Label-Aware Automatic Verbalizer for Few-Shot Text Classification in Mid-To-Low Resource Languages

By
Thanakorn Thaminkaew[1]
Piyawat Lertvittayakumjorn[2]
Peerapon Vateekul[1]

[1]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
[2]Google, United States

6472031921@student.chula.ac.th, piyawat@google.com, peerapon.v@chula.ac.th

Outline

1. Introduction & Motivation
2. Proposed Method: Label-Aware Automatic Verbalizer (LAAV)
3. Experiments & Baseline Details
4. Results and Additional Analyses
5. Conclusion

1. Introduction: Methods for Text Classification

Traditional Full-Finetune

Input Text: A great movie!!!

[CLS] A great movie!!!

Pre-trained Model (RoBERTa-base)

Fine-Tune

[CLS]
last hidden layer

Classifier
logistic regression

Fine-Tune

1: Positive
0: Negative

Limitations

- Require a lot of label data
- Require high computational power (GPUs) / Larger LM

Prompt-Based Learning

Input Text: A great movie!!!

A great movie!!! It was **[MASK].**

Pre-trained Model (RoBERTa-base)

Fine-Tune

Great: Positive
Terrible: Negative

Advantages

- No additional parameter
  - Less label data
  - Less computational power / Smaller LM

1. Introduction: Prompt-Based Learning for Text Classification

Prompt-Based Learning

Input Text: A great movie!!!

A great movie!!! It was **[MASK]**

Template

Pre-trained Model
(RoBERTa-base)

Fine-Tune

Great: Positive ✓
Terrible: Negative

Verbalizer

Different Templates

SNLI (entailment/neutral/contradiction)		mean (std)
<S_1> ? [MASK] , <S_2>	Yes/Maybe/No	**77.2 (3.7)**
<S_1> . [MASK] , <S_2>	Yes/Maybe/No	76.2 (3.3)
<S_1> ? [MASK] <S_2>	Yes/Maybe/No	74.9 (3.0)
<S_1> <S_2> [MASK]	Yes/Maybe/No	65.8 (2.4)
<S_2> ? [MASK] , <S_1>	Yes/Maybe/No	62.9 (4.1)
<S_1> ? [MASK] , <S_2>	Maybe/No/Yes	60.6 (4.8)
Fine-tuning	-	48.4 (4.8)

Different Verbalizers

Template	**Label words**	**Accuracy**
SST-2 (positive/negative)		mean (std)
<S_1> It was [MASK] .	great/terrible	**92.7 (0.9)**
<S_1> It was [MASK] .	good/bad	92.5 (1.0)
<S_1> It was [MASK] .	cat/dog	91.5 (1.4)
<S_1> It was [MASK] .	dog/cat	86.2 (5.4)
<S_1> It was [MASK] .	terrible/great	83.2 (6.9)
Fine-tuning	-	81.4 (3.8)

1. Motivation

Prompt-Based Learning

Input Text: A great movie!!!

A great movie!!! It was **[MASK]**
Template

Pre-trained Model
(RoBERTa-base)
Fine-Tune

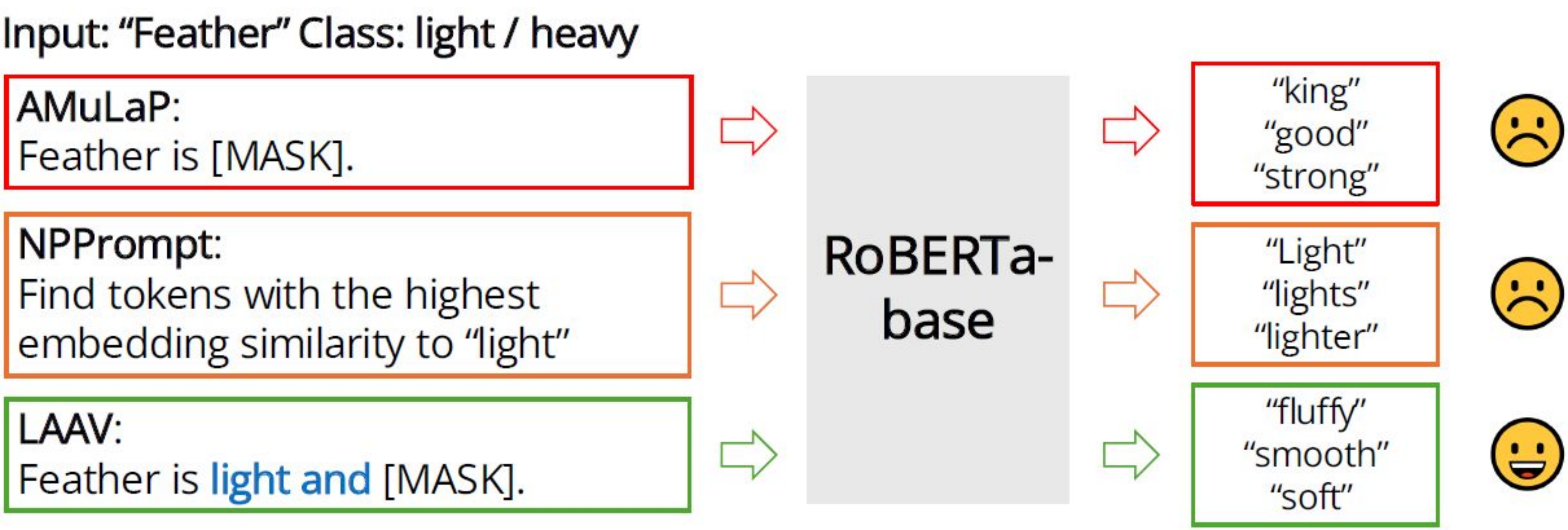
Great: Positive ✓
Terrible: Negative
Verbalizer

Automatically selecting words from the language model for the verbalizer does not ensure relevance to the classes of interest.

For mid-to-low resource languages, language models may have received less comprehensive training data. This, combined with the previously mentioned issue, can negatively impact classification accuracy.

2. Proposed Method: Label-Aware Automatic Verbalizer (LAAV)

- ✓ An effective verbalizer by adding a class label and the conjunction “and”
- ✓ This approach leverages class labels to prompt the model to generate more relevant words, which is crucial for mid-to-low resource languages.



2. Proposed Method: Label-Aware Automatic Verbalizer (LAAV)

Explanation

Objective: Classify input text x to a class $y \in Y$ using a pre-trained language model.

Verbalizer Construction:

- Each class y_i is represented by a set of k tokens, denoted as $S(y_i)$.
- Tokens are selected from the sub-word vocabulary V_M of the language model M.
- Apply a LAAV template to training examples x with ground truth label y_i

$$T_{y_i}(x) = [x]\ [prompt][y_i]\ and\ [MASK]$$

- Next, let M predict the probability of each $v \in V_M$ for the $[MASK]$ of $T_{y_i}(x)$
- Given D as the training dataset and p_M as the probability predicted by M, the token score of v for class y_i is

$$s(v, y_i) = \sum_{(x,y_i)\in D} p_M([MASK] = v \mid T_{y_i}(x))$$

- Define $S(y_i)$ as a set of k tokens with the highest $s(v, y_i)$. Ensure that each token v is assigned to only one class:

$$y_i = argmax_{y\in Y} s(v, y_i)$$

Fine-Tuning:

- The log-probability of class y_i for an input x:

$$L(y_i|x) = \frac{1}{k} \sum_{v\in S(y_i)} log p_M([MASK] = v|T_{y_i}(x))$$

- Fine-tune the language model on D using cross-entropy loss.

$$loss = -\sum_{(x,y)\in D} \sum_{y_i\in Y} I(y, y_i)\cdot L(y|x)$$

where $I(y, y_i) = 1$ if $y = y_i$; otherwise, 0.

Prediction:

- During validation and testing, the predicted label $\hat{y}$ for an input x is

$$\hat{y} = argmax_{y_i\in Y} L(y_i|x)$$

3. Experiments: Datasets and Pre-trained Models

Dataset (language)	Pre-trained language model	Labels and verbalizer template	
SmSA [1] (Indonesian)	IndoBERT [5]	Labels	[negatif, netral, positif] => [negative, neutral, positive]
		LAAV Template	" komentar ini adalah + [y]+ "dan" + [MASK]."
		AMuLaP / Training Template	" komentar ini adalah [MASK]."
Shopee Reviews [2] (Tagalog)	Tagalog RoBERTa [6]	Labels	[napakasama, masama, karaniwan, mahusay, napakahusay] => [very bad, bad, average, good, excellent]
		LAAV Template	" ito ay + [y] + "at" + <mask> reivew."
		AMuLaP / Training Template	" ito ay <mask> reivew."
Wisesight Sentiment [3] (Thai)	WangchanBERTa [7]	Labels	[ลบ, กลาง, บวก, คำถาม] => [negative, neutral, positive, question]
		LAAV Template	"เป็นความเห็นเชิง + [y] + "และ" + <mask>"
		AMuLaP / Training Template	"เป็นความเห็นเชิง<mask>"
Students' Feedback [4] (Vietnamese)	PhoBERT [8]	Labels	[tiêu cực, trung lập, tích cực] => [negative, neutral, positive]
		LAAV Template	" Nó là + [y] + "và" + <mask>."
		AMuLaP / Training Template	" Nó là <mask>."

3. Experiments: Implementation Details

- Randomly selected 1, 2, 4, or 8 samples per class for training and validation.
- Repeated the process 5 times with different seeds for robustness.
- Used the Adam optimizer [9] with a learning rate of 1e-5 and employed early stopping with a 100-epoch limit.
- Set the Number of Representative Tokens (k) to 32, as determined by the experiment on the right.

Sample Size	1	2	4	8
SmSA (Indonesian)				
1	41.7 (2.1)	40.9 (6.5)	59.9 (10.0)	58.6 (5.6)
4	44.2 (7.4)	46.6 (11.2)	58.0 (8.8)	58.9 (6.9)
8	41.1 (10.3)	45.8 (6.6)	59.4 (9.3)	55.9 (10.5)
16	41.9 (11.5)	43.9 (8.5)	61.0 (6.7)	57.6 (10.0)
24	44.2 (10.3)	46.3 (4.9)	**61.1 (6.0)**	**59.1 (7.6)**
32	**45.3 (9.9)**	**46.7 (4.7)**	**61.1 (7.6)**	58.5 (10.9)
40	45.2 (9.3)	**46.7 (4.1)**	60.9 (7.3)	58.3 (12.0)
Shopee Reviews (Tagalog)				
1	19.1 (2.1)	26.6 (1.1)	25.7 (4.7)	30.6 (3.1)
4	22.9 (3.6)	26.6 (4.2)	29.4 (2.8)	32.8 (2.0)
8	24.9 (3.5)	28.9 (2.2)	30.7 (3.7)	32.9 (2.6)
16	24.7 (3.0)	29.4 (2.6)	31.4 (3.5)	**33.3 (2.2)**
24	24.8 (5.1)	29.7 (2.4)	31.1 (3.4)	33.0 (2.4)
32	**25.5 (5.0)**	**30.5 (1.3)**	**31.6 (3.7)**	32.6 (2.8)
40	23.1 (6.8)	30.2 (1.2)	31.5 (3.5)	32.0 (3.1)
Wisesight sentiment (Thai)				
1	23.0 (4.9)	29.8 (7.4)	32.9 (5.8)	38.1 (4.4)
4	25.8 (3.3)	29.9 (8.8)	34.4 (5.5)	42.0 (3.7)
8	25.7 (4.3)	34.1 (7.8)	37.5 (4.5)	41.3 (5.5)
16	**25.9 (4.8)**	33.9 (6.0)	36.4 (6.0)	40.1 (5.6)
24	24.3 (5.2)	**34.3 (5.0)**	35.1 (4.7)	41.9 (6.1)
32	**25.9 (5.9)**	31.5 (7.6)	**38.1 (4.5)**	**42.1 (5.8)**
40	**25.9 (5.8)**	34.2 (7.4)	38.0 (5.6)	37.4 (9.0)
Students' Feedback (Vietnamese)				
1	39.7 (10.5)	50.7 (8.5)	64.1 (4.2)	64.7 (3.4)
4	50.4 (11.5)	55.0 (4.4)	64.3 (0.9)	68.4 (3.9)
8	47.8 (11.5)	60.0 (3.8)	65.6 (3.0)	68.6 (2.7)
16	49.2 (12.5)	60.0 (4.5)	67.0 (3.3)	68.8 (2.4)
24	50.3 (11.5)	62.0 (3.4)	67.9 (3.5)	69.1 (1.7)
32	**53.6 (10.7)**	**61.7 (3.8)**	67.9 (2.8)	**69.5 (1.9)**
40	52.5 (9.2)	61.5 (2.6)	**68.1 (3.0)**	69.2 (2.1)

Table 5: Macro-F1 results along with their standard deviation in the parentheses tested on four datasets when using LAAV with a different number of tokens to represent each label varying from 1, 4, 8, 16, 24, 32, and 40. The best results are marked in **bold**.

3. Baseline Details

- **PET [10]:** Manually selecting a token to represent each class.
- **WARP$_V$ [11]:** Representing each class with a trained continuous vector.
- **PETAL [12]**: Searching for the most suitable representative token.
- **AMuLaP [13]**: Searching for multiple suitable representative tokens using an unmodified template.
- **NPPrompt [14]**: Using a set of tokens with the highest embedding similarity to the manual label as representative tokens.
- **LLM-ICL [15]**: Unlike other baselines that involve fine-tuning, we augmented the prompt template with examples for each few-shot learning scenario, enabling in-context learning (ICL).

DataMind
Data Mining Group
Machine Intelligence and Knowledge Discovery Lab
Chulalongkorn University

4. Results

Baseline Results:

- LLM-ICL: Promising in extreme few-shot settings but less effective with more examples.
- PET: The strongest baseline overall, leveraging label names as representative tokens.

Baselines vs. Proposed Method:

- **LAAV**: Consistently outperforms other baselines in almost all settings. When averaging all sample sizes across four datasets:
  - Achieves a 5.7% absolute improvement in Macro F1 scores over PET.
  - Achieves a 6.7% absolute improvement in Macro F1 scores over AMuLaP.

Sample Size	1	2	4	8
SmSA (Indonesian)				
Traditional FT	42.5 (7.1)	43.9 (3.6)	48.1 (7.4)	52.2 (6.6)
PET	34.5 (9.8)	39.8 (7.5)	49.1 (8.4)	53.0 (7.0)
WARPv	37.5 (9.1)	43.9 (5.8)	50.9 (7.2)	52.2 (5.2)
PETAL	35.5 (8.8)	44.1 (6.9)	53.8 (6.2)	52.1 (8.2)
AMuLaP	38.7 (10.4)	44.5 (4.9)	58.9 (4.6)	58.3 (4.4)
NPPrompt	22.6 (6.2)	41.7 (7.1)	50.7 (6.4)	51.6 (8.4)
LLM-ICL	**49.4 (2.4)**	**54.1 (8.0)**	50.5 (1.6)	51.9 (0.9)
LAAV (ours)	45.3 (9.9)*	46.7 (4.7)	**61.1 (7.6)***	**58.5 (10.9)***
Shopee Reviews (Tagalog)				
Traditional FT	17.3 (4.5)	21.7 (3.9)	24.4 (3.8)	28.1 (5.0)
PET	18.3 (2.4)	20.6 (1.9)	22.8 (1.2)	24.0 (1.8)
WARPv	18.6 (2.4)	23.0 (1.3)	25.1 (2.1)	28.1 (2.7)
PETAL	17.8 (4.0)	26.9 (1.5)	26.8 (3.8)	30.2 (1.6)
AMuLaP	21.4 (6.0)	27.2 (3.5)	28.9 (5.8)	32.4 (3.3)
NPPrompt	13.9 (7.0)	18.0 (6.5)	17.9 (7.4)	26.9 (5.0)
LLM-ICL	**28.1 (0.7)**	28.7 (1.4)	28.1 (1.3)	28.8 (1.2)
LAAV (ours)	25.5 (5.0)*	**30.5 (1.3)***	**31.6 (3.7)***	**32.6 (2.8)***
Wisesight sentiment (Thai)				
Traditional FT	20.7 (4.3)	24.2 (5.5)	28.2 (4.2)	29.6 (5.4)
PET	23.8 (4.4)	31.0 (7.2)	34.5 (6.5)	41.0 (5.5)
WARPv	23.4 (5.7)	27.2 (5.9)	30.8 (4.2)	37.7 (2.8)
PETAL	20.5 (2.0)	26.5 (7.6)	30.8 (4.4)	37.1 (2.8)
AMuLaP	21.1 (5.4)	28.0 (10.6)	32.3 (5.6)	37.4 (8.9)
NPPrompt	25.3 (2.3)	26.2 (9.1)	31.0 (7.8)	37.0 (4.6)
LLM-ICL	17.7 (2.0)	19.1 (1.3)	21.4 (2.6)	23.2 (1.9)
LAAV (ours)	**25.9 (5.9)**	**31.5 (7.6)**	**38.1 (4.5)**	**42.1 (5.8)**
Students' Feedback (Vietnamese)				
Traditional FT	39.5 (7.1)	47.3 (8.7)	51.2 (10.1)	62.6 (1.6)
PET	49.3 (13.3)	60.7 (2.1)	65.5 (3.0)	68.7 (2.8)
WARPv	23.3 (3.5)	47.8 (7.6)	51.4 (8.3)	57.2 (2.6)
PETAL	21.1 (9.2)	38.3 (6.8)	49.1 (8.9)	57.7 (4.3)
AMuLaP	38.7 (13.6)	47.0 (10.9)	55.6 (11.2)	64.6 (2.1)
NPPrompt	25.5 (6.1)	39.5 (11.8)	37.0 (17.4)	40.0 (17.2)
LLM-ICL	41.5 (0.7)	41.5 (0.8)	41.5 (0.9)	41.9 (1.3)
LAAV (ours)	**53.6 (10.7)**	**61.7 (3.8)**	**67.9 (2.8)***	**69.5 (1.9)**

Table 1: Macro F1 results along with their standard deviations (in parentheses) tested on four datasets. The best results are marked in **bold**. A

4. Additional Analyses: Choices of conjunction

Setup:

- Explore other potential conjunctions as we used "and" for LAAV templates.
- Utilize the AMuLaP template to identify the initial $S(y_i)$ for each class and apply the following template.

$$T^{S}_{y_i}(x) = [x][prompt][y_i][MASK][v]$$

for all $v \in S(y_i)$ to each training example x labeled y_i

- Predict tokens that effectively serve as conjunction between y_i to v.
- Used the predicted tokens, referred to as Automatic, instead of “and” in the LLAV template.

Results:

- **"and"** consistently yields the best results across datasets, validating our initial LAAV template design.

Dataset	Top Translated Words	Automatic	“and”
SmSA	exchange, dough, mopped	42.7 (8.3)	**45.3 (9.9)**
Shopee Reviews	already, in, just	20.6 (3.2)	**25.5 (5.0)**
Wisesight sentiment	really, very, yes	24.8 (3.8)	**25.9 (5.9)**
Students’ Feedback	of, for, and	43.7 (6.5)	**53.6 (10.7)**

Data Mining Group
Machine Intelligence and Knowledge Discovery Lab
Chulalongkorn University

5. Conclusion

- Our proposed method, LAAV, constructs a better verbalizer by exploiting class labels to collect more relevant words.
- Experiments show that LAAV outperforms other existing verbalizers in few-shot text classification across four languages, even surpassing LLM with in-context learning.
- Comprehensive analysis highlights "and" as an effective conjunction for retrieving high-discriminative words, enhancing text classification performance.
- Future plans include applying LAAV to multilingual LMs and multilabel classification.

Data Mind
Data Mining Group
Machine Intelligence and Knowledge Discovery Lab
Chulalongkorn University

References

[1] Wilie et al. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. AACL-IJCNLP 2020.
[2] Nguyen et al. (2018). UIT-VSFC: Vietnamese students' feedback corpus for sentiment analysis. KSE 2018.
[3] Suriyawongkul et al. (2019). Pythainlp/wisesight-sentiment: First release.
[4] Riego (2023). shopee-reviews-tl-stars. Hugging Face Datasets.
[5] Wilie et al. (2020). Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. arXiv preprint.
[6] Cruz & Cheng (2021). Improving large-scale language models and resources for filipino. arXiv preprint.
[7] Lowphansirikul et al. (2021). Wangchanberta: Pretraining transformer-based thai language models.
[8] Nguyen et al. (2023). Seallms–large language models for southeast asia. arXiv preprint.
[9] Kingma & Ba (2014). Adam: A method for stochastic optimization. arXiv preprint.
[10] Schick & Schütze (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. EACL 2021.
[11] Hambardzumyan et al. (2021). WARP: Word-level Adversarial ReProgramming. ACL-IJCNLP 2021.
[12] Schick et al. (2020). Automatically identifying words that can serve as labels for few-shot text classification. COLING 2020.
[13] Wang et al. (2022). Automatic multi-label prompting: Simple and interpretable few-shot classification. NAACL-HLT 2022.
[14] Zhao et al. (2023). Pre-trained language models can be fully zero-shot learners. ACL 2023.
[15] Brown et al. (2020). Language models are few-shot learners. NeurIPS 2020.