실시간 콜센터 상담사 보조를 위한 주요 상담 발화 추출 요약 시스템

정이안○, 이용택, 김현목, 김영철
SK 주식회사
ianjung@sk.com, eyongt@sk.com, kimhm@sk.com, youngchol.kim@sk.com

An Extractive Summarization System for Real-time Call Center Agent

Ian Jung○, YongTaek Lee, Hyunmok Kim, Yongchol Kim
SK Inc.

요 약

인공지능 기술이 발전하며, 다양한 산업군에 사람의 업무를 보조하는 인공지능 시스템이 적용되고 있다. 그 중 콜센터 상담사의 상담 업무를 보조하는 자연어 처리 기술 역시 활발히 연구되고 있는 분야 중 하나이다. 콜센터 상담사 보조 시스템은 상담사를 보조하기에 앞서 고객과 상담사의 대화로 진행되는 상담이 어떤 내용인지 정확히 인식해야 한다. 이때, 시스템이 상담의 목적을 대표할 수 있는 발화를 판별한다면 상담 내용을 보다 명확히 인식할 수 있다. 본 논문은 구어체로 진행되는 상담 스크립트의 특징을 주목하여, 실시간으로 상담 내용을 분석하고, 중요한 의미를 가지는 발화를 인지하여 추출하는 모델을 제안한다. 실험 결과, 제안한 모델이 기존 추출 요약과 비교하여, 우수한 성능을 보였다. 본 논문에 제안한 모델을 적용하여, 주요 상담 발화를 추출하고, 관련된 상담 문서 검색, 상담 내용 분류 등에 적용할 수 있다.

주제어: 요약, 대화시스템, 구어체요약

1. 서론

최근 인공지능 기술이 발전함에 따라, 기존에 사람이 수행하던 업무 중 일부를 기계가 대신 수행하거나 보조하는 기술이 자연스럽게 함께 발전하고 있다. 이 중 콜센터에서 근무하는 상담사의 상담 업무를 보조하는 인공지능 기술 역시 연구되고 있는 분야 중 하나다. 기존 상담사의 업무를 보조하기 위해 통화 내용을 텍스트 형태의 상담 스크립트로 받아 적는 Speech-to-Text(STT), 상담 내용 중 주요 키워드 추출 혹은 상담 내용 요약, 상담 내용 분류 등이 다양한 기술이 적용되고 있다.

실제 콜센터를 운영하는 운영자나 상담사의 고충은 이러한 기술 발전과 괴리가 있다. 특히 단순 노동 시간 감축이 아닌, 실시간으로 수행되는 상담의 보조에 대한 요구가 있다. 센터의 업무 특성상 상담사의 업무 숙련도에 따라 업무 처리에 소요되는 시간과 정확성의 차이가 발생한다. 그뿐만 아니라, 콜센터는 상담사의 이직률이 높은 편이기 때문에 신규 상담사가 신속하고 정확한 상담을 수행할 수 있도록 보조할 기술이 필요하다[1]. 이를 위해 시스템은 현재 진행되는 상담이 어떤 내용인지 이해하고 있어야 한다.

자연어 처리 연구 분야 중 추출 요약은 텍스트 정보에서 중요한 정보를 담고 있는 문장을 선별하여 요약을 생성하는 방법이다. 이 접근법을 통해 상담에서 주요 발화를 추출하여 시스템에 상담 내용을 이해시킬 수 있다. 그러나 기존의 연구는 대부분 문어체 데이터인 뉴스 기사 혹은 위키피디아 문서를 사용한다는 한계가 있다. 본 논문은 콜센터 상담 스크립트에 등장하는 텍스트가 고객과 상담사 사이에서 진행되는 구어체라는 특징에 주목하여, 이에 적합한 데이터셋을 구축하고 주요 발화를 추출 요약하는 시스템을 소개한다.

제안하는 모델 학습을 위해 공개 데이터셋 'AIHub 상담 음성 데이터셋[2]'을 사용하였다. 'AIHub 상담 음성 데이터'는 녹취록이 전사된 상담 스크립트와 발화자 정보 그리고 발음 전사가 아닌 표준 표기법 등이 제공되고 있다. 본 논문에선 콜센터 상담 내용을 분석하는 데 필요한 정보를 추가로 태깅하여 데이터를 구축하였다. 우선, 콜센터 상담 발화는 다음과 같은 네 가지 유형으로 정의하였다. 고객 요청, 혹은 고객의 요청 사항을 확인하는 질문과 이에 대한 답변, 그리고 두 유형에 해당하지 않는 평서문 유형이 있다. 또한, 상담 내용을 이해하기 위해 의미가 있는 발화도 있지만, '네', '아', '안녕하세요 ○○○입니다.'와 같이 분석에 있어 불필요한 발화가 존재한다.

제안한 모델의 검증을 위해, 정확도 및 F1-Score를 평가 지표로 사용하여, 추가 구축한 데이터 태깅 정보의 효용성과 주요 상담을 적절하게 추출하였는지, 그 효과성을 확인하였다. 실험 결과, 본 논문에서 제안한 모델이 상담 내용 중 주요 발화를 가장 잘 추출하였음을 확인할 수 있었다.

2. 관련 연구

최근 자연어 처리 연구에 딥러닝이 적용되면서, 뉴스 기사 또는 위키피디아와 같은 문서뿐만 아니라 회의록, 상담 내용의 자동 요약 혹은 카테고리 분류에 대한 수요가 증가하고 있다. 회의록과 상담 데이터는 구어체 데이터라는 특징이 있다. 구어체 텍스트 요약 데이터셋으로 DialogSum[3], SAMSum[4], JDDC Corpus[5] 등이 있지만, 한국어 데이터가 아닌 영어 혹은 중국어 데이터이기 때문에 한국어에 적용하기 어렵다는 한계가 있다. DialogSum[3]과 SAMSum[4]은 일반 대화 요약 데이터셋이며, JDDC Corpus[5]는 각 상담의 상담 의도(예. 주소지 변경, 환불 정책 문의 등)와 상담사 챗봇이 생성할 적절한 응답을 구축한 데이터셋이다.

[6]와 [7]은 콜센터 시스템에 자연어 처리를 수행할 때 현재 발화의 주제를 인지하고, 관련 있는 발화를 한 조각으로 만드는 '주제 조각화(Topic Segmenatation)' 과정을 진행한다. 동일한 주제 조각에 있는 발화들은 내재적으로 유사한 주제를 포함한다. [6]는 인접한 발화들이 주어졌을 때, 주제 조각화가 수행될지 그 여부를 판단한다. 판단 알고리즘은 판단을 시작할 발화를 중심 c로 정의하고, 좌우 d개의 발화를 사용하여, 좌측 발화 묶음과 우측 발화 묶음의 유사성을 비교하며 탐욕적으로 주제 조각화를 수행한다. 좌우 발화 묶음이 유사하다고 판단되면, 주제 조각화가 수행되지 않으며, 다음 발화를 검사한다. 이때 조각화가 지나치게 수행되지 않도록 매 발화를 검사하지 않고 k개만큼의 발화를 건너뛴 다음, 다시 주제 조각화 수행 여부를 검사한다. 또한, 한 주제 조각이 지나치게 많은 발화를 포함하는 문제를 해소하기 위해 한 주제 조각에 최대 R개의 발화를 수용한다. [6]와 [7]은 주제 조각화를 통해 분석된 주제 정보를 반영하여, 각각의 태스크를 수행한다. [6]는 실시간 상담 챗봇이 응답을 생성하기에 앞서 이전까지 주어진 대화의 주제를 인지하여, 응답을 생성한다. [7]은 콜센터 대화 텍스트 요약 모델로 다중 발화가 주어졌을 때, 모델은 의미적으로 유사한 발화로 쪼개어 주제 조각화를 진행한 다음, 각 조각별로 중요한 정보를 인지하고, 요약문을 생성한다.

자연어 처리 연구 중 텍스트의 영화에 대한 호불호, 상품에 대한 만족도 분석과 같은 주어진 텍스트의 감성을 분류하는 연구가 있다. 감성 분류 태스크는 입력으로 자연어 텍스트를 주었을 때 긍정, 부정, 중립과 같은 감성 유형을 분류한다[8][9].

추출 요약은 주어진 문서에서 중요한 문장을 추출하는 요약 방법이다. 기존의 연구로는 주어진 문서에서 최초로 등장하는 3문장을 요약 문장으로 추출하는 LEAD-3, 주어진 문장의 추출 여부를 분류 문제로 정의하여 접근한 [10]가 있다. 중요한 문장을 추출하되 중복성을 방지하기 위하여 Maximal Marginal Relevance(MMR)[11][12]을 적용한다. 또한, 주어진 문장 외, 부가 정보를 활용하여 요약의 품질을 높인다[13][14].

	학습용	검증용	평가용	총계
발화문	47,469	11,868	1,235	60,572
상담 스크립트 수	1,515	166	36	1,717
주제 조각수	3,829	424	102	4,355

표 1. 구축 데이터의 학습용/검증용/평가용 데이터 내 발화문, 상담 스크립트 수, 주제 조각 수

대분류	소분류	데이터 수	비율
발화 유형	질문	26,379	43.55%
	답변	13,508	22.30%
	평서문	20,685	34.15%
무의미 여부	무의미	16,521	27.27%
	유의미	44,051	72.73%
주요 발화의 발화자	상담사	131	2.13%
	고객	6,009	97.87%
주요 발화의 발화 유형	질문	5,495	89.50%
	답변	145	2.36%
	평서문	500	8.14%
한 주제 조각별 평균 발화수			13.91 발화
한 주제 조각별 평균 주요 발화 수			1.41 발화
한 주제 조각별 평균 주요 발화 비율			9.92%

표 2. 구축 데이터의 레이블별 통계 상세

3. 상담 스크립트 발화 유형 분류 및 추출 요약 데이터 구축

본 논문은 구어체 대화 데이터를 요약하는 데이터셋을 구축하기 위해 상담 스크립트 공개 데이터셋인 'AIHub 상담 음성 데이터[2]'에 추가 태깅을 진행하였다.

우선, 상담을 진행할 때 화자(고객, 상담사)는 질문, 답변, 그리고 두 가지 유형에 해당하지 않는 평서문, 총 세 가지 유형의 의도로 대화를 진행한다는 특징이 있다.

또한 구어체 특성상 '예', '네', '아', '그렇군요'와 같은 단순한 호응에 해당하는 발화가 존재한다. 이와 같은 발화는 대화 맥락을 이해하는 데에 있어 불필요한 정보에 해당한다. 그뿐만 아니라, '안녕하세요. ○○○입니다', '감사합니다.'와 같은 인사말 역시 상담의 시작과 끝을 알리는 용도일 뿐 상담 내용 분석에 사용되지 않는 무의미한 문장이다. 따라서 우리는 각 발화에 대해 발화 유형을 세 가지로 분류하며, 동시에 해당 발화가 이 상담 내용을 이해할 때 유의미한지 그 여부를 분류하였다.

'AIHub 상담 음성 데이터[2]'의 대화 스크립트의 또 다른 특징은 하나의 상담 스크립트 내에서 여러 가지 주제로 대화하는 상황이 발생한다는 점이다. 가령, 카드를 분실했을 경우, 분실 신고 및 재발급, 배송 방법과 같은 다양한 주제로 상담이 진행될 수 있다.

이 특징에 주목하여 우리는 각 발화에 주제를 명시하도록 하였으며, 각 발화에 주제 번호를 부여하였다. 주제 번호는 단순히 한 스크립트 내에서 주제 전환이 있음을 인지하기 위해 태깅된 정보이며, '카드 분실'과 같은 명시적인 텍스트 레이블은 수행하지 않았다.

마지막으로 한 스크립트 내에서 상담의 목적 혹은 요

청 사항 등 대표성을 가질 수 있는 문장을 태깅하였다. 이러한 발화를 주요 발화로 정의하며, 이는 기존의 추출 요약 데이터셋의 요약 문장에 해당한다.

추가적으로, 해당 데이터는 동일한 상담 스크립트를 서로 다른 발화자가 녹취한 데이터가 다수 존재한다. 이 때문에 STT 전사 스크립트는 상이하나, 실제로 동일한 상담 스크립트일 가능성이 있으므로, 동일한 상담을 중복 태깅하는 일을 줄이고자, ROUGE 함수[15]를 활용해 중복 스크립트를 제거하였다.

중복 데이터 정제 후, 총 1,717개 상담 스크립트, 60,572개의 발화에 대해 태깅 데이터를 구축하였다. **표 1**과 **표 2는** 태깅을 수행한 후 데이터에 대한 정보이다.

'AIHub 상담 음성 데이터셋' 은 학습 데이터(Train)와 검증 데이터(Validation)를 공개하였다. 그러나, 검증용 데이터가 학습 데이터와 비교하여 충분하지 않았기 때문에, 본 논문에서는 기존의 학습 데이터를 학습용, 검증용 데이터로 나누어 태깅 작업을 진행하였고, 기존의 검증용 데이터는 평가용 데이터로 활용하였다.

태깅 결과, 무의미한 발화가 전체의 약 27%를 차지한다. 이는 우리가 분석을 수행할 때 노이즈 데이터에 해당하는 부분으로 정의할 수 있다. 또한, 분석에 필요하지 않기 때문에, 이를 제거하는 것만으로도 연산량을 감소시킬 수 있다. 상담 스크립트 내에서 주요 발화의 특징을 살펴본다면, 대부분 고객의 발화에 해당하며 발화 유형을 질문 형태에 해당한다는 것을 확인할 수 있다.

그림 1. 실시간 주요 상담 발화 추출 시스템.

4. 실시간 상담 스크립트 주요 문장 요약 모델

본 논문에서 제안하는 방법은 상담 발화를 4가지 측면에서 분석을 수행한 다음, 결과를 결합하여 주요 발화를 최종적으로 추출한다(**그림 1**).

입력으로 STT 모델을 통해 실시간으로 주어지는 상담 발화 텍스트를 사용한다. 모델은 가장 먼저 [6]의 방법을 따라 주제 조각화를 수행한다. 그다음, 각 발화의 유형과 무의미한 문장을 판별하여 분류한다. 주제 조각이 주어졌을 때, 기존의 추출 요약 모델을 사용하여 각 주제 조각 내의 각 발화의 중요도 점수를 측정한다. 마지막으로, 모델은 주제 조각 내에서 주요한 발화를 상기 분석 결과를 참고하여 추출한다.

4.1 발화 유형 및 무의미 발화 여부 분류

현재 발화를 분류할 때, 텍스트 정보만을 사용하여 분류하는 방법이 일반적으로 사용된다. 그러나, 상담 데이터 발화는 서로 다른 발화자 간에 주고받는 정보이며, STT 특성상 한 문장의 발화가 분리된 텍스트 정보로 제공될 수도 있다. 또한, 이전 발화에 따라 동일한 텍스트가 서로 다른 유형이 될 수도 있다. 가령, '네' 라는 단어는 단순 호응에 해당할 수도 있으나, 질문에 대한 답변일 수도 있다. 따라서, 본 논문은 이전 발화의 정보를 사용하여 발화 유형을 분류하고, 내용 분석에 사용되지 않을 무의미한 발화를 분류한다.

그림 2는 현재 발화 텍스트와 이전 발화 유형 정보를 사용하여 현재 발화 유형을 분류하는 모델이다. 임베딩 레이어를 통과한 이전 발화 유형 정보는, 언어 모델을 통과한 발화 텍스트 정보와 결합하여, 현재 발화의 유형을 분류하게 된다.

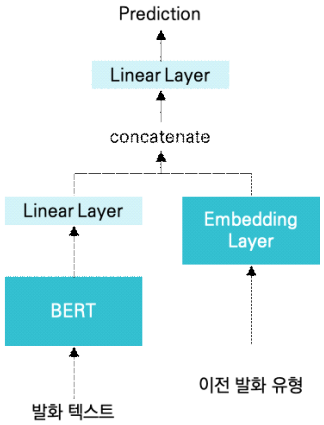

그림 2. 이전 발화 유형을 참고하여 현재 발화의 발화 유형 및 무의미 여부를 분류하는 모델 구성.

4.2 각 발화의 중요도 점수 측정 모델

기존 추출 요약 모델의 경우 문서 전체 문맥에 대한 추출 확률을 기반으로 수행하였다. 본 논문은 각 주제 조각의 발화들을 단일 문서로 가정한다. 즉, N개의 주제 조각이 주어졌을 때 중요도 점수 산출 모델의 입력되는 정보는 M개의 발화를 갖는 n번째 주제 조각 $TC_n = [uttr_n^1, uttr_n^2, \cdots, uttr_n^M]$이 된다. 이때 $uttr_n^i$는 n번째 주제 조각의 i번째 발화를 의미한다.

본 논문에서는 **BERTEXTSUM+Classifier**[10] 모델을 사용하였으며, 중요도 점수로 해당 모델이 추출 요약문으로 사용될 확률을 사용하였다.

4.3 문장 필터링 및 요약 문장 선택 방법

주제 조각화, 발화 유형 분류, 무의미 여부 그리고 중요도 점수 정보가 분석된 후, 모델은 모든 정보를 결합하여 최종적으로 실시간 상담 스크립트의 주요 발화를 추출한다.

각 주제 조각별로 주제를 나타낼 수 있는 문장을 선택할 때 **4.2**에서 도출된 중요도 점수를 가장 우선으로 한다. 이때, 제약 조건은 **4.1**의 결과를 기반으로 한다.

한 주제 조각 내에서 현재 문맥을 이해함에 있어 의미가 없는 문장을 선택할 필요가 없으므로, 무의미한 문장으로 분류된 발화는 최종 선택에서 제외한다.

5. 실험

5.1 실험 설정

모든 모델의 Epoch은 10, Weight Decay는 0.1, Learning Rate는 5e-5, 각 발화의 최대 길이는 128로 설정하여 학습하였다. 사전 학습된 언어 모델은 내부적으로 학습한 토크나이저와 **BERT**[16]를 사용하였다.

5.1.1 주제 조각화 실험 설정

표 1에 따르면, 단일 상담 내에서 평균 2.5개의 주제 조각이 존재한다는 걸 알 수 있다. 즉 평균적으로 두 번 혹은 세 번의 주제 전환이 이루어진다. 본 논문은 이 특징에 주목하여, [6]의 주제 조각화를 수행하고, 이에 대한 효과성을 확인하는 실험을 수행한다.

주제 조각화의 언어 모델은 언어에 구애받지 않는 BERT기반 문장 임베딩 모델인 **LaBSE**[17]를 사용하였고, 학습용 데이터의 통계를 기반으로 검증용 데이터셋에서 가장 실제와 비슷한 결과를 도출하는 파라미터 $d=2$, $R=10, k=3$으로 설정하였다.

5.2 데이터셋 설명 및 전처리

'AIHub 상담 음성 데이터[2]' 는 표준 발성에서 벗어나거나, 같은 전사에 대해 두 가지 이상 발음이 가능할 경우 표준어법에 맞게 표기한 '철자 전사'와 발성된 내용을 소리 값에 가깝게 표기한 '발음 전사' 두 가지를 병행하여 표기한다. 본 논문은 STT 특성상 모든 발음을 한글로 표기하기 때문에 '(1999년)/(천 구백 구십 구 년)' 과 같은 전사 데이터에서 발음 전사만 추출하여 사용하였다. 또한, 마침표, 물음표와 같은 특수기호도 STT에서 생성하지 않기 때문에 한글을 제외한 모든 특수기호도 제거하였다.

5.3 실험 결과

5.3.1 발화 유형 분류

주어진 발화 유형을 분류하기 위해 일반적으로 사용하고 있는 분류기 모델(언어 모델+Linear Classifier)을 비교 모델로 사용하였다.

실험 결과 현재 발화 정보만 사용하는 모델보다, 이전 발화 유형을 사용했을 때 성능이 높았다(**표 3**의 **A-1, A-2, B-1, B-2**). 즉, STT 전사 데이터이자 구어체인 상담 텍스트는 이전 발화의 정보가 의미 있게 사용될 수 있다는 사실을 확인할 수 있다.

구분	태스크 분류	입력	Accuracy	F1-Score
A-1	발화 유형 분류 (질문/답변/평서문)	현재발화	83.64%	80.44%
A-2		현재발화 +이전발화 유형	86.64%	84.77%
B-1	발화 유형 분류 (질문/비질문)	현재발화	90.12%	89.76%
B-2		현재발화 +이전발화 유형	**91.90%**	**91.57%**
C-1	무의미 여부 분류	현재발화	78.04%	74.07%
C-2		현재발화 +이전발화 유형	**78.85%**	**75.28%**
D-1	발화 유형 분류 (질문/답변/ 평서문/무의미)	현재발화	67.61%	61.63%
D-2		현재발화 +이전발화 유형	70.85%	67.25%
E-1	발화 유형 분류 (질문/비질문/ 무의미)	현재발화	73.20%	72.61%
E-2		현재발화 +이전발화 유형	**77.41%**	**76.95%**
F	A-2 * C-2	현재발화 +이전발화 유형	68.32%	63.81%
G	B-2 * C-2	현재발화 +이전발화 유형	72.46%	68.93%

표 3. 발화 유형 및 무의미 발화 여부 분류 모델의 실험 결과. 모델 입력으로 현재 발화만을 사용하는 경우와 현재 발화와 함께 이전 발화의 유형을 참고하는 경우를 구분하여 평가를 수행하였다.

5.3.2 무의미 발화 여부 분류

발화 유형 분류 모델과 마찬가지로 무의미한 발화 여부를 분류하는 모델 역시 이전 발화의 유형 정보를 주었을 때, 향상된 성능을 보여준다(**표 3**의 **C-1, C-2**). 즉, 유사한 형태의 텍스트일지라도 이전 발화 정보에 따라 무의미 여부가 바뀐다는 특징을 확인할 수 있다.

추가적으로, 우리는 발화 유형 분류와 무의미 발화 여부를 두 단계에 걸쳐 구분하는 방법 외에도 분류 모델을 하나로 통합하여 유의미한 질문·답변·평서문과 무의미한 발화로 구분하는 모델을 추가로 학습하였다(**표 3**의 **D-1, D-2**). 이에 대한 비교 모델로 상대적으로 우수한 성능을 보였던 **표 3**의 **A-2** 모델과 **C-2** 모델의 성능을 곱하였다(**표 3**의 **F**). 비교 결과 두 단계로 분리하여 학습하는 것보다는 단일 모델로 통합하여, 무의미한 문장을 판별할 때 더 효과적임을 확인하였다.

5.3.3 발화 유형 '질문' 통합 후 분류

표 2의 통계에 따르면 대부분 주요 발화는 질문 유형의 발화에 해당했기 때문에, 질문 유형에 해당하는 발화를 구분하는 모델을 추가로 학습하였다(**표 3**의 **B-1, B-2, E-1, E-2**). 실험을 위해 답변과 평서문 유형의 발화를 비질문 유형으로 통합하였다, 실험 결과, 최종적으로 질문 유형 여부를 분류하는 모델이 가장 높은 성능을

보였다. 이 모델의 유효성은 5.3.5와 **표 5**에서 다시 한 번 확인할 수 있다.

5.3.4 주제 조각화 및 주요 발화 추출

상담 스크립트 수가 1,700여 개이며, 이 중 평가용 데이터는 36개이기 때문에 검증용 데이터에서의 성능을 병행하여 표기하였다.

또한 주제 조각화 과정의 효과성을 검증하기 위해, 주제 조각화가 이상적으로 수행된 상황을 가정하고, 학습 및 평가를 진행하였다. 비교 모델로 **LEAD-3**의 이론을 참고하여 각 주제 조각의 첫 발화를 추출하는 모델(LEAD-1)을 사용하였다. 또한, 주제 조각화 없이 상담 스크립트 전체에서 주요한 발화를 추출하는 모델도 함께 학습시켰다(**표 4**의 **BERTEXTSUM + Classifier(no topic)**)

실험 결과, 주제 조각화가 이상적으로 수행된다면 보다 높은 성능을 얻을 수 있음을 확인하였다.

	Accuracy	F1-Score
검증용 데이터		
LEAD-1	87.44%	71.87%
BERTEXTSUM[10] +Classifier	**89.61%**	**76.62%**
BERT + biGRU + Classifier	85.52%	49.08%
BERTEXTSUM +Classifier(no topic)	87.08%	57.17%
평가용 데이터		
LEAD-1	89.43%	73.74%
BETRTEXTSUM +Classifier	**90.81%**	**75.54%**
BERT + biGRU + Classifier	70.34%	50.29%
BERTEXTSUM +Classifier(no topic)	90.64%	61.32%

표 4. 주요 발화 추출 모델 성능

5.3.4 End-to-End 통합 테스트

다양한 정보를 결합하여 주요 발화를 추출하는 제안한 모델의 효과성을 검증하기 위하여 다양한 모델로 성능을 비교하였다.

BERTEXTSUM 주제 조각화를 수행하지 않고, 주요 발화를 추출하는 BERTEXTSUM + Classifier 모델이다. 중복문 제거에 사용되는 MMR[11] 수식을 적용한 결과도 함께 확인하였다. MMR을 적용하여, 최대 3개 발화 혹은 중요도 점수가 상위 10% 이내의 발화만 선택하였다.

주제 조각화 + LEAD-1 주제 조각화 수행 후 최초 발화를 선택하는 LEAD-1 방법을 적용하였다.

주제 조각화 + BERTEXTSUM 주제 조각화 수행 후 중요도 점수가 가장 높은 1개 발화를 선택하였다.

주제 조각화 + 발화 유형 분류 + BERTEXTSUM 주제 조각화 수행 후 주요 발화를 추출할 때 발화 유형 정보를 참고하여 선택하였다. 선별 기준으로 발화 유형 분류 결과 질문 유형의 발화만 선택하거나, 무의미한 발화는 배제하였다. 추출 발화수 설정은 중요도 점수가 높은 1개 발화 또는 중요도 점수가 0.5 이상인 발화를 선택하도록 하였다. 또한, 분류 결과에 따라 배제되는 방법 외에 질문으로 예측한 확률과 중요도 점수를 가중합하여 선택하는 방법도 적용하였다.

주제 조각화 + 발화 유형 분류 + 무의미 발화 여부 분류 + BERTEXTSUM 발화 유형 분류 모델과 무의미 발화 여부를 분류하는 모델을 각각 학습시켜 적용하였다.

Model		Accuracy	F1-Score
BERTEXTSUM		**90.64%**	61.32%
BERTEXTSUM+MMR(sent=3)		85.73%	55.15%
BERTEXTSUM+MMR(ratio=0.1)		85.73%	55.94%
주제 조각화 + LEAD-1		79.07%	47.64%
주제 조각화 + Highlight		83.54%	58.78%
주제 조각화 + 발화 유형 분류 모델 + BETRTEXTSUM (발화 유형 : 질문 / 비질문 / 무의미)	a	83.29%	**62.50%**
	b	84.93%	59.87%
	c	86.91%	60.43%
	d	80.43%	60.87%
주제 조각화 + 발화 유형 분류 모델 + 무의미 발화 여부 분류 모델 + BETRTEXTSUM (발화 유형 : 질문/비질문)		85.43%	58.07%

표 5. 평가용 데이터에서 End-to-End 통합 테스트 결과.

Model	Accuracy	F1-Score
BERTEXTSUM	87.08%	57.17%
Proposed Model	84.11%	**66.93%**

표 6. 검증용 데이터에서 End-to-End 통합 테스트 결과.

표 5는 평가용 데이터에서 상기 모델에 평가를 수행한 결과이다. 이 중 **표 5-a**는 질문 유형의 발화 중 중요도 점수가 가장 높은 1개 발화를 추출하였고, **표 5-b**는 유의미한 발화 중 중요도 점수가 가장 높은 1개 발화를 추출하였다. **표 5-c**는 질문 유형의 발화 중 중요도 점수가 0.5 이상인 모든 발화를 추출하였고, **표 5-d**는 질문으로 예측한 확률과 중요도 점수를 가중합하였을 때 가장 점수가 높은 1개 발화를 선택한 결과이다.

실험 결과 평가용 데이터에서 제안한 모델이 F1-Score에서 가장 높은 성능을 보였다. 평가용 데이터가 36개라는 비교에 충분하지 않은 적은 수치이기 때문인지 기존의 모델과 비교하여 눈에 두드러지는 성능 차이를 발견하기 어렵다. 때문에 **표 5**에서 가장 우수한 성능을 보인 **표 5-a** 모델과 **BERTEXTSUM**의 검증용 데이터에서의 성능을 비교하였다. 비록 검증용 데이터이기 때문에 정확한 비교는 어렵지만, 평가용 데이터와 검증용 데이터 모두 제안한 모델의 성능이 기존 요약 모델보다 높은 성능을

보이고 있어, 제안한 모델의 적합성을 확인할 수 있다.

표 4와 비교했을 때, 주제 조각화에 의해 성능이 저하됨을 알 수 있다. 다만, 주제 조각화가 이상적이지 않을 때, 발화 유형 및 무의미 발화 여부 분류 모델이 성능 향상에 기여한다는 사실을 확인할 수 있다. 향후 상담 스크립트의 주제 조각화를 잘 수행하는 알고리즘 연구가 필요하다.

6. 결론

본 논문은 실시간 주요 상담 발화 추출 시스템을 소개하였다. 제안한 모델이 우수한 성능을 보여, 구어체 대화 데이터 요약에 적합함을 입증하였다.

또한, 이 모델은 주요 발화를 추출하기 때문에 요약 외에도 추출된 주요 발화를 기반으로 관련된 내부 문서 검색 등 다양한 상담사 보조 시스템에 적용 가능할 것으로 기대한다.

향후 연구로는 더욱 많은 상담 데이터를 확보하여, 모델을 효과성을 검증할 계획이다. 또한 제안한 모델을 기반으로 영어 대화 요약 데이터셋에 적용하여 대화 시스템에 적합한 요약 모델을 연구할 계획이다.

참고문헌

[1] 박종무, 박동수, 이재강, 안성익, “콜센터 상담사의 감정노동행동과 이직의도에 관한 실증적 연구,” 한국경영학회 융합학술대회, pp. 1512-1530, 2014.

[2] AIHub 상담음성 데이터, (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=100) [Accessed:2022-04-04]

[3] 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, pp. 5062-5074, 2021.

[4] 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, pp. 70-79, 2019.

[5] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou, “The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service”, In Proceedings of The Language Resources and Evaluation Conferences, pp. 459-466, 2020.

[6] Xu Yi, Hai Zhao, and Zhuosheng Zhang, “Topic-aware multi-turn dialogue modeling,” In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.

[7] Abedelkadir Asi, Song Wang, Roy Eisenstadt, Dean Geckt, Yarin Kuper, Yi Mao, and Royi Ronen, “An End-to-End Dialogue Summarization System for Sales Calls,” In Proceedings of NAACL-HLT, Industry Track Papers, pp. 45-53, 2022.

[8] 박광현, 나승훈, 신종훈, 김영길, “BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정,” 한국정보과학회 학술발표논문집, pp. 584-586, 2019.

[9] AIHub 감성 대화 말뭉치, (https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=86)[Accessed:2022-04-04]

[10] Yang Liu, “Fine-tune BERT for extractive summarization”, arXiv preprint arXiv:1903.10318, 2019.

[11] Jaime G Carbonell and Jade Goldstein, “The use of mmr and diversity-based reranking for reordering documents and producing summaries,” 1998.

[12] 전재원, 황현선, 이창기, “딥러닝과 Maximal Marginal Relevance를 이용한 2단계 문서 요약,” 제31회 한글 및 한국어 정보처리 학술대회, pp. 297-300, 2019.

[13] Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He, “Keywords-guided abstractive sentence summarization,” In Proceedings of the AAAI, Vol. 34, No. 05, pp. 8196-8203, 2020.

[14] Chih-Wen Goo and Yun-Nung Chen, “Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts”, IEEE Spoken Language Technology Workshop, pp. 735-742, 2018.

[15] Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” In Text Summarization Branches Out, pp. 74-81, 2004.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Vol. 1, pp. 4171-4186, 2019.

[17] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang, “Language-agnostic BERT Sentence Embedding,” In Proceedings of ACL, Vol. 1, pp. 878-891, 2022.