**3. 시스템 설계**

**3.1 설계 블록도**

시스템 블록도는 다음과 같다. 뉴스 음성 데이터를 입력 받고 해당 데이터를 VAD알고리즘으로 자른다. 여기서 잘리는 구간의 길이는 1초 이상, 10초 이하로 제한하였는데, 이는 화자 분리 모델의 학습 데이터의 길이와 통일하기 위함이다. 또한 VAD알고리즘을 사용하는 이유는 최대한 음성을 문장 단위로 끊어서 STT(speech-to-text)과정에서 오류를 줄이기 위함이다. 이렇게 잘린 음성에 대해 화자 분리를 적용하여, 화자 별로 음성을 분리하였다. 다음 화자 인식 모델을 통해 같은 화자가 발화한 음성끼리 모두 모아 각각의 음성 별로 STT(speech-to-text)를 적용하여 합치면, 전체 음성에 대한 대본이 만들어 진다. 해당 대본을 요약 모델의 입력 값으로 하면 음성에 대한 요약문이 생성되게 된다. 요약 모델은 AI HUB의 ‘방송 콘텐츠 대본 요약 데이터’를 훈련 데이터와 평가 데이터를 8:2로 나눠 kobert를 학습시킨 모델로, 대본을 입력 값으로 넣으면 요약문을 출력해준다.

음성 요약

요약 모델

STT(speech-to-text)

대본

화자 인식

화자 분리 적용

음성 전 처리

뉴스 음성 데이터

**그림1. 전체 설계 블록도**

**4. 평가 지표**

요약문을 평가하는 지표로 ROUGE, METEOR, BLANC 3개를 사용하였다. ROUGE와 meteor는 참조 요약문을 사용하는 대표적인 요약문 평가 지표이고, BLANC는 참조 요약문을 사용하지 않는 요약문 평가 지표이다.

**4.1 ROUGE**

ROUGE는 Recall-Oriented Understudy for Gisting Evaluation의 준말로 언어 요약 모델의 성능을 평가하는 지표이다. 사람이 만든 참조 요약문과 모델이 만든 요약문을 대조하여 요약의 점수를 계산한다.

**4.1.1 ROUGE-N**

ROUGE-N은 N-gram의 수를 나타내고, 본 논문에서는 통상적으로 사용되는 ROUGE-1과 ROUGE-2를 사용하였다.

(1)

(2)

**4.1.2 ROUGE-L**

ROUGE-L은 common sequence중 가장 긴 것을 매칭한다. N- gram과 달리 순서나 위치 관계를 고려한 평가 지표이다.

(3)

(4)

**4.2 METEOR**

METEOR는 Metric for Evaluation of Translation with Explicit Ordering의 준말로 사람이 만든 참조 요약문과 모델이 만든 요약문을 대조하여 성능 점수를 계산한다. ROUGE와 다르게 precision과 recall에 가중치를 곱한 조화평균에 penalty를 곱하는 방법으로 점수를 측정한다.

(5)

(6)

**4.3 BLANC**

BLANC는 요약문의 점수를 평가하는 지표이다. 기존 ROUGE, METEOR와 기타 많이 사용되는 평가 지표와는 다르게 참조 요약문을 사용하지 않는 평가지표이다. 해당 지표는 언어 모델의 성능을 평가하는 masked token task를 사용하여 요약문의 성능을 평가한다. Masked token task는 문장안의 임의의 단어를 mask로 만들고 언어 모델을 통해 해당 mask를 주변 맥락을 고려하여 복원하는 작업을 의미한다. BLANC 평가 지표에는 BLANC help와 BLANC tune이 있는데 BLANC help를 사용하여 시스템을 평가하였다. BLANC help는 원래 전체 문장에 대해 masked token task를 수행하고, 요약문을 참고한 뒤 masked token task를 진행하여 점수의 향상 정도를 점수로 환산한다.

**Ⅲ. 분석 결과**

**1. 비교 대상 시스템**

화자 분리가 음성 요약에 효과가 있는지 정확하게 판단하기 위해서 본 시스템과 비슷한 조건으로 기존 음성 요약모델을 제작하여 baseline으로 설정하였다. 요약 모델은 본 시스템과 같은 모델인 kobart를 학습하여 제작하였다. 학습 데이터의 경우 같은 데이터 셋인 AI HUB의 ‘방송 콘텐츠 대본 요약 데이터’에서 화자 부분을 삭제하여 제작하였다. 비교 시스템의 시스템 블록도는 다음과 같다. 뉴스 음성을 입력으로 받아 STT(speech-to-text)를 수행하고 요약 모델을 통해 요약문을 얻는다.

뉴스 음성 데이터

STT(speech-to-text)

텍스트

요약 모델

음성 요약

**2. 요약 모델 성능 비교**

두 시스템이 각각 사용되는 요약 모델의 성능을 비교하였다. AI HUB의 ‘방송 콘텐츠 대본 요약 데이터’중 훈련으로 사용되지 않은 20%의 데이터를 사용하여 모델의 성능을 평가하였다. 성능 평가의 결과는 다음과 같았다. ROUGE의 경우 recall과 precision의 조화 평균을 표기하였다.

**표 1. 모델 성능 비교**

|  |  |  |
| --- | --- | --- |
|  | 화자 분리 X | 화자 분리 O |
| ROUGE-1 | 0.354 | 0.557 |
| ROUGE-2 | 0.257 | 0.299 |
| ROUGE-L | 0.313 | 0.399 |
| METEOR | 0.334 | 0.532 |
| BLANC | 0.228 | 0.307 |

전체 지표에 대해서 화자 분리를 적용한 요약 모델(대본 → 요약)이 화자 분리를 적용하지 않은 요약 모델(텍스트 → 요약)에 비해 모든 지표에 대해서 더 나은 성능을 보이는 것을 확인할 수 있었다.

3. 시스템 비교

위 두 모델을 적용한 각각의 시스템에 대해서 성능 비교를 진행하였다. real world data를 사용하기 위해 유튜브에서 뉴스를 60개 다운로드 하여 평가 데이터로 사용하였다. 해당 데이터는 앵커 혼자 말하는 단일 화자 음성 30개와 앵커, 기자, 인터뷰이 등 여러 명의 화자가 등장하는 다 화자 음성 30개로 구성 되어있다. 이렇게 데이터를 2가지 유형으로 구분하는 이유는 화자 분리가 음성 요약에 도움이 되는지 정확하게 판단하기 위함이다. 성능 측정 결과는 다음 표와 같았다. 대신 해당 데이터는 참조 요약문이 없기 때문에 BLANC만을 이용하여 요약의 품질을 평가하였다.

**표 2. 시스템 성능 비교**

|  |  |  |
| --- | --- | --- |
|  | 화자분리 X | 화자분리 O |
| 다 화자 음성 | 0.253 | 0.278 |
| 단일 화자 음성 | 0.263 | 0.222 |

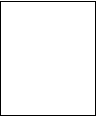
위 결과를 보니 다 화자 음성의 경우에는 화자 분리를 적용한 시스템이 더 좋은 요약 결과를 내는 것을 확인할 수 있었다. 하지만 단일 화자 음성의 경우에는 오히려 화자 분리를 적용하지 않은 시스템이 더 좋은 결과를 만들었다. 해당 이유를 분석해보니 음성에서 대본을 만드는 과정에서 화자 분리 및 화자 인식 과정에서 발생하는 오차로 인해 오히려 단순하게 STT(speech-to-text)를 적용하는 기존 시스템이 비교 더 좋지 않은 결과가 나온 것을 확인할 수 있었다. 그럼에도 불구하고 다 화자 음성의 경우에서 기존 시스템에 비해 좋은 요약문이 생성된 점으로 미루어 보아 화자 분리가 요약문 생성에 도움이 된다는 점을 추측할 수 있었다.

**Ⅳ. 결론**

기존 음성 요약의 경우 음성을 텍스트로 변환한 뒤 요약을 수행한다. 해당 시스템에 화자 분리를 적용하여 음성에서 화자 별로 발화 내용을 구분해 주었고, 대본 형식을 사용하여 이를 표시해 주었다. 이 대본을 사용하여 요약을 생성하는 시스템을 만들고, 기존 방식대로 만들 시스템을 만든 뒤 두 시스템의 성능을 비교하였다. 화자가 한명인 음성의 경우에는 기존의 방식대로 만든 시스템이 더 좋은 성능을 보였는데 이런 결과가 나온 이유를 분석하니, 화자 분리 및 화자 인식 단계에서 모델의 정확도 문제로 만들어지는 대본의 결과가 부정확했고 이 문제가 요약문에 반영된 것으로 사료된다. 다만 화자가 여러 명인 음성에 대해서는 화자 분리를 적용한 시스템이 더 좋은 요약문을 생성했는데, 이는 화자를 구분해 주는 것이 요약문 품질 향상에 어느정도 도움이 된다는 것으로 생각할 수 있다.

**참고 문헌**

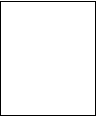
1. Yujun Wen, Hui Yuan and Pengzhou Zhang, "Research on keyword extraction based on Word2Vec weighted TextRank," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 2109-2113
2. Kim, H.-J., Cho, S., & Kang, P. (2014, February 15). KR-WordRank : An Unsupervised Korean Word Extraction Method Based on WordRank. Journal of Korean Institute of Industrial Engineers. Korean Institute of Industrial Engineers.
3. Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2020). SummEval: Re-evaluating Summarization Evaluation.
4. Vasilyev, O., Dharnidharka, V., & Bohannon, J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 11-20. arXiv preprint arXiv:2002.09836.
5. Kadagadkai, S., Patil, M., Nagathan, A., Harish, A., & MV, A. (2022). Summarization tool for multimedia data. Global Transitions Proceedings, 3(1), 2-7.
6. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. arXiv preprint
7. R. Sharma, S. Palaskar, A. W. Black and F. Metze, "End-to-End Speech Summarization Using Restricted Self-Attention," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8072-8076, doi: 10.1109/ICASSP43922.2022.9747320.
8. Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. [Fill in the BLANC: Human-free quality estimation of document summaries](https://aclanthology.org/2020.eval4nlp-1.2). In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pages 11–20, Online. Association for Computational Linguistics.
9. Shafey, Laurent & Soltau, Hagen & Shafran, Izhak. (2019). Joint Speech Recognition and Speaker Diarization via Sequence Transduction. 396-400. 10.21437/Interspeech.2019-1943.
10. A. Vartakavi, A. Garg and Z. Rafii, "Audio Summarization for Podcasts," 2021 29th European Signal Processing Conference(EUSIPCO), Dublin, Ireland, 2021, pp. 431-435, doi: 10.23919/EUSIPCO54536.2021.9615948.
11. C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong, "Attention Is All You Need In Speech Separation," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 21-25, doi: 10.1109/ICASSP39728.2021.9413901.

 **신 원 철**

 2017 ~ 현재 인하대학교 정보통신공학과 학사 과정 재학중

2024년 2월 졸업 예정

관심 분야는 백엔드 개발

**나 현 희**

 2016 ~ 현재 인하대학교 정보통신공학과 학사 과정 재학중

2024년 2월 졸업 예정

관심 분야는