Cross-Lingual Transfer for Semantic Role Labeling in Indonesian

Bariza Haqi* & Masayu Leylia Khodra

Informatics Magister Program, School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jalan Ganesa No. 10 Bandung, 40132, Indonesia
*E-mail: haqibariza@gmail.com

Abstract. Semantic role labeling is a semantic analysis task that aims to identify the semantic relationships within a sentence, such as who did what to whom, where, when, and so on. Current semantic role labeling (SRL) models for the Indonesian language still face challenges in achieving strong performance due to the limited availability of annotated corpora, especially compared with English SRL models. Therefore, this paper develops an Indonesian SRL model using cross-lingual transfer. This approach addresses the data scarcity problem in Indonesian SRL by leveraging the availability of annotated English-language corpora. The method uses multilingual models and SRL datasets from both English and Indonesian. The multilingual models used in this study are XLM-R and mT5, both in base and large configurations. The datasets include Universal PropBank Indonesia and Gojali's dataset for Indonesian, and CoNLL-2012 for English. Evaluation was conducted using test data from Universal PropBank Indonesia and Gojali's dataset. Among all developed models, XLM-R large with cross-lingual transfer achieved the best performance, with an F1 score of 0.916 on Gojali's dataset and 0.858 on the combined Universal PropBank Indonesia and Gojali datasets.

Keywords: *cross-lingual transfer; multilingual language model; semantic role labeling; low-resource NLP; annotated corpora; CoNLL-2012.*

1 Introduction

Semantic role labeling (SRL) is an approach in semantic analysis that aims to identify semantic relationships in sentences, such as who did what to whom, where, when, etc. [1]. SRL assigns roles or labels to words or phrases such as 'agent', 'experiencer', and 'instrument' [2]. SRL has been shown to be useful in downstream NLP tasks such as information extraction [3], natural language inference [4], and summarization [5].

One notable Indonesian SRL model was developed by Gojali [6], which uses a span-based SRL model based on the architecture introduced by Li et al. [7], which uses an LSTM-based approach for span representation and achieved an F1 score of 0.798.

Received June 27th, 2025, Revised April 30th, 2026, Accepted for publication June 17th, 2026.

However, Indonesian SRL resources remain limited, with fewer than 17,000 annotated instances available across existing datasets. In contrast, high-resource languages such as English provide substantially larger datasets, such as CoNLL-2012 [8] with over 315,000 annotated entries. This data disparity poses a significant challenge for developing high-performance SRL models for the Indonesian language.

Since SRL models are typically trained using supervised learning, the availability of large-scale annotated corpora has a significant impact on model performance. Prior work has shown that even state-of-the-art SRL models struggle when trained on limited data [9], while constructing annotated corpora is costly and time-consuming.

To address this challenge, several approaches have been proposed, among which cross-lingual transfer has shown promising results. In particular, Oliveira et al. [10] demonstrated that transferring knowledge from high-resource languages using multilingual transformer models can significantly improve SRL performance in low-resource settings.

Cross-lingual transfer leverages transfer learning by fine-tuning a multilingual model on data from a high-resource language and adapting it to a low-resource language. This approach is effective for SRL, where multilingual transformer models such as mBERT [11] and XLM-R [12] are trained on English data and then adapted to target languages. Oliveira et al. [10] demonstrated that this approach improves SRL performance in low-resource settings, with XLM-R outperforming mBERT. However, the effectiveness of cross-lingual transfer for Indonesian SRL has not been systematically investigated, and it remains unclear which multilingual transformer models are most suitable for this task.

This paper investigates cross-lingual transfer for Indonesian SRL using multilingual transformer models, including XLM-R and mT5 [13]. The models are first fine-tuned on English SRL data from CoNLL-2012 and subsequently adapted to Indonesian using Universal PropBank Indonesia [14] and Gojali's dataset. This approach aims to address data scarcity and improve SRL performance in low-resource settings.

The main contributions of this paper are as follows: (1) we adapt a cross-lingual transfer framework for Indonesian SRL; (2) we provide a comparative evaluation of multilingual transformer models under a unified experimental setup; and (3) we demonstrate that cross-lingual transfer improves SRL performance for Indonesian.

2 Proposed Method

Transformer-based architectures were employed in this study due to their strong performance in semantic role labeling and their ability to capture contextual representations effectively. Prior work has shown that transformer models outperform traditional approaches such as LSTM and RNN in SRL tasks [10], [15], making them suitable for cross-lingual transfer.

The multilingual models used in this study were XLM-R and mT5. XLM-R was selected for its robust cross-lingual representation capabilities and strong performance in multilingual settings [10]. In contrast, mT5 was included to evaluate an encoder-decoder multilingual transformer model adapted to the token-level classification framework used in this study.

As a baseline, the monolingual IndoBERT model was used, which was fine-tuned solely on Indonesian data. IndoBERT was chosen due to its strong performance across Indonesian NLP tasks [16]. In addition, multilingual models without cross-lingual transfer were included as baseline comparisons to assess the impact of transfer learning.

The model architecture generally follows the approach proposed by Oliveira et al. [10], with the exception that Viterbi decoding was not employed in the present study. Instead, the task was formulated as token-level classification, in which BIO labels were predicted directly from the model outputs without an additional structured decoding step. This design choice was made to preserve a consistent implementation setting across all evaluated models and to ensure compatibility with the token-level BIO label evaluation framework adopted in this work. The effect of structured decoding on performance was therefore not investigated separately.

Three datasets were used in this study: CoNLL-2012 as the English source dataset, and Universal PropBank Indonesia and Gojali's dataset as the Indonesian target datasets. CoNLL-2012 was selected because of its large scale and its use in prior cross-lingual SRL studies [8]. CoNLL-2012 was derived from OntoNotes 5.0 and is therefore subject to OntoNotes/LDC licensing requirements. Universal PropBank Indonesia and Gojali's dataset were used according to their respective usage conditions. As this study relied on existing annotated SRL datasets, no new private data were collected, although cross-lingual transfer may carry domain, linguistic, or annotation biases from English to Indonesian. The Indonesian datasets were used to evaluate cross-lingual transfer under both combined-dataset and Gojali-only settings.

This study adopted a span-based argument annotation approach, which is suitable for low-resource settings where dependency parsing resources may be limited. Predicate identification and disambiguation were not included; instead, predicate positions were assumed to be provided in both the training and the test data.

The cross-lingual transfer process is illustrated in Figure 1. The model was first fine-tuned on English data and subsequently adapted to Indonesian datasets. The resulting model was then used for inference to evaluate the effectiveness of the proposed approach.

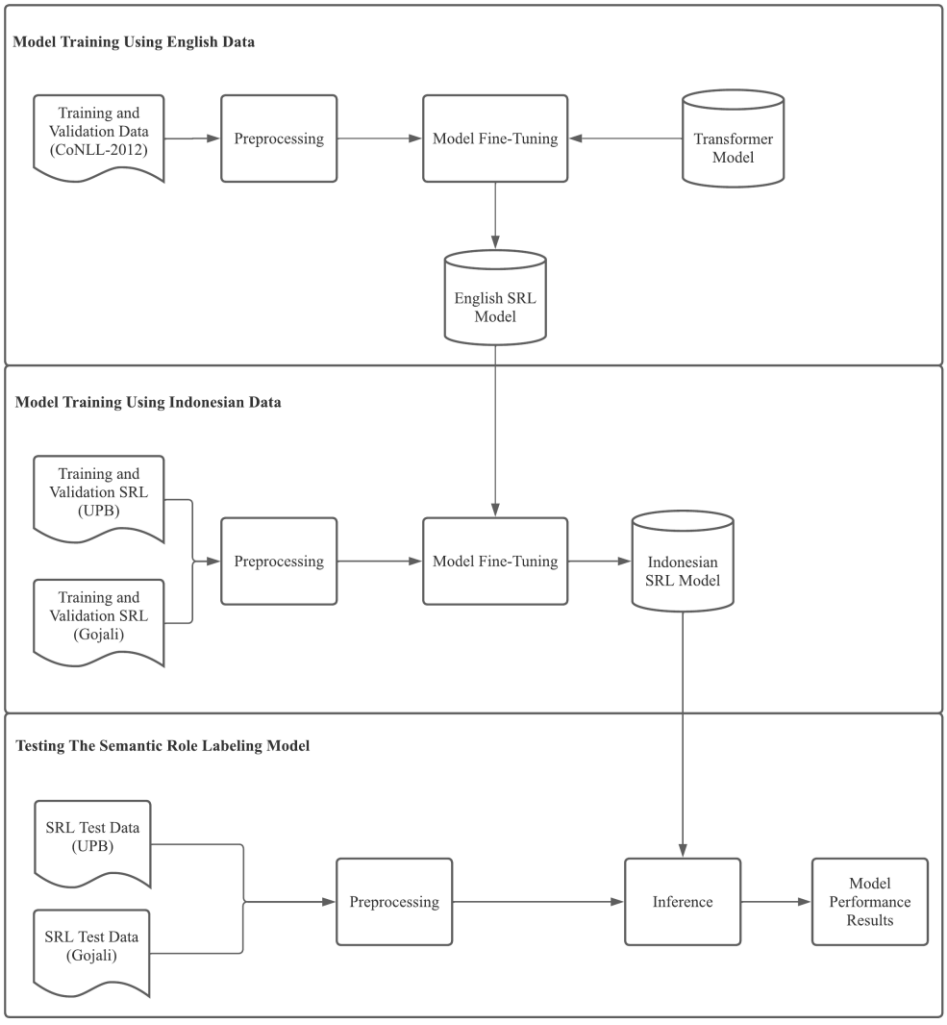

Figure 1 Process flow of the cross-lingual transfer method.

2.1 Preprocessing

Each dataset underwent a preprocessing procedure to extract SRL annotations from the original data. Since span-based argument annotation was adopted, all labels were converted into the BIO format during preprocessing.

The unified label set followed Gojali's dataset and consisted of 23 SRL labels, covering predicates, core arguments, and adjunct roles. Labels from CoNLL-2012 and Universal PropBank Indonesia were mapped to this inventory when direct equivalents existed, such as A0 to ARG0 and ARGM-TMP to AM-TMP. Labels without equivalents, such as R-ARG0 in Universal PropBank Indonesia, were converted to O rather than merged with other roles, to avoid ambiguous mappings, although this may have removed some information from the original annotations. Table 1 summarizes the label mapping used to harmonize the three datasets.

Table 1 Label mapping used for dataset harmonization.

Gojali's Data	UP Indonesia	CoNLL-2012
REL	No name	V
ARG0	A0	ARG0
ARG1	A1	ARG1
ARG2	A2	ARG2
ARG3	A3	ARG3
ARG4	A4	ARG4
AM-LOC	AM-LOC	ARGM-LOC
AM-TMP	AM-TMP	ARGM-TMP
AM-GOL	AM-GOL	ARGM-GOL
AM-PRD	AM-PRD	ARGM-PRD
AM-CAU	AM-CAU	ARGM-CAU
AM-DIR	AM-DIR	ARGM-DIR
AM-DIS	AM-DIS	ARGM-DIS
AM-EXT	AM-EXT	ARGM-EXT
AM-NEG	AM-NEG	ARGM-NEG
AM-ADV	AM-ADV	ARGM-ADV
AM-ADJ	AM-ADJ	ARGM-ADJ
AM-MOD	AM-MOD	ARGM-MOD
AM-MNR	AM-MNR	ARGM-MNR
AM-LVB	AM-LVB	ARGM-LVB
AM-COM	AM-COM	ARGM-COM
AM-REC	AM-REC	ARGM-REC
AM-PRP	AM-PRP	ARGM-PRP

The datasets were divided into training, validation, and test sets. For CoNLL-2012 and Universal PropBank Indonesia, predefined dataset splits were used. For

Gojali's dataset, the split ratio was 60:20:20 with a random seed of 43. The result of the dataset splitting is shown in Table 2.

Table 2 Comparison of the distribution of each dataset.

Data	Training	Validation	Test
CoNLL-2012	253,070	35,297	26,715
Universal PropBank Indonesia	8,128	1,065	1,008
Gojali's Data	4,001	1,334	1,334

After the data had been split, each sentence in the dataset was tokenized using the tokenizer corresponding to the selected model. As all models employ subword tokenization, each word may be split into multiple subwords. Consequently, additional labels are assigned to the resulting subwords to ensure alignment with the original word-level labels.

Predicate positions were marked using square brackets for mT5 and binary predicate indicators through token type IDs for IndoBERT and XLM-R, following each model's input mechanism. These strategies support a unified token classification framework, although their individual effects were not evaluated separately. The tokenized inputs, including input IDs and attention masks, were then padded or truncated to a fixed length.

2.2 Model Fine-Tuning

All models follow the same task-specific prediction framework, consisting of a pre-trained transformer-based component, a linear classification layer, and a softmax function, as illustrated in Figure 2.

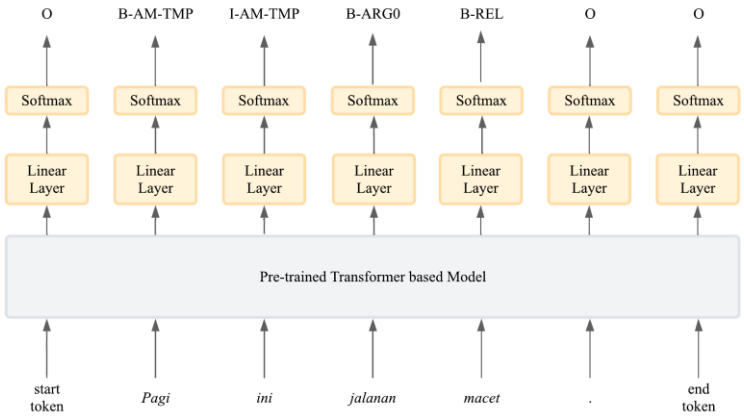

Figure 2 Model architecture.

Figure 3 illustrates the model prediction flow. In preprocessing, each sentence is processed using the corresponding Transformer tokenizer to produce word or subword input IDs. Labels are then adjusted for token–label alignment: special tokens receive the O label, while subsequent subwords or punctuation inherit the corresponding I label. For reference, the Indonesian input used in Figures 2 and 3, "*Jalanan macet pagi ini*", means "The roads are congested this morning."

In the modelling stage, the input IDs are fed into the transformer model, which produces logits, raw prediction scores for each label. During postprocessing, these logits are passed through a softmax function to select the label with the highest probability, resulting in the final SRL labels.

In the mT5 model, only the encoder component is used to align its implementation with the token-level classification framework adopted for the other evaluated models. The encoder output representations are then passed to a token classification layer for BIO label prediction.

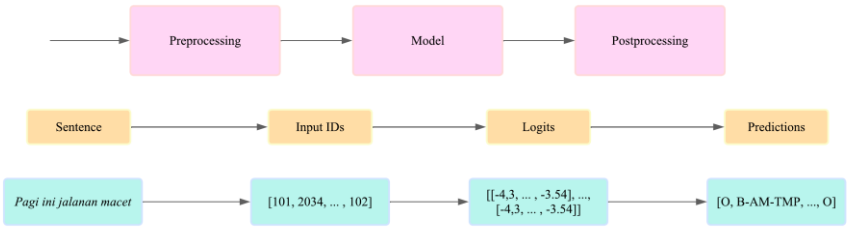

Figure 3 Process flow of the model.

2.3 Inference

This stage evaluated model performance on the test data using micro-averaged F1 as the primary evaluation metric. The evaluation was performed at the token level by comparing each predicted BIO label with its corresponding gold BIO label across all tokens. Following the evaluation setting used in the experiments, this paper reports micro-F1 as the primary metric. In addition to the overall F1 score, per-label F1 scores are reported to offer a more detailed analysis of performance across semantic role categories.

Due to limitations in computational resources, each experimental setting was conducted using a single training run. Therefore, the reported results should be regarded as single-run measurements, and the present study did not assess performance variability across multiple configurations.

3 Performance Evaluation Results and Discussion

Two evaluation settings were conducted. The first involved training and testing all models using the combined dataset of Gojali and Universal PropBank Indonesia. The second involved training and testing the model configuration that obtained the highest observed F1 score in the first evaluation, namely XLM-R large with cross-lingual transfer, using only Gojali's dataset.

The experiments were implemented in Python 3 using PyTorch v1.13 and the Hugging Face Transformers library. The main hyperparameters were a learning rate of 0.0001, a maximum length of 256, and the Adam optimizer. Indonesian fine-tuning used 8 epochs with a batch size of 32, while English fine-tuning used 4 epochs with a batch size of 64. Other training parameters followed the default settings. Training was conducted on Tesla V100 GPUs with 32 GB of memory. Configurations requiring more memory were trained using two Tesla V100 GPUs

In the first evaluation setting, cross-lingual transfer produced consistent improvements across all evaluated multilingual models, as shown in Tables 3 and 4. The effect was particularly evident for the base models, which benefited more substantially from English pre-tuning. Among these, the mT5 base showed the largest improvement, increasing from 0.768 to 0.831, while the XLM-R base improved from 0.839 to 0.855. For the large models, the gains were more modest: XLM-R large improved from 0.850 to 0.858, and mT5 large improved from 0.826 to 0.844. These results suggest that cross-lingual transfer is especially beneficial for model configurations with lower initial performance. Although the gain for XLM-R large over IndoBERT large was small, XLM-R large with cross-lingual transfer achieved the highest overall F1 score of 0.858 under the evaluation setting adopted in this study.

Overall, IndoBERT, XLM-R, and mT5 all achieved competitive performance on Indonesian semantic role labeling. Although mT5 uses a different predicate-marking strategy from IndoBERT and XLM-R, its results indicate that the model remains effective within the unified token classification framework used in this study.

Table 3 Test set F1 scores of baseline models.

Model	F1 Score
IndoBERT *base*	0.826
IndoBERT *large*	**0.856**
XLM-R *base*	0.839
XLM-R large	0.850
mT5 *base*	0.768
mT5 *large*	0.826

Table 4 Test set F1 scores of cross-lingual transfer models.

Model	F1 Score
XLM-R *base cross-lingual transfer*	0.855
XLM-R *large cross-lingual transfer*	**0.858**
mT5 *base cross-lingual transfer*	0.831
mT5 *large cross-lingual transfer*	0.844

The second evaluation further examined the XLM-R large model with cross-lingual transfer, as this configuration achieved the highest overall F1 score in the first evaluation setting. In this experiment, the model was trained and evaluated using only Gojali's dataset. As shown in Table 5, the model achieved an overall F1 score of 0.916 on the test data, which is higher than its performance in the combined-dataset setting using Gojali and Universal PropBank Indonesia. This result suggests that differences in annotation coverage, span-labeling consistency, and label distribution between the two Indonesian datasets may affect model performance when the datasets are combined.

Table 5 Per-label F1 scores of XLM-R large with cross-lingual transfer on Gojali's dataset.

Label	Support (n)	F1 Score
REL	1,336	1
ARG0	2,335	0.930
ARG1	5,025	0.936
ARG2	1,201	0.848
ARG3	185	0.848
ARG4	88	0.910
AM-LOC	724	0.876
AM-TMP	1,025	0.900
AM-GOL	34	0.745
AM-PRD	5	0.667
AM-CAU	282	0.897
AM-DIR	20	0.635
AM-DIS	137	0.843
AM-EXT	56	0.743
AM-NEG	86	0.946
AM-ADV	308	0.809
AM-ADJ	5	0.000
AM-MOD	131	0.956
AM-MNR	47	0.774
AM-LVB	53	0.928
AM-COM	67	0.786
AM-REC	11	0.933
AM-PRP	272	0.870
Overall	13,433	0.916

As shown in Table 6, Universal PropBank Indonesia contains a much larger number of O labels than non-O labels, while Gojali's dataset contains more non-

O labels than O labels. This indicates that Universal PropBank Indonesia has sparser argument annotation, whereas Gojali's dataset provides denser SRL labeling. The difference in labeled-token distribution may introduce inconsistency during training, and it may partly explain why the model performs better when trained and evaluated only on Gojali's dataset.

Table 6 Distribution of O and non-O labels in the Indonesian datasets.

Data	Number of O labels	Number of non-O labels
UP Indonesia	191,153	86,036
Gojali's Data	34,530	66,186

At the label level, the model achieved high F1 scores for roles such as REL, ARG0, ARG1, ARG4, AM-NEG, AM-MOD, AM-LVB, and AM-REC, but performance remained uneven across labels. The lowest result was observed for AM-ADJ, with an F1 score of 0.000. As shown in Table 7, AM-ADJ is rare in both Indonesian datasets, with only 8 instances in Universal PropBank Indonesia and 19 in Gojali's dataset. However, rarity alone does not explain performance, since AM-REC also has a few instances but achieves a high F1 score. This suggests that label ambiguity and linguistic cues also affect performance.

Table 7 Label distribution across the SRL datasets.

Label	UP Indonesia	Gojali's Data	CoNLL-2012
REL	10,188	6,715	315,082
ARG0	12,483	10,712	437,576
ARG1	31,368	24,409	1,437,554
ARG2	16,095	6,246	483,333
ARG3	306	694	28,285
ARG4	218	434	21,665
AM-LOC	2,915	3,398	82,380
AM-TMP	6,979	5,523	178,115
AM-GOL	29	194	5,608
AM-PRD	125	33	28,011
AM-CAU	506	1,707	46,245
AM-DIR	325	158	12,873
AM-DIS	244	717	33,369
AM-EXT	26	211	4,845
AM-NEG	270	465	14,697
AM-ADV	923	1,584	153,083
AM-ADJ	8	19	8,777
AM-MOD	755	673	27,924
AM-MNR	758	328	66,678
AM-LVB	1	259	627
AM-COM	103	302	1,982
AM-REC	0	56	185
AM-PRP	1,411	1,349	28,011
Total	86,036	66,186	3,416,905

Table 8 shows an AM-ADJ error case in which the phrase "*atas kemauan sendiri*" should be labeled as AM-ADJ but was predicted as AM-CAU. This indicates that AM-ADJ may be difficult to distinguish from other adjunct roles, highlighting the remaining effects of class imbalance and label ambiguity in Indonesian SRL.

Table 8 Error case for the AM-ADJ label.

Sentence	Relevant Span	Gold Label	Predicted Label
"*Mereka kembali ke Jayapura atas kemauan sendiri, setelah melihat situasi keamanan di Jayapura semakin stabil.*"	"*atas kemauan sendiri*"	B-AM-ADJ, I-AM-ADJ, I-AM-ADJ	B-AM-CAU, I-AM-CAU, I-AM-CAU

4 Conclusion and Future Work

This paper investigated cross-lingual transfer for Indonesian SRL using multilingual transformer models. The results show that using English CoNLL-2012 data improved performance across the evaluated models, with XLM-R large achieving the best result. However, the findings should be interpreted with consideration of the potential domain mismatch between the English and Indonesian datasets.

Despite these improvements, challenges remain in Indonesian SRL. Cross-lingual transfer improves performance across the evaluated multilingual models, but the lower result in the combined-dataset setting shows that combining Indonesian SRL datasets does not always improve performance. This may be related to differences in annotation coverage and O/non-O label distribution between Universal PropBank Indonesia and Gojali's dataset. Span-labeling differences may also contribute, but further dataset-level analysis is needed to confirm this. The model also struggles with rare and ambiguous roles such as AM-ADJ, which were not predicted correctly in the test set. Future work should focus on improving annotation alignment and addressing class imbalance for rare semantic roles.

Acknowledgement

This research was funded by Hibah P2MI ITB (Project ID: STEI.PPMI-1-58-2026).

References

[1] Villodre, L.M., Carreras, X., Litkowski, K.C. & Stevenson, S., *Semantic Role Labeling,* An introduction. Computational Linguistics, 34, pp.145-159, 2008.

[2] Jurafsky, D. & Martin, J.H., *Speech and Language Processing* (3rd ed., draft), 2003.
[3] Niklaus, C., Cetto, M., Freitas, A. & Handschuh, S., *A Survey on Open Information Extraction,* In Proceedings of COLING, pp. 3866-3878, 2018.
[4] Zhang, Z., Wu, Y., Hai, Z., Li, Z., Zhang, S., Zhou, X. & Zhou, X., *Semantics-aware BERT for Language Understanding*, In Proceedings of AAAI, 2019.
[5] Gunawan, Y.H.B. & Khodra, M.L., *Multi-document Summarization using Semantic Role Labeling,* In Proceedings of ICAICTA pp. 1-6, 2020. doi: 10.1109/ICAICTA49861.2020.9429055.
[6] Gojali, F., *Span-based Semantic Role Labeling for Indonesian News Summarization (Unpublished Undergraduate Thesis),* Institut Teknologi Bandung, 2022.
[7] Li, Z., He, S., Zhao, H., Zhang, Y., Zhang, Z., Zhou, X. & Zhou, X., *End-to-end Semantic Role Labeling,* In Proceedings of AAAI, pp. 6730-6737, 2019. doi:10.1609/aaai.v33i01.33016730.
[8] Pradhan, S., Moschitti, A., Xue, N., Uryupina, O. & Zhang, Y., *CoNLL-2012 Shared Task,* Modeling Multilingual Coreference, in Proceedings of CoNLL. pp. 1-40, 2012.
[9] Li, T., Jawale, P.A., Palmer, M. & Srikumar, V., *Structured Tuning for Semantic Role Labeling,* In Proceedings of ACL, pp. 8402-8412. 2020.
[10] Oliveira, S., Loureiro, D., & Jorge, A., *Improving Portuguese SRL with Transformers,* In Proceedings of DSAA, 2021. doi:10.1109/DSAA53316.2021.9564238 .
[11] Devlin, J., Chang, M.W., Lee, K. & Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers,* in Proceedings of NAACL-HLT, 2019.
[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V., *Cross-lingual Representation Learning at Scale*, in Proceedings of ACL, pp. 8440-8451, 2020.
[13] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. & Raffel, C., *mT5: A Multilingual Text-to-text Transformer*, in Proceedings of NAACL-HLT, pp. 483-498, 2021.
[14] Jindal, I., Rademaker, A., Ulewicz, M., Linh, H., Nguyen, H., Tran, K.N., Zhu, H. & Li, Y., *Universal Proposition Bank 2.0,* in Proceedings of LREC, pp. 1700-1711, 2022.
[15] Shi, P. & Lin, J.J., *Simple BERT Models for SRL,* arXiv preprint arXiv,1904,05255, 2019.
[16] Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z.Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S. & Purwarianti, A., *IndoNLU: Benchmark for Indonesian NLP,* in Proceedings of AACL-IJCNLP, pp. 843-857, 2020.