GLUE

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

집현전 17조
류지은 이기창 이노아

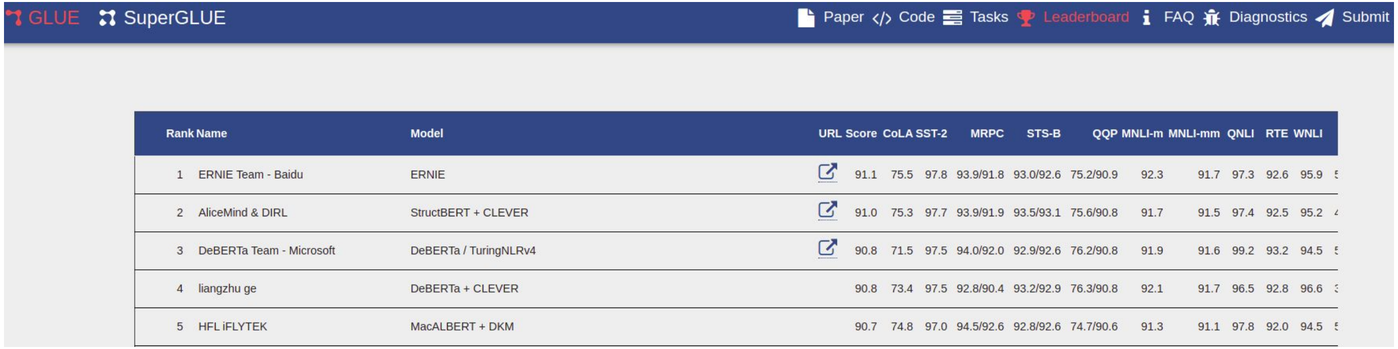
The idea of General Language Understanding Evaluation (GLUE)

보편적이고 유연하며 강건한, 사람의 언어 이해 능력과 달리, 많은 머신러닝 모델들은 **특정** 태스크에 대해서만 트레이닝되었고 다른 데이터에 대해선 어려워했다.

이에 GLUE(글루)를 소개한다. GLUE는 해당 모델의 퍼포먼스를, 현존하는 **여러** 자연어 이해, Natural Language Understanding (NLU) 태스크들에 대해 체크하는 데이터셋 모음이다.

GLUE benchmark link : https://gluebenchmark.com/

- Question answering, Sentiment analysis, and Textual entailment 포함 여러 NLU 태스크 평가 가능
- 모델 평가, 비교, 그리고 분석을 할 수 있는 온라인 플랫폼
- 아홉개의 기존 다양한 데이터셋들의 모음
- world knowledge와 logical operators같은 보편적인 챌린지가 포함된, 직접 만든 분석 툴도 포함



Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	
1	ERNIE Team - Baidu	ERNIE	↗	91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	5
2	AliceMind & DIRL	StructBERT + CLEVER	↗	91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	4
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	5
4	liangzhu ge	DeBERTa + CLEVER		90.8	73.4	97.5	92.8/90.4	93.2/92.9	76.3/90.8	92.1	91.7	96.5	92.8	96.6	3
5	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	5

Related Work

Multi-task learning

Collobert et al. (2011)은 POS tagging, chunking, NER, 그리고 SRL을 같이 배우게 하려고 함
문장 이해를 위한 shared component와 함께 multi-task 모델을 사용했다

Multi-task learning with cross-task sharing mechanisms

2017년도에는 핵심 NLP 태스크들로부터 label을 사용해서 탐색한다.
그리고 자동으로 multi-task 학습을 시키기 위한 cross-task 공유 매커니즘을 배운다.

Development of general NLU systems

일반적인 NLU 시스템 개발을 위한 많은 작업들은 sentence-to-vector와 레이블이 없는 데이터의 활용,
그리고 레이블된 데이터, 또는 이들의 조합에 집중하고 있다.

Standard evaluation practice

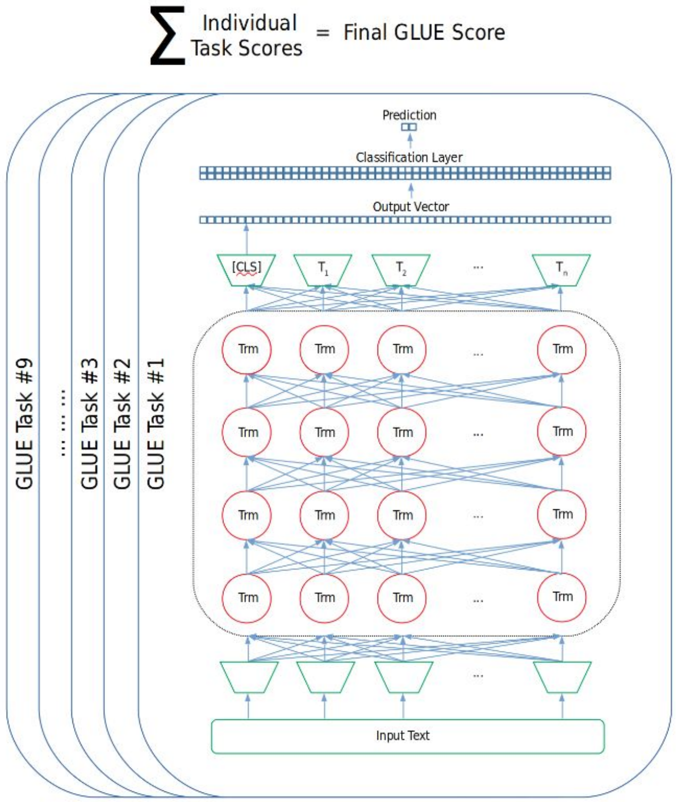
위와 같은 맥락에서 표준 평가 기능이 떠오르고 있다.
2017, 2018년도에 등장한 SentEval을 보면, GLUE와 같은 점은 기존의 classification task에 의존한다는 점
GLUE와 다른 점은 SentEval은 오로지 sentence-to-vector 인코더만 평가한다는 점이다.

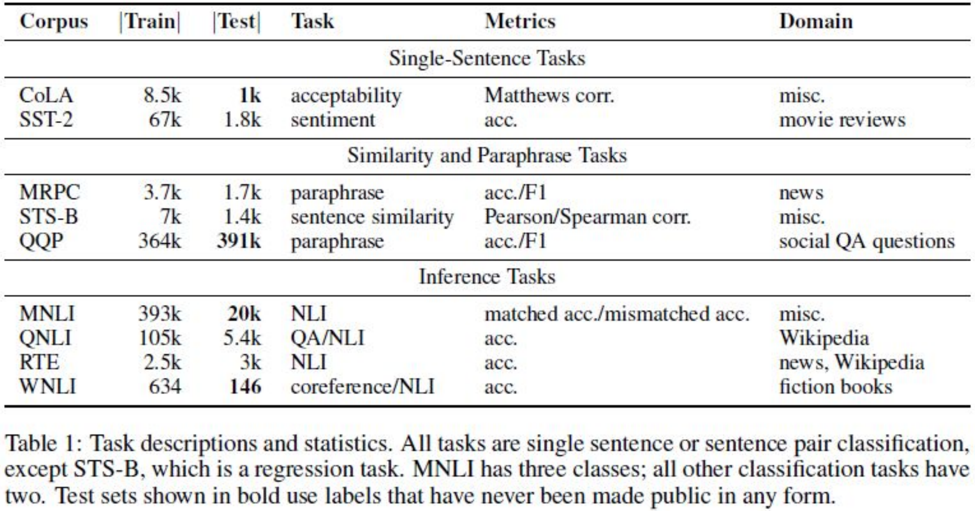
GLUE

Cross-sentence contextualization과 alignment는 MT, QA, NLI 등의 태스크에서 SOTA를 달성하는데 중요하다.
GLUE는 이 방법들을 발전시키도록 디자인되어 있으며, 모델이 문장에 대한 명확한 표현이 없어도 쓸 수 있는
Model-agnostic , 다양함과 어려움을 동시에 체크할 수 있도록 한, 벤치마크 데이터셋.

또한 GLUE는, decaNLP 벤치마크에는 부족한 리더보드와 에러 분석 툴킷을 갖고 있다.

Tasks





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.

▲ Task Outline

◄ GLUE score outline

Single-Sentence Tasks

CoLA (The Corpus of Linguistic Acceptability)

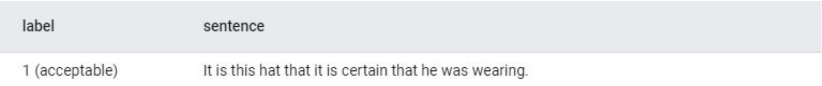
- 문법적 오류 존재 여부 (이진분류)
- Task: Acceptability
- Domain: misc.
- Metric: Matthews Correlation Coefficient
  - 이진 분류를 위한 quality measure

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- e.g.

They made him angry. 1 = acceptable

This building is than that one. 0 = unacceptable



label	sentence
1 (acceptable)	It is this hat that it is certain that he was wearing.

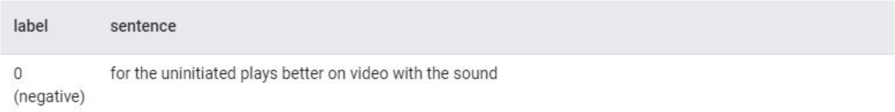
Single-Sentence Tasks

SST-2 (The Stanford Sentiment Treebank)

- 영화 리뷰에 대한 감성 분류 (이진분류)
- Task: Sentiment
- Domain: Movie Reviews
- Metric: Accuracy
- e.g.

  the movie as a whole is cheap junk and an insult to their death-defying efforts = 0

  the movie is funny , smart , visually inventive , and most of all , alive . = 1



label	sentence
0 (negative)	for the uninitiated plays better on video with the sound

Similarity & Paraphrase Tasks

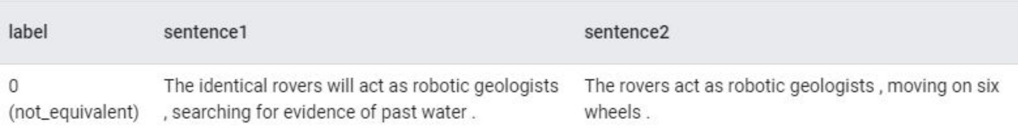
MRPC (The Microsoft Research Paraphrase Corpus)

- 온라인 뉴스 데이터에서 문장 pair들이 의미적으로 동일한지 평가
- Task: Paraphrase
- Domain: News
- Metric : F1 score & Accuracy (class imbalance of 68% pos)
- e.g.

"Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ."

"The island reported another 35 probable cases yesterday , taking its total to 418 ."

label = 1



label	sentence1	sentence2
0 (not_equivalent)	The identical rovers will act as robotic geologists , searching for evidence of past water .	The rovers act as robotic geologists , moving on six wheels .

Similarity & Paraphrase Tasks

QQP (The Quora Question Pairs Dataset)

- Quora 커뮤니티에서 문장 pair 들이 의미적으로 동일한지 평가
- Metric : F1 score & Accuracy (class imbalance of 63% neg)
- Task: Paraphrase
- Domain: Social QA Questions
- e.g.

How can I be a good geologist?

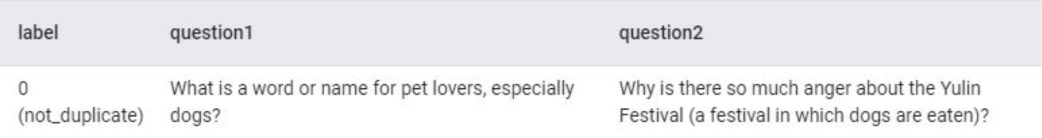
What should I do to be a great geologist?

1 = similar

How can I increase the speed of my internet connection while using a VPN?

How can Internet speed be increased by hacking through DNS?

0 = not similar



label	question1	question2
0 (not_duplicate)	What is a word or name for pet lovers, especially dogs?	Why is there so much anger about the Yulin Festival (a festival in which dogs are eaten)?

Similarity & Paraphrase Tasks

STS-B (The Semantic Textual Similarity Benchmark)

- 뉴스 헤드라인, 영상, 이미지 캡션 등에서 가져온 문장 pair 들의 의미적 유사도 평가
  - Similarity score from 1~5
- Task: Sentence Similarity
- Domain: Misc.
- Metric: Pearson correlation coefficient
- e.g.

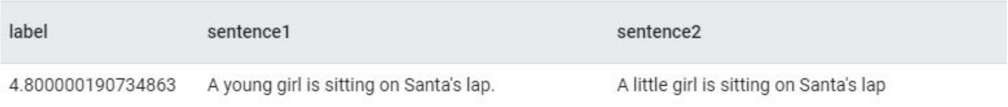
Elephants are walking down a trail | A herd of elephants are walking along a trail.' = 4.6 = very similar

'Thai police use tear gas against protesters | Brazil leader breaks silence about protests' = 1.0 = not similar

'A man is making a bed. | A woman is playing a guitar.' = 0.0 = not similar

'A man is playing a guitar and singing on a stage | A man is playing the electric guitar.' = 3.8 = somewhat similar

'The woman is holding the hands of the man. | The woman is checking the eyes of the man.' = 2.4 = somewhat similar



label	sentence1	sentence2
4.800000190734863	A young girl is sitting on Santa's lap.	A little girl is sitting on Santa's lap

Inference Tasks

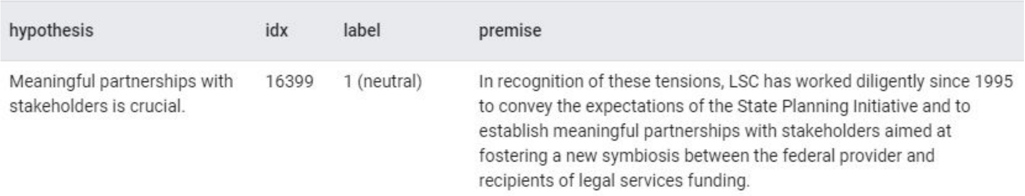
MNLI (The Multi-Genre Natural Language Inference Corpus)

- 크라우드소싱된 entailment 문장 pair 데이터셋
  - *Entailment, Contradiction, Neural* → on premise sent. & hypothesis sent.
- Task: NLI
- Domain: Misc.
- Metric: matched accuracy/mismatched acc.
- Evaluate on both matched(in-domain) & miss-matched(cross-domain)
- SNLI corpus (550k) for auxiliary training
- e.g.

Jon walked back to the town to the smithy. | Jon traveled back to his hometown.' = 1 = neutral

'Tourist Information offices can be very helpful. | Tourist Information offices are never of any help.' = 2 = contradiction

"I'm confused. | Not all of it is very clear to me.' = 0 = entailed



hypothesis	idx	label	premise
Meaningful partnerships with stakeholders is crucial.	16399	1 (neutral)	In recognition of these tensions, LSC has worked diligently since 1995 to convey the expectations of the State Planning Initiative and to establish meaningful partnerships with stakeholders aimed at fostering a new symbiosis between the federal provider and recipients of legal services funding.

Inference Tasks

QNLI (Question-answering NLI)

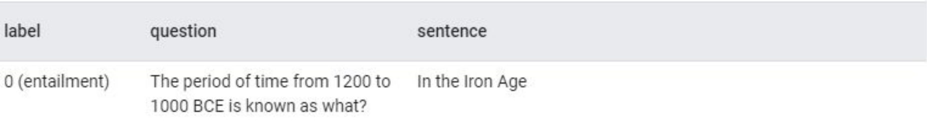
- Question-Paragraph pair 의 QA 데이터셋을 통해 paragraph 가 answer를 포함시키고 있는 지를 평가
  - 원래 answer 가 필히 존재해야 한다는 가정을 제외시킨 버전

- Task: QA, NLI
- Domain: Wikipedia
- Metric: Accuracy
- eg.

"How was the Everton FC's crest redesign received by fans? | The redesign was poorly received by supporters, with a poll on an Everton fan site registering a 91% negative response to the crest.' = 0 = answerable

'In what year did Robert Louis Stevenson die? | Mission work in Samoa had begun in late 1830 by John Williams, of the London Missionary Society arriving in Sapapali'i from The Cook Islands and Tahiti." = 1 = not answerable

'What is essential for the mating of the elements that create radio waves? | Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field.' = 0 = answerable



label	question	sentence
0 (entailment)	The period of time from 1200 to 1000 BCE is known as what?	In the Iron Age

Inference Tasks

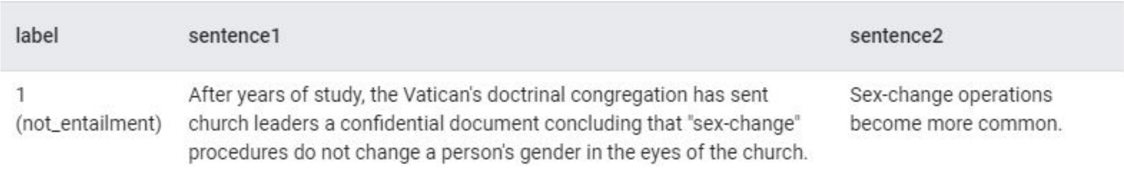
RTE (The Recognizing Textual Entailment)

- RTE1,2,3,5 의 annual textual entailment challenge 를 합친 데이터셋
- Two-class split으로 수정 (neutral & contradiction → not_entailment)
- Task: NLI
- Domain: Wikipedia, News
- Metric: Accuracy
- eg.) In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members. | Yunus supported more than 50,000 Struggling Members.' = 0 = entailed

  "In 1997, Tyson bit off part of one of Evander Holyfield's ears in their rematch that led to Tyson's disqualification and suspension. In 2002, at a melee at a news conference in New York, Tyson bit champion Lennox Lewis' leg. | Tyson bit off part of one of Evander Holyfield's ears in 2002." = 1 = not entailed

  "The marriage is planned to take place in the Kiev Monastery of the Caves, whose father superior, Bishop Pavel, is Yulia Timoshenko's spiritual father | Yulia Timoshenko is the daughter of Bishop Pavel.' = 1 = not entailed

  'With its headquarters in Madrid, Spain, WTO is an inter-governmental body entrusted by the United Nations to promote and develop tourism | The WTO headquarters is located in Madrid, Spain.' = 0 = entailed



label	sentence1	sentence2
1 (not_entailment)	After years of study, the Vatican's doctrinal congregation has sent church leaders a confidential document concluding that "sex-change" procedures do not change a person's gender in the eyes of the church.	Sex-change operations become more common.

Inference Tasks

WNLI (The Winograd NLI)

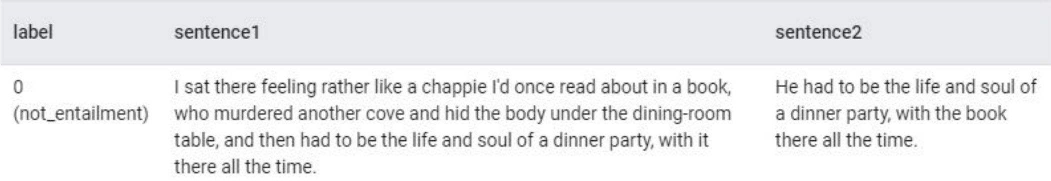
- Winograd Schema Challenge Reading comprehension task → 대명사가 지칭하는 명사 찾기
  - 여기서 단순화시켜서 명사가 대체된 문장의 entailment 여부 평가

- Development set가 adversarial... → 어떤 모형에 대해서 성능 문제가 심하다
- Task: conference, NLI
- Domain: fiction books
- Metric: Accuracy
- eg.

'Lily spoke to Donna, breaking her concentration | Lily spoke to Donna, breaking Lily's concentration." = 0

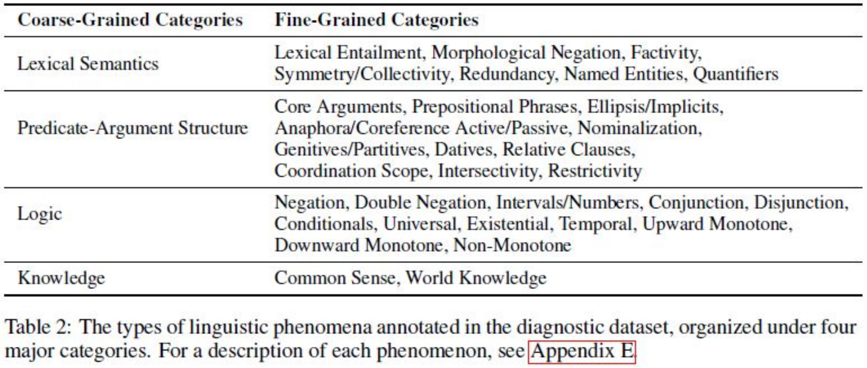
'I put the cake away in the refrigerator. It has a lot of leftovers in it. | The refrigerator has a lot of leftovers in it.' = 1

"Susan knows all about Ann's personal problems because she is nosy. | Ann is nosy.' = 0



label	sentence1	sentence2
0 (not_entailment)	I sat there feeling rather like a chappie I'd once read about in a book, who murdered another cove and hid the body under the dining-room table, and then had to be the life and soul of a dinner party, with it there all the time.	He had to be the life and soul of a dinner party, with the book there all the time.

Diagnostic Dataset



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.

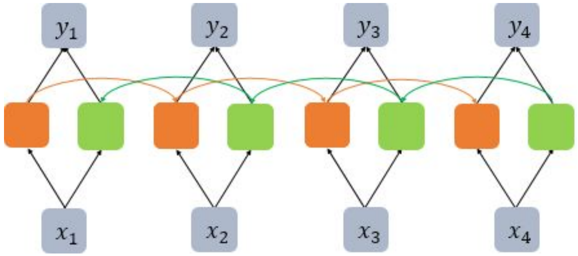
- Small, curated test set for NLI tasks

- NLI task 를 활용해서 해당 phenomena 의 tag에 대해서 이해를 가질 수 있음

  → still does not reflect overall performance so generalizable X

- 하지만 특정 phenomena 에 대한 analysis tool 로써 활용 가능
  - Error analysis
  - Qualitative model comparison
  - Development of adversarial examples

Baselines - Architecture



BiLSTM

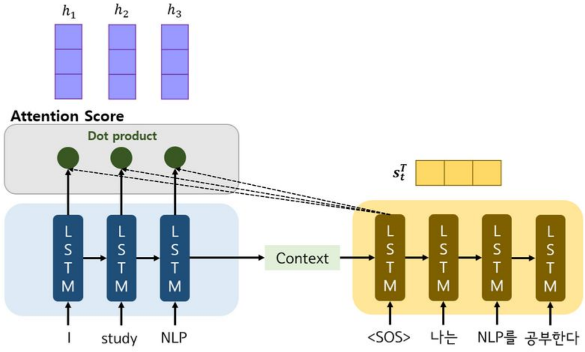
- 2-Layer (Dim=1500 per direction) with Maxpooling
- GloVe word Embeddings (Dim=300)

Tasks

- Single sentence tasks ⇒ [Sentence to Vector → Classifier]
- Sentence-pair tasks ⇒ [Sentence to Vector로 $u; v; |u - v|; u * v$ 를 생성한 뒤에 Classifier에 통과시킴]
- Classifier : MLP(Dim=512)

Attention

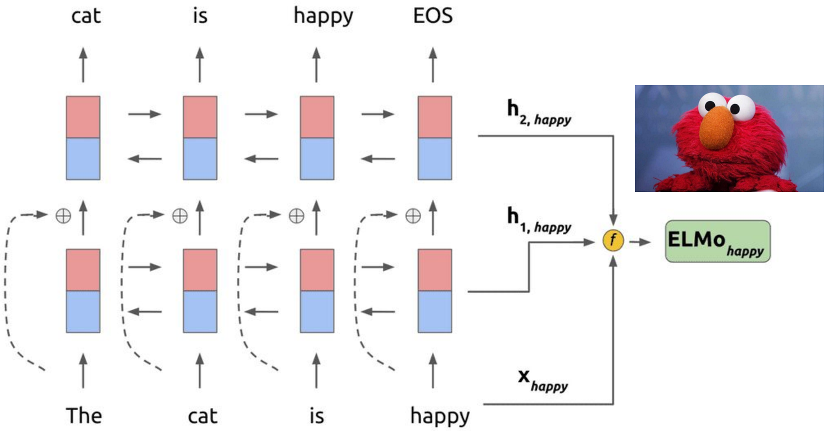
- Sentence-pair tasks 에서 문장간 단어의 관계 파악을 위해 Attention 적용



Baselines - Pre-training

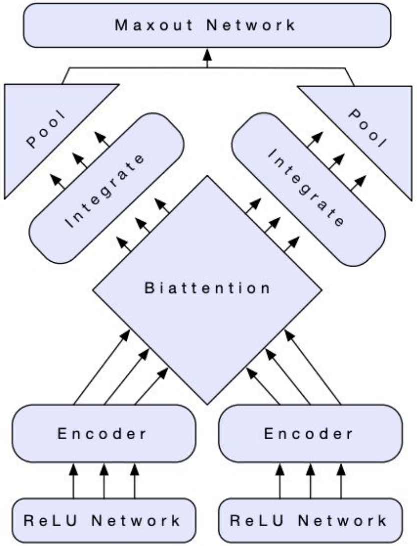
ELMo

- 각 층에서 발생하는 임베딩 벡터의 선형 결합을 통해서 단어의 문맥적 의미를 반영한 임베딩 표현 생성



CoVe

- Eng-to-German 번역 데이터셋으로 훈련한 2-layer BiLSTM 인코더



Baselines & Experiment

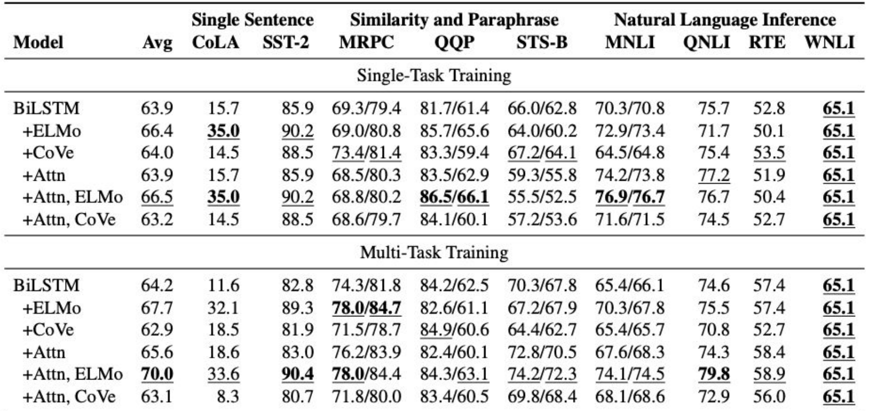
Training

- Optimizer : Adam
- Initial learning rate : 0.0001
- Batch size : 128
- Stop : learning rate가 0.00001에 도달하거나 , validation check이 5번 이상 개선되지 않을 경우 학습 중지

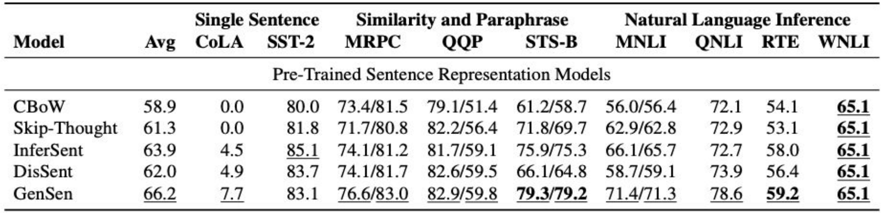
Sentence Representation Models

- GloVe, Skip-Thought, InferSent, DisSent, GenSen 과 비교

Experiment



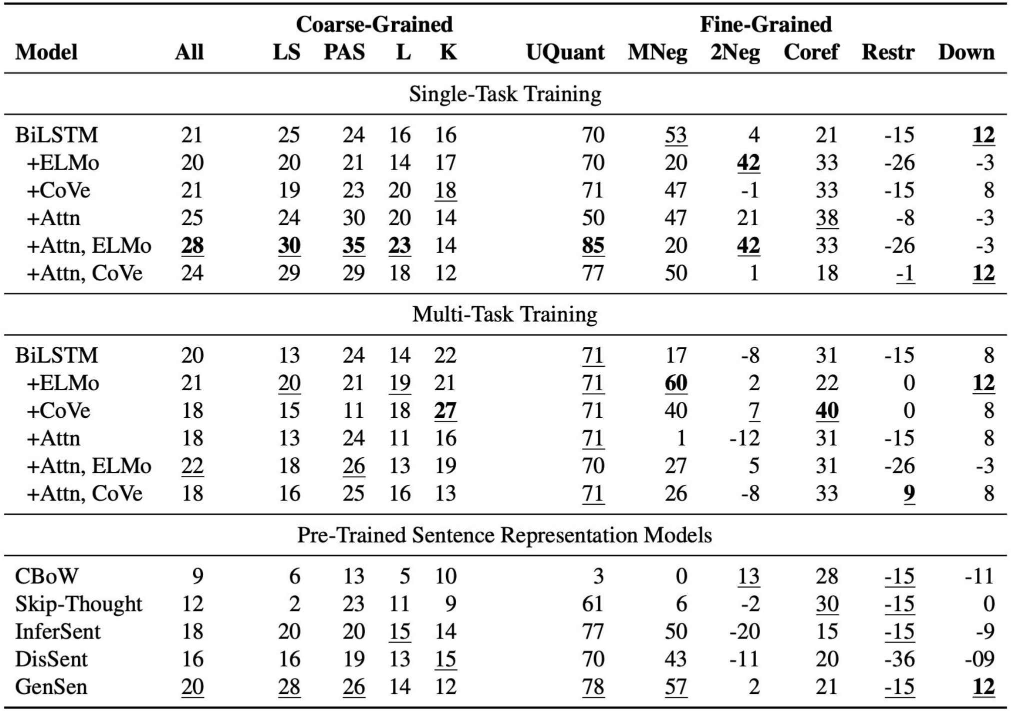
		Single Sentence		Similarity and Paraphrase			Natural Language Inference			
Model	**Avg**	**CoLA**	**SST-2**	**MRPC**	**QQP**	**STS-B**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Single-Task Training										
BiLSTM	63.9	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	75.7	52.8	**65.1**
+ELMo	66.4	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	71.7	50.1	**65.1**
+CoVe	64.0	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	75.4	53.5	**65.1**
+Attn	63.9	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	77.2	51.9	**65.1**
+Attn, ELMo	66.5	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	76.7	50.4	**65.1**
+Attn, CoVe	63.2	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	74.5	52.7	**65.1**
Multi-Task Training										
BiLSTM	64.2	11.6	82.8	74.3/81.8	84.2/62.5	70.3/67.8	65.4/66.1	74.6	57.4	**65.1**
+ELMo	67.7	32.1	89.3	**78.0/84.7**	82.6/61.1	67.2/67.9	70.3/67.8	75.5	57.4	**65.1**
+CoVe	62.9	18.5	81.9	71.5/78.7	84.9/60.6	64.4/62.7	65.4/65.7	70.8	52.7	**65.1**
+Attn	65.6	18.6	83.0	76.2/83.9	82.4/60.1	72.8/70.5	67.6/68.3	74.3	58.4	**65.1**
+Attn, ELMo	**70.0**	33.6	**90.4**	**78.0**/84.4	84.3/63.1	74.2/72.3	74.1/74.5	**79.8**	58.9	**65.1**
+Attn, CoVe	63.1	8.3	80.7	71.8/80.0	83.4/60.5	69.8/68.4	68.1/68.6	72.9	56.0	**65.1**



		Single Sentence		Similarity and Paraphrase			Natural Language Inference			
Model	**Avg**	**CoLA**	**SST-2**	**MRPC**	**QQP**	**STS-B**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	72.1	54.1	**65.1**
Skip-Thought	61.3	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	72.9	53.1	**65.1**
InferSent	63.9	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	72.7	58.0	**65.1**
DisSent	62.0	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	73.9	56.4	**65.1**
GenSen	66.2	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	78.6	**59.2**	**65.1**

- Multi-Task Learning > Single-Task Learning
- Attn, Pre-training(ELMo, CoVe) > 적용하지 않은 경우

Analysis



Model	All	Coarse-Grained				Fine-Grained					
		LS	PAS	L	K	UQuant	MNeg	2Neg	Coref	Restr	Down
Single-Task Training											
BiLSTM	21	25	24	16	16	70	53	4	21	-15	**12**
+ELMo	20	20	21	14	17	70	20	**42**	33	-26	-3
+CoVe	21	19	23	20	18	71	47	-1	33	-15	8
+Attn	25	24	30	20	14	50	47	21	38	-8	-3
+Attn, ELMo	**28**	**30**	**35**	**23**	14	**85**	20	**42**	33	-26	-3
+Attn, CoVe	24	29	29	18	12	77	50	1	18	-1	**12**
Multi-Task Training											
BiLSTM	20	13	24	14	22	71	17	-8	31	-15	8
+ELMo	21	20	21	19	21	71	**60**	2	22	0	**12**
+CoVe	18	15	11	18	**27**	71	40	7	**40**	0	8
+Attn	18	13	24	11	16	71	1	-12	31	-15	8
+Attn, ELMo	22	18	26	13	19	70	27	5	31	-26	-3
+Attn, CoVe	18	16	25	16	13	71	26	-8	33	**9**	8
Pre-Trained Sentence Representation Models											
CBoW	9	6	13	5	10	3	0	13	28	-15	-11
Skip-Thought	12	2	23	11	9	61	6	-2	30	-15	0
InferSent	18	20	20	15	14	77	50	-20	15	-15	-9
DisSent	16	16	19	13	15	70	43	-11	20	-36	-09
GenSen	20	28	26	14	12	78	57	2	21	-15	**12**

Diagnostic set

- Coarse-Grained
  - LS(Lexical Semantics)
  - PAS(Predicate-Argument Structure)
  - L(Logic)
  - K(Knowledge and Common sense)

- Fine-Grained
  - UQuant(Universal Quantification)
  - MNeg(Morphological Negation)
  - 2Neg(Double Negation)
  - Coref(Anaphora/Coreference)
  - Restr(Restrictivity)
  - Down(Downward Monotone)

언어학적인 능력을 측정하기 위한 Diagnostic set 에 MNLI 분류기를 적용하였을 때의 R_3 Coefficients (Target - Prediction)

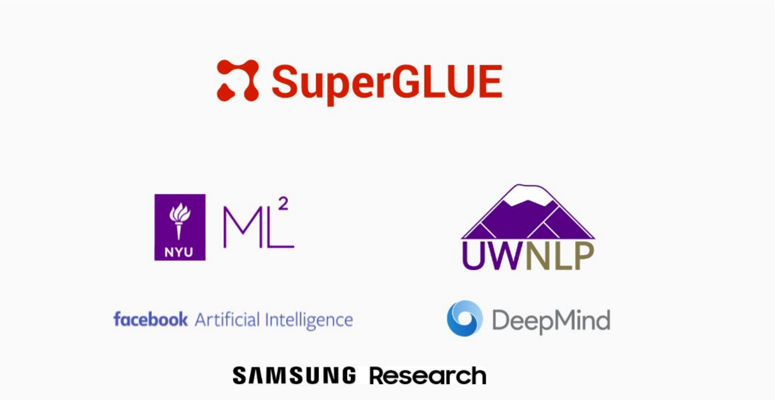
- Multi-Task Learning > Single-Task Learning
- Attn, Pre-training(ELMo, CoVe) > 적용하지 않은 경우

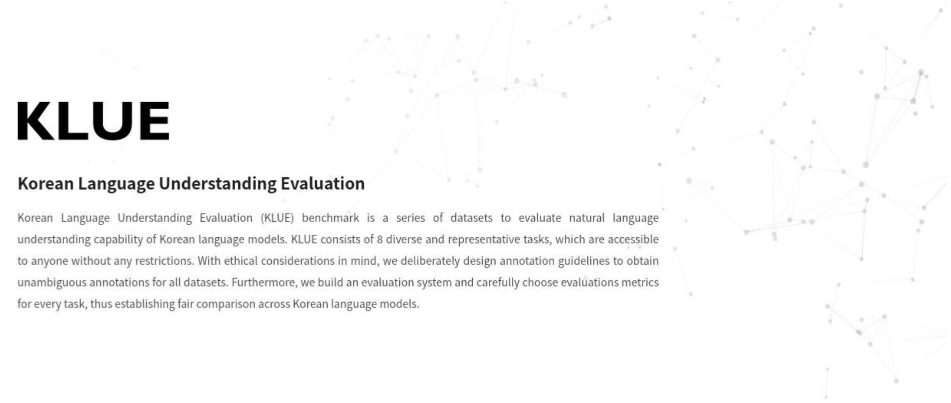
- Fine-Grained 에서는 Attention을 적용하였을 때 성능이 저하되는 현상
  ▶️ Representational Capacity의 증가가 Overfitting을 발생

Conclusion

- GLUE Task를 종합적으로 학습한 모델이 각 태스크를 독립적으로 학습한 모델을 조합한 것보다 좋은 성능을 기록
- Attention 이나 Pre-training(ELMo, CoVe)를 학습한 모델이 그렇지 않은 모델보다 더 좋은 성능을 보였으나 아직 성능이 더 좋아질 여지가 있음
  - 언어학적 현상(Linguistic Phenomena)에 대해서 모델이 실패하는 경우가 종종 있었기 때문
- NLU Tasks 를 종합적으로 평가하기 위한 Objective를 설계하는 일은 여전히 미궁 속에 있음
  - 하지만 해당 논문에서 제시한 **GLUE 가 좋은 밑거름**이 될 것으로 생각함

Coming..





KLUE

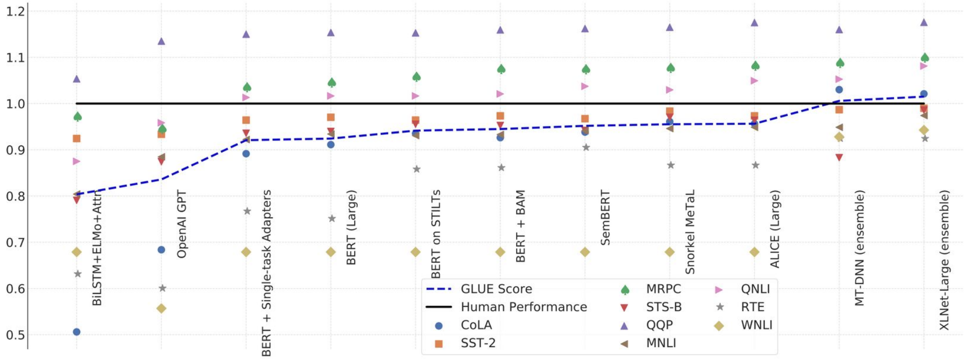
Korean Language Understanding Evaluation

Korean Language Understanding Evaluation (KLUE) benchmark is a series of datasets to evaluate natural language understanding capability of Korean language models. KLUE consists of 8 diverse and representative tasks, which are accessible to anyone without any restrictions. With ethical considerations in mind, we deliberately design annotation guidelines to obtain unambiguous annotations for all datasets. Furthermore, we build an evaluation system and carefully choose evaluations metrics for every task, thus establishing fair comparison across Korean language models.

SuperGLUE

Abstract

- (18년도에 BERT, GPT 및 19년도에 이를 활용한 모델이 뛰어난 성능을 기록하면서) NLU Task에 대한 자연어 모델의 엄청난 발전이 있었음
- GLUE 가 NLU Task를 측정할 수 있는 몇 가지 Metric 을 제시하였으나 이미 (비전문가) Human performance를 뛰어 넘음



- 그래서 더 어려운 **SuperGLUE 데이터셋을 제시**하고자 함

(SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems) 를 참조하였습니다.

SuperGLUE

GLUE와의 차이점

- 쉬운 Task는 버리고, 어려운 Task는 남기고 :)
- 새로운 형태의 Task를 추가하였음
  - Coreference Resolution(상호 참조), QA(Question Answering)
- 비교를 위한 Human Baseline 와 Code 를 제공

Tasks

- **BoolQ (Boolean Questions) : Question Answering**
  - 주어진 문단에 대한 질문에 (예, 아니오)로 대답하는 문제
- **CB (CommitmentBank) : NLI(Natural Language Inference)**
  - 전제(Premise)에 대해 가설(Hypothesis)이 함의(Entailment)인지 중립(Neutral)인지 모순(Contradiction)인지를 판단
- **COPA (Choice of Plausible Alternatives) : Question Answering (Casual reasoning)**
  - 전제에 주어지면 질문이 주어지고 Label/Alternative 중 1개의 답을 선택(Binary Classification)
- **MultiRC (Multi-Sentence Reading Comprehension) : Question Answering**
  - 단락에 대한 질문과 답이 주어지면 True/False 중 1개의 답을 선택(Binary Classification)
- **ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset) : Question Answering (Multiple-choice)**
  - 단락과 Entity가 주어지면 빈칸에 들어갈 단어를 Multiple-Choice
- **RTE (Recognizing Textual Entailment) : NLI(Natural Language Inference)**
  - 전제(Premise)에 대해 가설(Hypothesis)이 함의(Entailment)인지 아닌지(Not Entailment)를 판단 (GLUE에도 있는 Task)
- **WiC (Word-in-Context) : Word Sense Disambiguation**
  - 다의어(Polysemous Word)를 포함한 문맥이 2가지 주어질 때 두 다의어가 같은 뜻(True)인지 아닌지(False)를 판단
- **WSC (Winograd Schema Challenge) : Coreference Resolution**
  - 대명사와 명사가 주어지면 참조(Reference)가 올바른지(True) 아닌지(False)를 판단

KLUE: Korean Language Understanding Evaluation

- 한국어 NLU Task 특성을 위한 8가지 데이터셋 마련
  1. Topic Classification
  2. Semantic Textual Similarity : 의미 유사성 분석
  3. Natural Language Inference
  4. Named Entity Recognition
  5. Relation Extraction : 관계 추출
  6. Dependency Parsing : 의존성 분석
  7. Machine Reading Comprehension
  8. Dialogue State Tracking : 대화 상태 추적

KLUE: Korean Language Understanding Evaluation

Sungjoon Park*
Upstage, KAIST
sungjoon.park@upstage.ai

Jihyung Moon*
Upstage
jihyung.moon@upstage.ai

Sungdong Kim*
NAVER AI Lab
sungdong.kim@navercorp.com

Won Ik Cho*
Seoul National University
tsatsuki@snu.ac.kr

Jiyoon Han†
Yonsei University
clinamen35@yonsei.ac.kr

Jangwon Park
jangwon.pk@gmail.com

Chisung Song
daydrilling@gmail.com

Junseong Kim
Scatter Lab
junseong.kim@scatterlab.co.kr

Thank you!