Semantics-aware BERT for Language Understanding (SemBERT)

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, Xiang Zhou
Shanghai Jiao Tong University & CloudWalk Technology
zhangzs@sjtu.edu.cn, will8821@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

云从科技 — 定义智慧生活 —
上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

Introduction

Semantics-aware BERT (SemBERT):

- incorporate explicit contextual semantics from pre-trained semantic role labeling
- capable of explicitly absorbing contextual semantics over a BERT backbone
- obtains new state-of-the-art or substantially improves results on ten reading comprehension and language inference tasks.

Motivation:

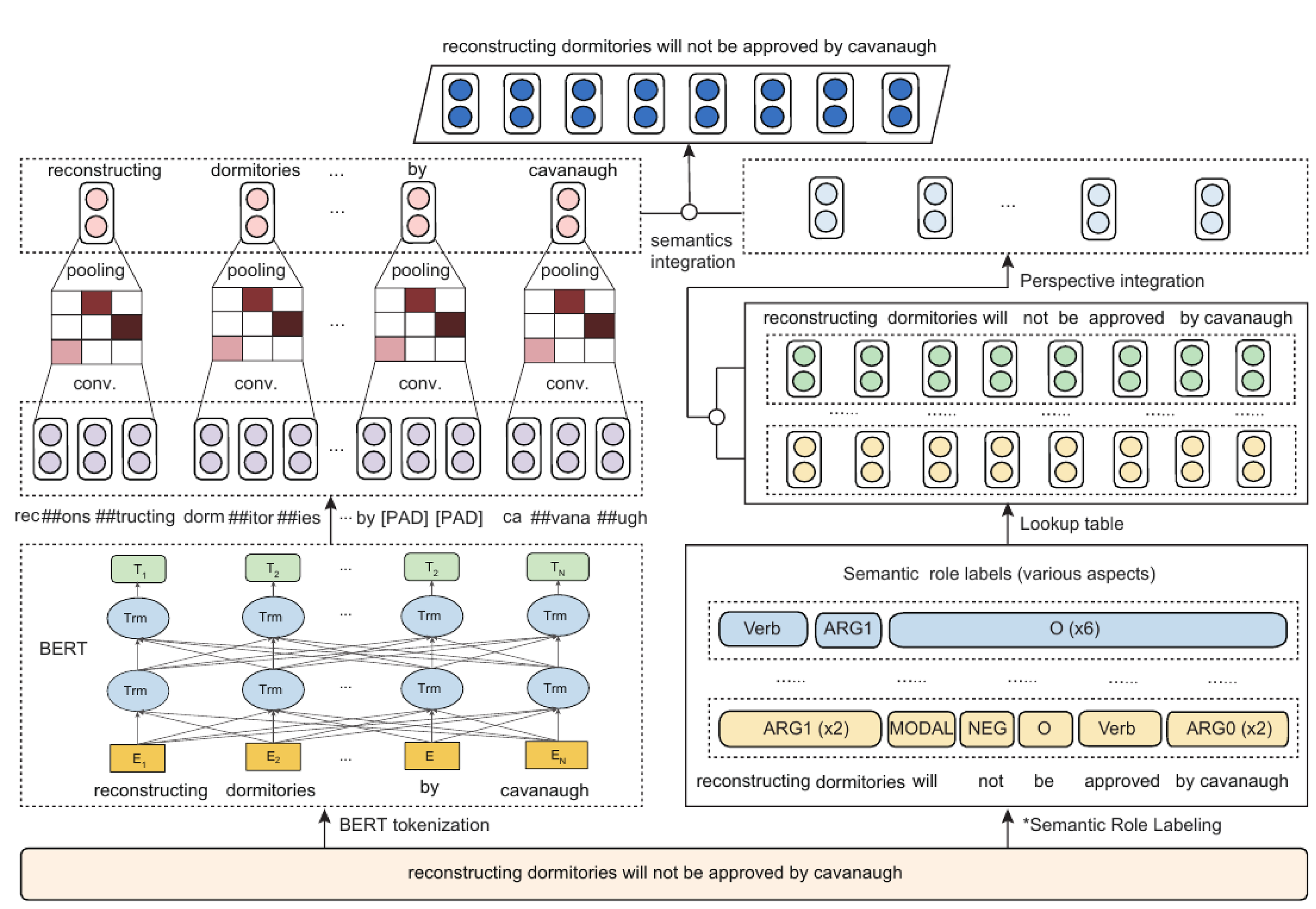
- Pre-trained language models rarely consider incorporating structured semantic information.
- Deep learning models might not really understand the natural language texts and vulnerably suffer from adversarial attacks.
- NLU tasks share the similar task purpose as sentence contextual semantic analysis.

Paper Link: https://arxiv.org/abs/1909.02209
Code Link: https://github.com/cooelf/SemBERT

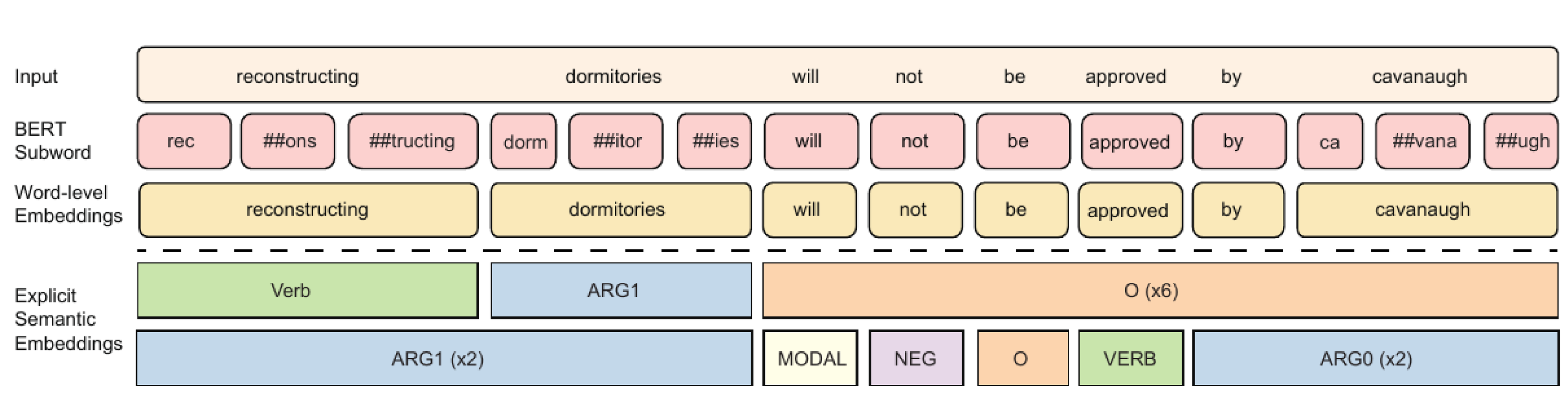
Method

SemBERT comprises three parts:

- Semantic Role Labeling
  - Annotate the input sentences
  - fetch multiple predicate-derived structures of explicit semantics
- Encoding
  - Vectorize and obtain the contextual representations of both the sentence and label sequences.
- Integration
  - the sentence representations and semantic embedding are concatenated to form the joint representation for downstream tasks



For the text, {*reconstructing dormitories will not be approved by cavanaugh*}, it will be tokenized to a subword-level sequence, {*rec, ##ons, ##tructing, dorm, ##itor, ##ies, will, not, be, approved, by, ca, ##vana, ##ugh*}. Meanwhile, there are two kinds of word-level semantic structures,
[ARG1: reconstructing dormitories] [ARGM-MOD: will] [ARGM-NEG: not] be [V: approved] [ARG0: by cavanaugh]
[V: reconstructing] [ARG1: dormitories] will not be approved by cavanaugh



Experiments

Datasets: 10 NLU benchmark datasets involving natural language inference, machine reading comprehension, semantic similarity and text classification.
Tasks: GLUE, SNLI, SQuAD2.0
Baseline: BERT

Method	Classification		Natural Language Inference			Semantic Similarity			Score
	CoLA (mc)	SST-2 (acc)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)	-
Leaderboard (September, 2019)									
ALBERT	69.1	97.1	91.3/91.0	99.2	89.2	93.4	74.2	92.5	89.4
RoBERTa	67.8	96.7	90.8/90.2	98.9	88.2	92.1	90.2	92.2	88.5
XLNET	67.8	96.8	90.2/89.8	98.6	86.3	93.0	90.3	91.6	88.4
In literature (April, 2019)									
BiLSTM+ELMo+Attn	36.0	90.4	76.4/76.1	79.9	56.8	84.9	64.8	75.1	70.5
GPT	45.4	91.3	82.1/81.4	88.1	56.0	82.3	70.3	82.0	72.8
GPT on STILTs	47.2	93.1	80.8/80.6	87.2	69.1	87.7	70.1	85.3	76.9
MT-DNN	61.5	95.6	86.7/86.0	-	75.5	90.0	72.4	88.3	82.2
$BERT_{BASE}$	52.1	93.5	84.6/83.4	-	66.4	88.9	71.2	87.1	78.3
$BERT_{LARGE}$	60.5	94.9	86.7/85.9	92.7	70.1	89.3	72.1	87.6	80.5
Our implementation									
$SemBERT_{BASE}$	57.8	93.5	84.4/84.0	90.9	69.3	88.2	71.8	87.3	80.9
$SemBERT_{LARGE}$	62.3	94.6	87.6/86.3	94.6	84.5	91.2	72.8	87.8	82.9

GLUE

Model	EM	F1
#1 BERT + DAE + AoA†	85.9	88.6
#2 SG-Net†	85.2	87.9
#3 BERT + NGM + SST†	85.2	87.7
U-Net (Sun et al. 2018)	69.2	72.6
RMR + ELMo + Verifier (Hu et al. 2018)	71.7	74.2
Our implementation		
$BERT_{LARGE}$	80.5	83.6
$SemBERT_{LARGE}$	82.4	85.2
$SemBERT^{*}_{LARGE}$	84.8	87.9

SQuAD2.0

Model	Dev	Test
In literature		
DRCN (Kim et al. 2018)	-	90.1
SJRC (Zhang et al. 2019)	-	91.3
MT-DNN (Liu et al. 2019)†	92.2	91.6
Our implementation		
$BERT_{BASE}$	90.8	90.7
$BERT_{LARGE}$	91.3	91.1
$SemBERT_{BASE}$	91.2	91.0
$SemBERT_{LARGE}$	92.3	91.6

SNLI

Results:

- GLUE: outperforms all the previous state-of-the-art models in literature
- SQuAD2.0: outperforms all the published works and achieves comparable performance with a few unpublished models from the leaderboard
- SNLI: achieves a new state-of-the-art on SNLI benchmark and even outperforms all the ensemble models

Analysis

Parameter Comparisons

- Without multi-task learning like MT-DNN, our model still achieves remarkable results.

Model	Params (M)	Shared (M)	Rate
MT-DNN	3,060	340	9.1
BERT on STILTs	335	-	1.0
BERT	335	-	1.0
SemBERT	340	-	1.0

The influence of the max number of predicate-argument structures

- The modest number would be better.

Number	1	2	3	4	5
Accuracy	91.49	91.36	91.57	91.29	91.42

Model Predictions

- potential to guide the model to produce meaningful predictions

Question	Baseline	SemBERT
What is a very seldom used unit of mass in the metric system?	The ki	metric slug
What is the lone MLS team that belongs to southern California?	Galaxy	LA Galaxy
How many people does the Greater Los Angeles Area have?	17.5 million	over 17.5 million