Text Classification with Negative Supervision

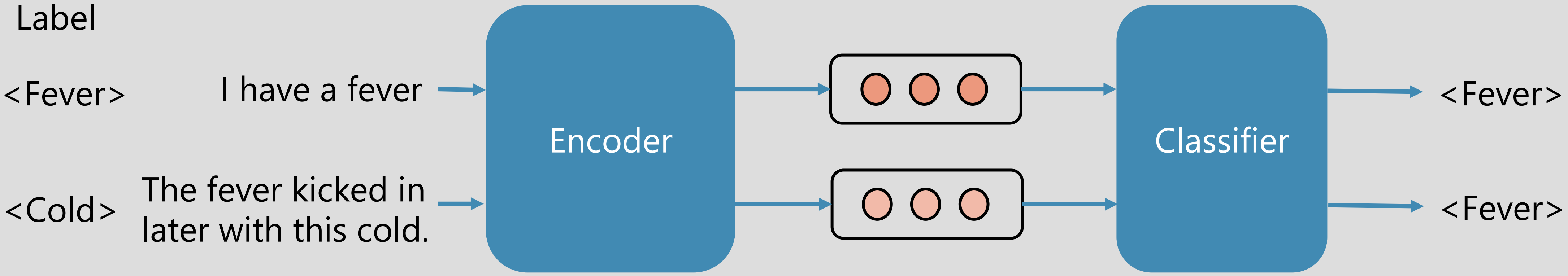
Sora Ohashi[*], Junya Takayama[*], Tomoyuki Kajiwara[**], Chenhui Chu[**], Yuki Arase[*]

[*] Graduate School of Information Science and Technology, Osaka University

[**] Institute of Datability Science, Osaka University

A Challenge in Text Classification

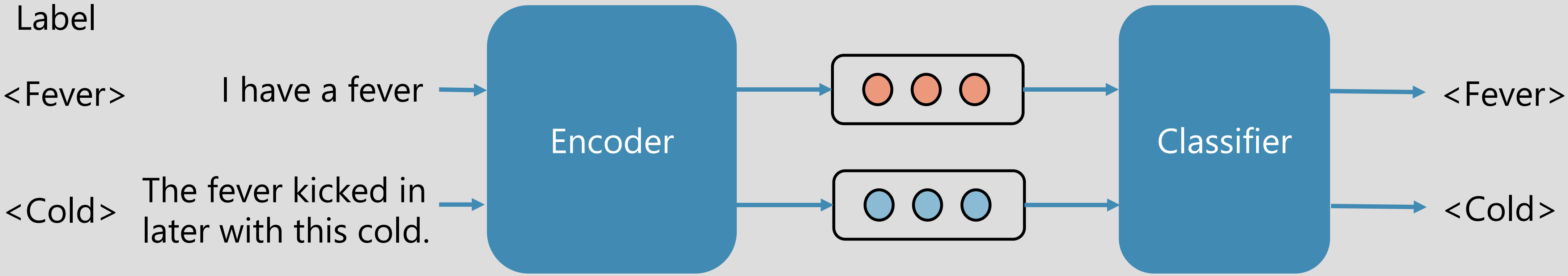
- Generally, text classification models have two components:
  - An encoder: generates a vector representation (e.g. BERT[1])
  - A classifier: predicts a label for an input (e.g. Feedforward Neural Network)
- Misclassify inputs of similar meanings into the same category
  - Generate similar representation of inputs that have similar meaning even if these have different labels



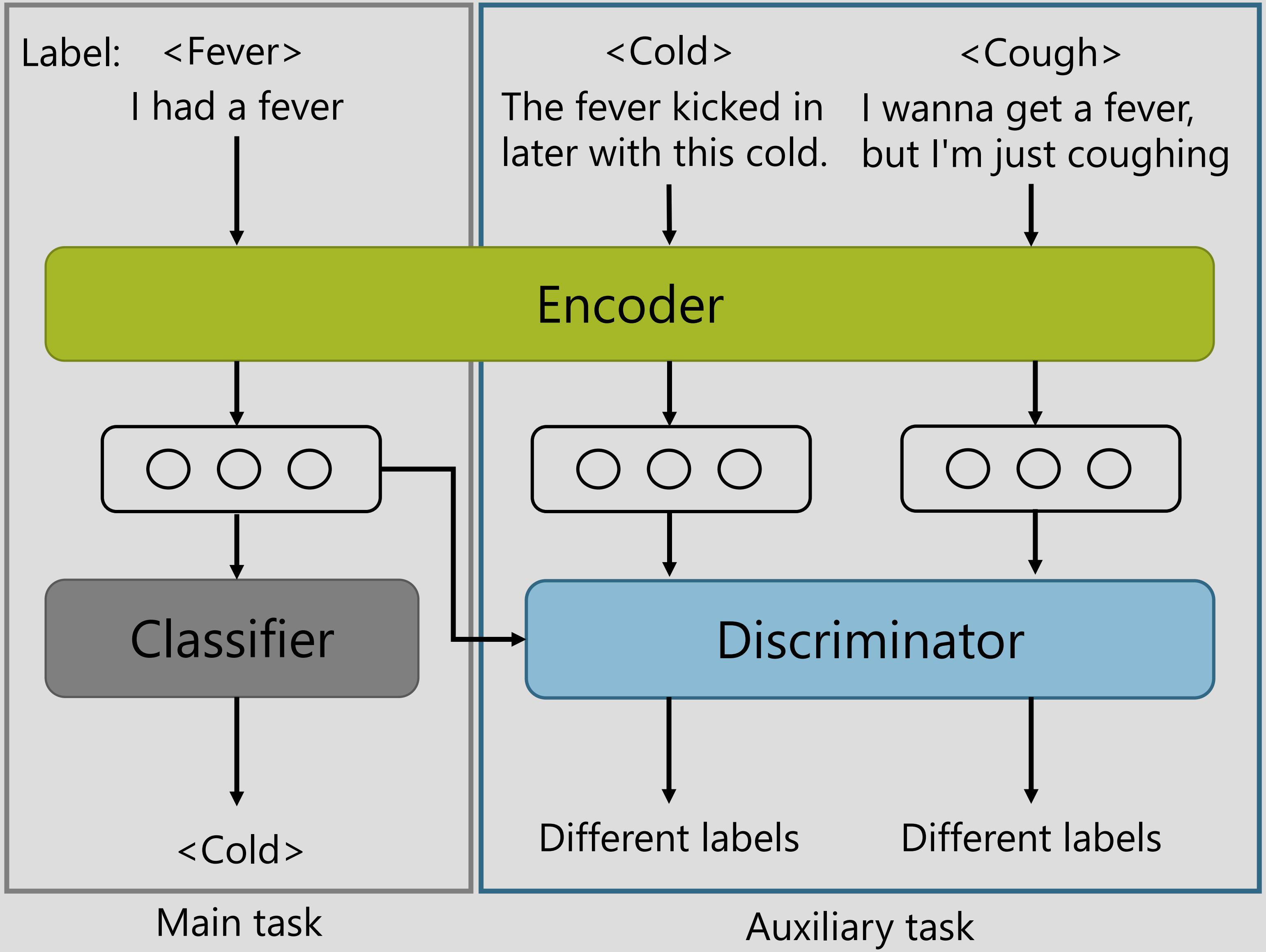
[1] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019

Approach: Negative Supervision

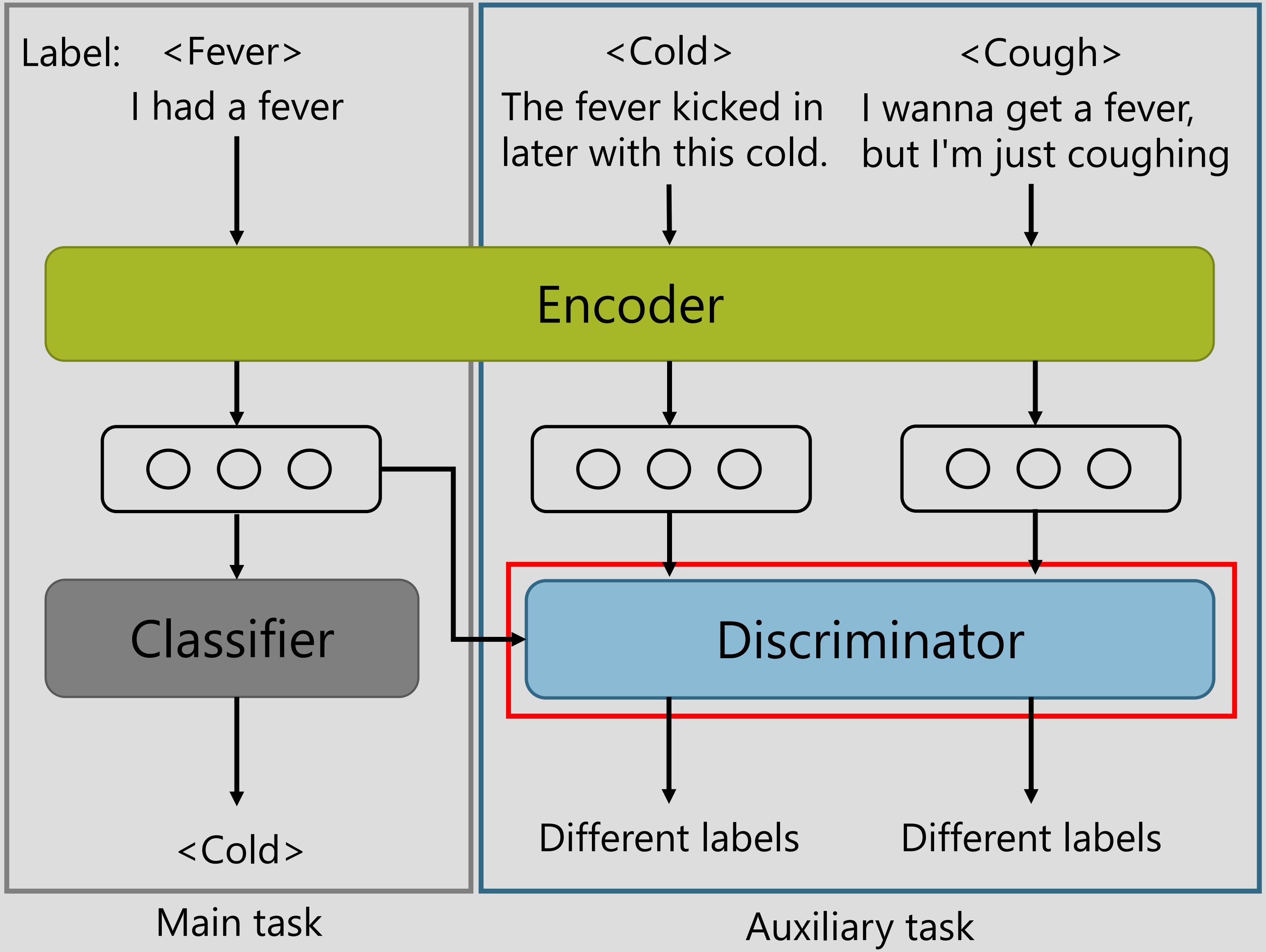
- Utilize negative examples to enable negative supervision of the encoder
- Generate distinct representations for inputs of different labels



Proposed Method

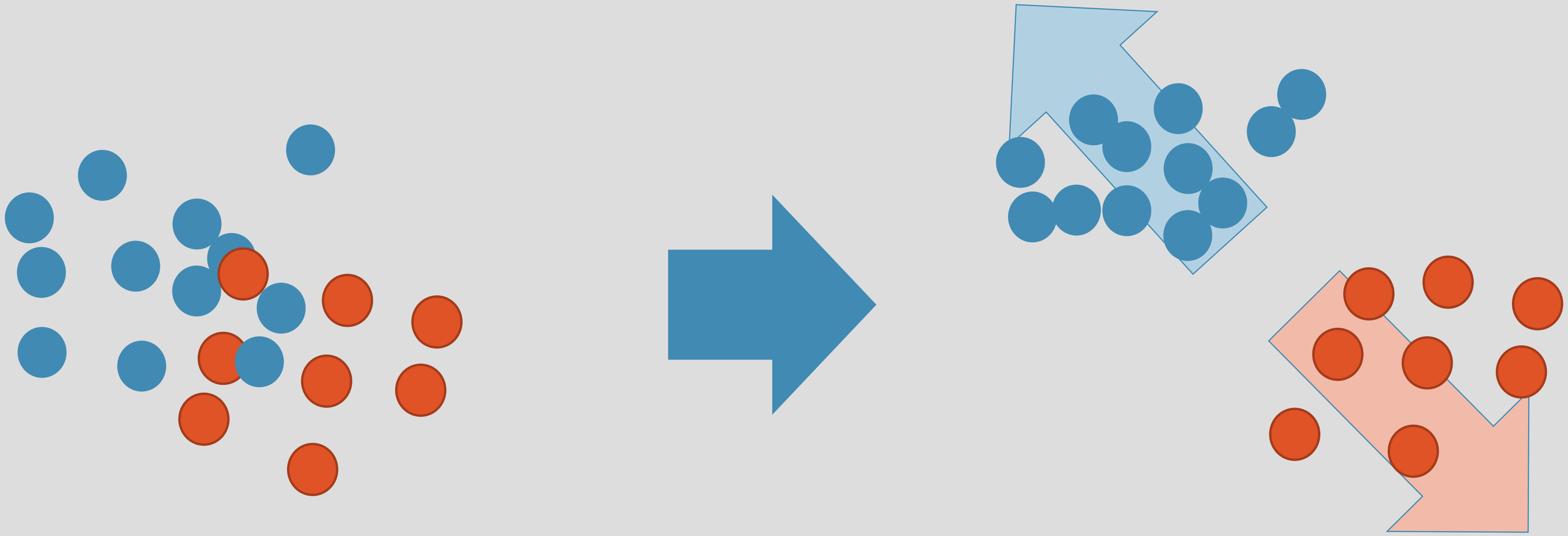


Proposed Method



Discriminator

- Discriminate whether two inputs have the same label or not
- Representations that have different labels become dissimilar through training (Negative Supervision)



The loss function of the auxiliary task

- Variations of the loss function
  - AAN (Auxiliary task with all negative examples)

$$L_a = \frac{1}{n}\sum_i s_i\,, \qquad s_i = 1 + \cos(\boldsymbol{v}_m, \boldsymbol{v}_{a_i})$$

  - AM (Auxiliary task with the margin loss)

$$L_a = \max\left(0, \delta - s_k + \frac{1}{n-1}\sum_{i \neq k} s_i\right)$$

    - k: the index of the positive example
    - δ: The margin
    - $\boldsymbol{v}_m, \boldsymbol{v}_{a_i}$: The vector representation of the main (auxiliary) task

Experiments

Datasets

Dataset	Task	Type	# of labels
MR	Sentence polarity	Single-label	2
SST-5	Fine-grained sentence polarity	Single-label	5
TREC	Classification of question types	Single-label	6
MedWeb	Classification of disease	Multi-label	8
arXiv	Document classification of fields of papers	Multi-label	40

Models

Name	Negative Supervision	Encoder
Baseline	None	BERT[1], HAN[2]
AAN	✓	
AM	✓	

[1] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019
[2] Yang et al., Hierarchical Attention Networks for Document Classification, NAACL, 2016

Results

	MR	SST-5	TREC	MedWeb (ja)	MedWeb (en)	MedWeb (zh)	arXiv
Baseline	86.5	54.0	97.0	86.1	83.1	86.9	36.0
AAN	86.8	53.0	96.9	87.1	83.6	86.4	36.4
AM	86.4	52.9	97.2	86.5	83.2	87.1	36.3

- Our method improves the performance on any dataset except SST-5
- Negative supervision works in most conditions

Results

	MR	SST-5	TREC	MedWeb (ja)	MedWeb (en)	MedWeb (zh)	arXiv
Baseline	86.5	54.0	97.0	86.1	83.1	86.9	36.0
AAN	86.8	53.0	96.9	87.1	83.6	86.4	36.4
AM	86.4	52.9	97.2	86.5	83.2	87.1	36.3

- Our method improves the performance on any dataset except SST-5
- Negative supervision works in most conditions

Conclusion

- We introduced the negative supervision to prevent the model from misclassification of text that has similar meaning
- Our method improves the performance on
  - Both single- and multi-label classifications
  - Sentence and document classifications
  - Classifications in three different languages
- We intend to consider semantic relations between class labels in the future