creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학전문석사 학위 연구보고서

키워드 추출을 이용한 대화 생성 요약 성능 향상

Improvement of Dialogue Summarization Using Keyword Extractor

2023년 2월

서울대학교 공학전문대학원
응용공학과 응용공학전공
유총재

키워드 추출을 이용한 대화 생성 요약 성능 향상

Improvement of Dialogue Summarization Using Keyword Extractor

지도교수 정 교 민

이 프로젝트 연구보고서를 공학전문석사 학위 연구보고서로 제출함

2023년 2월

서울대학교 공학전문대학원

응용공학과 응용공학전공

유총재

유총재의 공학전문석사 학위 연구보고서를 인준함

2023년 2월

위원장 _____ 곽 우 영 _____ (인)

위 원 _____ 정 교 민 _____ (인)

위 원 _____ 한 훈 _____ (인)

초 록

대화 생성 요약은 여러 화자가 말하는 대화에서 중요한 내용을 포함하는 요약문을 생성하는 것을 목표로 한다. 대화 생성 요약에서는 한번 언급된 단어에 대해서 대명사로 칭하거나, 특별히 지칭하지 않고 계속 해당 주제에 대해서 얘기하는 구어체적 특성으로 인하여 대화의 주제를 포착하기 어렵거나, 요약문을 생성하여도 핵심적인 정보를 담지 못하는 경우가 많다. 문서 요약 알고리즘은 대화 요약에서도 요약문을 생성할 수 있지만 대화의 특성을 살린 요약을 만드는 경우는 많지 않은 것이 실정이다.

따라서 본 연구에서는 키워드를 효율적으로 사용하여 의미적으로는 요약이 맞으나 실제 중요한 정보는 누락되는 문제를 개선하기 위하여 기존 알고리즘에 키워드를 추가하여 요약본만으로 대화의 주제를 쉽게 포착하는 모델을 제안한다. 이 모델은 사전 훈련된 키워드 추출기를 사용하여 대화의 키워드를 추출한다. 본 연구는 키워드를 이용하여 대화 생성 요약의 성능을 향상시킬 수 있는 시스템을 제안한다.

연구한 결과, 3개의 벤치마크 데이터셋에 대해서 수행된 다양한 실험들은 제안된 방법이 기존 대화 생성 요약 모델보다 성능이 우수함을 보여준다. 또한 제안된 모델은 훈련 데이터셋이 부족한 상황에서 성능을 더 향상시킬 수 있다. 또다른 중요한 특징 중 하나는 기존의 대화 생성 요약 모델의 변경 없이, 입력 데이터만을 수정하여 성능을 향상 시킬 수 있다.

주요어 : 대화 요약, 생성 요약, 키워드 추출

학번 : 2021-27215

목 차

I. 서론 **1**

1.1 연구 동기 1

1.2 연구의 개요 3

1.3 연구 보고서의 구성 5

II. 관련 연구 **6**

2.1 대화요약 6

2.2 대화 요약 데이터 8

2.2.1 DialogSum 9

2.2.2 SAMSum 10

2.2.3 TweetSumm 11

2.3 사전 학습 모델 12

2.3.1 Transformer 12

2.3.2 BERT 14

2.3.3 BART 16

2.3.4 T5 18

2.4 키워드 추출기 19

2.4.1 RaKUn 19

2.4.2 RAKE 20

2.4.3 YAKE 22

2.4.4 PKE 23

2.4.5 KeyBERT 24

2.5 평가 지표 26
2.5.1 BLEU 26
2.5.2 ROUGE 26
2.5.3 BERTScore 28
2.6 키워드 기반 요약 28

III. 본론 30
3.1 기존 모델 분석 30
3.2 개선 모델 제안 31

IV. 실험 결과 및 평가 36
4.1 데이터셋 비교 36
4.2 실험 상세 37
4.2.1 키워드 추출기 선택 전략 38
4.2.2 키워드 순서 40
4.3 실험 결과 42
4.4 평가 42
4.4.1 다른 데이터셋에 대한 실험 43
4.4.2 시간 비용 비교 44
4.4.3 KADS에 대한 키워드 검증 44
4.4.4 적은 데이터에서의 성능 비교 45

V. 결론 47
5.1 요약 및 결과 47
5.2 의의 및 한계 48
5.3 향후 계획 48

참고 문헌 **49**

Abstract **55**

그 림 목 차

그림 1. 최근 5개년의 대화 요약 논문의 카테고리 2
그림 2. RepSum이 제안하는 모델의 전반적인 구조 7
그림 3. Transformer의 Attention 구조 13
그림 4. BERT의 입력값 표현 14
그림 5. BART의 기본 원리 17
그림 6. BART의 5가지 noise 기법 17
그림 7. RaKUn의 구조도 20
그림 8. RAKE의 구조도1 21
그림 9. RAKE의 구조도2 21
그림 10. YAKE의 구조도 22
그림 11. keyBERT의 구조도 1 24
그림 12. keyBERT의 구조도 2 25
그림 13. keyBERT의 구조도 3 25
그림 14. BERTScore의 계산 방법 28
그림 15. 기존 대화 요약 시스템의 구조 30
그림 16. 개선된 키워드 기반 대화 요약 시스템 33
그림 17. 제안된 키워드 기반 대화 요약 시스템의 전반적인 아키텍처 33

표 목 차

표 1. 대화 요약의 예시. 빨간색은 키워드를 의미한다. 3
표 2. DialogSum의 데이터 예시 9
표 3. SAMSum의 데이터 예시 10
표 4. TweetSumm의 데이터 예시 11
표 5. 키워드 기반 문서 요약 시스템의 성능 비교 29
표 6. 대화 생성 요약 데이터셋 비교 36
표 7. BART와 T5를 이용한 DialogSum에 대한 요약 성능 비교 37
표 8. 각 키워드 추출기의 기본 매개 변수만을 적용한 KADS의 DialogSum에 대한 성능 비교 38
표 9. 각 키워드 추출기가 DialogSum에 대해서 최대 성능을 나타내는 매개 변수를 적용한 모델의 성능 비교 40
표 10. DialogSum에서 키워드를 정확도와 언급 순서 순으로 정렬한 모델의 성능 비교 41
표 11. BART와 T5를 이용한 DialogSum에 대한 KADS의 요약 성능 비교 42
표 12. SAMSum과 TweetSumm 데이터셋에 대한 기본 모델과 KADS 모델간의 성능 비교 43
표 13. DialogSum에서 기본 모델과 KADS 모델간의 학습 시간 비교 44
표 14. DialogSum에 대해서 무작위 키워드와 KADS 키워드를 적용한 모델의 성능 비교 44
표 15. KADS와 무작위 키워드 방식을 적용한 요약 예시 . . . 45

표 16. DialogSum에서 무작위로 샘플링된 데이터의 절반과
전체 데이터로 훈련된 시스템과의 성능 비교 46

제 1 장

서론

1.1 연구 동기

COVID-19로 인해 사회에 다양한 변화가 생겼다. 재택 근무가 확산됨에 따라 화상 회의가 증가하고 온라인 미팅의 수도 증가했다[1]. 이러한 변화들로 인하여 온라인에서 제공되는 데이터의 양을 고려할 때 회의 기록의 핵심 사항을 신속하게 결정해야 하는 경우가 많아졌다[2]. 따라서 대화 및 회의에 대한 요약의 필요성이 점차 증가하였지만 기존의 요약은 대부분이 뉴스나 발표 등과 같은 화자가 한명인 일방적인 정보 전달의 문체를 보였기 때문에 대화에 효과적인 성능을 보이는 모델을 찾는 것이 어려웠다.

대화는 문서와 달리 핵심정보가 보통 여러 화자의 여러 발화에 분산되어있는 경우가 많아 정보의 밀도가 낮다. 또한 대화는 한 명의 작성자가 의도와 목적을 가지고 써내려가는 글과는 달리 두 명 이상의 화자가 자유롭게 말을 하기 때문에 본질적으로 예측하기 어려운 주제 전환이 잦고, 지시어의 사용이 많으며, 다양한 상호작용이나 화자들의 특성에 따른 특수한 어휘 사용이 많다.

그림 1는 최근 5개년의 대화 요약 논문의 카테고리를 보여준다. 대화라고 해도 그림 1처럼 회의, 채팅, Email 등의 다양한 대화가 있지만 배경마다 목적성이 다르기 때문에 특정 분야에 집중되지 않고 있다. 일상 대화나 전화 통화는 불용어가 많이 쓰이고 대명사가 많이 등장하는 반면

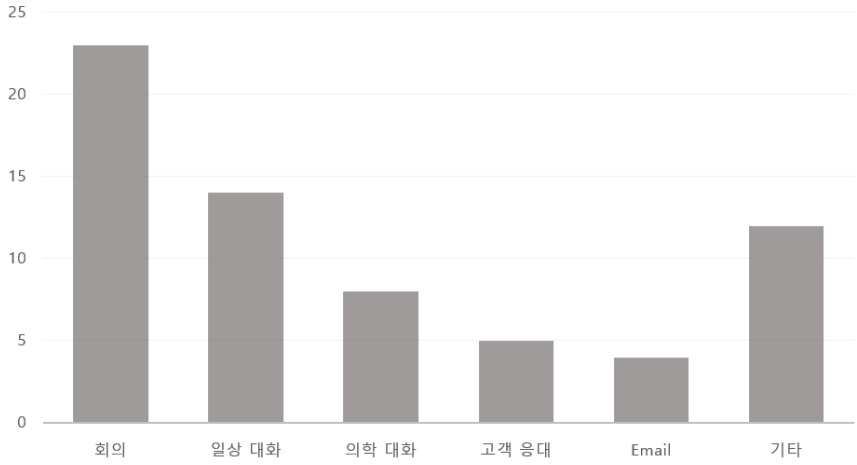

그림 1: 최근 5개년의 대화 요약 논문의 카테고리

회의는 주제가 정해져있거나 화자의 역할이 정해져있기 때문에 요약 모델에 따라 서로 다른 결과물을 보여준다. 대화라는 범주 내에서도 목적에 따라 요약의 성능이 서로 다르게 나오는 특성 때문에 기존의 데이터셋을 이용하여 요약 시스템을 적용하는 것이 어려워졌다.

대화의 데이터가 늘어남에 따라 사람들은 좀 더 빠르게 대화의 목적을 파악할 필요가 생김에 따라 연구자들은 TweetSumm[3], SAMSum[4], DialogSum[5]과 같은 여러 작업별 대화 요약 데이터셋을 제안하였다. 그러나 이러한 시스템의 데이터의 대화량은 일반적으로 이전 요약 데이터보다 적어 작업별 데이터셋을 통해 시스템을 훈련하더라도 시스템은 종종 화자의 정보를 포착하지 못하거나 대화 주제를 오해하고 매우 단순한 형태의 문장을 생성하는 한계가 있다[6].

예를 들어, 표 1에서 기존 시스템은 대화의 단편적인 사실만을 표시하는 제한적인 요약을 생성한다. 그러나 실제 대화에서 사용되는 유의미한 단어가 누락되어 해당 대화의 상세 정보를 파악하는 것이 불가능하다.

표 1: 대화 요약의 예시. 빨간색은 키워드를 의미한다.

대화
Person1: What makes you think you are able to do the job? **Person2**: My major is Automobile Designing and I have received my master's degree in science. I think I can do it well. **Person1**: What kind of work were you responsible for in the past employment? **Person2**: I am a student engineer who mainly took charge of understanding of the mechanical strength and corrosion resistance of various materials.
요약본
Person1 is interviewing Person2 about Person2's ability and previous experience.
기본 모델을 이용한 요약본
Person1 interviews Person2.
KADS를 이용한 요약본
Person1 asks Person2's major, the past work, and the reason to do the job.

1.2 연구의 개요

본 연구에서 주어진 대화의 가장 중요한 부분을 포착하고 사람들이 대화의 주요 내용과 상황을 빠르게 이해하는 데 도움이 되고자 한다. 기존의 생성 요약과 마찬가지로 입력 대화에서 추상적 대화 요약 생성 작업에 중점을 둔다[7].

이 작업을 처리하는 가장 간단한 방법의 하나는 CNN/DM[8] 또는 XSum[9]과 같이 널리 사용되는 데이터셋에 대해 훈련된 기존 요약 시스템을 사용하는 것이다. 그러나 뉴스 기사나 학술 논문과 같이 잘 구성된 문서에 대한 요약을 생성하는 것과 다르게 주어진 대화에 대한 요약을 생성하려면 구어체 텍스트의 속성에대해 추가적으로 고려해야 한다[10].

구어체의 대표적인 특징은 여러 화자의 동시다발적 발화로 구성되는 경우가 많다는 것이다. 이러한 특성은 일반적으로 독자가 화자의 정보를 파악하거나 대화의 주제를 파악하기 어렵게 만든다. 또한 여러 화자의 긴 대화에서는 주제가 시시각각 변할 수도 있고, 단순한 대명사로 지칭하며 주제에 대해 끊임없이 대화한다. 그러다보니 일반적인 요약 방식으로는 이러한 특징을 잡기 힘들다는 한계가 있다. 따라서, 본 연구에서는 이러한 문제를 해결하기 위해 키워드 추출기를 사용하여 기존 요약 모델에 키워드 정보를 효율적으로 활용하는 Keyword-Aware Dialogue Summarization(KADS) 시스템을 제안하려 한다.

KADS는 먼저 keyBERT[11]와 같은 사전 훈련된 키워드 추출기를 사용하여 대화에서 키워드를 추출한다. <keyword>를 스페셜 토큰으로 사용하여 키워드를 추가하고, 키워드와 대화 텍스트 사이에 구분 토큰은 <s>을 삽입하여, 입력 텍스트를 구성한다. 그리고, 입력 텍스트를 BART[12]와 같은 사전 훈련된 언어 모델을 이용하여 요약문을 생성한다. 널리 사용되는 3개의 대화 요약 데이터셋인 DialogSum, SAMSum, TweetSumm에 대해서 실험한 결과는 제안된 KADS가 ROUGE[13] 지표에서 기준 시스템보다 개선됨을 증명하였다. 이 방법은 ROUGE-L에서 약 2.7%의 개선을 보였고, BERTScore[14] 지표에서도 개선됨을 확인하였다. 또한 정성적 분석을 위해 추출된 키워드가 실제 요약문에서 유의미한 요약본을 생성한다는 것을 표 1을 통해 확인할 수 있다. 표 1의 예시에서는 키워드 *"degree"*와 *"responsible"*는 *"major"*와 *"reason to the job"*을 생성하는 데 사용되었다. 또한 대화 요약에 가장 적합한 키워드 추출기를 찾기 위해 다양한 키워드 추출기를 탐색한다.

또한, 대화 요약 작업에서 자주 발생하는 데이터셋이 부족한 상황에서의 KADS 시스템의 성능을 검증하였다. 이러한 상황에서는 오히려

KADS를 적용한 요약이 더 높은 성능 향상을 보였다. 본 연구의 주요 기여는 다음과 같이 요약될 수 있다:

- 대화의 핵심 정보를 효율적으로 활용하는 새로운 키워드 기반 요약문 생성 시스템을 제안한다.
- KADS 시스템이 세 가지 벤치마크 데이터셋에서 기존 방법보다 성능이 우수함을 증명한다.
- 최적의 사용법을 찾기 위해 대화 요약 작업에 대한 다양한 키워드 추출기의 사용법을 탐색한다.
- 데이터셋이 부족한 조건에서 제안된 KADS의 효과를 비교한다.

1.3 연구 보고서의 구성

제1장에서는 연구를 시작하게 된 동기와 배경 및 대략적인 연구 내용에 관해 서술하였다. 제2장에서는 배경지식 및 관련 연구로 대화 요약과 키워드 기반 요약에 관해 설명하였다. 제3장에서는 키워드 기반 대화 요약의 구조에 관해 서술하고, 키워드 추출에 관해 설명하였다. 이를 기반으로 키워드 추출을 이용한 대화 생성 요약인 KADS의 구조를 확인할 수 있다. 제4장에서는 실험에 대한 다양한 결과를 분석하였다. 최종적으로 도출한 향상된 구조를 구성하기까지 어떠한 과정을 거쳤는지, 연구 결과를 통해 알 수 있고, 다양한 데이터셋과 환경에서 제안하는 모델이 성능을 향상시킬 수 있다는 것을 기술하였다. 마지막 제5장에서는 본 연구의 결론과 의의에 대해 논의하였다.

제 2 장

관련 연구

KADS 시스템을 구성하고 평가를 통해 성능 향상을 확인하기 위해서는 다음과 같은 관련 연구가 선행되어야 한다.

2.1 대화요약

대부분의 대화 요약 데이터셋은 가장 많이 쓰이는 언어인 영어로 되어 있는 경우가 대부분이다. 또한 대화라는 것이 사생활에 밀접한 데이터인 특징 때문에 일상 대화에 대한 데이터가 거의 없다. 이러한 특징 때문에 대화 요약에 대한 훈련 데이터셋이 부족하여 대화 요약을 위한 학습이 어려운 편이다. 좋은 요약이란 원본보다 짧은 데이터로 빠르게 핵심 주제 및 중요한 정보를 파악할 수 있는 데이터를 의미한다[15]. 또한 좋은 요약은 모든 문장에 의미 있는 정보가 포함되어 있지 않다는 점을 고려하여, 원본 텍스트를 대체한다는 할 수 있어야 한다[16].

좋은 요약의 전제 조건을 만족시키기 위해 대화 요약 모델을 제안한 RepSum[17]도 부족한 데이터셋의 한계로 unsupervised 방식[18]을 제안하였지만, 이러한 방식은 평가하는 것이 매우 까다롭다. 그림 2은 Repsum이 제안하는 모델의 전반적인 구조를 나타낸다. 이 방식은 좋은 요약본은 원문의 데이터들을 충분히 가지고 있으므로, 다양한 분야에서 원래의 대화를 충분히 대체할 수 있다는 점에 착안하여 요약의 성능을 높이려 한다. 이 방식은 대화를 기반으로 요약을 생성한 후, 원본 대화와 요약본

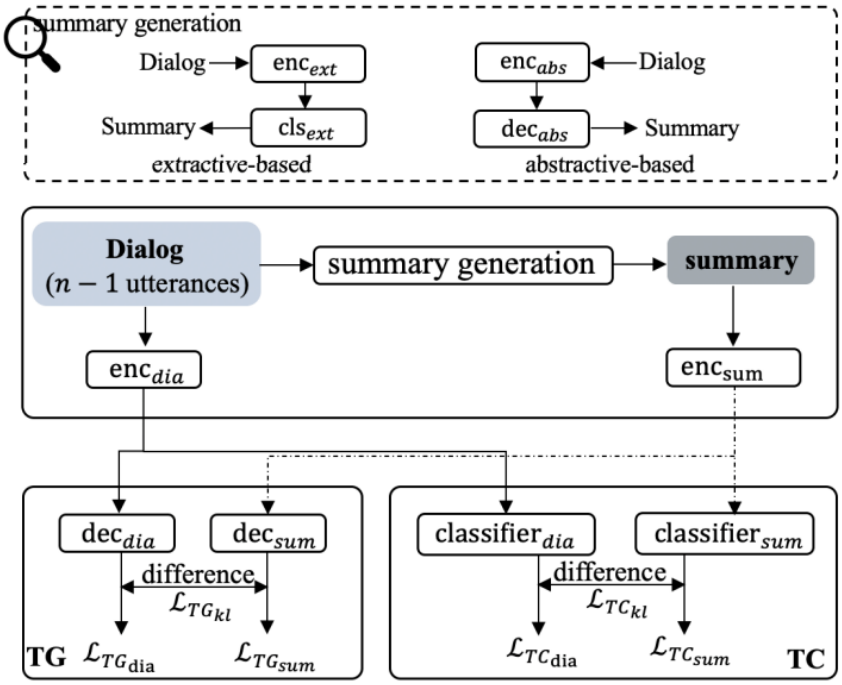

그림 2: RepSum이 제안하는 모델의 전반적인 구조[17]

에서 각각 N번째 문장을 예측한다. 이 둘의 차이를 최소화하도록 학습을 진행하는 방식을 사용하였고, 결과적으로 unsupervised의 데이터에서는 성능의 향상을 보였지만, 실제 요약본이 있는 데이터에서는 취약함을 보여주었다.

모델을 개선하는 방법뿐만 아니라 다른 방식과의 앙상블을 통해 요약 성능을 개선시키는 방법이 있다. 예를 들어, 어떤 모델은 구조를 인식하거나, 기반 지식을 추가함으로써 성능을 향상시키기도 한다[19]. 대표적인 예로 대화 구조와 기반 지식을 갖춘 추상적 대화 요약을 위한 새로운 구조인 Knowledge Graph Enhanced Dual-Copy network (KGEDC)이 있다. 소스 텍스트와 추출된 기반 지식 모두에 따라 조건화된 생성을 강제하기

위한 디코딩 프로세스를 적용하였다. 기존 모델보다 높은 ROUGE 점수를 보여주었고, 정성적인 평가로도 유의미한 결과를 도출하기도 하였다. 하지만, 이러한 방식은 기존에 사용 중인 모델에 융합하는 것이 쉽지 않고, 때때로 유의미한 키워드를 없애서 요약본의 의미가 줄어드는 경우가 발생하기도 한다[20]. 따라서, 기존의 모델에 쉽게 융합할 수 있으며, 성능이 유의미하게 향상될 수 있는 방법이 필요하다.

위와 같은 한계 때문에 대화 요약의 유의미한 성능 향상을 확인하는 것은 쉽지 않다. 기존 모델을 개선하는 방식은 적은 데이터셋의 한계때문에 사전 학습을 진행하며 유의미한 성능 향상을 기대하기 힘들다. 또한 다른 방식과의 앙상블도 기존 모델에 융합하여 사전 학습을 시키는 방식임에도 성능 향상이 더 큰 편이다. 두 방식 모두 요약의 성능은 향상되지만 기존에 요약 모델을 활용중인 상황에서는 해당 방식을 적용하는 것은 쉽지 않다. 따라서, 본 연구는 기존의 모델은 변경하지 않고, 키워드를 이용하여 데이터 입력에만 변화를 주어, 각 모델에 최적화된 키워드 추출기를 적용하여 성능을 향상할 방안을 제안할 필요가 있다.

2.2 대화 요약 데이터

unsupervised 방식의 단점을 극복하기 위해 요약본이 있는 데이터셋을 이용한 대화 요약의 성능을 개선하고 측정할 필요가 생겼다. 요약본이 있는 데이터셋은 supervised 방식[21]으로 정량적인 측정이 가능하다는 큰 장점이 있는 한편 데이터셋를 만드는일에 시간이 많이 드는 단점이 있다. 현재 대화 요약은 정량적 측정이 가능한 다양한 데이터셋이 나오고 있는데, 그중에서도 DialogSum[5], SAMSum[4], TweetSumm[3]이 가장 흔하게 쓰이고 있다.

2.2.1 DialogSum

DialogSum은 대화 요약을 위해 정제된 대화로만 만들어진 대규모 레이블링 데이터셋이다. 대규모 데이터셋은 뉴스 요약을 위한 심층 신경 모델에 관한 연구를 쉽게 하였다. 대화의 요약은 고객 서비스 관리 및 약물 추적을 포함한 다양한 실제 시나리오에서 유용하게 사용될 수 있다. DialogSum은 음성,특수 담화 구조,상호 참조 및 생략 방식,실용주의 및 사회 상식과 관련한 대화와 같은 다양한 과제에서의 대화 요약을 제공한다. DialogSum은 향후 대화 요약에서 고려해야 할 추상적이며 복잡한 참조에서의 문제 해결 방향성을 제시하였다. 이를 위해 실제 시나리오 대화 요약을 위한 대규모 데이터셋을 처음으로 제안하였다.

표 2: DialogSum의 데이터 예시 [5]

대화
Person1: Good morning. I wonder whether you have got an answer from your superior. **Person2**: Yes, we had a meting about it yesterday afternoon. **Person1**: What's the answer? **Person2**: We decided that we could agree to your price, but we are a bit worried about the slow delivery. **Person1**: Let me see. I quoted your delivery in three months, didn't I? **Person2**: Yes, but we hope that the wool could reach us as soon as possible. **Person1**: I thought you would. So I rang Auckland last night. As you are our biggest customer, they agreed to ship the order on the first vessel available that will leave Auckland next month. **Person2**: Good, if you agree we'll draft the agreement right away and sign it then. **Person1**: By all means
요약본
Person1 and Person2 agree to sign an agreement since Person1 could speed up the delivery as Person2 hopes.

DialogSum은 총 13,460개의 대화로 구성된 데이터셋을 제공해주며, 해당 데이터는 누구나 다운받을 수 있게 제공되고 있다. 데이터는 모두 정제되어 있으며 명확한 의사소통 패턴과 의도를 가지고 있기 때문에 쉽

게 의미를 파악할 수 있다. 표 2를 통해 이와 같은 특징을 확인할 수 있고, 적절한 길이의 원본과 요약본을 제공하여 자동 요약을 생성하여 성능을 확인하는데 유용하게 사용될 수 있다.

2.2.2 SAMSum

SAMSum은 대표적인 구어체 데이터셋으로, 현시점에서 가장 많은 대화 개수 셋을 제공하고 있다. 여러 모델을 테스트하여 뉴스 기사 코퍼스에서 얻은 결과와 비교한 뒤 자동 요약에 대해 제기되는 과제를 조사하는 방식을 제안한다. 일반적으로 뉴스와 같은 기사의 요약이 대화 요약보다 성능이 더 좋았지만, 모델 생성 대화 요약이 모델 생성 뉴스 요약보다 더 높은 ROUGE 점수를 달성하는 것을 보여주었다. 하지만, 정성적 평가에서는 대조적인 결과를 보여주었다.

SAMSum은 총 16,369개의 대화로 구성된 데이터셋을 제공한다. 데이터 개수는 가장 크지만, 정제된 데이터와 정제되지 않은 데이터가 혼재되어 요약본의 성능이 고르지 않다. 대화 요약의 요약본을 최초로 생성한

표 3: SAMSum의 데이터 예시 [4]

대화
Leo: BTW what are those pics? **Ryan**: Pics from Italy!!! :):):):))))))))) **Leo**: Yeah. They seem nice. (‘A‘) **Ryan**: That’s all???? I need more reactions!!!!!!!!!! **Leo**: I’m tied to this office and working like a slave. AM I SUPPOSED TO SAY Ï AM SO JEALOUS!!!!!!!!?!!!
요약본
Ryan is in Italy while Leo is working hard and wishing he could win the lottery

데이터셋이지만, 표 3과 같이 채팅 대화를 기반으로 만들어졌기 때문에,

대화의 품질은 떨어진다.

2.2.3 TweetSumm

TweetSumm은 고객 관리 대화의 영역에 초점을 맞춘 대화 요약 데이터셋이다. Kaggle Customer Support on Twitter를 기반으로 만들어졌으며,

표 4: TweetSumm의 데이터 예시 [3]

대화
Customer: flight1234 from Miami to LaGuardia smells awful. We just boarded. It's really really bad. **Agent**: Allie, I am very sorry about this. Please reach out to a flight attendant to address the odor in the aircraft. **Customer**: They're saying it game in from the last flight. They have sprayed and there's nothing else they can do. It's gross! **Agent**: I'm very sorry about the discomfort this has caused you for your flight! **Customer**: It's not just me! Every person getting on the flight is complaining. The smell is horrific. **Agent**: Oh no, Allie. That's not what we want to hear. Please seek for one of our crew members on duty for further immediate assistance regarding this issue. Please accept our sincere apologies. **Customer**: They've brought maintenance aboard. Not a great first class experience :(**Agent**: We are genuinely sorry to hear about your disappointment, Allie. Hopefully, our maintenance crew can fix the issue very soon. Once again please accept our sincere apologies for this terrible incident. **Customer**: Appreciate it. Thank you! **Agent**: You are most welcome, Allie. Thanks for tweeting us today. **Customer**: They told us to rebook, then told us the original flight was still departing. We got put back on 1234 but are now in the 1st row instead of the 3rd. Can you get us back in seats 3C and 3D? **Customer**: My boyfriend is 6feet tall and can't sit comfortably at the bulkhead. **Agent**: Unfortunately, our First Class Cabin is full on our 1234 flight for today, Allie. You may seek further assistance by reaching out to one of our in-flight crew members on duty.
상담원이 작성한 요약본
Customer complains about smell in flight. Agent updated the customer to seek further assistance by reaching out to one of their in-flight crew members on duty.
자동으로 작성된 생성 요약본
Customer is complaining about bad smell in his flight. Agent informed to contact in-flight crew member on duty for further assistance.
자동으로 작성된 추출 요약본 **Customer**: Flight1234 from Miami to LaGuardia smells awful.They told us to rebook, then told us the original flight was still departing. **Agent**: Unfortunately, our First Class Cabin is full on our 1234 flight for today, Allie. You may seek further assistance by reaching out to one of our in-flight crew members on duty.

각 대화는 사람이 직접 생성한 주석이 수반되기 때문에, 3개의 추출 요약과 3개의 추상 요약을 제공한다. 해당 데이터셋을 기반으로 생성한 요약본은 3개의 요약과 비교하여 평가할 수 있고, 추출 요약은 키워드 추출과 연관 지어서 사용할 수 있다는 장점이 있다. 표 4와 같이 고객 서비스 채팅

시나리오기반으로 고객과의 상담이 종료된 직후, 상담원이 직접 요약본을 생성하기 때문에 일반적으로 요약본의 품질이 좋다. 하지만, 상담원들이 개인의 이익을 위해 작성하기 때문에 요약본이 의도적으로 잘못된 경우가 존재하거나 상담원들의 의도에 따라 품질의 편차가 발생한다는 큰 단점이 존재한다.

2.3 사전 학습 모델

회의, 고객센터 통화내역, 인터뷰 등 분야별로 다양한 대화 요약 모델이 적용되어 있다. 기존의 모델을 개선하는 다양한 방식이 지속적으로 연구되고 있으나, 실제 사용하고 있는 분야에서는 기존 모델을 제거하고 성능이 신규 모델로 대체하는 것은 쉽지 않다. 본 연구에서는 이러한 상황에 맞춰, 실제 모델의 변경점 없이 대화 요약 성능이 개선될 수 있는 방안을 연구하려 한다. 기존 모델을 대체하지 않고 성능을 향상시키기 위해서는 사전 학습이 완료된 모델을 기반으로 대화 요약의 성능을 향상시킬 수 있어야 한다. 따라서, 사전 학습 모델이 무엇이냐에 따라 성능이 어떻게 향상되는지 확인할 필요가 있고, 어떤 사전 학습 모델을 사용할 수 있는지 연구가 필요하다. 통상적으로 많이 사용되는 대화 요약 사전 학습 모델로는 State-of-the-art(SOTA)를 달성한 BART와 T5가 있다. BERT 이후로 SOTA의 기준이 바뀌었고 그에 따라 현재 사전 학습 모델이 사용하는 기본적인 구조 및 BERT에 대한 연구가 선행되어야 한다.

2.3.1 Transformer

Transformer[22]는 RNN, LSTM 없이 입력된 문장을 병렬로 한번에 처리하는 특징을 가진 모델이라고 할 수 있다. 이 모델이 제안된 이후로

대부분의 SOTA 모델들은 Transformer 구조에 기반하여 구현되어 있다. Transformer는 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망이라고 할 수 있다. 기존의 seq2seq처럼 Encoder에서 입력 시퀀스를 입력받고, Decoder에서 출력 시퀀스를 출력하는 Encoder-Decoder 구조를 유지하지만, 단위가 N개 존재할 수 있다는 점이 다르다.

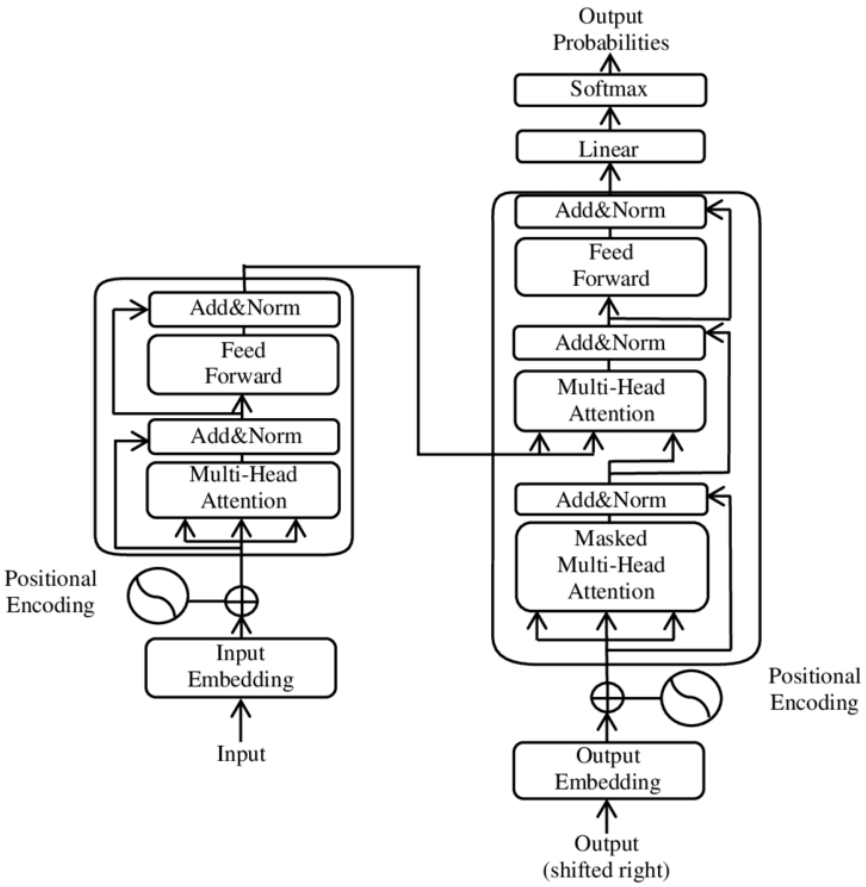

그림 3: Transformer의 Attention 구조[22]

특히, 그림 3와 같이 Attention 메커니즘만을 사용하는 Transformer

구조는 기계 번역, 구문 분석 등 모든 자연어 처리 분야에서 높은 성능을 보여주고 있으며, Image 분야에서도 SOTA를 달성하고 있다. Transformer는 일괄처리가 되지 않는 회귀 구조를 피하고, Self-Attention만으로 입, 출력의 Representation을 계산하기 때문에 병렬화가 우수하고, 입,출력 사이의 Global Dependency를 학습하여 SOTA를 달성하였다.

2.3.2 BERT

BERT[23]는 구글에서 개발한 자연어 처리 사전 훈련 기술이며, 특정 분야에 국한된 기술이 아니라 모든 자연어 처리 분야에서 높은 성능을 내는 범용적인 언어 모델이다. 11개 이상의 자연어 처리 과제에서 BERT가 최첨단 성능을 발휘한다고 하지만 그러한 결과가 나오는 정확한 이유와 로직은 상세히 알려지지 않았다. 그런데도, BERT 이전에 자연어 처리에 사용되던 선행 모델인 앙상블 모델보다 더 좋은 성능을 내고 있어서 많은 관련 연구가 진행되고 있다. BERT 모델의 특징은 학습 데이터가 부족하면 모델의 성능 또한 저하될 수 밖에 없는 딥러닝의 단점을 대용량 코퍼스를 통해 사전 학습된 모델을 사용하여 극복하였다는 것이다.

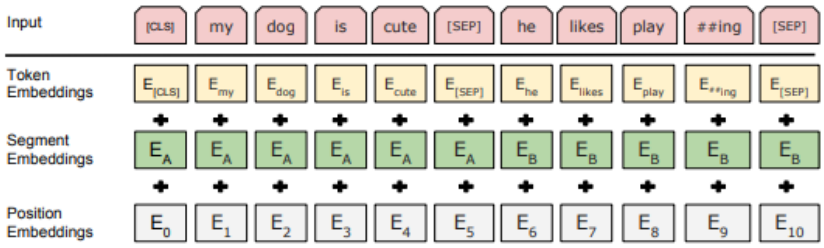

그림 4: BERT의 입력값 표현

BERT는 그림 4과 같이 입력값을 embedding할 때 총 3개의 레이어를

구현하여 합산한 결과를 이용한다. embedding이란 언어의 벡터화를 의미하는 과정으로, 언어를 벡터로 변환하여 각 단어의 의미와 단어 사이의 관계도 등을 계산할 수 있게 도와준다. 단어의 의미를 잘 표현하는 벡터로 표현하는 embedding된 단어들이 훈련과정에서 좋은 성능을 내는 것이다. 이 embedding 과정에서 BERT를 사용한다. BERT는 크게 Token, Segment, Position이라는 3개의 embedding 과정을 통해서 단어를 벡터화한다.

2.3.2.1 Token embedding

각 문자 단위로 embedding을 한 후 자주 등장하면서 가장 긴 길이의 하위 단어를 하나의 단위로 만드는 것을 의미한다. 자주 등장하지 않는 단어는 다시 하위 단어로 만드는 것을 지칭한다. 이는 이전에 자주 등장하지 않았던 단어를 전부 OOV(모르는 단어로 인해 문제를 푸는 것이 까다로워지는 상황)로 처리하여 모델링의 성능을 저하했던 OOV 문제도 해결할 수 있다.

2.3.2.2 Segment embedding

Token 화한 단어들을 다시 하나의 문장으로 만드는 작업을 의미한다. BERT에서는 두 개의 문장은 구분자를 넣어 구분하고 두 문장을 하나의 Segment로 지정하여 입력한다. 한 Segment를 512개의 하위 단어 길이로 제한하는데, 한국어는 보통 20개의 하위 단어로 이루어지고 대부분 문장은 60개의 하위 단어가 넘지 않게 한다. 학습하는 데이터의 언어,주제 등의 특징에 따라 Segment 크기를 조절할 수 있다.

2.3.2.3 Position embedding

Token의 순서대로 인코딩하는 것을 의미한다. BERT는 Transformer의 Encoder, Decoder 중 Encoder만 사용한다. Transformer와 같은 Self-Attention구조는 입력의 위치를 고려하지 않고, 입력 Token의 위치 정보를 고려한다. 따라서, Transformer 모델에서는 Sigmoid 함수를 이용하여 Position embedding을 사용하고, BERT도 이처럼 Position embedding을 수행한다.

BERT는 위의 3가지 embedding을 합치고, Layer 정규화와 Dropout을 적용하여 입력으로 사용한다. BERT는 간단한 접근만으로 자연어 처리 전 분야에 아우르는 SOTA를 달성하였다. 본 연구가 사용하는 BART또한 BERT에서 파생된 사전 학습 모델 중 하나이며, 요약 분야에서 SOTA를 달성하였다.

2.3.3 BART

BART는 페이스북에서 발표한 모델로, 기존 Transformer 구조에 새로운 사전 학습 모델을 제시하여 자연어 생성 및 이해에서 SOTA 성능을 인정받았다. BART는 seq2seq 구조로 만들어진 denoising auto-encoder이다. BART는 손상된 텍스트를 입력받아 bidirectional 모델로 인코딩을 하고, 정답 텍스트에 대한 likelihood를 autoregressive 디코더로 계산한다. 그림 5을 통해 간략하게 설명할 수 있다. BART는 다양한 noising 기법을 평가함으로써, 원본 문장을 랜덤하게 섞였다가 텍스트 복구 기법을 사용할 때에 가장 좋은 성능을 보인다. 그림 6는 5가지 noise 기법을 간략하게 보여준다. 각 기법은 아래와 같이 요약될 수 있다.

- Token Masking : BERT처럼 랜덤 Token을 masking하고 이를 복구

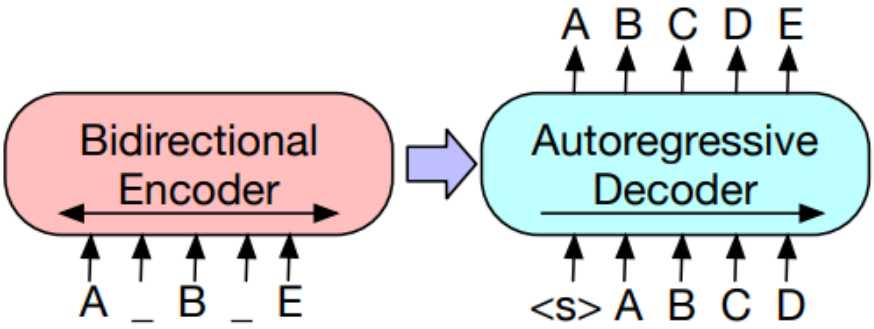

그림 5: BART의 기본 원리[12]

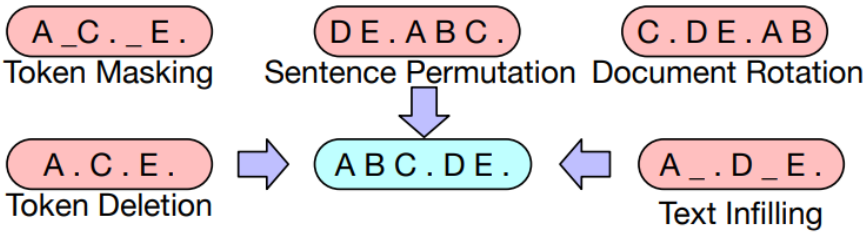

그림 6: BART의 5가지 noise 기법[12]

하는 방식이다.

- Token Deletion : 랜덤 Token을 삭제하고 이를 복구하는 방식이다. Masking과는 달리 어떤 위치의 Token이 지워졌는지 알 수 없다.
- Text Infilling : 포아송 분포를 따르는 길이의 텍스트 span을 생성하여 이를 하나의 mask token으로 masking한다. 즉, 여러 Token이 하나의 mask token으로 바뀔 수 있고 길이가 0인 경우에는 mask token만 추가될 수도 있음을 의미한다.
- Sentence Permutation : 문서를 문장 단위로 나눠서 섞는 방법이다.
- Document Rotation : Token 하나를 정해서 문장이 해당 Token부터

시작하게 한다. 모델이 문장의 시작을 구분하게 한다.

이 기법은 문장 생성에서 특히 효과적이기 때문에 요약문을 생성하는 과제에서 SOTA 성능을 보여주고 있다.

이름이 BERT와 비슷한 것처럼 구조적으로 비슷하나, BERT와는 차이점이 있다. BART는 Decoder의 각 레이어가 Encoder의 마지막 히든 레이어에 추가로 cross-attention을 수행해주고 있다. 또 다른 차이점으로 BERT는 단어 예측 전에 추가적인 feed-forward 방식을 사용하지만, BART는 추가적인 레이어를 사용하지 않고 같은 크기 대비 BERT보다 10% 많은 파라미터를 가지고 있다.

2.3.4 T5

T5는 Text-to-Text Transfer Transformer 모델로서, 텍스트 데이터를 입력하고 출력하는 방식으로 각종 작업을 해결한다. 번역, 요약, 질의응답 등의 모든 자연어 영역에서 입력과 출력이 항상 텍스트라는 점에서 출발한 T5는 기존 BERT의 단점을 보완하면서 다양한 영역에서의 성능 향상을 보여주었다. T5 모델은 기본적인 Transformer Encoder-Decoder 구조를 기반으로 한다. Encder와 Decoder 모두 multi-head self-attention layer와 feed forward network 및 residual skip connection, dropout 기법을 사용하고 있다. 즉, 현재 나와 있는 Transformer 관련 구조와 기법을 다양한 방법으로 실험한 후, 가장 좋은 조합의 구조와 학습 방법을 사용하여 현재 대다수의 성능 지표에서 SOTA를 달성하는 좋은 모델이 바로 T5이다.

2.4 키워드 추출기

키워드 추출기는 오래전부터 연구 되어 오던 분야이고, 다양한 방식으로 SOTA 모델을 제안하고 바뀌어 왔다. BERT가 제안된 이후, 키워드 추출 분야에서도 많은 변화가 있었다. 어떤 분야에 적용하느냐에 따라 SOTA를 달성하는 모델이 다르기 때문에, 각 키워드 추출기의 원리를 파악하고 대화 요약에 적합한 방식을 선택하는 것이 중요하다.

우선 다양한 키워드 추출기를 키워드 기반 요약 시스템의 주요 구성요소로 사용하기 위해 분석한 각 키워드 추출이기에 관해 설명한다. 최근에는 keyBERT[11]와 같은 다양한 키워드 추출기가 높은 성능을 나타내지만, Keyword-Aware Dialogue Summarization(KADS)시스템은 대화 요약에 적합한 키워드 추출기를 찾아야 한다.

2.4.1 RaKUn

Rank-based Keyword Extraction via Unsupervised Learning and Meta Vertex Aggregation(RaKUn)[24]은 비지도 학습 및 메타 정점 집계를 통한 순위 기반 키워드 추출 방식을 제안한다. RaKUn은 간단한 방법으로 효과적인 방식을 사용한다.

그림 7은 이 방식을 함축적으로 설명해 준다. 먼저, 텍스트 자체를 하나의 그래프로 변환한다. 입력 데이터에서 단어 간의 연속성을 이용하여 붙어있는 단어일수록, 이웃 노드가 될 가능성이 크게끔 그려진다. 이렇게 만들어진 그래프를 이용하여 각 단어 간의 이웃 노드 개수 등을 이용하여 가중치를 계산한 후, 각 그래프 간에 점수가 높은 노드끼리 병합을 진행한다. 이 과정에서 가중치는 잠재적으로 유지하되 단어 간의 유사성을 기반으로 최종적으로 가중치가 가장 높은 단어를 추출하여 키워드로 제안

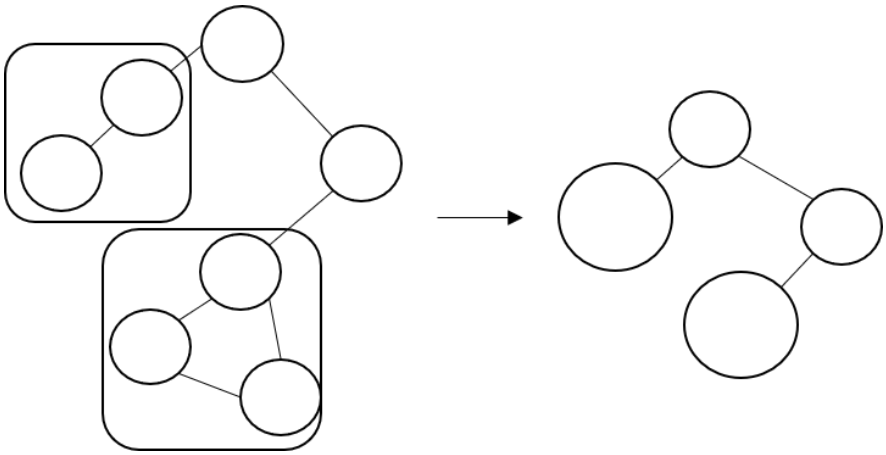

그림 7: RaKUn의 구조도

한다. 이러한 방식은 매우 간단하고 가장 빠른 방법의 하나이다. 실제로, Facebook의 fasttext 데이터셋에서 가장 높은 성능을 보여주기도 하였다.

2.4.2 RAKE

Rapid Automatic Keyword Extraction(RAKE)[25]는 동적 수집에 응용할 수 있도록 개별 문서에서 작동하는 효율적인 키워드 추출 방법이다. 이 방법은 여러 유형의 문서, 특히 특정 문법 규칙을 따르는 문서를 효과적으로 처리하기 때문에 새로운 도메인에도 적용될 수 있다. 크게 2가지로 작동 원리를 설명할 수 있다.

먼저, 그림 8의 1번 과정을 통해 후보 키워드가 텍스트에서 추출된다. 문서에서 중요하게 여겨지는 단어들이며, 이 때 문법적으로만 유의미한 단어인 불용어들은 제외한다. 불용어와 더불어서 키워드 후보군에 영향을 미치는 파라미터는 3가지가 있다. 첫번째는 단어로 간주할 수 있는 최소한의 문자열의 수, 두번째는 어구가 포함되는 단어들의 최대 개수, 그리고 세번째로 전체 입력에서 몇 번 등장하는가에 따라 후보군이 결

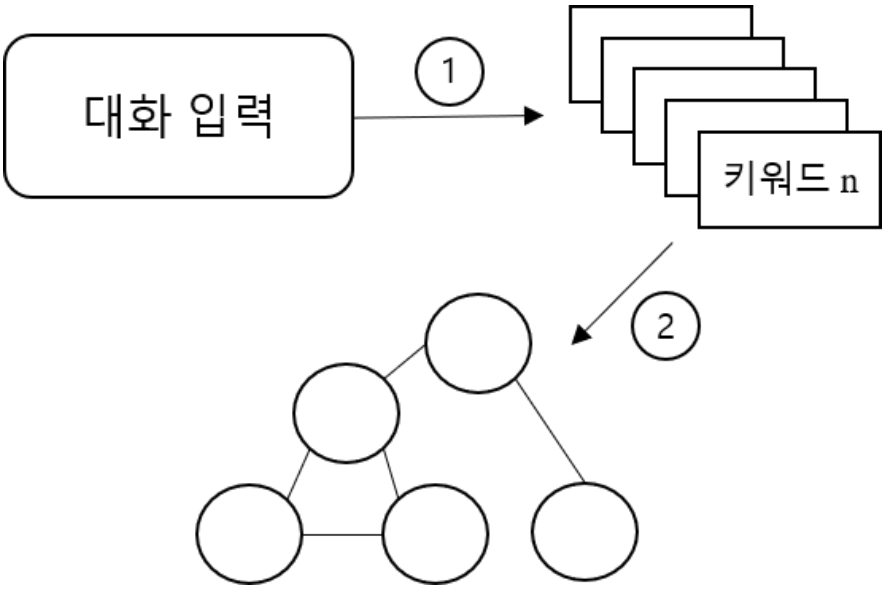

그림 8: RAKE의 구조도1

정된다. 그 다음 키워드는 노드가 되며, 다른 키워드가 함께 나타난 경우 이웃 노드라고 지정하여 2번과 같은 그래프가 완성된다. 그래프에는 키워드에 관련된 단어가 함께 나타나는 횟수에 따라서 가중치가 부여된다. 또한, 그래프에는 꼭짓점에 대한 연결이 포함되어 있다(각 단어는 자신과 함께 후보 키워드에 나타난다). 다음 과정으로 그림 9와 같이 그래프의 각

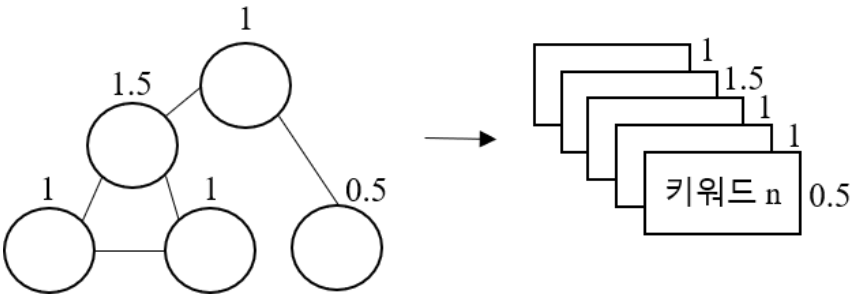

그림 9: RAKE의 구조도2

단어에 점수를 매긴다. 여기서 각 후보 키워드의 점수는 구성 단어들의 점수 합이다. 이러한 단어가 키워드 일부가 될 수 있으므로 후보 키워드에는 불용어가 포함되지 않는다. 이 알고리즘은 텍스트에서 불용어와 결합한 키워드 쌍을 찾아 기존 키워드 집합에 추가한다. 추가될 텍스트는 최소 두 번 이상 표시가 되어야 하며, 새 키워드의 점수는 구성원 키워드의 합이다. 점수를 기반으로 가장 높은 순으로 N개의 키워드를 추출한다.

RAKE는 빠른 속도를 보여주며 이미 다양한 분야에서 사용되고 있다. 특히 한국어 처리에 관한 연구도 활발하게 진행되어 이를 기반으로 한 한국어 키워드 추출기도 제안되고 있다.

2.4.3 YAKE

Yet Another Keyword Extractor(YAKE)[26]는 텍스트의 가장 중요한 키워드를 선택하기 위해 단일 문서에서 추출된 통계 텍스트 기능에 의존하는 경량 비지도 자동 키워드 추출 방법이다.

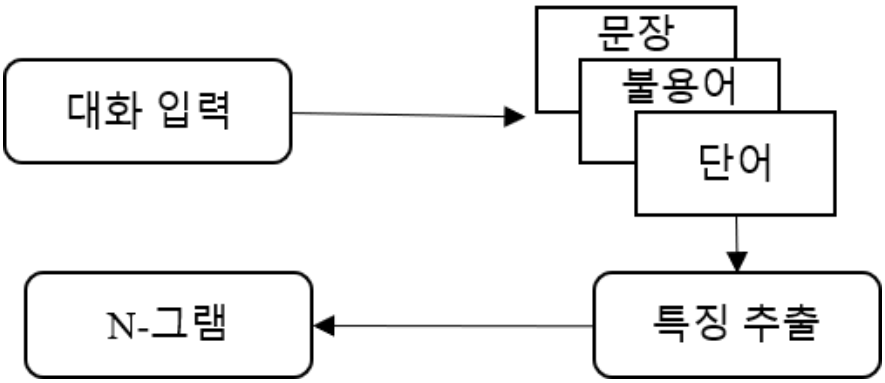

그림 10: YAKE의 구조도

이 방법은 그림 10과 같이 총 4단계로 키워드를 추출한다. 첫번째로 텍스트는 문장, 불용어, 토큰으로 나뉜다. 둘째, 텍스트를 인식하고 태그를

지정하여 불용어를 식별한다. 이 알고리즘은 문서의 단어에 대해 대소문자 구분, 용어 위치, 용어 빈도 정규화, 문맥에 대한 용어 관련성 및 용어 등 다른 문장의 다섯 가지 통계적 특징을 계산한다. 셋째, 이전 단계의 특징을 인공 방정식과 단일 점수로 결합하여 n-gram을 생성하고 키워드 점수를 계산한다. 각 n-그램은 구성원 점수를 곱하여 점수를 매기고 n-gram 길이의 영향을 줄이기 위해 정규화된다. 불용어는 영향을 최소화하기 위해 다르게 취급된다. 그런 다음, 알고리즘은 유사한 키워드를 제거하고 더 관련성이 높은 키워드(점수가 낮은 키워드)를 유지한다. 유사성은 Levenshtein 유사성[27], Jaro-Winkler 유사성[28] 또는 시퀀스 매칭기를 사용하여 계산된다. 마지막으로, 키워드 목록은 점수를 기준으로 정렬된다.

별도의 토큰화가 필요하지 않기 때문에 서비스 내 단위 모듈로서의 성능에 있어서 큰 장점이기 때문에 YAKE또한 매우 많이 쓰이는 키워드 추출기 중 하나이다. 게다가 개별 키워드의 지표를 동시에 보여주기 때문에 추출된 키워드뿐만 아니라 다른 키워드도 사용자가 고려하여 선택할 수 있다. 또한, 하이퍼 파라미터 최적화가 상대적으로 쉬우므로 손쉽게 키워드 후보군을 얻기 위해 많이 사용된다.

2.4.4 PKE

Python Keyphrase Extraction(PKE)[29]는 오픈 소스 파이썬을 기반으로 하는 키워드 추출 툴킷이다. 현재 YAKE를 포함한 여러 키워드 추출 모델을 구현하고 있다. PKE는 각 구성 요소를 쉽게 수정하거나 확장하여 새로운 접근 방식을 개발할 수 있는 종단 간 키워드 추출 파이프라인을 제공한다. PKE는 또한 최첨단 키워드 추출 접근 방식을 쉽게 벤치마킹할 수 있으며 SemEval-2010[30] 데이터셋에 대해 훈련된 감독 모델과 함께 제공된다. 사실상 PKE 자체가 키워드 추출 모델을 제안하지는 않지만, 다른

키워드 추출 모델들을 다양한 방식으로 앙상블 하거나 성능 비교를 하는 등의 이점이 있어 많이 사용되고 있다.

2.4.5 KeyBERT

KeyBERT[11]는 BERT 임베딩과 간단한 코사인 유사성을 사용하여 문서 자체와 가장 유사한 문서의 하위 단계를 식별하는 자체 컨텍스트 검색 시스템을 기반으로 하는 키워드 추출기이다. 구조를 간단하게 살펴보자면 다음과 같다.

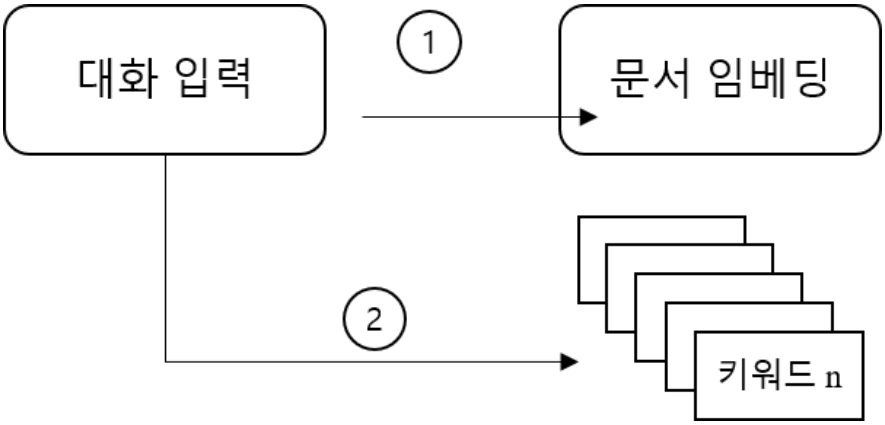

그림 11: keyBERT의 구조도 1

첫째로, 그림 11의 1번 과정과 같이 입력 대화를 사전 학습된 BERT를 사용하여 고정 크기 벡터로 임베딩한다. 2번 과정을 통해서 TF-IDF 벡터화 또는 Count 벡터화와 같은 Bag of Words 기법을 사용하여 키워드를 추출한다. keyBERT 이전의 키워드 추출기는 대부분 이 단계를 최적화하여 키워드를 추출한다. 이를 수식화하면 아래와 같이 요약할 수 있다. 문장

S를 BERT에 입력하여, 다음과 같이 상황별 특징 벡터 W를 얻는다.

$$W = \{\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_n\} = BERT(S) \tag{2.1}$$

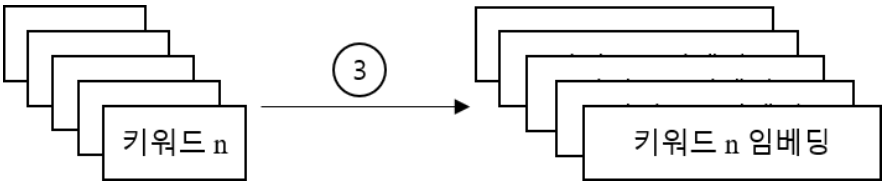

그림 12: keyBERT의 구조도 2

2번 과정에서 추출된 키워드는 그림 12의 3번 과정을 통해서 대화에 포함시키는데에 사용된 모델과 동일한 고정 크기 벡터로 임베딩한다.

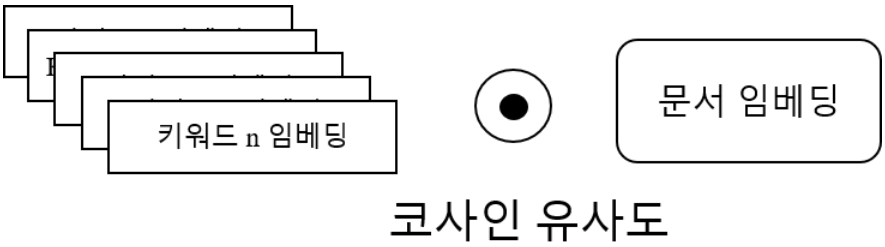

그림 13: keyBERT의 구조도 3

이제, 키워드와 문서는 동일한 벡터로 임베딩 되었다. 두 가지 정보를 이용하여 그림 13과 같이 코사인 유사성을 계산한다. 이 과정에서 코사인 유사성 점수가 가장 높은 키워드를 추출한다. 즉, 아래와 같은 수식으로 요약하여 설명할 수 있다.

$$Sim_i = cos(\mathrm{w}_i, W) \tag{2.2}$$

여기서 Sim_i는 단어 i의 단어 임베딩 벡터 w_i와 문장 임베딩 벡터 사이의 코사인 유사성이다. 키워드 후보가 추출되면, 인접 키워드 규칙을 통해 핵심 키워드를 추출한다.

아이디어 자체는 간단한 편이지만 고전적인 키워드 추출 방식에 BERT를 결합하여 기존 모델보다 훨씬 더 높은 성능을 보여준다.

2.5 평가 지표

대화 요약 생성은 자연어 처리의 한 영역으로 평가가 명확한 분야가 아니다. 자연어는 정성적인 평가에 의존성이 높은편이지만, ROUGE, BLUE 또는 BERTScore 등 다양한 정량 평가 지표가 사용되고 있다.

2.5.1 BLEU

Bilingual Evaluation Understudy(BLUE)[31]는 기계 번역과 같이 자연어 생성영역에서 결과와 사람이 직접 생성한 참조문이 얼마나 유사한지 비교하여 성능을 측정하는 방법이다. 문장의 길이와 단어의 중복을 고려하여 참조 문장과 생성 문장 사이의 겹치는 정도를 계산하는 지표라고 요약할 수 있다. 언어에 제약이 없다는 장점이 있지만, 참조 문장과 생성 문장의 정확한 매칭을 하기 때문에 요약과 같은 영역에서는 오히려 성능이 높지 않다. 사용하기 쉽고 한글에도 쉽게 적용이 가능하지만, 요약 생성의 지표로는 사용하기 힘든 특징때문에 본 연구에서는 채택하지 않았다.

2.5.2 ROUGE

Recall-Oriented Understudy for Gisting Evaluation(ROUGE)[13]는 텍스트 요약 모델의 성능 평가 지표이다. ROUGE는 요약,번역 등 자연어

생성 모델의 성능을 평가하기 위한 지표이며, 모델이 생성한 요약본 또는 번역본을 사람이 미리 만들어 놓은 참조본과 대조해 성능 점수를 계산한다. 개별 단어의 겹치는 빈도를 이용하는 것은 BLEU와 동일하지만, ROUGE는 Recall과 Precision을 이용하여 계산한다. ROUGE에서 Recall과 Precision은 아래와 같이 정의된다.

$$\begin{aligned} Recall &= \frac{\text{겹치는 단어 개수}}{\text{참조본의 총 단어 개수}} \\ Precision &= \frac{\text{겹치는 단어 개수}}{\text{생성본의 총 단어 개수}} \end{aligned} \tag{2.3}$$

ROUGE지표는 세부 측정 방법에 따라 ROUGE-N, ROUGE-L과 ROUGE-S 등으로 구분되어 사용된다. 첫 번째로 ROUGE-N은 N-gram Recall을 기반으로 계산하는 방식을 의미한다. N-gram에 대한 ROUGE-N은 다음과 같이 수식화 할 수 있다.

$$ROUGE-N = \frac{\sum_{S \in \text{참조본}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \text{참조본}} \sum_{gram_n \in S} Count(gram_n)} \tag{2.4}$$

두번째로 ROUGE-L은 가장 길게 매칭된 Recall을 구하는 것을 의미한다. 연속되는 단어가 아니어도 된다. 생성된 문장과 참조 문장이 완전히 일치하지 않아도, 떨어져 있는 단어의 연속성을 인식하기 때문에 체감상으로 정성적인 평가와 흡사한 점수를 보여준다. 마지막으로 ROUGE-S는 최대 2칸 내에 위치하는 단어 쌍의 Recall을 계산하는 방법이다. 이어지지 않아도 되므로 상대적으로 거리에 영향을 덜 받는다. DialogSum에서도 사용된 ROUGE-L은 요약 모델의 성능 측정 지표로 많이 사용된다. 본 연구에서도 ROUGE-L을 성능지표로 삼았다.

2.5.3 BERTScore

BERTScore[14]는 참조 문장과 생성 문장을 각각 BERT를 이용하여 유사도를 측정한다. 계산 방법은 아래와 같다. 그림 14를 보면, 참조문

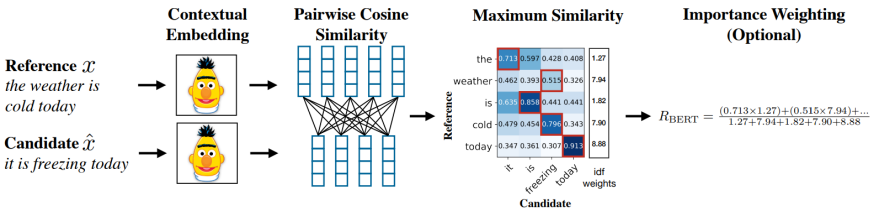

그림 14: BERTScore의 계산 방법[14]

과 생성문을 BERT를 이용하여 Contextual Embedding를 구한 후 코사인 유사도를 이용하여 최대 유사도를 측정한다. 측정된 최대 유사도 값을 이용하여 Recall과 Precision을 계산한다. 기존의 방식과 달리 BERT를 이용하여 단어를 벡터화하기 때문에 다른 단어여도 유사도에 따라 지표가 높게 나올 수도 있다. 이에 따라 빈출도가 낮은 단어가 오히려 유사성 검증에 핵심이 되는 경우도 있기 때문에 IDF값을 추가로 이용하는 방법 또한 제안되기도 하였다. 정성적인 평가가 ROUGE-L보다도 좋게 나오지만, 요약문 자체가 BERT를 이용한 BART로 생성되기 때문에 BERTScore로는 지표 점수가 더 높게 나오는 경향이 있어서 본 연구에서는 주요 성능지표로 사용하지 않았다.

2.6 키워드 기반 요약

추출 요약에서 키워드가 도움이 된다는 것은 기존 연구들에세 많이 증명되어 왔다[32]. 최근에는 문서에서의 생성 요약에서도 키워드가 유의미하게 적용된다는 것이 연구되기도 하였다.

표 5: 키워드 기반 문서 요약 시스템의 성능 비교[33]

모델	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Baseline	36.44	15.66	33.42
KIGN	**38.95**	**17.12**	**35.68**

BART와 같은 모델이 텍스트 생성에서 뛰어난 성능을 보이지만, 요약 생성이라는 것은 해당 문장이 어떠한 목적이 있느냐에 따라 결과물이 다르게 나온다. 따라서, Key Information Guide Network(KIGN)[33]은 핵심 정보가 부족한 요약문이 생성되는 것을 극복하기 위하여 추출 모델을 이용하여 문장에서 키워드를 얻어내고, 인코딩하는 과정에서 키워드를 이용하여 생성 과정을 유도하는 방식을 제안하기도 하였다. 이러한 방식은 표 5에서와 같이 상당한 성능의 개선을 보여주었다.

이러한 이유로 키워드가 문서에서의 불필요한 정보를 줄여줄 수 있다고 보았다. 이러한 방식으로 키워드를 이용한 문장 요약은 Encoder를 개선하여 키워드가 생성요약에서도 성능을 향상시킨다는 연구 결과를 보였었었다[20].

다른 연구에서는 키워드 기반의 추출기가 몇 가지를 선택하는 구조를 제안하기도 하였다[34]. 해당 연구에서는 입력 문서에서 두드러진 문장 집합을 추출한 다음, 요약에 적합하게 정렬하여 최종적으로 향상된 요약을 생성하는 방식을 제안하였었다. 이러한 키워드 기반의 요약 방식들은 모두 기존 모델과 비교하여 향상된 성능을 보여준다. 하지만, 이러한 키워드 기반의 요약은 모두 문서 기반으로만 연구가 진행되었고, 다른 도메인에서는 유의미한 성능 향상을 보여주지 못하는 한계점을 지닌다. 따라서 본 연구에서는 키워드 추출기가 요약에 유의미한 성능 향상을 보여주는 것을 확인하고, 모델의 변경 없이 키워드 추출기를 융합하는 것을 제안한다.

제 3 장

본론

3.1 기존 모델 분석

그림 15: 기존 대화 요약 시스템의 구조

일반적인 대화 요약의 경우 그림 15와 같은 구조로 동작한다. 만약 대화 요약을 위한 데이터셋이 자체적으로 충분히 있는 경우에는 해당 데이터셋을 기반으로 학습하지만, 대화 요약의 경우 일반적으로 데이터셋이 충분치 않기 때문에 사전 학습 모델을 이용하여 대화 요약 결과물을 생성한다. 사전 학습된 모델이라고 해도 세부 파라미터는 환경에 맞춰서 설정하므로 별도의 설정이 필요한 데다가, 한번 설정을 맞추면 일반적으로 다시 바꾸는 것은 쉽지 않다.

추출 요약과는 다르게 생성 요약의 경우, 대화 요약 결과물 또한 전적으로 정량적 평가로만 할 수 없으므로 정성적인 평가로 이어지는 경우가 대부분이다. 결국에는 정량적으로 성능이 안 좋아도 정성적으로 결과가 좋은 경우가 많이 발생하기도 한다.

위와 같은 이유로, 확연한 성능 개선을 보여주는 SOTA 모델이 제안되지 않는 이상, 한번 설정된 사전 학습 모델이 변경되는 것은 쉽지 않다. 따라서, 본 논문에서는 사전 학습 모델을 변경하지 않는 수준에서 대화

요약 결과물의 성능을 향상할 방안을 제안하고자 한다.

3.2 개선 모델 제안

기존 연구에서는 seq2seq 모델에서 Encoder와 Decoder 모두에 키워드를 적용하여 문장 요약 방법을 제안하였다. 입력 문장과 키워드 사이의 상호 작용을 조사하여 소스 정보를 필터링하기 위해서 키워드를 이용하여 선택적 인코딩 전략을 사용하였는데, 이러한 작업은 다중 작업 학습과 키워드 중심의 요약이 기본 모델보다 더 나은 성능을 달성하는 것을 보여주었다. 하지만 이 방식은 키워드를 추출하는 과정이나 근거가 명확지 않고, 해당 키워드가 유의미한 키워드인지에 대한 검증이 충분치 않기 때문에 키워드가 성능 개선에 영향을 끼친 것인지 알 수가 없었다. 따라서 본 연구에서는 키워드가 성능 개선에 확실한 영향을 끼치는 것인지 확인하기 위하여 이를 분리하여 융합하는 방법을 제안한다.

주어진 대화가 n개 있다고 할 때, 대화 D는 $D = \{(u_1), (u_2),...,(u_j)\}$로 정의할 수 있다. 대화 요약의 목적은 요약문 D에 대한 대화 요약문 S를 생성하는 것이다. 즉, 대화 요약은 조건부 확률 $P(S|D;\theta)$을 최대화하는 시스템을 훈련하는 것을 목표로 한다고 할 수 있다. 요약 외에도 본 연구에서는 E의 추가 입력 키워드로 대화 요약 문제를 공식화한다. 대화 요약을 개발하기 위해 BART[12]와 같은 사전 훈련된 언어 모델을 기반으로 seq2seq 모델을 훈련한다. 본 연구는 각 대화 D에서 키워드 K를 추출하여 대화 요약 시스템의 입력에 넣어서, 키워드 기반 대화 요약 시스템을 구축하였다. 간단히 말해서 키워드 기반대화 요약의 최종 목표는 $K = E(D)$일 때, 조건부 확률$P(S|D,K;\theta)$을 최대화하는 것이라고 할 수 있다.

본 연구는 키워드 추출기를 사용하여 알고리즘 1과 같은 방식으로 입

Algorithm 1 키워드를 추출하여 입력하는 알고리즘 흐름도

Input : 대화 입력 D, 추출된 키워드 K
Output : 대화에서 추출된 키워드 S^*

```
S* ← []
for each d ∈ D do
    K ←d로부터 추출된 키워드
    S ←< keyword >
    for each k ∈ K do
        S = S + k + < /s >
        S*.add(S)
    end for
end for
return S*
```

력 텍스트를 변경하고, 사전 훈련된 언어 모델을 사용하여 개선된 요약을 제안한다. 키워드 기반 대화 요약 시스템의 전반적인 흐름을 그림 16를 확인할 수 있다. 먼저 키워드 추출기를 이용하여 D에서 키워드를 추출한다. 추출된 키워드는 대화에서 언급된 순서에 따라 재정렬하고, 이를 기반으로 입력 텍스트에 키워드 K를 추가해준다. 완성된 입력 텍스트는 기존에 사용되던 요약 모델을 통해 요약문 S를 생성한다.

크게 보면, 본 연구의 시스템은 사전 훈련된 언어 모델을 기반으로 *키워드 추출*과 *키워드 인식 요약기*로 구성된다.

본 연구는 챕터 2.4에 설명된 다양한 키워드 추출기의 정보를 효율적으로 활용하는 키워드 기반 요약 시스템을 제안하고자 한다. 각 키워드 추출기의 목적이나 성능에 따라 키워드 기반 대화 요약 모델의 성능 또한 유동적으로 변하는 것을 확인할 수 있었다. 예를 들어 본 논문에서 제안하는 방식은 DialogSum 데이터셋을 기반으로 성능을 최대화하였기 때문에, 실제 적용하는 타겟에 따라 다른 키워드 추출기가 더 성능이 좋을 수도 있다. 본 연구가 제안한 요약 시스템은 seq2seq 기반 사전 훈련된 언어 모델

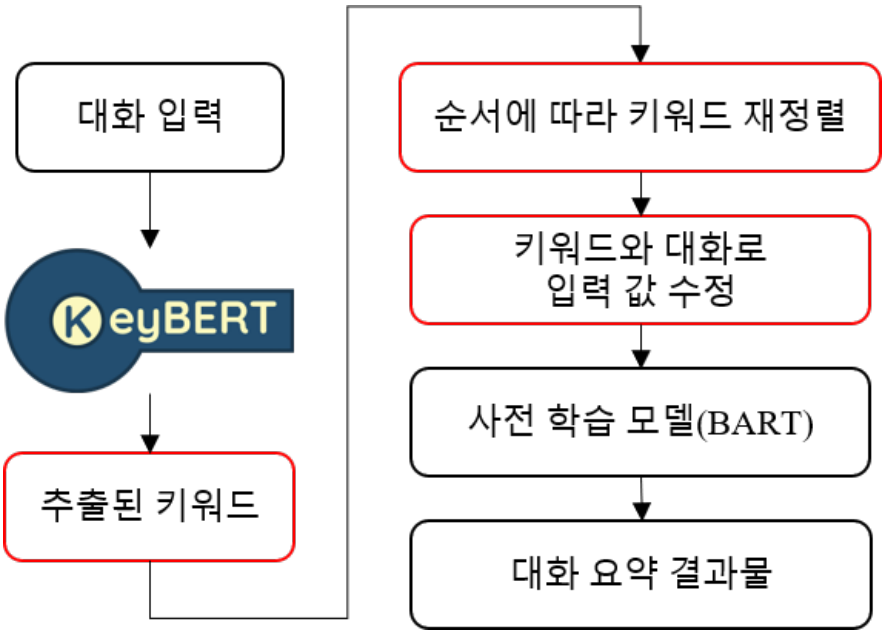

그림 16: 개선된 키워드 기반 대화 요약 시스템

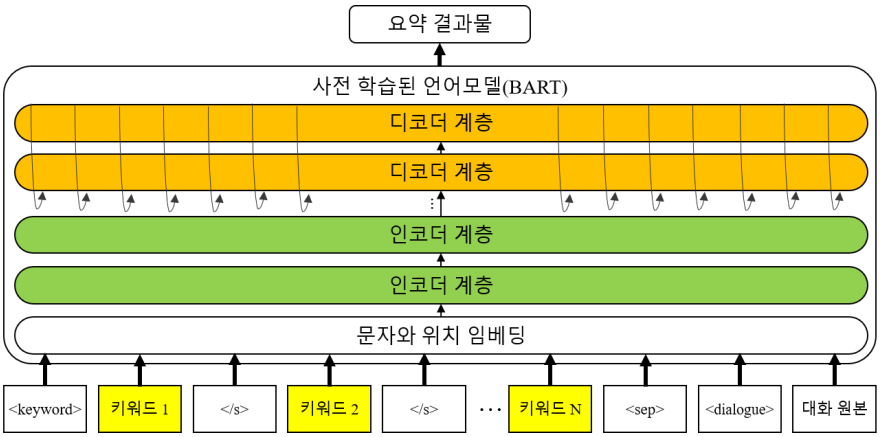

그림 17: 제안된 키워드 기반 대화 요약 시스템의 전반적인 아키텍처

BART를 기반으로 한다.

본 연구는 제안된 키워드 기반 요약 시스템의 전반적인 아키텍처를

그림 17에서 설명한다. 입력 대화상자에 특수 토큰으로 키워드를 추가한 후 사전 훈련된 언어 모델을 사용하여 내부적으로 내장하고, Encoder-Decoder와 seq2seq를 통해 추가 모델 변경 없이 새로운 요약 대화상자를 생성한다. 먼저 아래와 같은 대화 상자 D를 사용하여 키워드 추출기에서 키워드 $K = \{k_1, k_2, ..., k_m\}$를 추출한다.

$$K = \{\mathrm{k}_1, \mathrm{k}_2, ..., \mathrm{k}_m\} = E(D) \tag{3.1}$$

추출된 키워드를 그대로 사용하는 경우에는 성능 향상이 크지 않았다. 이후 대화의 구어체적 특성이 영향을 끼친다고 가정하여, 대화에서 키워드가 언급된 순서대로 키워드를 재정렬한 결과, 더 높은 성능 향상을 확인할 수 있었다. 결과적으로 키워드 K를 추출한 후, D에 언급된 순서에 따라 아래와 같은 알고리즘 2에 의해서 다시 정렬한다.

Algorithm 2 키워드 정렬 알고리즘

Input : 대화 입력 D
Output : 대화에서 추출된 키워드 K^*
1: $K \leftarrow$키워드 추출기로부터 추출한 키워드 리스트
2: $O \leftarrow \{\}$
3: **for each** $k \in K$ **do**
4: $O \leftarrow$D로부터 획득한 k의 위치
5: **end for**
6: $K^* \leftarrow$O에 저장된 k의 위치값을 기준으로 정렬한 리스트
7: **return** K^*

그런 다음 아래 수식과 같이 키워드 K를 대화 D와 합쳐서 요약 시스템에서 키워드를 활용하기 위한 BART의 입력으로 구성한다. 더 상세하게 설명하자면 먼저 입력 텍스트에 특수 토큰으로 *keyword* 및 *dialogue*을 추가한다. 키워드 추출기를 사용하여 n개의 키워드를 추출한 후 각 키워드가 입력에 분리 토큰인 s을 추가하여 n개의 키워드를 순서대로 배치한다. 그

리고 키워드 끝에 *sep* 토큰을 추가한 다음, 대화 원본을 *dialogue* 토큰과 함께 추가한다. 마지막으로 사전 훈련된 언어 모델 BART를 이용하여 주어진 대화에 대한 요약 S를 생성한다.

$$S = \{\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_n\} = BART([K, D]) \tag{3.2}$$

제 4 장

실험 결과 및 평가

제안된 모델을 실험을 통해 검증하였다. 챕터 4.1에서는 실험 데이터셋을 선택한 과정에 대해서 서술하였다. 챕터 4.2에서는 사전 학습 모델 및 키워드 모델을 선택하기 위한 실험을 진행하였다. 또한 키워드의 경우 추가적인 파라미터 조정이 필요하였고, 키워드의 순서에 대한 실험을 진행하였다. 챕터 4.3에서는 최종적으로 선택된 사전 학습 및 키워드 모델을 이용하여 KADS 모델을 구성하여 기본 모델과 성능을 검증하였다. 챕터 4.4에서는 KADS와 기본 모델간의 세부 성능 검증을 위해 다양한 환경에서 실험을 진행하였다. 또한 성능 향상에 비해 증가하는 학습 시간을 확인하였고, 키워드가 실제 성능에 유의미한 영향을 끼친것인지 실험을 진행하였다.

4.1 데이터셋 비교

표 6: 대화 생성 요약 데이터셋 비교

데이터셋	스타일	시나리오	대화 개수	데이터 사이즈
DialogSum	구어체	일상 대화	13,460	1.8M (token)
SAMSum	문어체	온라인	16,369	1.5M (token)
TweetSumm	문어체	온라인	1,100	1.8M (token)

우리는 세 가지 공개된 대화 요약 데이터셋인 DialogSum, SAMSum 그리고 TweetSumm을 비교하였다[35]. 표 6에 따르면, DialogSum의 대화

수는 TweetSum보다 10배 이상 많았으며, 대화 요약을 위한 대규모 신경망 모델을 훈련하는 데 도움이 될 수 있다. 단순히 대화 개수만 비교하면 SAMSum이 3,000개 정도 더 많지만, DialogSum의 평균 대화 길이가 SAMSum보다 평균적으로 39.8%가량 더 길다. 또한 DialogSum의 관점이 본 연구의 실질적인 취지와 부합하는 언어 스타일, 즉 구어체라는 점이 중요했다.

DialogSum을 이용하는 것은 다운스트림 애플리케이션의 관점에서 매일 말하는 대화를 요약하면 비즈니스와 개인 요구 사항 모두에 도움이 될 수 있다. 또한, 방법의 관점에서 DialogSum이 큰 규모의 긴 대화 데이터를 가지고 있어 신경망 모델을 이용한 대화 요약 연구를 쉽게 할 수 있다. 마지막으로, 대부분의 대화 데이터셋은 부적절하게 잘려있거나, 충분히 길지 않은 대화 또는 채팅 대화를 기반으로 하는 무용의 일상 대화로 특징지어진다. 그러나 DialogSum은 이러한 한계를 극복하기 위하여 유의미한 대화를 정제하여 실제 대화 데이터셋을 제공한다. 따라서 우리는 DialogSum을 이용하여 키워드 기반 대화 요약의 성능을 측정하게 되었다.

4.2 실험 상세

표 7: BART와 T5를 이용한 DialogSum에 대한 요약 성능 비교

모델	ROUGE-1	ROUGE-2	ROUGE-L
BART			
base	44.8874	19.6440	37.0678
large	46.1996	21.0814	38.8086
T5			
base	41.5242	16.4631	33.6254
large	42.1325	17.3326	34.4822

앞서 챕터 2.3에서 비교해보았듯이 요약 모델은 널리 사용되는 사전 훈련된 언어 모델 2개에 대해서 각기 다른 데이터셋 2개씩을 선택하여 4개의 성능을 비교해보았다. BART는 노이즈 제거 자동 인코더 작업을 사용하여 대규모 말뭉치에서 사전 훈련된 인코더 디코더 트랜스포머 모델이다. 그리고 T5[36]는 모든 NLP 작업을 텍스트 대 텍스트 문제로 취급하며, 같은 모델, 목표, 훈련 절차 및 디코딩 프로세스를 다양한 다운스트림 작업에 적용할 수 있게끔 학습된 모델이다.

BART와 T5를 이용하여 DialogSum에 대해 요약 성능을 비교해 보았다. 표 7에서 알 수 있듯이, BART-large 모델은 그중 가장 우수한 ROUGE 성능 지표 점수를 제공했다. DialogSum에는 학습 12,460개, 유효성 검사 500개 및 테스트 500개 세트로 구분되는 13,460개의 대화 상자가 포함되어 있다. 대화 요약을 위해 비교 결과 가장 성능이 높게 나오는 BART-large를 사용하고 5,000개의 학습 데이터와 200개의 유효성 데이터로 미세 조정한다. 학습 비율은 $3e^{-5}$으로 설정하였다.

4.2.1 키워드 추출기 선택 전략

표 8: 각 키워드 추출기의 기본 매개 변수만을 적용한 KADS의 DialogSum에 대한 성능 비교

모델	ROUGE-1	ROUGE-2	ROUGE-L
기본 모델	46.1996	**21.0814**	38.8086
KADS			
KeyBERT	44.5321	20.2534	37.2117
RAKUN	44.9626	20.0242	37.2444
RAKE	**45.6349**	20.1357	**38.1242**
YAKE	45.8326	20.0143	37.7641
PKE	44.9253	19.9257	37.8242

기존 키워드 추출기는 가장 많이 언급되는 N개의 단어, TF-IDF, TextRank를 결합하여 키워드를 얻는 방식을 사용했다. 그러나 BERT 이후, BERT를 사용하는 키워드 추출기의 성능이 기존 방법[37]보다 우수하다는 연구 결과가 나와 키워드 추출기의 성능이 향상되었다.

대화의 키워드는 키워드 추출기를 사용하여 추출되며 BART 요약을 수행하기 위한 특수 토큰으로 지정된다. 이 과정에서 현재 잘 알려지고 통상적으로 많이 사용되는 키워드 추출기들의 성능을 비교하여, 어떤 키워드 추출기를 KADS에 적용하느냐가 매우 중요하다. 다양한 키워드 추출기의 경우 각 도메인에서 최고의 성능을 얻기 위해 설정된 기본 매개변수로 설정되어 있다. 따라서 먼저 기본 매개 변수만 사용하여 성능을 비교하였다. 그 결과, 6개의 추출기를 선택하였고, 표 8를 통해 그 결과를 확인하였다. 표 8에서 보이듯 RAKE 추출기가 최고의 성능을 발휘했다. 그러나 각 추출기의 매개 변수는 해당 도메인에서 최고의 성능을 보이는 값일 뿐, DialogSum에 대해서는 해당 변수를 변경함에 따라 성능이 달라졌다. 따라서, 각각에 대해 최고의 성능을 나타내는 매개 변수로 실험을 진행할 필요가 있었다.

키워드 추출기가 유사한 결과를 생성하는 데는 이유가 있다. 키워드를 다양화한다면, 그것들은 문서를 집합적으로 나타낼 가능성이 작다. 따라서 결과를 다양화하려면 키워드의 정확성과 다양성 사이의 미묘한 균형이 필요하다. 본 연구는 결과를 다양화하기 위해 두 가지 알고리즘을 사용했다.

- 최대합 유사성[38]
- 최대 한계 관련성[39]

본 연구는 키워드 후보 간의 유사성을 최소화하면서 후보자의 문서 유사

성을 극대화하고자 한다. 상위 20개 키워드를 선정해 서로 가장 덜 닮은 5개를 선택하였다. 중복성과 텍스트 요약 작업에 대한 효과의 다양성을 최소화하는 최대 한계 관련성(MMR) 방법을 사용하여 결과를 다양화하였다. 각 키워드 추출기에 해당 방법을 적용하고 매개 변수를 세부조절하여 성능을 비교해 보았다.

표 9: 각 키워드 추출기가 DialogSum에 대해서 최대 성능을 나타내는 매개 변수를 적용한 모델의 성능 비교

모델	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
기본 모델	46.1996	21.0814	38.8086
KADS			
KeyBERT	**46.8676**	**21.9106**	**39.1413**
RAKUN	45.8668	20.9832	38.3474
RAKE	46.2899	21.0008	38.9808
YAKE	46.2077	20.9943	38.5658
PKE	45.8123	20.7648	38.3878

표 9에 나타난 바와 같이, keyBERT가 최적화된 후 성능이 가장 높았다. 따라서, DialogSum에 최적화된 keyBERT를 기반으로 KADS 모델을 설계하였다.

4.2.2 키워드 순서

키워드 추출기로 얻은 대부분 키워드를 가장 높은 정확도 순서로 나열하였다. 하지만 대화의 경우, 한 번 언급된 단어가 대명사로 지속적으로 언급되며 중요도가 높아지는 경우가 빈번하다. 이러한 대화의 특징때문에 단순히 키워드의 정확도만으로는 성능 향상이 미비할 것이라고 추측하였다. 따라서, 본 연구는 추출된 키워드를 다양한 방법으로 정렬하여 실험을 진행하였다.

표 10: DialogSum에서 키워드를 정확도와 언급 순서 순으로 정렬한 모델의 성능 비교

모델	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
기본 모델	46.1996	21.0814	38.8086
KADS			
정확도	46.8676	21.9106	39.1413
언급 순서	**47.2237**	**22.1353**	**39.8665**
언급 순서(역방향)	47.4271	21.9129	38.8780
빈도순	47.1117	21.2716	38.4360
랜덤	47.2033	21.7366	38.8307

5가지 경우로 키워드를 정렬하였다.

- 키워드 추출기에서 추출된 정확도 기준으로 정렬
- 단어가 언급된 순서대로 정렬
- 단어가 언급된 순서의 반대로 정렬
- 단어가 많이 등장한 순서대로 정렬
- 랜덤으로 정렬

대화의 특성상 처음 언급된 단어가 이후에도 중요할 수 있다고 생각하여 언급된 순서로 먼저 정렬을 하였다. 마찬가지로 마지막에 언급하는 단어가 더 중요할 수도 있기 때문에 역순으로도 정렬을 하였다. 또한 키워드 추출기에서 단어의 횟수에 따른 가중치로 키워드를 추출하겠지만, 실제 대화에서 언급이 많이 된 순서대로 정렬을 하였다. 마지막으로 추출된 키워드를 무작위로 섞어서 10번의 실험을 거쳐 평균을 계산하였다. 그 결과 표 10를 통해 대화에서 키워드가 언급된 순서대로 정렬하면 요약의 성능을 향상시킬 수 있음을 확인하였다.

4.3 실험 결과

표 11: BART와 T5를 이용한 DialogSum에 대한 KADS의 요약 성능 비교

모델	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
기본 모델				
BART-base	44.8874	19.6440	37.0678	0.9175
BART-large	46.1996	21.0814	38.8086	0.9189
T5-base	41.5242	16.4631	33.6254	0.9152
T5-large	42.1325	17.3326	34.4822	0.9154
KADS				
BART-base	45.9874	20.9440	38.1678	0.9190
BART-large	**47.2237**	**22.1353**	**39.8665**	**0.9192**
T5-base	44.2605	18.8368	36.2043	0.9162
T5-large	45.2232	18.9618	37.7235	0.9164

표 11에 기본 모델과 KADS를 이용하여 BART 및 T5에 적용한 대화 요약 성능을 제시하였다. BART를 기준 시스템으로 사용하고 T5[40]에 대해선 추가 실험을 수행했다. 그 결과 본 연구는 KADS가 모든 경우에 개선을 보였다는 것을 확인할 수 있었다. 표 11에 나타난 바와 같이 기존 BART-large에서 약 2.7%, T5-base에서 7.6%, T5-large에서 9.4%의 성능 향상을 보였다. 추가로 BERTScore에 대한 지표도 검증해본 결과, 약 0.03%의 성능 향상을 보였다. 두 가지 지표를 모두 확인한 결과 기본 모델의 성능이 떨어질수록 개선 효과가 높은 것으로 나타났다.

4.4 평가

KADS가 다양한 환경에서 성능이 향상되는 지 확인하고, 그에 따른 학습 시간 증가량을 실험하였다. 챕터 4.4.1에서는 KADS가 DialogSum 뿐만 아니라 다른 대화 데이터셋에서도 유의미한 성능 향상을 보이는지

검증하였다. 챕터 4.4.2에서는 성능이 향상됨에 따라 학습시간이 얼마나 증가하는 지 실험하였다. 챕터 4.4.3에서는 키워드가 대화 요약에서 성능에 영향을 끼치는 게 맞는 지 실험을 통해 검증하였다. 챕터 4.4.4에서는 대화 요약의 한계인 적은 데이터셋에서도 성능이 향상될 수 있는 지 실험하였다.

4.4.1 다른 데이터셋에 대한 실험

표 12: SAMSum과 TweetSumm 데이터셋에 대한 기본 모델과 KADS 모델간의 성능 비교

모델	ROUGE-1	ROUGE-2	ROUGE-L
SAMSum			
기본 모델	51.9170	27.6903	43.3052
KADS	**52.0063**	**27.9083**	**43.4162**
TweetSumm			
기본 모델	42.2314	19.2241	35.5624
KADS	**42.3342**	**19.3244**	**35.7624**

요약 기법은 도메인에 따라 다르게 적용될 수 있으며, 이러한 특성 때문에 각 도메인에 대한 최상의 성능이 다르다[41]. 키워드 적용은 SAMSum 및 TweetSum 데이터셋에 사용된다. SAMSum은 간결한 대화와 의미 없는 말이 자주 등장하는 것이 특징인 정제되지 않은 대화 형태였다. 반면에 TweetSumm의 고객 상담 내용 중 상당 부분은 이미 요약되어 있었다. 그러나 표 12에 나타난 바와 같이 SAMSum과 TweetSumm에 대한 KADS 모델의 적용은 성능 향상을 보였으나, 변화는 미미하였다.

표 13: DialogSum에서 기본 모델과 KADS 모델간의 학습 시간 비교

모델	Total time	Step per second
기본 모델	1574.0059	2.4750
KADS	**2556.2950**	**3.0470**

4.4.2 시간 비용 비교

KADS방식을 적용하기 위해서는 키워드 추출이라는 과정이 추가되어야 한다. 키워드 추출기를 적용하면 성능이 향상될 수 있지만, 훈련 시간은 성능 향상에 반비례하였다. 일반적으로 성능 및 균형에 관해 이야기할 때 메모리, 시간 및 스토리지를 언급한다. 그러나 메모리 및 스토리지 규모가 증가함에 따라 시간 비용이 상대적으로 많이 증가했다. 따라서 이를 성능[42]와 비교하였다. 표 13을 통하여 키워드 추출기를 적용하면 훈련 시간이 늘어나고, 이는 사용하는 추출기에 따라 크게 달라지므로 적절한 모델을 선택하는 것이 필수적이라는 것을 확인하였다.

4.4.3 KADS에 대한 키워드 검증

표 14: DialogSum에 대해서 무작위 키워드와 KADS 키워드를 적용한 모델의 성능 비교

모델	ROUGE-1	ROUGE-2	ROUGE-L
기본 모델	46.1996	21.0814	38.8086
무작위	44.3636	20.0146	37.1262
KADS	**47.2237**	**22.1353**	**39.8665**

자연어 처리는 블랙박스로 처리되기 때문에 결과가 개선되었다는 것만으로는 완벽하지 않다. 따라서 키워드가 실제로 요약 성능에 유의미한 성능 변화를 일으키는 지 확인이 필요하였다. 표 14을 통해 키워드가 무

작위로 추출되어 특수 토큰으로 입력되는 것과 KADS간의 성능을 비교하였다. 키워드를 특수 토큰으로 입력하면 특정 BART 매개 변수[43]의 무게가 증가할 수 있다는 것을 발견했다. 본 실험은 무작위 키워드 선택으로 요약 성능 측정을 10번 실행함으로써 평균 점수를 계산하였다. 때때로 일부 대화는 KADS와 유사한 성능을 보여주지만, 일반적으로 성능이 저하되었다.

결과적으로, 우리는 KADS에 비해 성능이 떨어지는 것을 관찰했다. 무작위 키워드와 KADS의 요약 결과를 눈으로 비교하여, 정량적인 평가뿐만 아니라 표 1 대화에 대해서 정성적인 부분에서도 실제 성능이 향상되었다는 것을 표 15을 통해서 확인할 수 있었다.

표 15: KADS와 무작위 키워드 방식을 적용한 요약 예시

요약본
Person1 is interviewing Person2 about Person2's ability and previous experience.
기본 모델
Person1 interviews Person2.
KADS
Person1 asks Person2's major, the past work, and the reason to do the job.
무작위 키워드
Person2 says I am a student engineer who mainly took charge of understanding of the mechanical strength and corrosion resistance of various materials. I think I can do it well.

4.4.4 적은 데이터에서의 성능 비교

대화 데이터의 경우, 정제된 자료가 거의 없고 대부분 정제되지 않은 상태로 제공되거나 데이터의 절대적인 수가 적어서 학습 중에 상당한 어려움이 있었다. 또한 본 연구에서 사용된 DialogSum의 데이터의 양도 충분하지 않았다. 그러나 실제로는 대화 요약이 더 적은 데이터[44]로 시

표 16: DialogSum에서 무작위로 샘플링된 데이터의 절반과 전체 데이터로 훈련된 시스템과의 성능 비교

모델	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
기본 모델			
전체	46.1996	21.0814	38.8086
절반	44.0717	20.1916	36.6705
KADS			
전체	**47.2237**	**22.1353**	**39.8665**
절반	45.2720	20.5406	38.1776

도되기도 한다. 따라서 데이터가 적어도 키워드 기반 방식인 KADS가 효율적인지 확인할 필요가 있었다.

표 16와 같이 기존 데이터의 절반만 무작위로 선택하여 성능 결과를 비교한 결과, KADS 방식이 더 적은 데이터로 더 큰 효과가 있었다. 즉, 데이터가 적은 대화 영역에서는 KADS와 같은 방식으로 유의미한 성능 향상을 기대할 수 있다.

제 5 장

결론

5.1 요약 및 결과

키워드 정보를 효율적으로 활용하여 대화 요약 시스템의 성능을 향상하는 대화 요약 시스템인 KADS를 제안하였다. 기존의 대화 요약 모델들은 문서 요약 모델을 기반으로 연구되었고, 대화 요약문도 생성하는 것이 가능하였다. 하지만 대화의 구어체적 특징 때문에 평가 지표 점수는 높아도 실제 요약 결과물은 많은 한계를 보였다. 대화 요약의 접근 방식에는 추출과 생성 요약이 포함된다. 이를 기반으로 키워드나 문맥 등을 고려한 다양한 대화 요약 모델이 제안되었지만, 성능 개선에 비하여 향후 발전 가능성이 모호한 경우가 많았다. 하지만 본 실험 결과를 통해 추상 요약의 성능이 키워드에 의해 향상될 수 있음을 확인하였다.

또한, 기존의 모델들을 활용한 다양한 대화 요약 플랫폼이 상용화된 시점에서 얼마 되지 않는 성능의 개선을 위하여 기존의 모델을 바꾸는 것은 현실적으로 불가능하다고 보았다. 따라서 본 연구는 KADS가 데이터셋에 따라 키워드를 학습시키지 않고도 사전 훈련된 키워드 추출기를 사용하여 성능을 향상할 수 있음을 입증했다. 이 과정에서 사전 학습 모델의 성능이 입력값의 변화만으로 성능이 변할 수 있음을 실험을 통해 확인하였다. 특히 대화 요약의 경우, 정확도가 높은 키워드일수록 요약의 성능 자체가 향상될 수 있음을 증명하였다. 본 연구는 단순히 키워드를 추출하는 데에 국한되지 않고 키워드의 순서 또한 의미가 있음을 알아냈으며,

성능이 좋지 않은 대화 요약 또는 데이터셋이 부족한 상황에서 오히려 더 높은 성능 향상을 보여주었다.

5.2 의의 및 한계

데이터셋이 부족하고 구어체적인 특성 때문에 대화 요약의 성능이 일정치 않다는 점을 고려했을 때, 본 연구는 실제 대화 요약을 사용하는 환경에서 더 큰 성능 향상을 체감할 수 있음을 시사한다. KADS의 방식은 기존 모델의 변경 없이 성능이 향상될 수 있음을 입증한 것이고, 이는 상용화된 다양한 대화 요약 플랫폼에서 진입 장벽 없이 쉽게 접근 및 연구할 수 있다는 가능성을 의미한다. 하지만 각 데이터셋에서 어떤 키워드 추출기가 최적의 성능을 보이는지 찾아야 하며, 이 과정에서 각 키워드 추출기의 세부 파라미터를 조절하는 과정이 필요하다.

5.3 향후 계획

향후 우수한 키워드 추출기가 새롭게 제시된다면 KADS의 성능을 추가로 향상할 수 있다. 즉, 대화 요약에서는 사전 처리가 기존의 문서 요약보다 더 중요하다는 것을 시사한다. 더 나아가 키워드 추출기와 대화 요약이라는 비슷하지만 다른 영역의 자연어 처리 분야를 융합하여 성능을 향상할 수 있다는 점은 향후 자연어와 관련된 다양한 연구의 앙상블을 기대할 수 있다.

실제 현업에서 사용되는 대화 및 회의록 요약의 성능을 향상시키기 위해 새롭게 제안된 키워드 추출기를 적용한 KADS를 도입할 예정이다. 추가로, Language Model(LM)을 이용한 음성인식 분야와 키워드를 융합하여 인식률을 높일 수 있는 방안을 추가 연구할 계획이다.

참고 문헌

[1] H. Pratama, M. N. A. Azman, G. K. Kassymova, and S. S. Duisenbayeva, "The trend in using online meeting applications for learning during the period of pandemic covid-19: A literature review," *Journal of Innovation in Educational and Cultural Research*, vol. 1, no. 2, pp. 58–68, 2020.

[2] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Dialoglm: Pretrained model for long dialogue understanding and summarization," *arXiv preprint arXiv:2109.02492*, 2021.

[3] R. He, L. Zhao, and H. Liu, "TWEETSUM: Event oriented social summarization dataset," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 5731–5736, International Committee on Computational Linguistics, Dec. 2020.

[4] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, (Hong Kong, China), pp. 70–79, Association for Computational Linguistics, Nov. 2019.

[5] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsumm: A real-life scenario dialogue summarization dataset," *CoRR*, vol. abs/2105.06762, 2021.

[6] D. Lee, J. Lim, T. Whang, C. Lee, S. Cho, M. Park, and H.-S. Lim, "Capturing speaker incorrectness: Speaker-focused post-correction for abstractive dialogue summarization," in *Proceedings of the Third Workshop on New Frontiers in Summarization*, pp. 65–73, 2021.

[7] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[8] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, (Berlin, Germany), pp. 280–290, Association for Computational Linguistics, Aug. 2016.

[9] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 1797–1807, Association for Computational Linguistics, Oct.-Nov. 2018.

[10] S. Lee, K. Yang, C. Park, J. Sedoc, and H. Lim, "Who speaks like a style of vitamin: Towards syntax-aware dialoguesummarization using multi-task learning," 2021.

[11] M. Grootendorst, "Keybert: Minimal keyword extraction with bert.," 2020.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

[13] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," *CoRR*, vol. abs/1904.09675, 2019.

[15] D. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational linguistics*, vol. 28, no. 4, pp. 399–408, 2002.

[16] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1–7, 2016.

[17] X. Fu, Y. Zhang, T. Wang, X. Liu, C. Sun, and Z. Yang, "RepSum: Unsupervised dialogue summarization based on replacement strategy," 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)*, (Online), pp. 6042–6051, Association for Computational Linguistics, Aug. 2021.

[18] J. Liu, D. J. D. Hughes, and Y. Yang, *Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs*, p. 2313–2317. New York, NY, USA: Association for Computing Machinery, 2021.

[19] L. Zhao, Z. Yang, W. Xu, S. Gao, and J. Guo, "Improving abstractive dialogue summarization with conversational structure and factual knowledge," 2021.

[20] H. Li, J. Zhu, J. Zhang, C. Zong, and X. He, "Keywords-guided abstractive sentence summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8196–8203, 2020.

[21] E. Collins, I. Augenstein, and S. Riedel, "A supervised approach to extractive summarisation of scientific papers," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL*

2017), (Vancouver, Canada), pp. 195–205, Association for Computational Linguistics, Aug. 2017.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[24] B. Skrlj, A. Repar, and S. Pollak, "Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation," *CoRR*, vol. abs/1907.06458, 2019.

[25] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory*, vol. 1, pp. 1–20, 2010.

[26] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[27] Wikipedia contributors, "Levenshtein distance — Wikipedia, the free encyclopedia," 2022. [Online; accessed 8-November-2022].

[28] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.," 1990.

[29] F. Boudin, "pke: an open source python-based keyphrase extraction toolkit," in *COLING*, 2016.

[30] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley, "Semeval-2010 task 1: Coreference resolution in multiple languages," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 1–8, 2010.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[32] N. Zhong, J. Liu, and Y. Yao, *Web intelligence*. Springer Science & Business Media, 2003.

[33] C. Li, W. Xu, S. Li, and S. Gao, "Guiding generation for abstractive text summarization based on key information guide network," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 55–60, 2018.

[34] Y. Liu, Q. Jia, and K. Zhu, "Keyword-aware abstractive summarization by extracting set-level intermediate summaries," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 3042–3054, Association for Computing Machinery, 2021.

[35] X. Feng, X. Feng, and B. Qin, "A survey on dialogue summarization: Recent advances and new frontiers," *CoRR*, vol. abs/2107.03175, 2021.

[36] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019.

[37] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, (Barcelona, Spain), pp. 404–411, Association for Computational Linguistics, July 2004.

[38] T. Werner, "A linear programming approach to max-sum problem: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1165–1179, 2007.

[39] S. Guo and S. Sanner, "Probabilistic latent maximal marginal relevance," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 833–834, 2010.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[41] M.-Y. Kan and K. McKeown, "Information extraction and summarization: Domain independence through focus types," 1999.

[42] S. G. Baker, "The summary test tradeoff: a new measure of the value of an additional risk prediction marker," vol. 36, 2017.

[43] N. B. Carnegie and J. Wu, "Variable selection and parameter tuning for bart modeling in the fragile families challenge," *Socius*, vol. 5, p. 2378023119825886, 2019.

[44] Y. Zou, B. Zhu, X. Hu, T. Gui, and Q. Zhang, "Low-resource dialogue summarization with domain-agnostic multi-source pretraining," *arXiv preprint arXiv:2109.04080*, 2021.

Abstract

Improvement of Dialogue Summarization Using Keyword Extractor

Chongjae Yoo
Graduate School of Practical Engineering
Seoul National University

Abstractive dialogue summarization aims to generate a short passage that contains important content for a particular dialogue spoken by multiple speakers. In abstractive dialogue summarization systems, capturing the subject in the dialogue is challenging owing to the properties of colloquial texts. Moreover, the system often generates uninformative summaries. In this paper, we propose a novel keyword-aware dialogue summarization system (KADS) that easily captures the subject in the dialogue to alleviate the problem mentioned above through the efficient usage of keywords. Specifically, we first extract the keywords from the input dialogue using a pre-trained keyword extractor. Subsequently, KADS efficiently leverages the keywords information of the dialogue to the transformer-based dialogue system by using the pre-trained keyword extractor. Extensive experiments performed on three benchmark datasets show that the proposed method outperforms the

baseline system. Additionally, we demonstrate that the proposed keyword-aware dialogue summarization system exhibits a high-performance gain in low-resource conditions where the number of training examples is highly limited.

Keywords : abstractive summarization, dialogue summarization, keyword extraction

Student Number : 2021-27215