

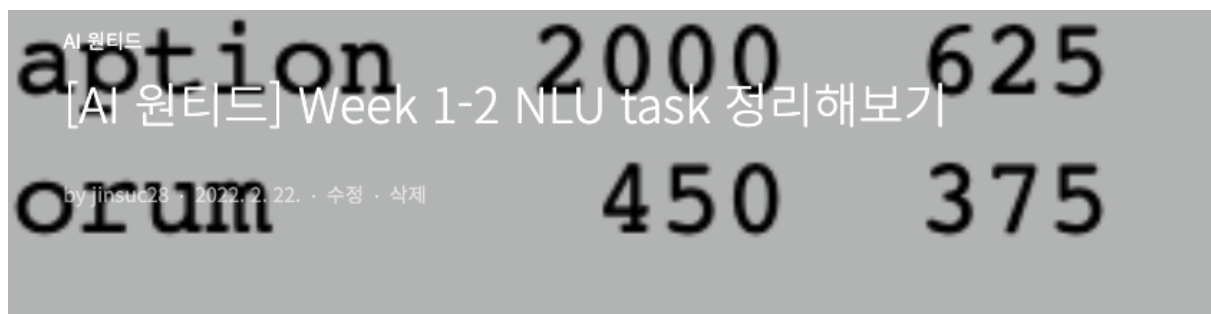
## - 본인 아이디 및 닉네임

jinsuc28(블로그 아이디), 6팀\_최진수

## - 게시글 URL

<https://jinsuc.tistory.com/7>

## - 게시글 캡처



## 0. 들어가며

NLP(Natural Language Understanding)는 크게 NLU(Natural Language Understanding)와 NLG(Natural Language Generation)로 나눌 수 있습니다. 물론, NLU와 NLG 사이의 task도 존재합니다.

오늘은 NLU task 중 **Semantic Textual Similarity**를 알아보겠습니다.

## 1. [NLU]Semantic Textual Similarity 정리해보기

### -Task 설명

문장 간의 유사도를 측정하는 것으로 비슷한 문장끼리 군집을 형성하게 할 수 있습니다.

예) 블로그 글을 유사도 측정을 통해 군집을 형성해하고 유저가 검색을 했을 때 관련도가 높은 글들을 제공해줄 수 있습니다.

## -데이터 세트

Semantic Textual Similarity on **STS Benchmark**가 대표적이며 **GLUE**(General Language Understanding Evaluation)에서

제공하는 데이터 셋이다. **2012년부터 2017년** 사이에 만들어진 데이터들로 구성되어있습니다.

	train	dev	test	total
news	3299	500	500	4299
caption	2000	625	625	3250
forum	450	375	254	1079
total	5749	1500	1379	8628


news, caption, forum 3개의 데이터 세트

※ caption은 영상의 이해를 돕기 위해 텍스트로 표시되는 자막입니다.

genre	file	years	train	dev	test
news	MSRpar	2012	1000	250	250
news	headlines	2013-16	1999	250	250
news	deft-news	2014	300	0	0
captions	MSRvid	2012	1000	250	250
captions	images	2014-15	1000	250	250
captions	track5.en-en	2017	0	125	125
forum	deft-forum	2014	450	0	0
forum	answers-forums	2015	0	375	0
forum	answer-answer	2016	0	0	254

데이터 별 상세 정보

<http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark>



### STSBenchmark - stswiki

STS benchmark dataset and companion dataset STS Benchmark comprises a selection of the English datasets used in the STS tasks orga...

ixa2.si.ehu.eus

## -SOTA 모델 두가지

### 1. SMART-RoBERTa Large

GPT-3처럼 대용량 파라미터를 이용하여 모델의 성능을 향상 시키는 모델들이 등장하면서 나온 방법론입니다. 모델에 크기는 유지하면서 성능을 극대화 시키는 것을 목표로 합니다.

저자는 이를 위해서 **Smoothness-inducing regularization**과 **Bregman proximal point optimization**를 제안합니다. 이 방법론을 적용하여 기존 **MT-DNN 모델로 SOTA**가 되었습니다.

Model / #Train	CoLA 8.5k	SST 67k	MRPC 3.7k	STS-B 7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	Score	#param
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	89.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	-
Ensemble Models												
RoBERTa <sup>1</sup>	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89.0	48.7	88.5	356M
FreeLB <sup>2</sup>	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1/90.7	98.8	88.7	89.0	50.1	88.8	356M
ALICE <sup>3</sup>	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7/90.2	99.2	87.3	89.7	47.8	89.0	340M
ALBERT <sup>4</sup>	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	99.2	89.2	91.8	50.2	89.4	235M*
MT-DNN-SMART <sup>†</sup>	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0/90.8	99.2	89.7	94.5	50.2	89.9	356M
Single Model												
BERT <sub>LARGE</sub> <sup>5</sup>	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN <sup>6</sup>	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5 <sup>8</sup>	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0/91.7	96.7	92.5	93.2	53.1	89.7	11,000M
SMART <sub>RoBERTa</sub>	65.1	97.5	93.7/91.6	92.9/92.5	74.0/90.1	91.0/90.8	95.4	87.9	91.8 <sup>8</sup>	50.2	88.4	356M

출처: SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization 본문 중

<https://paperswithcode.com/paper/smart-robust-and-efficient-fine-tuning-for>



## Papers with Code - SMART: Robust and Eff...

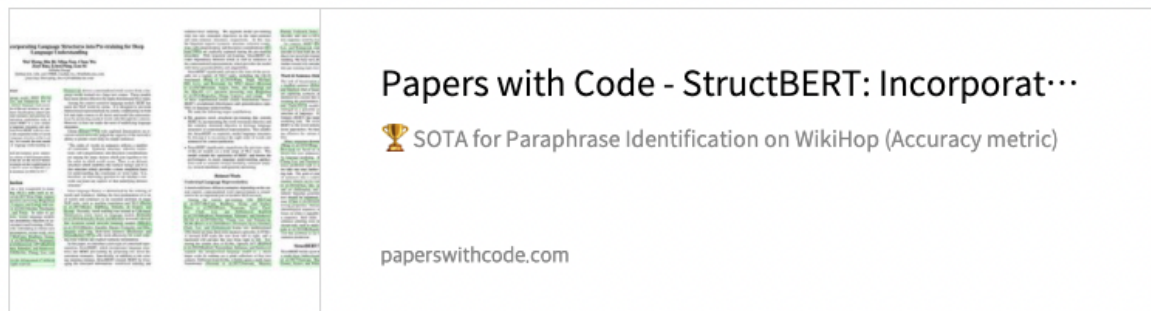
🏆 SOTA for Semantic Textual Similarity on MRPC (Accuracy metric)

[paperswithcode.com](https://paperswithcode.com)

## 2. StructBERTRoBERTa ensemble

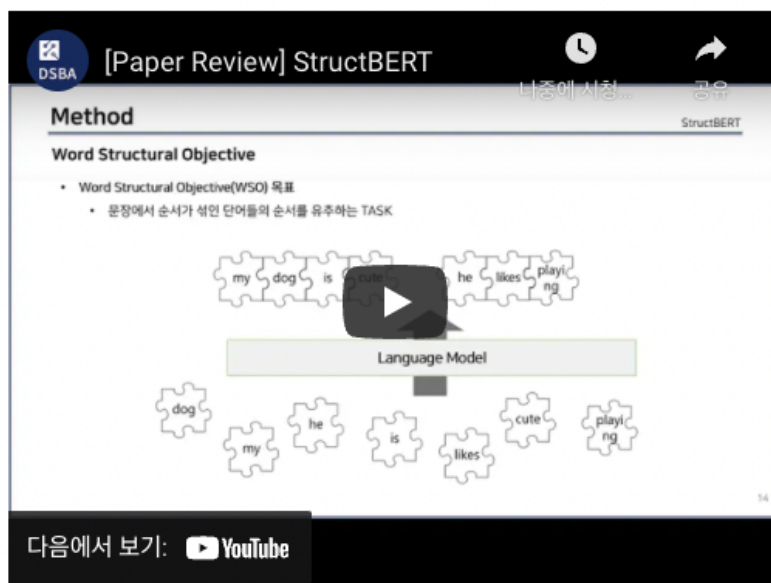
기존 BERT모델에서 **사전학습 단계 방법론을 변경**하여서 모델 성능을 향상 시킨 모델입니다. 사전학습 단계에서는 **MLM**(Mask Language Model),**SSO**(Sentence Structural Objective),**WSO**(Word Structural Objective)가 추가됩니다.

<https://paperswithcode.com/paper/structbert-incorporating-language-structures>



※ 참고자료

<https://youtu.be/iAqAhzToR6w>



### 손지아

한글 데이터들을 이용한 자연어 sota들을 잘 정리해주신 것 같습니다. 한국에서 바로 적용가능한 모델들을 정리해주신 것 같아 좋았습니다.

### 정태호

구체적으로 모델을 어떻게 평가할 것인지에 대한 설명을 적어주신 부분이 정말 좋았습니다. 전체적으로 모델의 인풋과 아웃풋 그리고 평가 방법을 쉽게 이해할 수 있었던 것 같아요.

### 현승환

간결하게 잘 작성해 주신 것 같아요. SST 데이터 중에도 2.5로 서로 다른 버전이 있다는 것을 잘 알게된 것 같습니다.