대화 요약 생성을 위한 한국어 방송 대본 데이터셋

김봉수○, 전혜진, 전현규, 정혜인, 장정훈

와이즈넛
{usgnob, hjjun, eddie14, hijung, jhjang}@wisenut.co.kr

KMSS: Korean Media Script Dataset for Dialogue Summarization

Bong-Su Kim○, Hye-Jin Jun, Hyun-Kyu Jeon, Hye-in Jung, Jung-Hoon Jang
Wisenut Inc.

요 약

대화 요약은 다중 발화자와 발화문으로 이루어진 멀티턴 형식의 문서에 대해 핵심내용을 추출하거나 생성하는 태스크이다. 대화 요약 모델은 추천, 대화 시스템 등에 콘텐츠, 서비스 기록에 대한 분석을 제공하는데 유용하다. 하지만 모델 구축에 필요한 한국어 대화 요약 데이터셋에 대한 연구는 부족한 실정이다. 본 논문에서는 생성 기반 대화 요약을 위한 데이터셋을 제안한다. 이를 위해 국내 방송사의 대용량 콘텐츠로부터 원천 데이터를 수집하고, 주석자가 수작업으로 레이블링 하였다. 구축된 데이터셋 규모는 6개 카테고리에 대해 약 100K이며, 요약문은 단문장, 세문장, 2할문장으로 구분되어 레이블링 되었다. 또한 본 논문에서는 데이터의 특성을 내재화하고 통제할 수 있도록 대화 요약 레이블링 가이드를 제안한다. 이를 기준으로 모델 적합성 검증에 사용될 디코딩 모델 구조를 선정한다. 실험을 통해 구축된 데이터의 몇가지 특성을 조명하고, 후속 연구를 위한 벤치마크 성능을 제시한다. 데이터와 모델은 aihub.or.kr에 배포 되었다.

주제어: 대화 생성 요약, 방송 대본 데이터, 모델 적합성 검증

1. 서론

문서 요약(Document Summarization)은 문장의 시퀀스들로 이루어진 긴 입력 문서의 핵심 내용을 간추리기 위해, 주요 문장을 추출하거나 생성하는 자연어 처리 태스크이다.

데이터 기반 방법론에 따라 문서 요약 태스크를 해결 하기 위해, 대용량-고품질의 데이터를 확보하는 것은 중요한 문제이다. 기존 연구들은 주로 온라인 뉴스 보도를 수집하고 제목, 부제목 등을 요약문으로 사용하는 방식으로 데이터 부족 문제를 해결해왔다.

최근에는 소셜미디어, 메신저, 회의록 등 다양한 매체에서 수집된 대화체 데이터에 대해 문서 요약 태스크를 수행야할 필요가 생기고 있다. 예를 들어 Q&A 서비스의 상담 기록에서 상담사와 고객의 대화를 요약하면 서비스 수요 정보를 얻을 수 있다. 또 상담사의 발화만을 요약하여 서비스 품질 정보를, 고객의 발화를 요약하여 서비스 만족도 정보를 얻을 수 있다.

대화 요약(Dialogue Summarization)은 둘 이상의 발화자(Speaker)와 발화문(Utterance)으로 구성된 입력 문서를 이해하고, 단일 발화자의 발화문(Monologic Text)으로 요약하는 태스크이다. 대화 요약 데이터는 일반적인 문서요약 데이터와 다르게 발화문 간 구조, 발화자 간 관계, 대화 주제, 상식 등의 암묵적, 메타적인 정보가 내재되어 있다. 이 정보를 요약 태스크에 필요한 자질로 사용할 수 있도록, 대화 요약 모델이 설계될 필요가 있다. 또한 대화요약 태스크는 일반요약 태스크의 하위 분야로서, 기존 일반 요약 모델을 기준 벤치마크로 사용할 수 있다.

일반 문서요약과 마찬가지로 대화요약 태스크를 수행할 때도 대용량-고품질의 데이터셋을 필요로 한다. 하지만 대화체의 데이터는 온라인 뉴스 보도에 비해 수집, 정제, 주석 작업이 어렵고 복잡하다. 예를 들어 수집 단계에서 메신저, SNS, 회의록 등의 출처에서 대용량 데이터를 확보하기 위해 소유자의 라이센스 문제를 해결해야 한다. 정제 단계에서는 발화자, 등장인물과 관련하여 비식별화, 저작권 등 법적인 문제에 대한 해결책이 있어야 한다. 또 온라인 뉴스 보도에 비해 다수 존재하는 욕설, 혐오표현 등의 편향도 제거해야 한다. 마지막으로 주석 과정에서 참고할 만한 요약문의 부족, 레이블링 기준의 부족은 입력문-요약문 형식의 데이터셋 구축을 어렵게 한다.

본 논문에서는 풍부한 한국 방송 콘텐츠의 대본(Script) 자원을 활용하여 대용량의 원천 데이터를 수집한다. 방송 콘텐츠 대본 자체에는 참고할 수 있는 요약정보가 없기 때문에 주석자를 모집하고 수작업으로 레이블링 한다. 주석자에게는 어노테이션 도구와 명확한 레이블링 가이드라인이 제공된다. 마지막으로 가공 및 주석된 데이터는 모집된 검수자를 통해 주기적으로 검증된다.

최종적으로 구축된 대화 요약 데이터셋의 규모는 약 100K이며, 대화 요약 데이터셋과 모델 간의 적합성을 검증하기 위해 대화 요약 모델을 학습하고 검증한다. 모델은 제안하는 레이블링 가이드라인에 따라 특징지어지는 데이터 특성을 가정하여 선정된다. 본 논문의 기여는 아래와 같다.

1) 한국어 대화 요약 데이터 구축
2) 한국어 대화 요약 데이터 레이블링의 기준 제안
3) 데이터 특성을 고려한 모델 선정 과정 제시

4) 후속 연구를 위한 벤치마크 성능 제시

2. 관련 연구

본 절에서는 문서 요약 및 대화 요약 태스크에 필요한 대용량의 데이터셋을 확보하고, 모델을 통해 검증하는 것을 목표로 하였던 기존 연구를 살펴본다.

기존 국내 연구들은 주로 온라인 뉴스 보도를 수집 및 가공하여 데이터 부족 문제를 해결했다. 온라인 뉴스 보도는 요약문으로 참고 할 만한 정보가 많기 때문에 적은 공수로 고품질 대용량의 데이터를 수집할 수 있다. 예를 들어 [1-2]는 뉴스 보도에 이미 작성된 요약문이 존재하는 경우 노이즈를 제거하고 이를 입력문에 대한 요약문으로 사용하였다. 또한 [3-4]는 뉴스 보도의 중요한 정보는 문서 초기에 등장한다는 가정 하에 첫 문장이나 첫 문단을 요약문으로 사용하였다.

뉴스 보도 외의 도메인에서 수집된 원천 데이터를 활용하여 연구한 사례도 있다. [5-6]은 온라인 QnA 서비스의 질문-제목을 모델의 입력문-요약문으로 사용하였고, [7]은 위키 사이트의 문서 내에 존재하는 키워드를 단어 수준의 레이블로 사용하였다.

[1-7]의 연구는 제안하는 모델의 성능을 검증하기 위해 부수적인 목표로 데이터셋을 구축하였다. [8-9]는 요약 데이터셋 자체의 구축 관점에서 뉴스 보도 원천 데이터를 수집하고 모델 적합성 검증을 수행했다. [8]은 언론사가 소셜미디어에 게시한 뉴스 보도의 경우 기사 링크에 들어가지 않아도 기사 내용을 파악할 수 있도록 짧은 요약을 제공한다는 점을 활용하여 요약문을 레이블링하였다.

국내에서 연구된 한국어 문서 요약 데이터셋은 주로 온라인 뉴스 보도와 같이 단일 발화자 문서에 대해 구축되었다. [10]처럼 국회 회의록 원천 데이터에 대한 대화 요약 관련 연구가 있었지만, 레이블링의 어려움으로 약 36개의 회의에 대해 데이터셋이 구축되었다. 해외의 경우 다양한 도메인에 걸쳐 대화 요약 데이터셋과 모델이 활발하게 연구되고 있다.

AMI[11], ICSI[12]은 각각 제품 설계 관련 회의록과 컴퓨터 과학 연구소 내 회의록을 원천 데이터로 수집하고, 수작업으로 레이블링 하였다. [11]은 141개 [12]는 59로 소량의 데이터셋으로 구축되었다.

Samsum[13], Dialaogsum[14]은 일상 대화 도메인의 요약 데이터셋이다. [13]은 다소 짧은 대화(대화 당 약 94개의 토큰)으로 구성되었다. [14]는 영어 스피킹 사이트에서 원천 데이터를 수집하였고 학교, 직장, 여가 등 더 다양한 일상 생활 주제의 대화로 구성되어 있다. 두 데이터 모두 수작업으로 요약문이 작성되었다.

Ubuntu[15], ForrumSum[16], Nyc[17]는 인터넷 게시물 QnA 서비스 등에서 원천 데이터를 수집한다. [15-16]은 ubuntuforums.org 및 tripadvisor.com에서 100개의 스레드를 수집하여 수작업으로 레이블링 하였다. [17]은 인기있는 웹사이트 281개의 출처로부터 게시글을 스크랩하고 사이트당 최대 200개의 대화를 샘플링 및 수작업 레이블링을 수행하였다.

Dr.summarize[18], SOAPnotes[19]는 환자와 의사 간의 의료 대화 데이터를 수집하여 구축되었다. [18]은 의료플랫폼에서 약 25,000건의 대화를 수집하고, Pointer Generator 모델을 통해 생성된 요약문을 의사가 직접 평가하여 레이블링 하였다. [19]는 녹음 스크립트를 입력문으로, 의사가 작성한 반정형의 임상 요약본인 SOAP note를 요약문으로 간주하여 구축한 데이터이다.

LiveStream[20], Podcast[21]는 실시간 방송 스트리밍으로부터 구축한 데이터셋이다. [20]은 소셜 미디어 플랫폼인 Behance.net에서 5,000개의 스트리밍 비디오를 수집하였으며, 총 500시간이 넘는 비디오를 5분 단위로 분할하여 주석을 달았다. 각 클립에는 평균 51개의 발화와 460개의 단어가 포함되어 있다. [21]는 60,000시간 분량의 팟캐스트 스크립트를 수집하여 요약였고, 스크립트는 코미디, 사회, 문화, 뉴스, 정치 등의 다양한 장르로 구성되어 있다. 크리에이터가 작성한 에피소드 설명을 참조 요약으로 사용하였다.

MediaSum[22], SummScreen[23]은 방송 데이터를 통해 데이터를 수집한다 [22]는 NRP와 CNN의 인터뷰 내용을 수집하고, 개요와 주제 설명을 생성적 요약문으로 사용한 463.6K건의 데이터이다. 각 대화는 평균 30개의 발화와 6.5명의 화자, 1,553.7개의 단어로 구성되어 있으며, 주석에는 평균 14.4개의 Word가 포함되어 있다. [23]은 TV시리즈 대본에 사람이 주석을 작성하여 구축된 데이터셋이다. 대본은 화자의 이름이 있는 발화나 장면이나 행동에 의한 설명으로 구성되어 있다. 데이터셋 품질을 보장하기 위하여 요약문과 입력문에 등장하는 등장인물의 중복 비율을 유지하고, 등장인물 정보가 있는 발화문의 라인의 수를 기준으로 분할을 수행했다.

본 연구에서는 [20-23]와 같이 방송 콘텐츠를 대화 요약 데이터셋 구축을 위한 원천 데이터로 사용한다. 데이터 정제 단계에서 [20]처럼 하나의 방송은 고정 길이의 입력문서들로 분할 되고, 다양한 카테고리 주제로 분류 되어 각각에 대해 레이블링 가이드라인을 구축한다. 또한 참고 할 수 있는 요약문 자원이 없기 때문에 [20, 23]과 같이 수작업으로 요약문을 작성한다. 최종적으로 구축된 데이터의 규모는 약 100K이다.

모델 적합성 검증 단계에서 기존 연구는 시퀀스 데이터 모델링에 적합한 것으로 알려진 LSTM[24], Transformer[25] 구조가 주로 사용되었으며, 최근에는 문서 요약 태스크에서 좋은 성능을 보인다고 알려져 있는 Copy-Mechanism[26] 기반의 구조가 사용되고 있다. 본 연구에서는 이러한 접근에 따라 Transformer 인코더-디코더 구조의 모델을 사용하여 베이스라인 성능을 제시한다. 뿐만 아니라 대화 요약 데이터셋의 특성을 가정하여 개선된 구조를 추가로 선정하여, 적합성 검증을 수행하고, 데이터셋의 특성을 조명한다.

3. 대화 요약 데이터 구축

본 절에서는 데이터의 수집-가공-주석의 전반적인 과정에 대해서 설명한다. 먼저 구축할 대화 요약 데이터셋과 관련된 원천 데이터를 수집하기 위해서 KBS 미디어 콘텐츠의 라이센스를 구매하였다. 구매한 방송 콘텐츠 원천 데이터는 인터넷 플랫폼의 스트리밍 또는 파일형식으로 확인할 수 있으며, 대본 파일이 같이 제공된다.

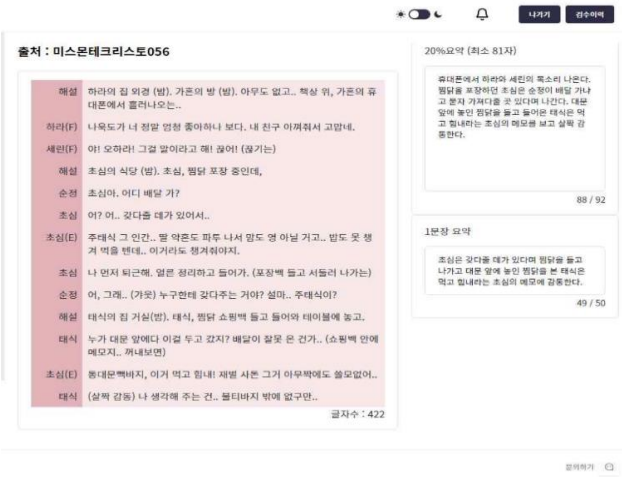
본 연구에서는 수집과 정제과정을 같이 처리하기 위하여 자사 제품인 BICrawler를 사용하였다. 먼저 수집 과정에서

띄어쓰기를 포함하여 5 음절 길이 미만의 문서는 반려되고, 문서의 최대길이는 1500 음절로서 그 이상의 문서는 분할-수집된다. 그리고 총 6개의 카테고리에 대해서 분류-수집되고 카테고리명은 각각 가족관련방송, 현대드라마, 역사극, 시사, 교양지식, 예능이다. 각 카테고리에 특성에 따라 발화량이 많은 가족 관련 방송, 현대드라마, 역사극의 경우는 장면(Scene) 단위로 획득된 원시데이터를 분할하여 사용한다.

수집 과정과 별개로 정제 과정에서의 분할 정제 길이는 띄어쓰기를 포함해서 500~1000 음절이다. 반면 발화량이 적은 시사, 교양지식, 예능의 경우 300~800 음절의 정제 길이 제한을 갖는다 마지막으로 원천 데이터의 전처리 과정을 거쳐 json 파일 포멧으로 변환되고, 형식 오류가 있는 데이터를 다시 한번 확인하여 제거한다.

데이터 수집과 정제 단계에서는 데이터 비식별화 작업을 수행하게 된다. 의뢰한 법률 기관의 자문에 따라 구축된 비식별화 가이드를 참고하여 추후 문제가 될 만한 등장인물, 고유명사, 욕설, 차별적 표현 등을 입력문에서 제거하거나 대체한다.

수집 및 정제가 완료된 원천 데이터의 주석작업에는 선발된 주석자가 배정되었고, 레이블링 검수 작업을 위해 선발된 언어학자 5명이 검수자로 배정된다. 두 그룹의 작업자들은 어노테이션 툴을 사용하여 협업하게 된다. [그림1]은 어노테이션 툴의 UI 이다.



[그림1 어노테이션 툴]

주석자가 원천 데이터를 수령하면 프로젝트와 함께 작업이 매칭되고, UI에 진입하면 오른쪽 요약 부분이 비어있게 된다. 작업자는 왼쪽에 있는 대화체의 원문을 확인하고, 주어진 레이블링 가이드에 맞춰 요약문 작성 작업을 수행 후 제출한다. 레이블링은 하나의 입력문에 대해 단문장, 세문장, 2할문장 요약문으로 길이 별로 작업을 수행해야한다. 이때 최종적으로 문장길이별 요약문의 목표 구축 수는 전체 입력문 100%에 대해 단문장 100%, 세문장 50%, 2할요약 50%의 요약문 구축을 목표로 한다.

주기적으로 레이블링이 완료된 데이터에 대해 검수자는 입력문-요약문을 아래와 같은 항목으로 검수한다.

1) 다양성 검수: 수집된 데이터의 카테고리의 비율을 균일하게 하기 위한 검수 항목이다. 분할-수집된 원천 데이터의 카테고리별 분포가 균등하지 않기 때문에 오차범위 ±5%의 분포를 유지하는 것을 목표로 한다.

2) 요약형태분포 검수: 목표한 입력문-요약문 길이의 비율이 적절한지 검수한다. 단문장 요약문은 20~50 음절, 세문장 요약문은 40~150 음절, 2할 요약문은 입력문 길이 대비 17~22% 음절인지 확인한다.

3) 구문정확성 검수: 최종 데이터셋으로 사용 가능한 지 검수하는 항목이다. 원천 데이터 수집 후 json 포멧으로 정제, 주석된 데이터에 대해 키값이 적절하게 기입되었는지 확인한다.

4) 의미정확성 검수: 마지막으로 의미정확성은 정성적으로 입력문-요약문 사이에서 주제와 의미가 유지되는지를 수작업으로 검수한다. 입력문-요약문 사이에 완전히 같은 문장이 포함되어 있지는 않은지, 핵심 문장과 키워드가 공통으로 포함 되어있는지, 시제와 문체가 일치하는지를 확인한다. 또 요약문의 주술관계와 조사가 적절한지, 입력문에 없는 배경지식이나 주관적 의견이 포함되어있는지 등을 확인한다.

구축된 데이터가 얼마나 대화 요약 모델에 적합한지 확인하기 위해, 모델과 데이터셋 간의 적합성 검사 또한 주기적으로 수행된다. 모델 검증은 데이터를 학습, 검증, 테스트 데이터셋으로 각각 8:1:1 스플릿해서 사용한다. 사용되는 모델과 실험에 대한 상세는 5절과 6절에서 소개한다.

방송 대본 원천 데이터를 통해 최종적으로 구축된 대화 요약 데이터는 약 10만건이다. 구축된 데이터와 검증에 사용된 모델은 AIHUB(www.aihub.or.kr)에 공개되어 있다.

4. 방송 대본 데이터의 특성

3절에서는 대화 요약 데이터셋을 구축하기 위해 KBS 미디어의 원천데이터를 수집-정제하고, 레이블링 및 검수작업을 통하여 데이터셋을 구축하는 일련의 과정을 설명하였다. 최종적으로 구축된 데이터의 카테고리와 요약문 길이별 규모는 [표1]과 같다.

[표1 카테고리별 데이터 규모]

데이터	입력문	단문장(분포)	세문장	2할요약
가족 관련 방송	643.6	20,000(20%)	9,786	10,214
현대 드라마	626.1	20,000(20%)	10,500	8,502
역사극	659.7	14,000(14%)	7,633	6,650
시사	787.8	17,000(17%)	8,792	8,491
교양지식	506.8	13,000(13%)	6,975	6,175
예능	762.6	16,000(16%)	10,478	9,168
전체	**671.2**	**100,000(100%)**	**54,164**	**50,200**

데이터셋의 전체 규모는 총 10만건이며, 카테고리 별로 13%~20% 정도의 분포를 가지고 있다. 평균 입력문의 길이는 띄어쓰기를 포함하여 가족관련 방송, 현대 드라마, 역사극이 650자, 시사와 예능은 750자, 교양지식 500자 정도이다. 전체 카테고리에 대해 입력문의 평균 길이는 671.2자 정도임을 [표1]에서 확인 할 수 있다.

[표2]에서는 본 연구에서 구축된 데이터셋(KMSS)과 기존에 연구된 데이터셋을 정량적 특성 관점에서 비교하고 있다. 대화 수, 턴 수, 화자 수 칼럼은 각각의 대화요약 데이터셋의 입력문의 수, 발화문의 수, 발화자의 수이며 모두 평균화 되었다.

[표2 기존 데이터와의 비교]

데이터명	대화 수	턴 수	화자 수	도메인
Samsum[13]	16.5K	9.9	2.2	미팅
MediaSum[22]	463.6K	30.0	6.5	인터뷰
SummScreen[23]	26.9K	≒330.0	28.3	방송
KMSS(fm_drama)	20K	15.8	4.1	방송(가족관련)
KMSS(fs_drama)	20K	15.3	4.2	방송(현대극)
KMSS(history)	14K	13.7	4.7	방송(역사극)
KMSS(c_event)	17k	5.9	2.0	방송(시사)
KMSS(culture)	13K	3.6	1.7	방송(교양지식)
KMSS(enter)	16K	5.5	2.0	방송(예능)
KMSS	**100K**	**10.3**	**3.2**	**방송**

먼저 대화 요약 태스크에서 널리 쓰이고 있는 Samsum Corpus와 비교 하였을 때 턴 수와 화자 수는 큰 차이가 없지만, KMSS에 더 많은 입력문이 포함되어 있다. 반면 방송 콘텐츠 도메인의 MediaSum과 SummScreen는 대화 수와 턴 수에서 KMSS보다 더 큰 입력문의 규모를 가지고 있다. 하지만 장면을 위주로 수집-분할된 KMSS에 비해 에피소드 단위로 수집된 기존의 방송 요약 데이터셋은 통제되지 않은 턴 수, 화자 수를 가지고 있다. 적당한 턴 수와 화자 수는 데이터 셋을 통해 학습된 대화 요약 모델이 실용화, 일반화 되는데 있어서 중요한 요소 중 하나이다. KMSS만의 특징으로는 다수의 카테고리에 대해 비교적 균일한 규모로 데이터셋이 수집되었고, 길이 별 세 가지 유형의 요약문이 작성되었다는 점이 있다.

[표3]은 가족 관련 방송 카테고리로 분류-주석된 데이터의 예시이다. 각각의 발화 문장은 개행과 문장 앞부분의 **연호], 박변]**과 같은 발화자명으로 구분되어 있고, 소설 문체와 같은 형식이 아닌 정형화된 발화문들이 나열되어 있는 형식이다. 대본 데이터의 특성상 아래 예시의 **(열변이 당황스럽다)**처럼 인물의 행동이나 상황에 대한 묘사가 발화문에 포함되어 있을 수 있다.

[표3 데이터 예시]

입력문:
연호]제가 요즘 제일 힘든 게 뭔지 아세요? 학부모들이 너무 우리 애, 우리 애, 그래서 얼마나 힘든지 몰라요. 선생님 못 믿어서 어딜 가든 다 따라다녀요! 전부들 자기 인생은 없구 아이한테 올인이에요!
박변](열변이 당황스럽다).
연호]일일이 다해주면 애들이 어떻게 앞으로 자기 힘으로 사나요?
해설]어디서 많이 듣던 말. 잠시 좀 묘한 기분이 드는 연호. 그러나 다시 말을 잇는다.
연호]그리구, 전 박변호사님이 이렇게 고리타분 하신 줄은 몰랐어요. 제 직업두 소중해요. 너무 함부로 말씀하시는 거 아니에요?
박변]함부로라뇨. 아닙니다. 전 다만 제 소신을 말씀 드린 거 뿐입니다..
연호](답답한 듯 창 밖을 바라보는데)
박변](다이어리 꺼내며) 아참, 잊어버리기 전에...부모님 인사 일정 말입니다..
연호](OL) 맨날 일정, 일정... 일정 밖에 모르세요?
박변]연호씨,
연호](스스로 당황) 죄송해요. 제가 오늘 두통이 심해서요... 이상하게 머리가 깨지는 거 같아요. (머리 감싸며).
박변]그러세요?(가만히 보다가) 혹시 무슨 지병 같은 거 있으십니까?
연호](한숨)

단문장 요약문:
자신의 직업을 함부로 말하지 말라는 연호에게 박변은 소신을 말했을 뿐이라고 하고는 부모님 인사 일정 이야기를 꺼낸다.

세문장 요약문:
연호는 박변 앞에서 요즘 학부모들 때문에 너무 힘들다며 열변을 토하고 박변은 그런 연호가 당황스럽다. 연호는 자신의 직업도 소중한데 함부로 말하는 것 아니냐고 하고 박변은 소신을 말했을 뿐이라고 답한다. 답답해 하는 연호 앞에서 박변이 부모님 인사 일정 이야기를 하자 연호는 맨날 일정밖에 모르냐며 화를 낸다.

2할 요약문:
연호는 자신이 힘든 점에 대해 열변을 토하고, 자신의 직업도 소중하다며 함부로 말하지 않을 것을 요구한다. 박변호사는 자신의 소신을 말했을 뿐이라고 하고는 부모님 인사 일정을 언급하지만 연호의 반응에 대화를 이어가지 못한다.

또 카테고리에 따라 등장인물 외에 **해설]**처럼 제3의 발화자가 존재할 수도 있으며, 입력문과 요약문에 발화자들이 언급되지 않은 경우 발화자명은 **화자1], 화자2]**와 같이 비식별화 되어 있다.

[표3]에 있는 요약문을 보면 입력문에 존재하는 핵심 키워드들이 단문장, 세문장, 2할 요약문 공통으로 포함되어 있고, 또 완전히 같지는 않지만 비슷한 문장들로 구성되어 있음을 알 수 있다. 이는 입력문에 존재하는 핵심 문장, 키워드를 기반으로 요약문이 구성되어 있다는 것을 뜻하며, 입력문 외의 배경지식에 독립적인 요약문이라는 것을 의미한다.

요약문의 표현 형식은 카테고리, 입력문의 형태, 대본 형식에 따라 아래와 같은 유형으로 분류되어 작성되어 있다.

1) 주체형: 발화자1 이(가) 000을 하고 있다.
2) 열거형: 000를 하고나서 000도 했으나, 결국 000 되었다.
3) 묘사형: 000 때문에 상황이 안 좋아져 분위기가 이상해졌다.

예를 들어 드라마나 가족 관련 방송 카테고리의 경우 등장인물간의 상황과 감정을 위주로 요약문이 작성되어 있다. 하지만 등장인물의 관계나, 대화 주제가 드러나지 않는 입력문이 많은 카테고리의 경우 상황을 묘사하거나 열거하는 형식으로 요약문이 작성되어 있다.

5. 적합성 검증 모델 선정

4절에서는 구축된 데이터셋을 정량적, 정성적 분석하였다. 대화 데이터, 방송 대본 데이터의 특성은 이미 수집-정제-주석 단계에서 제안하는 레이블링 기준에 따라 특징 지어진다. 이 점을 타겟팅하여 모델 적합성 검증 단계에서 데이터의 특성을 고려한 모델을 선정 할 수 있다.

본 절에서는 데이터의 특성을 고려하여 모델의 입력과 디코더 구조에 추가적인 개선 아이디어를 적용하였다. 기준

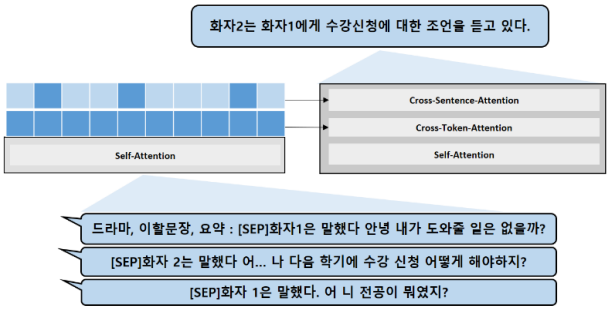
베이스라인 모델로 오픈소스인 KoT5(github.com/wisenut-research/KoT5)를 사용한다. 모델 선정 실험 및 모델 적합성 검증 실험에 사용된 모델은 아래와 같다.

T5[27]: Transformer 인코더-디코더 구조 기반의 모델로서 Text to Text로 모든 자연어 처리 태스크를 해결 하는 것을 목표로 한다. 모델의 입력에 Prefix를 넣고 풀고자 하는 태스크 정보를 명시해줌으로써 여러 태스크에 대한 Multi-task Learning 실험도 진행된 바 있다. T5는 비교적 단순한 MLM Pretrain task를 사용해 학습하기 때문에 모델 구조를 변경하기 쉽고, Down Stream Task를 설계하기가 상대적으로 쉽다.

T5 + Netural: 생성형 사전학습 모델에서 좀 더 자연스러운 인풋을 사용하여 Pretrain Task와 Finetuning Task의 간격을 줄이려는 기존의 아이디어들이 있었다. 본 연구에서 목표로 하는 대화요약 데이터셋에 이와 같은 아이디어를 적용할 수 있다. 예를 들면 대화체의 대화 요약 데이터셋의 인풋을 단일 발화자 문으로 관점을 변경하여 모델에 입력함으로써 성능 개선을 기대 할 수 있다.

T5 + Netural + Multi-task: 방송 대본을 통해 구축된 대화 요약 데이터는 다양한 카테고리의 입력문, 문장길이별 요약문을 가지고 있기 때문에, 실험 설계가 어려우며 구축된 학습 데이터셋을 충분히 사용하지 못한다. 이를 해결하기 위해 T5 모델 실험에 포함되어 있던 태스크 명시를 통한 Multi-task learning을 적용할 수 있다. 카테고리와 요약문 길이를 입력문 앞에 Prefix로 명시함으로써 모든 데이터셋을 사용해 하나의 모델을 구축할 수 있다.

T5 + Netural + Multi-task + Hierchy: 대화 요약 데이터셋은 발화자 뿐만 아니라 발화문에 대해서 모델링 될 필요가 있다. 대화의 흐름과 맥락은 항상 순차적으로 연결되지 않기 때문이다. 또 요약문은 입력문의 핵심 문장을 추출 및 변환한 결과이기 때문에, 핵심 문장 추출과 관련된 어텐션 스코어를 도입할 수 있다. 인코더에서는 토큰들의 표상에서 문장들에 대한 표상을 추출하며, 이를 위해 문장 단위 정보와 관련된 스페셜 토큰을 문장 앞부분에 삽입한다. 그 다음 두 가지 표상 모두를 사용해 디코더가 문장을 생성하는데 참조하는 어텐션 스코어를 제공한다.



[그림2 적합성 검증 모델 구조]

[그림 2]에서는 마지막 모델(T5 + Netural + Multi-task + Hierchy)을 묘사하였으며, 대화 요약 데이터셋 특성을 고려하여 고안된 아이디어가 모두 적용된 모델의 아키텍처이다.

6. 실험

본 절에서는 데이터셋의 특성에 맞게 고안된 모델들의 성능을 비교 분석하고 최종적으로 모델을 선정한다. 그 후 선정된 모델을 전체 카테고리와 요약문에 대해 학습 및 평가 하여 모델과 데이터간의 적합성을 평가한다. 데이터셋은 학습, 검증, 테스트 데이터셋으로 각각 8:1:1로 스플릿되어 사용된다.

평가 매트릭은 요약 태스크에서 사용되는 Rouge 1, 2, L을 사용한다. Rouge Unit의 기본적인 단위는 형태소 토큰이며, 형태소 분석에는 오픈소스 라이브러리인 Komoran을 사용하였다.

실험결과에 대한 [표4-표7]에서 Model열에 있는 모델 명은 5절에서 언급한 모델명의 생략된 표현이다. 예를 들어 T5~Hierchy는 T5, Natural, Multi-task, Hierchy의 아이디어가 모두 적용된 모델이다. 그리고 TrainSet과 TestSet이라는 열에서 표현된 데이터셋의 종류의 경우 {Category}_{Summary Length}의 형태로 되어있다. 예를 들면 all_single 은 모든 카테고리에 대한 단문장 요약 셋이며, enter_all은 예능 카테고리에 대한 단문장 요약 셋이다.

먼저 [표4]는 모델을 선정하는 기준이 되는 실험 결과로서 작은 사이즈의 모델인 T5-small을 사용하여 5절에서 고안된 모델들 중 대화 요약 데이터셋에 적합한 모델을 선정한다.

[표4 모델 선정 테스트]

Model	TrainSet	TestSet	Rouge(1 2 L)
T5	all_single	all_single	42.39 19.68 32.30
T5	all_3sent	all_3sent	53.24 27.10 37.43
T5	all_20per	all_20per	47.55 23.33 35.34
T5 ~ Natural	all_single	all_single	42.88 19.72 33.81
T5 ~ Natural	all_3sent	all_3sent	53.94 27.59 37.55
T5 ~ Natural	all_20per	all_20per	47.58 23.49 35.68
T5 ~ Mutli-task	all_all	all_single	43.23 20.11 34.24
T5 ~ Mutli-task	all_all	all_3sent	**54.03** 28.04 37.88
T5 ~ Mutli-task	all_all	all_20per	47.97 **23.66** 35.69
T5 ~ Hierchy	all_all	all_single	**43.42 20.31 34.28**
T5 ~ Hierchy	all_all	all_3sent	54.01 **28.07 37.89**
T5 ~ Hierchy	all_all	all_20per	**48.00** 23.65 **35.78**

[표4]의 실험결과에서 첫번째 행의 T5와 두번째 행인 T5-Natural의 성능을 비교했을때, 구조화된 대화 형식이 아닌 자연스럽게 인풋을 변경해서 넣었을 때 유의미한 성능 개선이 있는 것으로 보인다. 또한 세번째 행인 T5~Multi-task의 경우 그 이전 모델보다 훨씬 큰 폭의 성능향상이 있었다. 이는 카테고리나 요약문 길이별로 따로 학습하지 않고 구축된 데이터셋을 모두 사용해 학습했을 때 더 좋은 성능을 달성할 수 있음을 보여준다. 마지막으로 4행의 T5~Hierchy의 경우 대부분의 평가 매트릭에서 3행의 T5-Multi-task보다 좋은 성능을 달성하였다.

데이터의 특성에 따라 고안된 아이디어는 모델의 태스크 해결 능력을 향상시키는 것으로 보이며, 최종적으로 데이터 적합성 모델로 T5~Hierchy를 선정하였다.

[표5 모델 사이즈별 성능]

Model	TrainSet	TestSet	Rouge(1 2 L)
T5-small~Hierchy	all all	all all	48.00 23.65 35.77
T5-base~Hierchy	**all all**	**all all**	**49.69 25.40 37.54**

[표5]에서는 선정된 모델의 사이즈 별로 실험을 진행하였고, 파라미터 수가 많은 모델이 더 좋은 성능을 발휘하는 것을 확인할 수 있다. 더 큰 스케일의 데이터셋, 신경망 구조를 학습시키는 것은 모델의 성능을 향상시킨다는 근본적인 가정에 따라 이러한 결과는 합리적으로 보인다.

[표6 문장 길이별 성능]

Model	TrainSet	TestSet	Rouge(1 2 L)
T5-small~Hierchy	all all	all single	43.46 20.31 34.28
T5-base~Hierchy	**all all**	**all single**	**45.22 22.01 35.92**
T5-small~Hierchy	all all	all 3sent	54.01 28.07 37.89
T5-base~Hierchy	**all all**	**all 3sent**	**55.69 29.93 39.91**
T5-small~Hierchy	all all	all 20per	48.02 23.65 35.78
T5-base~Hierchy	**all all**	**all 20per**	**49.69 25.40 37.54**

[표6]에서는 요약문의 길이 별로 테스트를 진행하였으며, 세문장 요약이 가장 좋은 성능을 발휘하는 것을 알 수 있다. 세문장 요약문은 보통 단문장보다 길고 2할 문장보다 짧은 특성을 가지고 있다. 세문장 요약은 데이터 레이블링 하는 과정에서 추출된 2~5개의 핵심문장을 그대로 추출-변형하여 작성되는 경우가 많다. 이에 비해 단문장 요약은 정보가 함축되는 경우가 많고, 2할 요약은 입력문 길이 마다 상이한 요약문 길이를 가지고 있어 주석자 간에 일관되지 않은 레이블링이 존재할 가능성이 높다. 또한 Multi-Task Learning의 특성상 적절한 길이의 세문장 요약문에 대해 모델이 가장 잘 피팅되었을 가능성도 배제할 수 없다.

[표7 카테고리별 성능]

Model	TrainSet	TestSet	Rouge
T5-small~Hierchy	all all	fm_drama all	47.96 22.42 34.98
T5-base~Hierchy	**all all**	**fm_drama all**	**49.56 23.94 36.64**
T5-small~Hierchy	all all	fs_drama all	49.02 23.96 36.24
T5-base~Hierchy	**all all**	**fs_drama all**	**50.61 25.57 38.00**
T5-small~Hierchy	all all	history all	49.74 24.47 36.95
T5-base~Hierchy	**all all**	**history all**	**50.93 25.76 38.21**
T5-small~Hierchy	all all	c_event all	44.96 20.78 32.66
T5-base~Hierchy	**all all**	**c_event all**	**47.25 23.16 35.04**
T5-small~Hierchy	all all	culture all	46.70 23.81 35.61
T5-base~Hierchy	**all all**	**culture all**	**48.71 26.02 37.59**
T5-small~Hierchy	all all	enter all	49.52 26.96 38.62
T5-base~Hierchy	**all all**	**enter all**	**50.96 28.55 40.13**

선정된 모델이 모든 카테고리에 대해 어느정도 균일한 성능을 보이고 있음을 [표7]을 통해 알 수 있다. 가장 데이터셋의 크기가 적었던 시사(c_event), 교양지식(culture)의 경우 성능이 낮은 편에 속하고, 현대 드라마(fs_drama)과 가족 관련(fm_drama)의 경우 높은 성능을 보이는 것으로 보아 절대적인 데이터 셋의 양이 모델 성능에 영향을 미친다고 볼 수 있다.

하지만 작은 데이터셋 규모의 역사극(history)과 예능(enter)이 가장 좋은 성능을 보이는 것은 쉽게 이해가 되지 않는다. 이는 데이터셋 양 뿐만 아니라 고유한 카테고리별 특성도 중요함을 보여준다. 4절의 마지막에 언급했듯이 요약문의 형태는 주체형, 열거형, 묘사형이 있다. 드라마와 가족 관련 카테고리는 주체형과 묘사형으로, 역사극과 예능은 열거형으로 요약문이 대부분 작성되었다. 주체형과 묘사형은 모델이 발화자간의 관계를 고려해야하는 부담이 있으며, 열거형은 핵심 문장을 추출하는 것만으로 어느 정도 성능을 보장 할 수 있는 것으로 보인다.

6. 결론

본 논문에서는 한국어 대화 요약을 위해 방송 대본 데이터를 수집하고 가공하였다. 원천 데이터는 제안하는 레이블링 기준에 따라 수작업으로 레이블링 및 검수되었고 그 결과 총 100K의 한국어 대화 요약 데이터셋이 구축 되었다. 일반 문서 요약과 구별되는 대화 요약 데이터의 특성을 고려하여 대화 요약 모델을 고안하고 선정하였으며, 실험 설계를 구성하였다. 실험 결과 대화 요약 데이터셋의 특성에 맞게 모델을 고안하는 작업은 모델의 성능을 개선 시킬 수 있음을 확인함과 동시에 데이터가 학습에 적합함을 확인하였다. 최종적으로 선정된 모델로 실험한 결과 구축된 데이터의 특성을 다시 한번 확인할 수 있었으며, 후속 연구를 위한 유의미한 벤치마크 성능을 제시하였다. 사용된 데이터와 모델은 aihub.or.kr에 배포 되었다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.1711117120, 뉴럴-심볼릭(Neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

참고문헌

[1] 전재원, 황현선, 이창기, “딥러닝과 Maximal Marginal Relevance를 이용한 2단계 문서 요약”, 제31회 한글 및 한국어 정보처리 학술대회, 2019.

[2] 황현선, 이창기, 고우영, 윤한준, “복사메커니즘과 강화 학습을 적용한 BERT 기반의 문서 요약 모델”, 제32회 한글 및 한국어 정보처리 학술대회, 2020.

[3] Kyoung-Ho Choi and Chang-Ki Lee, “End-to-end Korean Document Summarization using Copy Mechanism and Input-feeding,” Journal of KIISE, Vol.44, No.5, pp.503-509, 2017.

[4] T.-H. Kim, K. A-Yeong, N. Yunseok, S.-B. Park, and S.-Y. Park, “Generation of news article dataset using lead for neural summarization model”, Proceedings of the KIISE Korea Software Congress, pp. 688-690, 2017.

[5] 김원우, 김선훈, 장헌석, 강인호, 박광현, “Pointer-

Generator Networks를 이용한 cQA 시스템 질문 요약”, 제30회 한글 및 한국어 정보처리 학술대회, 2018.
[6] Su-Jin Baek, “Multi-Document Summarization Method Based on Semantic Relationship using VAE”, The Society of Digital Policy and Management, Pages.341-347, 2017
[7] 이경호, 박요한, 이공주, “신문기사와 소셜 미디어를 활용한 한국어 문서요약 데이터 구축”, KIPS Trans. Softw. and Data Eng. Vol.9, No.8 pp.251~258 pISSN: 2287-5905.
[8] 권홍석, 고병헌, 박주홍, 이명지, 오재영, 허담, 이종혁, “요점만 남긴 신문 기사: 한국어 표제 형식 문서 요약 데이터셋”, 제32회 한글 및 한국어 정보처리 학술대회, 2020.
[9] 이재걸, 박성배, 이상조, “2단계 문장 추출 방법을 이용한 회의록 요약”, 한국지능시스템학회 논문지 2010, Vol. 20, No. 6, pp. 741-747.
[10] Jean Carletta, Simone Ashby, et al, “The AMI meeting corpus: A pre-announcement”. In International workshop on machine learning for multimodal interaction, pages 28-39. Springer, 2005.
[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus”. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)., volume 1, pages I-I.
[12] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer, “SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization”, In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 70-79, 2019.
[13] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang, “DialogSum: A Real-Life Scenario Dialogue Summarization Dataset”, In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062-5074.
[14] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau, “The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems”, In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285-294, 2015.
[15] Misha Khalman, Yao Zhao, and Mohammad Saleh, “ForumSum: A Multi-Speaker Conversation Summarization Dataset.” In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4592-4599
[16] Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra, “Summarizing online forum discussions-can dialog acts of individual messages help?”, In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 2127-2131.
[17] Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan, “Dr. summarize: Global summarization of medical dialogue by exploiting local structures”, In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3755- 3763
[18] Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton, “Generating SOAP notes from doctor-patient conversations using modular summarization techniques”, In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4958-4972
[19] J. Ulrich, G. Murray, and G. Carenini, “A publicly available annotated corpus for supervised email summarization”, In AAAI08 EMAIL Workshop, Chicago, USA. AAAI, 2008.
[20] Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu, “StreamHover: Livestream transcript summarization and annotation”, In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6457-6474
[21] m Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones, “100,000 podcasts: A spoken English document corpus”, In Proceedings of the 28th International Conference on Computational Linguistics, pages 5903-5917, 2020.
[22] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng, “MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization”, In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5927-5934
[23] Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. “SummScreen: A Dataset for Abstractive Screenplay Summarization”, In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8602-8615, 2022.
[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, Neural Computation 1997, 9(8):1735-1780.
[25] A. Vaswani, et al, “Attention Is All You Need”, Neural Information Processing Systems, 2017.
[26] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio, “Pointing the unknown words”, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 140-149, 2016.
[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, “Exploring the limits of transfer learning with a unified text-totext transformer”, Journal of Machine Learning Research, 21(140):1-67, 2020.