HPCLab Paper Review 2024

SUPER-NATURALINSRUCTION
Generalization via Declarative Instruction on 1600+ NLP Tasks

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi

2024.03.14

홍익대학교 소프트웨어융합학과

HPC Lab

Contents Table

Super-Instructions Dataset & Tk-Instruct Model

- ✓ **SUPER-NATURALINSTRUCTIONS**를 소개하여 자연어 처리(NLP) 모델이 작업 지시를 받을 때 얼마나 효과적으로 다양한 작업에 일반화 되는지 평가한다.
- ✓ 해당 데이터셋은 **1616가지 다양한 NLP 작업(Task)이** 포함되어 있으며 **Task의 유형은 76 가지**이다.
- ✓ 데이터의 기본 schema는 **Definition, Positive Example, Negative Example, Evaluation Instances**으로 구성되어 있으며 task 유형에 따라 추가되는 schema가 있다.
- ✓ 데이터셋의 용량은 약 **3GB**이다.
- ✓ **Tk-Instruct**라는 작업 지시를 따르도록 훈련된 **트랜스포머 기반**의 모델을 소개하며, 실험 결과 기존의 모델들보다 **9% 이상 우수한 성능**을 보여주었다.

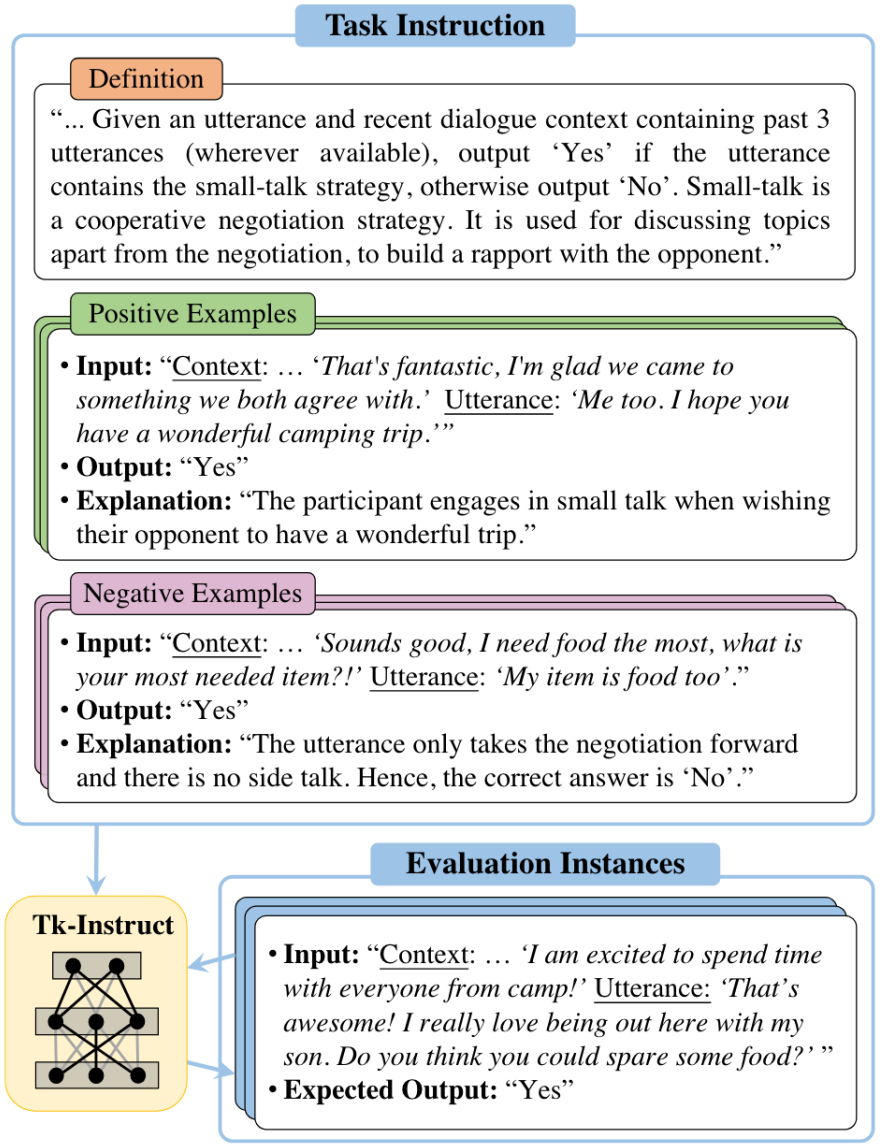
Characteristics of the dataset

1. Task instruction을 가지고 있다.
2. 오답인 경우와 이에 대한 설명이 포함되어 있다.
3. 55개의 다양한 언어로 구성되어 있다.
4. 데이터셋이 공개되어 있다.

Resource →	SUP-NATINST (this work)	NATINST (Mishra et al., 2022b)	CROSSFIT (Ye et al., 2021)	PROMPTSOURCE (Bach et al., 2022)	FLAN (Wei et al., 2022)	INSTRUCTGPT (Ouyang et al., 2022)
Has task instructions?	✓	✓	✗	✓	✓	✓
Has negative examples?	✓	✓	✗	✗	✗	✗
Has non-English tasks?	✓	✗	✗	✗	✓	✓
Is public?	✓	✓	✓	✓	✓	✗
Number of tasks	1616	61	269	176	62	–
Number of instructions	1616	61	–	2052	620	14378
Number of annotated tasks types	76	6	13	13*	12	10
Avg. task definition length (words)	56.6	134.4	–	24.8	8.2	–

Architecture of the Dataset

- ✓ Definition : 자연어로 주어진 작업을 정의한다. 이는 입력 텍스트에 대한 출력이 어떠한 형태를 기대하는지 정의한다.
- ✓ Positive Examples : 입력과 그에 대한 올바른 출력의 샘플이다. Output이 왜 정답인지에 대한 간단한 설명이 포함되어 있다.
- ✓ Negative Examaples : 입력과 그에 대한 부정확하거나 잘못된 출력의 샘플이며, Output이 왜 오답인지에 대한 간단한 설명이 포함되어 있다.
- ✓ Instances : input과 output으로 구성되었으며, 모델의 학습과 성능 평가를 위해 사용된다.



Example : En-Ko Translation

- **Definition** *[] 1 item*
  - 0 "Given a sentence in English, provide an equivalent paraphrased translation in Korean that retains the same meaning both through the translation and the paraphrase."
- **Input_language** *[] 1 item*
- **Output_language** *[] 1 item*
- **Instruction_language** *[] 1 item*
- **Domains** *[] 1 item*
- **Positive Examples** *[] 2 items*
  - 0
    - **input** "The NBA season of 1975 -- 76 was the 30th season of the National Basketball Association ."
    - **output** "National Basketball Association의 1975 - 76 시즌은 NBA의 30 번째 시즌이었다."
    - **explanation** "This is a correct and accurate translation from English to Korean because the translated paraphrase retains the main message that between the years 1975-1976, the 30th NBA season occurred."
  - 1
- **Negative Examples** *[] 2 items*
  - 0
    - **input** "In Paris , in October 1560 , he secretly met the English ambassador , Nicolas Throckmorton , asking him for a passport to return to England through Scotland ."
    - **output** "1560 년 10 월 그는 파리의 니콜라스 스록 모튼 (Nicolas Throckmorton) 대사와 비밀리에 만나 영국에있는 스코틀랜드로 돌아갈 여권을 요청했다."
    - **explanation** "This is not a proper translation from English to Korean because the resulting translation is not accurate to the original idea in which the locations of England and Scotland are not accurate."

Example : En-Ko Translation (Instance)

▼ **Instances** *[] 250 items*

▼ **0**

id "task777-c1c403f04671467d9b86db61b365a53c"

input "He was a scholar in Metaphysical Literature , Theology and Classical sciences ."

▼ **output** *[] 1 item*

0 "그는 형이상학 문학, 신학 및 고전 과학 학자였습니다."

▼ **1**

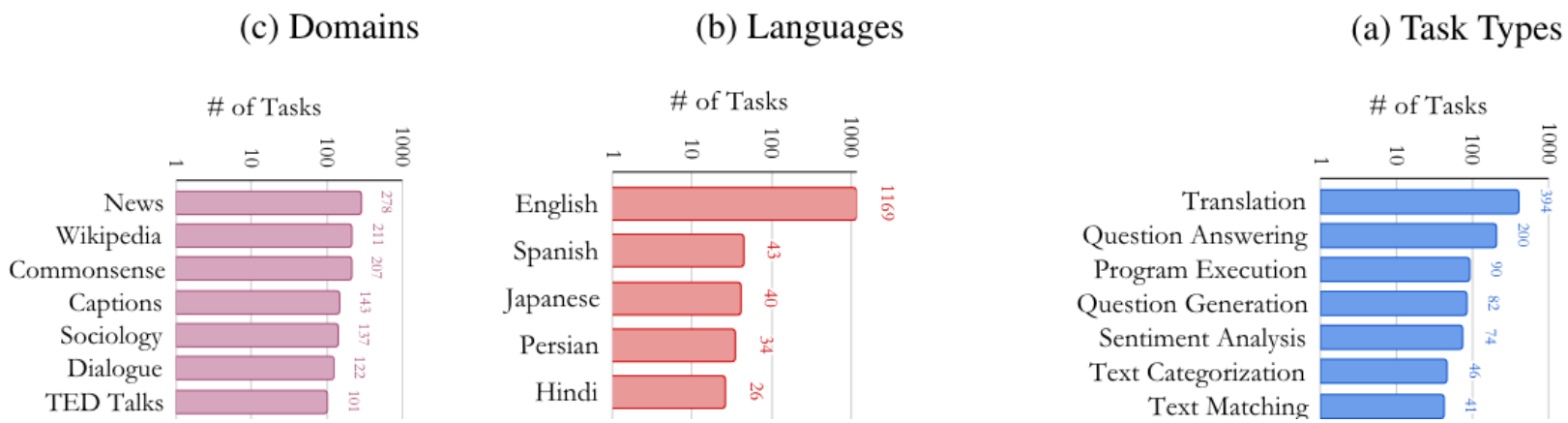
id "task777-8a32e1d68a7c4fe0965f6cc5d12b75e3"

input "The city sits at the confluence of the Snake River with the great Weiser River , which marks the border with Oregon ."

▶ **output** *[] 1 item*

Diversity of Tasks

- ✓ Task의 다양성을 더 잘 이해하기 위해 Task를 세 가지 지침에 따라 분류한다.
  - ✓ **Task Types** : Task의 유형. (Translation, Question Answering, Program Execution …)
  - ✓ **Languages** : Task에서 사용되는 언어. (언어 분류 시 input이 어떤 언어인지에 따라 분류. 한국어는 55개의 언어 중 27위.)
  - ✓ **Domains** : Task의 주제가 어떤 분야인지. (News, Wikipedia, Commonsense …)



Statistics of Super-Naturalinstruction

- ✓ Task 개수 : 1616
- ✓ Task 유형 : 76
- ✓ 언어 수 : 55
- ✓ 도매인(분야) 수 : 33
- ✓ 영어가 아닌 Task : 576
- ✓ Definition schema에서 평균 단어 개수 : 56.6
- ✓ Positive example의 평균 개수 : 2.8
- ✓ Negative example의 평균 개수 : 2.4
- ✓ Instance의 평균 개수 : 3106.0 (total 5백만)

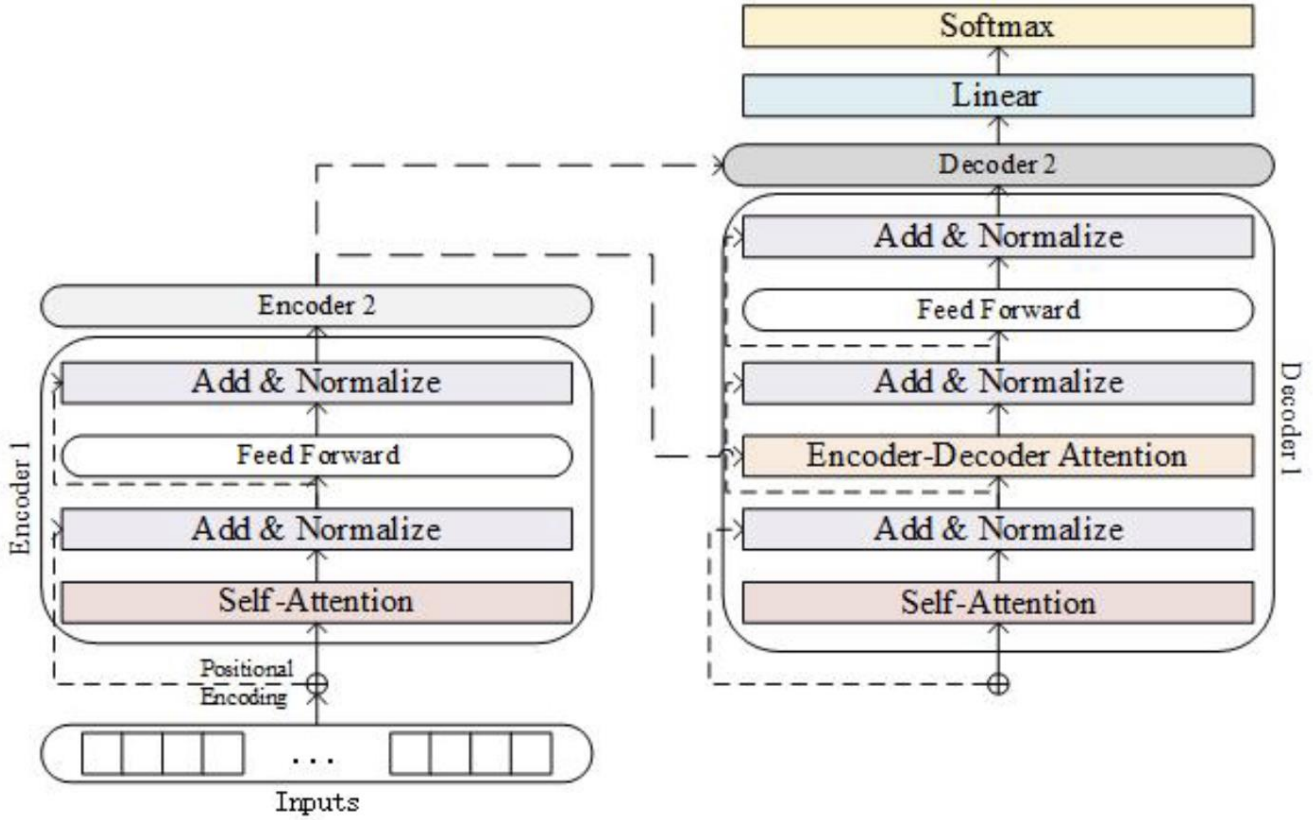
statistic	
# of tasks	1616
# of task types	76
# of languages	55
# of domains	33
# of non-English tasks	576
avg. definition length (words per task)	56.6
avg. # of positive examples (per task)	2.8
avg. # of negative examples (per task)	2.4
avg. # of instances (per task)	3106.0

Table 2: Statistics of SUP-NATINST.

Tk-Instruct Model

- ✓ **Pre-train된 T5모델**에 **Super-Naturalinstruction 데이터셋을 사용하여 fine-tuning** 시킨 모델이다.

- ✓ T5모델의 파라미터 개수 : 2.2억개
  <-> 기초적인 Transformer 모델의 파리미터 개수 : 6500만개



Architecture of T5(Text-to-Text Transfer Trasformer)

Evaluation Setup

- ✓ 총 1616개의 task 중에서, 154개(12가지 task type)의 task에서 최대 100개의 instance만 추출한다.
- ✓ 이를 통해 **총 15,310개의 테스트 instance가 생성**된다.
- ✓ **이러한 테스트 instance는 모델 학습에는 사용되지 않으며**, 이를 제외한 나머지 task가 학습에 사용된다.
- ✓ 154개의 task는 또한 **영어 task (119개)**, **영어가 아닌 task (35개)**로 나누어 성능을 평가한다.

Evaluation Metric – Rouge L

✓ **Rouge-L**은 텍스트 요약의 성능을 측정하는 지표로, 가장 긴 공통 부분열(LCS)을 활용한다.

원본 문장 (정답) : "The quick brown **fox jumped over** the lazy dog"

모델의 출력 문장 : "A **fox jumped over** a dog"

ROUGE_L = 0.333…

$$ROUGE\_L = \frac{\text{LCS 길이}}{\text{원본 문장 길이}}$$

Heuristic Baseline

- ✓ **Copying Demo Output** : 무작위 데모 예제의 출력을 복사하여 실제 정답과 비교해 Rouge-L을 평가한다.
- ✓ **Copying Instance Input** : 주어진 인스턴스의 입력을 복사하여 실제 정답과 비교해 Rouge-L을 평가한다.
- ✓ **Heuristic** 방법들은 주로 **모델을 훈련하지 않은 상태**에서 어떤 성능을 기대할 수 있는지를 확인하기 위한 평가 방법으로 사용된다.
- ✓ 이들은 모델의 훈련 후에 얻은 결과와 비교함으로써 모델의 효과를 평가하는 데 도움이 된다.

Performance of different methods

- ✓ **Heuristic Baselines < Pretrained LMs < Instruction-tuned Models 순**으로 우수한 성능 지표(Rouge-L)를 보인다.

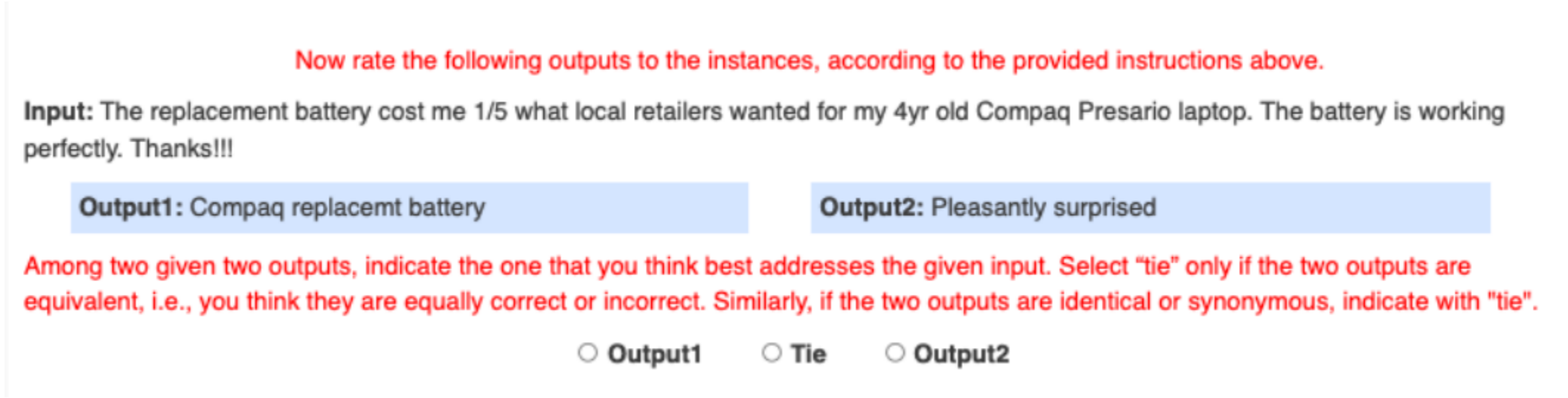
- ✓ 기존에 존재하던 T0, InstructGPT 모델보다도 Tk-instruct 모델과, mTk-INSTRUCT 모델이 영어 task, 다국어 task 각각에서 월등한 성능을 보인다.

- ✓ Tk-Instruct 모델은 상한 추정치인 지도 학습 결과와 비교했을 때 더 향상될 수 있는 가능성을 지니고 있다.

	Methods ↓ / Evaluation →	En	X-lingual
Heuristic Baselines	Copying Instance Input	14.2	5.4
	Copying Demo Output	28.5	50.3
Pretrained LMs	T5-LM (11B)	30.2	–
	GPT3 (175B)	45.0	51.3
Instruction-tuned Models	T0 (11B)	32.3	–
	InstructGPT (175B)	52.1	52.8
	T*k*-INSTRUCT (ours, 11B)	**62.0**	–
	mT*k*-INSTRUCT (ours, 13B)	57.1	**66.1**
Upper-bound (est.)	Supervised Training	74.3	94.0

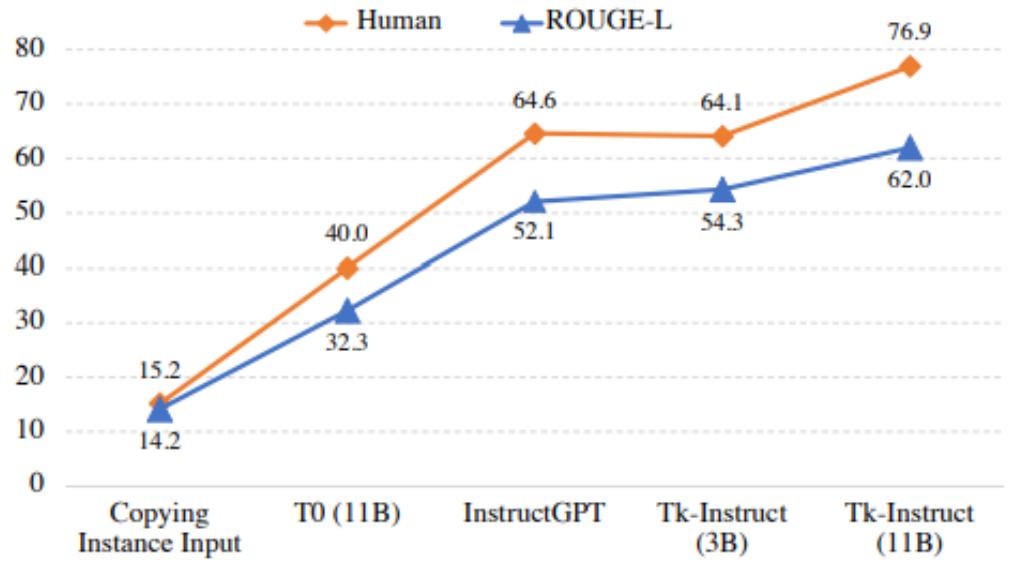
Human Evaluation

- ✓ 아래 그림은 Human Evaluation에서 사용되는 사용자 평가 채점기준의 일부이다.
- ✓ 모델 Output과 사람이 고른 output이 동일하면 점수 1을 얻고 다르면 0을 얻는다.
- ✓ 사람이 생각했을 때, Output이 문맥상 동일한 의미이거나 실제로 같다면 tie를 고르고 1점을 얻는다
- ✓ 평균을 낸 점수가 사람 평가의 지표이다.



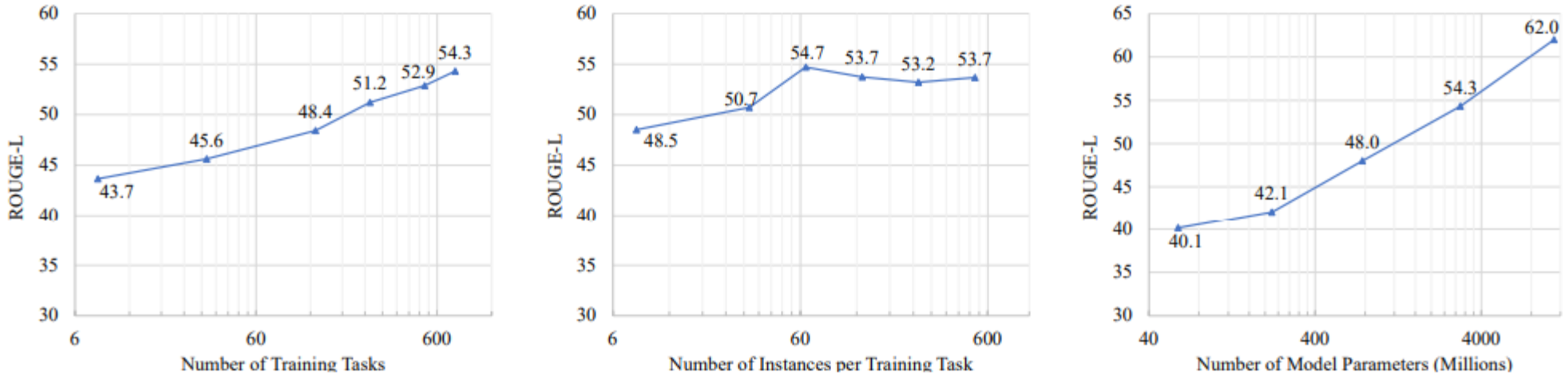
Human Evaluation

- ✓ 인간이 직접 모델의 출력 결과를 보고 정답과 비교해서 우월하거나 동등한지 평가한다.
- ✓ 인간 평가는 평가 작업 중에서 무작위로 선택된 60개의 task(평가 작업의 약 절반) 및 각 task의 10개의 무작위 인스턴스에서 수행된다. 인력 문제로 영어 task만 수행하였다.
- ✓ 인간 평가는 Rouge-L과 비교해봤을 때 비슷한 추이를 확인할 수 있다.



Scaling Trends of Generalization

- ✓ Task 수가 많아질수록 성능을 향상시킨다.
- ✓ Train Instance는 각 task당 64개가 넘어가면 학습 시간이 늘어나고 overfitting 위험을 증가시킨다.
- ✓ Pretrained 된 모델의 크기가 클 수록 성능을 향상시킨다.
- ✓ 모델의 크기가 작다면 task 수를 늘리는 것이 대안이 될 수 있다.



Scaling Trends of Generalization

✓ 아래의 표는 schema를 변화시키며 Training, Test Encoding을 진행했을 때의 Rouge-L 점수이다.

✓ Schema의 (k) 은 사용된 데이터의 개수를 나타낸다.

✓ 실험 결과 **Definition + 4개의 Positive** schema로 Training 했을 때 다양한 test case에 대하여 가장 높은 평균 점수를 나타낸다.

Testing Encoding → Training Encoding ↓	Task ID	Def	Pos (1)	Def + Pos (1)	Pos (2)	Def + Pos (2)	Def + Pos (2) + Neg (2)	Def + Pos (2) + Neg (2) + Expl	Pos (4)	Def + Pos (4)	Average
Task ID	21.2	33.3	16.8	30.9	23.0	33.7	33.9	31.6	26.0	36.4	33.9
Def	17.3	45.0	31.1	43.8	36.4	46.4	44.2	44.3	38.0	46.0	39.9
Pos (1)	10.9	22.1	43.9	47.8	46.6	49.2	46.2	43.4	46.6	49.5	43.1
Def + Pos (1)	11.1	42.2	43.8	52.4	47.4	53.3	53.1	51.8	47.8	53.7	44.5
Pos (2)	12.7	22.4	47.1	50.2	49.3	52.3	50.6	46.7	49.8	52.4	45.0
Def + Pos (2)	12.4	42.1	44.5	52.4	49.0	54.3	53.5	52.7	50.3	54.8	46.4
Def + Pos (2) + Neg (2)	14.0	42.3	43.6	51.8	48.6	53.5	54.3	50.2	49.6	53.8	45.9
Def + Pos (2) + Neg (2) + Expl	15.4	42.0	43.8	50.7	47.6	51.9	52.5	52.6	48.6	52.2	44.3
Pos (4)	11.0	23.9	45.6	49.8	49.0	51.7	49.5	47.5	49.8	51.3	44.5
Definition + Pos (4)	11.0	42.4	44.3	51.9	48.7	53.7	53.4	50.6	50.5	53.5	46.0

Scaling Trends of Generalization

- ✓ **Definition과 Positive Example을 결합**하여 학습시키면 성능 향상이 있다.
- ✓ Positive Example을 무작정 많이 추가하는 것은 성능 향상에 도움이 되지 않는다.
- ✓ Negative Example은 성능 향상에 미미한 영향을 미친다.
- ✓ Explanation을 추가하면 성능이 감소하는데, 이는 모델이 충분히 크지 않을 때 관측된 결과이다.

Testing Encoding → / Training Encoding ↓	Task ID	Def	Pos (1)	Def + Pos (1)	Pos (2)	Def + Pos (2)	Def + Pos (2) + Neg (2)	Def + Pos (2) + Neg (2) + Expl	Pos (4)	Def + Pos (4)	Average
Task ID	21.2	33.3	16.8	30.9	23.0	33.7	33.9	31.6	26.0	36.4	33.9
Def	17.3	45.0	31.1	43.8	36.4	46.4	44.2	44.3	38.0	46.0	39.9
Pos (1)	10.9	22.1	43.9	47.8	46.6	49.2	46.2	43.4	46.6	49.5	43.1
Def + Pos (1)	11.1	42.2	43.8	52.4	47.4	53.3	53.1	51.8	47.8	53.7	44.5
Pos (2)	12.7	22.4	47.1	50.2	49.3	52.3	50.6	46.7	49.8	52.4	45.0
Def + Pos (2)	12.4	42.1	44.5	52.4	49.0	54.3	53.5	52.7	50.3	54.8	46.4
Def + Pos (2) + Neg (2)	14.0	42.3	43.6	51.8	48.6	53.5	54.3	50.2	49.6	53.8	45.9
Def + Pos (2) + Neg (2) + Expl	15.4	42.0	43.8	50.7	47.6	51.9	52.5	52.6	48.6	52.2	44.3
Pos (4)	11.0	23.9	45.6	49.8	49.0	51.7	49.5	47.5	49.8	51.3	44.5
Definition + Pos (4)	11.0	42.4	44.3	51.9	48.7	53.7	53.4	50.6	50.5	53.5	46.0

Conclusion

✓ 해당 논문에서는 **다양한 유형의 NLP task와 그에 대한 지시사항**을 포함하는 대규모 벤치마크를 구축한다.

✓ Super-Naturalinstruction 데이터셋을 활용하여 Tk-Instruct 모델을 훈련시키고, **모델이 새로운 task를 성공적으로 수행할 수 있음을** 보여주었다.

Limitations

- ✓ 데이터셋은 다양성을 제공하지만, 특정 task 및 언어(영어)에 편향되어 있다.

- ✓ ROUGE-L이 효과적인 평가 지표로 작용하지 않는 특정 task가 있다.
  - ✓ Ex) Rewriting task, Error correction task과 같은 경우에 입력을 복사하면 높은 ROUGE-L 값을 얻는다.

Thank You