IndirectQA Model Analysis Using Transfer Learning

NYU DS-GA 1012
Natural Language Understanding

Anu-Ujin Gerelt-Od (ag0265)
Lakshmi Menon (lsm454)
Angela Marie Teng (at2507)

Task & Data

- IndirectQA task - understand indirect responses to naturally occurring boolean questions
- **Circa** Corpus with 34K question-answer-label pairs
- Relaxed: 4 labels; Strict: 6 labels
- BERT-based models, fine-tuned on BoolQ and MNLI
- Can be used to improve performance of conversational chatbots and AI agents

"I'd rather just go to bed": Understanding Indirect Answers

Annie Louis
Google Research, UK
annielouis@google.com

Dan Roth*
University of Pennsylvania
danroth@seas.upenn.edu

Filip Radlinski
Google Research, UK
filiprad@google.com

Label	RELAXED	
Yes	16,628	(48.5%)
No	12,833	(37.5%)
Yes, subject to some conditions	2,583	(7.5%)
In the middle, neither yes nor no	949	(2.8%)
Other	504	(1.5%)
N/A	771	(2.2%)

Table 8: Distribution of RELAXED gold standard labels. 'N/A' indicates lack of majority agreement.

Label	STRICT	
Yes	14,504	(42.3%)
No	10,829	(31.6%)
Probably yes / sometimes yes	1,244	(3.6%)
Yes, subject to some conditions	2,583	(7.5%)
Probably no	1,160	(3.4%)
In the middle, neither yes nor no	638	(1.9%)
I am not sure	63	(0.2%)
Other	504	(1.5%)
N/A	2,743	(8.0%)

Table 7: Distribution of STRICT gold standard labels. 'N/A' indicates lack of majority agreement.

Model	Accuracy for relaxed		Accuracy for strict	
	Original	*Replicated*	*Original*	*Replicated*
BERT-YN	87.8	83.3	84.0	87.3
BERT-BOOLQ-YN	87.1	85.6	83.4	82.1
BERT-MNLI-YN	88.2	86.4	84.8	82.6

Table 1: Replication results in comparison to original values

RoBERTa

- Replication study of BERT pretraining model that optimizes hyperparameters and training data size
- SOTA on GLUE, RACE, and SQuAD
- Aside from replicating BERT-MNLI-Circa code, we wanted to expand to other SOTA models and compare performance
- Longer training, bigger batches, removing next sentence prediction objective, training on longer sequences
- Dynamically changing masking pattern on training data

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] Myle Ott[*§] Naman Goyal[*§] Jingfei Du[*§] Mandar Joshi[†]
Danqi Chen[§] Omer Levy[§] Mike Lewis[§] Luke Zettlemoyer[†§] Veselin Stoyanov[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

	RoBERTa MNLI Strict Matched	**RoBERTa MNLI Relaxed Match**
Test Accuracy	0.87	0.90
Test F1 Score	0.86	0.89

T5

- Text-to-Text Transfer Transformer with input and output as text
- Trained on C4 corpus of English text ("Colossal Clean Crawled Corpus")
- Offers flexibility of applying the same model to different NLP tasks

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* CRAFFEL@GMAIL.COM
Noam Shazeer* NOAM@GOOGLE.COM
Adam Roberts* ADAROB@GOOGLE.COM
Katherine Lee* KATHERINELEE@GOOGLE.COM
Sharan Narang SHARANNARANG@GOOGLE.COM
Michael Matena MMATENA@GOOGLE.COM
Yanqi Zhou YANQIZ@GOOGLE.COM
Wei Li MWEILI@GOOGLE.COM
Peter J. Liu PETERJLIU@GOOGLE.COM
Google, Mountain View, CA 94043, USA

	T5 Strict Matched	**T5 Relaxed Match**
Test Accuracy	0.77	0.74
Test F1 Score	0.82	0.76

UnifiedQA

- T5 and BART- based architecture, pretrained on four different NLI tasks using 8 datasets
- Fine-tuned directly on Circa dataset for our task
- Saw better performance than original paper on Relaxed setting, but not on Strict

EMNLP-Findings'20

UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi[1] Sewon Min[2] Tushar Khot[1] Ashish Sabharwal[1]
Oyvind Tafjord[1] Peter Clark[1] Hannaneh Hajishirzi[1,2]

[1]Allen Institute for AI, Seattle, U.S.A.
[2]University of Washington, Seattle, U.S.A.

	UnifiedQA Strict Matched	UnifiedQA Relaxed Matched
Test Accuracy	0.747	0.897
Test F1 Score	0.717	0.892

Thank you!