

Task-Aware Representation of Sentences for Generic Text Classification



Kishaloy Halder



Alan Akbik



Josip Krapac



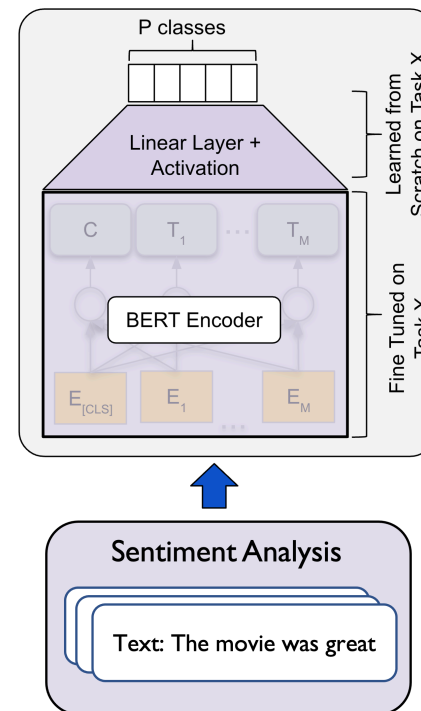
Roland Vollgraf

COLING 2020
December 8-13



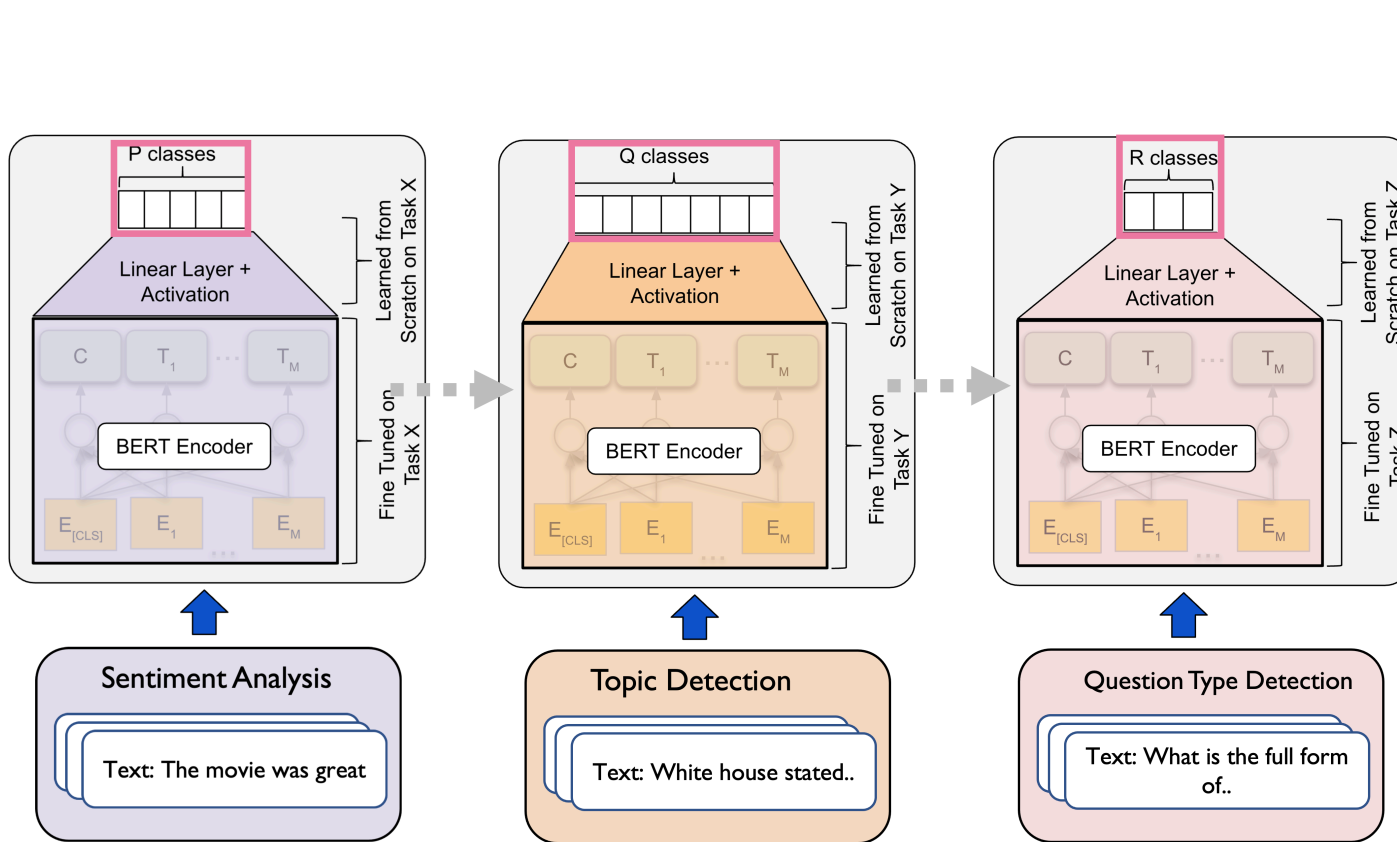
Text Classification: Forms and Standards

- Task: Classify textual documents into pre-defined classes
- Used in variety of downstream applications:
 - Sentiment Analysis, Topic Detection, Question Type



Standard Formulation
 $f: text \rightarrow \{0, 1\}^P$

Text Classification: Common Transfer Learning Practices



Task-Aware Representation of Sentences for Generic Text Classification

Standard Formulation

$$f: \text{text} \rightarrow \{0, 1\}^P$$

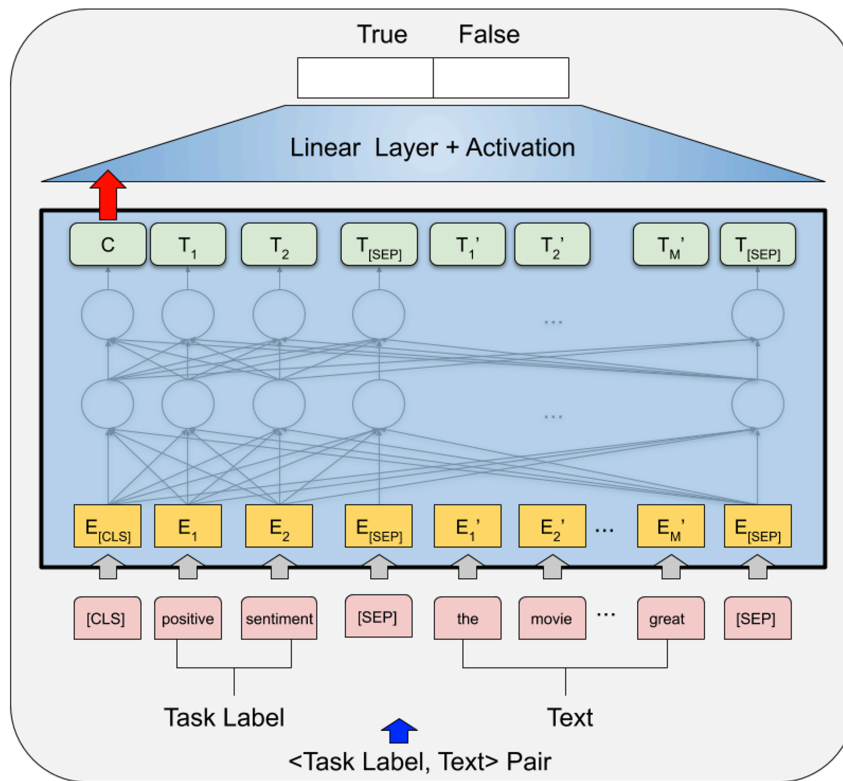
Limitation #1

- $P \neq Q \neq R$
- Information in the decoder can not be transferred

Limitation #2

- Class semantics are learned implicitly from the training samples
- Can not leverage label names e.g., "Politics"

Our Proposed Approach: TARS



Standard Formulation
 $f: \text{text} \rightarrow \{0, 1\}^P$

Our Formulation
 $f: \langle \text{task label}, \text{text} \rangle \rightarrow \{0, 1\}$

- ✓ Makes the *entire stack* independent of number of classes in a task
- ✓ Enables transfer of *all parameters* between tasks
- ✓ Encodes the label names *explicitly* from tasks

Sentiment Analysis

Text: The movie was great
 Label: positive sentiment

Topic Detection

Text: White house stated..
 Label: topic politics

Question Type Detection

Text: What is the full form..
 Label: question abbreviation

Task-Aware Representation of Sentences for Generic Text Classification

TARS: Working Principle

Our Formulation

$$f: \langle \text{task label}, \text{text} \rangle \rightarrow \{0, 1\}$$

Sentiment Analysis

Positive, Neutral, Negative

Training Sample

Text: "I enjoyed the movie a lot", Label: Positive

Transformed Input to TARS

"Positive Sentiment [SEP] I enjoyed the movie a lot", Label: 1

"Neutral Sentiment [SEP] I enjoyed the movie a lot", Label: 0

"Negative Sentiment [SEP] I enjoyed the movie a lot", Label: 0

Inference

1. Populate $\langle \text{task label}, \text{text} \rangle$ tuples from input text
2. Perform *argmax* over all classes

Complexity

Grows linearly with number of classes in a task

Research Questions



Can the Task-Aware formulation help in zero/few shot scenario?

How does the semantic distance between tasks affect the transfer learning capability?

How well does TARS memorize multiple tasks?

Training Data size vs Accuracy of a Typical Model

Task-Aware Representation of Sentences for Generic Text Classification

Experiments: Baseline Setup

Used Standard Datasets from Multiple Classification Task Types

Dataset	Type	#classes
TREC-6 (Li and Roth, 2002)	Question	6
TREC-50 (Li and Roth, 2002)	Question	50
YELP-FULL (Zhang et al., 2015)	Sentiment	5
AMAZON-FULL (Zhang et al., 2015)	Sentiment	5
AGNEWS (Zhang et al., 2015)	Topic	4
DBPEDIA (Zhang et al., 2015)	Topic	14

Evaluation of Transfer Learning Capability

- Train classification model on *source task* with all labeled samples
- Fine Tune the model on limited labeled samples from *target task*
- Compare Accuracy of Baseline models on the **entire test set** from *target task*

Model	Source Task	Target Task
BERT _{BASE}	No Access	Limited Access
BERT _{BASE} (ft)	Full Access	Limited Access
TARS	Full Access	Limited Access

Experiments: Baseline Comparison (In Domain)

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

Domain: Topic Classification									
DBPEDIA → AGNEWS					AGNEWS → DBPEDIA				
M	k	BERT _{BASE}	BERT _{BASE} (ft)	TARS	M	k	BERT _{BASE}	BERT _{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

M : Number of classes in target task

k : Number of labelled samples per class used for training

- TARS shows **impressive** improvement in zero/few shot scenarios
- The binary text classification is **effective** compared to other zero shot formulations

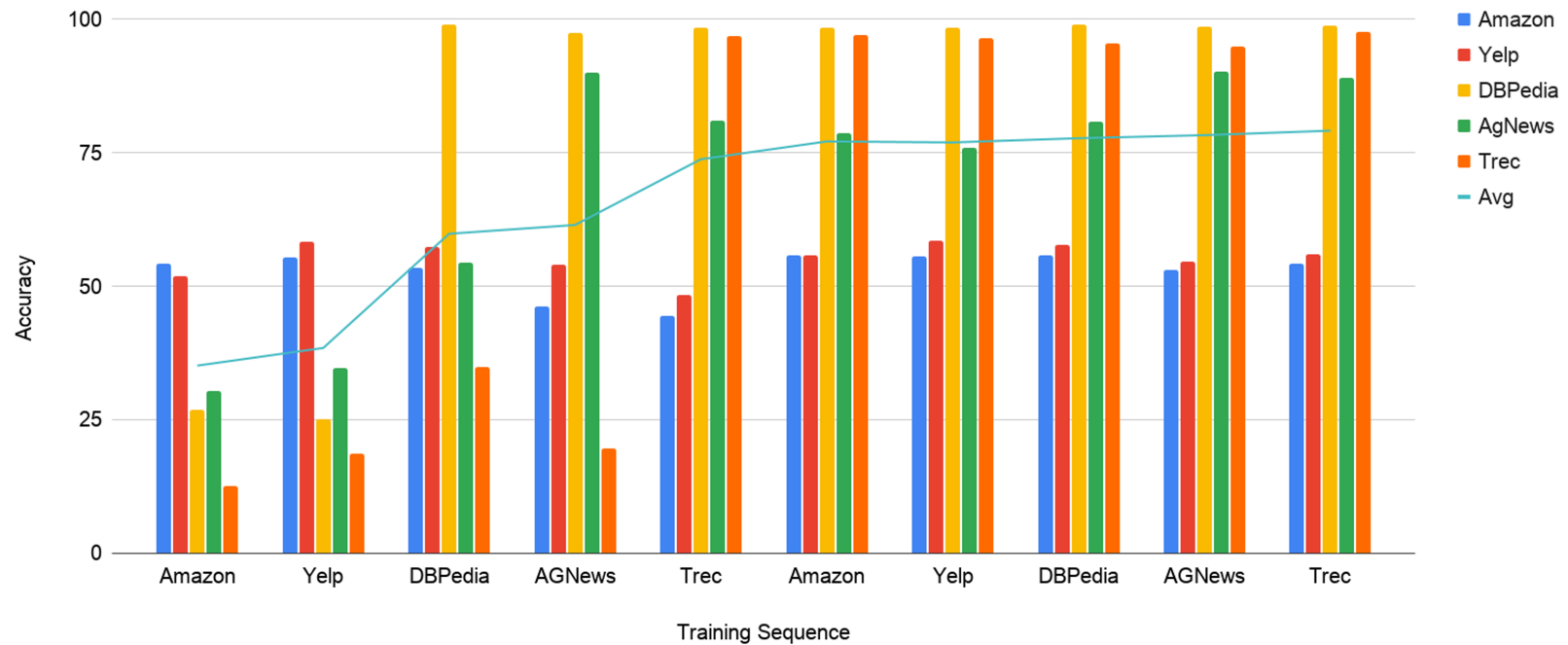
Model	Model Size	AGNEWS	DBPEDIA
GPT-2 (2019)	117M	40.2*	39.6*
TARS	110M	52.4	51.2

Experiments: Baseline Comparison (Cross Domain)

DBPEDIA (Topic) → TREC-6 (Question Type)					AMAZON-FULL (Sentiment) → AGNEWS (Topic)				
M	k	BERT _{BASE}	BERT _{BASE} (ft)	TARS	M	k	BERT _{BASE}	BERT _{BASE} (ft)	TARS
	0	–	–	43.0		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
6	4	43.5±3.2	45.3±3.0	62.7±2.2	6	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

- Cross domain transfer remains a **challenging** task
- If domains are very different (both nature of task, and formalism), knowledge from the *source task* becomes **less effective**

Experiments: Pushing TARS beyond Single Task



- Continue training a **single TARS model** on all the datasets
- Observe accuracy across all datasets after each training round
- The final model does not show **catastrophic forgetting**
- All 5 tasks can be performed well by the final model

Conclusion

- Proposed a **task label aware** formulation for text classification called TARS
- Allows **full transfer of model weights** on unseen tasks
- Performed experiments to track its **effectiveness in limited data scenario** i.e., zero/few shot
- TARS can perform **accurate zero shot predictions**
- Outperforms other transfer learning mechanisms in few shot cases
- TARS can encapsulate multiple tasks in a **single model**, does not show catastrophic forgetting
- **Ready-to-use** implementation available in flair

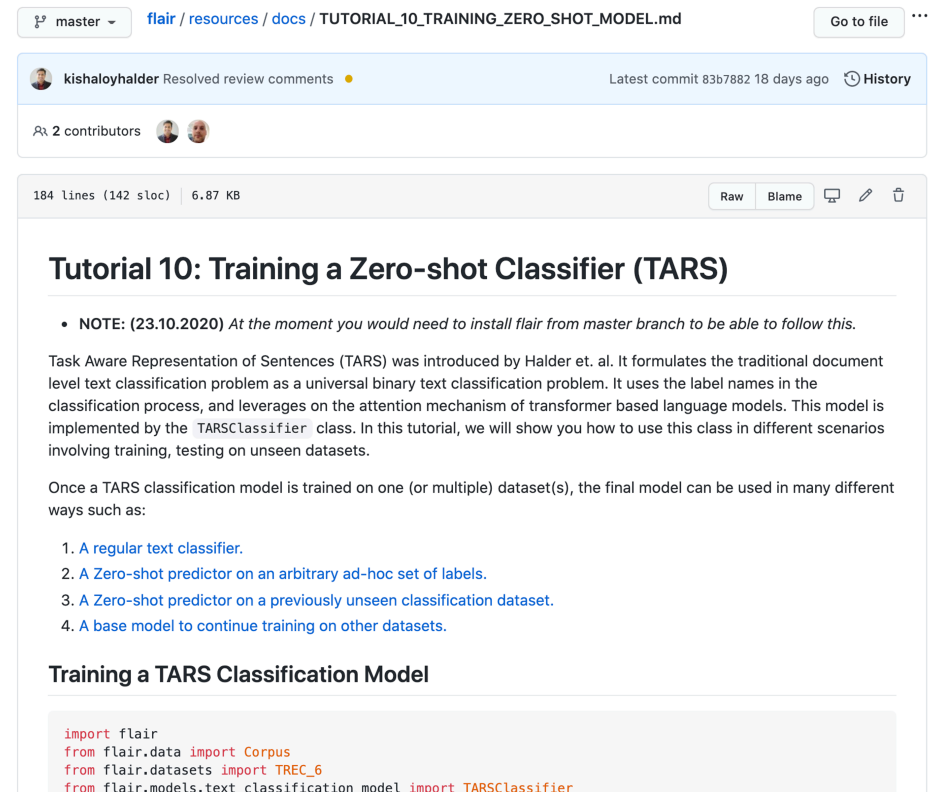
Thanks for listening!

Contact: kishaloy.halder@zalando.de

Paper: bit.ly/TARS-PDF

Code: bit.ly/TARS-CODE

Task-Aware Representation of Sentences for Generic Text Classification



master flair / resources / docs / TUTORIAL_10_TRAINING_ZERO_SHOT_MODEL.md Go to file

kishaloyhalder Resolved review comments Latest commit 83b7882 18 days ago History

2 contributors

184 lines (142 sloc) 6.87 KB Raw Blame

Tutorial 10: Training a Zero-shot Classifier (TARS)

- **NOTE: (23.10.2020)** *At the moment you would need to install flair from master branch to be able to follow this.*

Task Aware Representation of Sentences (TARS) was introduced by Halder et. al. It formulates the traditional document level text classification problem as a universal binary text classification problem. It uses the label names in the classification process, and leverages on the attention mechanism of transformer based language models. This model is implemented by the `TARSClassifier` class. In this tutorial, we will show you how to use this class in different scenarios involving training, testing on unseen datasets.

Once a TARS classification model is trained on one (or multiple) dataset(s), the final model can be used in many different ways such as:

1. A regular text classifier.
2. A Zero-shot predictor on an arbitrary ad-hoc set of labels.
3. A Zero-shot predictor on a previously unseen classification dataset.
4. A base model to continue training on other datasets.

Training a TARS Classification Model

```
import flair
from flair.data import Corpus
from flair.datasets import TREC_6
from flair.models.text_classification_model import TARSClassifier
```