GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

이번에는 GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding이라는 논문을 리뷰해보도록 하겠다. 해당 논문은 GLUE benchmark를 제안하는 논문이며, 이 GLUE는 GPT, BERT와 같은 pre-trained language model의 성능을 테스트할 때 사용된다.

원문 링크는 다음과 같다.

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

Introduction

인간이 언어를 general, flexible, robust하게 이해하는 것과는 달리, 대부분의 NLU(Natural Language Understanding) model들은 specific 한 task를 위해 설계되었으며, out-of-domain의 data에서는 성능이 잘 나오지 않는다.

만약 단순히 입력과 출력의 표면적인 대응관계를 감지하는것을 넘어서는, linguistic understanding을 가지는 model을 만들어야 한다면, **다른 domain에서 다양한 linguistic task를 실행하는 unified model을 만드는 것이 매우 중요하다.**

이러한 unified model을 만드는 방향의 연구를 촉진시키기 위해, 본 연구에서는 GLUE(General Language Understanding Evaluation) benchmark를 제시한다.

GLUE benchmark는 QA(Question Answering), Sentiment analysis, Textual entailment를

포함하는 NLU task의 모음이며, test 하는 model의 evaluation, comparison, analysis를 위한 온라인 플랫폼과 연동된다.

(해당 온라인 플랫폼은 아래와 같다.)

GLUE Benchmark

또한 **GLUE는 model의 architecture에서 single-sentence와 sentence-pair input을 처리하여 이에 대응하는 prediction을 만들어내는 것 이외에는 아무런 제약을 두지 않는다.**

GLUE에는 training data가 풍부한 task들도 있지만, training data의 개수가 제한된 task도 있으며, training set과 test set의 genre와 match하지 않는 task도 존재한다.

이러한 특징은 GLUE가 **sample-efficient learning과 다양한 task에서 효과적인 knowledge-transfer를 촉진하는 방법으로 언어적 지식을 학습하는 model을 선호**하게끔 한다.

GLUE에 포함되어 있는 dataset들은 기존에 존재하던 dataset들로부터 기인하였다고 하며, dataset중에서 4개는 benchmark가 공정하게 사용되는지 확인하는 데 사용되는 private-held test data를 제공한다.

(private test data로 evaluation하기 위해서는 결과를 위에 소개된 온라인 플랫폼에 제출해야한다.)

이에 더불어, GLUE는 model이 학습한 유형을 이해하고, 언어적으로 의미 있는 solution strategies를 장려하기 위해 set of hand-crafted analysis example을 제공한다. 해당 dataset은 model이 robust하게 task를 해결하려면 반드시 다뤄야 하는 common challenge(use of world knowledge and logical operator)에 초점을 맞추도록 설계되었다.

지금까지 나온 내용들을 정리하자면 다음과 같다.

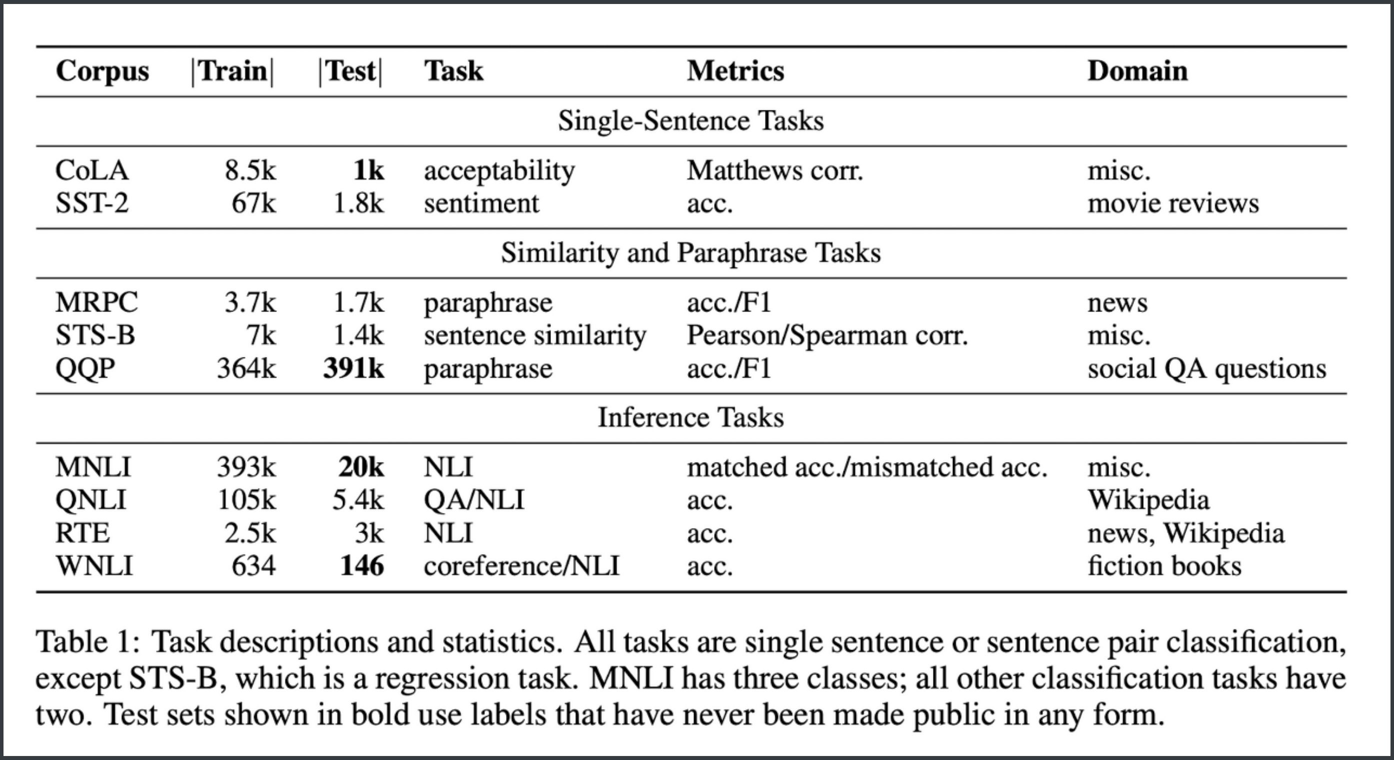
- **GLUE는 9개의 NLU task 모음이며, 각각의 task는 annotated dataset으로 구축되었으며 다양한 genre, dataset size, degrees of difficulty를 다루도록 선별되었다.**
- **privately-held test data를 기반으로 하는 online evaluation platform과 leaderboard가 존재하며, model-agnostic 하여 architecture의 제약을 두지 않는다.**
- **expert-constructed diagnostic evaluation dataset을 제공한다.**

Tasks

GLUE는 9개의 english sentence understanding task로 이루어져 있다. 해당 task들은 다양한 domain, data의 양, 난이도를 총망라한다. GLUE의 목적이 일반화 가능한 NLU system의 개발의 촉진이기 때문에, 저자들은 GLUE benchmark에서 model이 좋은 성능을 내려면 해당 **model이 약간의 task-specific 한 요소를 남겨놓는 상태에서 모든 task에 걸쳐 상당한 양의 knowledge(trained parameter)를 공유해야 하도록 설계하였다.**

GLUE에서 제공하는 task들에 대해, pre-training을 하지 않거나 별도의 외부 source를 이용한, 각각의 task를 위한 single model을 train 하여 GLUE benchmark를 통해 evaluate 하는 것도 가능하다. 그러나 저자들은 GLUE의 몇몇 task는 data의 개수가 제한된 data-scarce task이기 때문에, 이러한 task에서는 위와 같은 방법이 제 성능을 발휘하지 못할 것이라고 예측한다.

그러면 GLUE에서 제공하는 task들에는 어떤 것들이 있을까? task들의 대략적인 정보를 담고 있는 자료를 본 다음, 하나씩 살펴보도록 하겠다. GLUE의 task들은 다음과 같다.



Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

GLUE의 task는 크게 나눠보면 다음과 같이 나뉘게 되는데,

- Single-Sentence Tasks
- Similarity and Paraphrase Tasks
- Inference Task

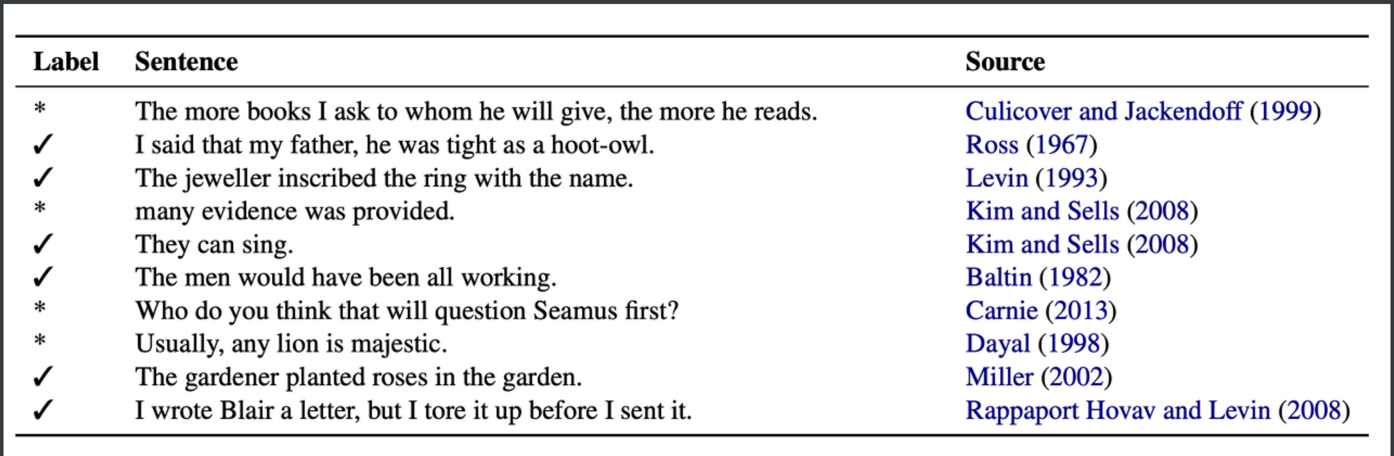
차례대로 하나씩 살펴보겠다.

Single-Sentence Tasks

CoLA

Corpus of Linguistic Acceptability(CoLA)는 언어 이론과 관련된 책과 저널 기사에서 추출된 english acceptability judgement로 이루어진 dataset이다. 각각의 example은 해당 영어 문장이 문법적인지 아닌지에 대해 annotated 된 sequence of words이다.

(아래의 도표는 CoLA의 example들이다.)



Label	Sentence	Source
*	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
✓	I said that my father, he was tight as a hoot-owl.	Ross (1967)
✓	The jeweller inscribed the ring with the name.	Levin (1993)
*	many evidence was provided.	Kim and Sells (2008)
✓	They can sing.	Kim and Sells (2008)
✓	The men would have been all working.	Baltin (1982)
*	Who do you think that will question Seamus first?	Carnie (2013)
*	Usually, any lion is majestic.	Dayal (1998)
✓	The gardener planted roses in the garden.	Miller (2002)
✓	I wrote Blair a letter, but I tore it up before I sent it.	Rappaport Hovav and Levin (2008)

Table 3: CoLA random sample, drawn from the in-domain training set (✓= acceptable, *=unacceptable).

CoLA에서는 evaluation metric으로 unbalanced binary classification에 대해 -1과 1 사이의 값(0은 uninformed guessing에 대한 성능을 나타냄)으로 evaluate 하는 Matthews correlation coefficient를 사용한다.

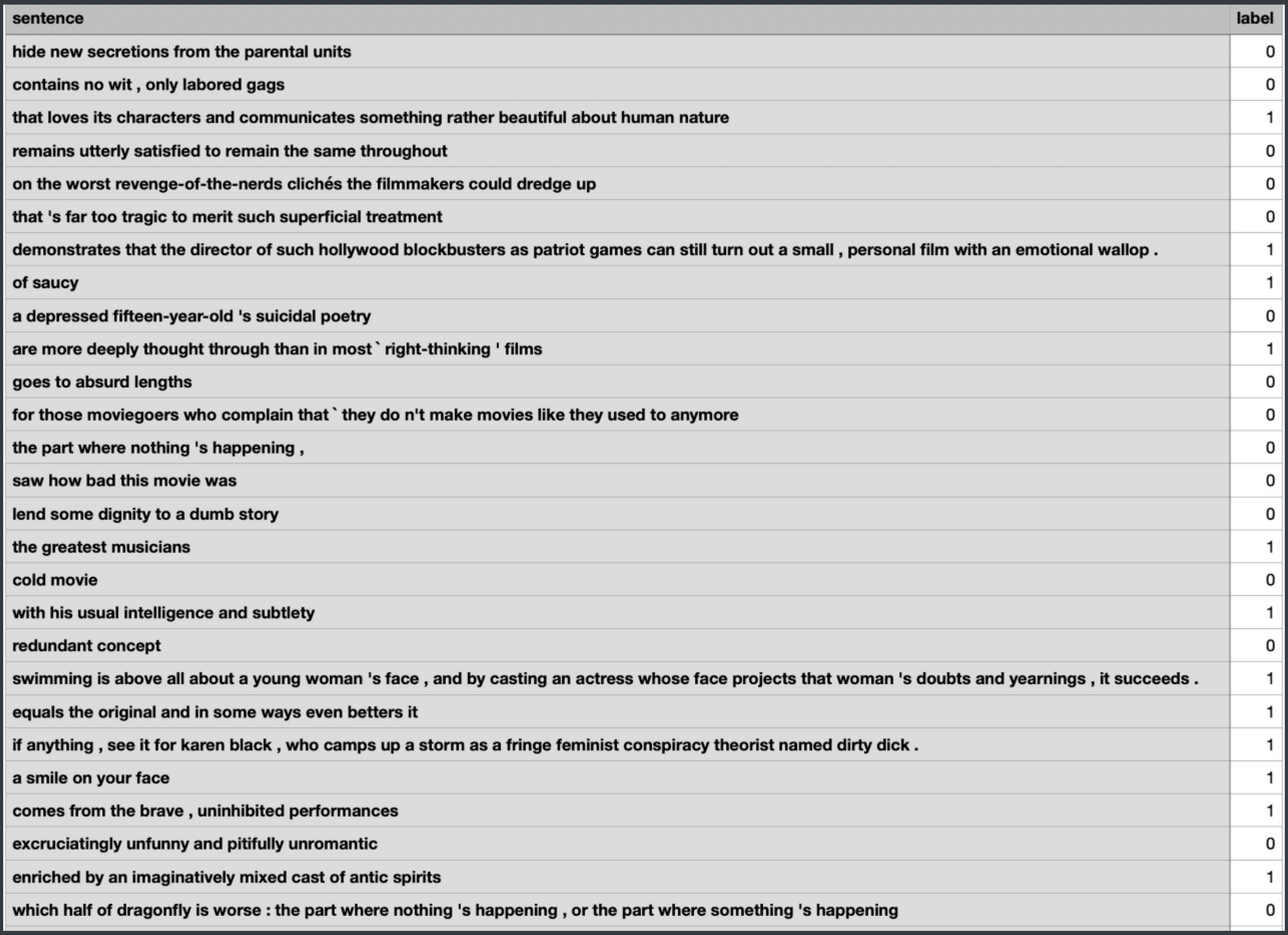
GLUE에서는 CoLA의 저자로부터 private label을 얻은 standard data set을 사용하며, test set에서 in-of-domain section과 out-of-domain section의 조합에서의 단일 성능 수치를 report 한다.

SST-2

Stanford Sentiment Treebank(SST-2)는 영화 리뷰 sentence와 해당 sentence에 대한 sentiment의 human annotation으로 이루어져 있다.

SST-2로 수행하는 task는 주어지는 sentence에 대해 sentiment를 예측하는 것이다. GLUE에서는 positive와 negative의 두 가지 class만 사용하며, sentence-level label만 사용한다.

SST-2의 예시는 다음과 같다.



sentence	label
hide new secretions from the parental units	0
contains no wit , only labored gags	0
that loves its characters and communicates something rather beautiful about human nature	1
remains utterly satisfied to remain the same throughout	0
on the worst revenge-of-the-nerds clichés the filmmakers could dredge up	0
that 's far too tragic to merit such superficial treatment	0
demonstrates that the director of such hollywood blockbusters as patriot games can still turn out a small , personal film with an emotional wallop .	1
of saucy	1
a depressed fifteen-year-old 's suicidal poetry	0
are more deeply thought through than in most ` right-thinking ' films	1
goes to absurd lengths	0
for those moviegoers who complain that ` they do n't make movies like they used to anymore	0
the part where nothing 's happening ,	0
saw how bad this movie was	0
lend some dignity to a dumb story	0
the greatest musicians	1
cold movie	0
with his usual intelligence and subtlety	1
redundant concept	0
swimming is above all about a young woman 's face , and by casting an actress whose face projects that woman 's doubts and yearnings , it succeeds .	1
equals the original and in some ways even betters it	1
if anything , see it for karen black , who camps up a storm as a fringe feminist conspiracy theorist named dirty dick .	1
a smile on your face	1
comes from the brave , uninhibited performances	1
excruciatingly unfunny and pitifully unromantic	0
enriched by an imaginatively mixed cast of antic spirits	1
which half of dragonfly is worse : the part where nothing 's happening , or the part where something 's happening	0

Similarity and Paraphrase Tasks

MRPC

Microsoft Research Paraphrase Corpus(MRPC)는 온라인 뉴스로부터 추출된 sentence pair와, 해당 sentence pair를 구성하는 sentence가 의미상으로 같은 sentence인지에 대한 human annotation으로 이루어져 있다.

MRPC는 imbalanced data이기 때문에(68% positive), accuracy와 f1 score 둘 다 사용하여 report 한다.

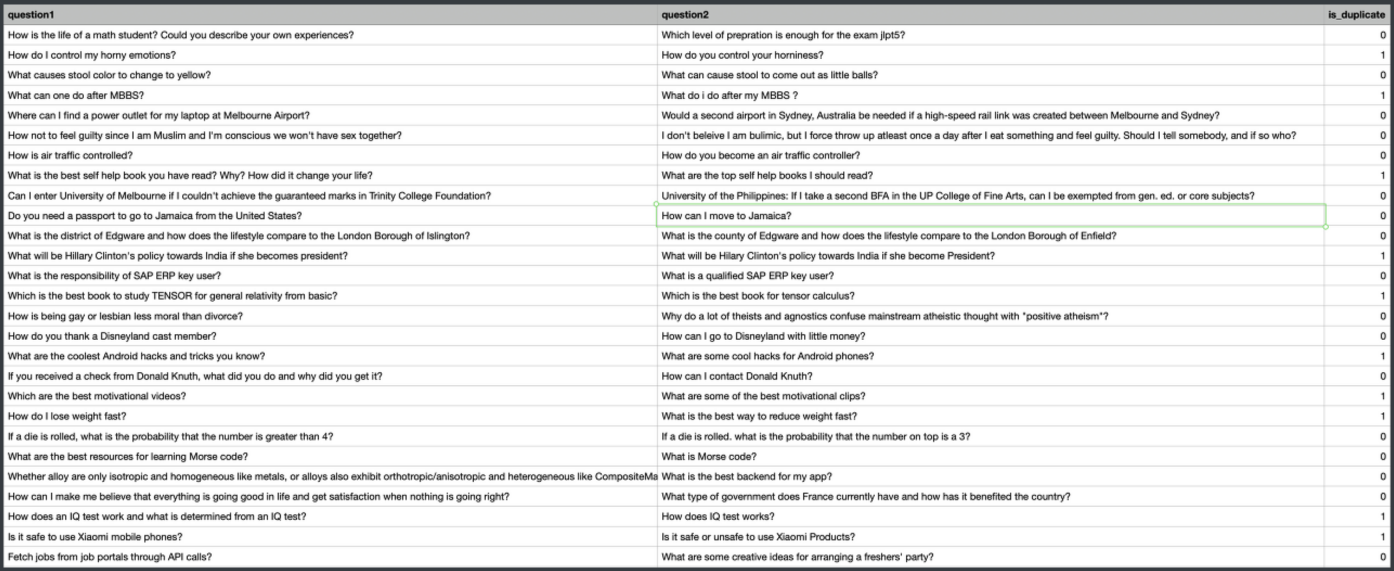
QQP

Quora Question Pairs(QQP)는 Question answering 웹사이트인 Quora에서 추출한 question pair의 모음이다.

MRPC의 경우처럼, sentence가 의미상으로 같은 sentence인지에 대한 human annotation으로 이루어져 있고 imbalanced data이기 때문에 (63% negative) accuracy와 f1 score 모두 사용하여 report 한다.

또한, GLUE에서는 QQP의 원작자로부터 받은, private label을 가진 standard test set을 사용하며, 이러한 test set은 training set과는 다른 label distribution을 보인다.

QQP의 예시는 다음과 같다.



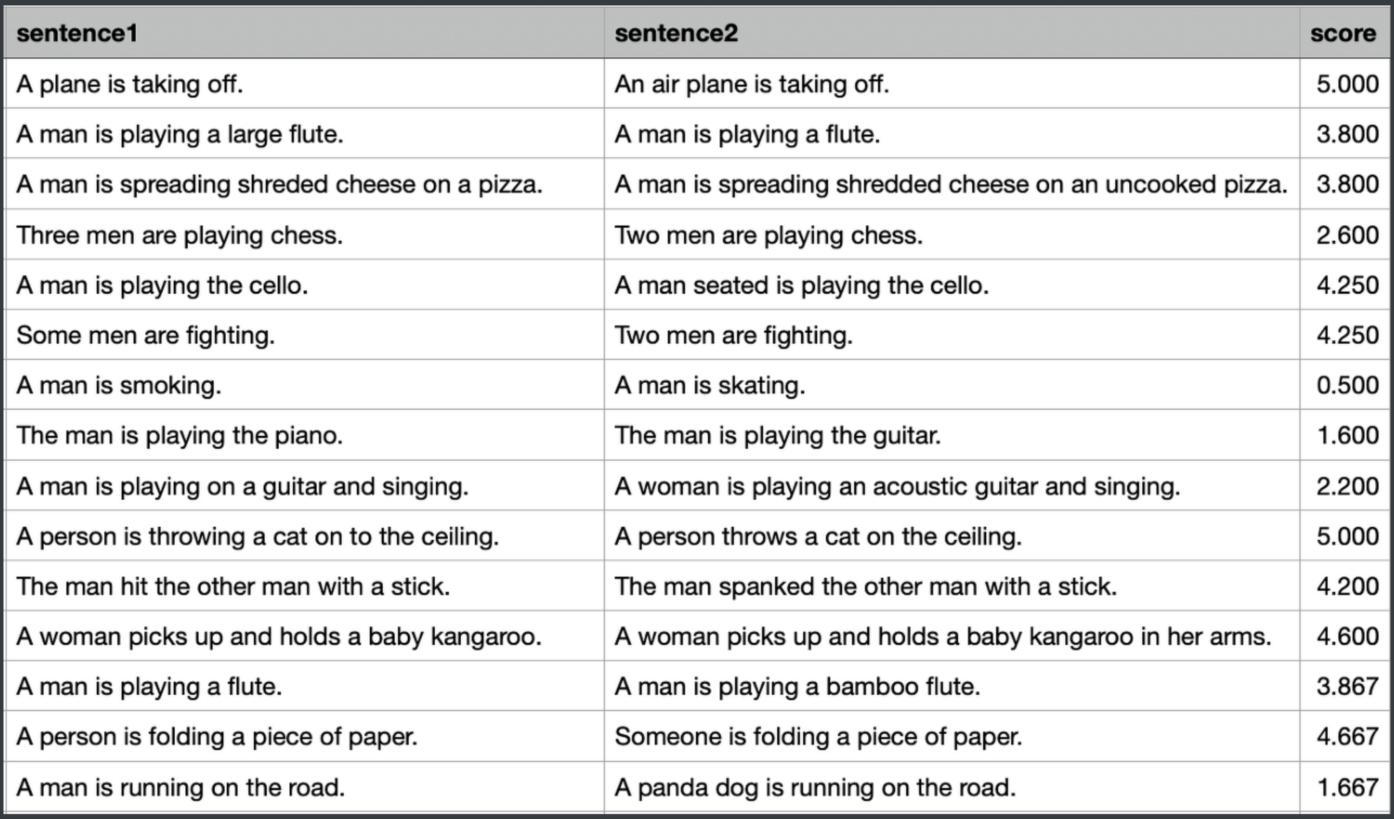
question1	question2	is_duplicate
How is the life of a math student? Could you describe your own experiences?	Which level of prepration is enough for the exam jlpt5?	0
How do I control my horny emotions?	How do you control your horniness?	1
What causes stool color to change to yellow?	What can cause stool to come out as little balls?	0
What can one do after MBBS?	What do i do after my MBBS ?	1
Where can I find a power outlet for my laptop at Melbourne Airport?	Would a second airport in Sydney, Australia be needed if a high-speed rail link was created between Melbourne and Sydney?	0
How not to feel guilty since I am Muslim and I'm conscious we won't have sex together?	I don't beleive I am bulimic, but I force throw up atleast once a day after I eat something and feel guilty. Should I tell somebody, and if so who?	0
How is air traffic controlled?	How do you become an air traffic controller?	0
What is the best self help book you have read? Why? How did it change your life?	What are the top self help books I should read?	1
Can I enter University of Melbourne if I couldn't achieve the guaranteed marks in Trinity College Foundation?	University of the Philippines: If I take a second BFA in the UP College of Fine Arts, can I be exempted from gen. ed. or core subjects?	0
Do you need a passport to go to Jamaica from the United States?	How can I move to Jamaica?	0
What is the district of Edgware and how does the lifestyle compare to the London Borough of Islington?	What is the county of Edgware and how does the lifestyle compare to the London Borough of Enfield?	0
What will be Hillary Clinton's policy towards India if she becomes president?	What will be Hilary Clinton's policy towards India if she become President?	1
What is the responsibility of SAP ERP key user?	What is a qualified SAP ERP key user?	0
Which is the best book to study TENSOR for general relativity from basic?	Which is the best book for tensor calculus?	1
How is being gay or lesbian less moral than divorce?	Why do a lot of theists and agnostics confuse mainstream atheistic thought with "positive atheism"?	0
How do you thank a Disneyland cast member?	How can I go to Disneyland with little money?	0
What are the coolest Android hacks and tricks you know?	What are some cool hacks for Android phones?	1
If you received a check from Donald Knuth, what did you do and why did you get it?	How can I contact Donald Knuth?	0
Which are the best motivational videos?	What are some of the best motivational clips?	1
How do I lose weight fast?	What is the best way to reduce weight fast?	1
If a die is rolled, what is the probability that the number is greater than 4?	If a die is rolled. what is the probability that the number on top is a 3?	0
What are the best resources for learning Morse code?	What is Morse code?	0
Whether alloy are only isotropic and homogeneous like metals, or alloys also exhibit orthotropic/anisotropic and heterogeneous like CompositeMa	What is the best backend for my app?	0
How can I make me believe that everything is going good in life and get satisfaction when nothing is going right?	What type of government does France currently have and how has it benefited the country?	0
How does an IQ test work and what is determined from an IQ test?	How does IQ test works?	1
Is it safe to use Xiaomi mobile phones?	Is it safe or unsafe to use Xiaomi Products?	1
Fetch jobs from job portals through API calls?	What are some creative ideas for arranging a freshers' party?	0

STS-B

Semantic Textual Similarity Benchmark(STS-B)는 news headline, video and image captions, natural language inference data에서 추출한 sentence pair의 모음이다.

각각의 pair에 대해서 1부터 5까지의 유사도 점수를 매긴 human-annotated label이 대응되며, STS-B를 통해 진행하는 task는 해당 유사도 점수를 predict 하는 것이다. 또한 이를 Pearson and Spearman correlation coefficients로 evaluate 한다.

예시는 다음과 같다.



sentence1	sentence2	score
A plane is taking off.	An air plane is taking off.	5.000
A man is playing a large flute.	A man is playing a flute.	3.800
A man is spreading shreded cheese on a pizza.	A man is spreading shredded cheese on an uncooked pizza.	3.800
Three men are playing chess.	Two men are playing chess.	2.600
A man is playing the cello.	A man seated is playing the cello.	4.250
Some men are fighting.	Two men are fighting.	4.250
A man is smoking.	A man is skating.	0.500
The man is playing the piano.	The man is playing the guitar.	1.600
A man is playing on a guitar and singing.	A woman is playing an acoustic guitar and singing.	2.200
A person is throwing a cat on to the ceiling.	A person throws a cat on the ceiling.	5.000
The man hit the other man with a stick.	The man spanked the other man with a stick.	4.200
A woman picks up and holds a baby kangaroo.	A woman picks up and holds a baby kangaroo in her arms.	4.600
A man is playing a flute.	A man is playing a bamboo flute.	3.867
A person is folding a piece of paper.	Someone is folding a piece of paper.	4.667
A man is running on the road.	A panda dog is running on the road.	1.667

Inference Task

MNLI

Muti-Genre Natural Language Inference Corpus는 crowd-sourcing으로 구축된 textual entailment annotation이 있는 sentence of pair이다.

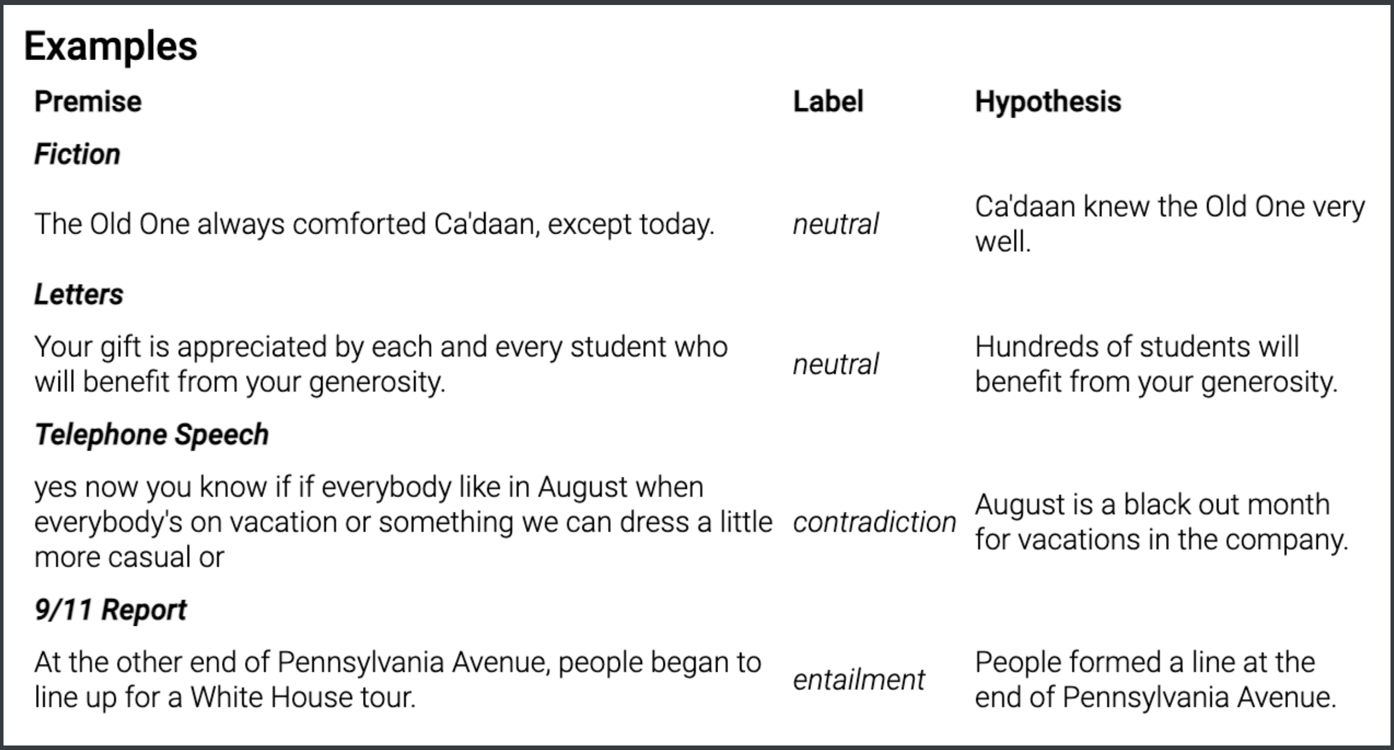
MNLI로 수행하는 task인 textual entailment는, premise(전제)와 hypothesis(가설)가 주어지며, premise(전제)가 hypothesis(가설)를 entail 하는지 predict 하는 task이다.

premise(전제)가 hypothesis(가설)를 entail 하면 entailment, contradict 하면 contradiction, 둘 다 아니면 neutral의 label로 predict 한다.

(여기서의 entail은 "함의"의 뜻이며, premise가 사실일 때 hypothesis의 사실을 보장하는 것을 의미한다. contradict의 경우에는 premise와 hypothesis가 서로 모순된다는 의미이다.)

Premise sentence들은 transcribed speech, fiction, government report를 포함한 10개의 다른 source로부터 수집되었다. 이 논문에서는 나오지 않지만, hypothesis를 구축할 때에는 수집된 premise를 바탕으로 crowd-sourcing을 진행하였다.

아래는 MNLI dataset의 예시이다.



Examples

Premise	Label	Hypothesis
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

QNLI

Question-answering NLI(QNLI)는 Stanford Question Answering Dataset(SQuAd)를 변형시킨 dataset이다.

우선, Stanford Question Answering Dataset(SQuAd)에 대해 알아보자. SQuAD는 question-paragraph의 pair로 구성된 question-answering dataset이다. paragraph는 자기 자신과 대응되는 question에 대한 answer를 포함하고 있다.

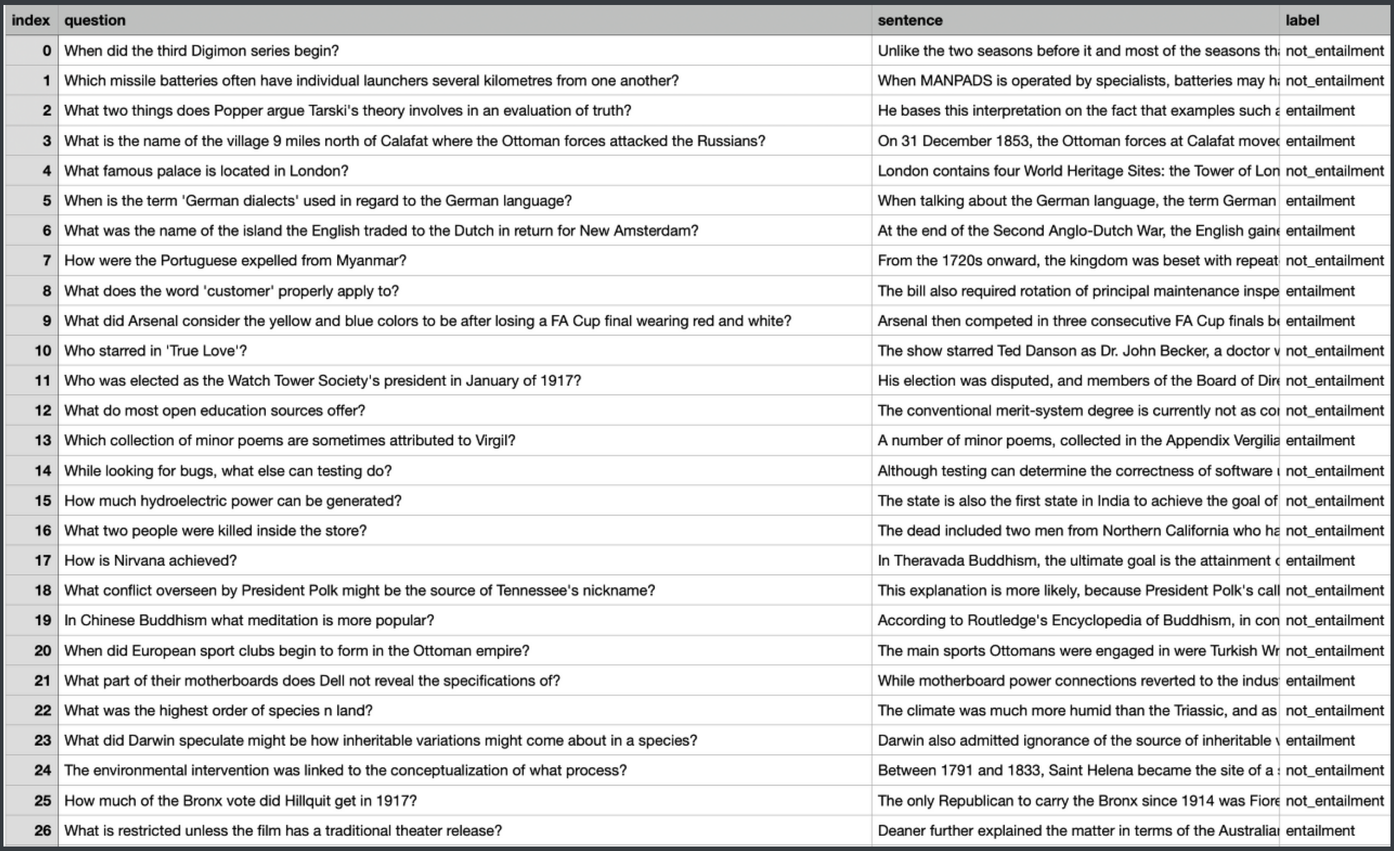
(paragraph는 wikipedia로부터 추출되었으며, question과 answer는 crowd-sourcing을 통해 생성되었다.)

SQuAD를 통해 수행하는 task는 QA task인데, QNLI는 SQuAD dataset을 변환하여 sentence pair classification으로 변환한다. 먼저, question과 이에 대응하는 paragraph를 이용하여 pair of sentence를 형성하고, 이렇게 형성된 pair of sentence들 중에서 lexical overlap(어휘적 중복)이 낮은 쌍을 필터링한다. (어휘적 중복이 낮은 쌍을 제거하는 것이 아닌, 중복이 낮은 쌍을 선별한다)

이러한 수정은 model이 옳은 answer를 선택할 필요를 제거하지만 **answer가 input속에 항상 있을 것이라는, 또한 lexical overlap(어휘적 중복)이 신뢰할 수 있는 단서일 것이라는 "단순화된 추정"도 제거해주는 효과가 있다.**

이후, 수정된 QNLI dataset을 통해서 paragraph가 question에 대한 answer를 포함하고 있는지에 대한 classification task를 수행하게 된다.

다음은 QNLI의 예시이다.



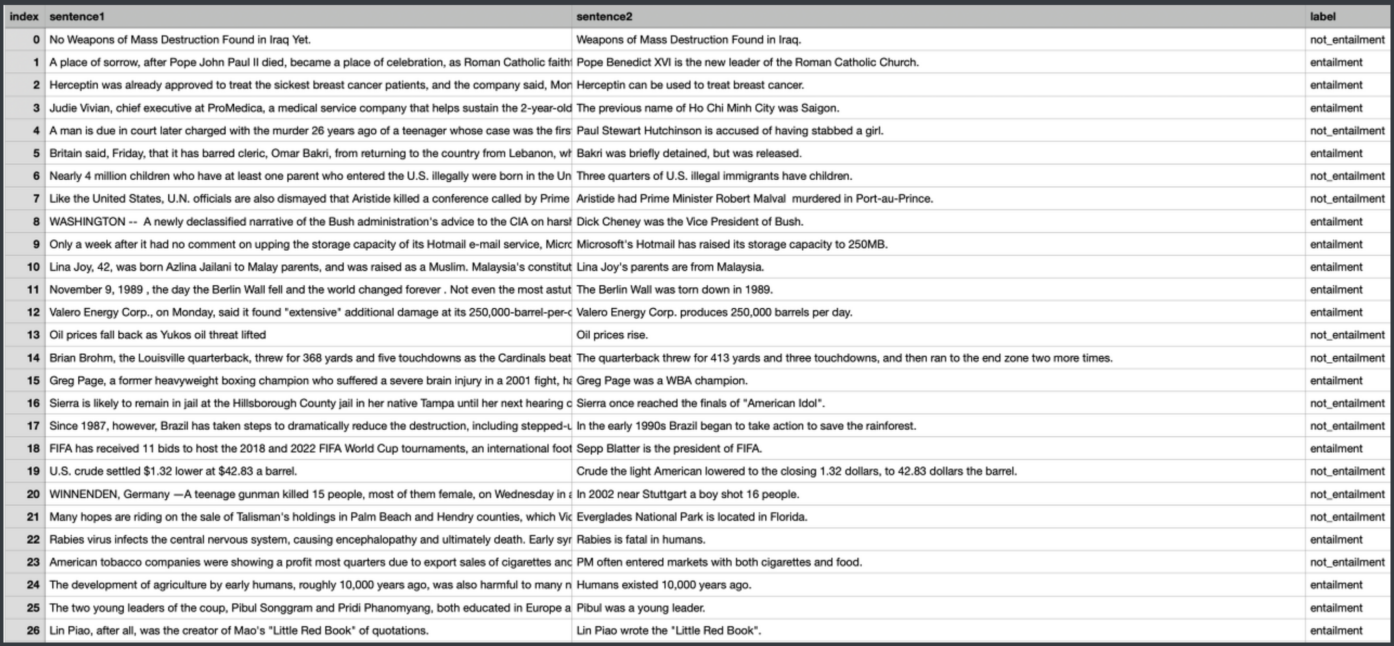
index	question	sentence	label
0	When did the third Digimon series begin?	Unlike the two seasons before it and most of the seasons th	not_entailment
1	Which missile batteries often have individual launchers several kilometres from one another?	When MANPADS is operated by specialists, batteries may h	not_entailment
2	What two things does Popper argue Tarski's theory involves in an evaluation of truth?	He bases this interpretation on the fact that examples such a	entailment
3	What is the name of the village 9 miles north of Calafat where the Ottoman forces attacked the Russians?	On 31 December 1853, the Ottoman forces at Calafat movec	entailment
4	What famous palace is located in London?	London contains four World Heritage Sites: the Tower of Lon	not_entailment
5	When is the term 'German dialects' used in regard to the German language?	When talking about the German language, the term German	entailment
6	What was the name of the island the English traded to the Dutch in return for New Amsterdam?	At the end of the Second Anglo-Dutch War, the English gaine	entailment
7	How were the Portuguese expelled from Myanmar?	From the 1720s onward, the kingdom was beset with repeat	not_entailment
8	What does the word 'customer' properly apply to?	The bill also required rotation of principal maintenance inspe	entailment
9	What did Arsenal consider the yellow and blue colors to be after losing a FA Cup final wearing red and white?	Arsenal then competed in three consecutive FA Cup finals be	entailment
10	Who starred in 'True Love'?	The show starred Ted Danson as Dr. John Becker, a doctor v	not_entailment
11	Who was elected as the Watch Tower Society's president in January of 1917?	His election was disputed, and members of the Board of Dire	not_entailment
12	What do most open education sources offer?	The conventional merit-system degree is currently not as co	not_entailment
13	Which collection of minor poems are sometimes attributed to Virgil?	A number of minor poems, collected in the Appendix Vergilia	entailment
14	While looking for bugs, what else can testing do?	Although testing can determine the correctness of software	not_entailment
15	How much hydroelectric power can be generated?	The state is also the first state in India to achieve the goal of	not_entailment
16	What two people were killed inside the store?	The dead included two men from Northern California who ha	not_entailment
17	How is Nirvana achieved?	In Theravada Buddhism, the ultimate goal is the attainment c	entailment
18	What conflict overseen by President Polk might be the source of Tennessee's nickname?	This explanation is more likely, because President Polk's call	not_entailment
19	In Chinese Buddhism what meditation is more popular?	According to Routledge's Encyclopedia of Buddhism, in con	not_entailment
20	When did European sport clubs begin to form in the Ottoman empire?	The main sports Ottomans were engaged in were Turkish Wr	not_entailment
21	What part of their motherboards does Dell not reveal the specifications of?	While motherboard power connections reverted to the indus	entailment
22	What was the highest order of species n land?	The climate was much more humid than the Triassic, and as	not_entailment
23	What did Darwin speculate might be how inheritable variations might come about in a species?	Darwin also admitted ignorance of the source of inheritable v	entailment
24	The environmental intervention was linked to the conceptualization of what process?	Between 1791 and 1833, Saint Helena became the site of a	not_entailment
25	How much of the Bronx vote did Hillquit get in 1917?	The only Republican to carry the Bronx since 1914 was Fiore	not_entailment
26	What is restricted unless the film has a traditional theater release?	Deaner further explained the matter in terms of the Australiar	entailment

RTE

Recognizing Texual Entailment(RTE) dataset은 annual textual entailment challenges로부터 추출되었다. 본 연구에서는 RTE1, RTE2, RTE3, RTE5를 혼합하였다고 밝힌다. (이 dataset들은 뉴스와 wikipedia로부터 구축되었다.)

또한 원문 dataset인 RTE들은 위에서 소개한 MNLI와 같이 entailment, neural, contradiction의 3개 class로 구분되어있었는데, GLUE에서의 RTE는 neutral과 contradiction을 not_entailment로 묶어 entailment와 not_entailment의 2개의 class로 수정하였다.

RTE의 예시는 다음과 같다.



index	sentence1	sentence2	label
0	No Weapons of Mass Destruction Found in Iraq Yet.	Weapons of Mass Destruction Found in Iraq.	not_entailment
1	A place of sorrow, after Pope John Paul II died, became a place of celebration, as Roman Catholic faith	Pope Benedict XVI is the new leader of the Roman Catholic Church.	entailment
2	Herceptin was already approved to treat the sickest breast cancer patients, and the company said, Mon	Herceptin can be used to treat breast cancer.	entailment
3	Judie Vivian, chief executive at ProMedica, a medical service company that helps sustain the 2-year-old	The previous name of Ho Chi Minh City was Saigon.	entailment
4	A man is due in court later charged with the murder 26 years ago of a teenager whose case was the firs	Paul Stewart Hutchinson is accused of having stabbed a girl.	not_entailment
5	Britain said, Friday, that it has barred cleric, Omar Bakri, from returning to the country from Lebanon, wh	Bakri was briefly detained, but was released.	entailment
6	Nearly 4 million children who have at least one parent who entered the U.S. illegally were born in the Un	Three quarters of U.S. illegal immigrants have children.	not_entailment
7	Like the United States, U.N. officials are also dismayed that Aristide killed a conference called by Prime	Aristide had Prime Minister Robert Malval murdered in Port-au-Prince.	not_entailment
8	WASHINGTON -- A newly declassified narrative of the Bush administration's advice to the CIA on harsh	Dick Cheney was the Vice President of Bush.	entailment
9	Only a week after it had no comment on upping the storage capacity of its Hotmail e-mail service, Micro	Microsoft's Hotmail has raised its storage capacity to 250MB.	entailment
10	Lina Joy, 42, was born Azlina Jailani to Malay parents, and was raised as a Muslim. Malaysia's constitut	Lina Joy's parents are from Malaysia.	entailment
11	November 9, 1989 , the day the Berlin Wall fell and the world changed forever . Not even the most astut	The Berlin Wall was torn down in 1989.	entailment
12	Valero Energy Corp., on Monday, said it found "extensive" additional damage at its 250,000-barrel-per-c	Valero Energy Corp. produces 250,000 barrels per day.	entailment
13	Oil prices fall back as Yukos oil threat lifted	Oil prices rise.	not_entailment
14	Brian Brohm, the Louisville quarterback, threw for 368 yards and five touchdowns as the Cardinals beat	The quarterback threw for 413 yards and three touchdowns, and then ran to the end zone two more times.	not_entailment
15	Greg Page, a former heavyweight boxing champion who suffered a severe brain injury in a 2001 fight, ha	Greg Page was a WBA champion.	entailment
16	Sierra is likely to remain in jail at the Hillsborough County jail in her native Tampa until her next hearing c	Sierra once reached the finals of "American Idol".	not_entailment
17	Since 1987, however, Brazil has taken steps to dramatically reduce the destruction, including stepped-u	In the early 1990s Brazil began to take action to save the rainforest.	not_entailment
18	FIFA has received 11 bids to host the 2018 and 2022 FIFA World Cup tournaments, an international foot	Sepp Blatter is the president of FIFA.	entailment
19	U.S. crude settled $1.32 lower at $42.83 a barrel.	Crude the light American lowered to the closing 1.32 dollars, to 42.83 dollars the barrel.	not_entailment
20	WINNENDEN, Germany —A teenage gunman killed 15 people, most of them female, on Wednesday in a	In 2002 near Stuttgart a boy shot 16 people.	not_entailment
21	Many hopes are riding on the sale of Talisman's holdings in Palm Beach and Hendry counties, which Vic	Everglades National Park is located in Florida.	not_entailment
22	Rabies virus infects the central nervous system, causing encephalopathy and ultimately death. Early syr	Rabies is fatal in humans.	entailment
23	American tobacco companies were showing a profit most quarters due to export sales of cigarettes anc	PM often entered markets with both cigarettes and food.	not_entailment
24	The development of agriculture by early humans, roughly 10,000 years ago, was also harmful to many n	Humans existed 10,000 years ago.	entailment
25	The two young leaders of the coup, Pibul Songgram and Pridi Phanomyang, both educated in Europe a	Pibul was a young leader.	entailment
26	Lin Piao, after all, was the creator of Mao's "Little Red Book" of quotations.	Lin Piao wrote the "Little Red Book".	entailment

WNLI

Winograd NLI(WNLI)는 Winograd Schema Challenge의 변형이다. (QNLI와 비슷한 느낌이다.)

Winograd Schema Challenge는 reading comprehension task이다. 이 reading comprehension task란, model이 pronoun(대명사)가 있는 문장을 읽고 해당 pronoun(대명사)의 referent(참조)를 선택 목록 중에서 선택해야 하는 task인 것이다.

이러한 reading comprehension task를 sentence pair classification task로 변환한다. 이를 위해 각 문장의 pronoun을 가능한 각각의 referent로 바꾸고, 원본 sentence와 함께 sentence of pair를 만든다.

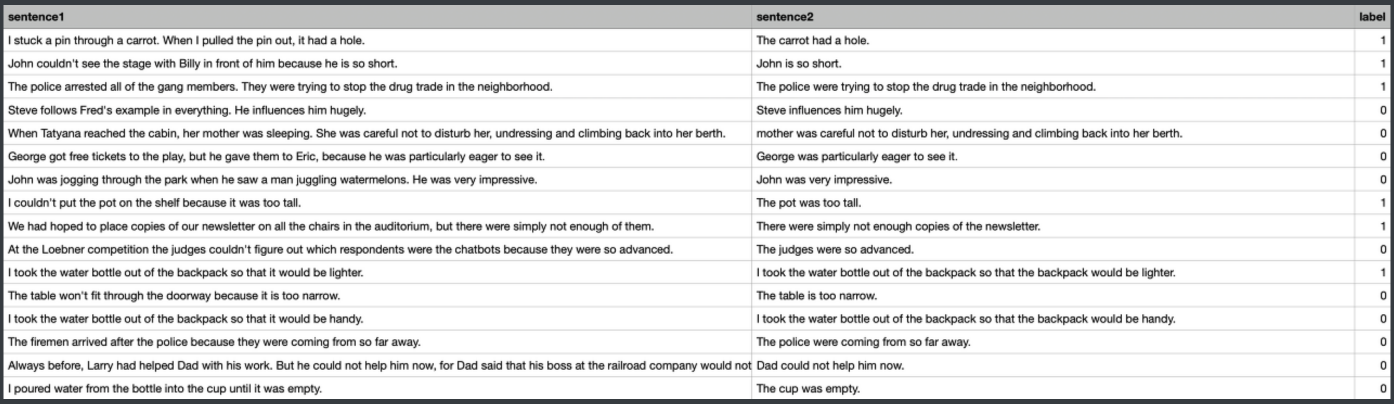
이후, 원본 sentence가 pronoun이 대체된 문장을 entail 하는지 predict 하는 task를 수행한다. (이때, entailment, not_entailment의 이진 분류이다.)

GLUE에서 WNLI dataset은 original corpus의 저자들이 privately 하게 공유한 small evaluation set을 사용한다.

WNLI의 training set이 balanced data인 것에 비해, test set은 imbalanced data이다. 또한, data quirk로 인해 development set은 adversarial(적대적)하여 간혹 training example과 development example이 공유되는 경우가 발생하는데, 이때 만약 model이 training example을 기억하고 있다면 development example에서는 잘못된 label을 예측하게 된다.

(즉, training set에 overfitting 되어있으면 development set에서는 성능이 크게 하락할 수 있다.)

아래는 WNLI의 예시이다.



sentence1	sentence2	label
I stuck a pin through a carrot. When I pulled the pin out, it had a hole.	The carrot had a hole.	1
John couldn't see the stage with Billy in front of him because he is so short.	John is so short.	1
The police arrested all of the gang members. They were trying to stop the drug trade in the neighborhood.	The police were trying to stop the drug trade in the neighborhood.	1
Steve follows Fred's example in everything. He influences him hugely.	Steve influences him hugely.	0
When Tatyana reached the cabin, her mother was sleeping. She was careful not to disturb her, undressing and climbing back into her berth.	mother was careful not to disturb her, undressing and climbing back into her berth.	0
George got free tickets to the play, but he gave them to Eric, because he was particularly eager to see it.	George was particularly eager to see it.	0
John was jogging through the park when he saw a man juggling watermelons. He was very impressive.	John was very impressive.	0
I couldn't put the pot on the shelf because it was too tall.	The pot was too tall.	1
We had hoped to place copies of our newsletter on all the chairs in the auditorium, but there were simply not enough of them.	There were simply not enough copies of the newsletter.	1
At the Loebner competition the judges couldn't figure out which respondents were the chatbots because they were so advanced.	The judges were so advanced.	0
I took the water bottle out of the backpack so that it would be lighter.	I took the water bottle out of the backpack so that the backpack would be lighter.	1
The table won't fit through the doorway because it is too narrow.	The table is too narrow.	0
I took the water bottle out of the backpack so that it would be handy.	I took the water bottle out of the backpack so that the backpack would be handy.	0
The firemen arrived after the police because they were coming from so far away.	The police were coming from so far away.	0
Always before, Larry had helped Dad with his work. But he could not help him now, for Dad said that his boss at the railroad company would not	Dad could not help him now.	0
I poured water from the bottle into the cup until it was empty.	The cup was empty.	0

Evaluation

model을 GLUE benchmark로 evaluation 하기 위해서는 제공되는 task에 대한 test data에 대해 model을 통해 나오는 결과를 GLUE 웹사이트에 업로드하여 점수를 책정받아야 한다. (이는 Kaggle이나 Dacon을 생각해보면 쉽게 이해할 수 있다.)

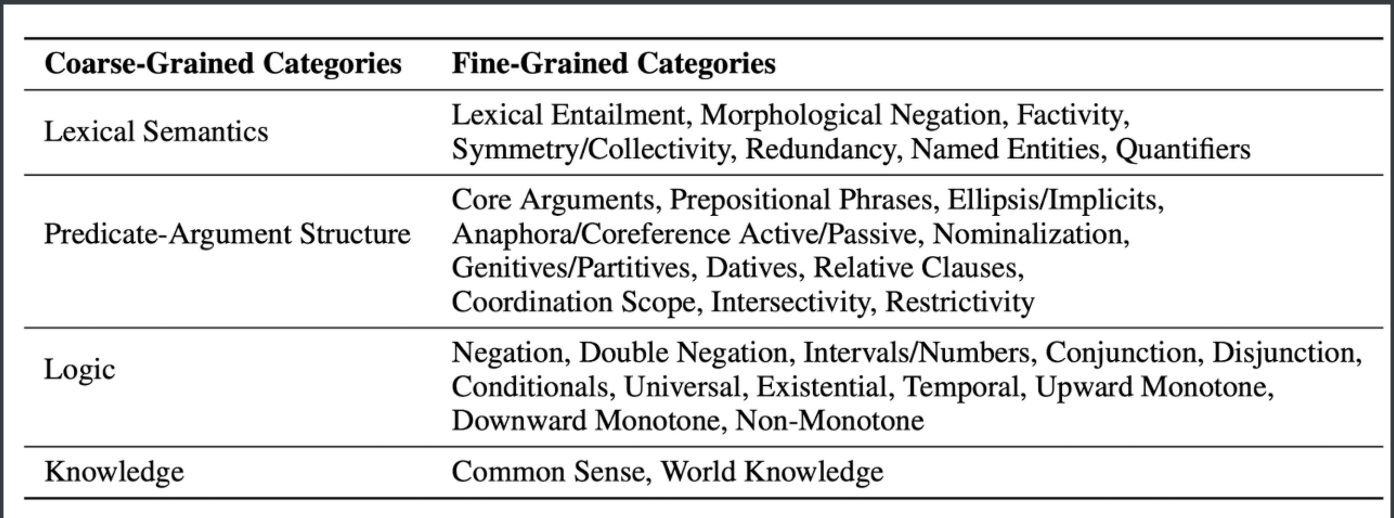
GLUE benchmark 웹사이트는 task별 score와 해당 task들의 macro-average score를 보여주고, 리더보드 상에서의 순위를 결정한다. accuarcy와 f1 score를 같이 사용하는 task의 경우, 전체 task의 macro-average score를 구할 때에는 task에 대한 두 metric의 unweighted average를 해당 task의 score로 계산한다.

추가적으로, GLUE benchmark 웹사이트에서는 diagnostic dataset의 fine-grained result와 coarse-grained result도 제공한다.

Diagnostic Dataset

GLUE에서는 model의 성능 분석을 위해 manually-curated 된 small test set을 제공한다. Main benchmark가 application 중심의 distribution of example을 반영하는 것과는 달리, diagnostic dataset은 저자들이 model이 capture 하기에 중요하고 흥미롭다고 생각하는 pre-defined set of phenomena에 중점을 둔다.

pre-defined set of phenomena은 다음과 같다.



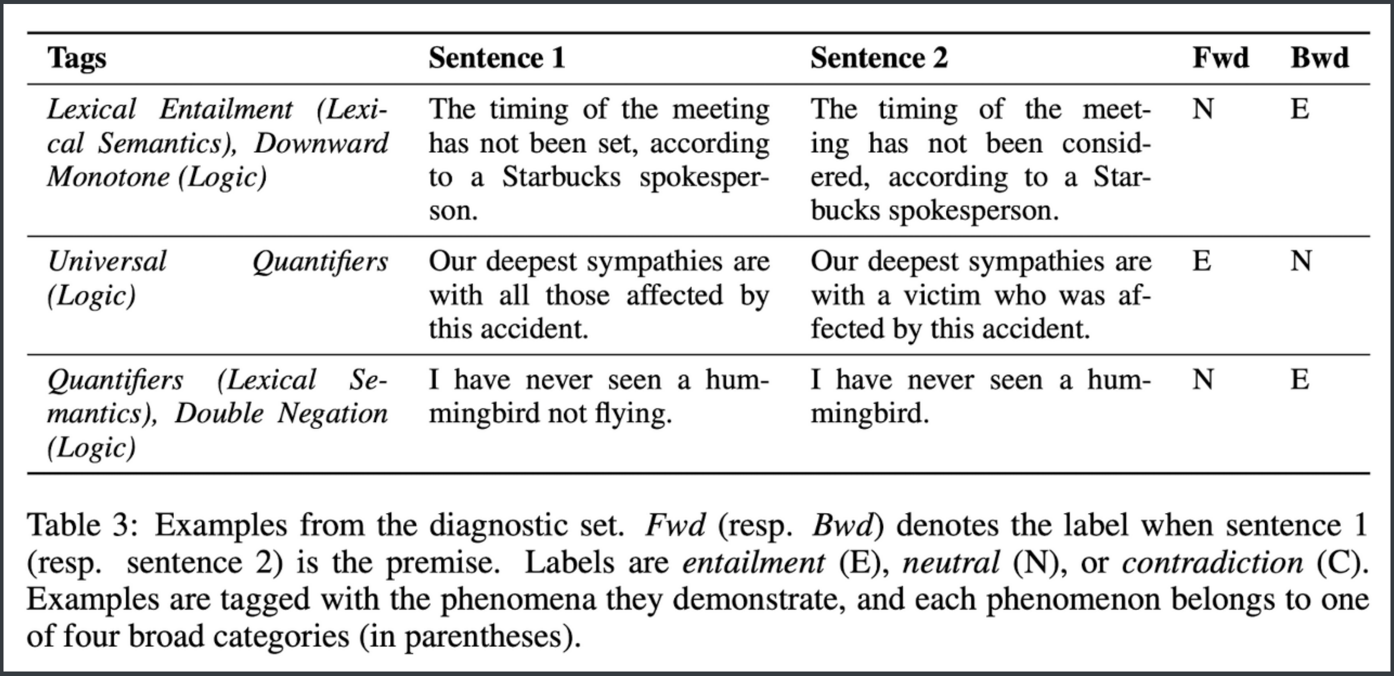
Coarse-Grained Categories	Fine-Grained Categories
Lexical Semantics	Lexical Entailment, Morphological Negation, Factivity, Symmetry/Collectivity, Redundancy, Named Entities, Quantifiers
Predicate-Argument Structure	Core Arguments, Prepositional Phrases, Ellipsis/Implicits, Anaphora/Coreference Active/Passive, Nominalization, Genitives/Partitives, Datives, Relative Clauses, Coordination Scope, Intersectivity, Restrictivity
Logic	Negation, Double Negation, Intervals/Numbers, Conjunction, Disjunction, Conditionals, Universal, Existential, Temporal, Upward Monotone, Downward Monotone, Non-Monotone
Knowledge	Common Sense, World Knowledge

Table 2: The types of linguistic phenomena annotated in the diagnostic dataset, organized under four major categories. For a description of each phenomenon, see Appendix E.

각각의 Diagnostic example은 NLI(Natural Language Inference) sentence pair와 증명된 phenomena로 이루어진다.

또한 저자들은 이러한 diagnostic dataset이 다양한 linguistic phenomena에 대한 예제를 생성하였고, 여러 domain에서 자연스럽게 발생하는 sentence들을 기반으로 하였기 때문에 합리적으로 다양한 dataset이라고 주장한다.

dianostic example의 예시는 다음과 같다.



Tags	Sentence 1	Sentence 2	Fwd	Bwd
Lexical Entailment (Lexical Semantics), Downward Monotone (Logic)	The timing of the meeting has not been set, according to a Starbucks spokesperson.	The timing of the meeting has not been considered, according to a Starbucks spokesperson.	N	E
Universal Quantifiers (Logic)	Our deepest sympathies are with all those affected by this accident.	Our deepest sympathies are with a victim who was affected by this accident.	E	N
Quantifiers (Lexical Semantics), Double Negation (Logic)	I have never seen a hummingbird not flying.	I have never seen a hummingbird.	N	E

Table 3: Examples from the diagnostic set. *Fwd* (resp. *Bwd*) denotes the label when sentence 1 (resp. sentence 2) is the premise. Labels are *entailment* (E), *neutral* (N), or *contradiction* (C). Examples are tagged with the phenomena they demonstrate, and each phenomenon belongs to one of four broad categories (in parentheses).

이에 대해 간단히 설명하자면, 우선 위에서 언급한 것처럼 NLI sentenc pair와 tag로 구성되어있고, 그 밖에 Fwd와 Bwd가 보인다.

이에 대해 설명을 하자면, Fwd는 sentence 1이 premise이고 sentence 2가 hypothesis일 경우의 label이고, Bwd는 그와 반대로 sentence 2가 premise이고 sentence 1이 hypothesis인 경우이다. 또한, label은 E(entailment), N(neutral) 그리고 C(contradiction)로 구성된다

저자들은 diagnostic set을 model의 전반적인 성능이나 downstream application에서의 generalization을 반영할 것으로 생각하지 않는다고 말한다. 그러면서, diagnostic set은 benchmark가 아닌 error analysis, qualitative model comparison, development of adversarial examples를 위한 분석 도구라고 덧붙인다.