공학 프로젝트 기획
번역기 성능 향상

자모 단위 변환 & 높임말, 낮춤말 변환

허재무, 김준태, 김주환, 김정희 2021.11.16(금)

Optimizer 변경

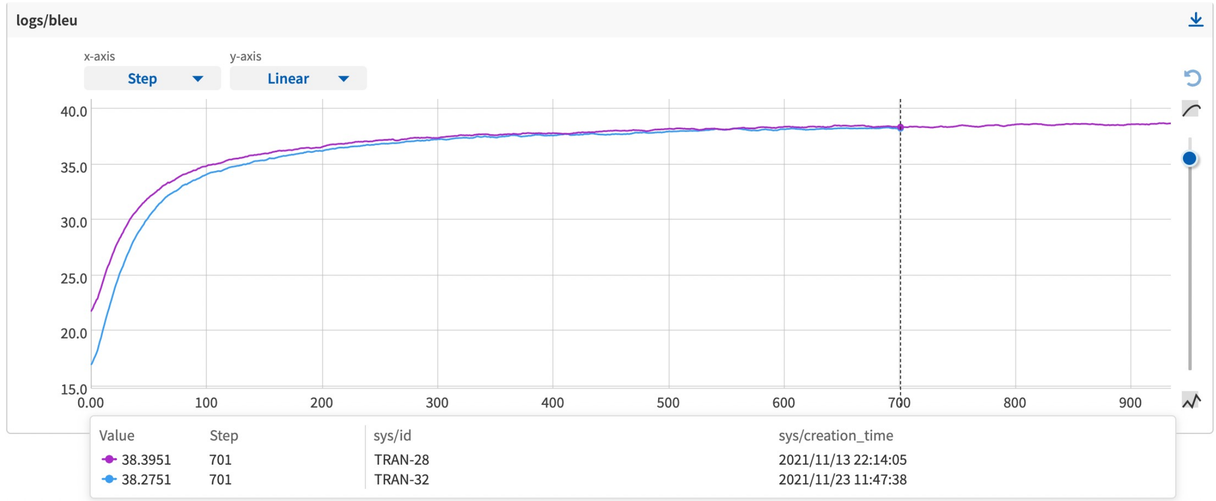
Optimizer 변경

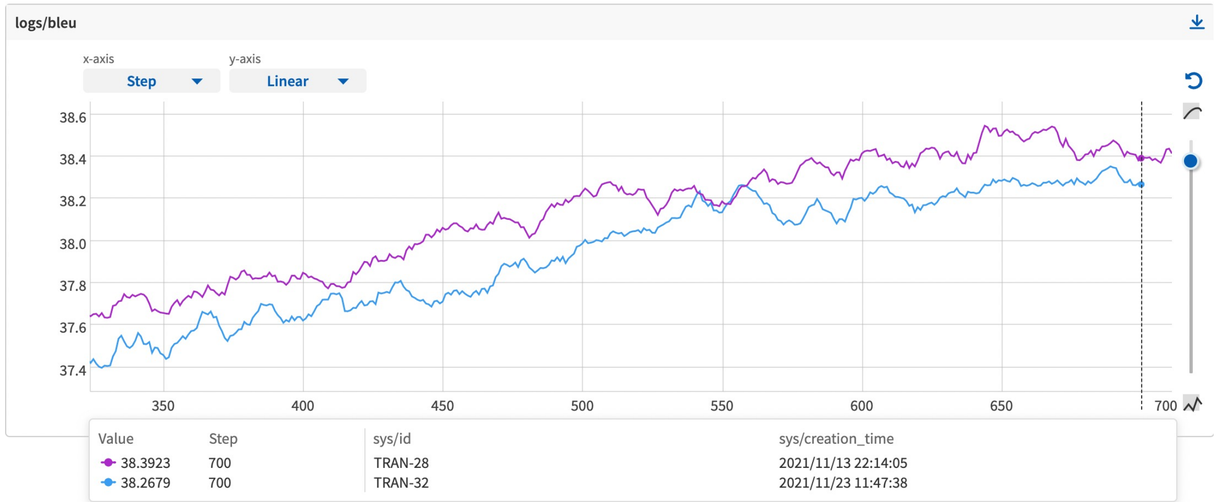
-optimizer(DiffGrad)

보라 - Adam

파랑 - DiffGrad

Adam이 성능이 더 좋은 것으로 보이나 조금 더 관찰할 필요가 있다고 생각이 됩니다.





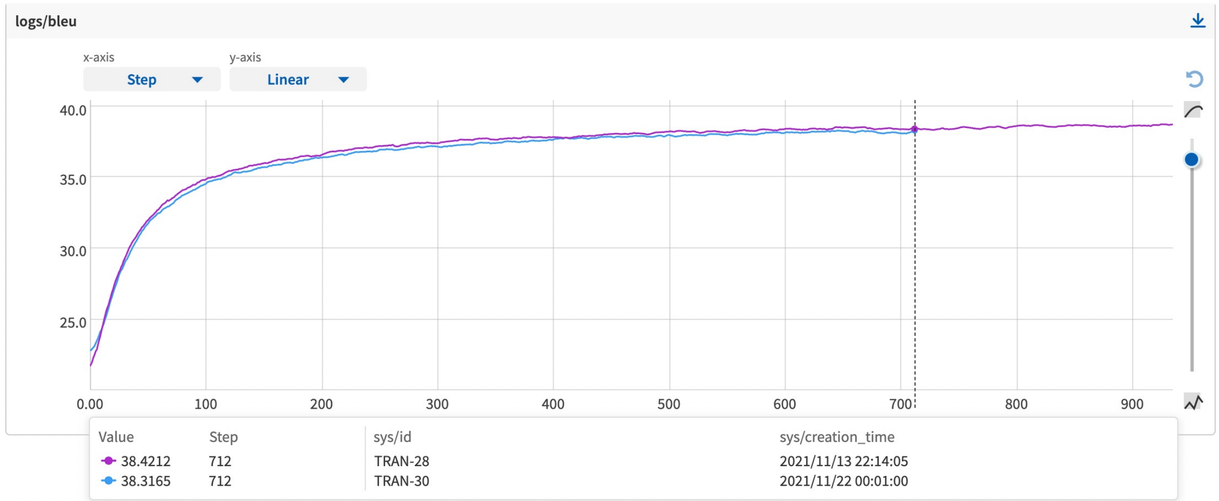
Optimizer 변경

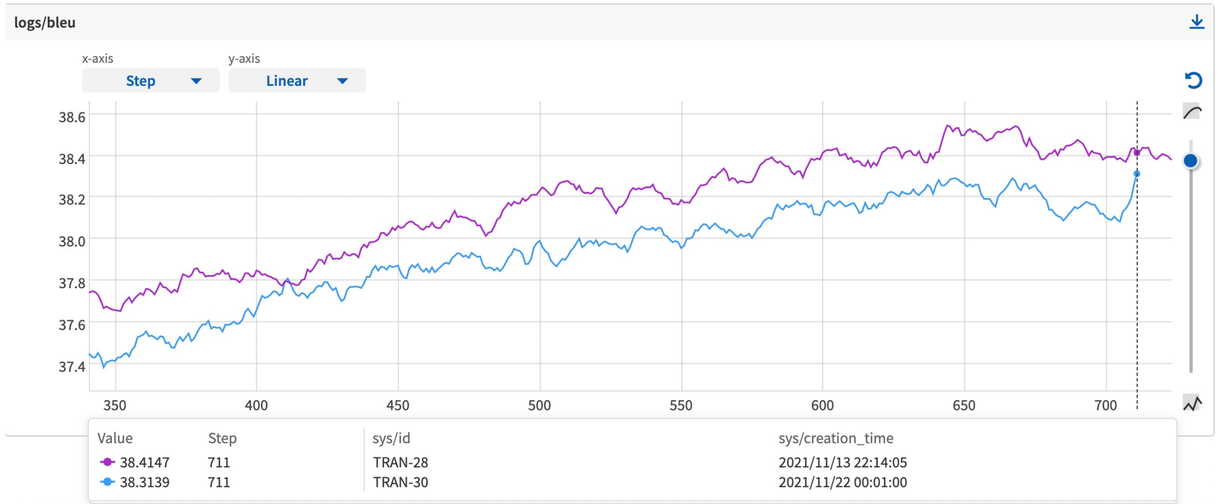
-optimizer(AdamP)

보라 - Adam

파랑 - AdamP

Adam이 성능이 더 좋은 것으로 보이나 조금 더 관찰할 필요가 있다고 생각이 됩니다.





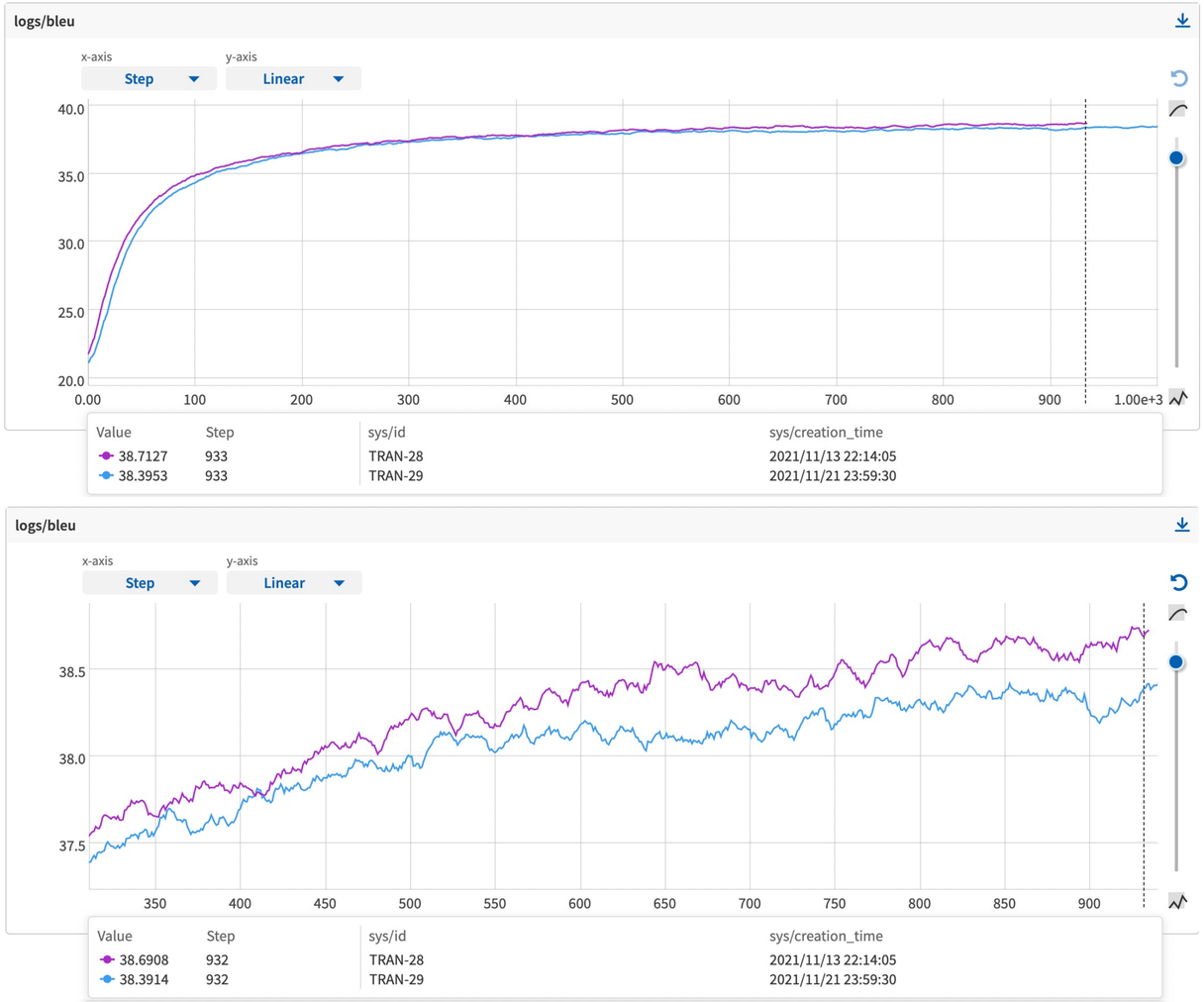
Optimizer 변경

-optimizer(AdaBelief)

보라 - Adam

파랑 - AdaBelief

Adam이 성능이 더 좋은 것으로
보이며 종료하였습니다.



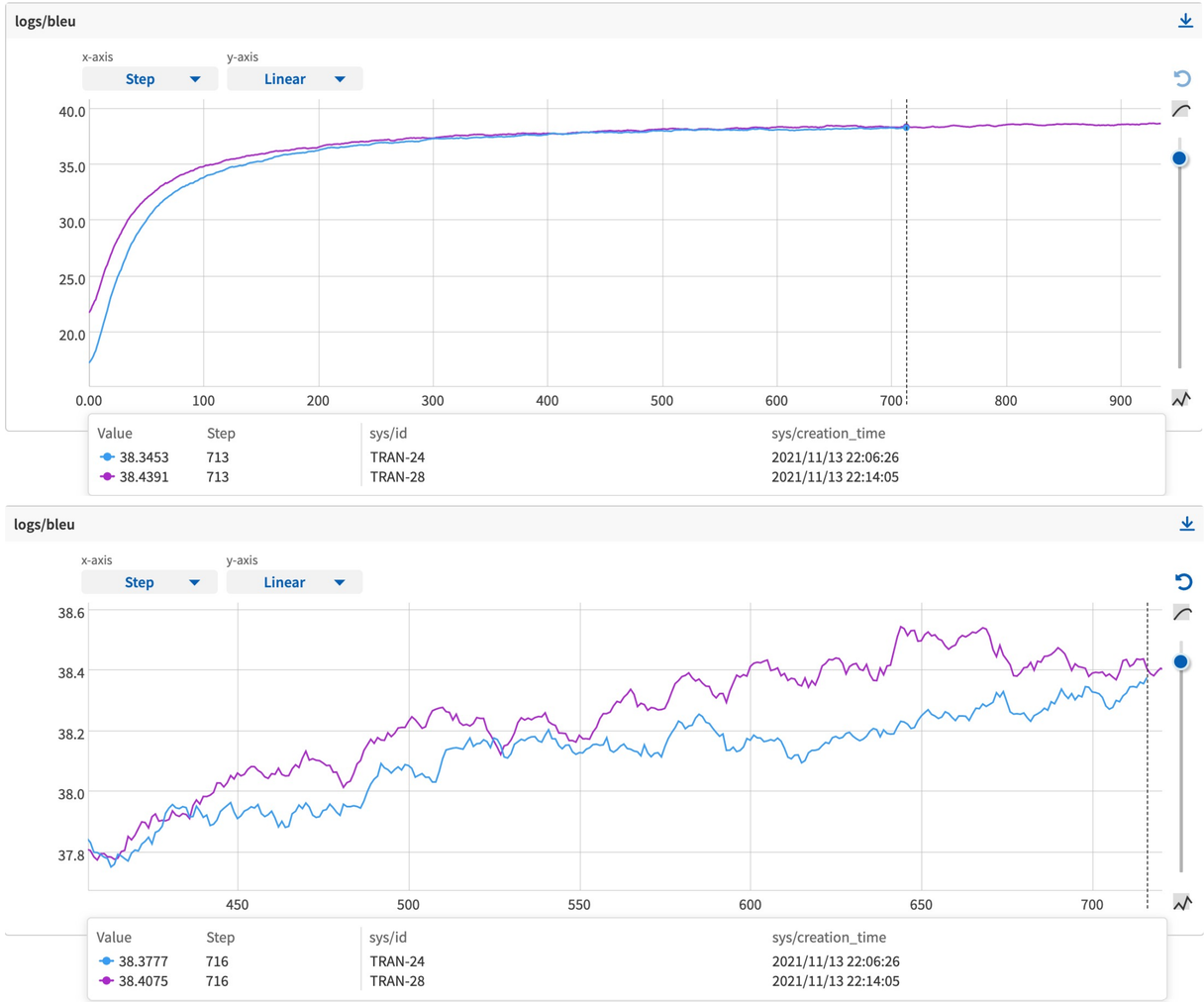
Optimizer 변경

-optimizer(Angulargrad)

보라 - Adam

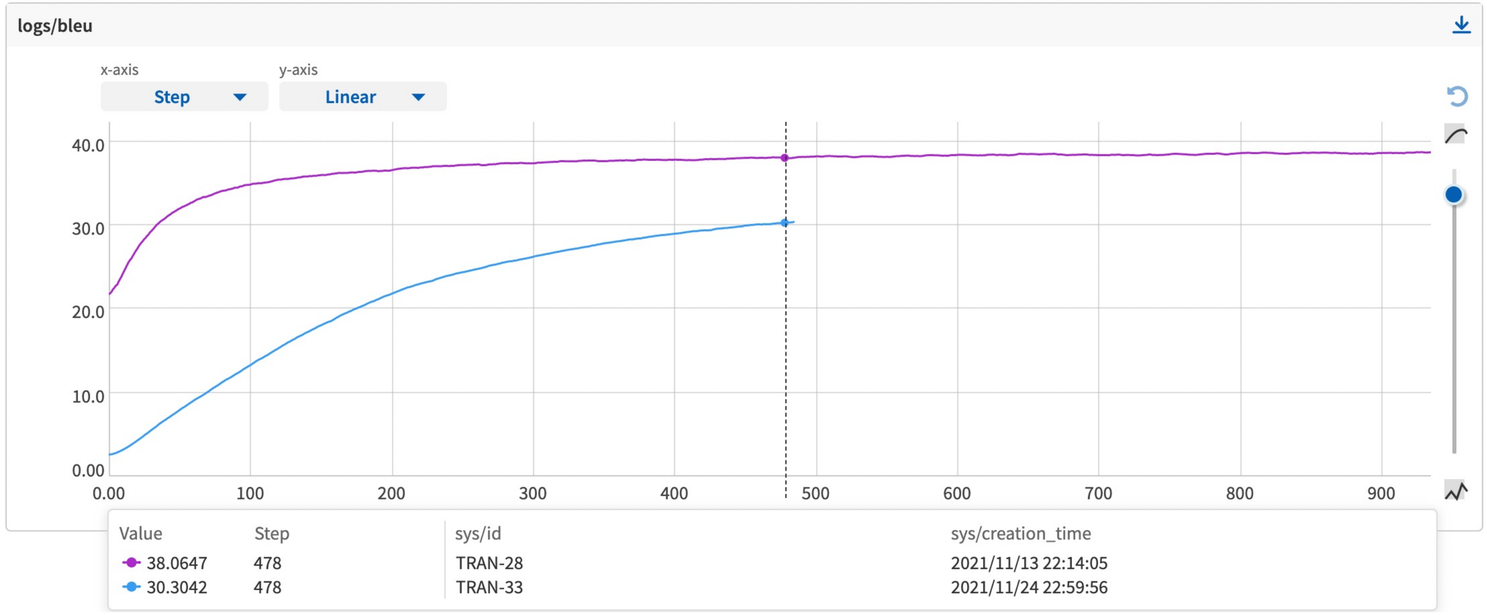
파랑 - Angulargrad

성능 상으로는 비슷한 것으로 보이나 Adam은 8일만에 해당 bleu score가 나온 반면 Angulargrad는 16일 만에 Adam과 비슷한 bleu score에 도달하는 것을 볼 수 있습니다.
고로 Adam을 사용하는 것이 더 좋다고 판단이 됩니다.



Optimizer 변경

-optimizer(Lamb)



보라 Adam 파랑 - Lamb

Adam이 성능이 더 좋은 것으로
보이며 종료하였습니다.

높임말 반말 변환

데이터 전처리

지난 미팅 후 개선 사항

- 수정하고 추가해야 했던 부분들

1.코퍼스 검사 후 나온 에러 문장 처리

2. 반말-> 높임말 변환의 개선

3. 영어의 격식 표현 처리 이슈

에러 문장 처리

지난 변환기의 문제점

지난 미팅 후, 데이터에서 발견한 오류가 있는 문장의 예시는 다음과 같음

기존의 문장	변환된 오류가 있는 문장
이것 봐, 우리 대원들이 입고 있는 방탄복이 이 녀석의 도움을 받았다는 걸 아나?	이것 봐요, 우리 대원들이 입고 있는 방탄복이 이 녀석의 도움을 받았다는 걸 아나세요?
가끔 시간이 엉뚱하게 바뀌었다가 다시 돌아와.	가끔 시간이 엉뚱하게 바뀌었다가 다시 돌아요와.
집에서 한 끼 간편하게 해결하려는 수요가 계속 늘고 있어서다.	집에서 한 끼 간편하게 해결하려는 수요가 계속 늘고 있어서습니다.
왜 우리가 자꾸 말씀드리냐면 경기도는 거리가 너무 멀다.	왜 우리가 자꾸 말씀드리냐면 경기도는 거리가 너무 멀ㅂ니다.

반말 -> 높임말 개선사항

- 문장의 예외적인 경우 처리 및 사전 개선

예외처리 사전을 세분화하였으며, 규칙에 따라 예외사전을 다르게 참고하도록 구현함

```
EXC_4_deal_2 = [

    ['ㄴㅣ',['ㅆ','ㅇㅓㅇㅛ'],'ㅇㅛ']
]

EXC_4_deal_3 = [

    ['ㄴㅡㄴㄱㅏ',['ㅗ','ㅗ','ㅏㅇㅛ'],['ㅜ','ㅜ','ㅓㅇㅛ'],['ㅏ','ㅏ','ㅣㅇㅛ'],'ㅇㅓㅇㅛ'],
    ['ㄷㅏ',['ㅆ','ㅆ','ㅅㅡㅂㄴㅣㄷㅏ'],['ㄴ','ㅂ','ㄴㅣㄷㅏ'],['ㄹ','ㅂ','ㄴㅣㄷㅏ'],'ㅂㄴㅣㄷㅏ']
]

EXC_4_deal_4 = [

    ['ㄱㅔ','ㄱㅔㅇㅛ','ㄱㅔ'],
    ['ㄷㅔ','ㄷㅔㅇㅛ','ㄷㅔ'],
    ['ㄱㅓㄹ','ㄱㅓㄹㅇㅛ','ㄱㅓㄹ']
]
```

~다

앞 글자 종성에 'ㅆ'이 있다면 '습니다'로 변경

시험을 봤다 => 시험을 봤습니다.

앞 글자 종성에 'ㄴ'이 있다면 'ㅂ니다'로 변경

도자기를 만든다. => 도자기를 만듭니다.

•
•
•

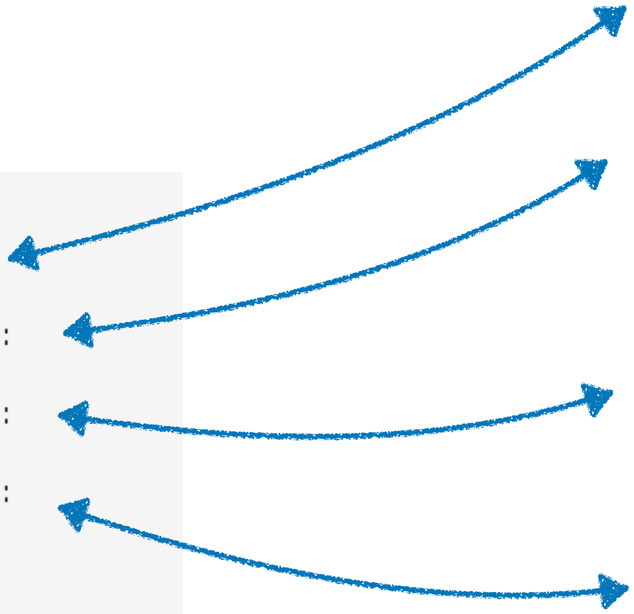
형태의 변화가 많은 형태소에 대해서는 예외적인 경우를 두어서 처리

반말 -> 높임말 개선사항

- 문장의 예외적인 경우 처리 및 사전 개선

예외사항에 대한 구별

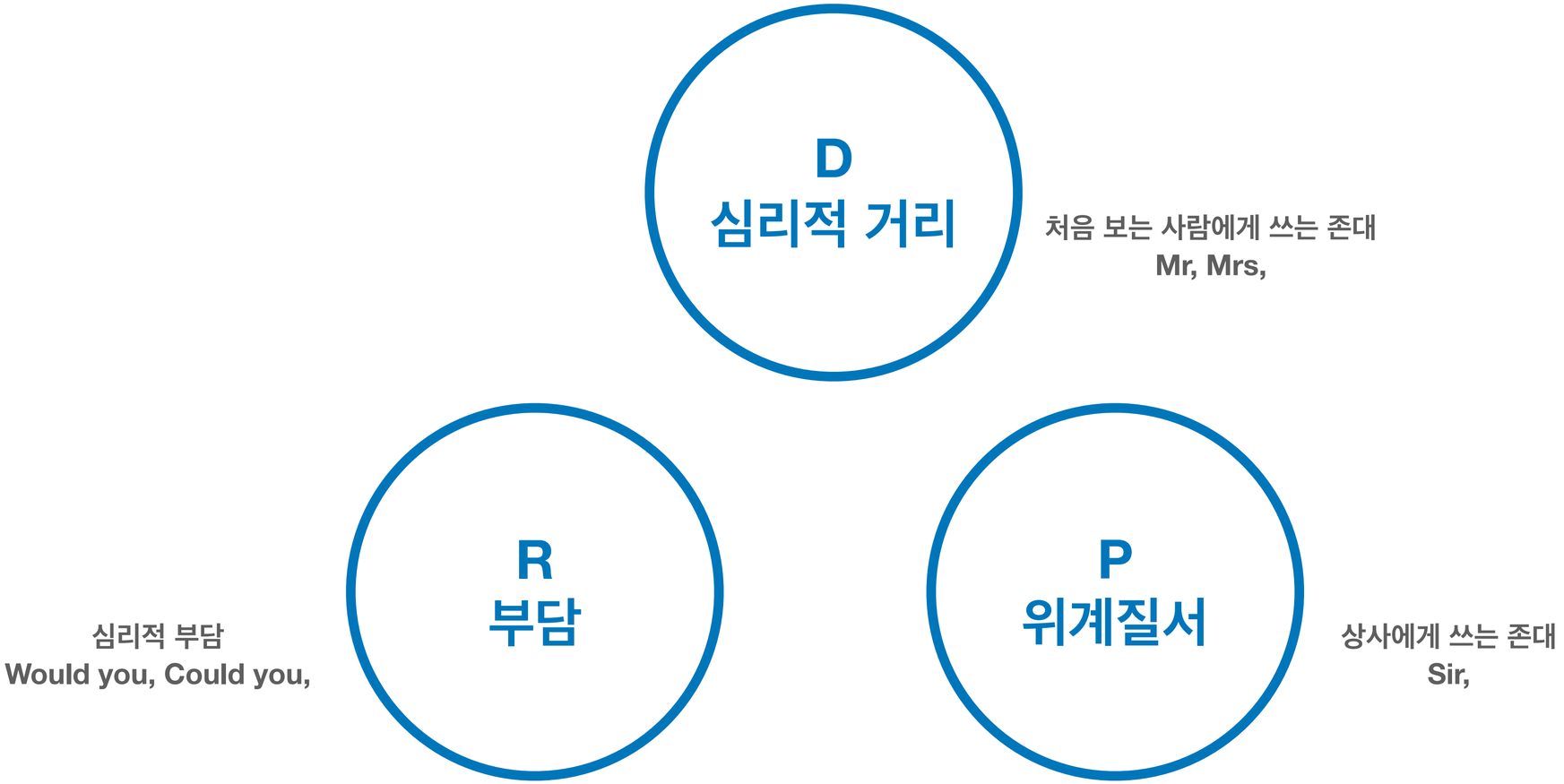
```
def isExcept(self, input):
    if input[1]=='special':
        return 1
    elif input[1]=='special-':
        return 2
    elif input[1]=='-special':
        return 3
    elif input[1]=='specialx':
        return 4
    else:
        return 0
```



1. **동사 파생 접미사와 종결어미가 오는 경우 종성에 ‘ㄴ’이 오면 ㅂ니다로, 아니면 ‘습니다’로 변경**
   Ex) 해본다 -> 해봅니다.
2. **앞에오는 단어을 신경써야하지만, 같은 형태소 안에서 해결이 되는 경우**
   했니? -> 했어요?(‘ㅆ’이 오는 경우 ‘-어요’를 붙인다.)
   자니? -> 자요?(종성에 아무것도 오지 않으면 ‘-요’를 붙임.)
3. **앞에 오는 것을 신경써야하며 하나의 형태소 안에서 해결이 안 되는 경우**
   ‘-다’의 경우 앞 글자 또는 종성에 따라 변화가 다양하기에 이를 고려하여 변환해야 함.
4. **같은 형태소의 같은 단어이지만, 문장의 특정 위치에서만 변환이 이루어져야 하는 경우**
   ‘-네.’, ‘-걸’ 같은 경우는 형태소 분석기에서 종결어미로 분류가 되지 않고 연결어미로 분류가 됨. 따라서 문장의 끝에 올 경우에만 변환을 하고 문장 중간에서 연결어미로 오는 경우는 변환하면 안되기에 예외사전을 만들었음.

영어 격식 표현을 고려한 변환

- 영어 격식 표현의 유형



영어 격식 표현을 고려한 변환

- 영어 변환 시도

이 중 위계질서, 심리적 거리 같은 경우 한국어의 존댓말 사용에서도 나타나는 것이기에(Mr. Mrs. Sir. 등등 명사에 붙는 형태로 나타남.) 이는 단순 치환으로 구현이 가능하다 판단했음.

```
convert_low_en('Mr. Kim invented gibbson\'s method')
```

```
" Kim invented gibbson's method"
```

허나, 부담의 경우, 영어에서는 would, could 등의 용언(변형의 범위 큼)의 형태로 나타나기 때문에 이를 어떻게 처리 할지에 대한 자문위원에게 문의를 했음.

영어 격식 표현을 고려한 변환

- 영어 변환 시도

자문위원

자문: Would, could 처럼 비교적 이에 대응하는 것을 찾기 쉬운 경우, 그냥 치환을 진행하는 추천하심. 변형이 심한 경우, 한번 코퍼스 상에서 그런 표현이 들어간 말들이 한국어 데이터에서 얼마나 많이 존댓말의 형태와 연결되는지를 파악해야 할 것 같다고 하셨음.

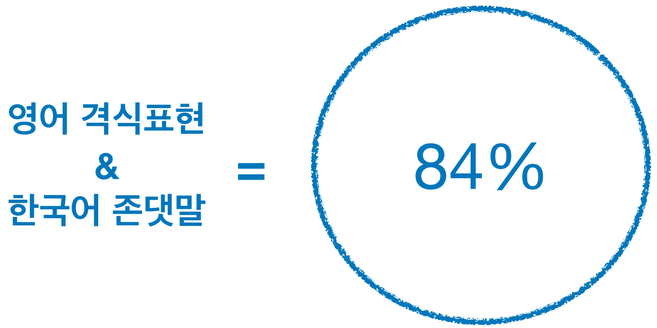
영어 격식 표현을 고려한 변환

- 영어 변환 시도

```
f0 = open('./hgu_clean.kr.shuf', 'r', encoding = 'utf-8')
f1 = open('./hgu_clean.en.shuf', 'r', encoding = 'utf-8')

total = 0
num = 0

while True:
    flag = 0
    line_kr = f0.readline()
    line_en = f1.readline()
    if not line_kr: break
    if line_kr[-1]=='₩n':
        line_kr = line_kr.replace('₩n','')
    if line_en[-1]=='₩n':
        line_en = line_en.replace('₩n','')
    if find_high_en(line_en)==1:
        total = total+1
        if find_high_kr(line_kr)==1:
            num = num+1

f0.close()
f1.close()

print(num/total*100)
```

```
84.82587064676616
```

그래서 간단한 코드를 만들어 이를 확인한 결과 84%라는 수치가 나왔음. 이는 일단 Mr. Mrs. 등의 경우도 고려 하였지만, 그럼에도 불구하고 수치가 많이 큰 것 같아 이 부분을 처리해야하는 필요성을 느꼈음.



영어 격식 표현을 고려한 변환

- 영어 격식 표현에 대한 논문

참고 논문

예를 들어, 'Mr' 'Mrs' 'Miss', 'James', 'Jim', 'Jimmy'와 같이 구분적 호칭이 있는 영어를 사용하는 사람은 의사소통에서 상대방과의 높고 낮음의 거리나 친소의 거리를, 이에 따라 구분하고 인식할 것이다. 연령이나 지위 등에 의한 존대 체계를 가진 한국어를 사용하는 사람은 연령이나 지위의 작은 차이라도 상대방과의 거리로 인식하고 구분하며, 이를 존대 체계에 따라 표현한다.

이원국 영어 공손의 거리 : 근접과 격원, 새한영어영문학회, 2005, 149 - 179

https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01057228