Task-Aware Representation of Sentences for Generic Text Classification

K Halder, Zalando SE
A Akbik, Humboldt-Universitat zu Berlin
J Krapac, Humboldt-Universitat zu Berlin
R Vollgraf, Humboldt-Universitat zu Berlin

Year of Publication :- 2020
Number of Citations :- 67

Introduction

Introduction

- State of the art for text classification use transformer with linear layer on top.
- Effective for different tasks.
- Suffers from conceptual limitations that affect usage in zero shot or few shot transfer learning.
- In a transfer learning setting, linear layer and the information on it need to be discarded when a new class is added.

Introduction

- Extending a classifier to predict a new class with very few training examples. This uses,
  - Information in the pre-trained decoder (linear layer).
  - Information provided by class labels.
- Evaluation of few shot and zero shot learning abilities of the method.
- TARS (Task-Aware Representation of Sentences).

Related Works

Related Works

- Transfer learning - Transferring knowledge from one learned task to another relies on exploiting similarities across tasks.
  - Question Answering [1].
  - Fine-tuning for Text Classification [2].
  - BERT[3] for query-based passage re-ranking[4].
  - Fine tuning BERT for text retrieval [5].
- Zero/few shot learning [6,7].

1. Min, S., Seo, M., & Hajishirzi, H. (2017). Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.
2. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
4. Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
5. Dai, Z., & Callan, J. (2019, July). Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 985-988).
6. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018). Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.
7. Geng, R., Li, B., Li, Y., Zhu, X., Jian, P., & Sun, J. (2019). Induction networks for few-shot text classification. *arXiv preprint arXiv:1902.10482*.

Methodology

Methodology

- Universal Binary Text Classification Formulation.
- Cross-Attention between Text and Label.
- Training and Prediction.
- Model Transfer.

Universal Binary Text Classification Formulation

- Goal of any text classification problem is to find a function.

$$f : text \rightarrow \{0,1\}^M \quad i.e., \quad f(t) = P(y_i|t) \, \forall \, i \in \{1...M\}$$

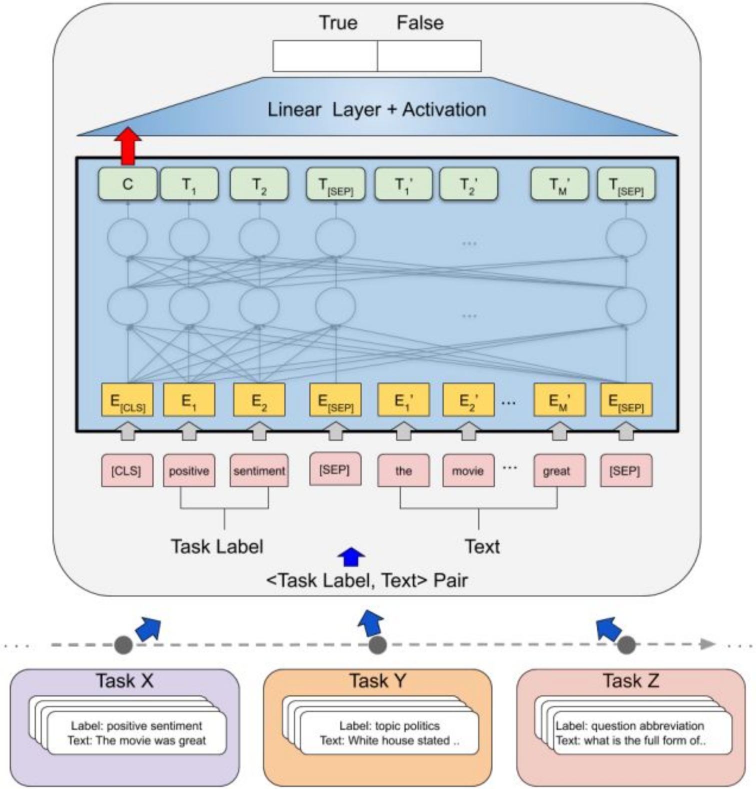
- Factorize the text classification problem into a generic binary classification task.

$$f :< task\, label, text > \rightarrow \{0,1\} \quad i.e., \quad f(\text{label}(y_i), t) = P(\text{ True} \mid y_i, t) \, \forall \, i \in \{1...M\}$$

```
<"positive sentiment", "I enjoyed the movie a lot">

<"topic politics", "The White House announced that [..]">
```

Universal Binary Text Classification Formulation



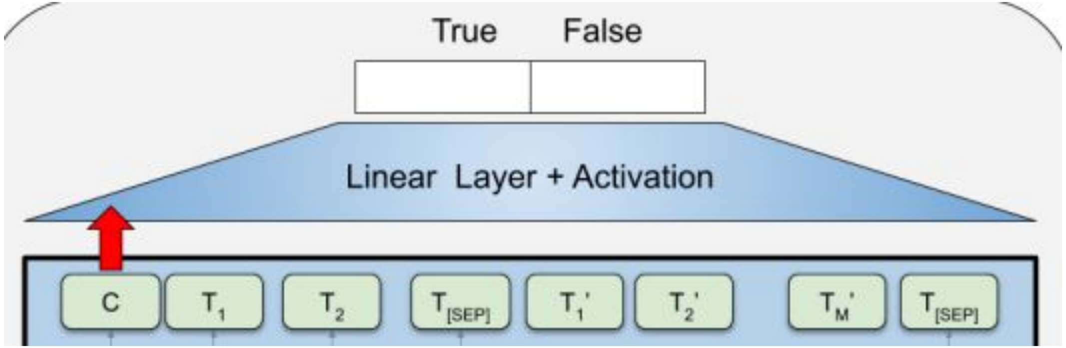
Cross-Attention between Text and Label.

- Additional input to the BERT [3] encoder in the form of class label.
- The encoder itself must learn to understand the connection between a class label and a given text.
- Cross-attention mechanism that transformer architectures supply is made use.
- [CLS], the class label, [SEP] and the text to classify.
- This input sequence is then passed through all self-attention layers in BERT.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cross-Attention between Text and Label

- [CLS]-token in the final layer as the task label dependent representation of the input text.
- Linear layer to project the H-dimensional tensor produced by the encoder into 2 real-valued logits. A softmax function is used to form a probability distribution over 2 classes i.e., True, and F.



Training and Prediction

- Populate M <task label, text> pairs for each sample text for a text classification task with M classes.
- Increased amount of the training data and thus the computational costs by a factor of M.
- During prediction, true/false predictions are done for all the possible M <Label,Text> pairs for the classification task.

<"positive sentiment", "I enjoyed the movie a lot"> → TRUE
<"negative sentiment", "I enjoyed the movie a lot"> → FALSE

Training and Prediction

- For multi-class problems, use the class with maximum confidence (for True)
- Conceptual drawback.
- Followed standard practice and use cross-entropy loss, and optimize all parameters using gradient descent.

Model Transfer

- The entire model (encoder and decoder) can be shared across tasks, as the encoder now performs the matching between label and text.
- Transfer learning to train a new tasks becomes equivalent to continuing to train the same model with different training data.
- Advantages in few-shot learning scenarios.

Model Transfer

- If there is enough similarity between tasks (e.g., the nature of the classification task, and/or word distributions), this formulation even enables a zero-shot scenario.
- Enables multi-task learning across corpora with different annotations as separate prediction heads for each task not required.
- Train the same model using tuples from different tasks and during prediction only request predictions for the labels required.

Computational Complexity

- Traditional text classification requires one forward pass per task for each input text.
- TARS requires M forward passes, one for each input text.
- The model parameters for different tasks are shared, so only one model for all tasks is kept in memory, while traditional models require a separate model for each task.
- Therefore TARS is more suited for training many tasks.

Experiments

Experiments

- How well is TARS able to transfer to new classification tasks with little training data?
- How does semantic distance between source and target task affect the transfer learning abilities of TARS?
- And what are the zeroshot capabilities of TARS?

Datasets

- 2 datasets for the task of topic detection.
  - AGNEWS [8], a corpus of news articles classified into 4 topics.
  - DBPEDIA [8], a corpus of 14 entity topics.
- One dataset in two variants for the task of classifying question types [9], namely TREC-6 with 6 coarse-grained and TREC-50 with 50 fine-grained.
- Two corpora for 5-class sentiment analysis, namely AMAZON-FULL [8] for product reviews and YELP-FULL [8] for restaurant reviews.

[8] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.

[9] Li, X., & Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Label Formulations

- In AGNEWS [8], and DBPEDIA [8] short class labels were manually curated so that they form individual meaningful words.
  - “Sci/Tech” was renamed to “Science Technology”.
  - “EducationalInstitution” to “Educational Institution”.
- For the sentiment analysis datasets, a numeric rating (1−5) is available along with each sample. They were formulated as textual description.

Transfer learning setup

- Source task and target task.
- The model for the source task is trained using the full dataset for the respective task.
- To evaluate transfer learning capabilities in few-shot and zero-shot scenarios, fine-tune the source model on the target task using only very limited numbers of training examples.
- Reporting accuracy for all the baseline models for different transfer scenarios.

Transfer learning setup

- Started with zero shot scenario, where the model does not see any training example from the target task (i.e., k = 0).
- Exposed the models to increasing number of randomly chosen samples per class from the target task (k = 1, 2, 4, ...)
- Observed how fast the competing models are able to leverage new labeled data.

Comparison

- TARS against two baselines:
  - BERTBASE [3]: Standard non-transfer learning variant in which a pre-trained BERT-model ('bert-base-uncased') is fine tuned with a linear classifier on top directly on the target task.
  - BERTBASE (ft): In this variant, BERT is fine tuned on the source task. The encoder weights are transferred to a new model and initialize a new linear layer, and fine-tune this model again on the target task. This covers the traditional transfer learning mechanism prevalent in the literature.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Results

Domain: Sentiment Analysis									
YELP-FULL → AMAZON-FULL					AMAZON-FULL → YELP-FULL				
M	k	BERT$_{BASE}$	BERT$_{BASE}$ (ft)	TARS	M	k	BERT$_{BASE}$	BERT$_{BASE}$ (ft)	TARS
	0	–	–	**51.8**		0	–	–	**50.6**
	1	21.8±1.7	27.5±6.5	**51.0**±0.3		1	22.5±3.2	28.0±5.3	**53.0**±0.3
	2	24.6±1.1	36.4±7.0	**52.7**±0.2		2	22.6±1.7	33.7±4.1	**52.2**±0.7
5	4	25.8±1.7	43.2±3.0	**52.3**±0.5	5	4	26.5±2.3	44.1±1.4	**52.0**±2.1
	8	25.4±1.8	45.0±1.1	**49.9**±1.7		8	31.9±2.0	46.5±2.0	**53.3**±1.1
	10	29.0±1.5	45.2±1.0	**51.6**±0.4		10	32.8±2.1	47.2±3.0	**52.5**±0.3
	100	50.7±0.9	53.2±0.4	**53.4**±0.4		100	53.9±1.8	55.8±0.5	**56.4**±0.7

Results

Domain: Topic Classification

DBPEDIA → AGNEWS					AGNEWS → DBPEDIA				
M	k	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ (ft)	TARS	M	k	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ (ft)	TARS
	0	–	–	**52.4**		0	–	–	**51.2**
	1	41.6±6.5	66.6±4.6	**72.1**±3.4		1	45.4±2.6	45.2±3.7	**76.6**±2.7
	2	56.0±3.3	69.8±2.7	**74.3**±4.5		2	76.4±2.4	66.0±4.2	**81.7**±3.8
4	4	70.8±5.6	78.5±2.3	**80.2**±0.9	14	4	**91.3±0.5**	84.4±2.7	90.1±1.3
	8	78.3±1.3	80.1±2.1	**81.0**±0.8		8	**96.5**±0.4	93.5±1.4	94.8±0.7
	10	80.1±2.9	82.0±0.6	**83.5**±0.2		10	**97.6**±0.3	95.8±0.1	96.6±0.2
	100	**87.8**±0.4	86.9±0.4	86.7±0.3		100	**98.7**±0.0	98.4±0.0	98.4±0.0

Results

Domain: Question Type Classification				
TREC-6 $\rightarrow$ TREC-50				
M	k	BERT$_{\text{BASE}}$	BERT$_{\text{BASE}}$ (ft)	TARS
50	0	–	–	**53.4**
	1	11.4$\pm$3.7	40.2$\pm$4.8	**57.2**$\pm$1.0
	2	29.1$\pm$4.7	74.5$\pm$1.4	**82.0**$\pm$2.6
	4	47.9$\pm$5.2	78.6$\pm$1.3	**82.7**$\pm$2.3
	8	64.4$\pm$1.6	81.6$\pm$1.5	**86.2**$\pm$2.9
	10	67.1$\pm$2.9	83.2$\pm$0.7	**85.1**$\pm$1.0
	100	89.6$\pm$0.6	91.3$\pm$0.2	**91.4**$\pm$0.5

Results

Model	Model Size	AGNEWS	DBPEDIA
GPT-2 (2019)	117M	40.2*	39.6*
TARS	110M	**52.4**	**51.2**

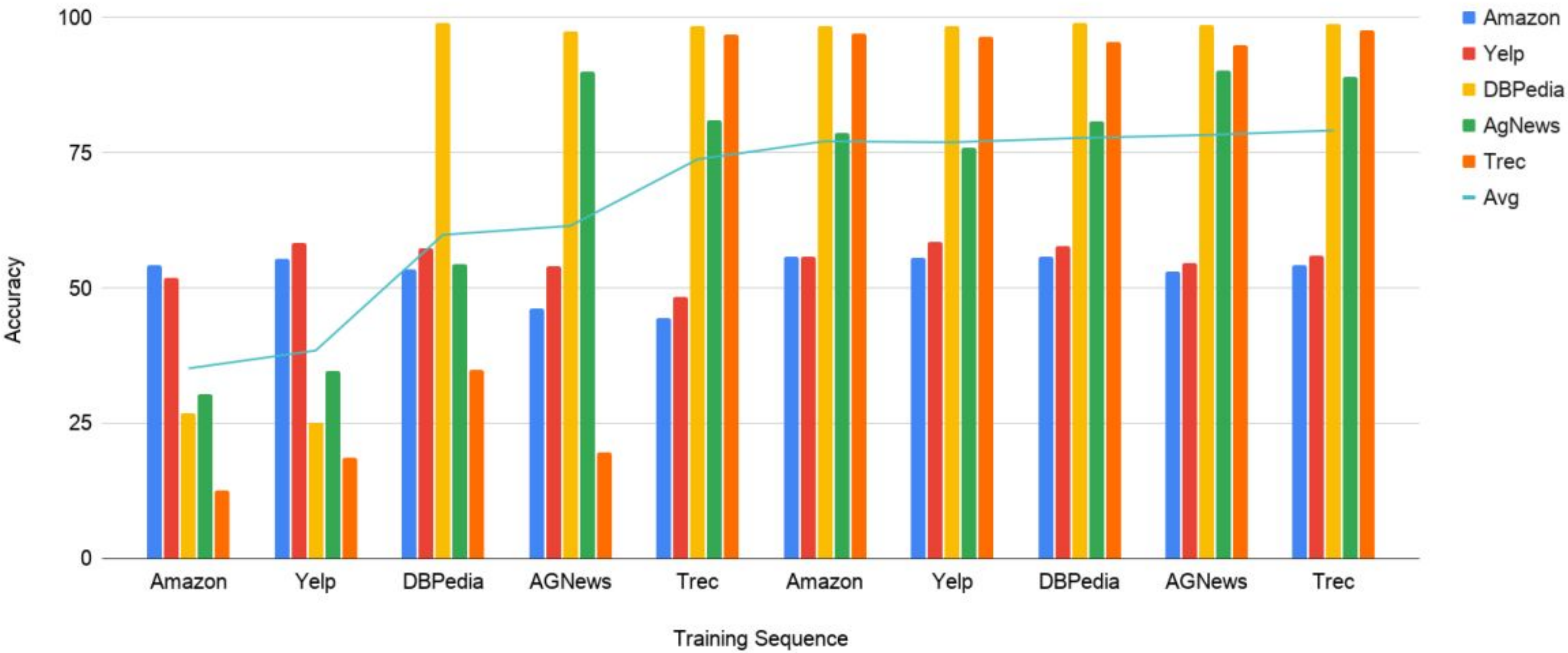
Results

Cross Domain Transfer									
DBPEDIA (Topic) $\rightarrow$ TREC-6 (Question Type)					AMAZON-FULL (Sentiment) $\rightarrow$ AGNEWS (Topic)				
M	k	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ (ft)	TARS	M	k	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ (ft)	TARS
6	0	–	–	**43.0**	6	0	–	–	**28.0**
	1	26.4±4.2	38.5±3.9	**45.7**±6.2		1	**43.8**±4.0	29.8±0.7	42.9±3.5
	2	36.9±6.0	32.8±7.1	**62.9**±5.7		2	**59.6**±1.1	37.1±4.3	49.5±1.0
	4	43.5±3.2	45.3±3.0	**62.7**±2.2		4	**70.4**±4.6	49.0±2.8	63.7±6.4
	8	56.4±3.1	57.2±1.8	**61.9**±1.9		8	**80.5**±0.3	57.4±0.8	79.2±0.2
	10	58.8±6.6	63.7±2.3	**64.7**±1.0		10	**81.4**±0.7	65.4±6.3	79.6±0.7
	100	92.5±0.8	**93.4**±1.0	91.6±0.9		100	**88.0**±0.1	86.9±0.4	86.6±0.6

Ablation Study: New Class Without Training Data added

Domain: Topic Classification w/ New Class Addition				
DBPEDIA-13 $\rightarrow$ DBPEDIA				
M	k	$\text{BERT}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ (ft)	TARS
14	0	–	–	**0.60**
	1	0.05	0.40	**0.72**
	2	0.58	0.73	**0.85**
	4	0.91	0.89	**0.96**
	8	0.93	0.91	**0.95**
	10	0.93	0.94	**0.96**
	100	0.98	0.98	**0.99**

Knowledge Retention Experiment



Conclusion

- Proposed TARS architecture to address key shortcomings of transfer learning approaches.
- The proposed TARS architecture captures the similarity between an input text and the task label to perform text classification.
- TARS is capable of making zero-shot predictions in multiple text classification tasks.
- TARS adapts to a new domain faster than competitive baseline models in few-shot learning settings.

References

1. Min, S., Seo, M., & Hajishirzi, H. (2017). Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.
2. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
4. Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
5. Dai, Z., & Callan, J. (2019, July). Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 985-988).
6. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018). Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.
7. Geng, R., Li, B., Li, Y., Zhu, X., Jian, P., & Sun, J. (2019). Induction networks for few-shot text classification. *arXiv preprint arXiv:1902.10482*.
8. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.
9. Li, X., & Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.