한국어 회의록 생성 요약을 위한 국회 회의록 요약 말뭉치 구축 연구

함영균[1]○, 강예지[2], 박서윤[2], 정용빈[1], 서현빈[1], 이이슬[1], 서혜진[3], 서샛별[3], 김한샘[2]
테디썸[1], 연세대학교[2], 국립국어원[3]
{hahmyg, ybjeong, errol.seo, yslee}@teddysum.ai, {yjkang5009, seoyoon.park, khss}@yonsei.ac.kr, {koreanshj, saetbyol}@korea.kr

Corpus Construction of National Assembly Minutes Summarization for Korean Abstractive Meeting Minutes Summarization

Younggyun Hahm○, Yejee Kang, Seoyoon Park, Yongbin Jeong, Hyunbin Seo, Yiseul Lee, Hyejin Seo, Saetbyol Seo, Hansam Kim
Teddysum, Yonsei University, National Institute of Korean Language

요 약

요약 연구의 주류는 아직 문서를 대상으로 하지만, 최근에는 회의 요약 연구에 대한 관심이 크게 높아지고 있다. 본 연구는 국립국어원 국어 빅데이터 구축 사업의 일환으로 국내에서 아직 연구되지 않은 국회 회의록 생성 요약에 대해 연구를 진행하였으며, 국회 회의록에 대한 생성 요약 데이터세트를 구축하였다. 또한 생성 요약 모델을 통해 구축된 데이터세트에 대한 정량 및 정성적 평가를 진행함으로써 국회 회의록 요약 데이터세트에 대한 평가 및 향후 생성 요약과 회의록 요약의 연구 방향을 모색하였다.

주제어: 회의록 요약, 추상 요약, 요약 말뭉치

1. 서론

자동화된 대화 요약은 강연이나 고객 서비스 대화, 회사의 업무 회의, 의사와 환자의 대화 등 실제 세계의 응용 시스템에 적용될 수 있는 매력적인 분야이다. 특히 회의 대화에 대한 요약은 비대면 회의 및 원격 근무가 증가하고 있는 최근 상황에서 중요하며 이러한 연구는 업무의 효율을 높일 수 있도록 해주며 회의록 작성에 드는 비용을 크게 절감할 수 있다.

그러나 일반적으로 여러 명의 화자가 참여하는 대화를 분석하는 것은 뉴스나 도서와 같이 한 명의 저자에 의해 기술되는 문서를 분석하는 것보다 상대적으로 어려운 과업이다. 회의 대화는 일반 대화에 비해 정제된 문장을 사용하는 경향이 있다. 그러나 정형화된 프레임을 가지지 않으며, 참여 발화자가 많고 다양한 주제 및 도메인을 다루기 때문에 회의 요약 데이터 구축 및 모델 개발에 어려움이 있다.

특히 국회 회의록의 경우 회의록 당 회의 참석자 인원이 평균 16명이며, 회의록 당 발화 개수가 평균 475개로 발화 참여자가 많으며, 한 번의 회의가 A4 문서 기준 20쪽 내외, 문서당 평균 어절 13,619개의 긴 대화들로 구성되어 있다[1]. 또한 회의의 각 단계가 구분되지 않으며, 각 주제가 독립적으로 구성되지 않고 회의 전반에 걸쳐 논의가 되기 때문에 요약문을 작성하는 데 특히 어려움이 있다. 따라서 회의 전체에 대한 구조 분석이 필수적이며, 회의 내에서 다루고 있는 주제들을 파악하고 각 주제에 따른 발화 간의 논변 구조를 파악하는 것이 선행되어야 한다.

이에 본 논문에서는 국내외에서 아직 수행된 바 없는 긴 대화들로 구성된 국회 소위원회 회의록을 대상으로 대화 요약문 구축 방법을 제시하는 바이다. 구축한 말뭉치는 베이스라인 실험을 통해 정량적 평가와 정성적 평가를 진행하였다.

2. 관련 연구

2.1. 회의 요약 데이터

기존의 생성 요약 연구는 저자에 의해서 잘 쓰여진 뉴스, 논문, 도서자료 등에 대한 문서 요약에 집중되어 있다. 해외에서는 CNN/Daily Mail 말뭉치[1,2], Gigaword 말뭉치[3], DUC 2004 Task1[4], Sentence Compression[5,6] 등의 말뭉치 등이 있으며 국내에서는 2019년 국립국어원에서 문서 요약 말뭉치(신문 기사 요약)[7]를, 2020년 한국정보화진흥원에서 요약데이터[2] 등을 구축하여 공개되었다.

대화에 대한 생성 요약 연구로는 해외에서는 회사에서의 회의 대화에 대한 요약을 제공하는 ICSI 데이터세트

1) '국립국어원 국회 회의록 말뭉치 2021'은 '03~'20년 국회 소위원회 회의록 5,395건을 대상으로 구축, 총 어절 73,478,080개, 문서당 평균 어절 13,619개로 구성되어 있다. 본 연구에서는 그 중 A4 문서 기준 20쪽 내외의 회의록 200건('09 1건, 2012~2020년 199건)을 선정, 총 어절 1,560,255개, 문서당 평균 어절 7,801개에 대해 요약 분석을 실시하였다.

2) https://aihub.or.kr/

[8]이나 AMI 데이터세트[9]가 구축되어 공개된 바 있으며, 이를 활용한 다양한 요약 모델 개발 연구가 최근 주목받고 있다[10,11].

ICSI 데이터세트는 학술적 성격을 띤 그룹 회의 녹취록 75건을 바탕으로 구축된 회의 데이터세트이다. 데이터세트에는 메타정보와 더불어 회의 본문, 참여자의 특징 등이 주석되었으며 회의 요약을 위한 기초적인 데이터세트 구축을 목적으로 구축되었다.

회의 요약을 다루는 대표적인 데이터세트인 AMI 데이터세트는 100시간 분량의 회의 녹화를 토대로 구축한 회의 요약 데이터세트며 주제에 따라 회의 내부의 대화들이 분할되어 있다. 주제별 대화 분할 후에는 각 대화에 대한 추출(extractive) 및 추상(abstractive) 요약이 수행되었다. 이에 따라 대화 요약에는 대화에 대한 추상 요약인 'ABSTRACT'와 회의를 통해 도출된 결정을 나타내는 'DECISION', 회의 내에서 제시된 안건이나 문제를 나타내는 'PROBLEMS/ISSUES' 그리고 회의를 통해 수행하게 된 후속 사항인 'ACTIONS' 주석을 포함하여 요약문을 구조화하였다. 본 연구에서도 요약문의 중요 정보를 주석하고 요약문을 구조화하고자 AMI의 대화 요약 레이블을 사용하여 회의록 내 분할된 주제에 대해 해당 정보들을 주석하였다.

최근에 연구된 회의 요약 데이터세트로는 QMSum[12]을 들 수 있다. QMSum의 경우 길이가 긴 회의에 대해 주제와 관련된 대화와 범위(span)만을 선택하여 요약하는 쿼리(query) 기반 요약을 채택하고 있으며, 쿼리를 입력하여 해당 쿼리와 관련된 대화의 추상 요약을 산출한다. QMSum은 232개 회의에 대한 쿼리(query)-요약(summary) 쌍 1,808개로 구성되어 있으며 데이터세트에는 제품 디자인 회의, 학술 세미나 회의, 그리고 의회 회의록이 사용되었다. 본 연구에서도 국회 회의록에 대한 정확한 요약을 위해 키워드를 작성하는 방법을 채택하였다.

대화 요약 데이터세트에도 담화 구조를 구조화하려는 시도들이 있었다. DIALOGUESUM[13]의 경우 대화를 요약한 데이터세트이나, 대화에서 드러나는 담화 관계(discourse relation)를 주석함으로써 대화 내 주요 사건들을 구조화하였다. 가령 대화 내 한 사건이 다른 사건의 원인이 될 경우 이를 'since'로 표현함으로써 담화 관계를 밝힐 수 있다. ConvoSumm[14]은 대화 요약 시 'issue-viewpoints-assertions' 프레임을 도입하여 요약문에 대한 구조화를 시도하였다. 특히 데이터세트에 포함된 토론 포럼(discussion forums) 게시판 게시물에 대해서는 각 스레드(thread)별로 담화 구조를 주석하기도 하였다.

국내의 경우에는 대화 요약을 위해 2020년 한국정보화진흥원에 의해 일상대화 및 토론 대화에 대한 요약데이터 구축 연구가 시작 단계에 와 있으나 대화의 말차례(turn)가 4~5회의 짧은 대화를 주로 다루고 있으며, 대화 데이터 구축 과정에서 정형화된 프레임(예: 토론의 주제, 문제 제기, 찬성, 반론, 결론 등의 구성요소가 사전에 정의된 상태로 구축된 가상의 대화 데이터)으로 구성된 하나의 주제에 대한 짧은, 그리고 인위적으로 만들어진 대화 요약 연구라는 점에서 한계가 있다.

본 연구는 한 번의 회의가 A4 문서 기준 20쪽 내외, 문서당 평균 어절 13,619개의 긴 대화들로 구성된 국회 소위원회 회의록을 대상으로 하여 현실의 회의, 즉 논쟁적 대화에 대해 요약하였다. 기존의 해외 연구(예: AMI 데이터세트)의 경우 회사에서의 회의(1개 대화가 평균 약 160 말차례로 구성)에 대한 요약 데이터세트를 구축한 바 있으나, 20쪽 내외의 길고 평균 4개 이상의 법안에 대한 설명과 질의응답으로 구성되었으며, 전문적인 영역을 다루는 국회 회의록과 같은 복잡한 데이터세트에 대한 연구는 아직 국내외에서 수행된 바 없다. 표 1은 본 연구와 기존 연구의 비교이다.

표 1 본 연구와 기존 연구의 비교

	문서요약	일상대화[3]	AMI[8]	**본 연구**
화자	1명(저자)	다자	다자	**다자**
주제	1개	1개	1개	**약 4.42개 (안건 별)**
분량	1페이지 가량	4~5 말차례	약 160 말차례	**약 475 말차례**

본 연구는 다음의 특징을 갖고 있다.

- 뉴스, 도서와 같은 독백과 달리 다자 대화를 요약의 대상으로 하였다.
- 전문적인 도메인(국회 회의록)을 다룬다. 이는 여러 안건(주제)에 대한 긴 대화이며, 각 주제는 독립적 대화로 구성되지 않고 회의 전반에 걸쳐 논의된다.
- 기존의 연구에 비해 긴 문서를 다루며, 기존의 일상대화 요약문이 1개 문장으로 작성된 것과 달리 여러 문장의 요약문으로 작성되었다. 그리고 주제에 대한 세부 요약문(문제, 결정사항, 후속조치 등)과 주제 요약문, 그리고 회의록 전체에 대한 대표 요약문으로 구성되어 회의의 구조를 반영하였다.

2.2. 생성 요약의 정량적 평가

생성 요약(summary generation)을 자동으로 평가하는 방법에는 보편적으로 ROUGE 점수가 활용된다[15]. ROUGE 점수는 재현율(recall)과 n-gram을 바탕으로 측정되는 요약 자동 평가 방법으로써, n-gram에 따라 uni-gram을 사용한 ROUGE-1, bi-gram을 사용한 ROUGE-2 혹은 n-gram 순서에 상관없이 텍스트 안에서 등장하는 최장 길이 문자열(longest sequence)을 바탕으로 하는 ROUGE-L이 ROUGE 점수에 포함된다.

2.3. 생성 요약의 정성적 평가

생성된 요약문에 대한 정성적인 평가는 주로 요약문에 대한 ROUGE 점수가 인간의 언어적 직관에 얼마나 부합하는지 측정하는 것을 목표로 하며, 요약 태스크의 정량적 평가에 대한 대표적인 연구로는 [16,17,18]이 있다. [16]연구에서는 2명의 언어 전문가로 하여금 생성 요약

3) 한국어 대화 요약, https://aihub.or.kr/

에 대해 -1부터 1까지의 점수를 매기도록 한 후, 둘 간의 점수를 cohen's kappa coefficient로 검증하였다. [17]에서는 기계로 생성된 요약에 대해 보다 구체적인 지표인 fluency, consistency relevance, coherence, discourse relation, coreference information, Intent identification을 활용하여 각 지표에 대해 -1부터 1까지의 점수를 매김으로써 정성적 평가를 진행하였다. [18]에서는 정성적 평가를 활용하여 데이터세트의 정확도를 제고하였다. 구체적으로는 기계가 생성한 후보 요약문에 대해 정성적 평가를 진행한 후 다시 이를 수정하여 학습 데이터로 사용하였으며, 이를 통해 최종 목적지향 대화 요약 데이터세트를 얻었다. 본 연구에서는 기초 연구임을 감안하여 가장 성능이 좋은 모델과 낮은 모델에 대해서 정성적 평가를 진행하였다.

3. 국회 회의록 요약 말뭉치 구축

국회 회의록은 각 안건에 대한 정부 측의 의견, 전문위원의 검토 사항 및 의견과 해당 안건에 대한 참석 위원들의 질의응답 및 의견, 안건에 대한 동의 및 비동의 등으로 구성되어 있으며, 안건과 관계없는 내용도 포함되어 있다. 국회 회의록 요약 말뭉치는 아래와 같은 과정을 거쳐 총 200건의 요약 말뭉치로 구축되었다.

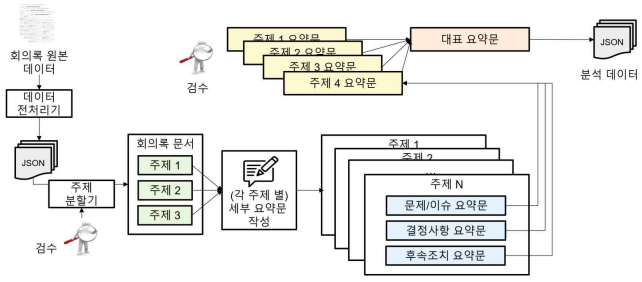

그림 1 국회 회의록 요약 말뭉치 구축 개념도

그림 1과 같이, 국회 회의록 요약 말뭉치는 국회 회의록의 입체적 구조, 즉 다수의 안건(주제)을 다루며, 각 안건에 대해 문제와 결정사항, 후속조치 등의 주요 정보가 포함되는 구조임을 고려하여 상향식으로 구축되었다.

먼저, hwp 파일로 된 국회 회의록에 대한 원시 말뭉치를 회의록의 구조를 바탕으로 전처리하여 발화와 메타정보를 추출하였다. 이후 주제 분할을 통해 주제별 대화들에 대해 각각의 요약문을 작성하였다. 이때, 주제별 요약문은 세부 요약문들과 주제 요약문으로 구성하여 주제에 대한 문제, 결정사항, 후속조치가 명확히 드러날 수 있도록 하였다. 마지막으로 각 주제에 대한 요약문을 토대로 회의록 전체에 대한 대표 요약문을 작성한다.

3.1. 회의록 선정 및 전처리

국회 회의록 요약 말뭉치 구축을 위해 다양한 도메인에서 말싸움을 최소한으로 포함하고 있는 회의록 200개를 선정하였으며 원자료인 hwp 형식의 파일에서 회의록 카테고리, 제목, 작성자, 안건, 발화자 목록, 발화 및 발화에 대한 정보들을 추출한 후 전처리 작업을 거쳤다.

3.2. 구축 과정

본 연구의 국회 회의록 요약문은 각각의 안건과 쟁점별로 요약하기 때문에 일반 신문기사 요약문과는 달리 핵심 문장 몇 개를 추출하여 환원하는 방식을 적용하기 힘들다. 따라서 본 연구에서는 회의록 내에 다루고 있는 안건과 관련한 이슈와 문제점, 결정사항, 후속조치 등에 관한 내용으로 세부 요약문 작성하고 이를 종합한 주제 요약문을 작성한 후 회의 전체에 대한 최종 대표 요약문을 작성하는 상향식 방식의 요약문을 작성하도록 하였다. 회의록 대화의 상향식 요약 절차 및 하나의 안건에 대한 요약문의 구조는 다음과 같다.

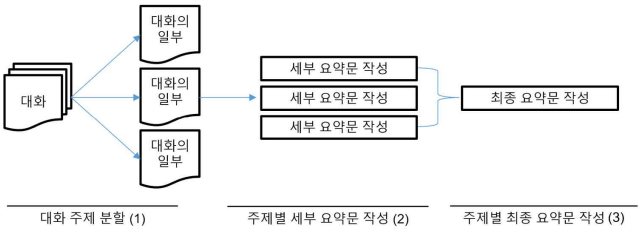

그림 2 논쟁적 대화의 상향식 요약 절차

먼저 한 회의록에서 논쟁의 대상이 되는 주제가 여러 개 나타날 경우 이를 주제별로 분할한 후 각 주제에서 논의의 대상이 되는 문제/이슈(Problems/Issues), 논쟁적 대화를 통해 도출된 결정사항(Decisions)과 결정된 내용의 후속조치나 과업사항(Actions)에 대한 각각의 세부 요약문을 작성한다. 이후 각각의 세부 요약문을 통합하여 핵심 정보를 추출한 후 주제 요약문을 작성하였다. 이때 주제 요약문은 논쟁적 대화의 시간 순서에 따라 취합하고, 세부 요약문의 내용이 중복될 경우 상위 개념으로 통합하여 기술하여 자연스러운 문장으로 작성한다. 각각의 요약문은 원본 회의 대화 중 어떤 발화들을 참조하여 구축되었는지를 명시하고, 해당 발화들의 아이디를 데이터세트에 포함한다.

마지막 단계에서는 각 주제들에 대한 주제 요약문들만을 작업자에게 제공하여 이를 하나로 묶어 회의 전체에 대한 대표 요약문을 작성한다.

3.3. 세부 요약문 구성

3.2장에서 기술한 것과 같이 쟁점별 세부 요약문은 문제/이슈(Problem/Issues), 결정사항(Decisions), 후속조치 및 과업 사항(Actions)으로 구성되어 있다. 문제/이슈 요약문은 논쟁적 대화에서 나타나는 요청사항, 문제제기, 제안 등의 내용이 포함되며 누가 어떤 문제를 제기하였는지, 또 근거가 드러난 경우 해당 근거를 문장에 포함하도록 한다. 결정사항 요약문은 논쟁적 대화를 통해 결정된 사항, 합의된 사항 또는 안건에 대한 동의/

비동의 등의 내용을 포함한다. 마지막으로 후속 조치 및 과업 사항 요약문은 논쟁적 대화를 통해 회의 참여자 또는 관계자가 수행하여야 할 과업이나 처리해야 하는 후속 조치를 포함한다.

마지막으로 세 가지 세부 요약문을 통합한 주제 요약문에는 안건, 이슈 및 논의사항, 결정사항을 모두 포함한 요약문으로 구성된다. 세부 요약문의 예시는 아래와 같다.

표 2 세부 요약문 유형별 예시

유형	세부 요약문 예시[4]
문제/이슈 (Problems /Issues)	전문위원은 먹는물관리법 일부개정법률안에 대해 주요 골자 및 수정 의견과 보완해야 할 사항들을 개괄하였다.
결정사항 (Decisions)	홍무18년 고려총통 국보지정 재심요청에 관한 청원은 보류되었다.
후속조치 (Actions)	의사일정 제6항 사회복지법인 특수학교의 학교법인 전환 청원에 대하여 교육부는 사후에 처리 경과를 소위원회에 보고할 것이다.

4. 실험

4.1. 실험 방법

본 연구에서는 생성 요약에 적합한 모델로 알려져 있는 BART 모델[19]의 한국어 오픈소스 버전[5]을 사용하여 학습을 수행하였다. 요약 모델은 BART에서 제공하는 트랜스포머의 인코더-디코더 기반 요약 방법을 그대로 사용하였으며, 각 발화를 구분하기 위해 발화와 발화 사이에 구분 기호(BART의 경우 〈/s〉 기호)를 사용하여 하나의 텍스트로 모델의 입력으로 사용하였다.

또한 [11]에서 알려진 바와 같이, 문서 요약 데이터로 요약 모델을 먼저 미세 조정(fine-tuning)하는 경우 요약의 품질이 좋아질 수 있다. 본 실험에서는 문서 요약 데이터를 사용하여 미세 조정된 모델을 추가로 사용하여 비교 실험하였다. 이를 위해 AI Hub의 문서 요약 데이터의 원문 데이터 40만 건(신문기사 30만 건, 기고문 6만 건, 잡지기사 1만 건, 법원 판결문 3만 건)을 활용하여 각각 추출요약 40만 건, 생성요약 40만 건, 총 80만 건의 요약문 데이터를 선정하였고, 이로부터 243,983건(3,282,214 문장, 50,717,051어절)의 요약 데이터를 추출하여 학습데이터로 사용하였다[6].

실험은 다음과 같이 진행되었다:

- KoBART: KoBART에 회의록 말뭉치 학습 후 평가
- KoBART+Doc: KoBART에 문서 요약 데이터 미세 조정 후 회의록 말뭉치 학습
- KoBART+speaker: KoBART에 화자 정보를 추가하여 회의록 데이터 학습
- KoBART+Doc+speaker: KoBART에 문서 요약 데이터 미세 조정 후 화자 정보를 추가하여 회의록 말뭉치 학습

이때 화자 정보를 추가하기 위해서 각 발화의 앞에 발화자의 이름을 특수기호(&)로 감싸 기입하여 하나의 문장으로 간주하였다. 이에 대한 예시는 아래와 같다.

표 3 화자 정보가 추가된 모델의 입력 예

모델명	입력 예시
KoBART	<s> 개정안은 안 5조의2 …
KoBART +speaker	<s> & 배용근 & 개정안은 안 5조의2 …

본 연구에서는 모델의 입력 길이 제한의 문제와 함께, 추출 요약은 수행하지 않는 조건에 따라 발화에서 요약에 관련 있는 중요한 발화들의 정보는 정답(ground truth)을 그대로 사용하여 요약에 관련 있는 발화들만 모델의 입력으로 사용하였다.

4.2. 데이터세트

실험 데이터세트로 200개 회의록에 등장한 887건의 주제(안건)에 대한 주제 요약문을 사용하였다. 실험을 위해 877건 중 709건을 무작위 선정하여 학습데이터로 사용하였고, 그 외 178건을 평가데이터로 사용하였다.

5. 평가

5.1. 정량적 평가

회의록 요약 모델을 평가하기 위해 4.1장에서 제시된 4개 모델에 대해 [11]의 방법론을 따라 ROUGE-1, ROUGE-2, ROUGE-SU4 평가를 수행하였다. 세 평가지표는 유니그램(unigram), 바이그램(bigram), 그리고 유니그램과 최대 4의 스킵(skip) 거리를 사용한 스킵-바이그램을 사용한다.

평가 결과는 표 4에서 확인할 수 있다.

표 4 국회 회의록 말뭉치로부터 자동으로 생성된 요약문에 대한 ROUGE-1, ROUGE-2, ROUGE-SU4 점수(F1)

모델	ROUGE-1	ROGUE-2	ROUGE-SU4
KoBART	22.7	15.7	13.6
KoBART +Doc	**23.0**	16.3	14.5
KoBART +speaker	20.5	13.1	11.4
KoBART +Doc+speaker	22.1	14.6	12.6

4) 국회회의록 “320교문(청원심사)소위01(13.11.18)”
5) https://huggingface.co/gogamza/kobart-base-v1
6) 해당 모델은 자체평가에서 ROUGE-1 34.50, ROUGE-2 21.51의 성능을 보였다.

정량적 평가에서는 문서 요약 데이터(뉴스)를 모델에 함께 학습하였을 때 성능 향상이 존재함을 확인할 수 있었다. 발화자의 이름을 추가할 경우 대화에서 어떤 발화자가 언급하였는지를 요약문에 명시하는 효과가 있지만 정량적 수치에서는 오히려 성능이 하락하였다. 발화자 정보를 요약 모델의 입력으로 사용하기 위한 보다 개선된 방법이 필요해 보인다.

5.2. 정성적 평가

정성적 평가는 ROUGE-1 성능이 가장 높게 나타난 KoBART+Doc 모델과, 가장 낮게 나타난 KoBART+speaker 모델의 결과를 바탕으로 진행하였다. 먼저 KoBART+Doc 모델에서는 정답에 나타나는 주제 및 키워드와 비교적 동일하게 요약문을 생성하고 있음을 확인할 수 있다. 즉, 입력 문장의 주제를 벗어나지 않고 표기와 형식에 오류가 없으며, 자연스러운 문장으로 잘 작성되어 있다. 표 5는 KoBART+Doc 의 생성 요약 예시 중 일부이다.

표 5 KoBART+Doc 생성 요약 예시

유형	내용
ground-truth	개발제한구역 내 축사 등 동식물 관련시설을 무단 용도변경한 위반행위자가 시정명령을 이행하겠다는 동의서를 제출하면 2020년 말까지 이행강제금 부과를 유예하려는 것에 대해 1년간 부과는 하되 징수를 1년간 유예하는 걸로 정리하였다. ... 제32항은 보다 심도 있는 심사를 위해서 소위원회에 계류하여 계속 심사하기로 하였다.
output	</s> 개발제한구역 내 축사 등 동식물 관련시설을 무단 용도변경한 위반행위자가 시정명령을 이행하겠다는 동의서를 제출하면 2020년 말까지 이행강제금 부과를 유예하려는 것에 대해 1년간 부과는 하되 징수를 1년간 유예하는 걸로 결정되었다. ... 제32항은 보다 심도 있는 심사를 위해서 소위원회에 계류하여 계속 심사하기로 하였다.</s>

그럼에도 불구하고 생성 요약 모델 특성상 일부 생성 요약 출력(output) 결과에서는 문장의 특정 부분을 반복하거나 입력 문장의 연속된 텍스트를 그대로 결과로 출력하는 치명적인 오류들도 다수 발견되었다. 이는 추상 요약을 생성하는 것이 까다로운 태스크이며, 추가적인 개선의 여지가 있음을 시사한다.

화자 정보를 추가하여 회의록을 학습한 KoBART+speaker 모델의 경우 표 6과 같이 생성 요약 결과에 전체 회의록에 대한 요약이나 화자별 발화에 대한 요약 대신 입력 문장에 등장한 인물의 이름이 포함된 부분만을 요약하는 오류가 나타났다. 즉, 화자 정보는 생성 요약에 유의미한 자질로 작용하지 않았으며, 오히려 '인명'에 주목하는 쪽으로 편향이 발생하여 모델의 성능을 떨어뜨리고 있다는 것을 확인하였다.

표 6 KoBART+speaker 생성 요약 예시

유형	내용
ground-truth	수석전문위원 류환민은 **윤명희 의원안**, ..., 정부안에 대해 타당하다는 검토의견을 보고했다. 입법조사관 장설희는 김우남 의원안과 **강석호 의원안**은 타당하고 ... 검토의견을 보고했다. ...
output	</s> **전문위원 윤명희**는 농촌진흥법이 개정되면서 기존 국유재산특례 관련 규정의 조문번호가 변경되었고... 입장을 밝혔고, **강석호 위원**은 ...특례 규정에 관한 내용의 동일성이 인정되므로...타당하다는 의견을 표명하였다.</s>

일반적으로 앞서 실험한 네 모델이 생성한 요약문은 치명적인 오류를 제외하고는 글의 유창성 측면에서 보았을 때 하나의 짜임새 있는 글로 느껴지나, 내용적인 측면에서는 사실관계가 일치하지 않기도 하였다. 또한 회의록의 요약 구조를 학습하여 형식적으로 출력하거나, 입력 문장을 그대로 참조하여 가져오거나, 같은 부분이 반복되는 등 추상 요약에 필수적인 패러프레이징을 수행하는 데 아직까지는 한계가 있음을 알 수 있었다.

표 7 생성 요약 오류 유형

유형	예시
사실 관계 불일치	</s> 전문위원 김상훈 의원이 대표발의한 송유관 절취방지 및 피해보상 등에 관한 ... 법률 일부개정법률안은 원안대로 가결되었다.</s> ⇒ 입력 문장에서는 가결되지 않음
요약문 형식(template) 출력	</s> 제천화재피해자지원 및 지원에 관한 법률 일부개정법률안과 관련하여 제천시가 무엇을 보고해야 되는지에 대한 **전문위원의 검토의견이 있었고**, ... 법률 일부개정법률안은 **원안대로 가결되었다**.</s> ⇒ 정답 요약문 작성 시 보편적으로 사용된 상투어를 그대로 생성 요약에 포함
입력 문장 베끼기	이미 취소를 하고 다시 환매를 다 한 다음에 또 산단을 지정하려면 또 다시 수용을 해 가지고 보상해야 하기 때문에 개정안에 대해서 수용하였다. ⇒ 입력 문장과 동일한 부분 출력
같은 부분 반복	</s> 의사일정 제1항부터 제14항까지 3건의 국제개발협력 기본법 일부개정법률안과 2건의 국제협력협력 기본법 일부개정법률안이 상정되었다...2건의 국제협력협력 기본법 일부개정법률안이 상정되었다. ....</s>

6. 결론

본 연구에서는 대한민국 국회 소위원회 회의록을 바탕으로 회의 요약에 대한 말뭉치를 구축한 후, 다양한 모델을 통해 검증하였다. 정량적 분석 결과 생성 요약 모델에 문서 요약 데이터셋을 활용해 미세 조정을 추가한

모델의 생성 요약 성능이 가장 좋은 것을 확인하였고, 결과의 정성적 분석을 통해 화자 정보 및 생성 요약 분야 자체에 관한 추가 연구 지점을 확인할 수 있었다. 국회 회의록 요약 말뭉치의 경우 기존의 유사한 회의 요약 말뭉치들에 비해 많은 참여자 수와 장시간의 발화, 그리고 다양한 주제를 포함하고 있다는 점에서 의의가 있으며, 향후 이를 통해 복잡한 회의 요약 혹은 생성 요약에 대한 연구가 진행될 수 있을 것으로 기대한다. 향후에는 화자 정보나 담화 구조, 논변 구조와 같이 회의록 요약에 있어 추가적인 자질로 사용할 수 있는 특징에 대한 연구를 진행할 예정이다.

사사

이 논문은 국립국어원의 ‘2021년 회의록 요약 말뭉치 연구 분석’(국립국어원 2021-01-14 / 발간등록번호 11-1371028-000864-01) 사업 수행 결과를 활용하여 작성되었습니다.
이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2022-0-00078, 의료지식 생성을 위한 설명가능한 추론 기술개발)

참고문헌

[1] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. “Abstractive text summarization using sequence-to-sequence rnns and beyond”. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280-290, 2016

[2] Abigail See, Peter J Liu, and Christopher D Manning. “Get to the point: Summarization with pointergenerator networks”. In Proceedings of the 55th Annual Meeting of the Association for Computational Lingistics, pages 1073-1083, 2017

[3] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. “Annotated gigaword”. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pages 95-100. Association for Computational Linguistics, 2012

[4] Paul Over and James Yen, “An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems”, In Proceedings of the Document Understanding Conference, 2004

[5] Katja Filippova and Yasemin Altun, “Overcoming the lack of parallel data in sentence compression”, In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2013

[6] Dimitrios Galanis, Ion Androutsopoulos, “A New Sentence Compression Dataset and Its Use in an Abstractive Generate-and-Rank Sentence Compressor”, In proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, 2011

[7] 국립국어원, 국립국어원 문서 요약 말뭉치(버전 1.0), https://corpus.korean.go.kr, 2020

[8] Adam Janin, et al. “The icsi meeting corpus”. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003

[9] Iain McCowan, et al. “The ami meeting corpus”, In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research,, 88:100, 2005

[10] Virgile Rennard, Guokan Shang, Julie Hunter, Michalis Vazirgiannis, “Abstractive Meeting Summarization: A Survey”, arXiv:2208.04163, 2022

[11] Chenguang Zhu, Ruochen Xu, Michael Zeng, Xuedong Huang, “A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining”, In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2020

[12] Zhong, Ming and Yin, Da and Yu, Tao and Zaidi, Ahmad and Mutuma, Mutethia and Jha, Rahul and Hassan Awadallah, Ahmed and Celikyilmaz, Asli and Liu, Yang and Qiu, Xipeng and Radev, Dragomir, “QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization”, In Proceedings of North American Association for Computational Linguistics (NAACL), 2021

[13] Yulong Chen, Yang Liu, Liang Chen and Yue Zhang, “DialogSum: A Real-life Scenario Dialogue Summarization Dataset”, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021

[14] Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, Dragomir Radev, “ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining”, In Proceedings of Association for Computational Linguistics 2021, 2021

[15] Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, In Proceedings of Association for Computational Linguistics 2021, 2021

[16] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, 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, Hong Kong, China. Association for Computational Linguistics, 2019

[17] Yulong Chen, Yang Liu, Liang Chen, Yue Zhang. “DialogSum: A real-life scenario dialogue summarization dataset”. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062-5074, 2021

[18] Zhao, Lulu, et al. “TODSum: Task-Oriented Dialogue Summarization with State Tracking”. arXiv preprint arXiv:2110.12680, 2021.

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, arXiv:1910.13461, 2019