의도적인 표기 변형이 이루어진 문장을 표준어로 교정하는 모델 구축

모라구해써?
2016100440 경영학과 홍서영
2017103309 생물학과 심미단

Contents

과제 개요

Needs

- 일상적인 채팅에서의 구어성이 두드러진 언어 사용
- e.g.) **밥은 먹었어?**라는 문장을 채팅에서 **밥은 머거써?** 또는 **밥은 먹어써?** 등으로 표기하는 것
- 이러한 언어 습관은 한국어 채팅 데이터 분석을 어렵게 한다고 판단
- 다양한 자동 문법 교정 프로그램에서도 이는 완벽히 교정되지 않고 있음
- 실제로 Google 대체 단어 suggestion에서 모해?는 뭐해?로 제안하지만 밥은 머금?은 대체어를 제안하지 못함
- 오탈자가 아니라 사용자가 의도적으로 표기 오류를 낸 것, OCR 연구와 오탈자 교정 과는 결이 다르므로 연구가 필요

Goals

의도적인 표기 변형이 이루어진 채팅체 문장을 표준어로 교정

과제 수행 과정

1) 데이터셋 구축 및 전처리

- 연구자 2인이 지인들과 주고받은 카카오톡 채팅 데이터를 추출해 사용
- 연구에 쓰인 채팅체 데이터는 **두 연구자의 카톡 데이터에서 공통으로 나온 단어**만 추출한 데이터
- 채팅체에 최적화된 띄어쓰기 framework(chatspace)를 활용해 문장을 단어로 분리
- 분리된 단어 중 표기가 올바르게 된 단어와 올바르게 되지 않은 단어를 분류했다.
- 전체 21923개의 단어 중 **2432개**가 의도적인 표기 오류를 포함한 case로 분류되었다.
    - e.g.) 그래써, 모해, 월욜, 했넴, 잤넹

과제 수행 과정

2) Edit Distance 도출

- 텍스트 유사도를 구하기 위한 척도로 **자모 분리 후의 Edit Distance**를 채택했다.
- 이는 오탈자 교정에 자주 사용되는 기법이다.
- 표기 오류가 있는 단어와 표기 오류가 없는 단어간의 텍스트 유사도를 구해 가장 유사한 표준어를 찾아 교정한다.
- 표기 오류가 있는 단어 전체에 대해, 구축한 구어체 데이터셋에서 수정 거리가 가까운 단어들을 리스트업했다.
- 한계점 발견
  - [모해] → [모래, 못해, 묘해, 오해, 토해, 뭐해]
  - 모두 자음 또는 모음 하나만 바뀌는 0.33333… 의 edit distance → 우선 순위 부여 불가
  - 변형 ‘패턴’을 학습하는 모델의 필요성

과제 수행 과정

3) 모델 학습

- 모델 학습에는 Seq2Seq with Attention 모델을 구현해 사용했다.
- 그리고, 표기 오류가 있는 단어를 교정해 라벨링했다.
- 더 나은 모델 학습을 위해 단어의 **초,중,종성을 분리**
- 종성이 없는 경우 **padding**을 넣었다.
- 적은 데이터를 보완하고, 일반화하기 위해 **n-gram** 기법을 적용해 10159개로 학습 데이터를 늘렸다.
- **최종 성능**
  - test accuracy 82.3% (400/486)
  - train accuracy 94.3% (8753/9284)
  - n-gram 제외 전체 데이터 89.3% (2171/2432)
  - n-gram 포함 전체 데이터 93.3% (9476/10159)

labelling

text	label
먹찌	먹지
왜웽	왜왜
자기얌	자기야
모해써	뭐했어
저기용	저기요
끄래	그래
대써	됐어
대지	돼지
에바여	에바야
샀냐구	샀냐고
귀용	귀엽
되어용	되어요
잤넹	잤네
라구	라고
머것냠	먹었냐
신기햄	신기해

초,중,종성 분리, padding, n-gram

ㅁㅗpㅎㅐp	ㅁㅝpㅎㅐㅆ
ㅗpㅎㅐpㅆ	ㅝpㅎㅐㅆㅇ
pㅎㅐpㅆㅓ	pㅎㅐㅆㅇㅓ
ㅎㅐpㅆㅓp	ㅎㅐㅆㅇㅓp

과제 수행 과정

4) 최종 변환 Algorithm 구축

- 두 방법으로 찾은 예측 교정어가 같은 경우가 2432개 중 1451개로, 약 59.7%정도 되었다.
- **모델과 edit distance의 결과가 다른 경우**에는 아래의 알고리즘을 적용했다.
    - edit distance output이 없는 경우: model output return
    - model output이 original text와 같은 경우: edit distance output return

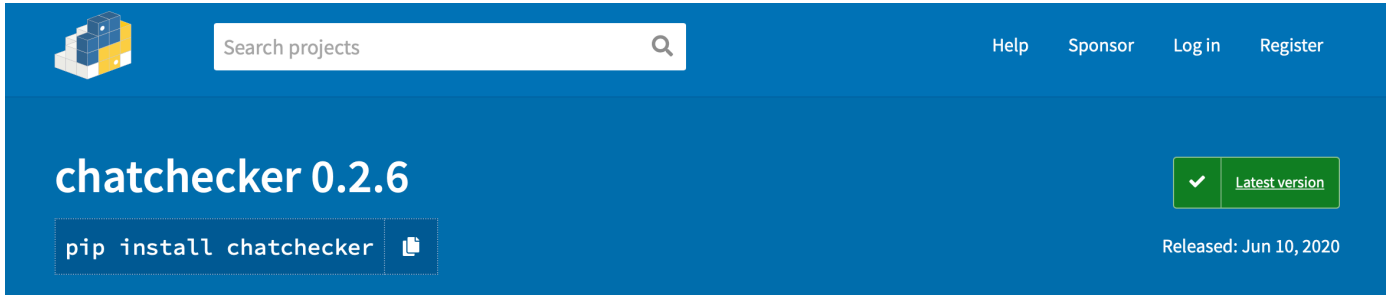
    (둘 다 결과가 있는데 다른 경우)

    - model output이 표준어 데이터셋에 있는 경우: model output return
    - model output이 표준어 데이터셋에 없는 경우: edit distance output return
        - e.g.) 찾아봐야자, 그랬, 교수님힘, 아버지가가, 신김해, 안들어가가서고

과제 수행 과정

5) 배포

작성한 코드 패키지화, **pip**를 활용해 배포



수행 결과물

https://github.com/seoyoungh/ko-chat-checker

Workflow

1. 입력으로 들어온 문장은 띄어쓰기 framework에 의해 단어로 나뉜다.
2. 각 단어들을 표기 오류가 없는 구어체 데이터셋과 대조 (약 13.7만개, '대화체' 데이터 직접 수집)
3. 데이터셋에 없는 단어를 교정이 필요한 단어로 분류한다.
4. 단어에 포함된 특수문자는 전처리를 위해 제거되지만, 단어 끝의 {,.!?} 문장 부호는 제거되지 않는다.
5. 교정이 필요한 단어에 대해 custom model 및 edit distance에 의해 교정된다.
6. 교정이 완료된 단어와 교정이 필요 없는 단어들은 다시 문장으로 조합되어 반환된다.
7. 이로써 의도적인 표기 오류에 대한 교정이 완료된 문장을 반환한다.
8. 입력으로 교정이 필요한 단어가 들어오는 경우에 대해서도 함수를 만들어 사용할 수 있도록 했다.

시연 영상

결론

활용 방안

가. 다른 한국어 채팅 데이터 분석에서의 전처리로 활용

나. 대화 음성 인식 분야에서 대화체가 반영된 텍스트 생성에 활용

다. 한국어를 배우는 외국인들이 채팅 및 온라인 게시글 해석에 활용

라. 사람처럼 말하는 챗봇 개발

의도적인 표기 오류가 포함된 채팅체 문장/단어를 Seq2Seq와 Edit Distance를 활용해 교정할 수 있었다.

하지만, 채팅체 사용에 따른 표기 오류 종류가 매우 다양하고, 데이터가 부족해 완벽히 교정하기엔 한계가 있었다.

유행 및 시대 변화에 따라 기존에 없던 새로운 채팅체는 계속 등장하고 있다.

채팅이 많이 이루어지는 현대 사회에서 채팅 데이터에 대한 연구 및 분석이 더 많이 이루어지길 제안한다.

감사합니다.