

# 공학 프로젝트 기획 번역기 성능 향상

자모 단위 변환 & 높임말, 낮춤말 변환

허재무, 김준태, 김주환, 김정희 2021.11.5(금)

# 번역기 설명

-단계

Preprocessing

Translating

Apply

# 기존의 preprocessing 단계

## -순서

data\_step1.sh -> kr,en file을 shuffle해주는 단계

data\_step2.sh -> file을 train, valid, test로 split해주는 단계

data\_step3.sh -> 문장을 tokenize해주는 단계

data\_step4.sh -> 단어를 symbolize해주는 단계

data\_step5.sh -> bpe생성 및 bpe적용하는 단계

data\_step6.sh -> aihub와 hgu\_clean 병합 및 shuffle하는 단계

# 기존의 preprocessing 단계

-`data_step1.sh`, `data_step2.sh`

`data-step1.sh` -> 문장의 순서를 random하게 shuffle해줌

`data-step2.sh` -> 파일을 train, valid, test로 나눠줌  
(Ex. `aihub.kr` -> `aihub.train.kr` , `aihub.valid.kr`, `aihub.test.kr`)

# 기존의 preprocessing 단계

## -data\_step3.sh

```
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr]  
[  
버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않자 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다.  
문재인 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다.  
이 같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다.  
구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다.  
약 2만 5000 달러의 학자금대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 워싱턴포스트에 말했다.  
서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다.  
도지사는 용자를 결정한 날부터 15일 안에 신청인에게 용자금을 교부하여야 한다.  
청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다.  
법 제14조 제1항 제3호에서 "대도시가 아닌 시 또는 군으로서 시·도 조례로 정하는 경우"란 영별표 1에 해당하는 정비계획 입안대상지역의 토지등소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다.  
한화토탈 관계자는 "SM와 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다.  
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr.tok]  
[  
버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않자 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다.  
문재인 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다.  
이 같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다.  
구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다.  
약 2만 5000 달러의 학자금대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 워싱턴포스트에 말했다.  
서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다.  
도지사는 용자를 결정한 날부터 15일 안에 신청인에게 용자금을 교부하여야 한다.  
청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다.  
법 제14조 제1항 제3호에서 "대도시가 아닌 시 또는 군으로서 시·도 조례로 정하는 경우"란 영별표 1에 해당하는 정비계획 입안대상지역의 토지등소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다.  
한화토탈 관계자는 "SM와 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다.]
```

# 기존의 preprocessing 단계

## -data\_step4.sh

```
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr.tok]  
[  
  버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않자 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다.  
  문재인 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다.  
  이같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다.  
  구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다.  
  약 2만 5000 달러의 학자금대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 워싱턴포스트에 말했다.  
  서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다.  
  도지사는 응자를 결정한 날부터 15일 안에 신청인에게 응자금을 교부하여야 한다.  
  청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다.  
  법 제14조 제1항 제3호에서 "대도시가 아닌 시 또는 군으로서 시·도조례로 정하는 경우"란 영별표 1에 해당하는 정비계획 입안대상 지역의 토지등소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다.  
  한화토탈 관계자는 "SM와 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다.  
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr.sym]  
[  
  버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않자 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다.  
  __P0 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다.  
  이같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다.  
  구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다.  
  약 2만 5000 달러의 학자금대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 __P0에 말했다.  
  서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다.  
  도지사는 응자를 결정한 날부터 15일 안에 신청인에게 응자금을 교부하여야 한다.  
  청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다.  
  법 제14조 제1항 제3호에서 "대도시가 아닌 시 또는 군으로서 시·도조례로 정하는 경우"란 영별표 1에 해당하는 정비계획 입안대상 지역의 토지등소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다.  
  __P0토탈 관계자는 "SM와 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다.]
```

# 기존의 preprocessing 단계

## -data\_step5.sh

```
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr.tok.sym
[  버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않자 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다 .
  __P0 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다 .
  이 같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다 .
  구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다 .
  약 2만 5000 달러의 학자금 대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 __P0에 말했다 .
  서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다 .
  도지사는 융자를 결정한 날부터 15일 안에 신청인에게 융자금을 교부하여야 한다 .
  청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다 .
  법 제14조 제1항 제3호에서 "대도시가 아닌 시 또는 군으로서 시·도 조례로 정하는 경우"란 영별표 1에 해당하는 정비계획 입안대상지역의 토지 등 소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다 .
  __P0도 탈 관계자는 "SM와 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다 .
[nmt21@gpu1:~/hgu_data/processed_testjamo$ head aihub.train.kr.tok.sym.10000sub.safe
[  버스 기사가 경찰에 진술한 내용에 따르면 당시 버스 기사는 자동문이 작동하지 않아 브레이크를 걸고 하차해 살펴보던 중 갑작스럽게 사고가 발생했다 .
  __P0 대통령도 "사고 원인을 알아보는 상황이지만 우리 기업이 참여하고 있는 만큼 정부가 나서야 한다"고 했다 .
  이 같은 장점이 알려지면서 공모전 경쟁률은 100대 1을 넘어섰다 .
  구청장은 민관협치 활성화를 위하여 필요한 조사·연구 및 교육 프로그램을 개발하고 민관협치 관련 홍보 등의 업무를 수행하기 위하여 노력하여야 한다 .
  약 2만 5000 달러의 학자금 대출이 있는 한 학생은 "스미스 말을 듣고 눈물이 차올랐다"며 "경제적으로 어려웠는데, 새 출발을 하게 됐다"고 __P0에 말했다 .
  서툰 것 가락질에 포크로 짜장면을, 손으로 탕수육을 집어 먹으면서도 "맛있다"를 연발했고, 앉을 자리가 없어 대기하는 손님까지 생길 정도였다 .
  도지사는 융자를 결정한 날부터 15일 안에 신청인에게 융자금을 교부하여야 한다 .
  청와대 민정수석실 공직감찰반의 감찰 범위를 사실상 제한하는 법안이 발의된다 .
  법 제1400조 제1항 제300호에 "대도시가 아닌 시 또는 군으로서 시·도 조례로 정하는 경우"란 영별표 100에 해당하는 정비계획 입안대상지역의 토지 등 소유자가 시장·군수에게 정비계획의 입안을 제안하는 경우를 말한다 .
  __P0도 탈 관계자는 "SM과 잔사유를 분리해 이송하는 DA205라는 설비에 이상이 생긴 것을 근본 원인으로 보고 있다"고 밝혔다 .
```

# 기존의 preprocessing 단계

## -data\_step6.sh

aihub파일과 hgu\_clean파일을 합쳐서 shuffle해주는 단계

# 변경된 preprocessing 단계

## -순서

low\_data.sh -> kr file을 모두 낮춤말로 변환하는 단계

data\_step1.sh -> kr, en file을 shuffle하는 단계

data\_step2.sh -> file을 train, valid, test로 split하는 단계

data\_step3.sh -> 문장을 tokenize하는 단계

data\_step4.sh -> 단어를 symbolize하는 단계

data\_step\_jamo.sh -> 문장을 자모 단위로 변환하는 단계

data\_step5.sh -> bpe생성 및 bpe적용하는 단계

data\_step6.sh -> aihub와 hgu\_clean 병합 및 shuffle하는 단계

# Jamo 단위 변환

# 자모 단위 변환

## -예시

```
[nmt21@gpu1:~/jamo_jeonghui$ cat test test.jamo
안녕하세요. 저는 김정희입니다.
ㅇㅏㄴㄴㅕㅇㅎㅏㄴㅅㅔㅇㅍㅇ. ㅈㅓㄴㄴㄱㅣㅁㅈㅓㅇㅎㅓㅇㅣㅂㅓ_|_ㄷㅓ_.
```

# 자모 단위 변환

## -기대 효과

data\_step1.sh -> kr, en file을 shuffle하는 단계

data\_step2.sh -> file을 train, valid, test로 split하는 단계

data\_step3.sh -> 문장을 tokenize하는 단계

data\_step4.sh -> 단어를 symbolize하는 단계

data\_step\_jamo.sh -> 문장을 자모 단위로 변환하는 단계

data\_step5.sh -> bpe생성 및 bpe적용하는 단계

data\_step6.sh -> aihub와 hgu\_clean 병합 및 shuffle하는 단계

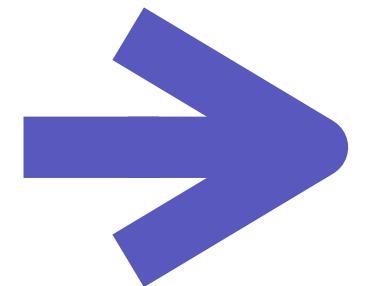
# 자모 단위 변환

## -bpe란?

oov(out of vocabulary)문제를 해결하기 위한 알고리즘

Bpe 적용후

단어  
low  
lower  
newest  
widest



사전

es

est

lo

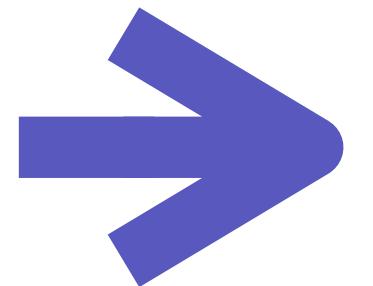
low

ne

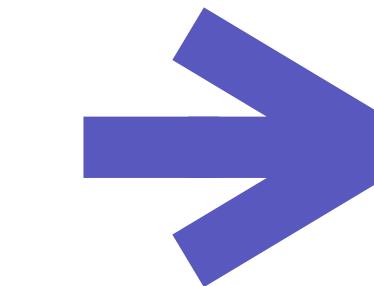
new

wi

wid



Input  
lowest



low, est로 인식함

# 자모 단위 변환

## -bpe 및 번역 예시

IP & Date: 221.164.46.182: 2021-11-05-15:48:37[k2e]  
Input: 정 흰 곧 집에 갈 것이다 .  
Symbolized: 정 흰 곧 집에 갈 것이다 .  
BPE: 정@0 흰 곧 집에 갈 것이다 .  
Translated: We will go home soon , Jeong@0 n@0 ae .  
Output: We will go home soon, Jeongnae.

IP & Date: 221.164.46.182: 2021-11-05-15:48:41[k2e]  
Input: 정 흰 곧 집에 갈 것이다 .  
Symbolized: 정 흰 곧 집에 갈 것이다 .  
BPE: 짜 | ㅇ @0 흐 - ㄴ ㄱ ㅗ ㄷ 짜 | ㅂ ㅇ ㅋ \_ ㄱ ㅏ ㄹ ㄱ ㅓ ㅅ ㅇ | \_ㄷ ㅏ \_ .  
Translated: Jeong-@0 hee will go home soon .  
Output: Jeong-hee will go home soon.

IP & Date: 221.164.46.182: 2021-11-05-15:51:46[k2e]  
Input: 정 흰 정말 대단 해 .  
Symbolized: 정 흰 정말 대단 해 .  
BPE: 정@0 흰 정말 대단@0 해 .  
Translated: It 's really amazing how we do .  
Output: It's really amazing how we do.

IP & Date: 221.164.46.182: 2021-11-05-15:51:43[k2e]  
Input: 정 흰 정말 대단 해 .  
Symbolized: 정 흰 정말 대단 해 .  
BPE: 짜 | ㅇ @0 흐 - ㄴ ㅈ ㅓ ㅇ ㅁ ㅏ ㄹ ㄷ ㅐ \_ ㄷ ㅏ \_ ㅇ @0 ㅎ ㅐ \_ .  
Translated: Jun@0 gh@0 ee is really amazing .  
Output: Junghee is really amazing.

IP & Date: 221.164.46.182: 2021-11-05-15:53:57[k2e]  
Input: 인 생 은 즐 거 워 .  
Symbolized: 인 생 은 즐 거 워 .  
BPE: 인생@0 은 즐거@0 워 .  
Translated: Life is a joy .  
Output: Life is a joy.

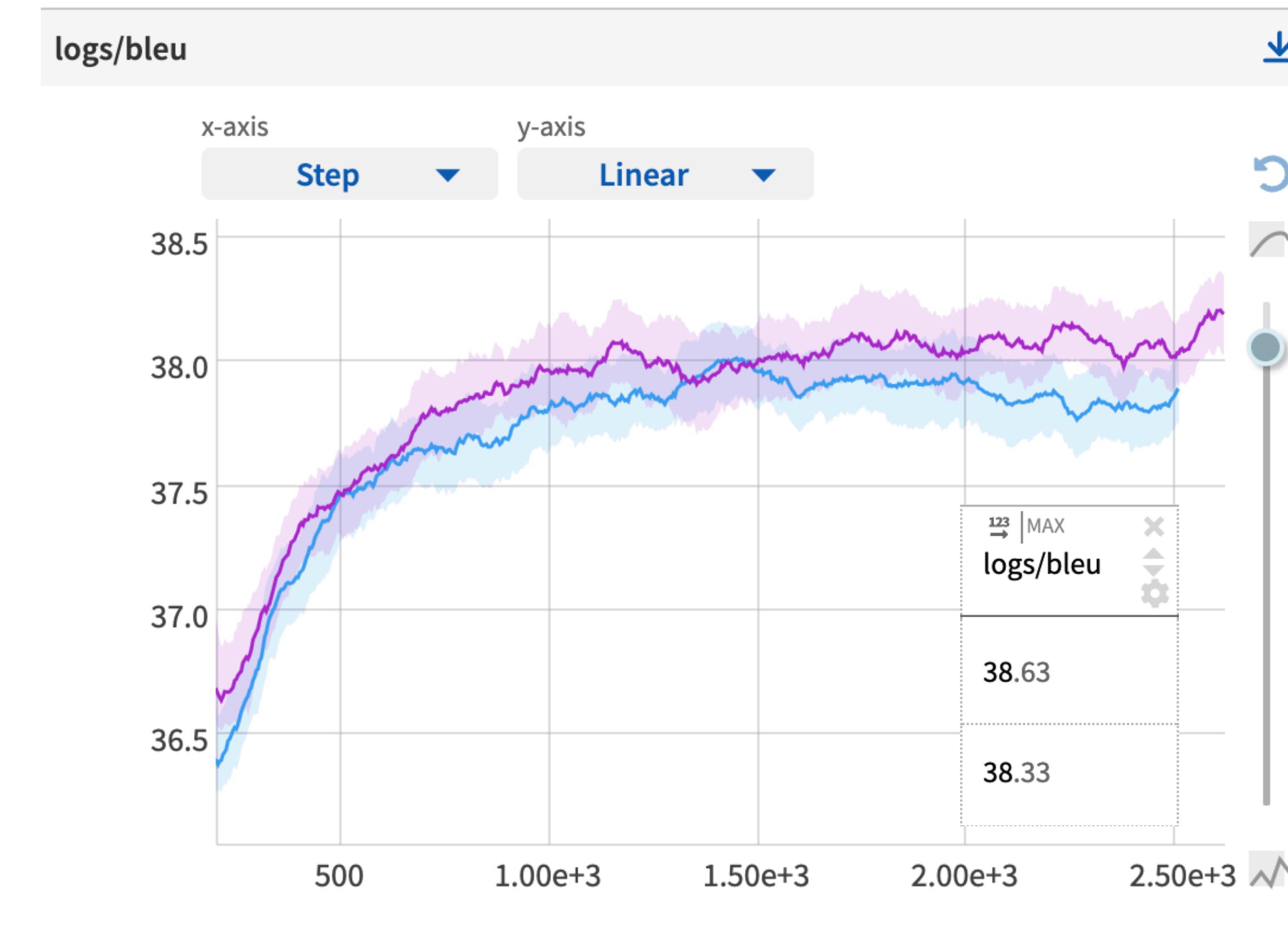
IP & Date: 221.164.46.182: 2021-11-05-15:54:06[k2e]  
Input: 인 생 은 즐 거 워 .  
Symbolized: 인 생 은 즐 거 워 .  
BPE: ㅇ | ㄴ ㅅ ㅐ ㅇ @0 ㅇ \_ ㄴ ㅈ \_ ㄹ ㄱ \_ ㅇ @0 ㅋ \_ .  
Translated: Life is full of pleas@0 ure .  
Output: Life is full of pleasure.

IP & Date: 221.164.46.182: 2021-11-05-15:47:05[k2e]  
Input: 형 젠 용 감 했다 .  
Symbolized: 형 젠 용 감 했다 .  
BPE: 형@0 젠 용@0 감@0 했다 .  
Translated: Bro@0 ther was b@0 ra@0 ve .  
Output: Brother was brave.

IP & Date: 221.164.46.182: 2021-11-05-15:47:12[k2e]  
Input: 형 젠 용 감 했다 .  
Symbolized: 형 젠 용 감 했다 .  
BPE: ㅎ ㅋ ㅇ @0 ㅈ @0 ㅋ \_ ㄴ ㅇ ㅍ ㅇ @0 ㄱ ㅏ ㅁ ㅇ @0 ㅎ ㅐ ㅆ ㄷ ㅏ \_ .  
Translated: He was b@0 ra@0 ve .  
Output: He was brave.

# 자모 단위 변환

## -성능 비교(bleu)



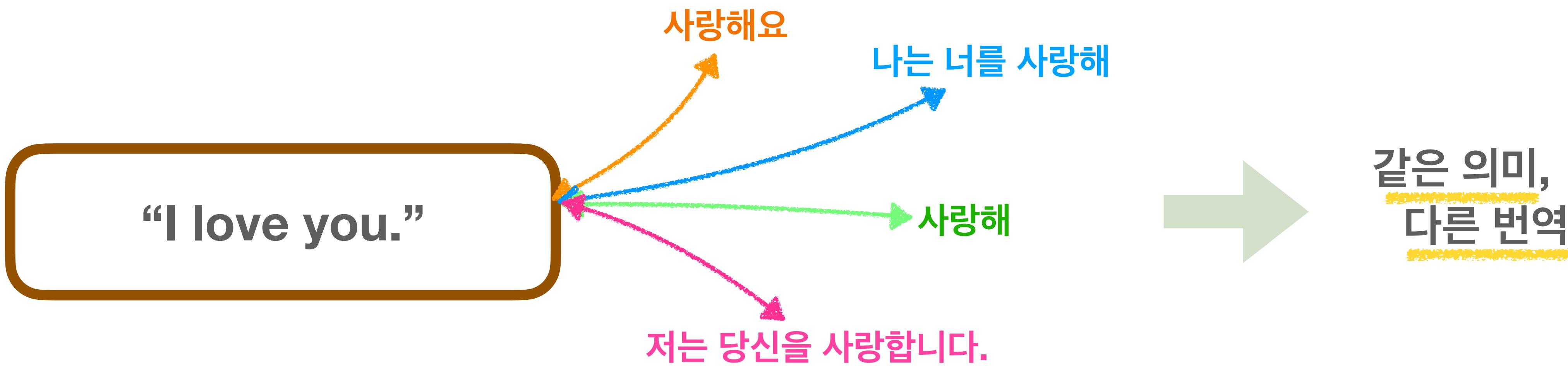
보라색 : 자모  
파란색 : 기존

높임말, 반말 변환

# 높임말 ↔ 반말 변환이 왜 필요한가?

목적 및 기대 효과

- 영어에 없는 한국어의 높임말 때문에 한가지 영어 문장에 대해 여러 한국어 문장이 번역이 가능합니다. 올바른 번역에 *BLEU score*가 낮게 나올 수 있습니다.



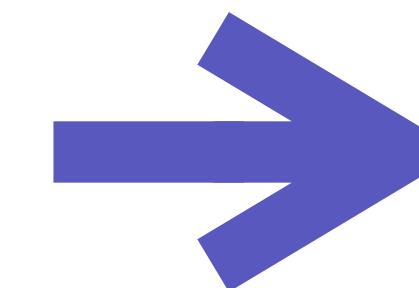
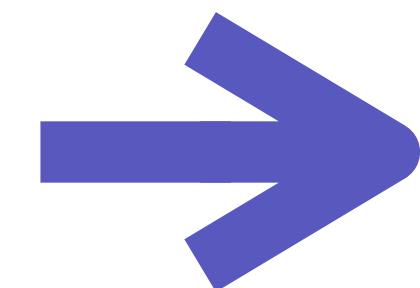
# 번역의 성능을 개선하는 것이 목표이다.

목적 및 기대효과

“밥 먹으러 갑시다.”

“부담이 되더라.”

“죄송해요. 많이 힘들었죠?”



“밥 먹으러 가자.”

“부담이 되더라.”

“미안해. 많이 힘들었지?”

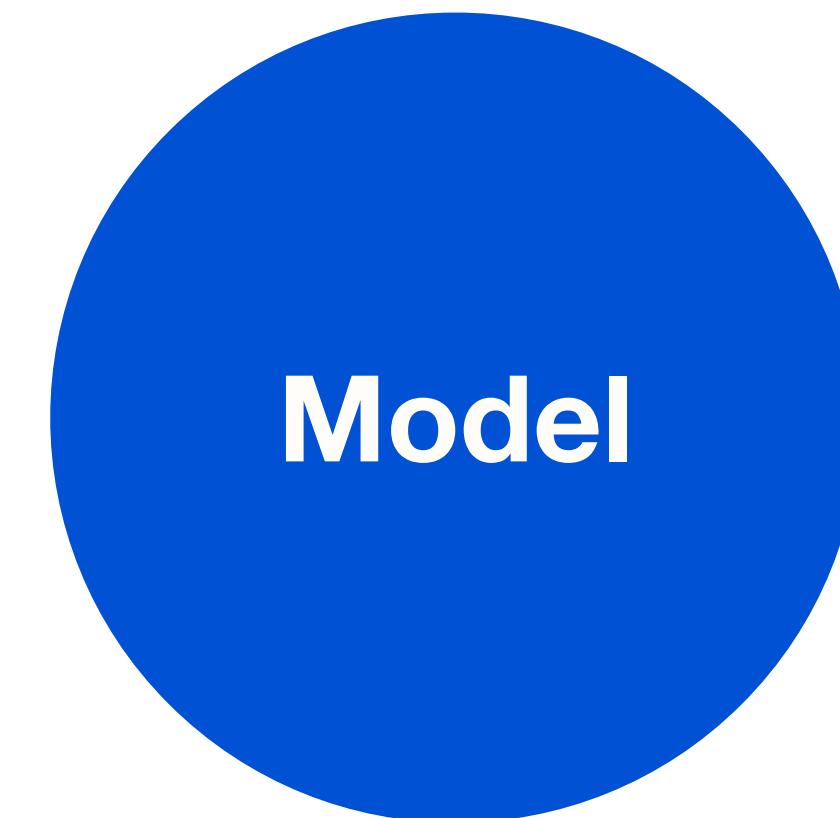
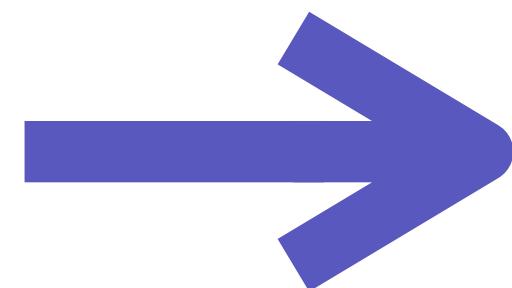
# 번역의 성능을 개선하는 것이 목표이다.

목적 및 기대효과

“밥 먹으러 가자.”

“부담이 되더라.”

“미안해. 많이 힘들었지?”



# 높임말 반말 변환하는 것이 성능을 개선한다고 가정

## 가정 및 구현

- 데이터를 반말 또는 높임말로 통일하기 때문에 번역상 생기는 문제를 해결하는데 기여할 수 있다고 생각
- 이때 데이터들을 변환하는 것을 `train`을 통해서 변환하는 것이 아닌 변환기를 구현하여 보다 가볍게 변환을 하는 것이 목표

# 어떻게 변화를 하였는가?

높임말 -> 반말 과정

바쁘고 힘들겠지만 즐거운 마음으로 돌아왔습니다.



'바쁘/VA', '고/EC', '힘들/VA', '겠/EP', '지만/EC',  
'즐거운/VA+ETM', '마음/NNG', '으로/JKB', '돌아왔/vv+EP', '습니다/EF'

# 어떻게 변화를 하였는가?

# 높임말 → 반말 과정

'바쁘/VA', '고/EC', '힘들/VA', '겠/EP', '지만/EC',  
'즐거운/VA+ETM', '마음/NNG', '으로/JKB', '돌아왔/VV+EP', '**습니다/EF**

# 어떻게 변화를 하였는가?

높임말 -> 반말 과정

'바쁘/VA', '고/EC', '힘들/VA', '겠/EP', '지만/EC',  
'즐거운/VA+ETM', '마음/NNG', '으로/JKB', '돌아왔/VV+EP', '**다/EF**'



바쁘고 힘들겠지만 즐거운 마음으로 돌아왔다.

# 어떻게 변화를 하였는가?

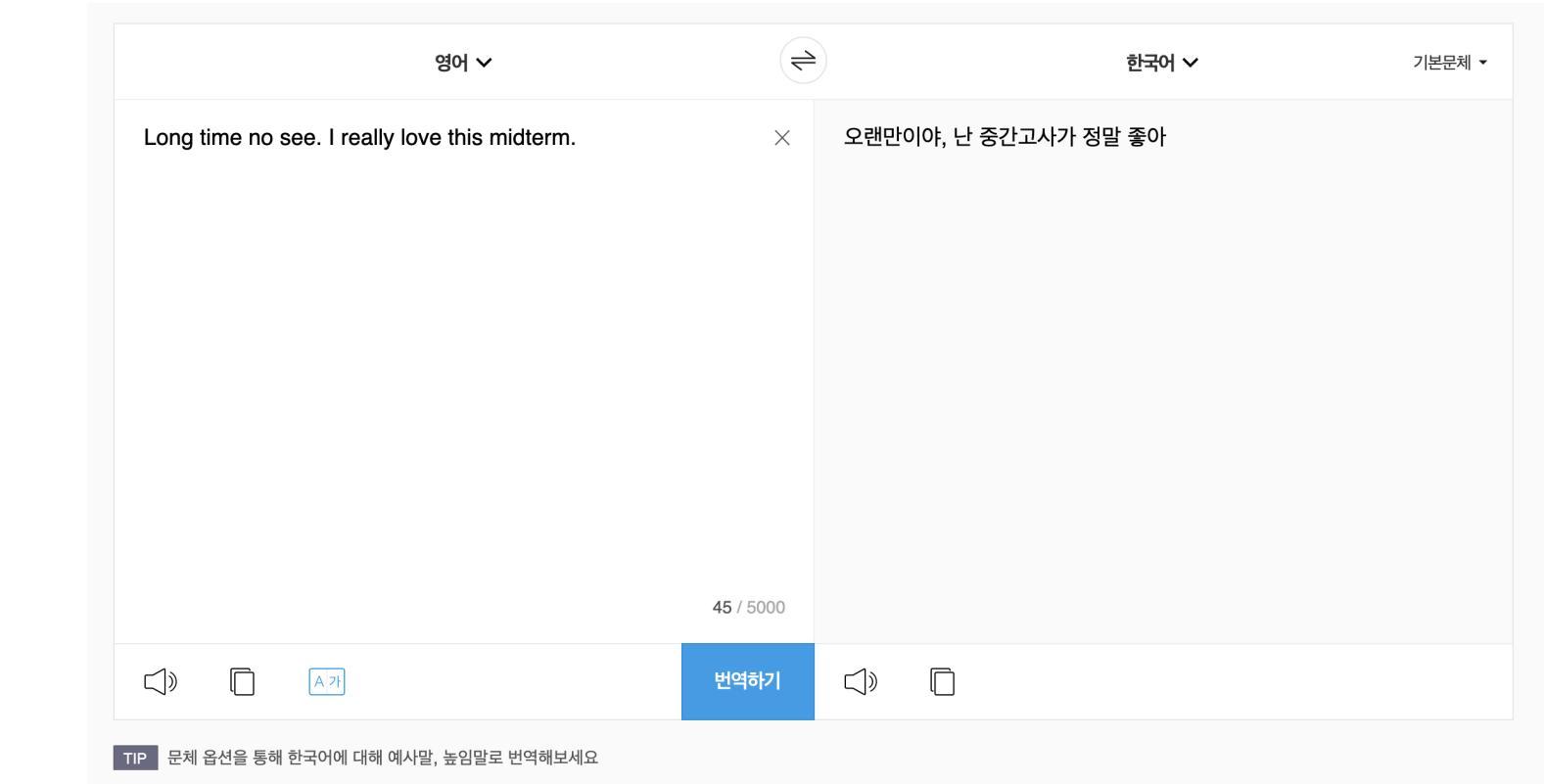
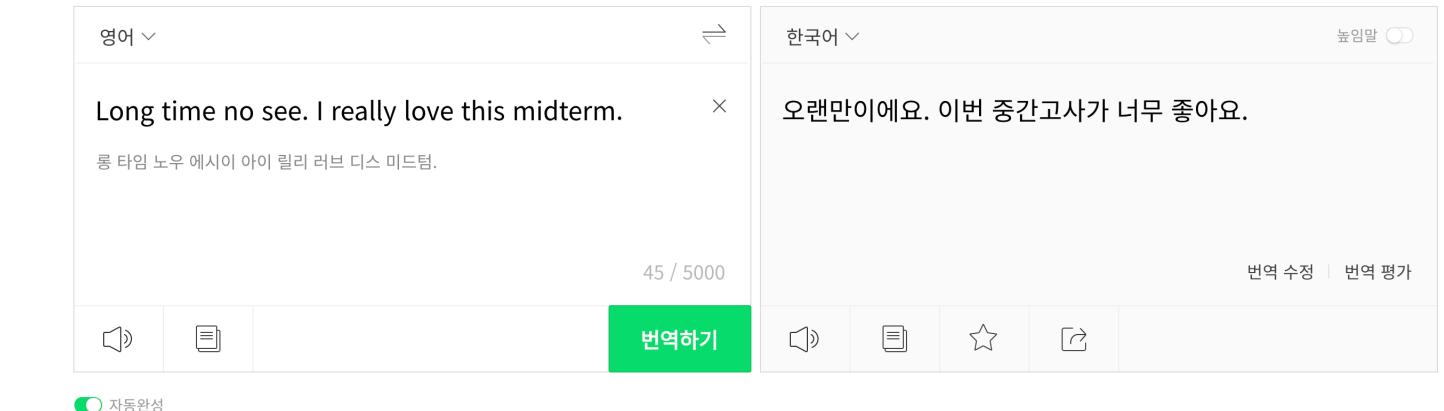
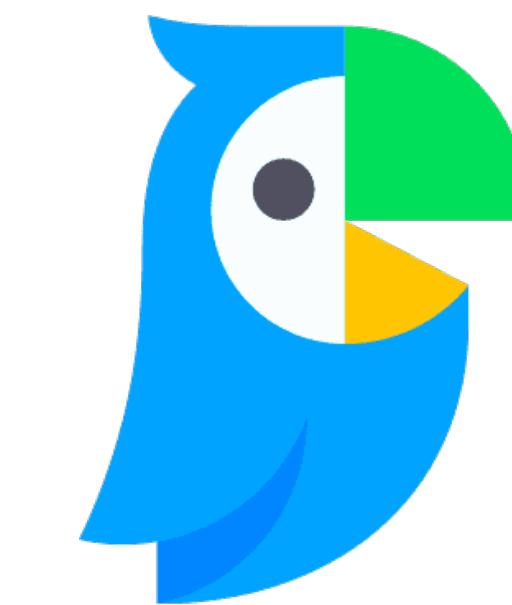
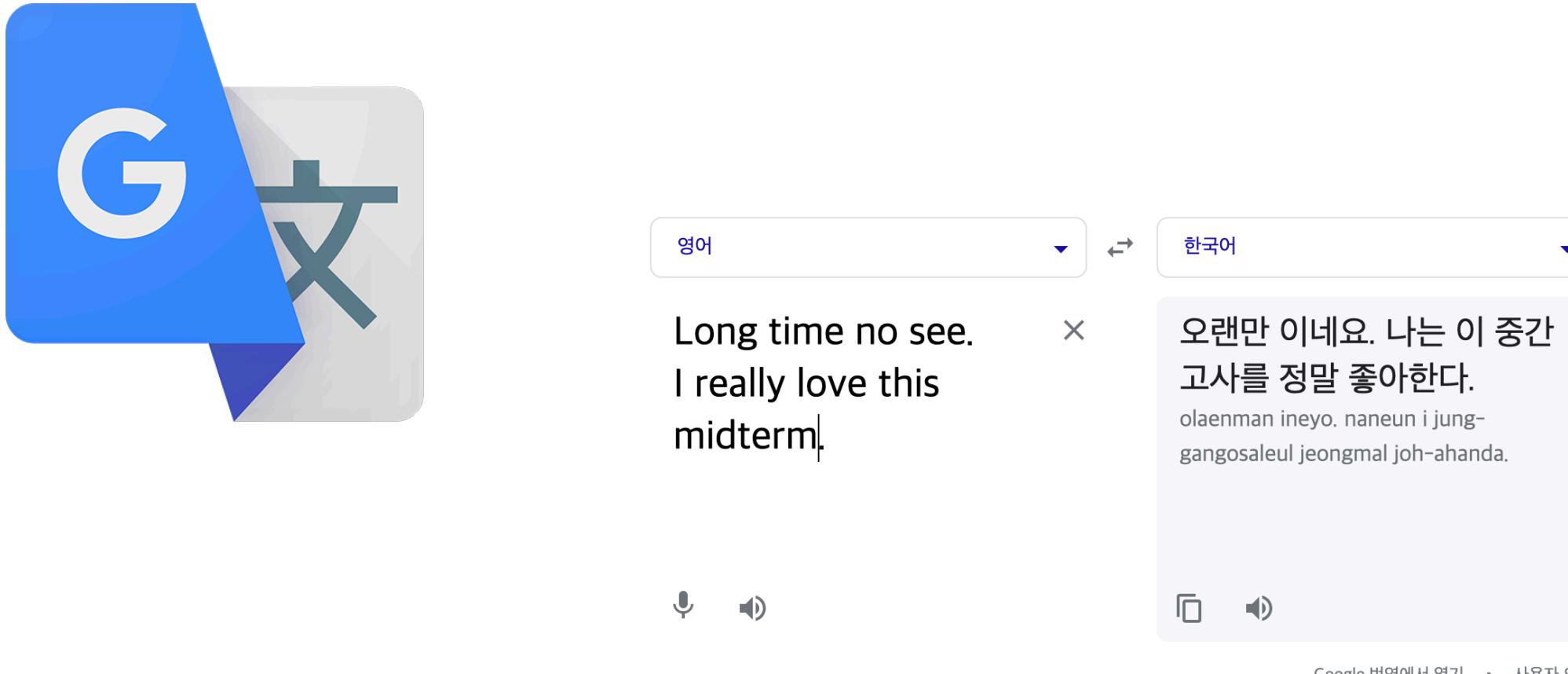
높임말 -> 반말 변화 예시

'바쁘/VA', '고/EC', '힘들/VA', '겠/EP', '지만/EC',  
'즐거운/VA+ETM', '마음/NNG', '으로/JKB', '돌아왔/VV+EP', '**다/EF**'

어떤 형태소인지, 또한 앞에 나온 형태소가 무엇인지에 따라 사전을 다르게  
구현해서 해당하는 형태소가 바뀌도록 구현하였다.

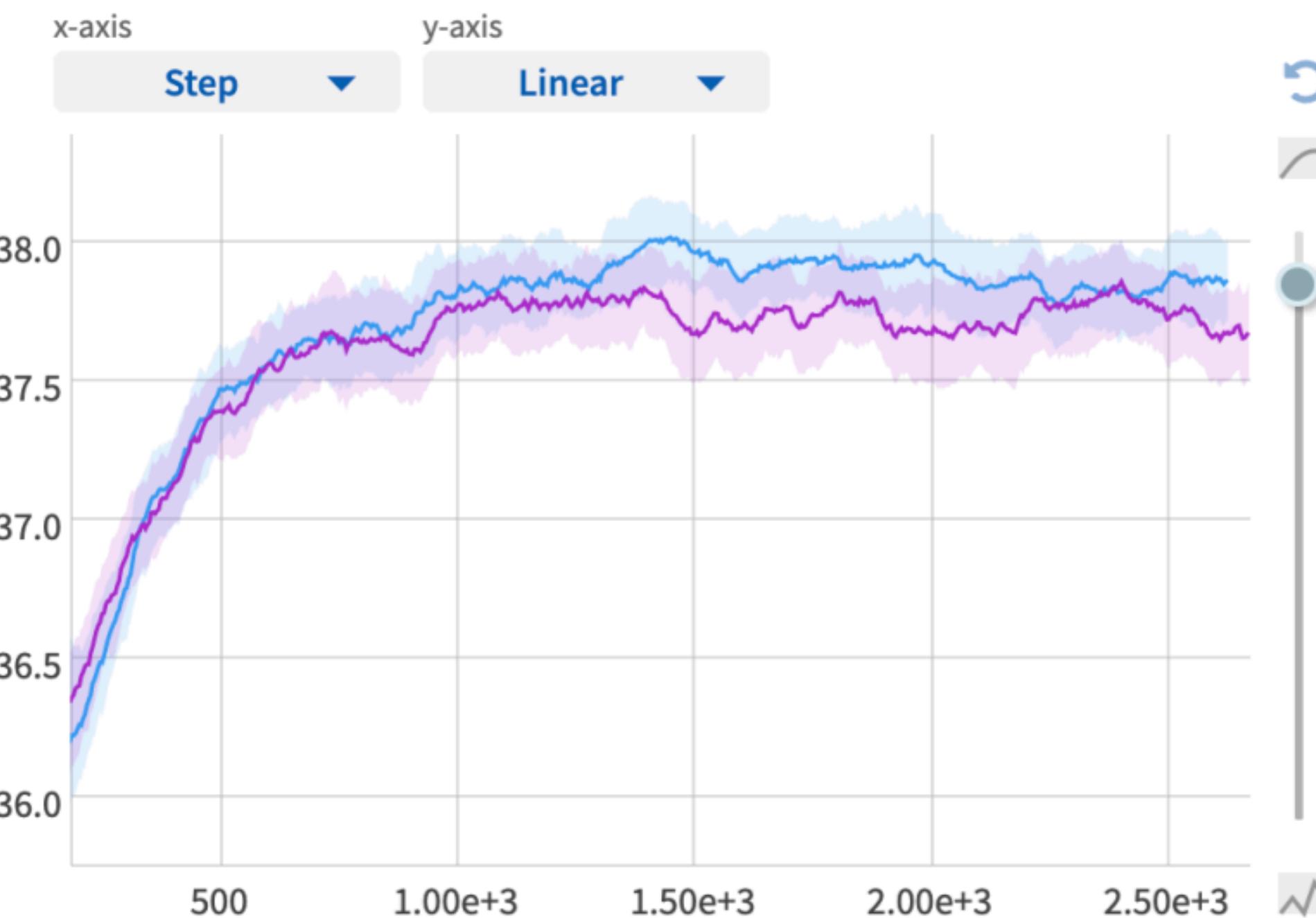
# 현존하는 번역기와의 차별화

## Google Translate, Papago, Kakao i 번역



# 높임말->반말 데이터로 Train을 해봤다.

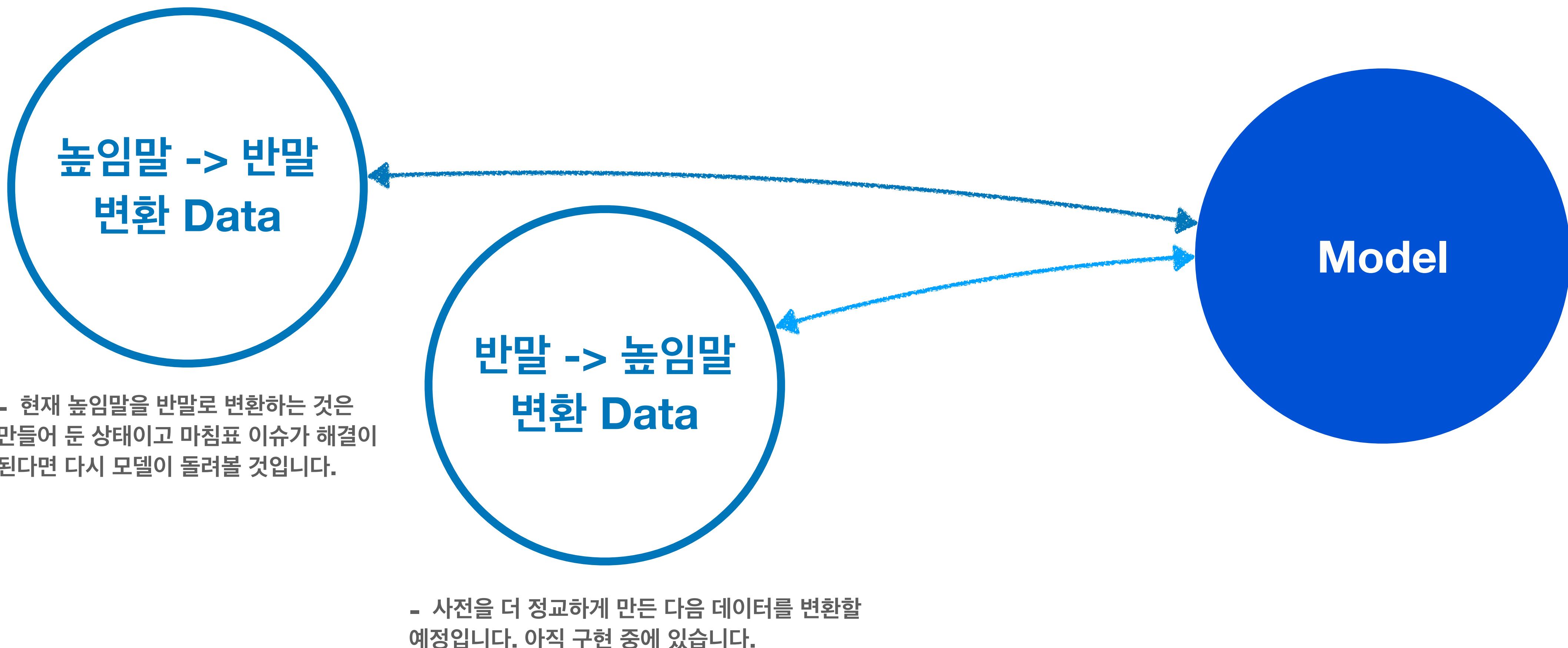
반말로 이루어진 Data로 Train



현재 Training하고 있는 데이터는 마침표 문제를  
해결하지 않은 데이터이다.

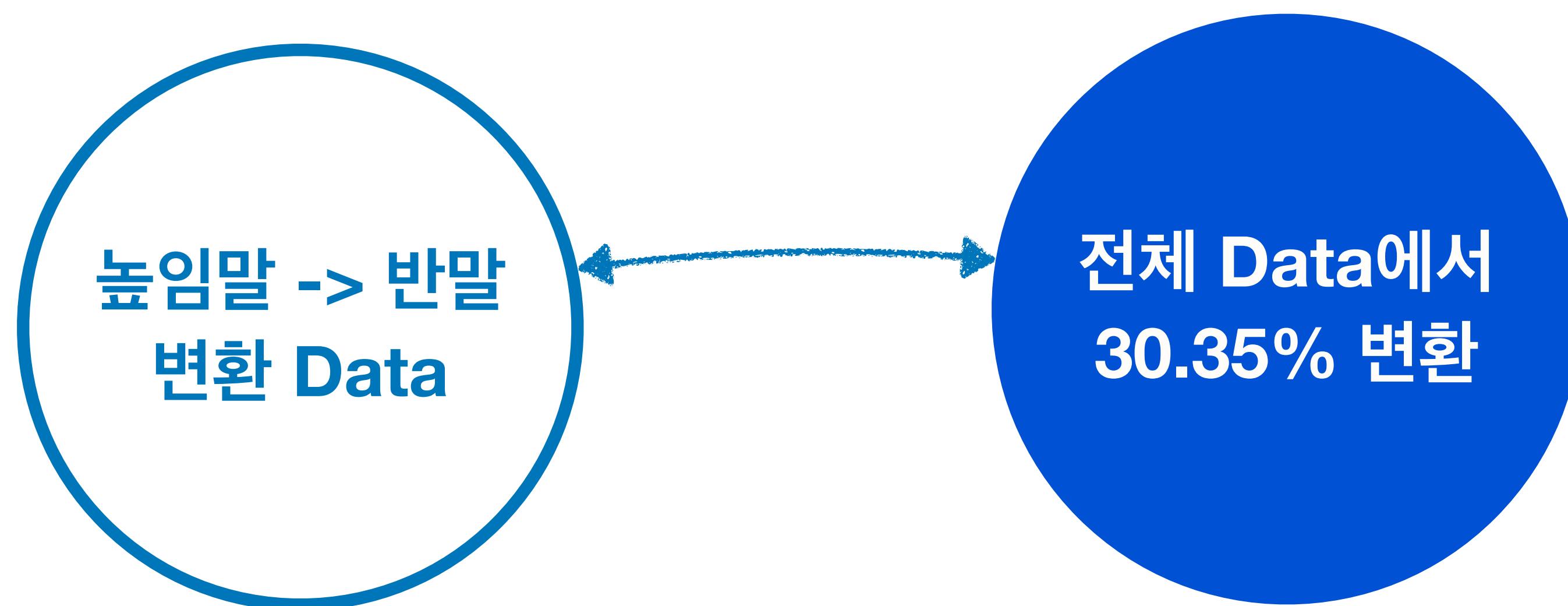
# 앞으로의 개선 방향 및 질문

## 계획 및 자문



# 앞으로의 개선 방향 및 질문

## 계획 및 자문



# 앞으로의 개선 방향 및 질문

## 계획 및 자문

보고된 상황을 사안의 중요도에 따라 계통을 통하여 보고(통보)한후 종결시까지 관리

법 제3조제1호나목에 해당되는 사회재난: 해당 재난대응 업무를 총괄·수행하는 국·소장

서울특별시 노원구립 체육시설 설치 및 운영에 관한 조례

맛에 신경 쓰기보다는 홍보를 더 많이 해서 그럴 거야,

주택재개발, 시장정비사업 등으로 전통시장 또는 전통상점가로서의 기능을 상실하여 협의회의 지정 취소 요청이 있을 경우

장기불참 등의 사유로 위원회의 직무를 수행하기가 부적당하다고 판단 될 경우

그 밖에 노인의 자립기반 조성 및 육성에 관한사항

고발 : 감사 결과 범죄 혐의가 있다고 인정되는 경우

그 밖에 구청장이 보호가 필요하다고 인정하는 아동

등록, 재등록 또는 폐기공인의 공인명 및 인영

기업의 국내·외 시장 판로개척을 위한 시장개척단 및 전문전시회 등

그 밖에 품위손상 등으로 그 직무수행에 부적당하다고 인정될 때

서울특별시 노원구 사회복지사 등의 처우 및 지위 향상을 위한 조례

사용자가 사용 예정일 9일 전부터 사용개시 7일 전까지 사용을 취소하는 경우 : 사용료의 100분의 80

학생의 진로 탐색·설계에 필요한 직업체험, 진로캠프, 진로특강 사업

서비스 제공 공무원이 착오·과실로 인하여 고객이 동일 건에 대하여 동일한 행정기관을 2회 이상 방문한 경우

이 마을은 임진강 수계의 상류지역으로 천혜의 수자원을 가진 살기좋은 마을로 신기한 버섯재배장 체험을 비롯한 과수원체험, 쑥떡만들기, 산림욕하기 등도 체험해 볼 수 있다