한국어 음성인식 오류 교정을 위한 N-Best 결과 기반 생성 모델

Generative Error Correction using N-Best Hypotheses for Automatic Speech Recognition of Korean

서민택 나승훈
전북대학교

나민수 최맹식 이충희
㈜엔씨소프트

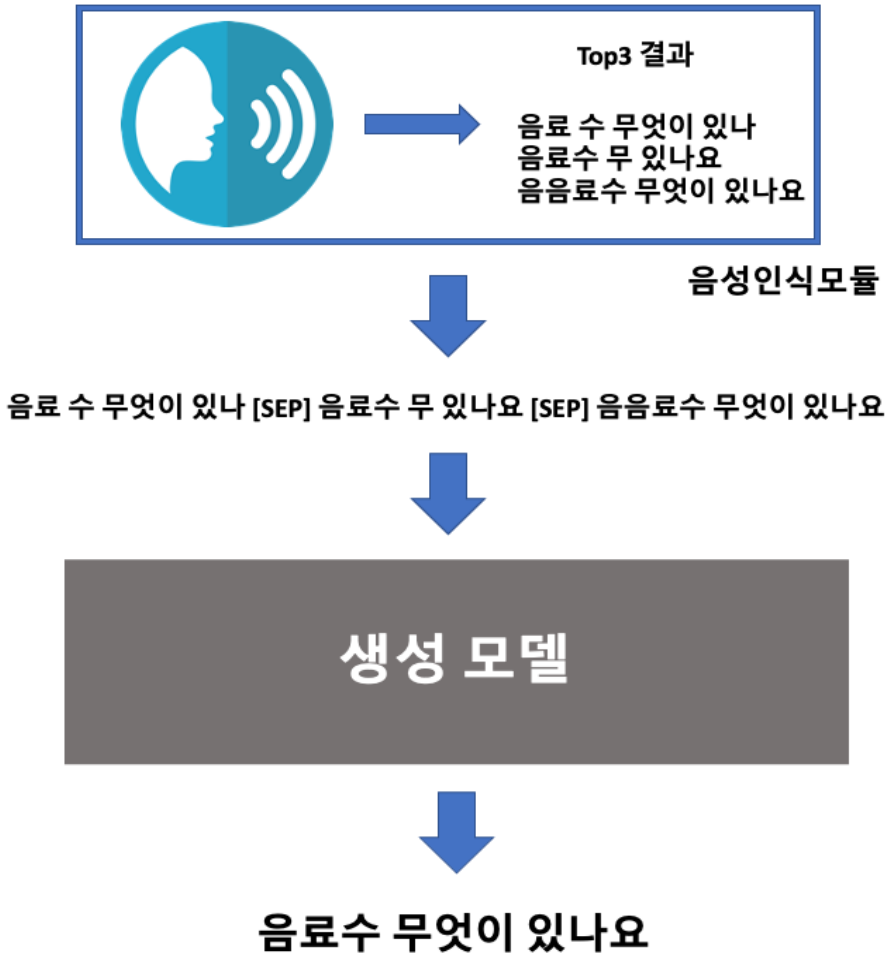
Introduction

- 최근 ASR 연구들은 어느정도 높은 수준에 도달하였으나, 아직 완벽한 서비스 제공에는 어려움이 존재.
- 일반적으로 ASR을 사용시 API를 활용하는 경우가 많아, 이에 독립적인 교정 방법 또한 중요
- 따라서 API에 독립적인 모듈을 통하여 ASR 모델에 종속되지 않고, 음성 인식 Quality를 높이고자 함

데이터 구축

- AIHUB 데이터를 통해서 대화 대본 말뭉치 확보
- TTS(Text-To-Speech)를 통해서 (대본, 음성) 쌍 획득
- STT(Speech-To-Text)를 통해서 (대본 , 음성 , Top N 결과) 쌍 획득

- N-Best 기반 생성 모델 오류 교정(Zhu et al., 2021)



- N-Best 기반 생성 모델 오류 교정(Zhu et al., 2021)

1 . Prefix에 민감한 T5대신 BART 구조 사용하여 학습

2 . Train과 Dev에는 좀 더 현실 반영을 위하여 ASR이 정답인 경우 필터 X

3 . Test는 수동으로 명사 오류만 선정

4 . 문장 부호 및 표기 차이(1월 -> 일월)는 무시함

명사 오류 예시
더치라떼는요
-> 덧칠할 때는요

5 . Top1 결과만 입력 시 맞춘 12개 중 Top 10의 결과를 입력할 경우 맞추지 못한 경우는 2개

Exact Match	ASR	Top1	Top10
Dev	55.9	62.6	**63.9**
Test	0	12	**19**

WER	ASR	Top1	Top10
Dev	0.21	0.18	**0.17**
Test	0.43	0.34	**0.32**

- 결과 분석

오류 : 북구청 -> 북극청

Top1의 경우 북구->북구청으로 추론 한 것으로 보임

Top10의 경우 10개 전부 북극청으로 나와 편향을 가지는 것 같음

원본	아 북구니까 **북구청**이죠
ASR결과	아 북구니까 **북극청**이죠 아 북그니까 **북극청**이죠 아 북끄니까 **북극청**이죠 아 북고니까 **북극청**이죠 다 북끄니까 **북극청**이죠 아 북구으니까 **북극청**이죠 아 북구가 **북극청**이죠 아 북끄으니까 **북극청**이죠 아 북끄가 **북극청**이죠 아 북구나니까 **북극청**이죠
Top1	아 북구니까 **북구청**이죠
Top10	아 북구니까 **북극청**이죠

- 결과 분석

오류 : 그란데 -> 그런데

Top1의 경우 문장 하나만으로 수정하지 못함

Top10의 경우 '그런데' 부분을 내부적으로 조합하여 그란데로 수정함.

원본	**그란데** 사이즈로 주세요
ASR결과	**그런데** 사이즈로 주세요 **그런데** 사이즈로주세요 **근데** 사이즈로 주세요 **그런데** 사이이즈로 주세요 **그** 사이즈로 주세요 **그런데데** 사이즈로 주세요 **그런데** 대 사이즈로 주세요 **그런** 사이즈로 주세요 **그런데** 데 사이즈로 주세요 **그런데란** 사이즈로 주세요
Top1	**그런데** 사이즈로 주세요
Top10	**그란데** 사이즈로 주세요

- 추가 예시


```
"DOMAIN": "민원",
"NUMBER": "7536",
"src": "아 도장이라면 인감도장이요?",
"hyp": [
  "다 도장이라면 인간도장이요",
  "아 도장이라면 인간도장이요",
  "다 도장이라면 인간 도장이요",
  "다 도장이라면 인간도장이여",
  "아 도장이라면 인간도장이여",
  "도장이라면 인간도장이요",
  "아도장이라면 인간도장이요",
  "다 도장이라면 인간도장이오",
  "다도장이라면 인간도장이요",
  "도장이라면 인간도장이요"
],
"pred": "아 도장이라면 인감도장이요",
```

```
"DOMAIN": "의류",
"NUMBER": "2394",
"src": "돼지나 부엉이를 가방이나 끈에 다는 건 어떠신가요?",
"hyp": [
  "돼지나 부엉이를 가방이나 꺼내 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 끄에 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 꺼내 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 끈 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 끄 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 꺼내에 다는 건 어떠신가요",
  "돼지나 부엉이를 가방이나 꺼내 다 건 어떠신가요",
  "돼지나 부엉이를 가방이나 끈 다 건 어떠신가요",
  "돼지나 부엉이를 가방이나 끄 다 건 어떠신가요",
  "돼지나 부엉이를 가방이나 꺼내에 다 건 어떠신가요"
],
"pred": "돼지나 부엉이를 가방이나 끈에 다는 건 어떠신가요",
```


항상 편향이 발생 하는것은 아니며, 단순 Enesemble 문제도 잘 해결함

결론

1 . ASR 모델에 독립적인 방법을 통해 성능이 향상됐으나, 절대적 성능 자체는 아쉬움

2 . ASR 모델이 만든 오류를 추가적 단서 없이 수정하는 모습을 보여줌

3 . MultiModal을 통해 음성 표상을 통해 표상을 강화한다면 추가 성능 향상이 기대됨