VTCC-NER at SemEval-2023 Task 6: An Ensemble Pre-trained Language Models for Named Entity Recognition

Quang-Minh Tran, Xuan-Dung Doan*
Viettel Cyperspace Center, Viettel Group, Vietnam
minhtq21,dungdx4@viettel.com.vn

Abstract

We propose an ensemble method that combines several pre-trained language models to enhance entity recognition in legal text. Our approach achieved a 90.9873% F1 score on the private test set, ranking 2nd on the leaderboard for SemEval 2023 Task 6, Subtask B - Legal Named Entities Extraction. Our code is available for further exploitation at: https://github.com/tqgminh/SemEval2023_LegalNER_VTCC.

1 Introduction

Named Entity Recognition (NER) is a key problem in natural language processing (NLP) that involves recognizing named entities in text and classifying them into specific types. Recent studies on NER have achieved many promising results. Most of state-of-the-art NER models are based on pre-trained language models, which have been trained on large text corpora and are effective for word representation.

However, Domain-Specific NER, such as Legal NER, remains a challenging task. Named entities in legal texts are more different and specific than named entities in general texts. For example, a person's name in legal texts may refer to a petitioner, respondent or lawyer. Therefore, some pre-trained language models for legal domain have been released for use not only in Legal NER, but also for other NLP tasks in legal texts.

In this work, we propose a model that combines a pre-trained language model and some techniques, such as CRF or dependency parsing, for NER in legal texts. An ensemble method is used to further boost the performance.

2 Related Works

NER in legal texts has been researched for many years and in many languages, but it still remains a challenging task. Dozier et al. (2010) defined five legal named entities in US legal texts (judges, attorneys, companies, jurisdictions and courts) and developed a named entity recognition and resolution system based on pattern-based and statistical models. Cardellino et al. (2017) trained a Named Entity Recognizer, Classifier and Linker by mapping the LKIF ontology to the YAGO ontology and developed a structured knowledge representation of legal concepts. Due to variant writing style and vocabulary across different languages and forms, some datasets have been published to create separate models, such as German LER (Leitner et al., 2020), LegalNERo (Păiș et al., 2021), LeNER-Br (Luz de Araujo et al., 2018), and Legal NER (Kalamkar et al., 2022).

Applying pre-trained language models boost the accuracy of NER tasks. BERT (Devlin et al., 2018) is a language model based on Transformer architecture, and was trained on two different tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). RoBERTa (Liu et al., 2019) is a robust version of BERT, and was trained on a much larger dataset and using a more effective training procedure. XLM-RoBERTa (Conneau et al., 2019) is based on RoBERTa architecture and was trained on the large multi-lingual dataset. DeBERTa-V3 (He et al., 2021) replaced MLM with Replaced Token Detection (RTD) and implemented a gradient-disentangled embedding sharing method. DeBERTa-V3 achieves state-of-the-art performance on many downstream NLP tasks.

Regarding the NLP tasks in legal texts including NER, some pre-trained language models have been trained exclusively in legal texts in order to enhance the performance. For instance, Legal-BERT (Chalkidis et al., 2020), based on BERT architecture, was trained on 12GB of diverse English legal texts and achieved better results than BERT with the same number of parameters on some Legal NLP tasks. InLegalBERT and InCaseLegalBERT (Paul et al., 2022) were initialized from Legal-BERT and

* Corresponding author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 415–419

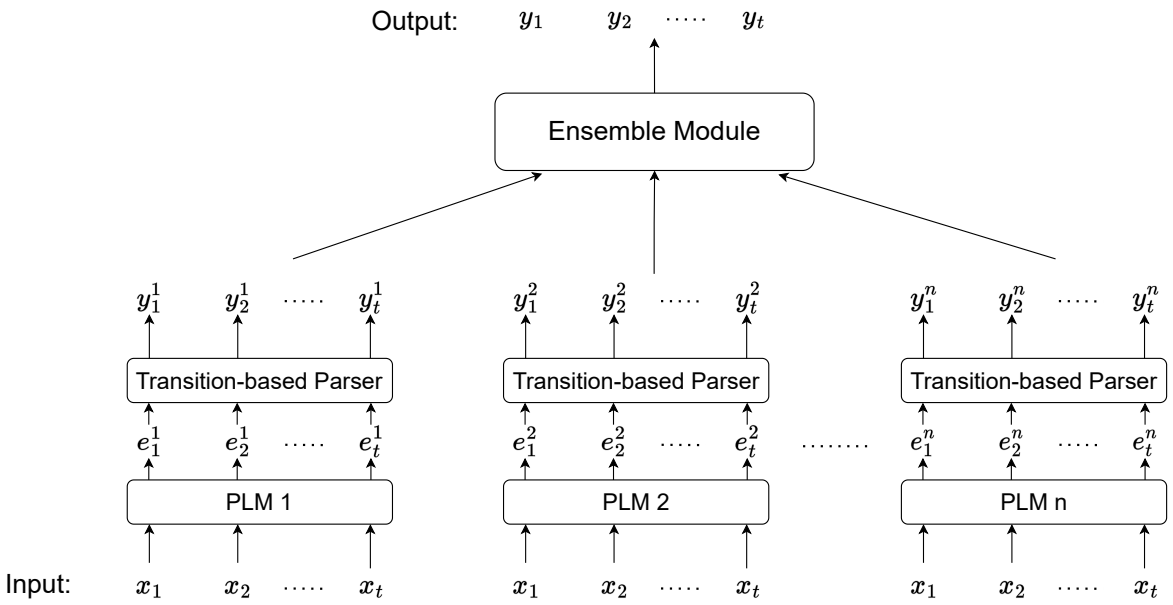

Figure 1: The archiecture of the proposed method

were trained on 27GB of Indian legal documents.

CRF (Collobert et al., 2011) is popular techniques used for NER. Many studies have stacked them on top of other methods and achieved impressive results, as demonstrated by Lample et al. (2016), Ma and Hovy (2016), Peters et al. (2018), Yang and Zhang (2018), Akbik et al. (2019).

3 Methodology

For the Legal NER Subtask in SemEval 2023 Task 6 (Modi et al., 2023), the architecture of the proposed method is illustrated in the Figure 1. Accordingly, we first execute the task with various benchmark pre-trained language models (PLMs) in this research field such as BERT, RoBERT, InlegalBERT, and so on. The output of aforementioned PLMs are then put into an ensemble module for extracting the final results. The processes of the proposed method are sequentially described as follows:

3.1 PLMs for legal name entity extraction task

We processed the benchmark dataset using two approaches. The *first approach* use the model based on the spaCy framework, following the work in Kalamkar et al. (2022). The *second approach* is our custom model using IOB format incorporating with CRF Layer and Dependency parsing as features in the model.

3.1.1 The first approach

With the input dataset, we adopt the pipeline proposed by Kalamkar et al. (2022). This pipeline used a transition-based parser, proposed by Honnibal and Johnson (2015), on the top of a PLM.

For the PLMs, we used RoBERTa, BERT, LegalBERT, InLegalBERT, and XLM-RoBERTa as the backbones of the pipeline, and all experiments were performed using the spaCy framework.

3.2 The second approach

The second approach is our custom model for the task. Specifically, given the input $X = x_1, x_2, ..., x_t$, each word in the input is assigned a different tag depending on whether it is the beginning (B-y), inside (I-y), or outside (O) of a named entity phrase $Y = y_1, y_2, ..., y_t$, respectively. For the NER task, we adopted the baseline model of SemEval 2022 (Malmasi et al., 2022). Furthermore, we also incorporated dependency parsing as a feature of the model. The PLMs were fed subword to obtain contextualized embedding information, which are then concatenated with the dependency feature and part-of-speech feature of each word. Subsequentially, the output are fed into a CRF layer to classify the label of each word in the sequence.

For the backbone PLMs, we selected various models such as XLM-RoBERTa, RoBERTa, DeBERTa, BERT, LegalBERT, InLegalBERT, and InCaseLegalBERT.

.

3.3 Ensemble Module

After obtaining results from two aforementioned approaches, we apply an ensemble strategy for each approach in order to improve the performance. Specifically, ensemble strategy combines multiple models through majority voting with weights. In particular, For each token x_i, we predict the final label y_i based on the majority label f of the models, which is formulated as follows:

$$y_i = f(w_1 h_1(x_i), w_2 h_2(x_i), \ldots, w_n h_n(x_i)) \quad (1)$$

where $h_j(x_i)$ denotes the output of token x_i by the PLM h_j. w_j denotes the weight of the PLM h_j, which can be set equally ($w_i = w_2 = ... = w_n$) or dynamically based on the validation process. Specifically, the values of the weights are discussed in the experiment section.

4 Experiment

4.1 Dataset

Our experiments were conducted on the Legal NER dataset, introduced by Kalamkar et al. (2022), which includes training and dev data. The dataset, collected from Indian court judgments, consists of two parts: preamble and judgment. Each preamble and judgment were split into sentence for training and evaluation. Named entities to be recognized belong to 14 labels.

4.2 Setting

For the first approach, we trained for 30,000 iterations with 4 sentences per batch. The gradient accumulation was set to 3. The optimizer used was Adam, with an initial learning rate of $5e-5$. We selected the checkpoint with the best performance on the dev set for each model. First, we combined five models with RoBERTa, BERT, LegalBERT, InLegalBERT, and XLM-RoBERTa as the backbone, with equal weights. Second, we assigned weights of 4, 2, 1, 1, and 1 to the RoBERTa, BERT, LegalBERT, InLegalBERT, and XLM-RoBERTa models, respectively, based on their performance on the dev set. We assigned higher weights to RoBERTa and BERT since they outperformed the other models, as shown in Table 1. Due to submission limitations, we only submitted the predictions of two highest models. As shown in Table 3, RoBERTa outperformed BERT, so we assigned a higher weight to RoBERTa in the ensemble.

For the second approach, we set the gradient accumulation to 4 and trained for 50 epochs. The optimizer was Adam, with an initial learning rate of $1e-4$. We saved the checkpoint of the model after each epoch and kept the ten latest checkpoints after training. The checkpoint with the best performance on the dev set was selected for inference. We used the Transkit (Nguyen et al., 2021) to obtain the dependency feature. Ensemble method with equal weights were then applied, similar to the first approach. All experiments were executed on a single 40G Nvidia-A100 GPU.

4.3 Results

Model	Micro-F1 (%)
RoBERTa + Transition-based Parser	89.29
BERT + Transition-based Parser	89.31
LegalBERT + Transition-based Parser	88.07
InLegalBERT + Transition-based Parser	88.77
XLM-RoBERTa + Transition-based Parser	88.55
Ensemble model with equal weights	**91.12**
Ensemble model with dynamic weights	89.91

Table 1: Evaluation on the dev set in the first approach

Model	Micro-F1 (%)
XLM-RoBERTa+CRF	90.25
DeBERTa-V3+CRF	90.59
RoBERTa+CRF	87.89
BERT+CRF	89.12
LegalBERT+CRF	87.95
InlegalBERT+CRF	88.94
InCaseLawBERT+CRF	88.57
XLM-RoBERTa+Dep+CRF	89.94
DeBERTaV3+Dep+CRF	**90.73**
RoBERTa+Dep+CRF	87.45
BERT+Dep+CRF	89.14

Table 2: Evaluation on the dev set in the second approach

Table 1 and Table 2 show the Micro-F1 score of each single and ensemble model on the dev set, based on the two following approaches. In the first approach, the model using BERT as its backbone achieved the highest performance. Our ensemble method helped improve the performance, with the highest score by implementing following the first way. In the second approach, the highest perfor-

mance was achieved by DeBERTa-V3, which incorporated dependency parsing as a feature and added a CRF layer on top (the model named DeBERTaV3+Dep+CRF).

Model	Micro-F1 (%)
RoBERTa+Transition-based Parser	90.9851
BERT+Transition-based Parser	88.8245
Ensemble model with equal weights	90.3567
Ensemble model with dynamic weights	**90.9873**

Table 3: Evaluation on the private test set in the first approach

Model	Micro-F1 (%)
DeBERTaV3+Dep+CRF	87.3298
Ensemble model with equal weights	86.9233

Table 4: Evaluation on the private test set in the second approach

Table 3 and Table 4 show the results of our submission to the competition evaluation on the private test set. Our highest score, 90.9873, was achieved by implementing the second way of the ensemble method with five models in the first approach. The ensemble model with the first approach outperformed the baseline model, which uses RoBERTa and incorporates a transition-based parser. However, ensemble models with the second approach do not improve the performance on the private test set. Furthermore, the models from the first approach achieved better performance than the second approach on the private test set

Table 5 shows the resulting score for the Legal NER task. Our team achieved the 2nd place ranking on the final leaderboard.

Rank	Team Name	Score
1	ResearchTeam_HCN	0.9120
2	**VTCC-NER (Ours)**	**0.9099**
3	DeepAI	0.9099
4	Jus Mundi	0.9007
5	Autohome AI	0.8833
6	uottawa.nlp23	0.8794
7	TeamUnibo	0.8743
8	DamoAI	0.8627
9	AntContentTech	0.8622
10	xixilu556	0.8611
11	PoliToHFI	0.8321
12	Ginn-Khamov	0.7265
13	TeamShakespeare	0.6670
14	UO-LouTAL	0.6489
15	Nonet	0.5532
16	Legal_try	0.5173
17	shihanmax	0.0186

Table 5: Final Leaderboard for Legal NER task

5 Conclusion

To summarize, we applied two approaches to solve the named entity recognition of SemEval 2023 Task 6, Subtask B. Specifically, we applied an ensemble method to improve performance in the first approach. In contrast, the model did not perform well on the private test in the second approach. For the future work of this study, we try to optimize the weight for each PLM in the ensemble module in order to improve the performance.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 9–18.

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.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of

deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named entity recognition and resolution in legal text*. Springer.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A dataset of german legal documents for named entity recognition. *arXiv preprint arXiv:2003.13016*.

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. *CoRR*, abs/1907.11692.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 313–323. Springer.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

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).

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pre-training transformers on indian legal text.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia. Association for Computational Linguistics.