INTISARI

PENGEMBANGAN MODEL NER BAHASA INDONESIA MELALUI DISTILASI LLM DENGAN *CHAIN-OF-THOUGHT PROMPTING*

Oleh
Dhanada Santika Putri
22/497239/PA/21407

Named Entity Recognition (NER) merupakan salah satu tugas penting dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengekstraksi entitas seperti nama orang, lokasi, dan organisasi dari dokumen teks tidak terstruktur. Namun, pengembangan model NER bahasa Indonesia yang optimal masih menghadapi kendala terkait keterbatasan data berlabel. Di sisi lain, proses pengembangan dataset NER dengan anotasi manual memerlukan alokasi waktu, biaya, serta sumber daya manusia yang substansial, sehingga tidak efisien secara operasional. Sebagai upaya untuk menjawab permasalahan tersebut, penelitian ini mengusulkan pendekatan *knowledge distillation* berbasis *Large Language Model* (LLM) untuk meningkatkan kinerja model NER tanpa ketergantungan penuh pada data berlabel manual. Dengan memanfaatkan LLM Gemini 2.0 Flash dan teknik *Chain-of-Thought* (CoT) *few-shot prompting*, dilakukan anotasi otomatis terhadap data mentah dari dataset IDNER-News-2K. Hasil anotasi ini digunakan untuk melatih dua arsitektur *student model*, yaitu BiLSTM-CRF dan IndoBERT-lite. Penelitian ini membandingkan empat strategi pelatihan guna mengidentifikasi pendekatan paling efektif dalam pemanfaatan data hasil distilasi. Hasil dari penelitian menunjukkan bahwa strategi pelatihan bertahap "*Simple Mix*" memberikan performa terbaik pada kedua arsitektur model dan secara konsisten melampaui strategi lainnya. Temuan ini menegaskan potensi *knowledge distillation* berbasis *Large Language Model* dengan dukungan teknik *Chain-of-Thought (CoT) prompting* sebagai solusi efektif untuk mengatasi keterbatasan data berlabel dalam pengembangan model NER bahasa Indonesia.

Kata kunci: *Knowledge Distillation*, *Named Entity Recognition*, *Large Language Model*, *Chain-of-Thought Prompting*, BiLSTM-CRF, IndoBERT-lite.

ABSTRACT

INDONESIAN NER MODEL DEVELOPMENT THROUGH LLM DISTILLATION WITH CHAIN-OF-THOUGHT PROMPTING

by
Dhanada Santika Putri
22/497239/PA/21407

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) aimed at extracting entities such as person names, locations, and organizations from unstructured text documents. However, achieving optimal performance in Indonesian NER models remains challenging due to the limited availability of labeled data. On the other hand, the process of developing NER datasets through manual annotation process requires a substantial allocation of time, cost, and human resources, making it operationally inefficient. To address this challenge, this study proposes a knowledge distillation approach based on Large Language Models (LLM) to improve NER performance without relying heavily on manually labeled data. By utilizing the Gemini 2.0 Flash LLM and Chain-of-Thought (CoT) few-shot prompting, automatic annotation is performed on raw data from the IDNER-News-2K dataset. The LLM-annotated data are used to train two student model architectures, BiLSTM-CRF and IndoBERT-lite. This study compares four training strategies to determine the most effective way to utilize distilled data. The results of the study indicate that the "Simple Mix" progressive training strategy achieved the best performance across both model architectures and consistently outperformed the other strategies. These findings highlight the potential of Large Language Model-based knowledge distillation, supported by Chain-of-Thought (CoT) prompting, as an effective solution to overcome the limitation of labeled data in the development of Indonesian NER models.

Keywords: Knowledge Distillation, Named Entity Recognition, Large Language Model, Chain-of-Thought Prompting, BiLSTM-CRF, IndoBERT-lite.