



■ 전공용어 변환 서비스, TermCorrector

소프트웨어융합학과
한주상, 선신욱, 한지훈

T TERM-CORRECTOR

CONTENTS

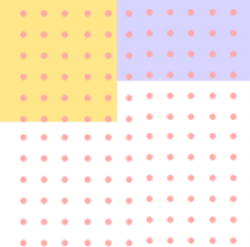
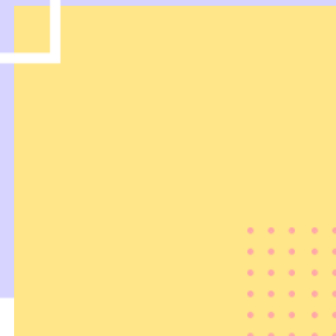
1 프로젝트 소개

4 시연 과정

2 모델 관련

5 결론

3 프로그램 구조




프로젝트 소개

- 많은 대학생들은 강의 내용을 녹음해두었다가 STT(speech-to-text) 변환하여 학습에 활용함
- 하지만 기존 STT 서비스는 일상 대화는 변환을 잘 해주지만, 전공 용어는 잘 변환해주지 못함

■ 기존 STT 서비스 한계

External Fragmentation



교수님

익스터널
컬그래테이션?

이스8
프로멧테이션?

익스터널
프라미테이션?

→ 전공 용어 변환에 한계

■ 생성형 AI 수정 한계

wrtm 대부분의 틀린 부분
인식 불가

답변


글에서 틀린 부분을 고쳐보겠습니다.

그러니까 아까도 얘기했지만 익스 마피그라미테이션을 해결할 수 있다.

ChatGPT 텍스트 요약을 자동으로
시도하여 중요 부분/단어 사라짐

자, 앞서 우리는 메모리 매이닝에 대해 다루었고, 지난 시간부터는 메모리 할당에 대해 이야기를 시작했습니다. 가장 기본적인 방식 중 하나는 파린선된 서버가 사용하는 방식입니다. 앞서 다 이야기한 예시 문 것지만, 메모리 주소 공간을 연속적인 형태로 나누는 것이 아니라, 파티션을 이용해 메모리 블록을 할당하는 것입니다. 이 방법은 물리적 메모리(Physical Memory)에 데이터를 연속적으로 저장 및 관리하는 방식입니다.

■ 학습시 불편함



- 강의 녹음을 공부에 활용하기 위해 사용자는 일일이 잘못된 단어가 무엇인지 고민해야 함
- 수정 시에도 예상 단어 선택지가 없어 모두 직접 입력해야 하며, 그 과정에서 오타도 발생 가능

-> 강의 PDF의 키워드들을 추가적으로 학습하여 전공 용어 인식률을 높이고 학습 활용도를 높임

M4A + PDF + W → 전공 용어에 대한 음성 텍스트 변환의 정확도 높이고 학습 자료 활용도 개선

모델 관련

■ 키워드 추출

- 사용자가 업로드한 강의 자료(PDF)에서 영어 단어에 한 해 전공 용어 키워드 추출
- 문장 단위로 토큰화 및 품사 태깅을 진행하여 NN(명사), NP(고유 명사), JJ(형용사)만 추출
- TF-IDF 방식은 일부 키워드를 놓칠 수 있는 한계가 있어 전공 용어가 아닌 키워드를 제외하는 방향으로 키워드 추출 진행

■ 데이터 증강

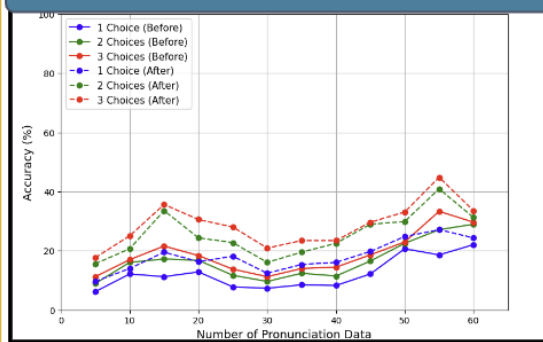
- 추출한 키워드를 g2pK 라이브러리를 사용하여 한글 발음으로 변환
- 자음과 모음에 맞는 노이즈 방식을 적용하여 글자의 초성/중성/종성 중 하나를 변형하는 방식으로 다양한 발음 변형 데이터 생성

	target	발음1	발음2	발음3	발음4	발음5
0	abstraction	에브스 트랙션	캐드스 트랙션	에브시 트랙션	에브스 트랙션	캐브스 트랙션
1	access	엑세스	엑떼스	엑세즈	엑세스	익세스
2	address	에드레스	에브레스	에드리스	캐드레스	에드레트
3	addressable	어드레서벌	에느레써벌	어드레서벌	에드레서벌	허드로서벌
4	allocated	엘러케이티드	엘러케히티드	엘러케이티드	엘러케에티드	엘러코이티드

■ 모델 학습

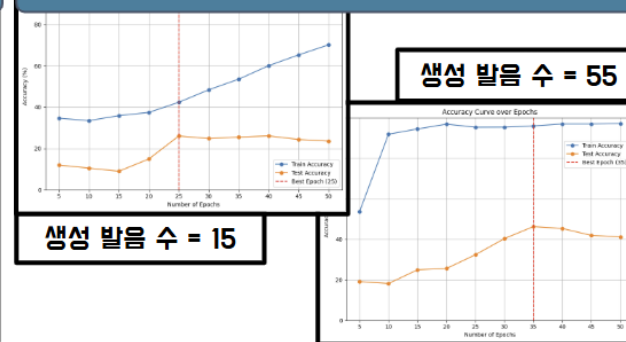
- 데이터 증강 기법으로 생성한 다양한 발음 데이터를 통해 fastText 모델 학습 진행

생성 발음 데이터 수 및 후처리에 따른 모델의 성능



- Top-1 Accuracy, Top-2 Accuracy, Top-3 Accuracy 3가지로 모델 성능 측정
- 발음 수 15개, 55개일 때 높은 정확도
- 후처리 진행 시, 높은 정확도

Epoch 변화에 따른 Train/Test Accuracy curve



- 발음 수 15의 경우 Epoch = 25에서 최적
- 발음 수 55의 경우 Test Acc은 살짝 높지만 Train data에 대한 overfitting 발생

모델 관련

■ 키워드 추출

- 사용자가 업로드한 강의 자료(PDF)에서 영어 단어에 한 해 전공 용어 키워드 추출
- 문장 단위로 토큰화 및 품사 태깅을 진행하여 NN(명사), NP(고유 명사), JJ(형용사)만 추출
- TF-IDF 방식은 일부 키워드를 놓칠 수 있는 한계가 있어 전공 용어가 아닌 키워드를 제외하는 방향으로 키워드 추출 진행

■ 데이터 증강

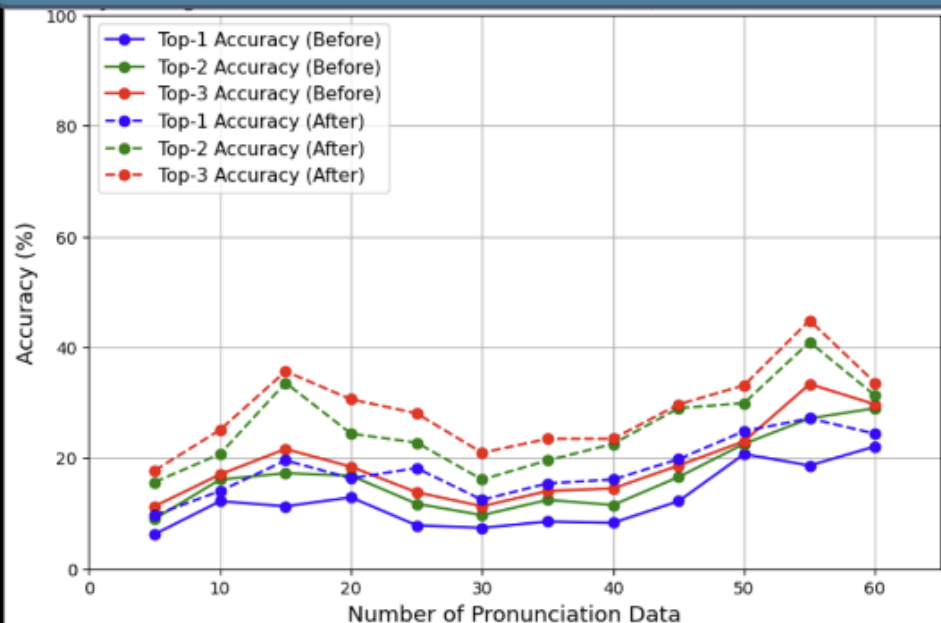
- 추출한 키워드를 g2pK 라이브러리를 사용하여 한글 발음으로 변환
- 자음과 모음에 맞는 노이즈 방식을 적용하여 글자의 초성/중성/종성 중 하나를 변형하는 방식으로 다양한 발음 변형 데이터 생성

	target	발음1	발음2	발음3	발음4	발음5
0	abstraction	애브스 트랙션	캐드스 트랙션	애브시 트랙셀	애브스 트랙션	캐브스 트랙션
1	access	엑세스	엑떼스	엑세쯔	엑세스	익쌔스
2	address	애드레 스	애브레 스	애드뢰 쓰	깨드레 스	애드레 트
3	addressable	어드레 서벌	어느레 써벌	어드레 서멸	애드레 서범	허드로 서벌
4	allocated	앨러케 이티드	앨러케 히티드	엘러께 이티드	앨러케 에티드	앨러코 이테드

■ 모델 학습

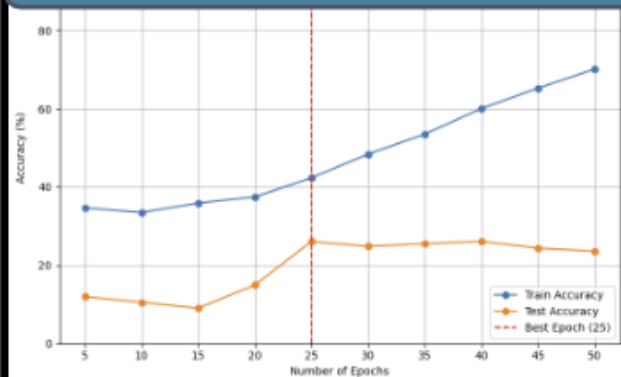
- 데이터 증강 기법으로 생성한 다양한 발음 데이터를 통해 fastText 모델 학습 진행

생성 발음 데이터 수 및 후처리에 따른 모델의 성능



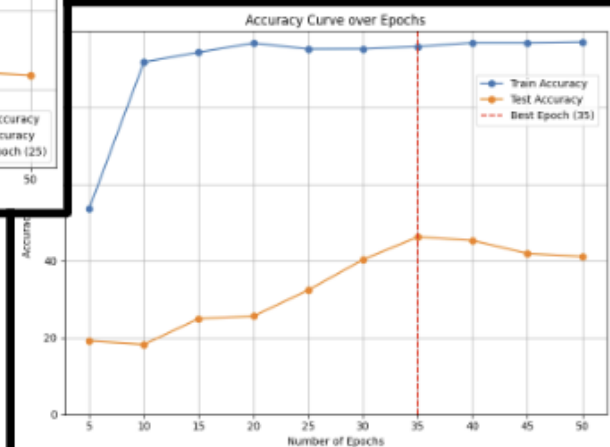
- Top-1 Accuracy, Top-2 Accuracy, Top-3 Accuracy 3가지로 모델 성능 측정
- 생성 발음 수 15개, 55개일 때 높은 정확도
- 후처리 진행 시, 정확도 향상

Epoch 변화에 따른 Train/Test Accuracy curve



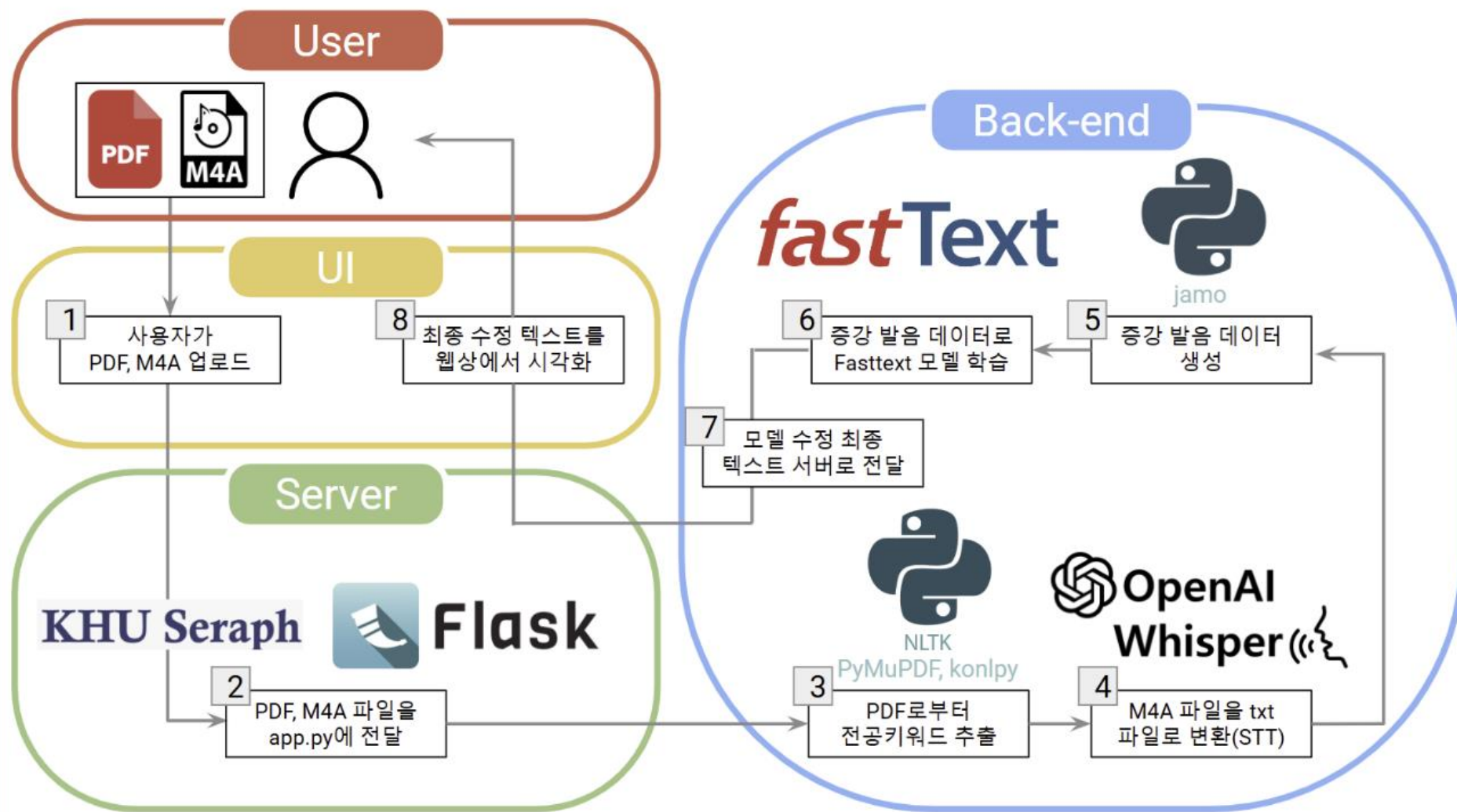
생성 발음 수 = 15

생성 발음 수 = 55



- 생성 발음 수 15의 경우 Epoch = 25에서 최적
- 생성 발음 수 55의 경우 Test Acc은 살짝 높지만 Train data에 대한 overfitting 발생 우려

프로그램 구조



1 메인 페이지



프로젝트 기능 소개

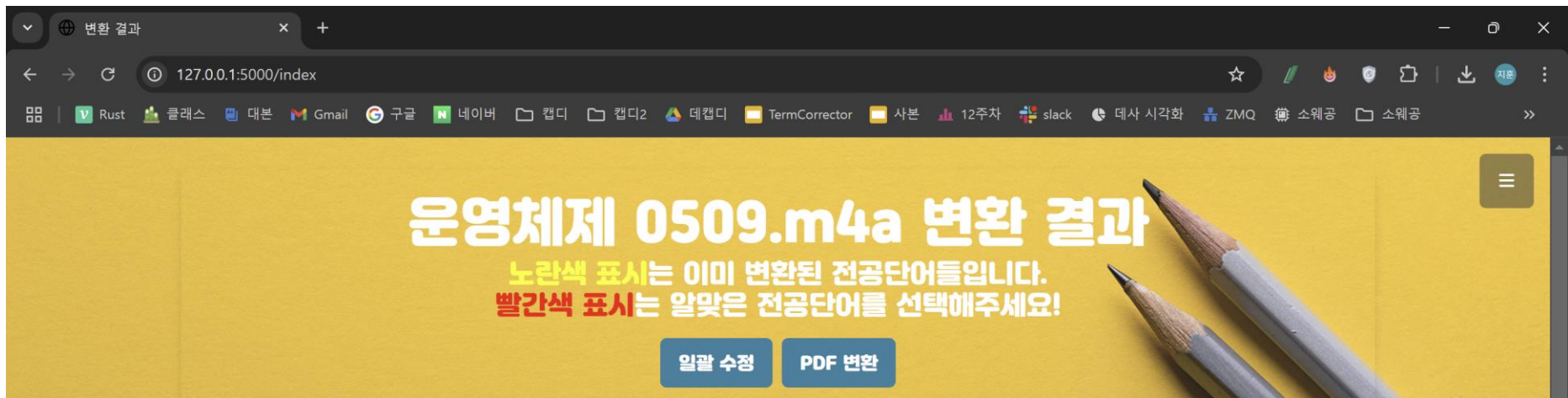
사용자가 수강한 강의 녹음 파일, 강의 자료 업로드

2 로딩 페이지



STT 변환, fastText 모델 적용동안 로딩
로딩이 완료되면 결과 페이지로 이동

3 결과 페이지



드레스

address

컨티비어스라게

contiguous

이렇

컨티비어소을

contiguous

만

하이라이트를

hierarchical

비

transparent

translation

트랜기션을 해가지고

segmentation

implementation

언론스테이션 정책에 따라서 저

오버이도가

overhead

결

어소시티프

associate

머

인스터덕션들이

instruction

순천

레이시는

ratio

중요

transparent

translation

트랜티슈 하는 게 아니라

paging

page

페이브에 있는 일

space

spatial

스페셜을 높이

꼴

수정

트랜션하는

translation

hole

이라고 하는 문제를

internal

external

fragmentation

mani

익스터라

프래규먼트스

익스터라

프래규먼트스

롯데이션,

fastText 모델이 수정한 단어 직접 확인
선택지 중 옳은 단어 선택 및 직접 입력 가능

4 PDF로 변환

운체 0509.m4a 변환 결과

문장 6 [00 : 06 : 29]

자, 그래서 어찌 됐든, external fragmentation이 문제니까, 이 external fragmentation을 어떻게 없앨 수 있느냐, 이것을 생각을 하다가 만들어낼게, 자, physical memory address를 단이 높이 크기만큼으로, 바쪽에서, 이 단이만큼으로만 쓸 수 있게, 이렇게 쪽에 있고, logical memory address 도, 저 physical memory 얻은 쪽에 놓은 그 단이만큼으로, 쪽이 가질 수 있는 것 같아요. 그래서 그 똑같은 단이만큼으로, 이제 헛단을 해주기 시작하는 거예요. 헛단을 해서 쓰는, 이런 방식을 찾게 하는 게, page table 입니다. 그래서 이제 logical memory 댕에 있는, 그러니까 우리가 잘 알고 있는, logical memory 댕에 있는, 요 파트는 이제 페이지라고 들어요. 그 다음, physical memory 또 같은 크기로 쪽해져 있는, 요 단이들을, 그러니까 회인이라고 들어요.

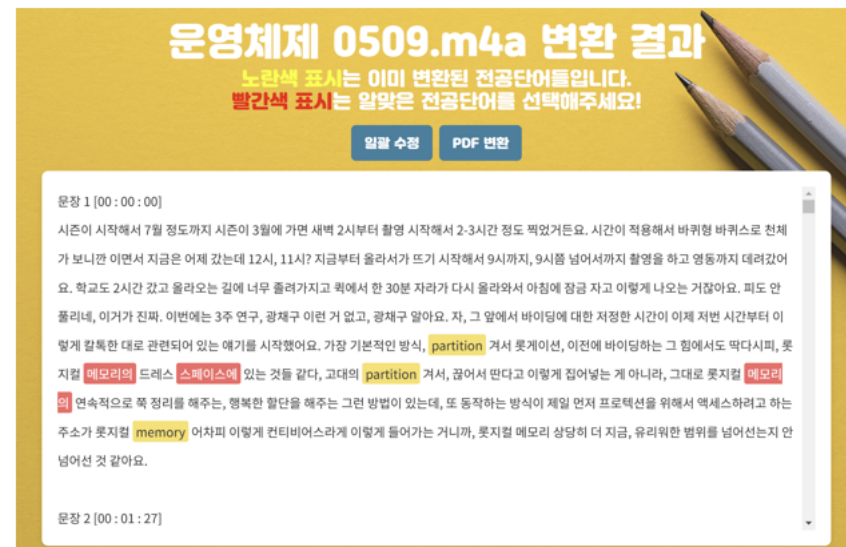
최종 변환 결과를 PDF로 저장
웹 뿐만 아니라 로컬 파일로 확인 가능

결론



강의 자료(PDF)에서 전공 용어 키워드를 추출하고,
데이터 증강 기법을 활용하여 전공 용어 변환 오류 수정

향후, 사용자가 유사한 domain의 여러 강의들을 업로드하여 학습하면, 이전 강의들의 전공 용어들이
모두 DB에 축적되어 전공 용어 변환 정확도 향상을 통해 더욱 유용한 학습 자료로서 활용 가능



웹상으로 수정 결과를 직관적으로 확인 가능,
바로 학습에 활용 가능

x

THANK YOU!

발표 들어주셔서 감사합니다

*