

Goorm 3rd Project

구어체-문어체 변환기

Team 1

양다경 구선혜 김유나 김현수 김현지 이용주



Contents

1. 프로젝트 개요: 구어체-문어체 변환기

- 주요 서비스 소개 및 기존 서비스와의 비교

2. 프로세스

3. 병렬 코퍼스 구축

- Google Translate API 이용
- 자체 훈련한 번역기 모델 이용

4. 문어체-구어체 분류기

5. 구어체-문어체 변환기

- KoBART
- KoT5

6. Demo

7. 자체 평가 및 추후 과제

1. 프로젝트 개요: 구어체-문어체 변환기

- 주요 서비스 소개 및 기존 서비스와의 비교



프로젝트 개요: 구어체-문어체 변환기

1. 입력된 문장이 구어체인지, 문어체인지 분류

예)

- 정기 시험이 엄청 어려웠어요. → 구어체
- 정기 시험은 매우 어려웠습니다. → 문어체

2. 구어체-문어체 변환

예)

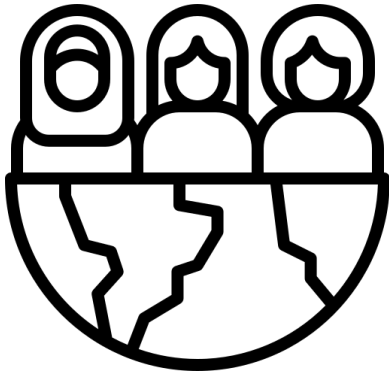
- 정기 시험이 엄청 어려웠어요. → 정기 시험은 매우 어려웠습니다.
- 정기 시험은 매우 어려웠습니다. → 정기 시험이 엄청 어려웠어요.

사용 스택

Google Colab Pro+	datasets-2.8.0	levenshtein- 0.20.9	matplotlib-3.2.2	numpy-1.21.6
pandas-1.3.5	transformers- 4.25.1	torch-1.13.0	torchinfo-1.7.1	wandb-0.13.7

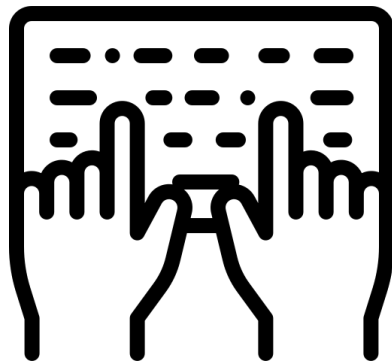


프로젝트 개요: 구어체-문어체 변환기



Target User

한국어가 모국어가 아닌
집단 및 개인



유저들의 작문을 도와
한국어 작문의 문턱 낮추기



상황에 맞는 적절한 문체 사용에
도움을 줌



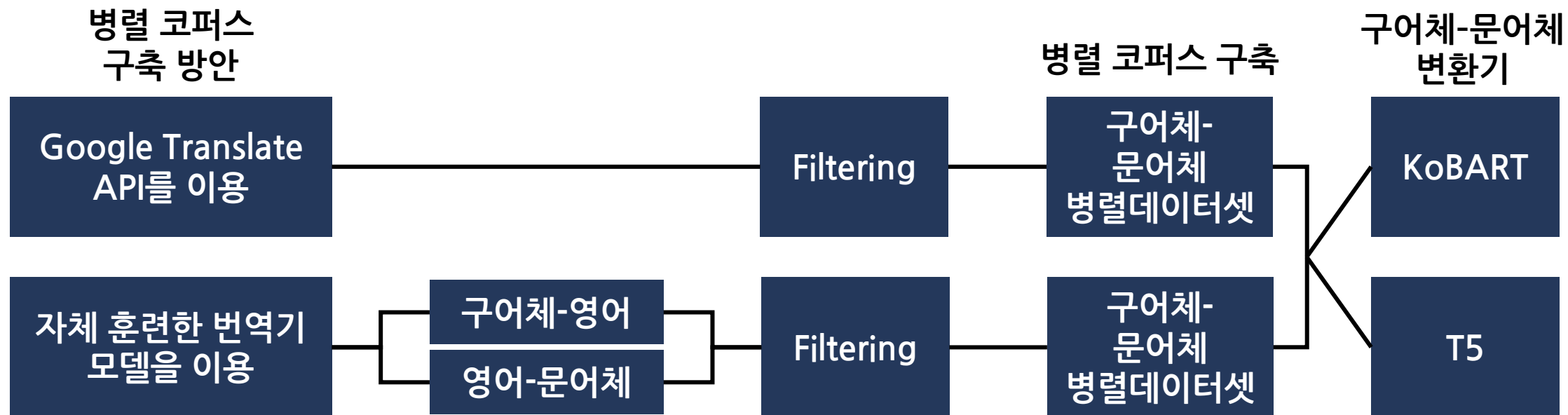
기존 서비스

서비스 지역	서비스	주요기능/특징
국내	카카오 I 번역 서비스 문체 변환	기본문체, 높임말, 예사말의 문체 설정이 가능하도록함.
	엑소브레인 구어체 언어분석 API	기존 문어체 언어 분석 기술을 고도화한 API.
	네이버 파파고 높임말 번역	번역 후, 높임말 on-off 기능으로 문체 설정이 가능하도록함.
국외	Grammarly	<ul style="list-style-type: none">문법과 철자에서 스타일과 어조에 이르기까지 포괄적인 writing assistant 서비스 제공.옵션을 설정하여 개인 맞춤형 스타일 변환이 가능
	Wordtune	casual, formal로 text에 대한 style 변환이 가능한 웹기반 writing assistant
	Sampling	캐주얼한 입력 텍스트가 주어지면 공식적인/전문적인 텍스트로 출력

2. 프로세스



프로세스



3. 병렬 코퍼스 구축

- Google Translate API 이용
- 자체 훈련한 번역기 모델 이용



병렬 코퍼스 구축

사용한 데이터셋

AI-HUB 한국어-영어 번역(병렬) 말뭉치

	분야	설명	수량
문어체	뉴스	뉴스 텍스트	80만
	정부 웹사이트/저널	정부/지자체 홈페이지, 간행물	10만
	법률	행정 규칙, 자치 법규	10만
	한국문화	한국 역사, 문화 콘텐츠	10만
구어체	구어체	자연스러운 구어체 문장	40만
	대화체	상황/시나리오 기반 대화 세트	10만

합계 : 160만 문장

Problem

프로젝트에 이용할 수 있는 구어체-문어체로 이루어진 병렬 코퍼스가 존재하지 않음.

병렬코퍼스 구축 방안

1. Google Translate API Round-Trip Translation

- 텍스트를 다른 언어로 번역하고, 그 결과를 원래의 언어로 다시 번역.

2. 자체 번역 모델

- 구어체-영어 / 영어-문어체 번역기를 이용하여 데이터셋 구축.

#train dataset	#validation dataset	#output dataset
350,000	50,000	100,000



3-1. Google Translate API

Round trip translation

Google Translate API에 Round trip translation 기법을 이용하여 구어체를 문어체로 변환하여 병렬코퍼스 생성.

	번역전	영어번역	한글번역
0	사과는 잘 씻은 뒤 껍질 채 먹는 게 좋대네요.	It is better to wash the apples and eat it.	사과를 씻고 먹는 것이 낫습니다.
1	내가 언제까지 거기 가면 됩니까?	How long can I go there?	거기에 얼마나 오래 갈 수 있습니까?
2	우리가 작성해서 전달해 드려야 하나요?	Should we write and deliver?	우리는 쓰고 전달해야합니까?
3	난 지금 행복해요, 세상이 나를 원하고 있죠.	I'm happy now, the world wants me.	나는 지금 행복합니다. 세상은 나를 원합니다.
4	그는 틀림없이 꽃 가게에 들렸을 겁니다.	He must have been in a flower shop.	그는 꽃 가게에 있었을 것입니다.
...
99995	그럼 컨셉을 약간 바꾸는 것도 좋은 방법이겠네요.	Then it's a good idea to change the concept a ...	그런 다음 개념을 조금 바꾸는 것이 좋습니다.
99996	객실 안에 비밀번호가 작성되어 있습니다.	A password is written in the room.	암호는 방에 작성됩니다.
99997	4번 게이트로 갔을 때 바로 찾을 수 있나요?	Can I find it right when I went to Gate 4?	게이트 4에 갔을 때 바로 찾을 수 있습니까?
99998	지금 3개나 빠진 상태인데 한꺼번에 이렇게 많이 빠져도 상관없는 건가요?	I'm missing 3 of them right now. Does it matte...	지금 3개를 놓치고 있습니다. 이만큼 한꺼번에 잃어도 상관없나요?
99999	저기 앞에 창가 쪽 좌석인데 제 친구랑 같이 앉으려고요.	It's a seat side in front of you, and I'm goin...	그것은 당신 앞에 좌석이 있고, 나는 내 친구와 함께 앉을 것입니다.

100000 rows × 3 columns

Google translate API를 이용한 구어체-문어체 병렬 코퍼스 구축



3-2. 자체 훈련한 번역기 모델

model	params	KorNLI acc	KorSTS spearman	NSMC acc	PD acc	NER acc	KorQuAD 1.0 f1	class
Multilingual BERT	172M	76.8	77.8	87.5	91.1	84.20(2022)	86.5	encoder
XLM-Roberta-Base	270M	80.0	79.4	90.1	92.6	83.9	92.3	encoder
KoBERT	92M	79.6	81.6	90.1	91.1	87.9	90.3	encoder
ETRI BERT(KoBERT)	110M	79.5	80.5	88.8	93.9	82.5	94.1	encoder
KoreALBERT Base	12M	79.7	81.2	89.6	93.8	82.7	92.6	encoder
KoreALBERT Large	18M	81.1	82.1	89.7	94.1	83.7	94.5	encoder
HanBERT-54kN	128M	80.3	82.7	90.2		87.7		encoder
KoGPT2 2.0	125M		77.8	89.1				decoder
KoBART-base	124M		81.66	90.24				en-de
kolang-t5-base	225M	77.1		88.8				en-de
KE-T5 SMALL	60M	73.41	77.9	97.9			87.9	en-de
KE-T5 BASE	247M	78.67	79.73	88.95			88.95	en-de
KE-T5 LARGE	770M	79.76	83.25	89.74			89.7	en-de
DistilKoBERT	28M		86.53	88.41		84.13	77.8	encoder
KoELECTRA-BASE-V3	124M	82.24	85.53	90.63		88.11	93.45	encoder
koelectra-small-v3		78.6	80.79	89.36		85.4	91.13	encoder
Kobigbird-bert-base	114M						94.7	encoder
KCBERT-BASE		74.85	75.57	89.62		84.34	84.39	encoder
KcBERT-Large		76.99	77.49	90.68		85.53	86.64	encoder

구어체-영문 번역 모델 encoder 선정 근거

대부분의 평가 지표에서
koelectra-base-v3의 성능이 높게 측정
Encoder-KoElectra 선정

Model	Train FLOPs	Params	CoLA	SST	MRPC	STS	QQP	MNLI	QNLI	RTE	Avg.
BERT	1.9e20 (0.27x)	335M	60.6	93.2	88.0	90.0	91.3	86.6	92.3	70.4	84.0
RoBERTa-100K	6.4e20 (0.90x)	356M	66.1	95.6	91.4	92.2	92.0	89.3	94.0	82.7	87.9
RoBERTa-500K	3.2e21 (4.5x)	356M	68.0	96.4	90.9	92.1	92.2	90.2	94.7	86.6	88.9
XLNet	3.9e21 (5.4x)	360M	69.0	97.0	90.8	92.2	92.3	90.8	94.9	85.9	89.1
BERT (ours)	7.1e20 (1x)	335M	67.0	95.9	89.1	91.2	91.5	89.6	93.5	79.5	87.2
ELECTRA-400K	7.1e20 (1x)	335M	69.3	96.0	90.6	92.1	92.4	90.5	94.5	86.8	89.0
ELECTRA-1.75M	3.1e21 (4.4x)	335M	69.1	96.9	90.8	92.6	92.4	90.9	95.0	88.0	89.5

Table 2: Comparison of large models on the GLUE dev set. ELECTRA and RoBERTa are shown for different numbers of pre-training steps, indicated by the numbers after the dashes. ELECTRA performs comparably to XLNet and RoBERTa when using less than 1/4 of their pre-training compute and outperforms them when given a similar amount of pre-training compute. BERT dev results are from Clark et al. (2019).

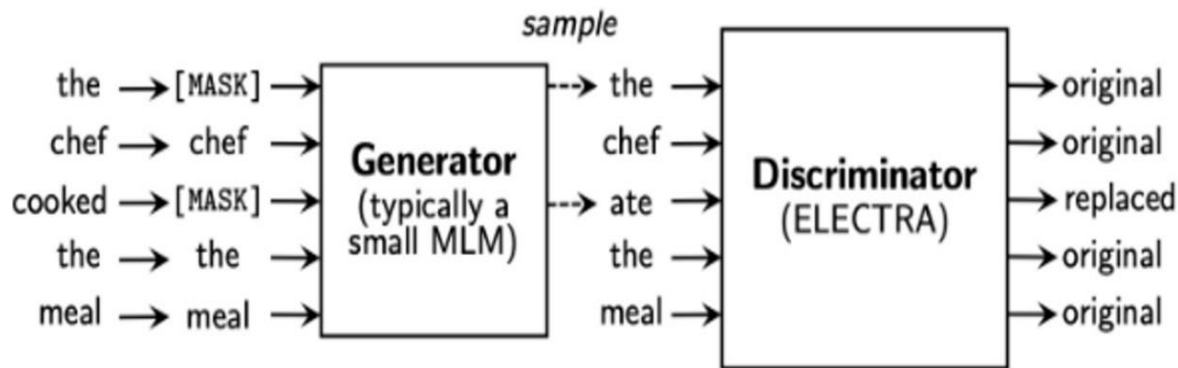
영문-문어체 번역 모델 encoder 선정 근거

ELECTRA: Pre-training Text Encoders as
Discriminators Rather Than Generators에서
제시한 모델 비교 성능
Encoder-ELECTRA 선정



3-2. 자체 훈련한 번역기 모델 ELECTRA & KoELECTRA

Generator에서 나온 token을 보고 discriminator에서 real token인지 fake token(generator)가 만든 토큰인지 판별.



KoELECTRA

학습 데이터 34GB 한국어 text

- 14GB - 뉴스, 위키, 나무위키
- 20GB - 모두의 말뭉치(신문, 문어, 구어, 메신저, 웹)

Replaced Token Detection

- MLM에서 일부 토큰을 MASK 처리하듯이 몇 토큰을 골라 작은 generator가 샘플링한 토큰으로 대체.
- 이후 discriminator가 토큰이 generator로부터 만들어졌는지 맞추도록함.

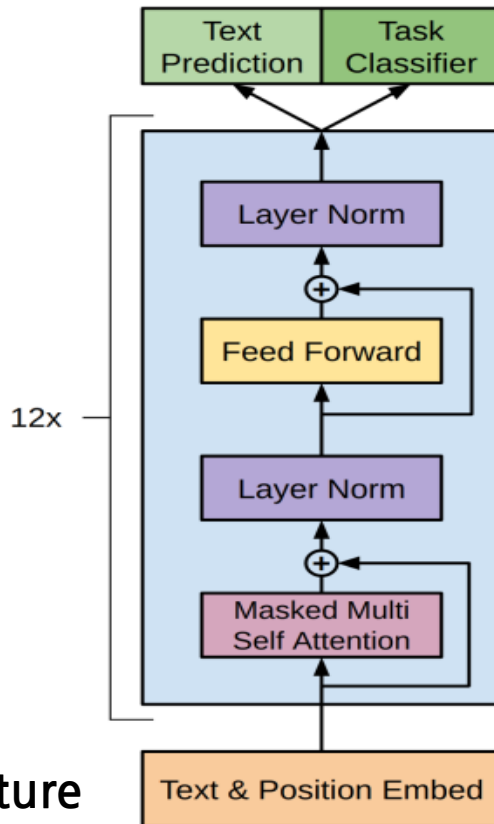
장점

- 모든 input token에 대해 학습할 수 있음.
- BERT와 비교했을 때 더 좋은 성능을 보임.

Hidden size	Parameters
768	110M

3-2. 자체 훈련한 번역기 모델 GPT2/KoGPT2

Hidden size	Parameters
768	125M(KoGPT2) 117M(GPT2)



- Transformer의 **decoder** 블록 구성.
- **1024개**의 token 처리 가능.
- fine-tuning 없이 적용.
- zero-shot learning.
- 대규모 데이터셋, large model, supervised learning.

Auto-Regression

각 token이 생성된 후에 입력 시퀀스에 더해지는 방식으로 동작.
새 시퀀스는 다음 단계에서 모델의 입력으로 들어감.

학습 데이터

KoGPT2 Character-level BPE tokenizer 사용

한국어 위키 백과, 뉴스, 모두의 말뭉치 v1.0, 청와대 국민청원

GPT2 Byte-level BPE tokenizer 사용

web text(reddit) 활용

+ 위키피디아 글 중복 제거 후 40GB 데이터



자체 훈련한 번역기 모델

구어체-영문 번역

Encoder monologg/koelectra-base-v3-finetuned-korquad
Decoder gpt2

	prediction	Korean
0	They say that it is good to wash apples well a...	사과는 잘 씻은 뒤 껍질 채 먹는 게 좋대요.
1	Until when should I go there?	내가 언제까지 거기 가면 됩니까?
2	Should we fill it out and deliver it?	우리가 작성해서 전달해 드려야 하나요?
3	I'm happy now, the world wants me.	난 지금 행복해요, 세상이 나를 원하고 있죠.
4	He must have been at the flower shop.	그는 틀림없이 꽃 가게에 들렸을 겁니다.
...
99995	Then it would be a good idea to change the con...	그럼 컨셉을 약간 바꾸는 것도 좋은 방법이겠네요.
99996	The password is written in the room.	객실 안에 비밀번호가 작성되어 있습니다.
99997	Can I find it right away when I go to gate 4?	4번 게이트로 갔을 때 바로 찾을 수 있나요?
99998	3 of them are missing, so is it okay to get th...	지금 3개나 빠진 상태인데 한꺼번에 이렇게 많이 빠져도 상관없는 건가요?
99999	It's a window seat over there, and I want to s...	저기 앞에 창가 쪽 좌석인데 제 친구랑 같이 앉으려고요.

100000 rows × 2 columns

자체 훈련 모델을 이용하여 구어체-영문 코퍼스 구축

영문-문어체 번역

Encoder google/electra-small-discriminator
Decoder skt/kogpt2-base-v2

	Unnamed: 0	spoken_ko	written_ko
0	0	사과는 잘 씻은 뒤 껍질 채 먹는 게 좋대요.	사과 잘 씻어 골고루 먹는 게 좋다고 한다.
1	1	내가 언제까지 거기 가면 됩니까?	언제까지 갈까.
2	2	우리가 작성해서 전달해 드려야 하나요?	우리가 이것을 풀어서 전달해야 한다.
3	3	난 지금 행복해요, 세상이 나를 원하고 있죠.	지금 행복하고, 세상이 나를 원한다.
4	4	그는 틀림없이 꽃 가게에 들렸을 겁니다.	꽃가게에 가봤어야 할 것 같다.
...
99995	99995	그럼 컨셉을 약간 바꾸는 것도 좋은 방법이겠네요.	그러면 개념을 바꿔보는 게 좋지 않을까 하는 생각이 들었다.
99996	99996	객실 안에 비밀번호가 작성되어 있습니다.	비밀번호는 방에 적혀있다.
99997	99997	4번 게이트로 갔을 때 바로 찾을 수 있나요?	제가 4호 출입을 할 때 바로 찾을 수 있는 것이냐?
99998	99998	지금 3개나 빠진 상태인데 한꺼번에 이렇게 많이 빠져도 상관없는 건가요?	이 중 3명이 빠져 있어 한꺼번에 다 빠져 나가는 게 말이 되냐는 것이다.
99999	99999	저기 앞에 창가 쪽 좌석인데 제 친구랑 같이 앉으려고요.	그곳에 있는 창가 좌석인데 친구와 앉아보고 싶다.

100000 rows × 3 columns

자체 훈련 모델을 이용하여 영문 - 문어체 코퍼스 구축



3-3. Filtering

필터링

1. 문장 길이를 이용한 필터링

- 문장 길이 **100** 이하로 제한.
- 입력 문장 길이의 90%보다 번역된 스타일 변환 문장길이가 짧은 경우 제거.

2. BLEU Score를 이용한 필터링

- **BLUE = 100**인 경우, **BLUE = 0**인 경우 제거.

3. KLUE-NLI를 이용한 필터링

- KLUE-NLI를 학습한 **RoBERTa Classifier**를 이용해 원문-스타일 변환 문장 간 의미가 같지 않은 쌍 제거.
- 문장 쌍의 관계가 의미가 같다고 분류된 경우에도 점수가 **0.8** 이하인 경우 제거

4. 코사인 유사도를 이용한 필터링

- 사전학습된 **Sentence Transformers**를 활용.
- 코사인 유사도를 이용하여 두 문장 간의 유사도 점수가 **0.9**보다 낮은 문장 쌍 제거.



Google Translate API + 자체 훈련한 번역기 모델 병렬 코퍼스 구축 + 필터링

1. Google Translate API

최종 코퍼스 - 10,111개

Unnamed: 0		spoken	written
0	0	난 지금 행복해요, 세상이 나를 원하고 있죠.	나는 지금 행복합니다. 세상은 나를 원합니다.
1	1	안녕하세요 올해가 돼서 매우 기쁘고요. 행복한 올해가 되세요! 안녕하세요, 나는 올해가되어 매우 기쁩니다. 행복한 한 해를 보냈습니다!	
2	2	여행은 예상 못 한 일이 벌어지기 때문에 환상적이에요.	여행은 예상치 못한 일이 일어나기 때문에 여행은 환상적입니다.
3	5	식당은 적절한 가격의 최상 서비스를 추구해요.	식당은 적절한 가격으로 최고의 서비스를 추구합니다.
4	6	저는 금연 실을 예약했는데 방에서 담배 냄새가 엄청나요.	나는 금연 공간을 예약했지만 방에 담배 냄새는 엄청납니다.
...
10106	16834	지금 우리가 어디쯤 걷고 있는지 확인할 수 있어?	우리가 지금 어디에서 걷고 있는지 볼 수 있습니까?
10107	16835	다른 학교에 가서 수업을 듣는 것입니다. 강의시간표에서 확인할 수 있어요.	다른 학교에 가서 수업을 듣습니다. 강의 일정에서 확인할 수 있습니다.
10108	16836	할인을 받을 방법이 있을까요?	할인받을 수 있는 방법이 있습니까?
10109	16837	실례하지만 수영시설이 어디 있나요?	실례하지만 수영 시설은 어디에 있습니까?
10110	16838	저희는 여기 맥주가 그렇게 맛있다고 들었는데요.	우리는 여기서 맥주가 너무 맛있다고 들었습니다.

10111 rows x 3 columns

Google Translate API를 이용하여 구어체-문어체 병렬코퍼스 구축

2. 자체 훈련한 번역기 모델

최종 코퍼스 - 3,655개

Unnamed: 0		spoken	written
0	0	바쁘신데 긴 글 읽어주셔서 정말 감사해요.	바쁜데도 긴 글을 읽으셔서 감사드립니다.
1	1	그는 배우가 아닌 가수로 살아왔어.	그는 배우가 아닌 가수로 살아왔다.
2	2	디자이너들은 재활용제품을 친환경 제품으로 바꾸기 위해 사용해요.	디자이너들이 재활용품을 활용해 친환경 제품을 바꾸기도 한다.
3	3	갑자기 왜 이런 문제가 발생하는 건가요?	왜 갑자기 이런 문제가 생기는 건가 말이다.
4	4	이 의견에 동의하지 않는 분 계신가요?	이 의견에 동의하지 않는 사람이 있을까.
...
3650	3650	그럼 마네킹이 입고있는 코트로 한번 입어볼 수 있을까요?	그렇다면 마네킹이 입었던 코트를 입어볼 수 있을까.
3651	3651	그런 경우는 거의 없지만 문제가 생기면 직원에게 물어보세요.	그럴 일이 거의 없는데 문제가 있으면 직원에게 물어본다.
3652	3652	여보세요, 지금 피자 주문하면 배달 얼마나 걸리나요?	지금 피자를 주문하면 배달까지 얼마나 시간이 걸릴까.
3653	3653	할인을 받을 방법이 있을까요?	할인받을 방법이 있는 것일까.
3654	3654	수술 시간이 오래 걸리나 봐요?	수술을 받기까지가 오래 걸린다.

3655 rows x 3 columns

자체 훈련 모델을 이용하여 구어체-문어체 병렬 코퍼스 구축

최종 Google Translate API + 자체 훈련한 번역기 모델 병렬 코퍼스

Number of Train Samples: 11,012 | Number of Validation Samples: 2,754

4. 문어체-구어체 분류기



문어체-구어체 분류기

KLUE-BERT-base 모델 선정

Klue 벤치마크 토픽 분류에서 최고 점수인 XLM-R-large는 코랩 자원 문제로 인해
정확도 2위 KLUE-BERT-base를 fine-tuning할 분류 모델로 선정.

#Train dataset	#Validation dataset	#Test dataset	Batch size
10,500	3,000	1,500	16

Epochs	Early_stopping_patience	Accuracy
20	5	0.988

5. 구어체-문어체 변환기

- KoBART
- KoT5

5-1. 구어체-문어체 변환기 BART/KoBART

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	44.16	21.28	40.90	45.14	22.27	37.25

BART 모델 선정 근거

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension의 Generation tasks 성능 비교에서 **BART**가 가장 좋은 성능을 보임.

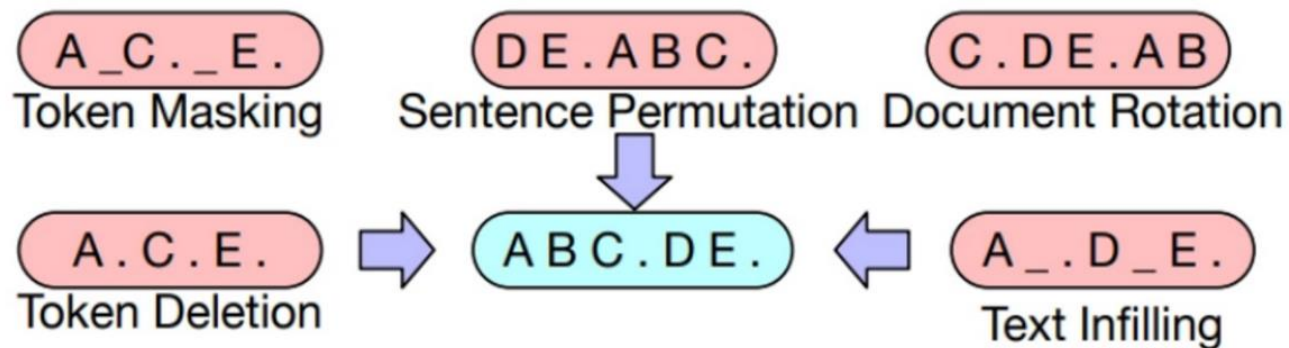
5-1. 구어체-문어체 변환기 BART/KoBART

Bidirectional and Auto-Regressive Transformers

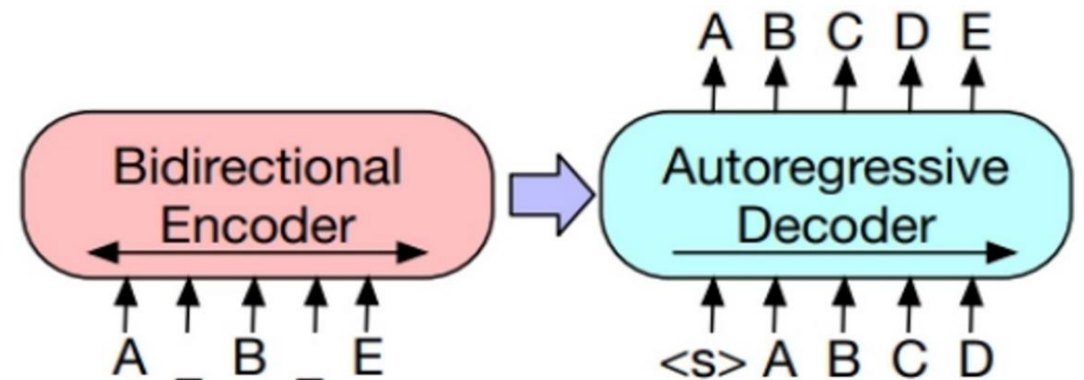
- seq2seq 구조로 만들어진 denoising auto encoder.
- 임의의 noising function으로 **손상된 텍스트를 복구**하도록 학습.
- 디코더의 출력과 원본 text의 loss를 줄이도록함.

학습 데이터

한국어 위키 5M
+ 다른 코퍼스(뉴스, 책, 모두의 말뭉치 V.10) 0.27B



문장 생성과 문맥 이해 task에 효과적



Hidden size	Parameters
768	124M



5-2. 구어체-문어체 변환기 T5

Text to Text Transfer Transformer

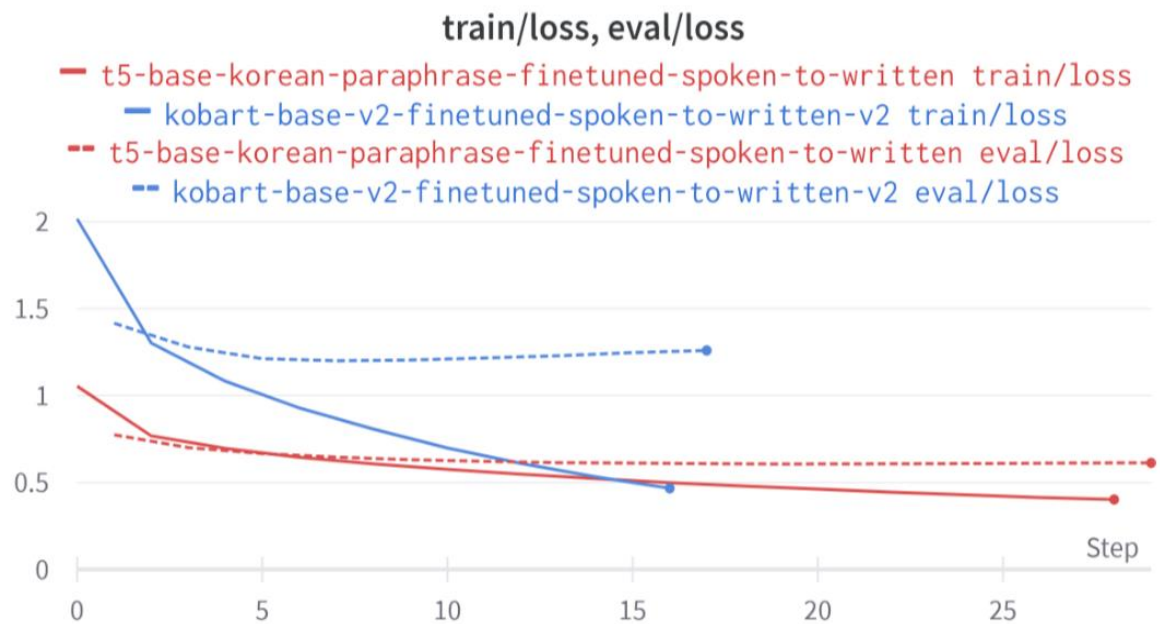
- 입력과 출력이 항상 텍스트 문자열인 text to text 프레임워크를 사용하며 모든 **NLP TASK 일반화**.
- 분류, 순차태깅, 기계번역, 문서 요약, QA 등 TASK를 **동일한 모델과 파라미터를 사용하여 범용적으로** 학습할 수 있음.
- 파라미터의 크기가 커졌지만 750GB가 되는 데이터를 소화.

Hidden size	Parameters
768	220M

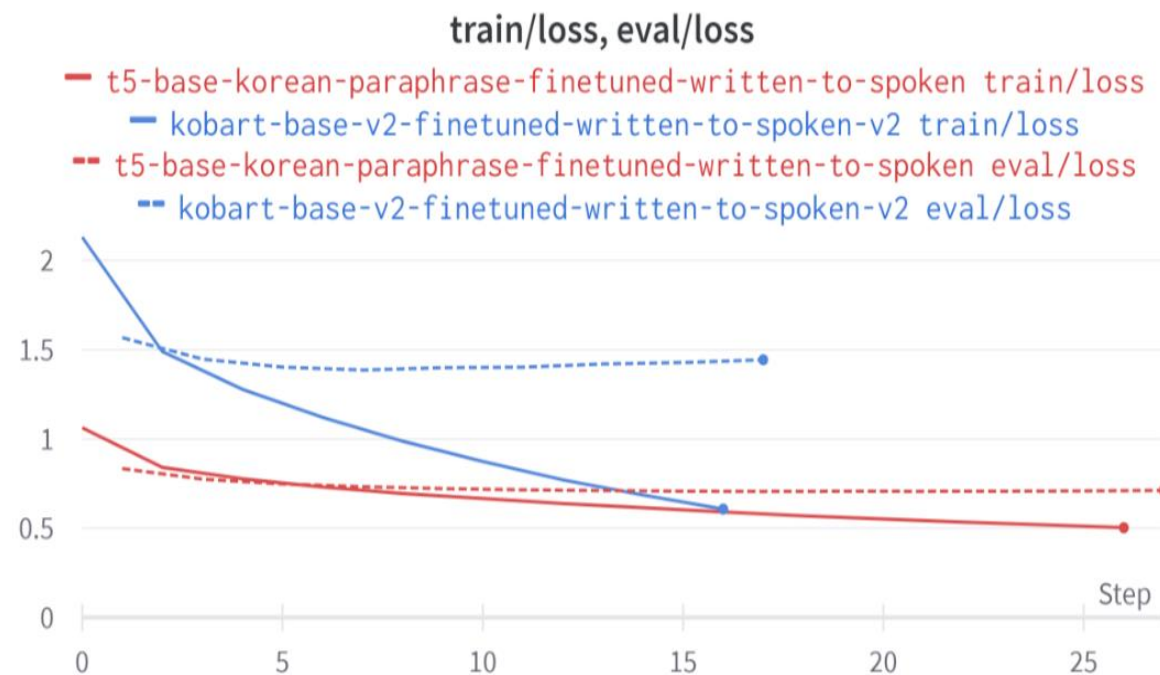
Replacing span

- 문장 내의 구간(span)을 하나의 센티넬 토큰으로 치환.
- 모든 단어에 대해서 별도의 [MASK] 토큰으로 가리는 Masking과 다르게, 구간 전체를 한 토큰으로 치환.
- 해당 구간에 대해서 고유한 토큰 [X], [Y] 등으로 치환.

5-3. 구어체-문어체 변환기 훈련 및 검증



T5와 KoBART train/validation loss
(구어체-문어체 변환기)



T5와 KoBART train/validation loss
(문어체-구어체 변환기)

5-3. 구어체-문어체 변환기 훈련 및 검증

	models	BLEU Score	Cosine Similarity
구어체→문어체	KoBART	37/100	0.932
	T5	40/100	0.942
문어체→구어체	KoBART	28/100	0.930
	T5	30/100	0.940

T5가 KoBART 모델보다 더 좋은 성능을 보임.



5-4. 정성적 검증 - Google Translate와 비교

구어체	문어체	
원문	Google Translate	구어체-문어체 변환기
인공지능(AI) 음성 인식 플랫폼의 활약이 그 어느 때보다 두드러져요.	인공지능(AI) 음성인식 플랫폼의 성능이 그 어느 때보다 두드러진다.	AI AI 음성 인식 플랫폼의 활약이 특히 두드러지고 있다.
기업은 고도로 개인화된 고객경험(CX)을 제공하여 보다 효율적으로 고객과 소통하길 원해요.	기업은 고도로 개인화된 고객 경험(CX)을 제공하여 고객과 보다 효과적으로 소통하기를 원합니다.	기업은 고도로 개인화된 고객 경험(CX)을 제공하여 보다 효율적으로 고객과 소통하기를 원합니다.
안녕. 오늘 점심으로 삼겹살을 먹을 예정이야.	안녕하세요. 오늘 점심으로 삼겹살을 먹겠습니다.	안녕하세요. 오늘 점심으로 삼겹살을 먹을 예정입니다.
수료식 끝나고 뭐해?	졸업식 끝나고 뭐해?	수료식 후 바로 뭐 합니까?
지금 뭐 듣고 계세요?	무엇을 듣고 있어요?	지금 무엇을 듣고 있습니까?
우주에는 외계인이 살고 있어.	외계인은 우주에 산다.	우주에는 외계인이 살고 있습니다.


6. Demo


- Spoken-Written-Converter





Demo


Spoken-Written Converter

Set Goal 

☒ Spoken 

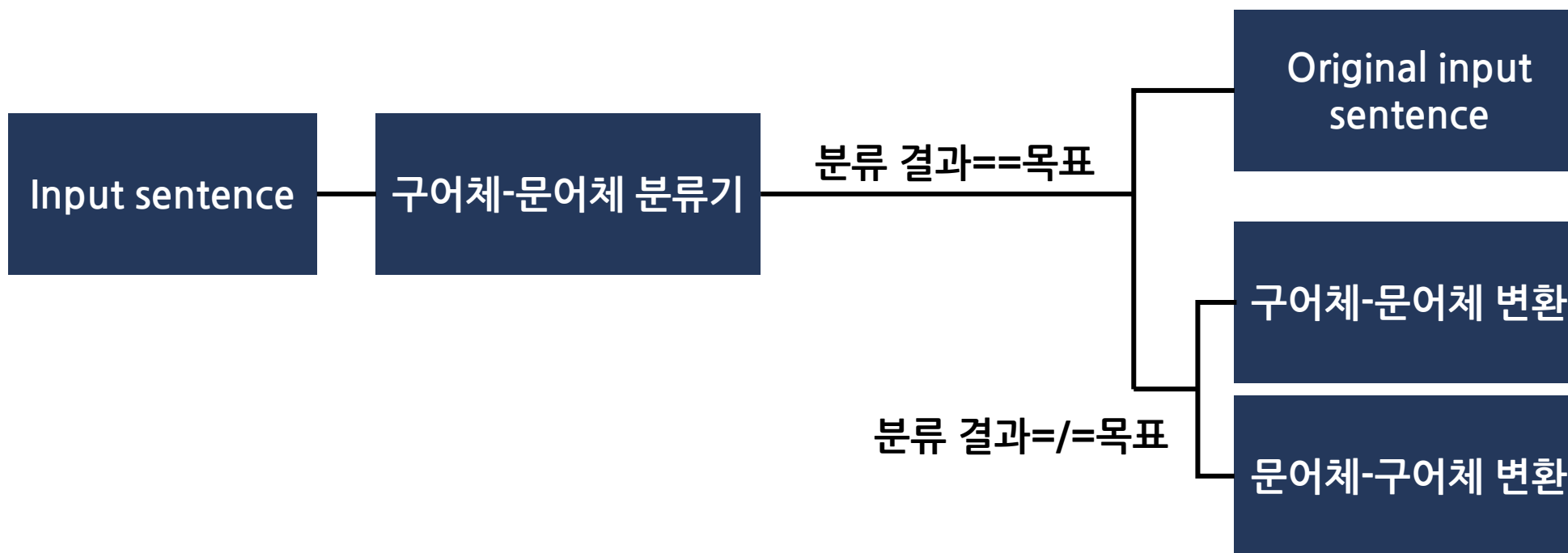
☐ Written 

Type your text here 

Convert 

Spoken-written-converter

Demo Architecture





Demo

원문	변환	goal
일기예보에 따르면 내일 비가 올 예정이다.	일기예보에 따르면 내일 비가 올 예정이에요.	spoken
일기예보에 따르면 내일 비가 올 예정이에요.	일기 예보에 따르면 내일 비가 내릴 것으로 보인다.	written

예시 문장

7. 자체 평가 및 추후 과제



자체 평가 및 추후과제

자체평가

구어체 -> 문어체는 대체적으로 변환이 잘 이루어지나, 문어체-> 구어체에서의 변환에 오류가 있음.

예)

- ‘-니다’라는 어미는 구어체/문어체 모두 사용이 될 수 있기에 엄격히 구분하기가 어려움.
- ‘나->저’와 같이 의미는 같으나 쓰이는 상황/상대가 다른 경우를 학습하지 못함.

추후과제

1. 데이터셋 품질 향상
 - 문장 부호가 없는 데이터, 인터넷 채팅체와같은 신조어 등 다양한 데이터셋 필요.
 - 데이터셋 크기 증가
2. 길이가 긴 코퍼스에 대해서도 변환이 가능하도록 보완.



역할 분담

양다경

- 아이디어 발표
- 구어체-문어체 병렬 데이터셋 필터링
- 구어체-문어체 변환기 구현 및 검증 (KoBART, T5)
- Streamlit 프로토타입 구현

구선헤

- 영어→한국어 문어체 번역기 구현 (ELECTRA-KoGPT2)
- 검증 메트릭 자료 조사

김유나

- 구어체-문어체 분류기 구현 (BERT)

김현수

- Google API를 이용한 구어체-문어체 병렬 데이터셋 구축
- Streamlit 프로토타입 구현

김현지

- AI-Hub 한국어-영어 번역(병렬) 말뭉치 데이터셋 분할
- Streamlit 프로토타입 구현

이용주

- 한국어 구어체→영어 번역기 구현 (KoELECTRA-GPT2)
- 최종 보고서 작성
- 최종 발표
- 사전학습 모델 자료 조사



References

<https://github.com/huggingface/notebooks/blob/main/examples/translation.ipynb>

<https://github.com/huggingface/notebooks/blob/main/examples/translation-tf.ipynb>

<https://huggingface.co/gogamza/kobart-base-v2>

<https://github.com/SKT-AI/KoBART>

<https://huggingface.co/gpt2>

<https://github.com/google-research/electra>

<https://github.com/monologg/KoELECTRA>

<https://huggingface.co/lcw99/t5-base-korean-paraphrase>

Streamlit

<https://spoken-written-converter.streamlit.app/>

Goorm 3rd Project

Any Questions?