06

NLP 모델의 평가지표와 한국어 NLP

06 NLP 모델의 평가지표와 한국어 NLP

오늘 학습을 통해 우리는

- 딥러닝 모델의 다양한 평가지표를 탐구합니다.
- 자연어처리 모델의 일반적인 성능 측정을 위한 벤치마크 데이터셋에 무엇이 있는지를 알아봅니다.
- 한국어 자연어처리의 어려움과 이를 극복하기 위한 다양한 노력들을 알아봅니다.

목차

NLP 모델의
평가지표와 한국어
NLP

01

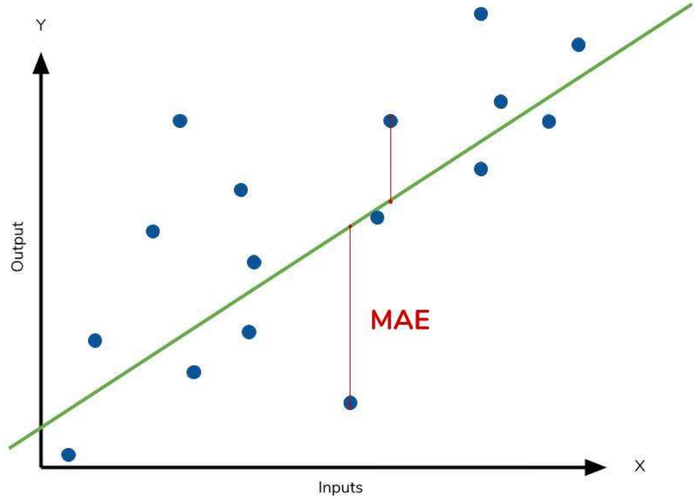
머신러닝 평가지표

머신러닝에서 평가지표의 역할

- 모델 간의 성능을 비교할 수 있는 객관적 수단
  - Task마다 주로 사용하는 평가지표가 상이함(Classification, regression 등)
- 학습 과정 모니터링
  - 모델이 학습하는 중 훈련 데이터셋과 검증 데이터에 대한 성능을 추적할 수 있음
  - 학습 중 발생하는 문제 또한 조기에 발견할 수 있음
- 모델 해석
  - 모델의 결과를 바탕으로 강점과 약점을 판단할 수 있음
  - 경우에 따라 모델의 신뢰도 또한 측정할 수 있음

평가지표의 다양성

- Regression 문제의 경우
  - 연속 변수 간의 차이를 계산
  - 예측값과 실제 값의 차이(Error)를 바탕으로 성능을 평가함
  - MAE(Mean Absolute Error 평균 절대 오차), MSE(Mean Squared Error 평균 제곱 오차)



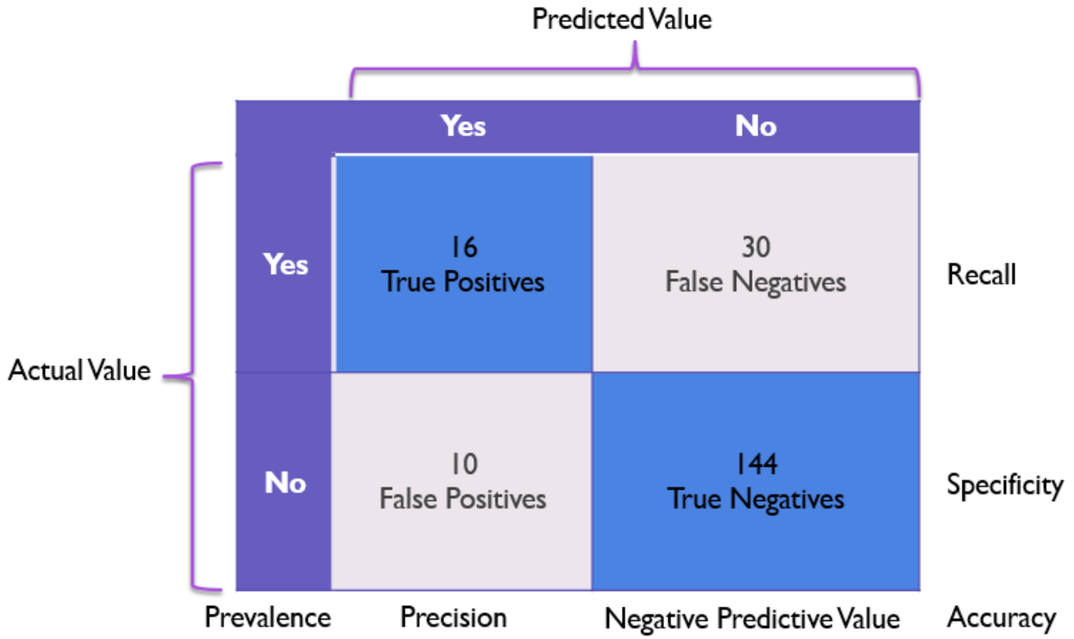
출처 : https://mizykk.tistory.com/102

평가지표의 다양성

- Classification 문제의 경우
  - 이산 변수 간의 차이를 계산
  - Accuracy, Precision, Recall, F1-score 등이 사용됨
  - 위의 평가지표들은 모두 Confusion Matrix를 바탕으로 계산할 수 있음

평가지표의 다양성 – Confusion Matrix

- 분류 문제에 관한 정답과 모델의 예측을 바탕으로, 모델의 성능을 확인할 수 있는 표
  - 모델이 이진 분류 문제를 풀었다고 가정할 경우,
  - 데이터의 정답과 모델의 예측 간 관계를 4개의 구역으로 표현(TP, TN, FP, FN)



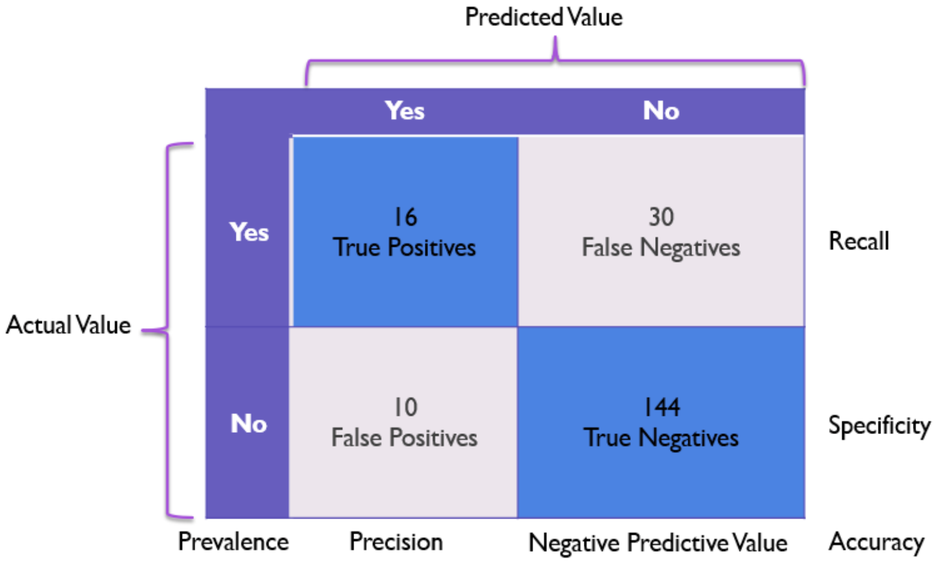
출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – Confusion Matrix

- 이진 분류에서, 클래스는 Positive와 Negative로 나뉨
  - Positive: 이진 분류 중, 실험자가 관심을 갖는 클래스
  - Negative: 실험자가 관심을 갖지 않는 클래스
- 정답과 예측의 일치여부는 True와 False로 나뉨
  - True: 정답과 예측이 일치
  - False: 정답과 예측이 불일치

평가지표의 다양성 – Confusion Matrix

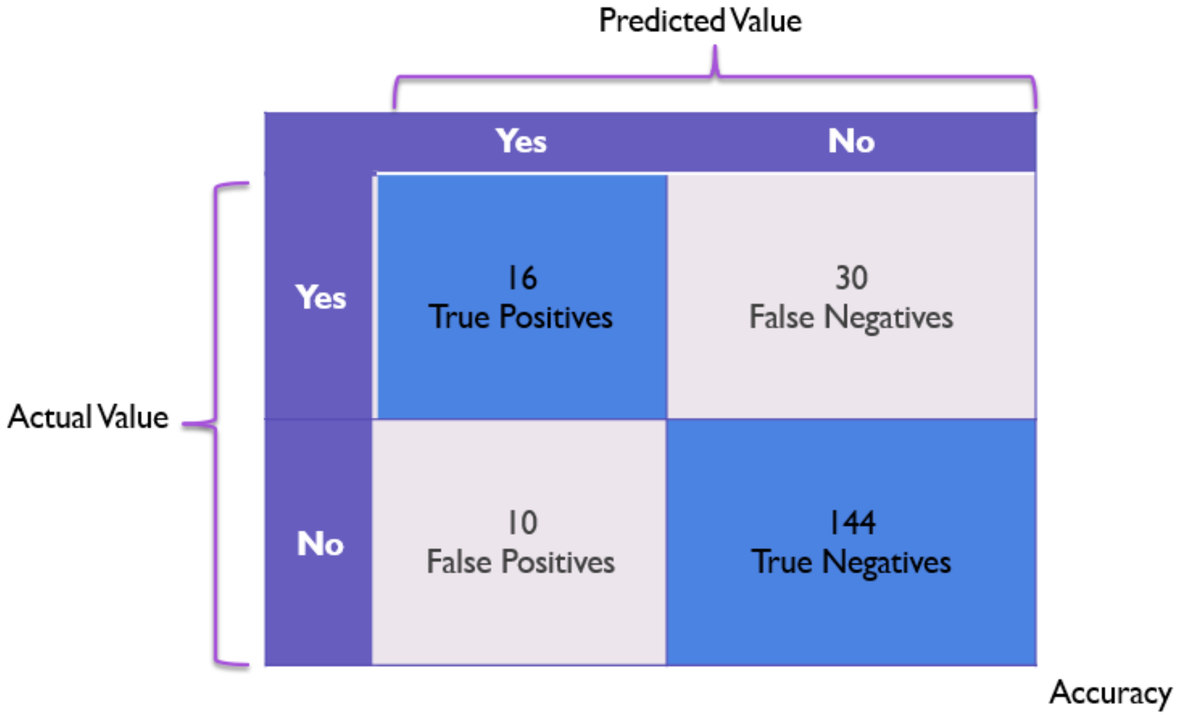
- 앞자리(일치 여부) + 뒷자리(예측 클래스)로 네 구역에 이름을 붙임
  - True Positive: Positive로 예측하여, 옳게 분류함
  - True Negative: Negative로 예측하여, 옳게 분류함
  - False Positive: Positive로 예측하여, 옳지 않게 분류함
  - False Negative: Negative로 예측하여, 옳지 않게 분류함



출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – Accuracy

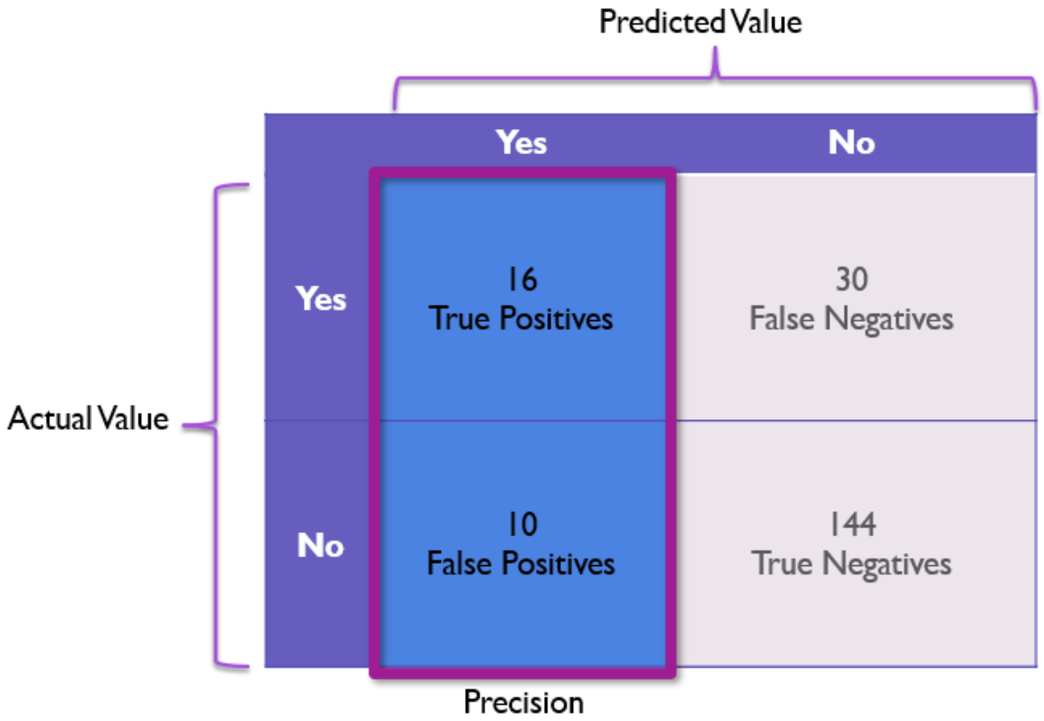
- 모든 예측 중, 정확한 예측의 비율
- TP + TN / TP + TN + FP + FN
- 장점: 쉽고 직관적인 지표
- 단점:
  - 클래스가 불균형할 경우, 모델의 실제 성능보다
    - 클래스 비율이 9:1인 데이터를 하나의 클래스로 전부 예측하는 경우
  - FP, FN을 고려하지 못함



출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – Precision(정밀도)

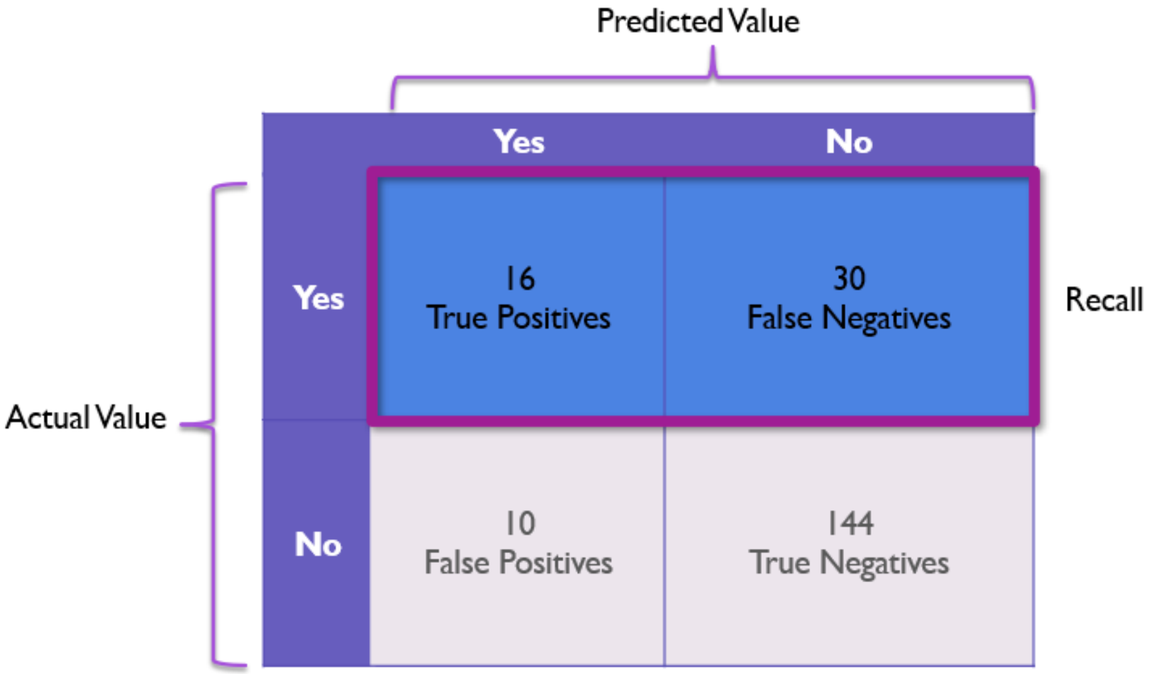
- 예측값이 양성인 것의 신뢰도
- TP / TP + FP
- Precision이 높을 수록, FP가 낮다는 뜻
- 단점
  - FN을 고려하지 못함
  - 정답이 양성이지만, 음성으로 예측한 경우를 감지하



출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – Recall(Sensitivity, 민감도)

- 모델이 양성 값을 찾아낼 수 있는 능력
- TP / TP + FN
- 민감도가 높다는 것은 FN이 낮다는 뜻
- 단점
  - FP를 고려하지 못함
  - 정답이 음성이지만, 양성으로 예측한 경우를 감



출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – Precision/Recall의 Trade-off

- Precision과 Recall은 Trade-off 관계
  - Precision = TP / TP + FP
  - Recall = TP / TP + FN
- 두 지표는 Accuracy의 단점을 보완하기에 좋으므로, 두 지표를 모두 사용하는 방법이 고안됨
  - F1-Score
  - ROC Curve와 AUC
  - PR Curve와 AP

출처 : https://accessibleai.dev/post/interpreting_confusion_matrixes/

평가지표의 다양성 – F1-Score

- Precision과 Recall의 조화평균
  - $\text{F1 Score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+ Recall}}$
- FP, FN에 모두 반응하는 평가지표
- 불균형한 데이터셋에도 효과가 좋음
- 그러나 두 지표를 동등하게 고려하므로, 어느 하나가 월등하게 좋을 경우, 나머지의 수치가 높지 않더라도 좋은 결과를 반환함

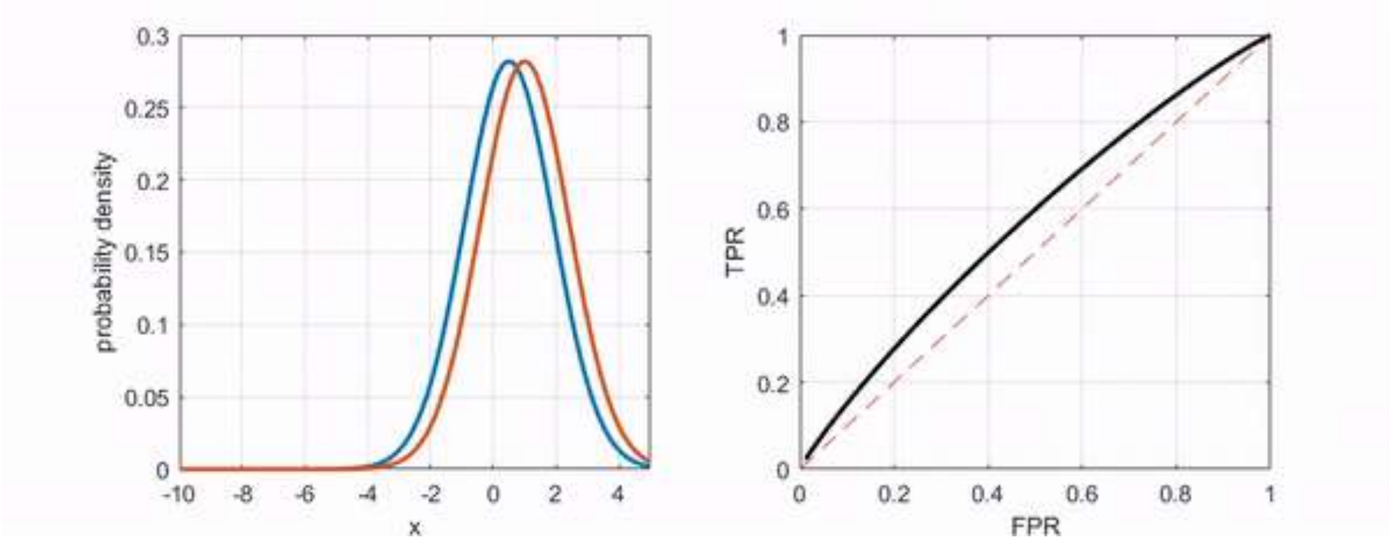
평가지표의 다양성 – ROC Curve와 AUC

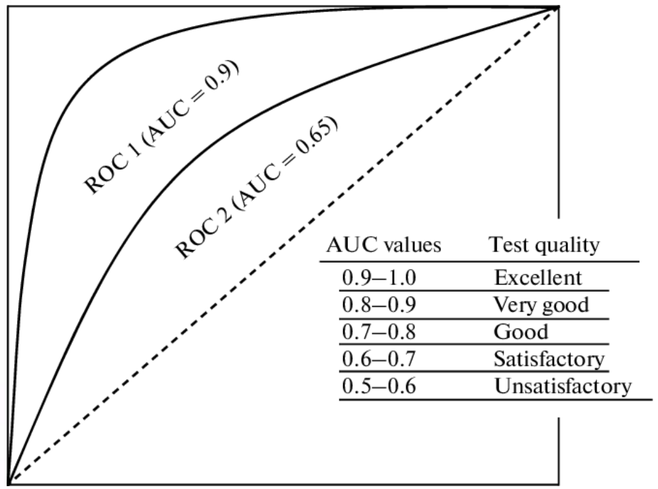
- 모델이 Positive/Negative 예측할 때의 기준치(Threshold)를 조정하면, Confusion matrix의 결과가 달라짐
  - 즉 TP와 FP의 비율이 바뀜

- Threshold의 변화에 따른 TPR(True Positive Rate)과 FPR(False Positive Rate)의 관계를 ROC Curve(Receiver Operating Characteristic, 수신자 조작 특성 곡선)라 함

평가지표의 다양성 – ROC Curve와 AUC

- 모델의 성능이 좋을 수록 곡선이 좌측 상단으로 휨
- 성능이 좋지 않을 수록, 우상향하는 직선의 형태에 가까워짐



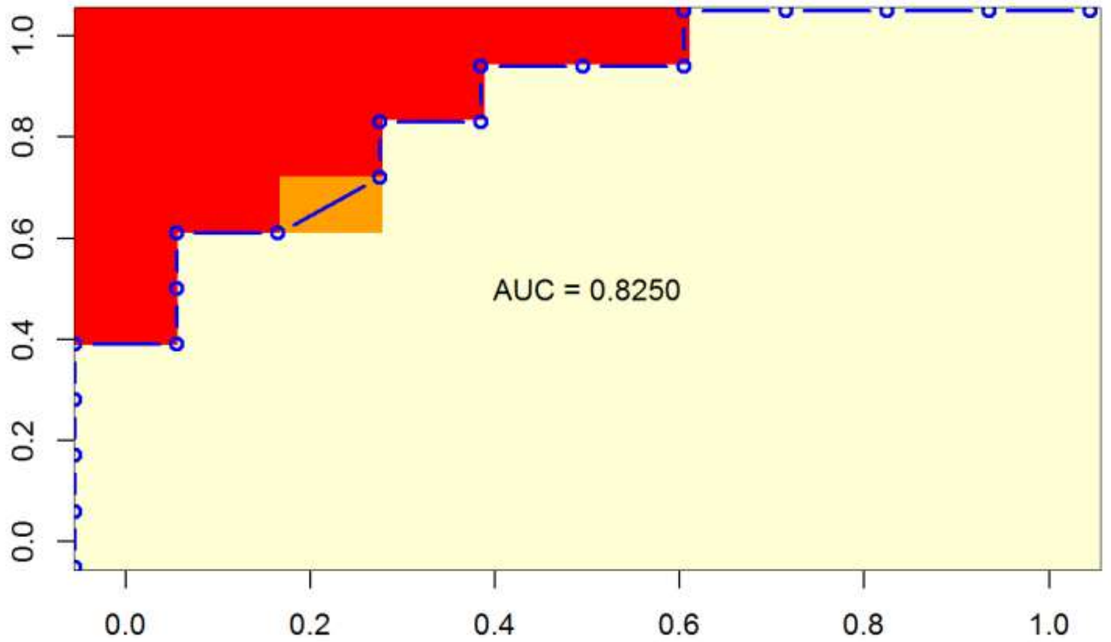


AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

출처 : https://angeloyeo.github.io/2020/08/05/ROC.html
https://www.researchgate.net/figure/An-example-of-ROC-curves-with-good-AUC-09-and-satisfactory-AUC-065-parameters_fig2_276079430

평가지표의 다양성 – ROC Curve와 AUC

- AUC(Area Under Curve)란 ROC Curve 아래의 면적을 의미하며 0-1사이 값으로 표현
  - AUC=1: 분류 성능이 완벽
  - AUC=0.5: 모델이 무작위로 분류하는 것과 같음
  - AUC=0: 모델이 완벽하게 잘못 분류



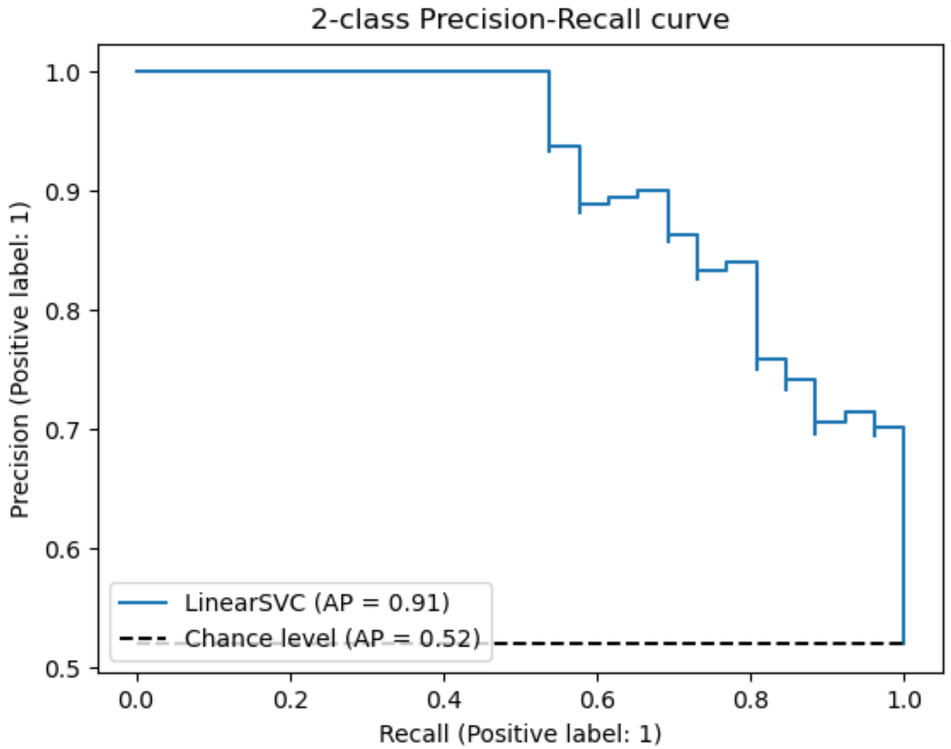
출처 : https://blog.revolutionanalytics.com/2016/11/calculating-auc.html

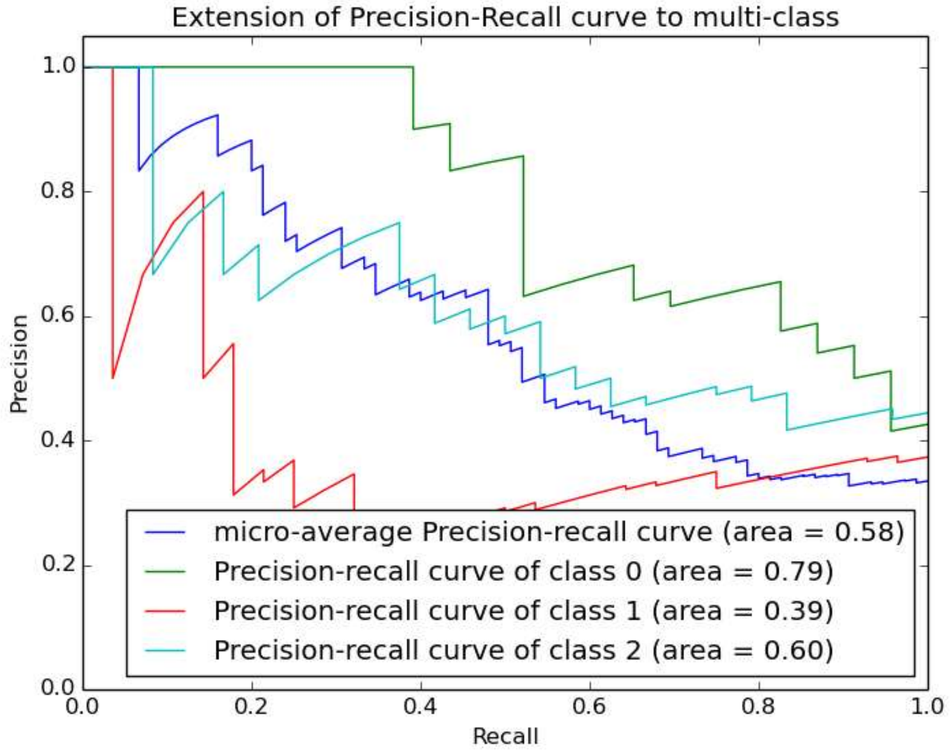
평가지표의 다양성 – PR Curve와 AP

- Precision – Recall은 Trade-off 관계이므로, 모델이 Positive/Negative 예측할 때의 기준치(Threshold)를 조정하면, Confusion matrix의 결과가 달라짐
  - 즉 FP와 FN의 비율이 바뀜
- Threshold의 변화에 따른 Precision과 Recall의 관계를 Precision-Recall Curve라 함
  - Precision – Recall curve에서는 두 지표가 모두 높을수록 우수한 모델을 의미
  - 그래프가 우상향하는 경우

출처: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

평가지표의 다양성 – PR Curve와 AP

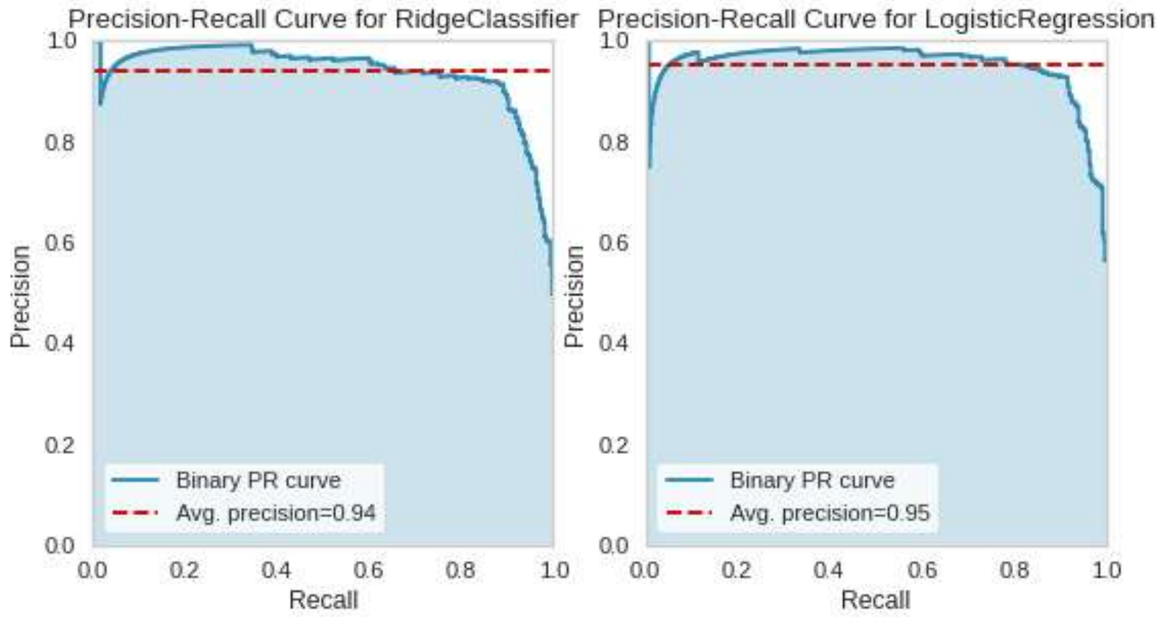




출처: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
https://scikit-learn.org/0.15/auto_examples/plot_precision_recall.html

평가지표의 다양성 – PR Curve와 AP

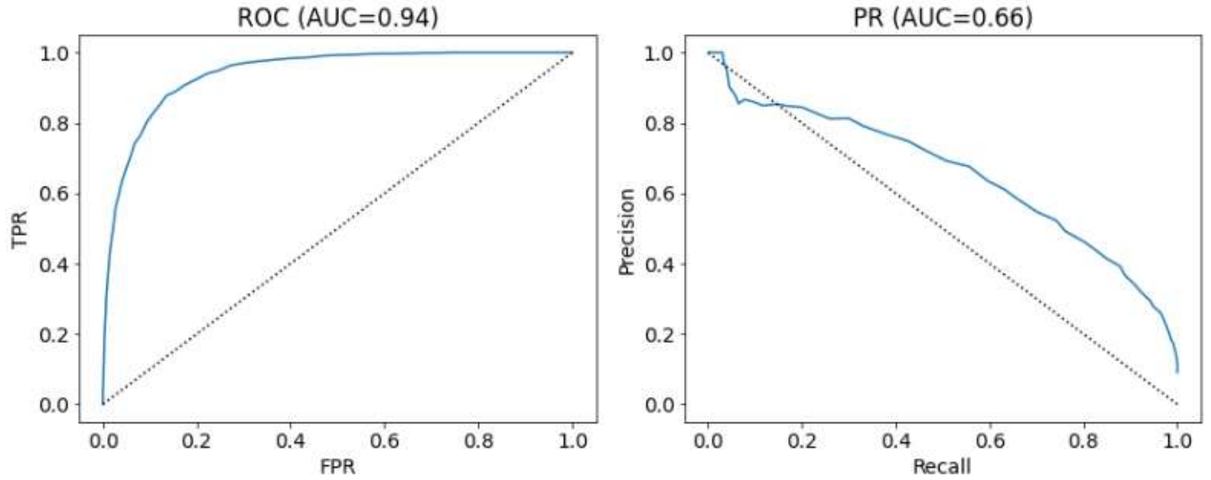
- 또는 PR-Curve의 아랫부분 면적을 구하면 모델의 성능을 평가할 수 있음
  - 이 면적을 AP(Average Precision)이라 부름
  - AP는 0-1 사이의 값으로 표현되며, 1에 가까울 수록 모델 성능이 우수함



출처 : https://www.scikit-yb.org/en/latest/api/classifier/prcurve.html

평가지표의 다양성 – PR Curve v.s. ROC Curve

- PR curve는 Precision과 Recall을 바탕으로 계산
  - 즉 FP와 FN 모두 엄격하게 평가하므로, 불균형한 데이터셋에 민감하게 반응함
- ROC curve는 TP와 FP의 비율을 바탕으로 계산
  - 즉 FN을 고려하지 않기에 모델 성능을 낙관적으로 평가함
  - 일반적인 모델 성능 평가에 유용함



출처 : https://data-minggeul.tistory.com/10

평가지표의 다양성

- 텍스트 생성 모델: 생성된 텍스트와 참조 텍스트 사이의 유사도를 측정
  - BLEU score, ROUGE, Cosine similarity 등
- 이미지 생성 모델: 주로 비지도 학습이므로 손실값 바탕의 평가지표를 활용
  - Reconstruction error, KL Divergence 등
- 추천 시스템: 사용자의 선호도를 벡터로 표현하기 때문에, 행렬 간 유사도를 측정
  - Cosine similarity, MAP(Mean Average Precision) 등

01 머신러닝 평가지표

BLEU Score

- BiLingual Evaluation Understudy
  - 기계번역(Machine translation)의 출력물이 얼마나 자연스러운지를 평가하는 지표
  - 즉 얼마나 사람의 번역과 유사한지를 측정
  - [0, 1]사이에서 표현
    - 0.4 이상이면 우수한 것으로 간주
- WMT English-German 데이터셋
  - Transformer: 28.4
  - Facebook FAIR: 43.1

BLEU Score

- BiLingual Evaluation Understudy
  - 번역문에 대해 두 가지 측면에서 평가를 함
    - n-gram 정밀도: 번역된 문장의 일부분(단어들의 연속)이 원래 문장과 얼마나 비슷한지
    - 브레버티 패널티(Brevity Penalty): 번역된 문장이 너무 짧으면, 점수를 조금 깎음

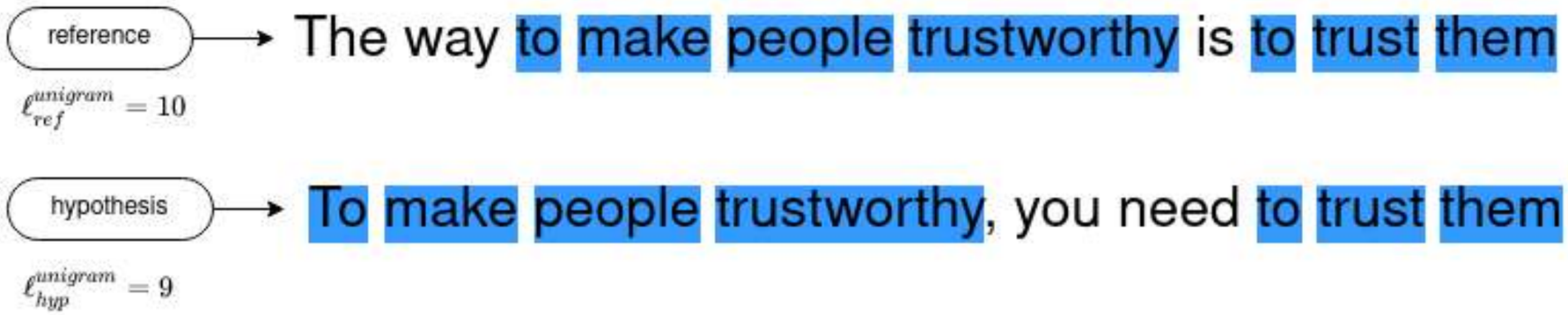
$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

    - BP = Brevity Penalty
    - w_n = n-gram에 대한 가중치
    - p_n = n-gram 정밀도(Precision)
  - 단어 재현 능력만 볼 뿐, 문장의 유사도나 의미는 전혀 고려하지 않음

01 머신러닝 평가지표

BLEU Score

- N-gram: 일반적으로 1부터 4까지 계산



n-gram	1-gram	2-gram	3-gram	4-gram
p_n	$\frac{7}{9}$	$\frac{5}{8}$	$\frac{3}{7}$	$\frac{1}{6}$

출처 : https://clementbm.github.io/theory/2021/12/23/rouge-bleu-scores.html

01 머신러닝 평가지표

ROUGE

- Recall-Oriented Understudy for Gisting Evaluation
  - 문서 요약 작업에서 사용되는 평가 지표
    - 요약된 문서와 정답 간의 유사도 측정
  - 세 가지 점수로 평가
    - 정밀도(Precision): 생성된 요약에서 정답과 일치하는 단어나 구의 비율
    - 재현율(Recall): 정답에서 생성된 요약과 일치하는 단어나 구의 비율
    - F1 점수: 정밀도와 재현율의 조화 평균

01 머신러닝 평가지표

ROUGE

- Recall-Oriented Understudy for Gisting Evaluation
  - 두 문서를 세 가지 관점으로 평가
    - ROUGE-N: n-gram 정밀도와 재현율을 기반
      - Ex)ROUGE-1은 1-gram (단어) 일치를, ROUGE-2는 2-gram 일치를 계산
    - ROUGE-L: LCS(Longest Common Subsequence) 기반
      - LCS: 두 문장 사이에서 순서를 변경하지 않고 얻을 수 있는 가장 긴 연속적인 단어의 시퀀스
    - ROUGE-S: 문장 내의 단어 순서를 고려하지 않는 skip-bigram을 기반
      - Skip-bigram은 문장 내의 두 단어 사이에 몇 개의 단어가 건너뛰어져 있더라도 그것들을 bigram으로 간주

02

GLUE와 KLUE

©Elice. All Rights Reserved

GLUE

- General Language Understanding Evaluation
  - 다양한 자연어 처리(NLP) 작업을 위한 9가지 벤치마크 데이터셋
- 목적
  - 자연어 처리 모델의 성능을 광범위하게 평가하기 위한 표준화된 방법 제공
  - 다양한 NLP 작업에 대한 모델의 성능을 종합적으로 비교
  - 연구 커뮤니티에 공통의 평가 기준 제공

출처 : https://gluebenchmark.com/

GLUE

- 9개의 다양한 NLP 작업 포함
  - 문장 관계 분석, 의도 분석, 감정 분석 등
- 다양한 난이도와 도메인의 작업 포함
  - Single-Sentence tasks: CoLA, SST-2
  - Similarity and Paraphrase tasks: MRPC, QQP, STS-B
  - Inference tasks: MNLI, QNLI, RTE, WNLI
- 모델의 범용성과 다양한 작업에 대한 적응성 평가

GLUE - CoLA

- Corpus of Linguistic Acceptability
  - 문장의 문법적 수용성을 평가하기 위한 데이터셋
  - 문장의 의미나 내용이 아닌, 순수하게 문법적 수용성에 초점
  - 딥러닝 모델이 언어의 미묘한 문법적 차이를 이해하는지 평가 가능
- 이진 분류 작업 (Binary Classification)
  - 문장이 문법적으로 수용 가능한지 (1)
  - 문법적으로 수용 불가능한지 (0) 판단
  - 평가지표: Matthew's Corr

0	"*"	"He gets angry, the longer John has to wait."
0	"*"	"He gets angry if John has to wait."
1	null	"The more that pictures of him appear in the news, the more embarrassed John becomes."
1	null	"The more pictures of himself that appear in the news, the more embarrassed John becomes."

출처 : https://huggingface.co/datasets/shivkumarganesh/CoLA/viewer/shivkumarganesh--CoLA/train?p=2

GLUE - SST

- Stanford Sentiment Treebank
  - 영화 리뷰의 감정을 분석하기 위한 데이터셋
  - 문장 전체뿐만 아니라, 부분 부분의 감정도 라벨링
  - 이진 분류 (Binary Classification)
    - 긍정적인 리뷰 (1)
    - 부정적인 리뷰 (0)
  - 5단계 감정 분류
    - 매우 부정적, 부정적, 중립, 긍정적, 매우 긍정적
  - 평가지표: Accuracy

"contains no wit , only labored gags "	0 (negative)
"that loves its characters and communicates something rather beautiful about human nature "	1 (positive)
"remains utterly satisfied to remain the same throughout "	0 (negative)
"on the worst revenge-of-the-nerds clichés the filmmakers could dredge up "	0 (negative)
"that 's far too tragic to merit such superficial treatment "	0 (negative)
"demonstrates that the director of such hollywood blockbusters as patriot games can still turn out a small , personal film with an emotional wallop . "	1 (positive)
"of saucy "	1 (positive)
"a depressed fifteen-year-old 's suicidal poetry "	0 (negative)
"are more deeply thought through than in most ` right-thinking ' films "	1 (positive)

출처 : https://huggingface.co/datasets/sst2

GLUE - MRPC

- Microsoft Research Paraphrase Corpus
  - 문장 간의 동일한 의미를 가지는지 평가하기 위한 데이터셋
  - 문장 간의 미묘한 차이와 유사성을 탐지하는 능력 평가
- 이진 분류 (Binary Classification)
  - 문장 쌍이 동일한 의미를 가지는지(1)
  - 동일한 의미를 가지지 않는지(0)
- 평가 지표: F1 / Accuracy

"Amrozi accused his brother , whom he called " the witness " , of deliberately…	"Referring to him as only " the witness " , Amrozi accused his brother of…	1	0	"equivalent"
"Yucaipa owned Dominick 's before selling the chain to Safeway in 1998 for $ 2.5…	"Yucaipa bought Dominick 's in 1995 for $ 693 million and sold it to Safeway for $…	0	1	"not equivalent"
"They had published an advertisement on the Internet on June 10 , offering the…	"On June 10 , the ship 's owners had published an advertisement on the Interne…	1	2	"equivalent"
"Around 0335 GMT , Tab shares were up 19 cents , or 4.4 % , at A $ 4.56 , having…	"Tab shares jumped 20 cents , or 4.6 % , to set a record closing high at A $ 4.57…	0	3	"not equivalent"
"The stock rose $ 2.11 , or about 11 percent , to close Friday at $ 21.51 on…	"PG & E Corp. shares jumped $ 1.63 or 8 percent to $ 21.03 on the New York Stock…	1	4	"equivalent"

출처 : https://huggingface.co/datasets/SetFit/mrpc

GLUE – STS-B

- Semantic Textual Similarity Benchmark
  - 문장 쌍 간의 의미적 유사도를 평가하기 위한 데이터셋
  - 문장 간의 의미적 유사도를 정량적으로 평가
- 회귀 분석 (Regression)
  - 문장 쌍의 유사도를 0~5 사이의 값으로 예측
  - 평가지표: Pearson-Spearman Corr

"if you don 't want to pay for adobe acrobat pro , as @ schultem mentions , latex can do this wit…	"if you don 't mind hosting your files online , slideshare is a good solution ."	0.8
"i don 't think there are likely to be any standards that address this issue specifically ."	"you 're going to find answers all over the map for this one (i.e. , there probably aren 't "…	2.4
"my answer would be depending on which gre are you referring to ?"	"the problem i see with the gres is that the scoring range is highly compressed ."	1

출처 : https://huggingface.co/datasets/mteb/sts15-sts

GLUE – QQP

- Quora Question Pairs
  - Quora 플랫폼에서 수집된 질문 쌍의 유사성을 평가하기 위한 데이터셋
  - 질문의 의미와 구조를 파악하는 능력 평가
  - 딥러닝 모델이 유사한 질문을 얼마나 잘 구별하는지 평가 가능
- 이진 분류 (Binary Classification)
  - 질문 쌍이 동일한 주제에 관한 것인지(1)
  - 동일한 주제에 관한 것이 아닌지(0)
  - 평가 지표: F1 / Accuracy

"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…	0	4	"not duplicate"
"How not to feel guilty since I am Muslim and I'm conscious we won't have sex…	"I don't beleive I am bulimic, but I force throw up atleast once a day after I eat…	0	5	"not duplicate"
"How is air traffic controlled?"	"How do you become an air traffic controller?"	0	6	"not duplicate"

출처 : https://huggingface.co/datasets/SetFit/qqp

GLUE - MNLI

- MultiNomial NLI
  - MultiNomial Natural Language Inference
  - 문장 간의 논리적 관계를 분석
- 다중 분류 (Multi-Class Classification)
  - 문장 A가 문장 B를 함축하는지(entailment)
  - 문장 A와 문장 B가 중립적인 관계인지(neutral)
  - 문장 A가 문장 B와 모순되는지(contradiction) 판단
  - 평가지표: Accuracy

"Conceptually cream skimming has two basic dimensions - product and geography."	"Product and geography are what make cream skimming work. "	1	0	"neutral"
"you know during the season and i guess at at your level uh you lose them to the nex…	"You lose the things to the following level if the people recall."	0	1	"entailment"
"One of our number will carry out your instructions minutely."	"A member of my team will execute your orders with immense precision."	0	2	"entailment"
"How do you know? All this is their information again."	"This information belongs to them."	0	3	"entailment"
"yeah i tell you what though if you go price some of those tennis shoes i can se…	"The tennis shoes have a range of prices."	1	4	"neutral"
"my walkman broke so i'm upset now i just have to turn the stereo up real loud"	"I'm upset that my walkman broke and now I have to turn the stereo up really loud."	0	5	"entailment"
"But a few Christian mosaics survive above the apse is the Virgin with the infant…	"Most of the Christian mosaics were destroyed by Muslims. "	1	6	"neutral"

출처 : https://huggingface.co/datasets/SetFit/mnli

GLUE - QNLI

- Question Natural Language Inference
  - Stanford Question Answering Dataset (SQuAD)에서 변환된 데이터셋
  - 질문과 해당 질문의 답변이 될 수 있는 문장 쌍으로 구성
  - 질문 응답 시스템의 중요한 구성 요소
  - 이진 분류 (Binary Classification)
    - 문장이 질문에 대한 답변인지(entailment)
    - 답변이 아닌지(not_entailment) 판단
  - 평가지표: Accuracy

"When did the third Digimon series begin?"	"Unlike the two seasons before it and most of the seasons that followed, Digimon…	1	0	"not entailment"
"Which missile batteries often have individual launchers several kilometres…	"When MANPADS is operated by specialists, batteries may have several dozen teams…	1	1	"not entailment"
"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 as the one described…	0	2	"entailment"
"What is the name of the village 9 miles north of Calafat where the Ottoman forces…	"On 31 December 1853, the Ottoman forces at Calafat moved against the Russian forc…	0	3	"entailment"
"What famous palace is located in London?"	"London contains four World Heritage Sites: the Tower of London; Kew Gardens;…	1	4	"not entailment"
"When is the term 'German dialects' used in regard to the German language?"	"When talking about the German language, the term German dialects is only used for…	0	5	"entailment"
"What was the name of the island the English traded to the Dutch in return for…	"At the end of the Second Anglo-Dutch War, the English gained New Amsterdam (New…	0	6	"entailment"

출처 : https://huggingface.co/datasets/SetFit/qnli

GLUE - RTE

- Recognizing Textual Entailment
  - 두 텍스트 간의 논리적 함축 관계를 인식하기 위한 데이터셋
- 이진 분류 (Binary Classification)
  - 한 문장이 다른 문장을 함축하는지(entailment)
  - 함축하지 않는지(not_entailment)
  - 평가 지표: Accuracy

"No Weapons of Mass Destruction Found in Iraq Yet."	"Weapons of Mass Destruction Found in Iraq."	1	0	"not entailment"
"A place of sorrow, after Pope John Paul II died, became a place of celebration, a…	"Pope Benedict XVI is the new leader of the Roman Catholic Church."	0	1	"entailment"
"Herceptin was already approved to treat the sickest breast cancer patients, and…	"Herceptin can be used to treat breast cancer."	0	2	"entailment"
"Judie Vivian, chief executive at ProMedica, a medical service company that…	"The previous name of Ho Chi Minh City was Saigon."	0	3	"entailment"
"A man is due in court later charged with the murder 26 years ago of a teenager…	"Paul Stewart Hutchinson is accused of having stabbed a girl."	1	4	"not entailment"
"Britain said, Friday, that it has barred cleric, Omar Bakri, from returning to the…	"Bakri was briefly detained, but was released."	0	5	"entailment"
"Nearly 4 million children who have at least one parent who entered the U.S.…	"Three quarters of U.S. illegal immigrants have children."	1	6	"not entailment"

GLUE - WNLI

- Winograd NLI
  - Winograd 스키마 챌린지를 기반으로 한 자연어 이해 평가 데이터셋
  - 대명사가 포함된 문장을 읽고, 대명사가 가리키는 대상을 원문에서 고르는 챌린지
  - 이진 분류 (Binary Classification)
    - 주어진 문장이 다른 문장의 대명사 참조를 올바르게 해결하는지 판단
  - 평가 방식: Accuracy

"I stuck a pin through a carrot. When I pulled the pin out, it had a hole."	"The carrot had a hole."	1	0	"entailment"
"John couldn't see the stage with Billy in front of him because he is so short."	"John is so short."	1	1	"entailment"
"The police arrested all of the gang members. They were trying to stop the dru…	"The police were trying to stop the drug trade in the neighborhood."	1	2	"entailment"
"Steve follows Fred's example in everything. He influences him hugely."	"Steve influences him hugely."	0	3	"not entailment"
"When Tatyana reached the cabin, her mother was sleeping. She was careful not…	"mother was careful not to disturb her, undressing and climbing back into her…	0	4	"not entailment"

출처 : https://huggingface.co/datasets/SetFit/wnli

KLUE

- Korean Language Understanding Evaluation
  - 한국어 자연어 처리(NLP) 작업들을 위한 평가 벤치마크
- 목적
  - 한국어는 그 구조와 특성이 영어와 다르기 때문에 별도의 평가 도구 필요
  - 한국어 모델의 성능과 범용성을 체계적으로 평가할 수 있는 플랫폼 제공
  - 연구자들 간의 경쟁을 통해 한국어 모델 성능 향상 도모

출처 : https://paperswithcode.com/dataset/klue

KLUE - DP

- Dependency Parsing
  - 의존 구문 분석 작업
  - 문장 내 단어들 간의 구문적 관계를 분석
  - 평가
    - 각 단어에 대해 올바른 의존 관계와 라벨을 예측
    - 평가지표: UAS(Unlabeled Attachment Score)와LAS (Labeled Attachment Score)

"해당 그림을 보면 디즈니 공주들이 브리트니 스피어스의 앨범이나 뮤직비디오, 화보 속 모습을 똑같이 재연했다."	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]	["해당", "그림을", "보면", "디즈니", "공주들이", "브리트니", "스피어스의", "앨범이나", "뮤직비디오,", "화보", "속", "모습을", "똑같이", "재연했다."]	["해당", "그림 을", "보면", "디즈니", "공주 들이", "브리트니", "스피어스의", "앨범 이나", "뮤직 비디오 ,", "화보", "속", "모습 을", "똑같이", "재연 하였 다 ."]

출처 : https://huggingface.co/datasets/klue/viewer/dp/train?row=0

KLUE – DST(WoS)

- 대화 상태 추적(Dialogue State Tracking) 작업
  - WOS(Wizard of Seoul)는 KLUE-DST에서 사용되는 대화 데이터셋의 이름
    - 여러 대화 세션들과 각 대화 턴에서의 상태 라벨 포함
    - 사용자와 시스템 간의 대화로 구성, 각 대화 턴마다 상태 정보 업데이트
  - 평가 지표
    - Joint Goal Accuracy
    - Slot Accuracy


```
[ "관광", "식당" ]
```

```
[ { "role": "user", "text": "서울 중앙에 있는 박물관을 찾아주세요", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙" ] }, { "role": "sys", "text": "안녕하세요. 문화역서울 284은 어떠신가요? 평점도 4점으로 방문객들에게 좋은 평가를 받고 있습니다.", "state": [] }, { "role": "user", "text": "좋네요 거기 평점은 말해주셨구 전화번호가 어떻게되나요?", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284" ] }, { "role": "sys", "text": "전화번호는 983880764입니다. 더 필요하신 게 있으실까요?", "state": [] }, { "role": "user", "text": "네 관광지와 같은 지역의 한식당을 가고싶은데요 야외석이 있어야되요", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes" ] }, { "role": "sys", "text": "생각하고 계신 가격대가 있으신가요?", "state": [] }, { "role": "user", "text": "음.. 저렴한 가격대에 있나요?", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-가격대-저렴", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes" ] }, { "role": "sys", "text": "죄송하지만 저렴한 가격대에는 없으시네요.", "state": [] }, { "role": "user", "text": "그럼 비싼 가격대로 다시 찾아주세요", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-가격대-비싼", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes" ] }, { "role": "sys", "text": "외계인의맛집은 어떠신가요? 대표 메뉴는 한정식입니다.", "state": [] }, { "role": "user", "text": "좋습니당 토요일 18:00에 1명 예약가능한가요?", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-가격대-비싼", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes", "식당-예약 요일-토요일", "식당-예약 시간-18:00", "식당-예약 명수-1", "식당-이름-외계인의맛집" ] }, { "role": "sys", "text": "가능합니다. 예약도와드릴까요?", "state": [] }, { "role": "user", "text": "넹 거기 주류는 판매하나요?주차는 가능한가요?", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-가격대-비싼", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes", "식당-예약 요일-토요일", "식당-예약 시간-18:00", "식당-예약 명수-1", "식당-이름-외계인의맛집" ] }, { "role": "sys", "text": "주류는 판매하고 있고 주차도 가능합니다. 더 궁금하신 점 있으신가요?", "state": [] }, { "role": "user", "text": "아니용", "state": [ "관광-종류-박물관", "관광-지역-서울 중앙", "관광-이름-문화역서울 284", "식당-가격대-비싼", "식당-지역-서울 중앙", "식당-종류-한식당", "식당-야외석 유무-yes", "식당-예약 요일-토요일", "식당-예약 시간-18:00", "식당-예약 명수-1", "식당-이름-외계인의맛집" ] }, { "role": "sys", "text": "감사합니다. 즐거운 여행되세요.", "state": [] } ]
```

KLUE - MRC

- 기계 독해 (Machine Reading Comprehension)
  - 주어진 문서 내용을 바탕으로 질문에 대한 답변을 추출하거나 생성하는 작업
  - 여러 문서 및 문서에 대한 질문-답변 쌍 포함
    - 답변은 문서 내의 특정 부분을 가리키거나, 문서 내용을 기반으로 생성됨
  - 평가 방법
    - 주어진 질문에 대해 올바른 문서 내 위치나 답변을 예측
    - 평가 지표: Exact Match (EM), F1 Score

title (string)	context (string)
"제주도 장마 시작 … 중부는 이달 말부터"	"올여름 장마가 17일 제주도에서 시작됐다. 서울 등 중부지방은 예년보다 사나흘 정도 늦은 이달 말께 장마가 시작될 전망이다.17일 기상청에 따르면 제주도 남쪽 먼 바다에 있는 장마전선의 영향으로 이날 제주도 산간 및 내륙지역에 호우주의보가 내려지면서 곳곳에 100㎜에 육박하는 많은 비가 내렸다. 제주의 장마는 평년보다 2~3일, 지난해보다는 하루 일찍 시작됐다. 장마는 고온다습한 북태평양 기단과 한랭 습윤한 오호츠크해 기단이 만나 형성되는 장마전선에서 내리는 비를 뜻한다.장마전선은 18일 제주도 먼 남쪽 해상으로 내려갔다가 20일께 다시 북상해 전남 남해안까지 영향을 줄
question (string)	**answers (sequence)**
"북태평양 기단과 오호츠크해 기단이 만나 국내에 머무르는 기간은?"	{ "answer_start": [478, 478], "text": ["한 달가량", "한 달"] }

출처 : https://huggingface.co/datasets/klue/viewer/mrc/train?row=0

KLUE - NER

- 개체명 인식(Named Entity Recognition)
  - 주어진 텍스트에서 특정 카테고리의 개체명(사람 이름, 기관 이름, 날짜 등)을 인식하고 분류하는 작업
  - 주어진 문장에서 올바른 개체명과 그 카테고리를 예측
  - 평가지표: Precision, Recall, F1 Score

sentence (string)	tokens (sequence)	ner_tags (sequence)
"특히 <영동고속도로:LC> <강릉:LC> 방향 <문막휴게소:LC>에서 <만종분기점:LC>까지 <5㎞:QT> 구간에는 승용차 전용 임시 갓길차로제를 운영하기로 했다."	["특", "히", " ", "영", "동", "고", "속", "도", "로", " ", "강", "릉", " ", "방", "향", " ", "문", "막", "휴", "게", "소", "에", "서", " ", "만", "종", "분", "기", "점", "까", "지", " ", "5", "㎞", " ", "구", "간", "에", "는", " ", "승", "용", "차", " ", "전", "용", " ", "임", "시", " ", "갓", "길", "차", "로", "제", "를", " ", "운", "영", "하", "기", "로", " ", "했", "다", "."]	[12, 12, 12, 2, 3, 3, 3, 3, 3, 12, 2, 3, 12, 12, 12, 12, 2, 3, 3, 3, 3, 12, 12, 12, 2, 3, 3, 3, 3, 12, 12, 12, 8, 9, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12]

출처 : https://huggingface.co/datasets/klue/viewer/ner/train?row=11

KLUE - NLI

- 자연어 추론 (Natural Language Inference)
  - 두 문장 간의 논리적 관계 (함축, 중립, 모순)를 판단하는 작업
  - 주어진 문장 쌍에서 올바른 논리적 관계 라벨을 예측
  - 평가지표: Accuracy

source (string)	premise (string)	hypothesis (string)	label (class label)
"NSMC"	"힛걸 진심 최고다 그 어떤 히어로보다 멋지다"	"힛걸 진심 최고로 멋지다."	0 (entailment)
"NSMC"	"100분간 잘껄 그래도 소닉붐땜에 2점준다"	"100분간 잤다."	2 (contradiction)
"NSMC"	"100분간 잘껄 그래도 소닉붐땜에 2점준다"	"소닉붐이 정말 멋있었다."	1 (neutral)

출처 : https://huggingface.co/datasets/klue/viewer/nli/train?row=12

KLUE - RE

- 관계 추출 (Relation Extraction)
  - 주어진 문장에서 두 개체명 간의 관계를 판단하는 작업
  - 정보 추출, 지식 그래프 구축, 질의 응답 등 다양한 NLP 작업의 핵심 요소
  - 평가 지표: Accuracy

sentence (string)	subject_entity (dict)	object_entity (dict)	label (class label)
"〈Something〉는 조지 해리슨이 쓰고 비틀즈가 1969년 앨범 《Abbey Road》에 담은 노…	{ "word": "비틀즈", "start_idx": 24, "end_idx": 26, "type": "ORG" }	{ "word": "조지 해리슨", "start_idx": 13, "end_idx": 18, "type": "PER" }	0 (no_relation)
"호남이 기반인 바른미래당·대안신당·민주평화당이 우여곡절 끝에 합당해 민생당(가칭)으로…	{ "word": "민주평화당", "start_idx": 19, "end_idx": 23, "type": "ORG" }	{ "word": "대안신당", "start_idx": 14, "end_idx": 17, "type": "ORG" }	0 (no_relation)
"K리그2에서 성적 1위를 달리고 있는 광주FC는 지난 26일 한국프로축구연맹으로부터 관중…	{ "word": "광주FC", "start_idx": 21, "end_idx": 24, "type": "ORG" }	{ "word": "한국프로축구연맹", "start_idx": 34, "end_idx": 41, "type"…	5 (org:member_of)
"균일가 생활용품점 (주)아성다이소(대표 박정부)는 코로나19 바이러스로 어려움을 겪고 있는 대구광역시에 행복박스를 전달했다고 10일 밝혔다."	{ "word": "아성다이소", "start_idx": 13, "end_idx": 17, "type": "ORG" }	{ "word": "박정부", "start_idx": 22, "end_idx": 24, "type": "PER" }	10 (org:top_members/employees)

출처 : https://huggingface.co/datasets/klue/viewer/re/train?row=3

KLUE - STS

- 의미적 텍스트 유사도(Semantic Textual Similarity)
  - 두 문장이 얼마나 의미적으로 유사한지를 판단하는 작업
  - 문장 쌍과 그에 해당하는 유사도 점수 포함
  - 회귀 문제
    - 유사도 점수는 주로 0(전혀 다름)에서 5(완전히 동일) 사이의 값
    - 평가 지표: Pearson-Spearman Corr

sentence1 (string)	sentence2 (string)	labels (dict)
"숙소 위치는 찾기 쉽고 일반적인 한국의 반지하 숙소입니다."	"숙박시설의 위치는 쉽게 찾을 수 있고 한국의 대표적인 반지하 숙박시설입니다."	{ "label": 3.7, "real-label": 3.714285714285714, "binary-label": 1 }
"위반행위 조사 등을 거부·방해·기피한 자는 500만원 이하 과태료 부과 대상이다."	"시민들 스스로 자발적인 예방 노력을 한 것은 아산 뿐만이 아니었다."	{ "label": 0, "real-label": 0, "binary-label": 0 }
"회사가 보낸 메일은 이 지메일이 아니라 다른 지메일 계정으로 전달해줘."	"사람들이 주로 네이버 메일을 쓰는 이유를 알려줘"	{ "label": 0.3, "real-label": 0.3333333333333333, "binary-label": 0 }

출처 : https://huggingface.co/datasets/klue/viewer/sts/train?row=0

KLUE – TC(YNAT)

- 주제 분류(Topic Classification)
  - YNAT (Yes or No Answer Topic)는 KLUE-TC에서 사용되는 데이터셋의 이름
  - 텍스트의 주제나 카테고리를 정확하게 분류하는 것은 뉴스 분류, 문서 관리, 추천 시스템 등 다양한 응용에서 중요
  - 주어진 문장 또는 문서에서 올바른 주제 라벨을 예측
  - 평가 지표: Accuracy

title (string)	label (class label)
"유튜브 내달 2일까지 크리에이터 지원 공간 운영"	3 (생활문화)
"어버이날 맑다가 흐려져…남부지방 옅은 황사"	3 (생활문화)
"내년부터 국가RD 평가 때 논문건수는 반영 않는다"	2 (사회)
"김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것"	2 (사회)

출처 : https://huggingface.co/datasets/klue/viewer/ynat/train

03

한국어 전처리

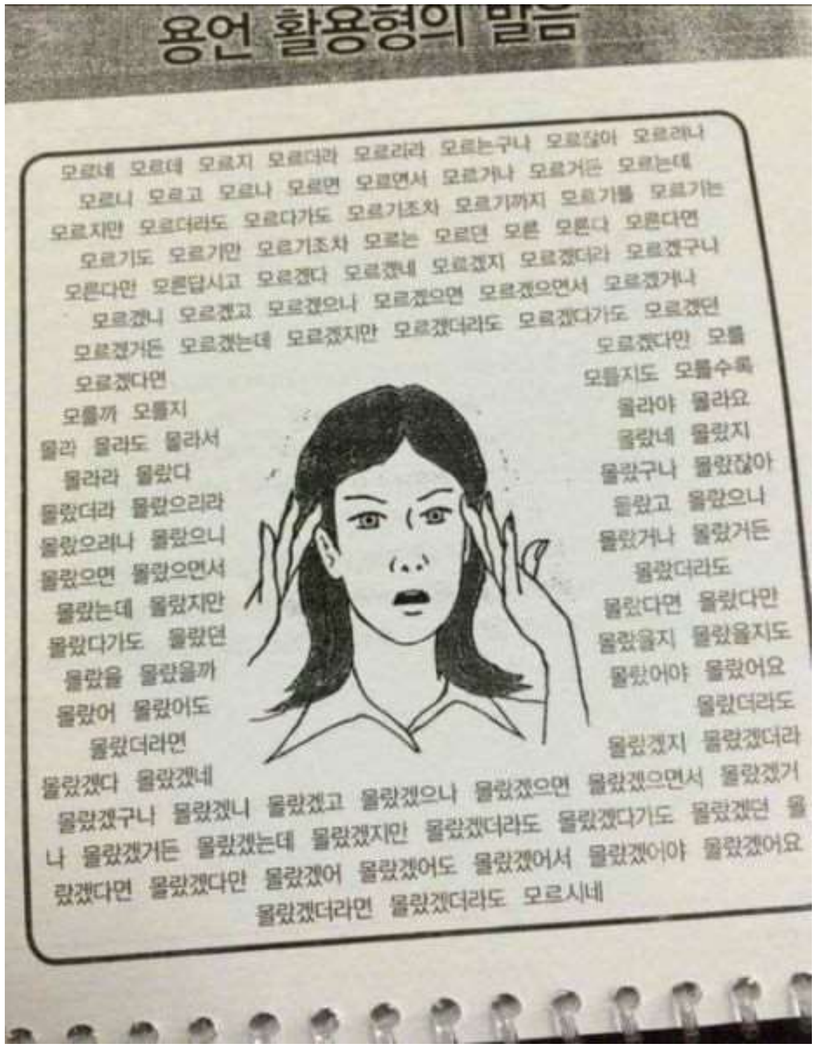
©Elice. All Rights Reserved

한국어 전처리의 어려움

- 한국어의 특성
  - 교착어: 한 단어 내에 여러 의미와 문법적 기능이 포함
    - 조사: 문장 내 단어의 역할을 지시
      - 주격 조사(이/가/께서), 목적격 조사(을/를/께) 등
    - 어미: 용언의 어간과 결합하여 다양한 의미 표현
      - 먹+(다/고/어/지만/으나/으며/겠)
    - 접사: 어근에 붙어 의미를 확장시킴
      - 날+(벼락/고기/것/파리)

한국어 전처리의 어려움

- 자연어처리 시 교착어의 문제점
  - 단어의 경계가 모호
    - 가시었겠더라만은 → 가 + 시 + 었 + 겠 + 더라 + 만 + 은
  - 동음이의어가 많음
    - 이: 벌레, 숫자, 주격 조사, 사람의 호칭, 지시대명사, 치아
  - 유래가 복잡하고 어근이 짧은 신조어가 많음
    - 뇌피셜, 대인배, 딸바보, 월급루팡, 지름신



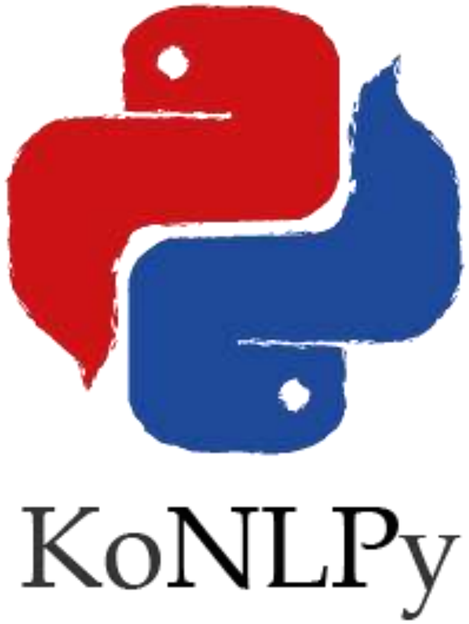
출처 : https://brunch.co.kr/@ifeelsunny/32

한국어 전처리의 어려움

- 공백 기반 단어 토큰화는 사용이 불가능
  - 단어보다 낮은 수준의 분석 필요
    - Subword tokenizer
    - 형태소 분석기
  - 불용어 제거가 단순하게 되지 않음
    - 동음이의어가 전부 제거될 수 있음

한국어 전처리의 어려움

- 대표적인 한국어 형태소 분석기
  - KoNLPy에서 오픈소스 데이터 제공
  - Hannanum: KAIST에서 만든 형태소 분석 및 품사 태깅 알고리즘
  - Kkma: 서울대학교에서 만든 형태소 분석기(Java)
  - Komoran: Shineware에서 만든 형태소 분석기(Java)
  - Mecab: 일본어 형태소 분석 및 품사 태깅 알고리즘을 한국어 판으로 업데이트
  - Twitter: 한국어 트위터 데이터를 바탕으로 만든 토크나이저



출처 : https://konlpy-ko.readthedocs.io/ko/v0.4.3/

한국어 전처리의 어려움

- 한국어 정규화 도구
  - PyKoSpacing: 띄어쓰기 오류를 교정
  - Py-Hanspell: 맞춤법과 띄어쓰기를 동시에 교정
  - soynlp: 신조어 및 속어를 포함한 텍스트의 정규화

한국어 전처리의 어려움

- 최근에는 Subword tokenizer을 사용함
  - 언어에 관계없이 단어를 분절하여 분석
  - 교착어, 굴절어, 고립어 관계 없이 모두 좋은 성능을 보임
- 대표적인 알고리즘
  - BPE(Byte-Pair Encoding)
  - WordPiece
  - SentencePiece

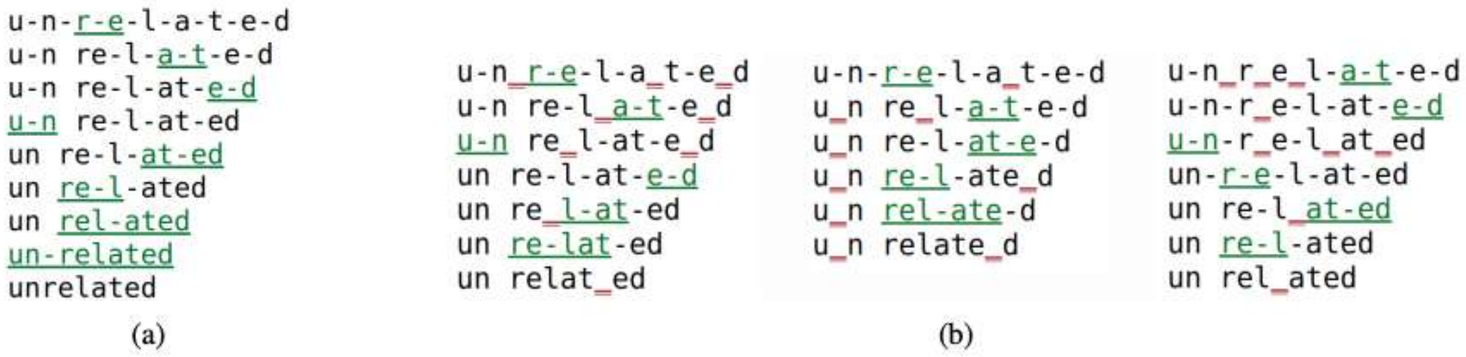


Figure 1: Segmentation process of the word '*unrelated*' using (a) BPE, (b) *BPE-dropout*. Hyphens indicate possible merges (merges which are present in the merge table); merges performed at each iteration are shown in green, dropped – in red.

출처 : https://blog.octanove.org/guide-to-subword-tokenization/

04

한국어 LLM

©Elice. All Rights Reserved

KoBERT

- SKTBrain에서 제공
- BERT (Bidirectional Encoder Representations from Transformers)의 구조를 기반으로 한국어 데이터로 사전 학습된 언어 모델
- 위키피디아, 뉴스 등의 코퍼스를 바탕으로 학습
- 데이터 기반 토큰화 기법 사용
  - 기존 대비 27%의 토큰만 사용
  - 2.6% 이상의 성능 향상

KoELECTRA

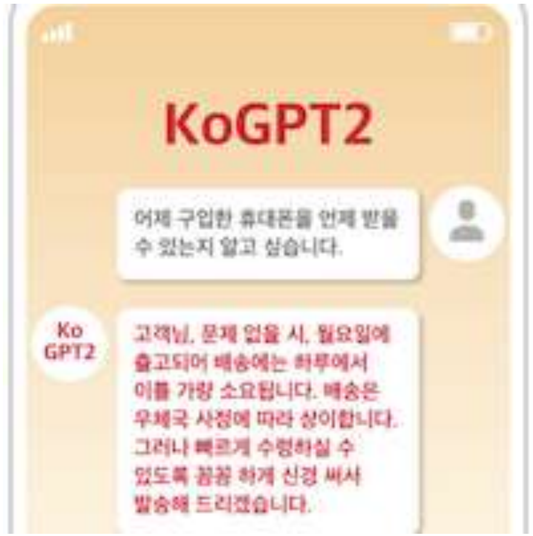
- Hugging Face의 transformers 라이브러리에서 손쉽게 사용하도록 배포
- 뉴스, 위키피디아, 나무위키 등의 문어체 데이터와 구어체 데이터(메신저 등)을 두구 학습
- KLUE 데이터셋에서 우수한 성능을 보임

	NSMC (acc)	Naver NER (F1)	PAWS (acc)	KorNLI (acc)	KorSTS (spearman)	Question Pair (acc)	KorQuaD (Dev) (EM/F1)	Korean-Hate-Speech (Dev) (F1)
KoBERT	89.59	87.92	81.25	79.62	81.59	94.85	51.75 / 79.15	66.21
XLM-Roberta-Base	89.03	86.65	82.80	80.23	78.45	93.80	64.70 / 88.94	64.06
HanBERT	90.06	87.70	82.95	80.32	82.73	94.72	78.74 / 92.02	**68.32**
KoELECTRA-Base	90.33	87.18	81.70	80.64	82.00	93.54	60.86 / 89.28	66.09
KoELECTRA-Base-v2	89.56	87.16	80.70	80.72	82.30	94.85	84.01 / 92.40	67.45
KoELECTRA-Base-v3	**90.63**	**88.11**	**84.45**	**82.24**	**85.53**	**95.25**	**84.83 / 93.45**	67.61

출처 : https://github.com/monologg/KoELECTRA

KoGPT

- SKT Brain과 카카오에서 각자 개발
  - SKT의 KoGPT2는 사내 업무 효율 증진이 목표
    - 챗봇, 문서 해독 등
  - 카카오의 KoGPT는 ChatGPT와 비슷한 생성형 언어 모델
    - 상용화를 통한 서비스 런칭이 목표
    - 성능은 ChatGPT에 비해 미미



출처 : https://sktelecom.github.io/project/kogpt2/, https://developers.kakao.com/product/kogpt

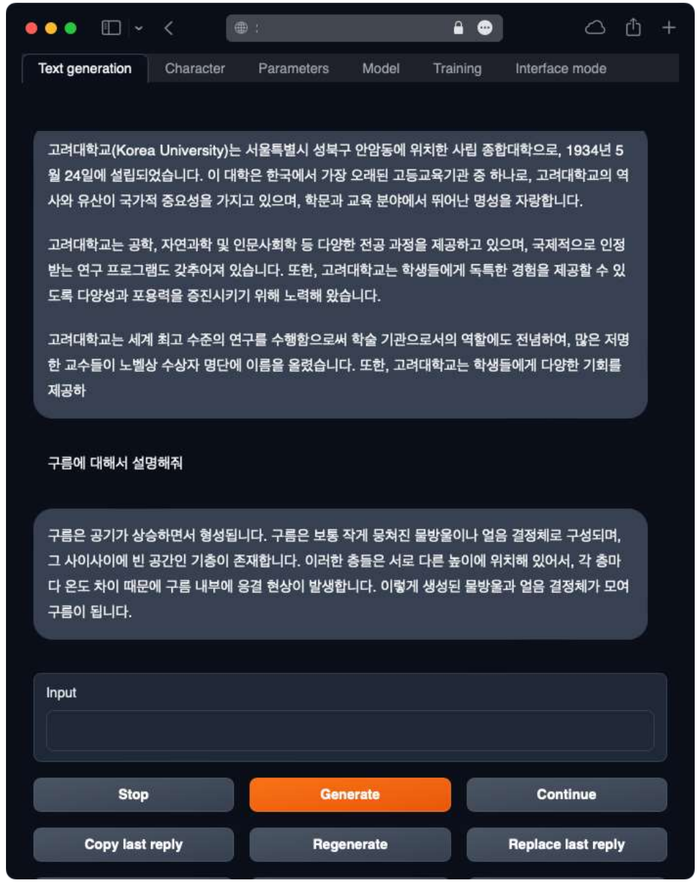
KoAlpaca

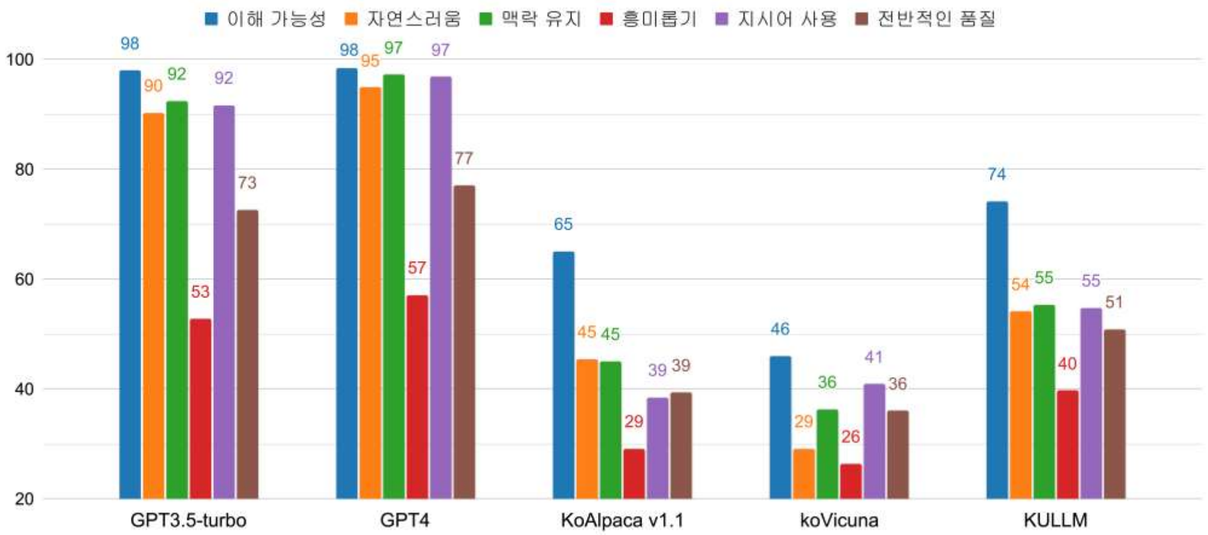
- Alpaca: Stanford에서 만든 LLaMA 모델의 개선 버전
- 같은 방식으로 LLaMA를 한국어로 미세조정한 모델
- Backbone
  - 한국어 모델: 다국어 모델인 Polyglot-ko를 사용
  - 영어+한국어 모델: LLAMA
- ChatGPT와 유사한 일반 언어 생성 모델

04 한국어 LLM

KKULM

- Korea University Large Language Model
- 다국어 언어모델인 Polyglot-ko를 Backbone으로 사용
- Hugging Face에서 손쉽게 사용 가능
- 한국어 대회에서 우수한 성능을 입증





출처 : https://github.com/nlpai-lab/KULLM