidirectional LSTM (Huang et al., 2015), Convolutional Neural Networks (CNN), and Iterated Dilated CNN (IDCNN) (Strubell et al., 2017). The recent emergence of transformers (Vaswani et al., 2017) has significantly impacted the NLP field, with promising results from NER models based on transformers. Numerous research areas are active in NER, including methods combining BERT, CRF, and BiLSTM (Dai et al., 2019; Jiang et al., 2019). These models can be trained on labeled NER datasets to recognize named entities in new texts.

Researchers (Chalkidis et al., 2019, 2020) investigate the application of BERT models in the legal domain, proposing adaptation strategies, introducing a new dataset comprising over 62,000 legislative documents and 4,000 subject matter labels, and demonstrating the potential of pre-trained language models for large-scale multi-label text classification in legal document categorization. Consequently, this paper's main contribution is the adoption of recent advances in heterogeneous models with a voting mechanism for ensembled entity annotation in legal texts.

2 System Overview

Figure 1 presents the architecture of the proposed system, which employs a combination of four heterogeneous models (Section 3) and selects results through a voting mechanism. The system setup process consists of the following steps: 1) creating four

[1] https://github.com/SuperEDG/Legal_NER

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1282–1286
July 13-14, 2023 ©2023 Association for Computational Linguistics

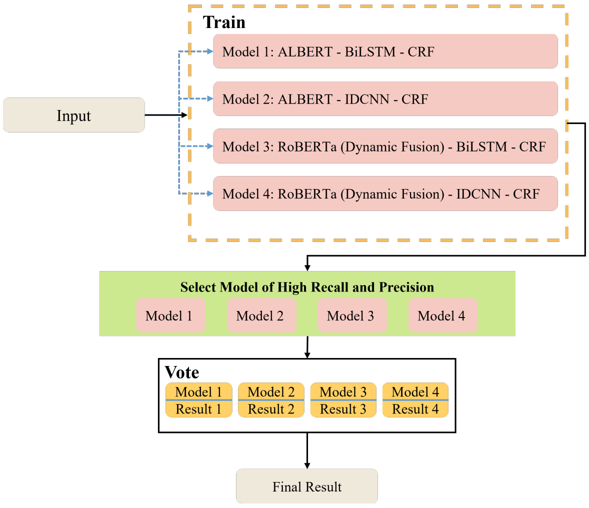

Train
Input
Model 1: ALBERT - BiLSTM - CRF
Model 2: ALBERT - IDCNN - CRF
Model 3: RoBERTa (Dynamic Fusion) - BiLSTM - CRF
Model 4: RoBERTa (Dynamic Fusion) - IDCNN - CRF
Select Model of High Recall and Precision
Model 1
Model 2
Model 3
Model 4
Vote
Model 1 Result 1
Model 2 Result 2
Model 3 Result 3
Model 4 Result 4
Final Result

Figure 1: Multi-model fusion system based on heterogeneous models BERT (ALBERT/RoBERTa) - BiLSTM/IDCNN - CRF with further results selection based on the voting mechanism

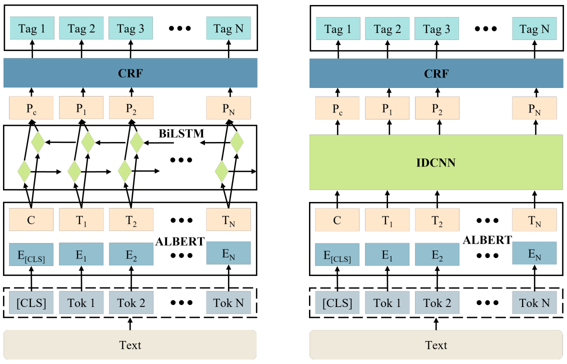
initial models; 2) training each model. The recall and precision metrics are considered to evaluate the performance of these models. The model that demonstrates the best performance based on these metrics is subsequently chosen for prediction. The heterogeneous model results are combined using a textual voting method to derive the final prediction. Ultimately, the voting process entails numerically polling labels from the four models and consolidating them into textual results.

3 Heterogeneous Models

In this paper, we consider *heterogeneous models* as three-tier designed NER annotators (see Figure 2). We consider BERT-based models in the first layer for learning and providing the contextual representation: ALBERT (Lan et al., 2019) and RoBERTa (Liu et al., 2019). The second layer incorporates BiLSTM/IDCNN to model the context. The third layer adopts CRF to constrain predicted labels to ensure validity. Sections 3.1 and 3.2 cover each layer's contents in greater detail.

3.1 ALBERT – BiLSTM – CRF

The ALBERT – BiLSTM – CRF architecture, illustrated in Figure 2a, comprises three main components: BERT for contextual word embeddings, BiLSTM for bidirectional processing of words to capture sequence information, and CRF for probabilistic modeling of label dependencies to generate a label sequence that maximizes joint probability. During inference, the Viterbi algorithm (Forney, 1973) is employed to predict the most likely label sequence. Backpropagation-based gradient descent is used during training to optimize the model by minimizing the negative log-likelihood of the gold standard label sequence. The architecture is organized into the following layers:



(a) ALBERT – BiLSTM – CRF (b) ALBERT – IDCNN – CRF

Figure 2: Architecture of the heterogeneous models

1. Input layer: Processes the string of `Text` tokens from the input sentence, which consists of words or sub-words. BERT converts each symbol E_i into a contextual embedding T_i that represents the token in the context of the complete sentence.

2. Hidden layer: Processes the embedding sequence through the BiLSTM layer, transforming each T_i into a processed embedding P_i, which captures the bidirectional sequence information.

3. Output layer: The CRF layer takes the processed embeddings P_i and generates the final tag sequence, with each tag represented by Tag_i. This sequence corresponds to the named object type for each label in the input sequence.

3.2 ALBERT – IDCNN – CRF

The ALBERT – IDCNN – CRF architecture (Figure 2b), described in Section 3.1, represents a model with the second layer, previously a BiLSTM layer, replaced by IDCNN (Strubell et al., 2017). This model consists of multiple convolutional layers that process BERT-based output to capture local contextual information. In this layer, the IDCNN transforms the T_i embeddings into processed embeddings P_i. The IDCNN layer includes both expanded convolutional and max-pooling layers. The expanded convolutional layer broadens

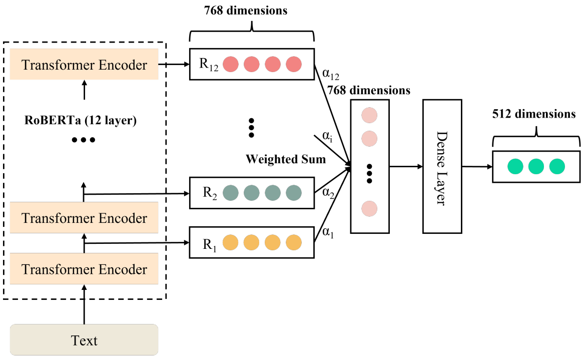
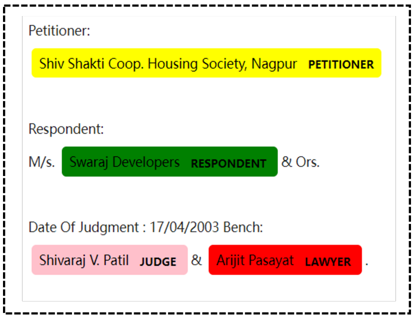


Figure 3: Architecture of Dynamic Fusion of RoBERTa



Petitioner:

Shiv Shakti Coop. Housing Society, Nagpur PETITIONER

Respondent:

M/s. Swaraj Developers RESPONDENT & Ors.

Date Of Judgment : 17/04/2003 Bench:

Shivaraj V. Patil JUDGE & Arijit Pasayat LAWYER .

Figure 4: An illustration of a named entity within a judgment, using distinct colors and text to represent various named entities in the judgment.

the network's receptive field, allowing it to capture a more comprehensive context around each token.

3.3 RoBERTa (dynamic fusion) – BiLSTM/IDCNN – CRF

Dynamic fusion is a technique that combines the outputs of multiple layers of a language model into a single representation (Yunqiu et al., 2022). In this work, we utilize dynamic fusion methods to enhance the efficiency of RoBERTa (Liu et al., 2019) for downstream tasks by leveraging the advantages of various layers. The process involves initializing the model for each layer (R_i). Each layer produces explicitly weighted representations (α_i). The weight values are then determined through training, and the representations produced by each layer are averaged in a weighted fashion. Formula 1 illustrates the initialization of the model, while Formula 2 demonstrates how a fully connected layer reduces the dimensionality to 512 dimensions before creating the dynamic weight fusion structure of the RoBERTa multilayer representation:

$$\alpha_i = Dense_{unit=1}(R_i) \tag{1}$$

$$o = Dense_{unit=512}(\sum_{i}^{n} \alpha_i \cdot R_i) \tag{2}$$

Our experiments consider RoBERTa$_{\text{base}}$ with $n = 12$ layers. The resulting vector o (Formula 2) is then passed to the second layer of the heterogeneous models. Figure 3 illustrates the architecture of the dynamic fusion process.

4 Dataset

In this paper, we utilize the Indian Judicial Decisions dataset (Kalamkar et al., 2022) for model fine-tuning and experimentation. Named entities are extracted from both the preamble and the main text of the judgment. The preamble of a judgment consists of formatted metadata, such as the names of the parties involved, the presiding judge, the representing lawyers, the date, and the court. A representative example is provided in Fig. 4, which contains four named entities: the petitioner (highlighted in yellow), the respondent (highlighted in green), the judge (highlighted in pink), and the lawyer (highlighted in red).

The training dataset includes 9,435 judgments and 1,560 preamble records, while the validation dataset contains 949 judgments and 125 preamble records. In total, the dataset comprises 14 distinct named entity types. The named entity counts are detailed in Table 1.

Named Entity	Training		Validation	
	Judgement	Preamble	Judgement	Preamble
Case Number	1040	0	121	0
Court	1293	1074	178	118
Date	1885	0	222	0
GPE	1398	0	183	0
Judge	567	1758	8	166
ORG	1441	0	159	0
Other Person	2653	0	276	0
Petitioner	464	2604	9	202
Precedent	1351	0	177	0
Provision	2384	0	258	0
Respondent	324	3538	5	310
Statute	1804	0	222	0
Witness	881	0	58	0
Lawyer	0	3505	0	589

Table 1: Details of the Indian legal named entities

Due to the computation complexity of the BERT model attention mechanism, the input sequence has a maximum length of 512 tokens. As such, we analyzed the distribution of sentence lengths within our training and validation sets. We found that the shortest sentence consisted of 14 words, while the longest contained 23,960 words. Figure 5 illus-

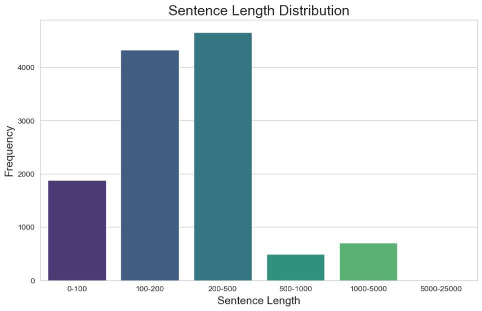


Figure 5: Distribution of Sentence Lengths: The graph represents the distribution of sentence lengths, categorized by length intervals. The x-axis denotes the sentence length, while the y-axis indicates the frequency of sentences within each length interval.

trates the distribution of sentence length intervals. Notably, 90.07% of the sentences were less than 500 words, with only a tiny fraction exceeding this limit. The most common sentence length interval was between 100 and 500 words, constituting the most significant proportion of the distribution.

5 Experimental Setup

BIO Format For our training process, we utilized the training and validation sets provided by the competition organizers. We used the trained model to analyze the test set and submitted three results to the competition. We consider a combined word sense and punctuation for data preprocessing in the splits. For instance, the initial word entry was `"1,31,37,500."`. The first split resulted in «`1,31,37,500.`», whereas the third split separates the prior into seven entries: «`1`»«`,`»«`31`»«`,`»«`37`»«`,`»«`500.`». After segmenting the sentences, we annotated the word using the BIO format, which represents the three possible states of a token in the annotated sequence: "Beginning," "Inside," and "Outside." For example, «`HongKong`»«`Bank`» are annotated as «`B-ORG`»«`I-ORG`». This process results in 29 labels within the material.

Hyper-parameter setting All considered language models are of the BASE size (Devlin et al., 2019). We configure the training process to span 50 epochs. Firstly, we restrict sentence lengths to a maximum of 500 tokens, as 90% of the sentences in the dataset are within this range, ensuring that the training results are not significantly impacted. Sentences exceeding 500 tokens are split into smaller segments with lengths under 500 tokens, allowing for batched training. We utilize a batch size of 4 for this process. Throughout the training period, we apply a dropout rate of 0.1 and a learning rate of 10^{-5}.

Named Entity	Test	F1 (%)
Case Number	61	68.1
Court	148	74.3
Date	111	69.2
GPE	92	64.8
Judge	87	74.0
ORG	80	65.6
Other Person	138	72.5
Petitioner	106	67.2
Precedent	89	58.4
Provision	129	72.3
Respondent	158	61.7
Statute	111	75.2
Witness	29	66.9
Lawyer	295	72.8
All	1631	68.8

Table 2: Results obtained by ALBERT for the manual split of Indian judical decisions dataset, separately for each named entity type and in total (All)

6 Results

Competition Result The final submission, based on the configurations outlined in Section 5, is the result of an ensemble of four heterogeneous models combined using a voting mechanism. This approach secured a sixteenth-place ranking out of 17 competitors, achieving a 51.73% F1 score.

Optimized Testing Result After the competition, we improved the system by replacing BERT with ALBERT in the BERT-BiLSTM/IDCNN-CRF model while keeping the other components unchanged. To evaluate the effectiveness of this change, we divided the dataset into validation and test sets based on the labels, allocating 50% of each of the 14 distinct named entity types to the test set and the remaining 50% to the validation set. This approach was taken since the original test set was not publicly available. The test results, an ensemble of the four heterogeneous models combined using a voting mechanism, are presented in Table 2 and show an average F1 score of 68.8%. Among the distinct named entity types, the *Statute* entity type achieves the highest F1 score (75.2%), while *Precedent* records the lowest (58.4%).

7 Conclusion

In this paper, we present a system for named entity annotation tailored explicitly for annotating objects in legal texts within Indian court judgments. The system employs a voting mechanism over a three-tier, heterogeneous NER model. We propose and utilize a diverse combination of model components, including BERT-based language models for input representation, variations of BiLSTM, and IDCNN encoders, all combined with a CRF module for hidden representation. To assess the performance of our proposed system, we conduct experiments using the Indian Judicial Decisions dataset provided by the competition organizers.

References

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5. IEEE.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Shaohua Jiang, Shan Zhao, Kai Hou, Yang Liu, Li Zhang, et al. 2019. A bert-bilstm-crf model for chinese electronic medical records named entity recognition. In *2019 12th international conference on intelligent computation technology and automation (ICICTA)*, pages 166–169. IEEE.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 Task 6: LegalEval: Understanding Legal Texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics (ACL).

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhang Yunqiu, Wang Yang, and Li Bocheng. 2022. Identifying named entities of chinese electronic medical records based on roberta-wwm dynamic fusion model. *Data Analysis and Knowledge Discovery*, 6(2/3):242–250.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480.