

Artificial Intelligence Laboratory

## Ch4. Downstream NLP Tasks with Transformers

Transformers for NLP

정보융합공학과 AI전공  
정주경

- 사람과 기계의 언어학습 방식
- 기존의 human baseline를 능가
- Transformer model 성능 측정 방법 : Downstream task

## - 용어 설명

NLU(Natural Language Understanding) : 자연어 표현을 기계가 이해할 수 있는 다른 표현으로 변환시키는 것.

즉, 단어나 문장의 형태를 기계가 인식하도록 하는 것이 아닌, 의미를 인식하도록 하는 것으로 언어를 읽고 통역하는 과정

Human baseline : NLU task에서 사람들의 성능

1. The human intelligence stack
2. The machine intelligence stack
3. Evaluating models with metrics
4. Defining the GLUE benchmark tasks
5. Defining the SuperGLUE benchmark tasks
6. Summary

# The human intelligence stack

- Layer 0

- Input : raw event들의 인식, 상황을 낱것 그대로 봄
  - Transduction Trial&error : 두가지 상황 사이의 연결고리를 만듦
- Transduction : 관찰된 특정 (학습) 데이터에서 특정 (테스트) 데이터를 추론하는 것

Ex)

As children, we were often forced to take a nap early in the afternoon<sup>Nx</sup>  
⇒ I haven't taken a nap, so it's not the afternoon : nap – afternoon

- Layer 1

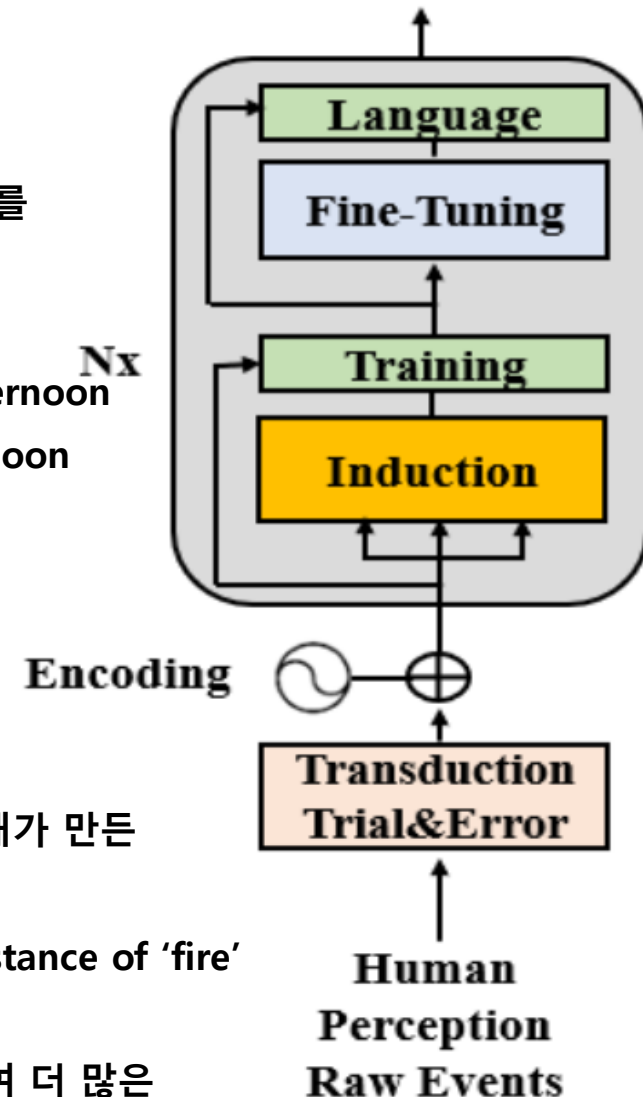
- Induction&Training : Transduction으로 연결고리를 만든 다음 일반화하고 추론

Ex) Trained\_realted events = {sunrise-light, sunset-dark, ...}

- Fine-tuning : 새로운 세대는 처음부터 학습할 필요없이 이전 세대가 만든 논리들 그대로 학습하거나 미세조정하여 학습

Ex) **fire** burns you -> candles, wildfires, volcanoes, and every instance of 'fire'  
layer 1은 많은 양을 학습하고 지식을 Fine-tuned

- Layer 0 -> Layer 1 -> Layer 0 ... 이러한 과정을 무한히 반복하여 더 많은 법칙과 정교화된 지식을 학습



# The machine intelligence stack

- Layer 0(sequence-based learning)

- Input : 음성 -> 텍스트의 변환 과정을 거친 간접적인 정보

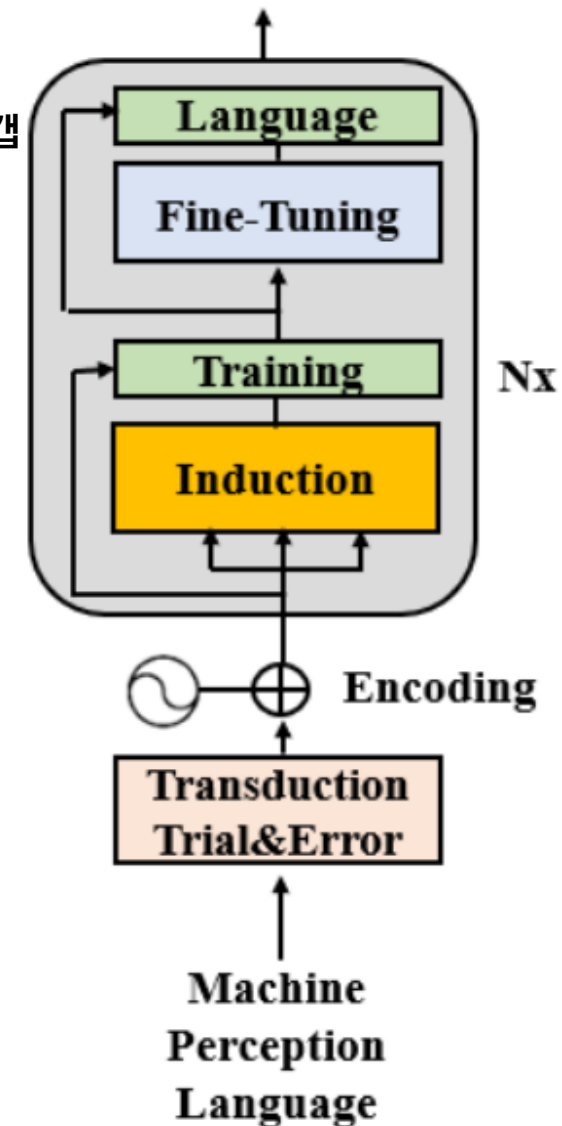
기계는 우리의 random quality language outputs에 의존해야하는 핸디캡

- Transduction Trial&error : 언어 sequence에서 함께 발생하는 모든 토큰(sub-words)들을 연결하는 transduction 수행

- Layer 1(Sub-layer, Transformer)

- Induction & Training : Transduction으로 만들어진 연결고리를 일반화하고 토큰들의 패턴 만듦

- 트랜스포머 모델은 이전 sequence 기반 학습 연산을 제외하고 모델의 시각을 높이기 위해 self-attention을 사용
- Attention sub-layer는 귀납적 사고 연산을 위해 수백만 개의 예들을 처리할 수 있다는 점에서 사람보다 이점
- 우리처럼, transduction과 induction을 통해 패턴을 찾음
- 모델에 저장된 파라미터와 함께 패턴들을 기억
- 상당한 데이터 양, 훌륭한 NLP Transformer 알고리즘, 컴퓨터 성능의 능력을 사용하여 언어 이해



## I Accuracy score

- 가장 실용적인 계산
- 단순히 true or false로 계산
- 실제값  $y$ 와 예측값  $\hat{y}$ 이 같을 경우 +1, 다를 경우 0
  - > 결과값을 다 더한 값에 sample의 개수  $n$ 을 나눈 것
- 전체 중 모델이 바르게 분류한 비율
- 모델의 예측 성능을 나타내는 매우 직관적인 평가 지표이지만 데이터의 구성에 따라 모델의 성능이 왜곡되게 평가될 수 있음

$$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i)$$

## F1-score

F1-score 불균일한 클래스 분포를 포함하는 데이터셋이 있을 때 도움이 될 수 있는 보다 유연한 접근 방식을 도입한다.

### Confusion Matrix :

Training을 통한 Prediction 성능을 측정하기 위해 예측 value와 실제 value를 비교하기 위한 표

- Precision : 예측값이 얼마나 정확한지

$$TP / TP + FP$$

- Recall : 실제 정답을 얼마나 맞췄는지

$$TP / TP + FN$$

ACTUAL VALUES

POSITIVE (1)

NEGATIVE (0)

PREDICTIVE VALUES

POSITIVE (1)    NEGATIVE (0)

	POSITIVE (1)	NEGATIVE (0)
POSITIVE (1)	TP	FN
NEGATIVE (0)	FP	TN

정밀도와 재현율 값의 조화평균으로 주로 분류 클래스 간의 데이터가 불균형이 심각할 때 사용

$$F1 \text{ score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

## Matthews Correlation Coefficient (MCC)

MCC는 class들의 사이즈가 다르더라도 훌륭한 binary metric을 제공

- -1은 모델이 클래스를 잘못 예측했다는 의미
- 0은 모델이 랜덤추측보다 잘 수행되지 않았음을 의미
- +1은 모델이 클래스를 바르게 예측했다는 의미

4개의 혼동 행렬 범주의 균형 비율을 고려하기 때문에 이진 분류 문제를 평가할 때 F1 score 및 정확도보다 더 많은 정보를 제공

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



## Transformer의 state-of-the-art 달성을 위한 3가지 조건

- A model (Bert, T5)
- A metric as described in the Evaluating models with metrics section of this chapter
- A dataset-driven task (GLUE, SuperGLUE)

## Benchmark Tasks

Benchmark는 어떤 대상의 성능을 측정할 때 기준이 되는 테스트이자 지표  
GLUE, SuperGLUE

SuperGLUE benchmark를 살펴봄으로 트랜스포머 모델의 성능 평가 과정을 알 수 있음

## General Language Understanding Evaluation (GLUE)

- NLU 모델들의 성능 평가를 위한 방법
- 하나의 모델로 여러 task들을 수행하고 그 값을 취합하여 최종 점수를 얻는 형태
- GLUE benchmark의 동기는 NLU가 다양한 task에 적용되어야 유용하다는 것을 보여주기 위해

The Corpus of Linguistic Acceptability

The Stanford Sentiment Treebank

Microsoft Research Paraphrase Corpus

Semantic Textual Similarity Benchmark

Quora Question Pairs

MultiNLI Matched

MultiNLI Mismatched

Question NLI

Recognizing Textual Entailment

Winograd NLI

Diagnostics Main

# The Corpus of Linguistic Acceptability (CoLA)

- 목표는 문장의 언어적 수용성을 판단하기 위해 NLP 모델의 언어적 능력을 평가하는 것이다.
- 문법적 수용성을 위해 주석이 달린 수천 개의 영어 문장 샘플이 포함되어있다.
- 문장이 문법적으로 수용할 수 없다면 그 문장은 0으로 분류되고 문법적으로 수용할 수 있다면 1로 분류 => 문장들이 문법적인지 아닌지를 분류한다.
- Measurement method는 MCC를 사용

Classification = 1 for 'we yelled ourselves hoarse.'

Classification = 0 for 'we yelled ourselves.'

- SST-2는 영화리뷰들을 포함한다.
- Datasets은 이진 분류를 넘어선 0 (negative) to n (positive) 범위로 감정을 분류할 수 있다.
- Measurement method는 Accuracy metric을 사용

```
#@title SST-2 Binary Classification
```

```
from transformers import pipeline
```

```
nlp = pipeline("sentiment-analysis")
```

```
print(nlp("If you sometimes like to go to the movies to have fun ,  
Wasabi is a good place to start."), "If you sometimes like to go to the  
movies to have fun , Wasabi is a good place to start.")
```

```
print(nlp("Effective but too-tepid biopic."), "Effective but too-tepid  
biopic.")
```

```
[{'label': 'POSITIVE', 'score': 0.999825656414032}] If you sometimes  
like to go to the movies to have fun , Wasabi is a good place to start
```

```
.
```

```
[{'label': 'NEGATIVE', 'score': 0.9974064230918884}] Effective but too-  
tepid biopic.
```

# Microsoft Research paraphrase Corpus (MRPC)

- 두 문장 쌍의 유사성을 판단하는 task
  - Paraphrase equivalent
  - Semantic equivalent
- Measurement method는 F1/Accuracy score method 사용

Sequence\_A = "The DVD-CCA then appealed to the state Supreme Court."

Sequence\_B = "The DVD CCA appealed that decision to the U.S. Supreme Court."




The DVD CCA appealed that decision to the U.S. Supreme Court. should be a paraphrase

not paraphrase: 8.0%

is paraphrase: 92.0%

# From GLUE to SuperGLUE





- SuperGLUE benchmark는 Wang et al (2019) 설계
- Wang은 처음에 General Language Understanding Evaluation (GLUE) benchmark 설계
- NLU 모델의 성능은 transformer의 등장으로 증가되어 졌고 사람 평균 수준을 초과하기 시작

Rank Name		Model	URL	Score
1	JDEExplore d-team	Vega v1		91.3
2	Microsoft Alexander v-team	Turing NLR v5		91.2
3	DIRL Team	DeBERTa + CLEVER		91.1
4	ERNIE Team - Baidu	ERNIE		91.1
5	AliceMind & DIRL	StructBERT + CLEVER		91.0
22	GLUE Human Baselines	GLUE Human Baselines		87.1

GLUE Leaderboard – March 2022

# Defining the SuperGLUE benchmark tasks





















- GLUE 문제는 불과 1년 만에 언어 모델의 성능이 인간 수준을 넘어버렸다는 것
- Wang은 이러한 GLUE의 한계를 보고 좀 더 어려운 NLU tasks를 위해 SuperGLUE 설계
- SuperGLUE의 목표는 NLU model이 fine-tuning으로 multiple downstream task들을 수행할 수 있다는 것을 보여주는 것

	Rank	Name	Model	URL	Score
+	1	Liam Fedus	ST-MoE-32B		91.2
	2	Microsoft Alexander v-team	Turing NLR v5		90.9
	3	ERNIE Team - Baidu	ERNIE 3.0		90.6
+	4	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4
+	5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3
	6	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8

SuperGLUE Leaderboard 2.0 – March 2022

# Defining the SuperGLUE benchmark tasks

- Wang은 SuperGLUE benchmark에 8개의 task들을 선택.
- 선택기준은 GLUE보다 엄격해졌다. 예를 들어 text을 이해하는 것 뿐만 아니라 추론이 필요
- 추론 수준은 최고수준의 전문가가 아니지만 성능 수준은 많은 human task들을 대체 가능

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

SuperGLUE tasks



## I Choice of Plausible Answers (COPA)

- 두가지 선택지 중 가장 그럴듯한 대답을 찾아내는 방법
- Dataset은 premise, question, 2개의 answer로 구성
- Metric : Accuracy

Premise: I knocked on my neighbor's door.

What happened as a result?

Alternative 1: My neighbor invited me in.

Alternative 2: My neighbor left his house.

## I BoolQ

- BoolQ는 Boolean yes or no answer task.
- SuperGLUE에 정의된 dataset는 자연적으로 발생하는 15,942개의 예가 포함되어있다.
- Dataset은 passage, question, answer(label)로 구성
- Matric : Accuracy

A raw sample of line #3 of the train.jsonl dataset contains a passage, a question, and the

```
{"question": "is windows movie maker part of windows essentials"  
"passage": "Windows Movie Maker -- Windows Movie Maker (formerly known  
as Windows Live Movie Maker in Windows 7) is a discontinued video  
editing software by Microsoft. It is a part of Windows Essentials  
software suite and offers the ability to create and edit videos as well  
as to publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr.",  
"idx": 2, "label": true}
```

## Commitment Bank (CB)

- 두 문장 간 논리적 관계 유형 판별 task
- Transformer model이 전제를 읽고 전제 위에 세워진 가설을 검증하도록 요청
- Transformer 모델은 가설을

전제에 대해 neutral(중립), entailment(함의), contradiction(모순) 이라고 분류한다.

- Dataset은 premise, hypothesis, label로 구성
- Metric : Avg. F1/Accuracy

The following sample, #77, taken from the training dataset, train.jsonl, shows how difficult the CB task is:

```
{"premise": "The Susweca. It means ''dragonfly'' in Sioux, you know.  
Did I ever tell you that's where Paul and I met?"  
"hypothesis": "Susweca is where she and Paul met,"  
"label": "entailment", "idx": 77}
```

## Multi-Sentence Reading Comprehension (MultiRC)

- MultiRC는 모델이 text를 읽고 몇 가지 가능한 선택지 중에서 고르는 task
- 모델은 text, 여러 개의 questions 그리고 0 (false) 또는 1 (true) label로 이루어진 각 질문에 대한 가능한 answers로 구성
- Metric : F1a / EM

Rolland's case.\" This is naive to the extreme -- Heston would not be president of the NRA if he was not kept up to date on the most prominent cases of gun violence. Even if he did not respond to that part of the interview, he certainly knew about the case at that point.

```
{"question": "Who was president of the NRA on February 29?",  
"answers": [{"text": "Charleton Heston", "idx": 173, "label": 1},  
{"text": "Moore", "idx": 174, "label": 0},  
{"text": "George Hoya", "idx": 175, "label": 0},  
{"text": "Rolland", "idx": 176, "label": 0},  
{"text": "Hoya", "idx": 177, "label": 0}, {"text": "Kayla", "idx": 178, "label": 0}], "idx": 27},
```

## Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD)

- Commonsense(상식)을 사용한 주어진 placeholder(빈 공간) 위치의 답 찾기
- Dataset는 70,000개 이상의 뉴스 기사에서 120,000개 이상의 쿼리가 포함
- Passage, query, answer로 구성
- Metric : F1/Accuracy

```
"text": "A Peruvian tribe once revered by the Inca's for their  
fierce hunting skills and formidable warriors are clinging on to their  
traditional existence in the coca growing valleys of South America,  
sharing their land with drug traffickers, rebels and illegal loggers.  
Ashaninka Indians are the largest group of indigenous people in the  
mountainous nation's Amazon region, but their settlements are so sparse  
that they now make up less than one per cent of Peru's 30 million
```

```
"entities": [{"start": 2, "end": 9}, ..., {"start": 711, "end": 715}]
```

```
{"query": "Innocence of youth: Many of the @placeholder's younger  
generations have turned their backs on tribal life and moved to the  
cities where living conditions are better",
```

```
"answers": [{"start": 263, "end": 271, "text": "Ashaninka"}, {"start": 601, "end":  
": 609, "text": "Ashaninka"}, {"start": 651, "end": 659, "text": "Ashaninka"}], "  
idx": 9}], "idx": 3}
```

## I Recognizing Textual Entailment (RTE)

- 두 문장 간 entailment 관계 판별 task
- Transformer model은 전제를 읽고 가설을 검증하고 entailment 가설 상태의 레이블을 예측  
즉 Entailment, not entailment 판별
- Premise, hypothesis, label로 구성
- Metric : Accuracy

**Premise :** The GDP report showed growth in business outlays advanced at a solid 8.9% pace

**Hypothesis :** The GDP was a disappointing report

**Label :** not\_entailment

## I Words in Context (WiC)

- 모호한 단어를 처리하는 모델의 성능을 테스트
- WiC에서 multi-task transformer는 두 문장을 분석하고 target word가 두 문장에서 같은 의미인지 결정한다.
- Word, 2개의 sentence, label로 구성
- Metric : Accuracy
- 먼저 target word가 지정된다: "word": "place"

```
"sentence1": "Do you want to come over to my place later?",  
"sentence2": "A political system with no place for the less prominent  
groups."
```

```
"idx": 0,  
"label": false,  
"start1": 31,  
"start2": 27,  
"end1": 36,  
"end2": 32,
```



## ■ The Winograd Schema Challenge (WSC)

- 주어진 대명사가 제시되는 특정 명사와 상호관계가 있는지 판단하는 task
- transformer가 잘 훈련되었다면 명확하지 않은 문제를 해결할 수 있어야 한다.
- Dataset는 대명사의 성별에 약간의 차이가 있는 문장이 포함
- 각 문장은 occupation, participant, pronoun을 포함.
- 문제를 풀기위해서는 대명사가 occupation이나 participant와 같은지 찾는 것
- Metric : Accuracy

```
{"text": "I poured water from the bottle into the cup until it was full.",
```

The WSC ask the model to find the target *pronoun* token number 10 starting at 0:

```
"target": {"span2_index": 10,
```

Then it asks the model to determine if "it" refers to "the cup" or not:

```
"span1_index": 7,
```

```
"span1_text": "the cup",
```

```
"span2_text": "it"},
```

For sample index #4, the label is true:

```
"idx": 4, "label": true}
```



Training set은 영어인데 만약 transformer model에게 English-French 번역에서 대명사 성별 문제를 해결하도록 요청하면 어떻게 될까? 프랑스어는 문법적 성별(여성형, 남성형)을 가진 명사의 철자가 다르다.

```
#@title Winograd
from transformers import pipeline
translator = pipeline("translation_en_to_fr")
print(translator("The car could not go in the garage because it was too big.", max_length=40))
```

```
[{'translation_text': "La voiture ne pouvait pas aller dans le garage parce qu'elle était trop grosse."}]
```

- *Elle*는 프랑스어로 그녀를 의미하는데 이것은 "it"의 번역이다. 남성형은 *il*
- *Grosse*는 big단어의 여성형 번역이다. 남성형은 *gros*이다.

사람 언어 표현 과정과 machine intelligence의 transduction 수행 방법의 차이를 분석했다.

Multi-task transformer의 성능을 어떻게 측정하는지 보았다.

Transformer를 사용한 NLU모델은 downstream task(Glue, SuperGlue)에서 최고의 성능을 보인다.  
기존의 GLUE Human baseline보다 높은 SuperGLUE task(BoolQ, CB, WiC)를 수행함으로써  
Transformer의 효율성을 증명

다음 챕터 Chapter 5. Machine Translation with the Transformer에서 번역 작업을 한 단계 더  
진행하여 Trax로 번역 모델을 구축할 것이다.