"Chatbot Service Based on National Assembly Minutes"
국회 회의록 기반 챗봇 서비스

지도교수 이 건 호

숭실대학교 산업정보시스템공학과
유수정, 조예은, 조인영

요약

본 연구는 국회 회의록의 방대한 데이터 속에서 유용한 정보를 효과적으로 검색하고 활용하기 위한 챗봇 서비스를 설계하고 구현하는 것을 목표로 한다. 정보 검색과 자연어 처리 기술을 결합하여 검색 정확도와 응답 적합성을 향상시켰으며, 사용자의 정보 접근성을 강화함으로써 국회 회의록이 정책 결정과 시민 참여 과정에서 더 큰 역할을 할 수 있도록 기여하고자 한다.이를 위해 데이터의 주요 토픽과 발언 내용을 파악하기 위해 워드클라우드 시각화를 활용하고, LDA(Latent Dirichlet Allocation) 토픽 모델링과 네트워크 분석을 수행하였다. BM25 알고리즘과 BERT 모델을 결합한 검색 구조를 채택하여 초기 검색 효율성을 확보하고, 문맥적 재정렬을 통해 응답의 신뢰도를 높였다. 또한, Gradio UI를 활용한 사용자 친화적 인터페이스를 구현하여 비전문가도 쉽게 활용할 수 있도록 지원하였다. 본 시스템은 공공 데이터의 가치와 활용성을 극대화하고, 민주적 정보 접근성을 향상시키는 데 기여할 것으로 기대된다.

목차

1. 서론
1.1 연구 개요 ---------- 3
1.2 연관 연구 ---------- 3

2. 문제 해결에 사용한 기법, 알고리즘
2.1 문제 정의 ---------- 4
2.2 알고리즘 소개 ---------- 4

3. 데이터 분석
3.1 데이터 소개 ---------- 7
3.2 데이터 전처리 ---------- 8
3.3 데이터 분석 ---------- 8

4. 서비스 구현
4.1 인공지능 기반 질의응답 서비스 ---------- 12
4.2 결과 해석 ---------- 14
4.3 한계점 및 개선사항 ---------- 14

5. 결론 및 향후 연구 방향
5.1 결론 ----------15
5.2 향후 연구 방향 ---------- 15

6. 참고문헌 ---------- 15

1. 서론

1.1 연구 개요

최근 인공지능(AI)과 자연어 처리(NLP) 기술이 급격히 발전하면서, 법률 및 공공 데이터를 효과적으로 활용하려는 시도가 활발히 이루어지고 있다. 특히, 법률과 정책 관련 데이터를 체계적으로 정리하고 분석해 사용자에게 정보를 제공하는 리걸테크(LegalTech) 분야는 효율성과 접근성을 높이기 위해 빠르게 성장하고 있다. 이러한 기술은 공공 데이터를 활용한 다양한 서비스로 이어지고 있으며, 국회 회의록과 같은 공공 기록은 민주주의의 투명성을 강화하고 시민의 알 권리를 충족시키는 데 핵심적인 역할을 하고 있다.[1]

국회 회의록은 국가의 주요 의사결정 과정과 정책 논의를 기록한 중요한 자료로, 연구자, 정책 입안자, 그리고 일반 시민에게 유용한 정보를 제공할 수 있는 잠재력을 가지고 있다. 국회도서관은 2021년부터 '국회회의록 빅데이터' 서비스를 통해 16대부터 21대 국회까지 약 2만 건의 회의록을 의원 별 발언 단위로 분리하여 총 1,200만여 건의 데이터를 제공하고 있다. 하지만 방대한 분량과 전문적이고 복잡한 언어로 작성되어 있어 원하는 정보를 효과적으로 탐색하고 이해하는 데 큰 어려움이 있다. 현재 국회 회의록은 발언자나 키워드 중심의 단순 검색 도구로 접근할 수 있지만, 이는 사용자가 자연어로 복잡한 질문을 입력했을 때 정확하고 맥락에 맞는 답변을 제공하지 못하는 한계를 가지고 있다. 특히, 시민들은 특정 사안에 대한 논의 내용을 찾기 위해 많은 시간을 소비해야 하며, 정책 입안자들은 필요한 정보를 신속히 확인하지 못해 정책 결정 과정에서 비효율성을 겪고 있다. 이를 위해 국내외적으로도 정부 데이터를 활용한 Q&A 서비스와 챗봇 개발 사례가 늘어나고 있지만 한계점이 명확히 드러나고 있다.[2]

또한, 행정안전부는 2024년 3월부터 정부와 자치단체를 대상으로 AI 기반의 자동회의록 작성 및 문서 인식 기능을 갖춘 행정업무 효율화 서비스를 시범 운영하고 있다. 이 서비스는 회의 내용을 녹화, 녹음한 파일로부터 문자를 자동 추출한 뒤 시간순으로 참석자와 회의 내용을 정리하여 보고서 형태로 제공하고 있다. 1시간 분량의 회의 영상 또는 음성 파일을 약 5분 만에 처리할 수 있어 회의록 작성에 소요되는 시간과 노력을 크게 줄일 수 있다.[3]

이러한 국내 사례들은 공공 데이터를 활용한 AI 기반 서비스의 가능성을 보여주고 있으며, 국회 회의록 데이터를 활용한 질의응답(Q&A) 서비스 개발은 공공 데이터의 가치를 높이고, 이를 시민과 정책 입안자가 더욱 효과적으로 활용하도록 지원할 수 있는 중요한 과제가 되고 있다.

본 연구는 국회 회의록 데이터를 기반으로 자연어 처리 기술과 정보 검색 기술을 결합한 챗봇 서비스를 개발하여 사용자가 원하는 정보를 신속하고 정확하게 제공받을 수 있는 시스템을 구현하는 것을 목표로 하고 있다. 이를 통해 국회 회의록의 활용성을 극대화하고, 정보의 비대칭성을 해소하며, 공공 데이터의 민주적 가치를 실현하는 데 기여하고자 한다.

1.2 연관 연구

본 연구에서, 문제를 다루기에 앞서 보다 정확한 연구를 진행하기 위해 관련성 있는 선행 연구들을 검토하였다. 이들 연구는 자연어 처리, 딥러닝 모델, 공공 데이터를 활용한 질의응답 시스템 개발에 있어 중요한 기술적 기반을 제공한다.

KoBERT와 KoGPT2 기반의 대형언어모델과 딥러닝을 통합한 리뷰 유용성 예측 모형에 관한 연구는 대형언어모델(LLM)과 딥러닝 기법의 통합적 활용

가능성을 보여준다. 해당 연구는 리뷰 데이터를 바탕으로 KoBERT, KoGPT2 모델과 CNN, LSTM 등의 딥러닝 모델을 결합하여 예측 성능을 극대화한 사례를 제시하고 있다. 연구 결과, 리뷰 유용성 예측 모형은 76.37%의 높은 성과를 나타냈으며, KoBERT와 KoGPT2 모델은 한국어 텍스트 데이터를 분석하고 문맥을 이해하는 데 있어 뛰어난 성능을 보였다. 이는 국회 회의록과 같은 방대한 한국어 텍스트 데이터를 효과적으로 활용할 수 있는 기술적 가능성을 시사하며, 본 연구에서 자연어 처리 모델의 활용성을 탐색하는 데 중요한 사례가 된다.[4]

딥러닝 기반의 세법 관련 질문 유형 분류 모델 연구는 세법 질문 데이터를 기반으로 KTL-BERT 모델을 활용하여 질문 유형을 분류하는 모형을 제안하고 있다. 국세청 홈택스의 상담 사례 데이터를 수집 및 전처리한 뒤, 법인세, 부가가치세, 양도소득세 등으로 분류하여 KTL-BERT 모델을 통해 질문 유형 분류 성능을 평가하였다. 연구 결과, 기존 벤치마크 모델 대비 높은 성능과 효율성을 나타냈으며, 딥러닝 기반의 자연어 처리 기술이 법률 및 공공 데이터 활용에서 중요한 역할을 할 수 있음을 입증하였다. 본 연구의 결과는 국회 회의록 데이터를 기반으로 질문-답변 시스템을 설계하는 과정에서 데이터 분류 및 처리 기술에 유용한 통찰을 제공하였다.[5]

마지막으로, 질의응답 기술과 법률 상담 도우미 시스템의 응용에 관한 연구는 정보 검색 및 질의응답 기술의 실용적 활용 가능성을 보여준다. 이 연구에서는 지식 기반 질의응답(knowledge-based QA)과 검색 기반 질의응답(search-based QA) 기법을 통해 콜센터 상담 도우미 시스템을 설계하였다. 지식 기반 질의응답은 온톨로지를 통해 구조화된 데이터를 활용하여 추론 과정을 거쳐 응답을 도출하고, 검색 기반 질의응답은 대규모 문서 집합에서의 검색과 순위화를 통해 답변을 제공한다. 이러한 기술은 실시간으로 사용자 질의에 응답할 수 있는 시스템 설계에 중요한 사례를 제시하며, 국회 회의록 기반 챗봇 서비스의 실시간 정보 검색 및 응답 기능 개발에 활용될 수 있다.[6]

2. 문제해결을 위한 알고리즘

2.1 문제 정의

현재 국회 회의록 빅데이터 서비스는 키워드 검색과 발언자별 데이터 분리 등 정보를 제공하는 데 초점을 맞추고 있지만, 사용자 질의의 맥락을 고려하거나 심층적인 질의응답을 지원하지 못하는 한계가 있다. 사용자가 원하는 정보에 접근하기 위해서는 복잡한 검색 조건을 조합하거나 다수의 문서를 직접 검토해야 하므로, 정보 검색의 효율성과 접근성이 떨어진다.

본 연구는 이러한 한계를 극복하기 위해 자연어 처리 기술과 정보 검색 기법을 결합하여, 사용자의 질문에 대해 관련성이 높은 국회 회의록 요약 결과를 제공하는 검색 기반 요약 서비스를 설계하고자 한다. 이를 통해 방대한 회의록 데이터에서 원하는 정보를 효율적으로 탐색할 수 이도록 지원하며, 국민과 국회 간 소통을 강화하고 국회의 입법 활동 투명성을 제고하는 것을 목표로 한다.

2.2 알고리즘 소개

2.2.1 워드클라우드 (Word Cloud)

워드클라우드는 텍스트 데이터에서 단어의 빈도를 시각적으로 표현하는 기법으로, 단어의 크기와 위치를 통해 데이터의 중요한 특징을 강조하는 도구이다. 자주 등장하는 단어일수록 더 큰 글자로 표시되며, 사용자가 데이터를 직관적으로 이해할 수 있도록 돕는다. 또한, 대량의 텍스트 데이터에서 키

워드를 빠르게 파악할 수 있는 유용한 방법으로, 특히 텍스트 분석 및 자연어 처리(NLP)분야에서 자주 활용된다.

[사진1. 워드클라우드 예시]

위는 “미세먼지+ 환경”을 키워드로 설정하여 기사 데이터를 크롤링한 후, 해당 데이터에서 단어의 빈도수를 기반으로 워드클라우드를 생성한 결과이다.

2.2.2 LDA(Latent Dirichlet Allocation)

Latent Dirichlet Allocation은 Blei et al.(2003)이 제안한 확률적 주제 모델로, 대규모 문서 집합에서 숨겨진 주제를 추출하는 데 널리 사용된다.
LDA는 문서가 다수의 주제로 구성되어 있으며, 각 주제는 단어 분포로 표현될 수 있다는 가정을 기반으로 한다.
LDA의 주요 구성 요소는 아래와 같다.

- 토픽 분포(Topic Distribution, Θ): 각 문서가 주제를 포함하는 확률 분포, Θ∼Dirichlet(α)
- 단어 분포(Word Distribution, ϕ) : 각 주제가 단어를 포함하는 확률 분포, ϕ∼Dirichlet(β)
- 단어 할당(Word Assignment, z): 문서 내 각 단어가 속하는 주제, z∼Multinomial(Θ)
- 단어 생성(Word Generation): 주제 z에 따라 단어 w 가 w∼Multinomial(ϕz)로 생성

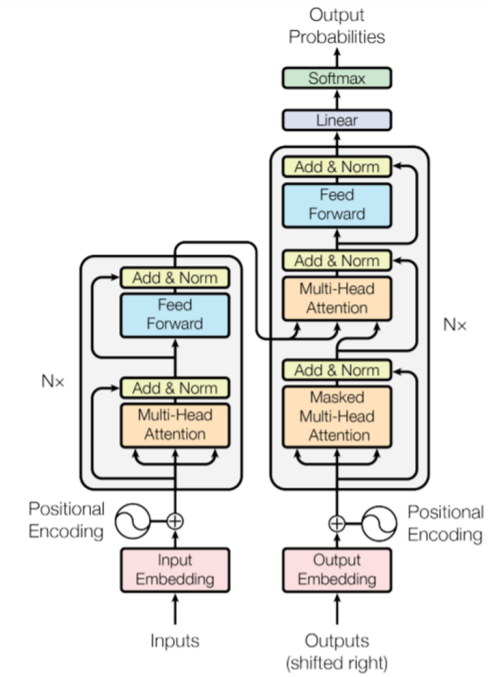
LDA는 문서 집합에서 숨겨진 주제를 학습하기 위해 변분 베이즈 방법 또는 깁스 샘플링과 같은 최적화 기법을 사용한다. 학습 과정에서 LDA는 문서의 토픽 분포와 단어 분포를 반복적으로 업데이트하며, 주어진 문서와 단어에 가장 적합한 확률 분포를 추정한다.
LDA는 방대한 텍스트 데이터에서 주제를 탐지하고, 문서 간 유사성을 분석하는 데 효과적이다.

2.2.3 네트워크 분석(Network Analysis)

네트워크 분석(Network Analysis)은 그래프 이론을 기반으로 한 분석 방법으로, 객체들 간의 관계를 모델링하고 이들의 상호작용을 분석하는 데 사용된다. 네트워크 분석은 주로 소셜 네트워크 분석, 웹 구조 분석, 금융 거래 네트워크, 추천 시스템, 물류 최적화 등 다양한 분야에서 중요한 역할을 한다. 네트워크 분석은 주로 노드(Node)와 간선(Edge)로 구성된 그래프 구조를 기반으로 하며, 네트워크의 특성과 구조를 파악하기 위해 여러 알고리즘을 적용한다.

2.2.4 KoBERT

KoBERT는 sk텔레콤에서 개발한 BERT 기반의 사전 학습 언어 모델로, 한국어 자연어 처리(NLP) 작업에 최적화되어 있다. BERT 모델의 구조를 기반으로 하며, 대용량의 한국어 말뭉치로 사전 학습되었다.
KoBERT는 BERT와 동일한 트랜스포머 기반 아키텍처를 사용하며, 양방향으로 문맥 정보를 학습한다.



[사진2. KoBERT 모델 구조]

한국어 문법적 특성과 토큰화를 반영하여 학습되었으며, 단어의 맥락을 효과적으로 파악할 수 있다. KoBERT는 텍스트 분류, 질의응답, 감성 분석, 문서 요약 등 다양한 NLP 작업에서 활용되며, 특히 한국어 특화 작업에서 높은 성능을 보여준다.

2.2.5 BM25

BM25(Best Matching 25)는 정보 검색(IR) 시스템에서 문서와 쿼리 간의 유사도를 측정하는 데 사용되는 가중치 기반 순위 알고리즘이다. BM25는 TF-IDF(Term Frequency-Inverse Document Frequency)모델을 확장하여, 단어의 빈도뿐만 아니라 문서의 길이나 특정 단어의 중요도를 고려하여 보다 정교한 순위 매기기를 수행한다.

BM25는 각 문서와 쿼리 간의 점수를 계산할 때, 쿼리의 각 단어가 문서에서 얼마나 중요한지 평가한다. 점수는 다음과 같은 수식으로 계산된다.

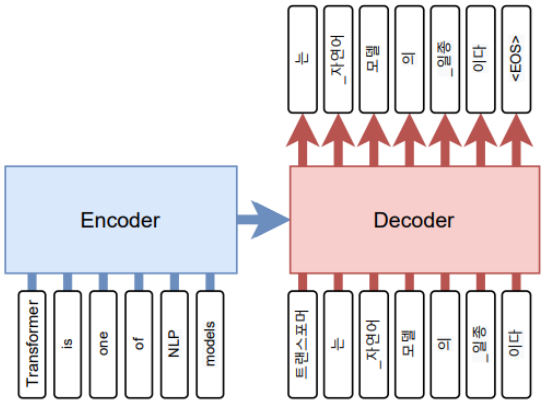
$$\text{score}(D,Q) = \sum_{i=1}^{|Q|} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avg|D|}\right)}$$

- f(qi ,D): 문서 D에서 단어 qi의 빈도
- |D|: 문서 D의 길이
- avg|D|: 모든 문서의 평균 길이
- k1, b: 하이퍼파라미터

BM25는 주로 검색 엔진, 추천 시스템, 문서 분류 등 다양한 정보 검색 분야에서 사용된다. 특히, 웹 검색에서 쿼리와 문서 간의 관련성을 평가할 때 강력한 성능을 보이며, 빠르고 효율적인 순위 매기기를 제공한다.

2.2.6 KoBART(Korean BART)

KoBART(Korean BART)는 한국어 자연어 처리(NLP) 작업을 위한 사전 훈련된 언어 모델로, BART(Bidirectional and Auto-Regressive Transformers)모델을 기반으로 한다. BART는 Facebook AI에서 개발된 모델로, 텍스트 생성 및 텍스트 변환 작업에 강력한 성능을 보이며, KoBART는 이 모델을 한국어에 맞게 조정한 변형 모델이다. KoBART는 인코더-디코더 구조를 사용하여 입력 텍스트를 양방향으로 이해하고, 이 정보를 바탕으로 텍스트를 생성하는 데 사용한다.



[사진3. KoBART 모델 구조]

특히 한국어 텍스트 요약, 질문-답변 시스템, 문장 완성과 같은 다양한 한국어 처리 작업에 뛰어난 성능을 보여준다. KoBART는 BERT와 GPT의 장점을 결합하여, 양방향 이해와 텍스트 생성을 동시에 처리할 수 있어, 다양한 자연어 처리 응용 분야에서 유용한 모델로 자리잡았다.

KoBART의 활용 예시로는 한국어 뉴스 기사 요약, 고객 서비스 챗봇, 한국어 문서에서의 정보 추출 등이 있다.

3. 데이터 분석

3.1 데이터 소개

AI-Hub에서 제공하는 '국회 회의록 기반 지식 검색 데이터'는 국회 회의록을 활용하여 인공지능 학습용 질의응답(Q&A) 데이터셋을 구축한 것이다. 이 데이터는 국회 회의록에서 발언을 질문과 답변으로 분리하여 총 44,033쌍의 질의응답 데이터를 포함하고 있다. 데이터는 15대부터 21대 국회까지의 회의록 11,827건을 기반으로 구성되어 있으며, 국정감사, 본회의, 소위원회 등 다양한 회의 유형을 포함한다. 질문 유형은 추출형과 단답형으로 구분되며, 각각 91.82%와 8.18%의 비율을 차지한다. 데이터셋은 국회 회의록 기반 챗봇 서비스 개발, Legal Tech 서비스, 법안 심사 및 논의 내용 분석 등 다양한 응용 가능성을 가지고 있다. 이 데이터는 지능형 입법활동 지원과 인공지능 연구 및 응용 서비스 개발에 기여하며, 국민의 정치에 대한 관심과 참여를 제고할 것으로 기대된다. 메타데이터 구조는 다음과 같다.

No.	항목명	타입	항목 설명
1	filename	string	원천데이터의 이름
2	original	string	회의록 원문의 URL
3	id	string	데이터의 고유 ID
4	date	string	회의가 열린 날짜
5	conference number	number	회의의 고유 번호
6	question number	number	질문의 고유 번호
7	meeting name	string	회의의 이름
8	generation number	string	국회의 대
9	committee name	string	위원회 명칭
10	meeting number	number	회의의 회 수
11	session number	number	회의의 차 수
12	agenda	string	회의 안건 내용
13	law	string	회의와 관련된 법안
14	qna type	string	질문 유형 (추출형/생성형)
15	context	string	전체 발언 원문
16	context learn	string	모델 학습을 위한 컨텍스트
17	context summary q	string	요약된 질문 내용
18	context summary a	string	요약된 답변 내용
19	questioner name	string	질문을 발언한 사람의 이름
20	questioner ID	number	질문자의 고유 ID
21	questioner ISNI	number	질문자의 국제표준 ID
22	questioner affiliation	string	질문자의 소속 기관 또는 정당
23	questioner position	string	질문자의 직위
24	question tag	number	질문 발언의 순서
25	question comment	string	질문 발언의 내용
26	question keyword	string	질문 발언의 키워드
27	answerer name	string	응답자의 이름
28	answerer ID	number	응답자의 고유 ID
29	answerer ISNI	number	응답자의 국제표준 ID
30	answerer affiliation	string	응답자의 소속 기관 또는 부처
31	answerer position	string	응답자의 직위
32	answer tag	Number	답변 발언의 순번
33	answer comment	String	답변 발언의 내용
34	answer keyword	string	답변 발언의 키워드

[표1. 메타데이터 구조]

3.2 데이터 전처리

3.2.1 워드클라우드, LDA

워드 클라우드와 LDA 토픽 모델링은 텍스트 데이터의 핵심 정보를 시각화 하거나 주제를 도출하는데 유용하다. 이 과정에서 데이터의 품질을 높이기 위해 전처리가 필수적이다. 다음 과정에 따라 전처리를 진행하였다.

1. 데이터 정제 : 데이터 정제 과정에서는 텍스트 데이터를 분석에 적합한 상태로 만들기 위해 불필요한 요소들을 제거한다.

   원본 텍스트

   "감사합니다, 위원장님. 예..이번 회의에서 논의될 법안은 중소기업기본법(일부 개정 법률안)입니다. 추가적으로, 자료를 보시죠."

   정제 후 텍스트

   "감사합니다 위원장님 예 이번 회의에서 논의될 법안은 중소기업기본법 일부개정법률안 입니다 추가적으로 자료를 보시죠"

2. 토큰화 (명사추출) : 토큰화는 텍스트 데이터를 분석에 유용한 단위(토큰)로 나누는 과정이다. 주제별로 명확한 논의가 이루어지는 텍스트에서는 명사 추출이 유용하다. koNLPy의 Okt를 이용하여 텍스트에서 명사만 추출하였다.

   명사 추출 결과

   [“위원장님”, “이번”, ”회의”, ”논의”, “법안”, "중소기업기본법", "일부개정법률안", "자료”]

3. 불용어 제거 : 불용어는 분석에 큰 의미가 없는 단어로, 텍스트에서 제거하여 중요한 단어만 남기는 과정이다. 기본 불용어 리스트를 사용하여 자주 등장하지만 의미 없는 단어와, 국회 회의록에 특화된 불용어를 추가로 정의하여 제거하였다.

   (예: “위원장님”,“이번”,“회의”,논의”,”자료”)

   불용어 제거 결과

   ["법안", "중소기업기본법", "일부개정법률안"]

3.2.2 BM25(검색 기반 알고리즘)

쿼리와 문서 간의 관련성을 평가하기 위해 텍스트 데이터를 수치형 벡터로 변환하는 단어 임베딩 과정을 거쳤다. 이를 위해 한국어에 최적화된 BERT 기반 모델인 KoBERT 모델을 활용하여 각 문서와 쿼리의 텍스트를 벡터 형태로 변환했다.

원본 텍스트

“민주평통이 새로운 결의를 가지고 출발하는 것 같다 하는 생각이 들어서 다행입니다. “

단어 임베딩 후 결과

tensor([-1.4002e-01,-2.8664e-01,...,-2.0858e-01, 5.0086e-02, 1.9962e-02]

3.3 데이터 분석

3.3.1 워드클라우드

각 회의록 문서의 주요 단어들을 시각적으로 분석하기 위해 워드클라우드 기법을 적용하였다. 회의의 유형인 국정감사, 본회의, 소위원회, 예산결산특별위원회, 특별위원회의 5종류의 문서로 나누어 각 문서 유형에서 자주 등장하는 단어를 추출하였다.

1. 국정감사

[그림1. 국정감사 워드클라우드]

워드 클라우드에서 "상황", "현재", "문제", "개선" 등의 단어는 각 부의 정책 집행 상태 점검, 문제 지적, 개선 방향을 제안하는 국정 감사의 전형적인 역할과 일치함을 확인할 수 있다. 또한 "경제", "지원", "계획" 같은 단어들이 강조되는 것을 볼 때, 회의록에서 경제 정책의 성과와 지원 방안에 대한 논의가 활발히 이루어졌을 가능성이 있음을 유추할 수 있다.

2. 본회의

[그림2. 본회의 워드클라우드]

"국민", "현장", "국가", "해결", "지원" 등의 단어들이 크게 나타나며, 국민의 삶과 국가적 역할에 대한 논의가 본회의의 주요 초점임을 보여준다.

3. 소위원회

[그림3. 소위원회 워드클라우드]

"예산", "지원", "자원" 등의 단어가 크게 나타나며, 소위원회의 주요 논의가 예산 배정과 사용의 타당성, 지원 정책, 그리고 자원 관리 및 배분에 초점이 맞추어져 있음을 시사한다. "법", "기준" 등의 단어들은 소위원회가 특정 정책이나 예산 집행에 대해 법적 타당성을 검토하고, 기준에 맞는 집행을 점검하고자 한 것으로 유추할 수 있다.

4. 예산결산특별위원회

[그림4. 예산결산특별위원회 워드클라우드]

"예산", "경제", "상황" 등의 단어들이 강조되어 있어, 이 회의에서 예산이 경제적 목표를 달성하는 데 얼마나 효과적으로 쓰였는지, 그리고 현재 상황에서 예산의 집행이 어떤 영향을 미치고 있는지에 대한 논의가 주요한 초점이었음을 알 수 있다. "경제", "지원", "국민" 등의 단어들은 예산이 국민 생활과 경제 전반에 미치는 영향을 평가하는 데 중점을 두었음을 시사한다.

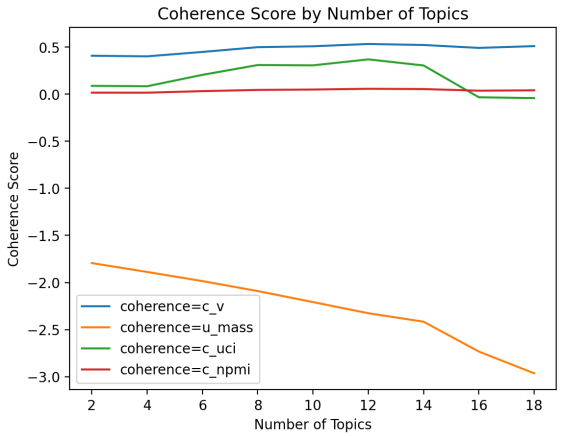
5. 특별위원회

[그림5. 특별위원회 워드클라우드]

"상황", "문제", "후보자", "개선" 등의 단어가 크게 보이는 것을 통해, 특별위원회에서는 문제 상황의 파악과 이에 대한 개선 방안 마련에 집중했음을 알 수 있다. 특히, 공직 후보자의 자격 검증과 관련된 논의도 중요한 역할을 한 것으로 유추할 수 있다.

3.3.2 LDA

국회 회의록 데이터에 LDA(Latent Dirichlet Allocation)모델을 적용하여, 문서에서 반복적으로 등장하는 주요 주제를 식별하였다.

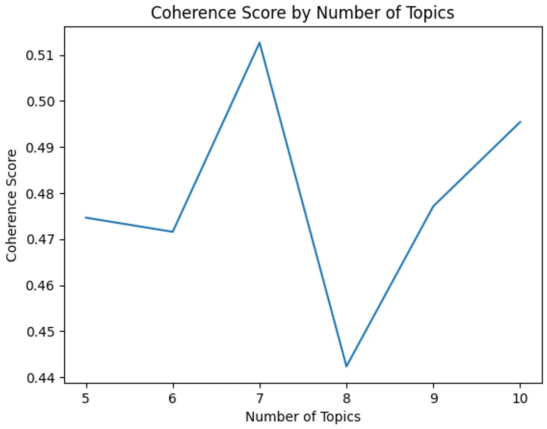


[그림6. Coherence score]

LDA 모델의 최적 주제 수를 결정하기 위해 coherence score를 사용하여 하이퍼 파라미터 튜닝을 진행하였다. coherence score는 주제의 품질을 평가하는 지표로, 높은 값은 주제가 의미 있고 일관성이 있음을 나타낸다. 여러 coherence metric(c_v, u_mass, c_uci, c_npmi)을 사용하여 모델을 평가하였다.

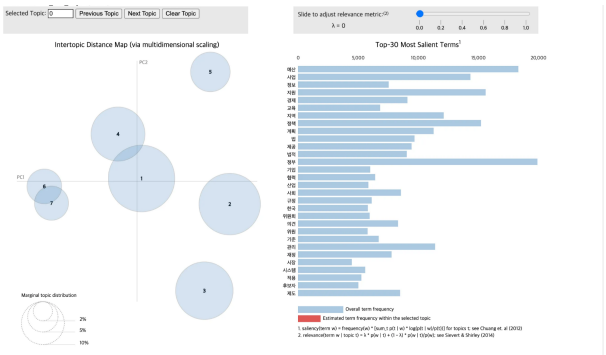
- c_v: 가장 널리 사용되는 coherence metric으로, 단어의 상관관계를 기반으로 주제의 품질을 평가
- u_mass: 문서 내 단어의 공기확률을 기반으로 coherence를 계산
- c_uci: 단어들의 상호 정보를 바탕으로 coherence를 평가
- c_npmi: 상호 정보 기반의 coherence로, 주제 간의 유사성을 측정

주제 수의 범위를 2부터 20까지 각 주제 수에 대해 모델을 학습한 후, 각 coherence metric에 대해 coherence값을 계산하였다. 그 결과, c_v metric이 가장 높은 coherence 값을 보였고, 이에 따라 c_v metric을 기준으로 주제 수 범위를 다시 설정하고, 추가적인 튜닝을 진행하였다.



[그림7. 최적의 토픽 수 결정]

튜닝 결과, 주제 수가 7일때 가장 높은 coherence score을 보였다. 주제 수를 7로 설정한 LDA 모델을 학습시키고, 각 주제에 대한 주요 단어들을 시각화 하였다.

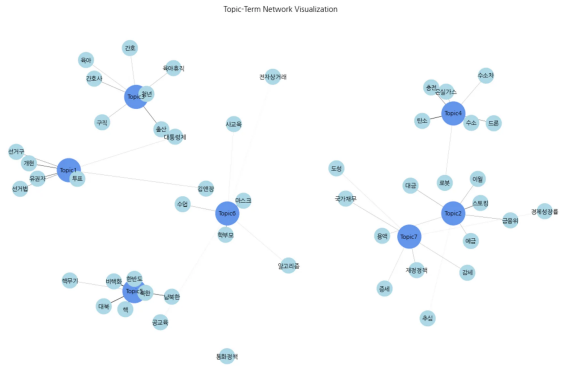


[그림8. LDA 결과]

Intertopic Distance Map은 각 주제의 관계를 2D 공간에서 시각화하여, 주제들 간의 거리가 가까운 경우 유사한 내용을 다루고 있음을 나타낸다. 특히 주제 1은 다른 주제들과 상대적으로 가까운 위치에 배치되어, 주요 중심 주제로 해석될 수 있다. Top-30 Most Salient Terms 그래프에서 '예산', '사업', '정보' 등 핵심 단어들이 각 주제의 주요 내용을 잘 반영하고 있음을 확인할 수 있다.

3.3.3 네트워크 분석(Network Analysis)

LDA분석을 통해 국회 회의록 데이터를 7개의 주제(topic)으로 분류하였다. LDA 분석 결과, 각 주제에서 상위 7개의 핵심 단어를 추출하여, 이를 기반으로 네트워크 분석을 진행하였다.



[그림9. 네트워크 시각화]

- Topic 1: 정치와 선거
- 주요 단어: 개헌, 투표, 선거구, 유권자, 김앤장, 선거법, 대통령제
- Topic 2: 금융과 채권
- 주요 단어: 스토킹, 이월, 대금, 예금, 용액, 금융위, 추심
- Topic 3: 출산과 육아
- 출산, 간호사, 육아휴직, 청년, 육아, 구직, 간호
- Topic 4: 환경과 기술
- 수소, 드론, 수소차, 온실가스, 로봇, 탄소, 충전
- Topic 5: 북한과 비핵화
- 비핵화, 북한, 핵, 대북, 남북한, 핵무기, 한반도
- Topic 6: 교육과 사교육
- 수업, 알고리즘, 사교육, 마스크, 공교육, 학부모, 전자상거래
- Topic 7: 경제와 재정 정책
- 국가채무, 감세, 재정정책, 증세, 경제성장률, 통화정책, 도성

각각의 토픽이 핵심 단어와 유의미하게 연관되어 있고, 핵심 단어들이 높은 연관성을 갖음을 확인할 수 있다.

[그림10. Gradio 인터페이스 구현]

3.3.4 국회 회의록 summary

대량의 국회회의록 문서 데이터셋을 활용해 요약 작업을 수행하였다. 대량의 텍스트 데이터를 효율적으로 요약하기 위해 한국어 특화 사전학습 모델인 KoBART를 사용하여 텍스트 데이터를 배치 단위로 요약하였다. Hugging Face의 transformers 라이브러리를 활용하여 gogamza/kobart-base-v2모델과 토크나이저를 로드하였다. 주요 설정은 다음과 같다.

- max_length : 512
- num_beams : 4
- length_penalty : 2.0
- temperature : 0.7
- top_k : 50
- top_p : 0.92

결과적으로 원문에 대한 요약이 다음과 같이 생성되었다.

원본 텍스트

"더군다나 사업비는 전액 공단이 부담하면서 관리감독은 할 수 없고 단지 수탁자인 철도청하고 협의 조정만 할 수 있다는 협약이 있다."

KoBART 요약 결과

"공단이 부담하면서 관리감독은 할 수 없고 없고 수탁자인 철도청하고 협의 협의 조정만 가능하다다다다"

요약문의 품질을 평가하기 위해 ROUGE(Recall-Oriented Understudy for Gisting Evaluation) 지표를 활용하였다. ROUGE는 자연어 처리, 특히 텍스트 요약의 품질의 평가하기 위해 사용되는 지표로 요약된 텍스트가 원문 텍스트와 얼마나 유사한지를 측정하며, 주로 n-gram의 중복 여부를 기반으로 계산된다.

- Precision : 모델이 생성한 요약 중 원문 텍스트와 일치하는 단어 비율

$$Precision = \frac{\text{일치하는 n-gram 수}}{\text{생성된 n-gram 수}}$$

- Recall : 원문 텍스트에서 모델 요약이 일치하는 단어 비율

$$Recall = \frac{\text{일치하는 n-gram 수}}{\text{참조 n-gram 수}}$$

- F1 Score : Precision과 Recall의 조화평균으로, 두 값의 균형을 평가

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

ROUGE-1	ROUGE2	ROUGE-L
0.4261	0.0235	0.3549

[표2. Rouge score]

KoBART모델이 생성한 요약은 단어 수준에서 어느 정도 성능을 보였으나, 문장 구성과 2-gram 유사성에서는 한계를 보였다. 특히, 요약 결과에서 불필요한 단어의 반복과 문장 간 일관성 및 구조적 흐름의 부족이 드러났다. 이러한 결과는 원본 텍스트가 일반적인 대화체가 아닌 국회에서 발언자들의 말을 기록한 텍스트이기 때문에, 사전 학습된 KoBART 모델이 이러한 특수한 언어적 스타일이나 문체를 충분히 반영하지 못했을 가능성이 있다. 따라서 모델이 국회 발언체의 특성에 맞는 문맥 이해와 요약을 수행하는 데 어려움을 겪었을 것으로 보인다. 이를 개선하기 위해 문장 간 일관성과 구조적 흐름을 강화하고, 모델의 문맥 파악 능력을 향상시키는 작업이 필요하다고 판단하였다.

4. 서비스 구현

4.1 인공지능 기반 질의응답 서비스

4.1.1 Summary fine tuning

국회 회의록 데이터를 요약한 결과를 더욱 매끄럽고 자연스럽게 정제하기 위해 GPT-2 모델을 Fine-Tuning하여 활용하였다. 요약 결과를 단순한 추출형 텍스트가 아닌 보다 자연스러운 문장으로 변환함으로써 사용자의 가독성과 이해도를 향상시키는 것을 목표로 한다.

생성형 언어 모델인 GPT-2는 사전 학습을 통해 대규모 텍스트 데이터를 기반으로 자연스러운 문장을 생성할 수 있으며 Fine-Tuning을 통해 국회 회의록 데이터에 최적화된 언어 생성 능력을 발휘할 수 있다. Bart 모델로 생성된 요약문과, 원 데이터에 존재하는 summary를 context 로 결합하여 모델을 학습하였다.

KoBART 요약 결과

"공단이 부담하면서 관리감독은 할 수 없고 없고 수탁자인 철도청하고 협의 협의 조정만 가능하다다다다"

GPT-2 정제 결과

"사업비를 공단이 부담하면서 관리감독은 할 수 없고 수탁장인 철도청과 협의 조정만 가능하다다"

또한 전 데이터에 중복단어가 존재하는 경우가 많아 규칙기반 텍스트 처리를 통해 불필요한 공백과

중복 단어 등 을 제거하였다. 이를 통해 후속 검색 시스템에서 더 나은 질의응답 (Q&A) 성능을 기대할 수 있다.

4.1.2 BM25 + Bert 검색기반

요약문이 정제된 후, 사용자가 입력하는 질의에 대해 가장 적절한 회의록 내용을 검색하고 응답하는 검색기반 질의응답(Q&A) 서비스를 구현하였다. 이 서비스는 BM25 알고리즘과 BERT 모델을 결합하여 검색 성능을 최적화하였다. 검색 시스템 설계 과정은 다음과 같다.

1단계 BM25를 통한 초기 검색: 사용자 질의(예:인천공항, 납세, 전기자전거 등)를 기반으로 국회 회의록 데이터에서 초기 검색 결과(상위 10개 문서)를 추출한다.

2단계 BERT를 통한 재정렬: BM25로 검색된 문서와 사용자 질의 간의 문맥적 유사도를 평가하여 최종 순위를 재정렬한다. 입력 데이터는 질의와 문서의 쌍으로 구성되며, 모델은 각 쌍에 대해 점수를 산출한다.

4.1.3 Gradio ui 인터페이스

BM25와 BERT를 결합한 검색 시스템을 사용자 친화적인 형태로 제공하기 위해 Gradio 기반 웹 인터페이스를 구현하였다. 이를 통해 사용자는 질문을 입력하고, 입력된 질문에 가장 적합한 회의록 데이터를 실시간으로 확인할 수 있다.

[그림10. Gradio 인터페이스 구현]

그림과 같이 “납세 관련된 회의록 알려줘”라고 질문하면 상위 5개의 회의록 문서의 핵심 내용을 요약한 텍스트와 BM 점수, BERT 유사도, 발언 날짜가 출력된다. 출력된 상위 2행만 확인하면 다음과 같다.

요약	BM25 점수	BERT 유사도
왜 납입관리자가 서울특별시장인가요? 서울특별시장이 납입관리자로 지정된 이유는 지방세를 중앙에서 모아 지역별로 차등 배율을 적용하기 위해 필요한 중앙 집중 관리 체계를 구축하기 위한 조치로, 공동세의 분배 방식에 대한 이해와 설명 부족으로 인한 혼란을 해결하기 위한 것입니다.	14.59	0.7279
대구지방국세청의 조세불복 환급금 및 과오납 환급금이 다른 지방청에 비해 큰 규모로 나타나는데, 이 현상의 원인 및 대구지방국세청의 대응에 대해 설명해 주십시오. 조세불복 환급금은 납세자가 불복하여 소송을 통해 패소한 경우로, 특히 2017년과 2018년에 큰 금액의 패소 사례가 있습니다. 과오납 환급금은 납세자가 신고 오류로 인해 잘못 납부한 세금을 경정청구로 반환하는 경우로, 대부분의 환급금은 과오납에 기인하며 약 2/3를 차지합니다.	15.49	0.7077

[표3. Gradio 출력 결과]

Gradio 기반 UI의 평균 질의당 처리 시간이 0.1초로 BM25와 BERT 결합에도 불구하고 응답 속도가 우수하다. 또한 직관적인 UI를 통해 비전문가도 쉽게 접근 가능하여 사용성이 매우 높다.

4.2 결과 해석

KoBART를 사용하여 국회회의록 데이터를 summary한 것의 미흡했던 점을 보완하기 위해 GPT-2를 활용하여 요약문을 자연스럽게 정제하였다. GPT-2의 Fine tuning을 통해 KoBART의 요약 결과의 중복 및 어색한 문장 구조를 간결하고 명확하게 개선했다. 사용자가 보다 자연스럽고 읽기 쉬운 문장을 접할 수 있어 회의록 정보에 대한 이해도를 높일 수 있음을 확인하였다.
이후, GPT2를 통해 정제된 summary를 활용해 BM25 알고리즘을 통해 초기 검색 단계에서 주요 문서를 추출하고, BERT를 활용한 문맥적 재정렬 과정을 통해 최적의 검색 결과를 제공하였다. 예시로 사용한 "납세 관련된 회의록" 질의와 BM25 상위 10개 문서 중 5개의 문서를 BERT가 재정렬 하였다. 유사도가 높은 상위 문서의 BM25 점수는 약 15점, BERT의 유사도 점수는 0.7이상으로 높은 유사도를 나타냈다.
이를 통해 단순 키워드 기반 검색의 한계를 극복하고, 문맥에 적합한 결과를 사용자에게 제공할 수 있음을 확인하였다.

4.3 한계점 및 개선사항

현재 서비스는 검색 기반의 질의응답 시스템을 통해 사용자의 질문에 대한 답변을 제공하고 있다. 이는 사용자가 입력한 질문과 회의록 내에서 일치하는 키워드나 구문을 검색하여 관련 내용을 제공하는 방식이다. 그러나, 검색 기반 시스템은 사용자가 질문을 정확히 입력해야만 유효한 답변을 찾을 수 있으며, 질문의 형태가 다를 경우 유용한 답변을 제공하기 어려운 한계가 있다.
따라서 향후 지식 기반 질의응답 시스템을 추가하여, 보다 심층적이고 맥락을 고려한 답변을 제공할 수 있어야 한다. 이를 위해, 문서나 데이터베이스 내의 구조화된 지식을 활용하여 사용자의 복잡한 질문에 대해서도 효율적으로 대응할 수 있도록 해야 한다. 또한 대화형 AI 서비스를 통해 사용자가 질문을 지속적으로 주고받을 수 있는 환경을 구축하고, QA pipeline을 구현하여 사용자의 질문을 정확히 이해하고, 필요한 정보를 실시간으로 추출하여 자연스러운 대화 방식으로 답변을 제공하는 기능을 추가해야 한다. 이는 단순히 한 번의 검색 결과를 넘어, 사용자와 지속적인 상호작용을 통해 점진적으로 적합한 답변을 도출하는 진화된 대화형 질의응답 시스템을 실현할 수 있을 것이다.

두번째 한계점으로는 회의록 데이터에 AI 모델이 학습하기에 어려운 비정형적 요소를 포함하고 있다는 점이다. 회의록은 주로 자연스러운 대화체와 다양한 표현들이 섞여 있어, 모델이 문맥을 파악하거나 의미를 정확히 해석하는데 어려움이 있을 수 있다. 또한, 발언자의 말투나 개별적인 표현 차이로 인해 모델의 학습 효율성이 떨어질 가능성이 있다.
따라서 이와 같은 문제를 방지하기위해 미리 회의록 데이터를 정제하고 표준화하여, AI 모델이 보다 쉽게 학습할 수 있도록 해야한다. 대화체의 비정형적인 표현을 정리하고 중복 표현이나 비속어 등을 필터링하고, 문맥에 맞는 표현으로 변환하는 작업을 거쳐야한다.

5. 결론 및 향후 연구 방향

5.1 결론

본 연구는 국회 회의록 데이터를 기반으로 BM25와 BERT를 결합한 검색 기반 질의응답 시스템을 설계하고, 이를 통해 검색 정확도와 응답 적합성을 높였다. BM25 알고리즘을 활용하여 초기 검색 단계를 간소화하고, BERT 기반 재정렬로 문맥적 적합성을 높임으로써 사용자에게 신뢰도 높은 검색 결과를 제공하였다. 또한, Gradio UI를 통해 비전문가도 쉽게 사용할 수 있는 사용자 친화적 인터페이스를 구현하여, 국회 회의록 데이터 활용성을 극대화하였다. 본 시스템은 법률 및 정책 데이터 검색의 새로운 가능성을 제시하며, 공공 데이터의 민주적 활용 가치를 실현하는 데 기여할 것으로 기대된다.

5.2 향후 연구 방향

- **키워드 분석 및 가중치 기반 검색 개선**
  현재 데이터에 포함된 키워드 정보를 활용하여 문서의 중요도를 평가하고, BM25 초기 검색 점수에 가중치를 반영하는 방법을 도입할 수 있다. 특정 키워드가 포함된 문서에 높은 가중치를 부여함으로써 정책적 중요성이 큰 데이터를 우선적으로 제공할 수 있다. 이를 통해 사용자 쿼리와 문서 간의 관련성을 더욱 정교하게 분석하고, 검색 결과의 품질을 높일 수 있다.
- **날짜 필터링 기능 추가**
  Gradio 기반 서비스에 질의 응답 시 "특정 날짜 이후의 회의록만 검색" 하는 기능을 추가할 수 있다. 이를 통해 사용자는 관심있는 시간대의 데이터에 더욱 집중할 수 있으며, 최신 정책 논의와 관련된 정보를 쉽게 찾을 수 있다.
- **대화형 AI 서비스로의 확장**
  현재 검색 기반 질의응답 시스템을 넘어, pipeline 형태의 지능형 질의 응답 시스템을 구축하고자 한다. 이 시스템은 사용자의 질의 의도를 더 깊이 이해하고, 맥락에 적합한 답변을 생성하여 보다 직관적이고 풍부한 정보를 제공한다. 또한, 상호작용형 챗봇으로 발전시켜 사용자가 후속 질문을 할 수 있는 기능을 추가할 수 있다. 이를 통해 사용자는 추가 정보를 요청하거나 초기 질의를 세분화하여 구체적인 정보를 받을 수 있으며, 시스템의 실용성을 더욱 높일 수 있다.

6. 참고문헌

[1] HM Company, "리걸테크란?", 2020.12.31

[2] 조정형, "국회 도서관, '국회회의록 빅데이터' 서비스 개시", 전자신문, 2021.09.01

[3] 행정안전부, "1시간 분량 회의도 AI가 5분만에 회의록으로 뚝딱", 공공누리, 2024.03.21

[4] 김은미 외 3인(2024), KoBERT와 KoGPT2 기반의 대형언어모델과 딥러닝을 통합한 리뷰 유용성 예측모형 : The Prediction of Review Helpfulness by Integrating Large Language Models and Deep Learning based on KoBERT and KoGPT2

[5] 유성준 (2021), 딥러닝 기반의 세법 관련 질문 유형 분류 모델 연구 Classification Model for the Type of Tax Law Related Questions based on Deep Learning

[6] 류기동 외 4인(2019), AI기반 콜센터 실시간 상담 도우미 시스템 개발 – N은행 콜센터 사례를 중심으로

[7] 크롤링한 기사 내용으로 워드클라우드 만들기 (https://blog.naver.com/cathx618/222426012141)

[8] Kichang Yang (2021),Soongsil University,
Transformer-based Korean Pretrained Language
Models: A Survey on Three Years of Progress