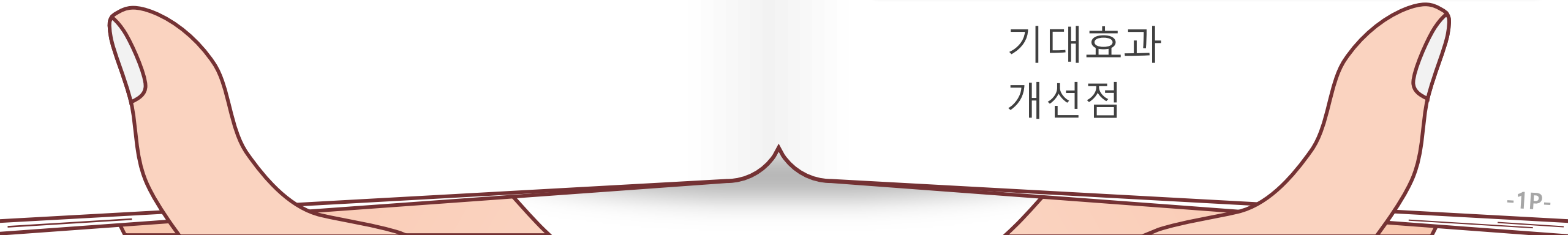
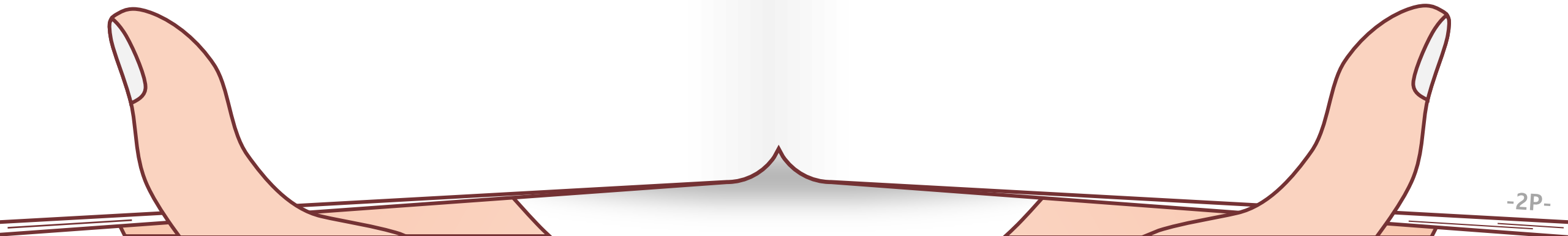
이정민
박민호

김민식
박승주
윤성식
이지평

CONTENTS

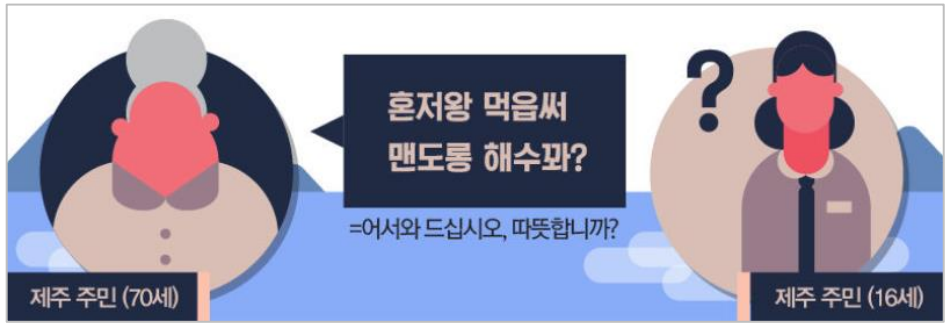


1. 주제 선정



1.주제선정 선정 배경

[사투리의 소멸]



제주도, 강원도 등 지에서는 사투리를 사용하는 비율이 낮아져 같은 지역의 젊은 층들도 지역 사투리를 알아듣지 못함

→ 지역 사투리의 소멸 가능성이 높아짐

[사투리 인식의 어려움]

기가 지니, 시리, 빅스비 등 음성인식 서비스에 사투리로 말을 걸 경우 말을 제대로 알아듣지 못하는 경우가 많음

→ AI가 표준어로 훈련됐기 때문

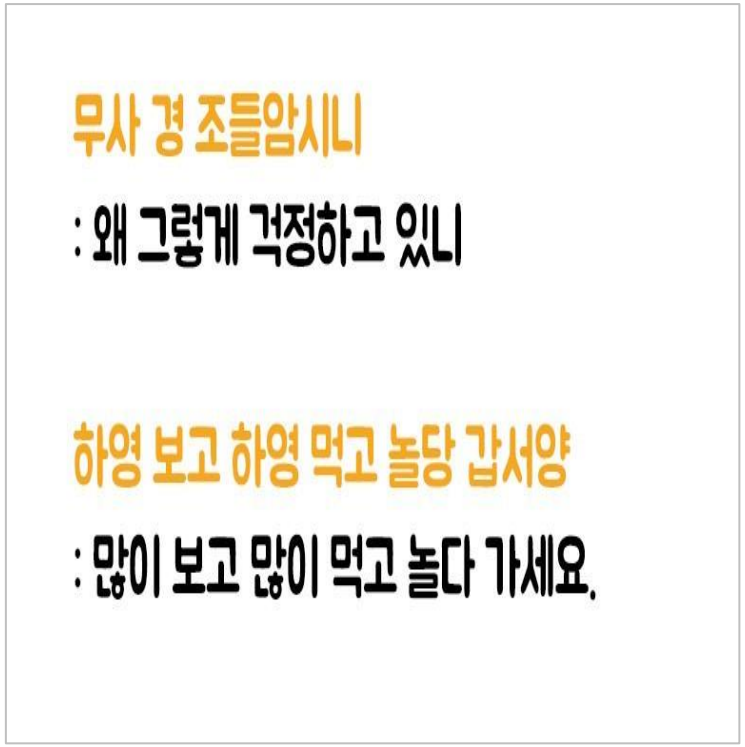
1.주제선정 주제

사투리 번역기

- 텍스트 데이터를 통해 표준어를 사투리로 바꿔 주는 번역기를 만들고자 함

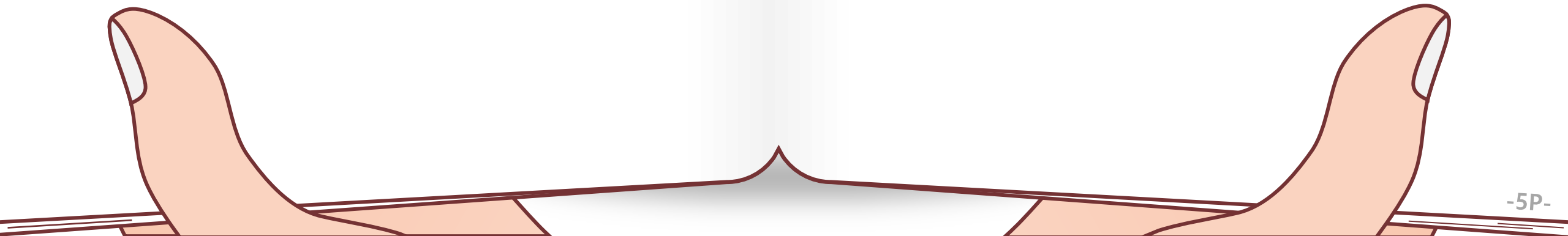
- 사투리는 어조에 따라 다양하게 분류될 수 있으나 텍스트데이터 만으로는 어조를 구분할 수 없음

- 특징이 뚜렷한 제주도 사투리를 우선적으로 사용하여 가능성을 평가함



<제주 사투리 예시>

2. 데이터 수집 / 전처리



2.데이터수집 사용데이터

한국어 방언 발화(제주도)

소개 | 다운로드

데이터셋명	한국어 방언 발화(제주도)			
데이터 분야	음성/자연어	데이터 유형		텍스트, 오디오
구축기관	솔트룩스	데이터 관련 문의처	담당자명	김민종(솔트룩스)
가공기관	이스트소프트		전화번호	02-2193-1656
검수기관	비투엔		이메일	mjkim@saltlux.com
구축 데이터량	432만	구축년도		2020년
버전	1.1	최종수정일자		2021.09.09
소개	방언(제주도)을 사용하는 일상 대화를 인식, 음성을 문자로 바꾸어주는 방언 발화 음성 데이터			

제주도 방언 데이터

- 2021년 6월 30일에 배포된 최신 데이터
- 방언에 대한 발화데이터와 표준어 발화데이터가 텍스트와 오디오 형식으로 존재
- 각 구역별 2000명 이상의 화자가 발화한 총 3000시간 이상 데이터 432만건 존재

2.데이터수집 데이터 전처리

내용	상세 설명	예시
개인정보 가명처리 **[&& 대체]**	(번호, 주소, 이름 등) 개인정보 비식별화 표시 예시: &name&	경허멍 이 저기양 &name4& 엄마 그
발화 소리(웃음, 박수 표시) **[제거]**	음성 발화 中 웃음•목청•박수•노래 부분 Tag	{laughing} 할 만해 {clearing}
(대화) 추정 문장 **[일부 제거]**	음성이 명확하지 않아 대화 내용을 추정해서 기입한 데이터	대충 점수 예상하고 ((가가지고))
특수 기호 **[제거]**	문장에서 의미를 가진 기호 (.?) 이외의 특수 기호 (@,*,-)	해그네 속에다 놩 @박스 맞아 옷을 만들어

2.데이터수집 데이터 전처리

내용	상세 설명	예시
선택 문제 (방언/표준어 대응쌍) **[일부 제거]**	데이터 中 "(방언 전사 형태) / (표준어 대응 쌍 형태)" 표시	아까 (집드레)/(집으로) (가라.)/(가더라.)
	선택 문제 형태 "()/()"꼴에서 벗어난 다양한 오류 존재	시간 많이 남아(있시난(있으니까) 지금은 좀
이/게 **[제거]**	선택 문제의 확장 Version(의미가 없는 단어의 선택 문제) - (게.)/(#게.), (이.)/(#이.)	서울(서)/(에서요)(게.)/(#게.)
짧은 문장 **[제거]**	문장의 길이가 짧은 데이터 존재	아

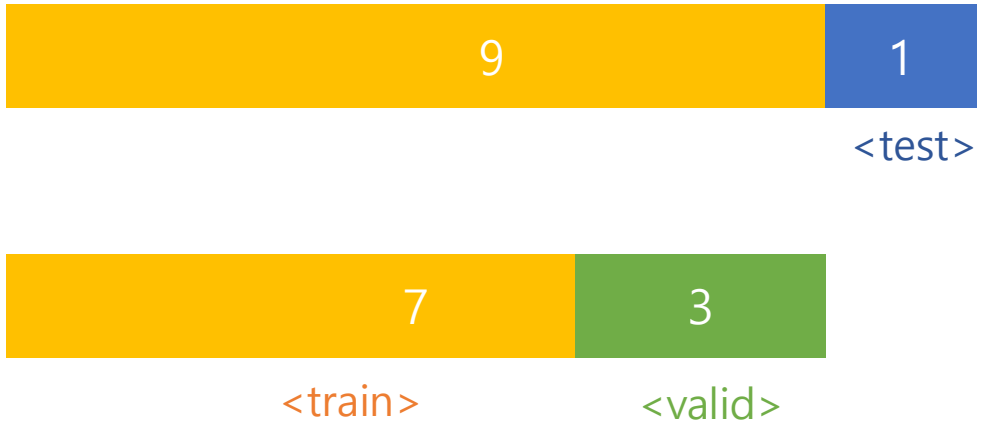
2.데이터수집 데이터 전처리

	방언 문장	표준어 문장
0	언니 만낭 반가워 아 오늘 제주 방언 에이 아이 데이터	언니 만나서 반가워 아 오늘 제주 방언 에이 아이 데이터
1	어 명절 설 명절 추석 명절 요로케 나누어서 해볼 거예	어 명절 설 명절 추석 명절 요로케 나누어서 해볼 거예요
2	자 그믄 이제부터 얘기해 보게예	자 그러면 이제부터 얘기해 봐요
3	어 그믄 언니네 설 명절 때 음식 어떵 해?	어 그러면 언니네 설 명절 때 음식 어떻게 해?
4	음식을 막 준비했던 기억이 나게	음식을 막 준비했던 기억이 나
...	...	...
1122200	가다가 뭐 하면 공항으로 보러 가게.	가다가 뭐 하면 공항으로 보러 가자.
1122201	밤까지는 가족이랑 있당	밤까지는 가족이랑 있다가
1122202	마무리 해그넹 저녁에 잠깐 나가크라.	마무리 해서 저녁에 잠깐 나갈게.
1122203	어 경 하라	어 그렇게 해라
1122204	난 차 어시난 우리 집 못가	난 차 없으니까 우리 집 못가

1122205 rows × 2 columns

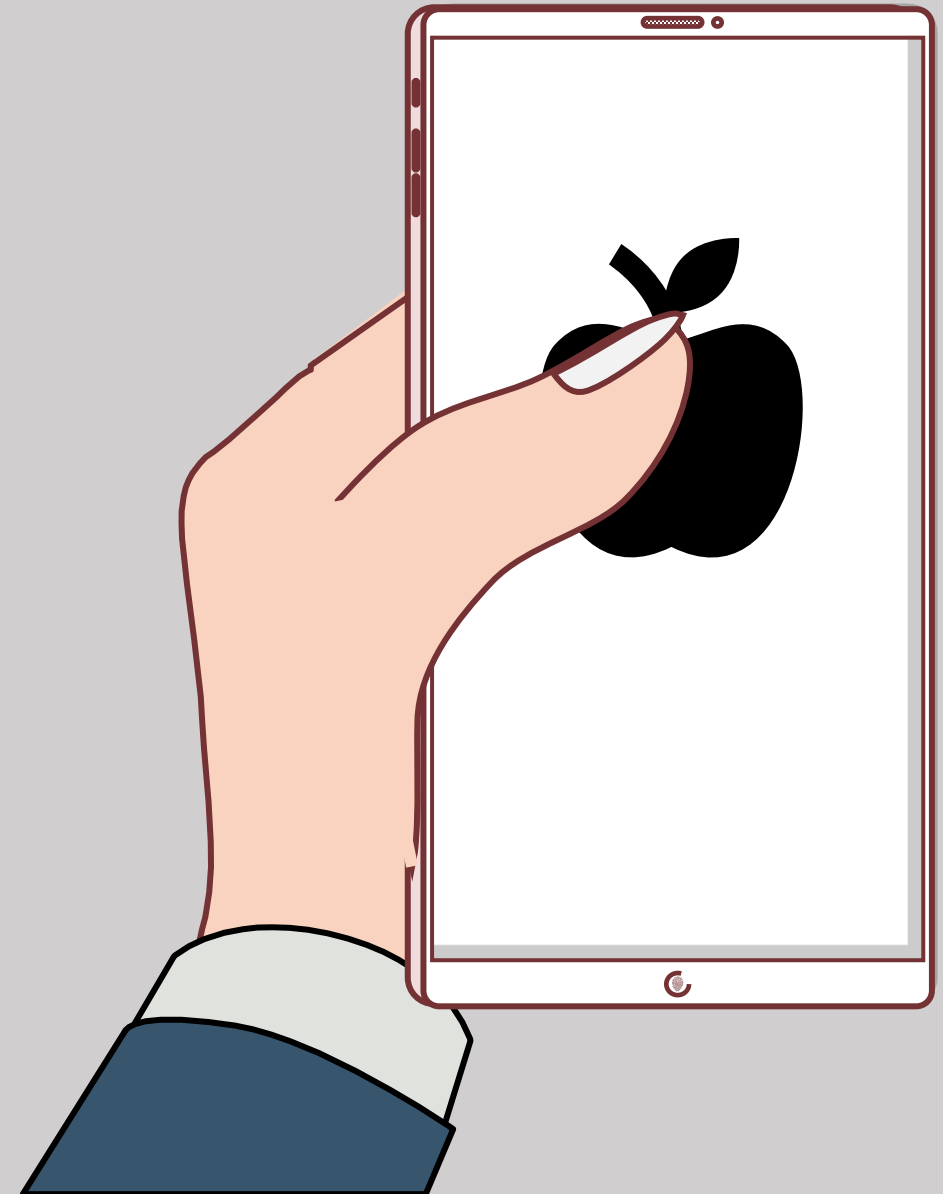
전처리 후 준비된 약110만개의 데이터를 사용

[Data Set 분할]



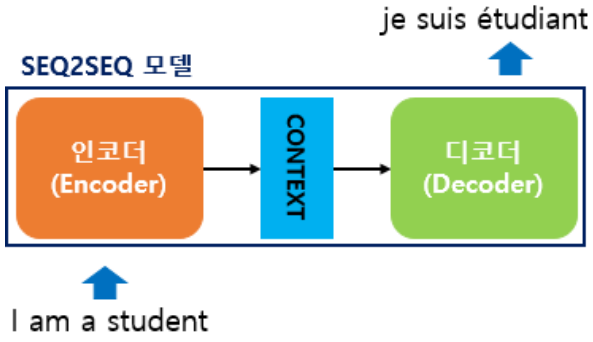
train과 test 비율은 9:1로 하여 dataset을 준비
train에서 다시 7:3으로 하여 validation set을 준비

3. 모델링



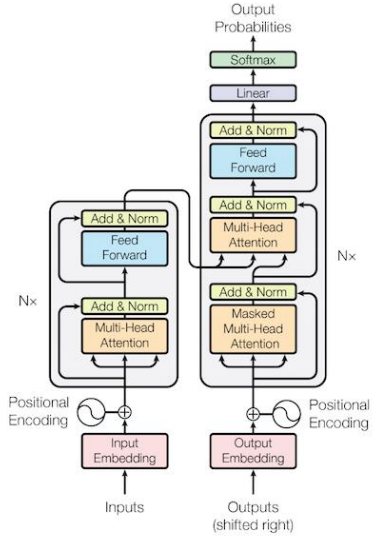
3.모델링 Translation model

Root Machine Translation

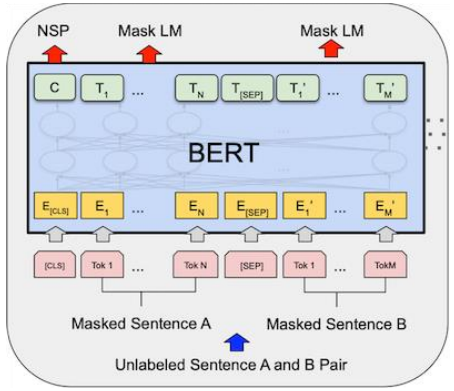


seq2seq

transformer

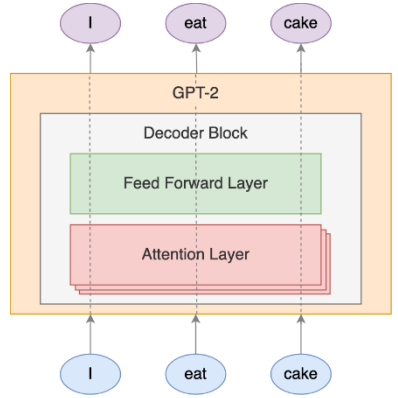


Pretrained Machine Translation



BERT

GPT



3.모델링 openNMT

OpenNMT

An open source neural machine translation system.

OpenNMT provides implementations in 2 popular deep learning frameworks:

OpenNMT-py

User-friendly and multimodal, benefiting from PyTorch ease of use.

- Documentation
- Pretrained models

OpenNMT-tf

Modular and stable, powered by the TensorFlow ecosystem.

- Documentation
- Pretrained models

The OpenNMT ecosystem also includes projects to cover the full NMT workflow:

CTranslate2

Fast inference engine for OpenNMT models.

Tokenizer

Fast and customizable text tokenization library with C++ and Python APIs.

nmt-wizard-docker

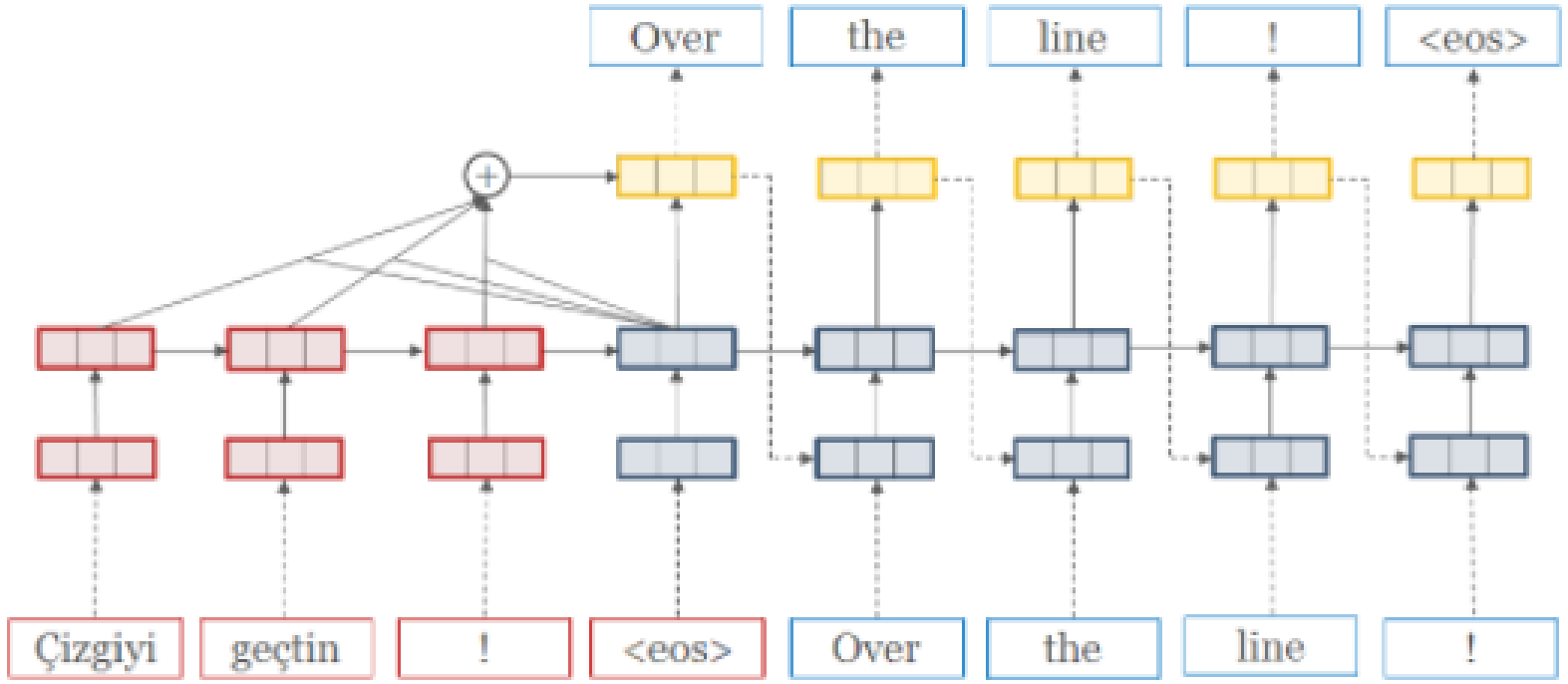
Docker-based wrapper for training and translating using a standardized interface.

nmt-wizard

Tasks launcher and monitor on remote platforms (SSH, EC2, etc.).

D&A Conference

3.모델링 openNMT



3.모델링 openNMT

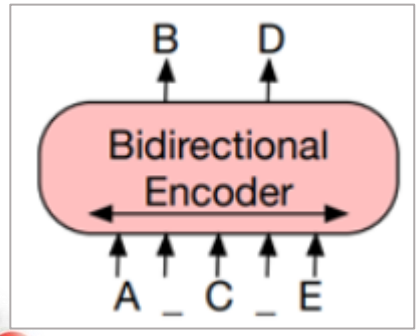
BPE

Sentence Piece

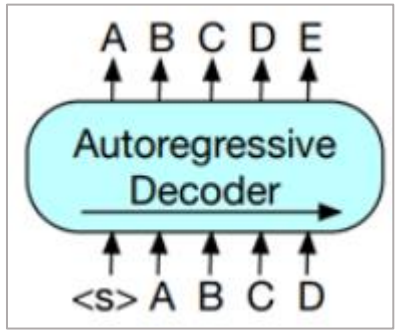
Onmt Tokenize

Mecab + SentencePiece

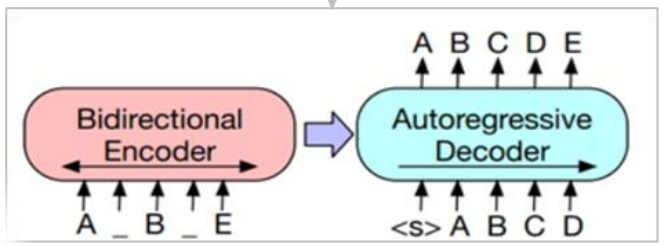
3.모델링 KoBART



<BERT>

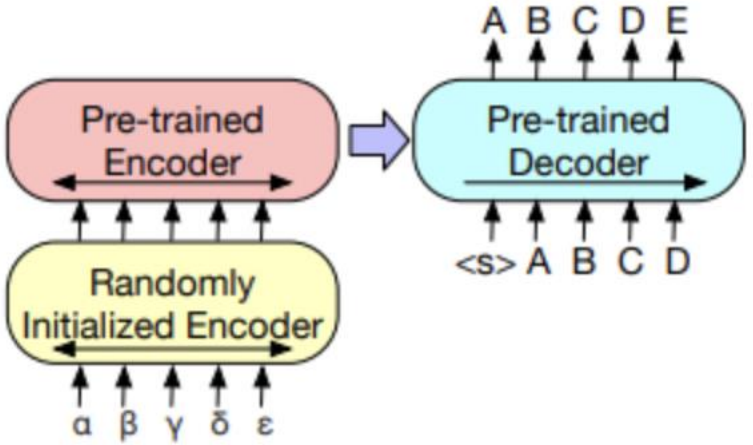


<GPT>



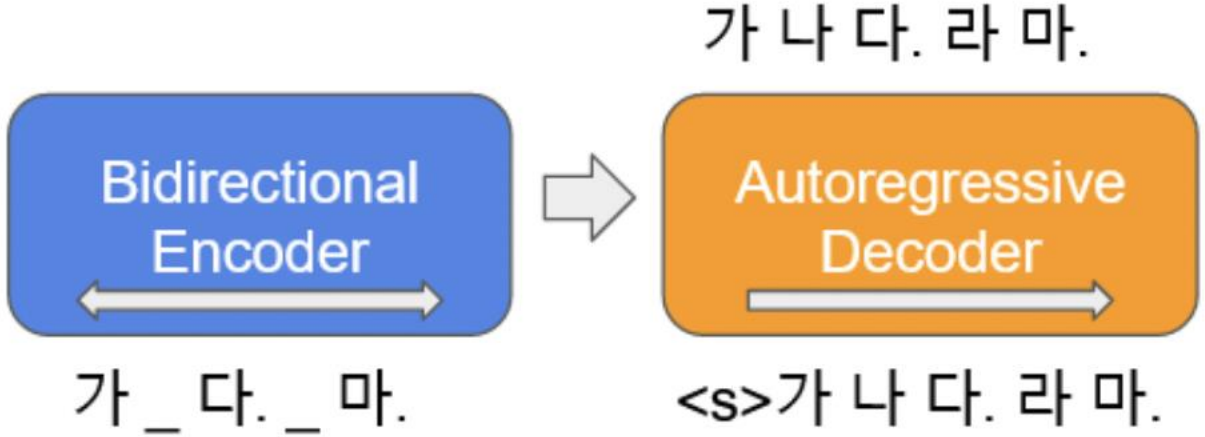
<BART>

3.모델링 KoBART

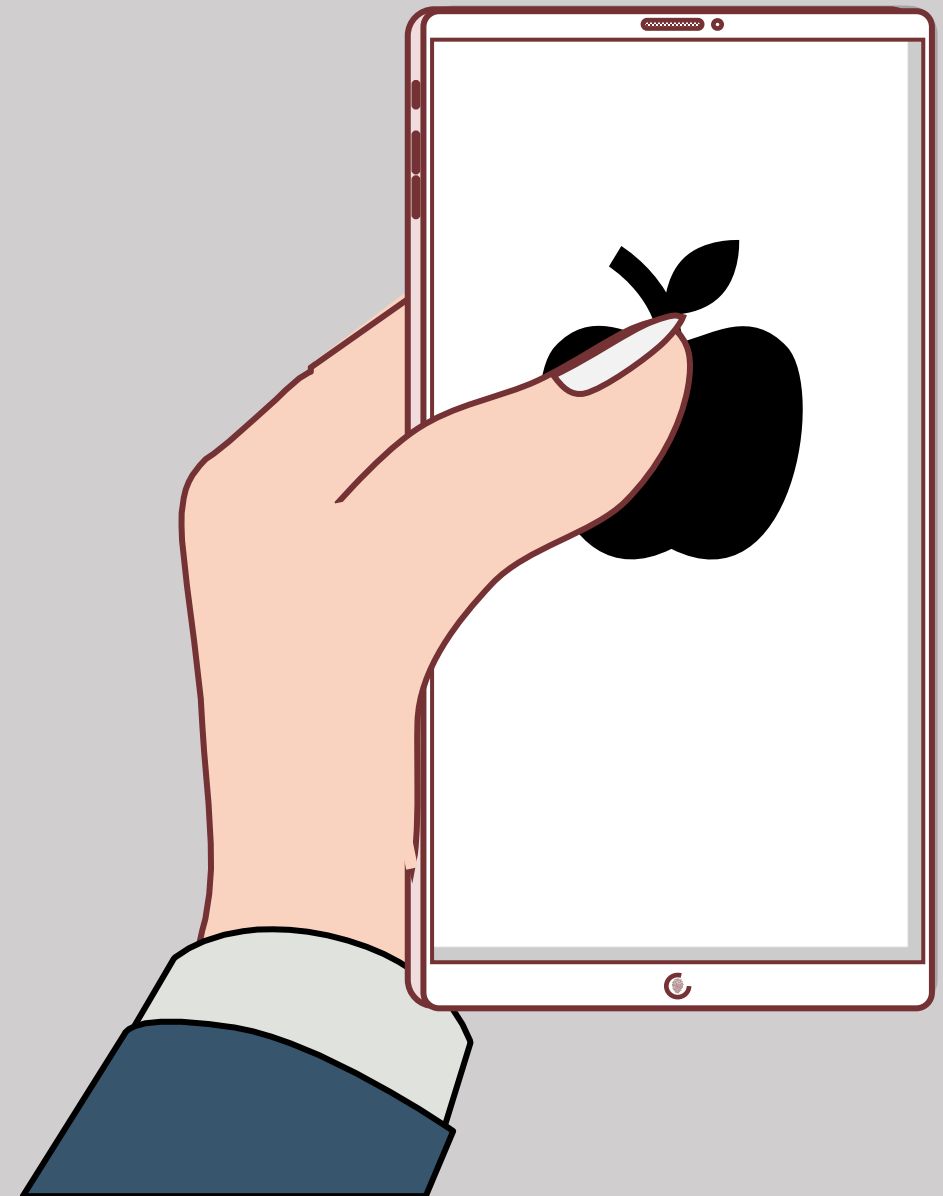


For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

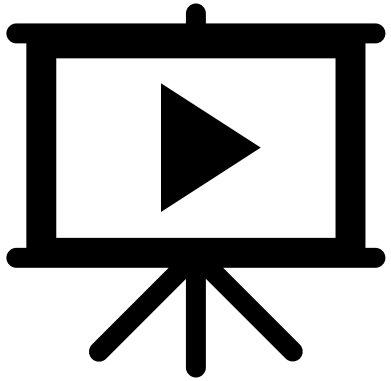
3.모델링 KoBART


가 나 다. 라 마.
Bidirectional Encoder
Autoregressive Decoder
가 _ 다. _ 마.
<s>가 나 다. 라 마.

4. 결론



4.결론 실제 번역



4.결론 실제 번역

번역된 사투리_OpenNMT

	ans	input	pred
0	뭐 겡 생물도 맨날 만점 받안. 만점 받았고 맨날 해나신디. 뭐 수능 망해부난	뭐 그래서 생물도 맨날 만점 받았어. 만점 받았고 맨날 했는데. 뭐 수능 망해서	뭐 그래서 입장 도 맨날 만점 받안 . 신경이 받았 고 맨날 해신디 . 뭐 수능
1	게난 근데 겅 하고 겅한 아이들 빼고는 사장이 완전 많아. 자영업이.	그니까 근데 그렇게 하고 그런 아이들 빼고는 사장이 완전 많아. 자영업이.	게난 근데 겅 하고 그런 아이들 빼고 사장이 완전 많아 . 자영업 이 .
2	그 이제 겅행 결국 나중에 영 동창회 때 보고 뭐 해고 했던 아이들은 공부 잘했던 ...	그 이제 그래서 결국 나중에 이렇게 동창회 때 보고 뭐 하고 했던 아이들은 공부 잘...	그 이제 그래서 결국 나중에 이렇게 각자가 때 보고 뭐 하고 했던 아이들은 공부 공...
3	왜냐하면은 잘 안 나가서. 잘 나가는 아이들은 공부 안 행 놀았던 그때 부지런히 놀...	왜냐하면은 잘 안 나가서. 잘 나가는 아이들은 공부 안 해서 놀았던 그때 부지런히 ...	왜냐하면 은 잘 안 나가서 . 잘 나가는 아이들은 공부 안 행 놀 그때 부지런히 잘...
4	어디 강 뭐 막 무슨 어디 회장도 하고 뭐도 막 햄서.	어디 가서 뭐 막 무슨 어디 회장도 하고 뭐도 막 해.	어디 강 뭐 막 무슨 어디 어디 하고 뭐 도 막 해 .
...	...	...	...
995	그니까 나가 촘아야지게 영 생각하멍 혼자 막 속으로 앓고 겅해신디	그니까 내가 참아야지 이렇게 생각하면서 혼자 막 속으로 앓고 그랬는데	게난 내가 참아야지 영 생각하멍 혼자 막 속으로 사주는 고 그래신디
996	그래도 이젠 좀 괜찮아진 거 닮 사람들 다.	그래도 이젠 좀 괜찮아진 거 같아 사람들 다.	그래도 이젠 좀 괜찮아 진 거 닮아 사람들 다 .
997	어 나 아까 막 음료수 먹어부난	어 나 아까 막 음료수 먹어서	어 나 아까 막 크리스마스 먹엉
998	배가 안 고픈게 오늘 저녁 뭐 나올 건지 궁금하다.	배가 안 고프네 오늘 저녁 뭐 나올 건지 궁금하다.	배가 안 움직이니까 네 오늘 저녁 뭐 나올 건지 삶이 .
999	어떵 확인을 해 볼 수 어서?	어떻게 확인을 해 볼 수 없어?	어떵 확인 을 해 볼 수 어서 ?

1000 rows x 3 columns

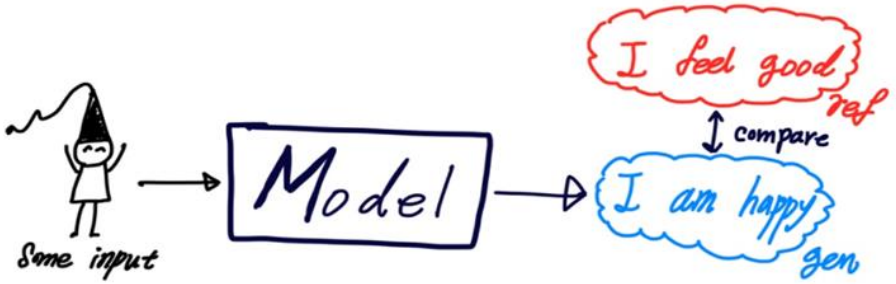
4.결론 실제 번역

번역된 사투리

_KoBART

	ans	input	pred
0	탁 잡아가고 영 영 이렇게 이렇게 서로 영	탁 잡아가고 이렇게 이렇게 이렇게 이렇게 서로 이렇게	탁 잡아가고 영 영 영 영 영 서로 영
1	그렇게 해서 엿도 만들고 이 경 했주게 그니까	그렇게 해서 엿도 만들고 그렇게 했지 그니까	경 행 엿도 만들고 이 경 했주 게난
2	아 식구는 막 우리 그디 그쪽에 우리 또	아 식구는 막 우리 거기 그쪽에 우리 또	아 식 아 식구는 막 우리 그디 그쪽에 우리 또
3	육 남매에다가 뭐 우리 식구밖에 어시난 이	육 남매에다가 뭐 우리 식구밖에 없으니까	육 남매에다가 뭐 우리 식구밖에 어시난 이
4	겅 해신디 우리 큰 집	그렇게 했는데 우리 큰 집	경 해 해신디 우리 큰 집
5	크 나 큰 집에서 우리 할머니도 인제 같이 사셔시난 이	크 나 큰 집에서 우리 할머니도 인제 같이 사셨으니까	크 나 크 큰 집에서 우리 할머니도 인제 같이 사셔시난 이
6	우리 큰 집은 큰 집 나름대로 했던 거 닮아	우리 큰 집은 큰 집 나름대로 했던 거 같아	우리 큰 우리 큰 집은 큰 집 나름대로 했던 거 닮아
7	용돈 용돈 할때 세뱃돈 줄 때 어른들은 명절만 되믄	용돈 용돈 할때 세뱃돈 줄 때 어른들은 명절만 되면	용돈 용돈 할때 세뱃돈 줄 때 어른들은 명절만 되면
8	세뱃돈 일일이 다 챙기잰 하민	세뱃돈 일일이 다 챙기려고 하면	세뱃돈 일일이 다 챙기젠 하면
9	그 봉투에 일일이 이름 적어가멍 누구	그 봉투에 일일이 이름 적어가면서 누구	그 봉투에 일일이 이름 적어가멍 누구

4.결론 BLEU score



얼마나 많은 **Generated Sentence**의 단어가
Reference Sentence에 포함되는가?

EX) Reference Sentence: I was referenced by human.
Generated Sentence: I was generated by the model.

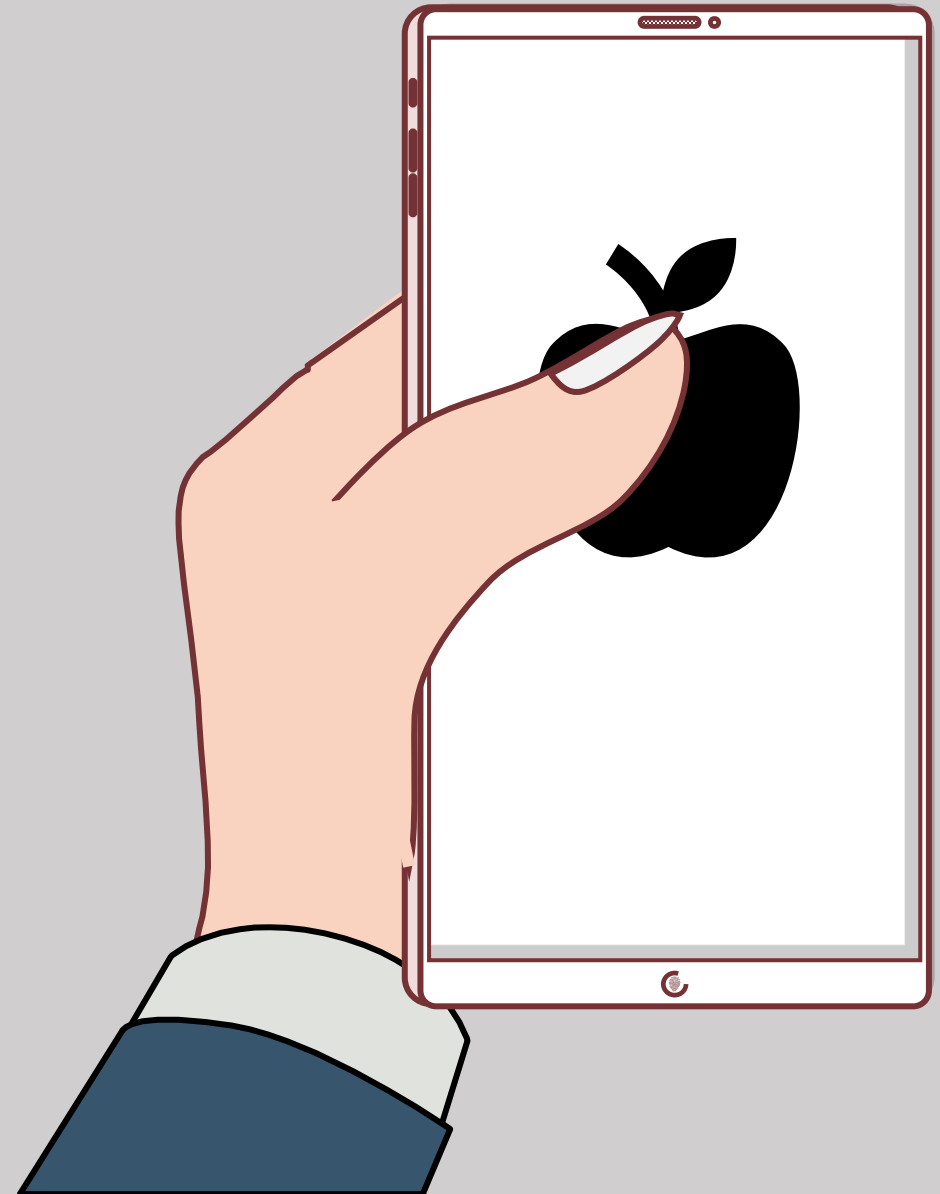
Bleu Score = 3/6

4.결론 BLEU score

BLEU 점수	해석
10점 미만	거의 의미 없음
10~19점	핵심을 파악하기 어려움
20~29점	요점은 명확하지만 많은 문법적 오류가 있음
30~40점	이해할 수 있는 양호한 번역
40~50점	고품질 번역
50~60점	매우 우수한 품질의 적절하고 유창한 번역
60점 초과	대체적으로 사람보다 우수한 품질

KoBART	OpenNMT
49.6	45.2

5. 향후계획



5.기대효과/개선점

기대효과

사투리 언어 보존

더 나은 AI 음성인식 서비스

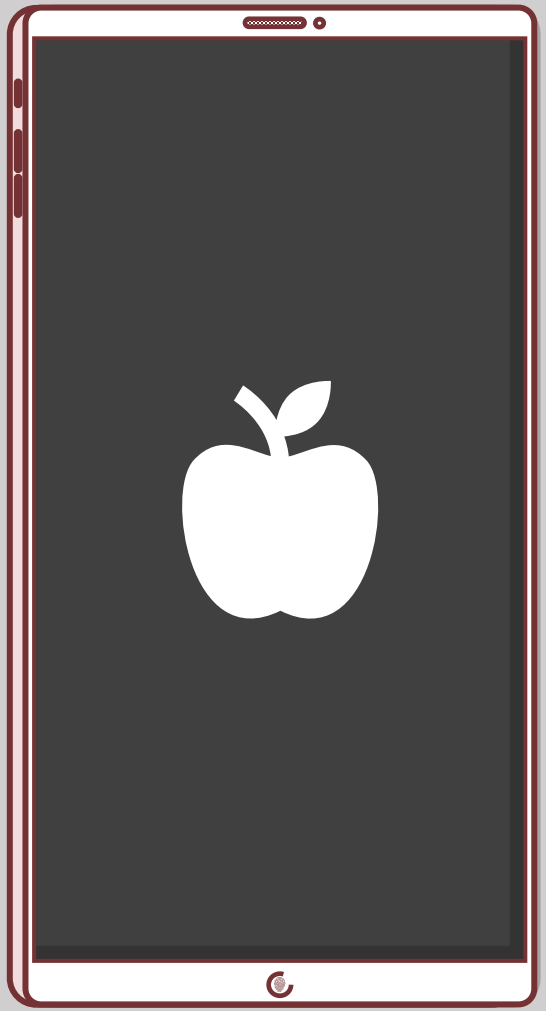
사투리를 사용하는 컨텐츠 및 영상매체에 대한 번역 개선

5. 기대효과/개선점 개선점

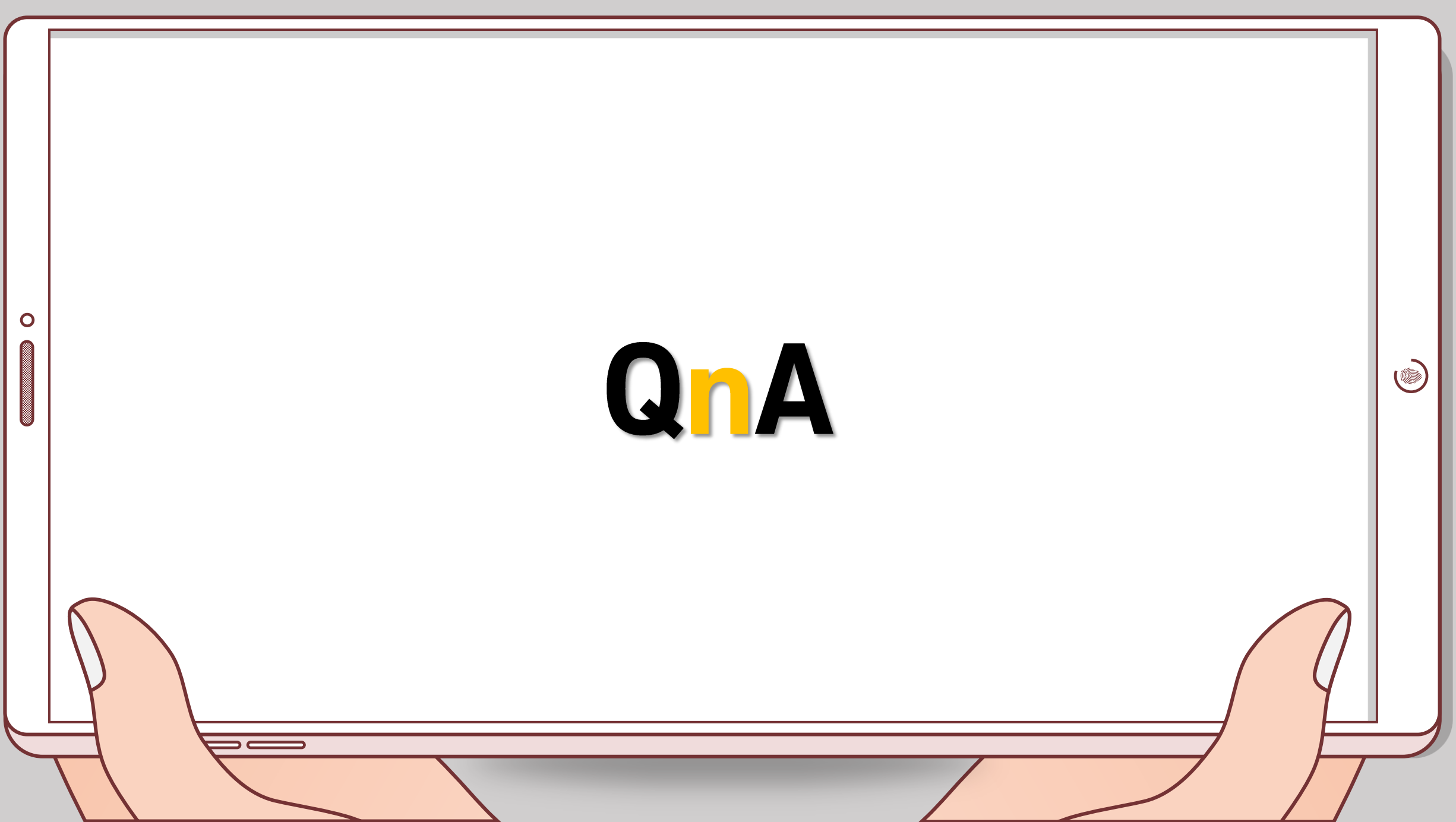
기술적 제약과 컴퓨팅 파워 제한으로
모델을 더 많은 epoch으로 학습 시키지 못함

→추후 더 많은 data및 epoch으로 학습 요망

제주도 사투리 뿐 아니라 다른 지역
사투리로 번역 가능하도록 개선



END


QnA

Appendix

<참고문헌>

https://dladustn95.github.io/nlp/BART_paper_review/

https://github.com/seujung/KoBART-translation

https://github.com/OpenNMT/OpenNMT-py

https://github.com/spongepad/Korean-Honorific-Translation

https://jrc-park.tistory.com/273

Appendix

내용	처리 방식	변경 전/후 예시
익명성 처리 (개인정보 가명처리)	개인정보 비식별화를 위한 표현 (&name&, &company-name&) 모두 "&&" 동일 처리 (대상: 주민등록 번호, 신용카드, 주소, 전화번호, 이름)	[전] - 어디 팔 건가? &company-name1&도 &company-name2&에 팔 건가 - 경허멍 이 저기양 &name4& 엄마 그
		[후] - 어디 팔 건가? &&도 &&에 팔 건가 - 경허멍 이 저기양 엄마 그
기타 소리 (웃음, 목청, 박수, 노래 처리)	음성 中 웃음•목청•박수•노래 태그 (예: {laughing}, {clapping}) 부분을 제거	[전] - 밀가룬지 쌀가룬지를 모르크라 {laughing} 근데 그거를 푹 - {laughing} 할 만해 {clearing}
		[후] - 밀가룬지 쌀가룬지를 모르크라 근데 그거를 푹 - 할 만해

내용	처리 방식	변경 전/후 예시
추측성 표시 처리 (음성 대화 처리 과정 中 잘 들리지 않아 추측)	추정 불가능하거나 일부만 추정 가능한 경우 해당 부분 제거 추정을 할 수 있는 경우 해당 부분을 살림	[전] [1] 추정 불가능: 하루종일 놀다 (()) [2] 일부 추정: ((x)) 하네 [3] 추정 가능: 그때 대충 점수 예상하고 ((가가지고))
		[후] [1] 추정 불가능: 하루종일 놀다 [2] 일부 추정: 하네 [3] 추정 가능: 나는 그때 대충 점수 예상하고 가가지고

Appendix

내용	처리 방식	변경 전/후 예시
★택 문제 (방언/표준어 대응 쌍 표시)	표준어 규정에 벗어난 방언을 "(방언 전사 형태) / (표준어 대응 쌍 형태)" 표시됨 방언 문장에서는 방언을 표준어 문장 대응되는 표준어를 선택	[전] - 방언: 아까 (집드레)/(집으로) (가라.)/(가더라.) - 표준어: 아까 (집드레) / (집으로) (가라.)/(가더라.) [후] - 방언: 아까 집드레 가라. - 표준어: 아까 집으로 가더라

내용	처리 방식	변경 전/후 예시
★택 문제 (방언/표준어 대응 쌍 표시)	[다양한 데이터 오류 존재] [1] (방언(표준어) 형태 [2] (방언)?(표준어) 형태 [3] (방언)/(표준어) 형태의 오류 [4] 방언, 표준어 중 대응이 안됨 (둘 중 한 곳만 택 문제 존재) [처리 방식] [1]와 [2] 최대한 살림 (방언 문장은 방언, 표준어 문장은 표준어) [3]과 [4] 방식은 제거	[예시] [1] 시간 많이 남아(있시난(있으니까) 지금은 좀 [2] 궁금한거 (있수꽈?)?(있어요?) [3-1] 거난 나 이거 고르켜 줄크메/주겠으니) (잘해봅서예 [3-2] 집에 전화 ()해도 가인게 나왕게)/() 전화 (받느냐? [4] 방언: 그 그 원두를 나 이제 요작에 /(않아) 표준어: 그 그 원두를 나 이제) 가야 될 건데 하나 사다 줘? 막 비싸지 (않애/(않아) [후] - [1] 방언: 시간 많이 남아있시난 지금은 좀 - [1] 표준어: 시간 많이 남아있으니까 지금은 좀 - [2] 방언: 궁금한거 있수꽈? - [2] 표준어: 궁금한거 있어요? - [3]과 [4] 해당하는 행을 지움

Appendix

내용	처리 방식	변경 전/후 예시
'이', '게' 단어 처리	이, 게에 관련된 단어 혹은 택 문제가 존재 제거하는 방향으로 진행(의미가 없음)	[전] 방언 문장: 서울(서)/(에서요)(게.)/(#게.) 표준어 문장: 서울(서)/(에서요)(게.)/(#게.)
		[후] (게.)/(#게.) 제거 후 택 문제 해결 방언 문장: 서울서 표준어 문장: 서울에서요.

내용	처리 방식	변경 전/후 예시
특수 기호 처리	특수 기호(@, *, -) 문장 속 제거	[전] - 미녕 짠거를 해그네 속에 다 놩 @박스 맞아 옷을 만들어
		[후] - 미녕 짠거를 해그네 속에 다 놩 박스 맞아 옷을 만들어
짧은 문장 처리	문장의 길이가 2 이하인 짧은 문장을 제거	[전] - 아
		[후] - 제거됨