인지과학, 제34권 제3호
Korean Journal of Cognitive Science
2023, Vol. 34, No. 3, 197~226
http://dx.doi.org/10.19066/cogsci.2023.34.3.003

오류 유형에 따른 생성요약 모델의 본문-요약문 간 요약 성능평가 비교*

이 승 수 강 상 우†
가천대학교 AI・소프트웨어학부

텍스트 생성요약은 자연어처리의 과업 중 하나로 긴 텍스트의 내용을 보존하면서 짧게 축약된 요약문을 생성한다. 생성요약 과업의 특성 상 본문의 핵심내용을 요약문에서 보존하는 것은 매우 중요하다. 기존의 생성요약 방법론은 정답요약과의 어휘 중첩도(Lexical-Overlap)를 기반으로 본문의 내용과 유창성을 측정했다. ROUGE는 생성요약 요약모델의 평가지표로 많이 사용하는 어휘 중첩도 기반의 평가지표이다. 생성요약 벤치마크에서 ROUGE가 49점대로 매우 높은 성능을 보임에도 불구하고, 생성한 요약문과 본문의 내용이 불일치하는 경우가 30% 가량 존재한다. 본 연구에서는 정답요약의 도움 없이 본문만을 활용해 생성요약 모델의 성능을 평가하는 방법론을 제안한다. 본 연구에서 제안한 평가점수를 AggreFACT의 라벨과 상관도 분석결과, 다음의 두 가지 경우 가장 높은 상관관계를 보였다. 첫 번째는 Transformer 구조의 인코더-디코더 구조에 대규모 사전학습을 진행한 BART와 PEGASUS 등을 생성요약 모델의 베이스라인으로 사용한 경우이고, 두 번째는 요약문 전체에 걸쳐 오류가 발생한 경우이다.

주제어 : 자연어처리, 텍스트 생성요약, 품질예측, 메타평가

---

* 이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2022R1A2C1005316).
이승수, 가천대학교 AI・소프트웨어학부 학생, E-mail: wlstjddl5916@gachon.ac.kr
† 교신저자: 강상우, 가천대학교 AI・소프트웨어학부, 경기도 성남시 수정구 성남대로 1342, 가천대학교 글로벌캠퍼스 AI관 419호
연구 분야: 자연어처리, E-mail: swkang@gachon.ac.kr

개요

본 연구는 문서 생성요약 모델의 성능을 측정하기 위한 비지도평가 방법을 제안한다. 본 연구에서는 요약문의 특성 별 평가를 위한 비지도평가 방법을 정의하고, 각 특성마다의 평가지표를 요약모델에서 발생하는 오류 유형 별로 확인한다. 요약문을 평가하기 위한 특성은 FFCI(Koto et al., 2020)를 참고해 비지도평가가 가능한 방법으로 변경한다. 또한 AggreFACT(Tang et al., 2022)의 오류 유형에 따라 수동평가와의 상관도를 측정한다.

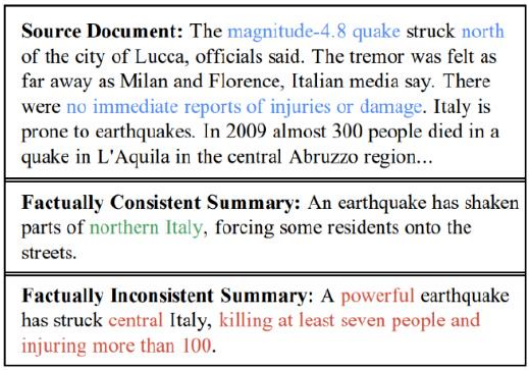
본 연구에서 제안한 비지도평가 방법을 통해 문서 생성요약에서 발생하는 Hallucination 현상의 정도를 측정한다. Hallucination은 생성요약(Kågebäck et al., 2014; Liu, 2019), 기계번역(Lee et al., 2018) 등의 텍스트를 생성하는 자연어생성(Naturul Lanuage Generation, NLG) 과제에서 부정확하거나(Schuster et al., 2021), 입력한 원문에는 없는(Ji et al., 2023) 정보를 생성하는 현상이다.

BART(Lewis et al., 2020)의 등장 이후, 자연어처리의 다양한 분야에 encoder-decoder 구조의 사전학습 기반 언어모델(Lewis et al., 2020; Zhang et al., 2020; Raffel et al., 2020)이 높은 성능을 보였다. 문서 생성요약(text summarization) 또한 활발히 연구가 이뤄지고 있는 자연어처리 분야이다. 문서요약은 긴 텍스트에서 핵심적인 내용을 보존하면서 보다 짧게 축약된 텍스트를 생성하는 자연어처리 분야다.(최경호&이창기, 2016; Dumais et al., 1995) 문서요약은 문제 정의에 따라 크게 원본 문서에서 중요한 몇 개의 문장을 선택하는 추출요약(extractive summarization)과 원문의 핵심 내용을 보존하면서 짧게 재구성하는 생성요약(abstractive summarization)으로 나눌 수 있다.(최경호 & 이창기, 2016; 김탁영 et al., 2022; Gudivada, 2018; Zhu, 2021) 추출요약은 원문에서 주요 문장을 선택하는 분류(classification) 또는 등급화(ranking)의 문제로 정의됐다(김탁영 et al., 2022). 생성요약은 요약할 문서를 입력받아 자연어 문장을 생성하는 언어 생성(natural language generation; NLG) 문제로 정의된다(김탁영 et al., 2022).

본문의 핵심내용을 보존해야 하는 문서요약 과제의 특성 상, 문서 생성요약 모델에서의 Hallucination은 모델의 신뢰도에 치명적이다. 최근에는 문서요약을 비롯한 다양한 자연어처리 분야에서 Hallucination의 발생정도를 측정하고(Thorne et al., 2018; Schuster et al., 2021; Gupta, Wu et al., 2021), 부정확한 텍스트를 그럴듯한 형태로 생성하려고하는 언어모델의 문제를 극복하기 위한 연구가 이뤄지고 있다.(Ji et al., 2023) 생성요약 벤치마크(Hermann et al., 2015; Narayan et al., 2018; See et al., 2017; Cachola et al., 2020)에서는 주로 정답 요약문과 lexical-overlap 기반 평가지표(Lin, 2004; Lavie & Agarwal, 2007; Papineni et al., 2002) 등을 통해 성능을 측정한다.

생성요약 평가지표에 대해서 매우 높은 성능을 보임에도 불구하고, 최근 연구에선 생성된 요약문의 약 30%는 본문과 내용이 일치하지 않는 문제가 존재하는 것으로 나타났으며(Z. Cao et al., 2018; Falke et al., 2019), 이는 아직도 도전적인 과제로 남아 있다(박은환 et al., 2021). 생성요약 결과 본문과 요약문 간의 내용이 일치하는 정도를 Factual consistency[1]라고 하며, 요약문에서

의 예시는 (그림 1)과 같다. 최근에는 생성요약 결과가 Facutally Consistent한 문제를 극복하기 위해, 본문-요약문 간의 내용이 일치하는 정도를 측정하는 Factual consistency 평가지표(Zhang, et al., 2020; Kryscinski et al., 2020; Vasilyev et al., 2020; Durmus et al., 2020)와 개선 방법론(S. Cao & Wang, 2021)에 대한 연구가 활발히 진행되고 있다. 하지만 정답요약에서 본문의 의역과 축약이 빈번한 추상요약 데이터셋의 특성 상(Narayan et al., 2018), 요약문에 대한 Factual consistency 여부에 대한 판단이 다소 주관적이고 편향될 수 있다는 문제(Falke et al., 2019)와 길이가 긴 본문에 대한 정답요약을 작성하는 데에 비용이 큰(Koto et al., 2020; Hardy et al., 2019) 문제 등이 존재한다. (Zhang et al. 2022; Zhao et al. 2022)은 자연어생성 모델의 확률을 특정 데이터셋에 적합하도록 조정(calibration)하는 방법으로 번역, 요약 등의 과제에서 높은 성능을 달성했다. 하지만 이러한 방법론은 학습에 사용되는 데이터셋이 편향되거나 정답이 잘못됐을 경우, 잘못된 정보를 그대로 학습하거나 증폭될 우려가 있다. 따라서 최근에는 생성요약 모델의 평가를 위한 비지도 평가 관련 방법론이 제안되고 있다.



(그림 1) 요약문에서 Factual Consistency가 발생한 예시

본 논문의 2장에서는 생성요약에 대한 소개 및 평가방법, 그리고 4장의 실험에서 사용한 AggreFACT(Tang et al., 2022) 데이터셋 등 관련 연구에 대해 설명한다. 3장에서는 본 논문에서 제안하는 생성요약 모델의 평가요소 및 비지도평가 방법에 대해 설명 한다. 4장에서는 생성요약 모델의 평가요소 별, 3장에서 제안한 비지도평가 방법론을 적용한 결과에 대해 설명한다.

1) 관련 연구들에서는 다음 단어들을 혼용하는 경향이 보여, 본 논문에서는 다음의 단어들을 Factual Consistency로 통칭한다: Faithfulness, Factuality

관련 연구

생성요약 벤치마크 및 생성요약 연구 동향

문서요약 벤치마크의 종류는 데이터 수집방법과 도메인, 그리고 요약 입출력의 형태에 따라 다양하다. 데이터 수집과정에서 사용한 한국어[2)3)], 다국어(Scialom et al., 2020; Ladhak et al., 2020) 등의 언어와 뉴스(Hermann et al., 2015; Narayan et al., 2018), SNS(B. Kim et al., 2019), 과학(Cohan et al., 2018), Dialogue(Gliwa et al., 2019) 등 도메인에 따라 다양한 버전이 있다. 본 논문에서는 일반적으로 가장 많이 사용하는 영어권의 문서요약 데이터만을 다룬다.

이외에도 문서요약의 분류는 본문의 수와 요약방법에 따라 다음과 같이 나뉜다. 문서요약은 요약의 입력으로 사용하는 원문의 개수에 따라 single document summarization과 multi document summarization으로 나뉜다. 문서요약은 생성하려는 텍스트 형태에 따라 keyword 혹은 sentence summarization, 또한 요약 과정에서 원문 외부의 정보를 얼마나 사용하는지에 따라 knowledge-poor 혹은 knowledge-rich summarization으로 나뉘는 등 다양한 구분이 있다.(Sizov, 2010) 본 논문에서는 문서 생성요약을 하나의 본문에 대해 외부 정보 없이 요약문장을 생성하는 과정으로 정의한다.

최근의 생성요약 모델은 자연어 생성(NLG) 모델을 전이학습하여 사용한다(박은환 et al., 2021). 생성요약에 사용하는 finetuning 방식에는 대조학습(Constrastive Learning)을 이용한 자기주도 학습(Self-supervised Learning) 방법(Liu & Liu, 2021; Liu et al., 2022)과 그래프와 같은 외부정보를 활용하는 Knowledge-Enhanced Text Generation 방법(Yu et al., 2022) 등이 있다.

기존 생성요약 모델의 평가방법

문서 생성요약 모델의 평가에는 일반적으로 생성한 요약문장과 정답문장을 비교하는 지도평가 방법을 사용한다(Narayan et al., 2018; Gliwa et al., 2019; B. Kim et al., 2019). 정답문장과 생성한 요약문 간의 비교는 두 문장 간의 어휘 중첩도(lexical-overlap)나 의미적 유사도(semantic textual similarity, STS)를 통해 비교한다.

어휘 중첩도를 측정하기 위한 평가지표로는 ROUGE(Lin, 2004)를 사용한다.(Koto et al., 2020) ROUGE를 측정하기 위해서 n-gram이나 최장 공통 부분 수열(Longest Common Subsequence, LCS) 단위로 단어의 중첩 비율을 측정한다. 문서요약의 평가지표로 ROUGE를 사용하는 경우, 연속되지 않은 문자열이나 동음이의어, 다의어 등에 대응하기 어렵다는 한계점이 있다.

따라서 어휘 중첩도 기반의 평가방법을 보완하기 위해 생성한 요약문에 대해 인간이 직접 평

2) DACON(한국어 문서 추출요약 AI 경진대회): https://dacon.io/competitions/ official/235671/data

3) 모두의 말뭉치(문서 요약 데이터): https://corpus.korean .go.kr/request/re ausetMain.do

가하는 수동평가(manual evaluation)를 함께 사용한다. 생성한 요약문에 대해 수동평가를 진행하는 경우, 크라우드 소싱 설문 플랫폼 Mechanical Turk[4]등을 통해 생성요약 결과에 대한 평가를 수량화한다.(Kryscinski et al., 2020; Maynez et al., 2020; Laban et al., 2022; Scialom et al., 2019)

요약과제에 대한 수동평가를 진행할 경우, 다음과 같은 한계점이 있다. 첫 번째로 요약과제의 경우, 본문의 길이가 길어 다른 자연어처리 과제에 비해 수동평가에 시간과 비용이 많이 든다. Mechanical Turk 등을 통해 데이터를 구축할 시에는 다른 과업 대비 50배 이상의 작업비용이 소요된다(Gabriel et al., 2021; Honovich et al., 2022). 두번째로, 생성한 요약문의 수동평가에는 작업자의 주관성이 포함될 수 있다. (Cachola et al., 2020; Narayan et al., 2018; See et al., 2017)에서는 수동평가의 객관성을 보장하기 위해 하나의 본문에 다수의 작업자를 투입했다. 작업자가 많아질수록 수동평가의 비용은 더욱 증가된다.

따라서 최근에는 생성요약 모델의 평가지표로써 어휘 중첩도인 ROUGE 점수와 함께, PLM 기반의 요약문에 대한 Factual consistency의 정도를 측정하는 평가모델의 점수를 사용한다.

Factual Consistency 평가지표의 종류

생성요약 모델에서 생성한 요약문의 입력된 본문에 대한 Factual consistency 평가지표는 평가모델의 학습에 사용된 데이터의 종류에 따라, 크게 의미적 유사도와 응용모델 기반의 평가방법으로 나뉜다. 자연어추론 모델 기반의 방법은 언어이해(Natural Language Understanding, NLU) 데이터로 학습된 자연어추론(Natural Language Infernce, NLI) 모델을 기반으로 본문과 요약문 간 의미적으로 유사한 정도를 측정한다. 응용모델 기반의 평가방법은 언어이해 분야(Wang et al., 2018)의 하위과제와 질의응답(Rajpurkar et al., 2016) 과제에 대해 높은 성능을 보인 모델의 점수를 본문-요약문 간 평가지표로 활용한다.

의미적 유사도 기반 평가방식은 모델의 유사도 비교 단위에 따라 토큰 단위와 문장 단위 평가 방법이 있다. 토큰 단위 평가방법에는 사전학습된 BERT(Devlin et al., 2018), BART(Lewis et al., 2020) 등의 의미적 표현 능력을 활용하여 토큰 단위 유사도를 점수로 사용하는 BERTScore (Zhang et al., 2020) 와 BARTScore(Yuan et al., 2021) 등이 있다.

응용 모델 기반 평가방식에는 언어모델의 NLU 사전학습 과제로 많이 사용되는 Cloze-Task 기반 평가 방식과 질의생성-질의응답 기반 평가방식이 있다.

Cloze-Task 평가는 언어모델 사전학습 과제로써 높은 성능을 보인 Text-Infilling 방법을 요약모델의 평가에 활용한다. (Devlin et al., 2018; Lewis et al., 2020) 등은 언어 모델의 사전학습 과제로써 본문의 빈칸을 채우는 Text-Infilling 과제가 효과적임을 증명했다. BLANC(Vasilyev et al., 2020)는 이러한 점에 착안해 Text-Infilling 방식을 요약문의 자동평가에 활용하는 Cloze-Task 기반 평가

4) Amazon Mechanical Turk: https://www.mturk.com/

방법을 제안했다. BLANC(Vasilyev et al., 2020)는 문서요약의 자동 평가지표로 활용하던 의미적 유사도나 질의생성-질의응답 방식 외에 PLM의 사전학습 objective로 많이 사용하는 MLM(Masked Language Modeling)(Devlin et al., 2018; Lewis et al., 2020; Zhang et al., 2020)을 활용했다. BLANC(Vasilyev et al., 2020)에서는 본문-요약문의 입력방식에 따라, 본문과 함께 요약문을 접합(concat)하여 사용하는 BLANC-help와 특정 데이터에 대해 모델을 finetuning하여 사용하는 BLANC-tune 방식을 제안했다.

질의생성-질의응답 평가는 주어진 본문 내에서 질문에 대한 정답을 찾는 질의응답과 기계독해 모델을 생성문장에 대한 평가점수로 활용한다. 질의생성-질의응답 기반의 평가지표의 종류로는 SummaQA(Scialom et al., 2019), SummaC(Laban et al., 2022), QAGS(Wang et al., 2020), FEQA (Durmus et al., 2020), QuestEval(Scialom et al., 2021), Qafacteval(Fabbri et al., 2022) 등이 있다.

SummaQA(Scialom et al., 2019)는 질의응답 데이터인 SQuAD에서 학습한 모델을 활용해 비지도 방식의 요약문 평가모델을 제안하였다. (Paulus et al., 2017; Pasunuru & Bansal, 2018; Arumae & Liu, 2019)에서는 질의생성-질의응답 모델 기반 평가모델을 다운스트림 과제와 같은 CNNDM 등의 요약 과제의 본문-정답요약 쌍을 활용해 finetuning하여 평가모델로 사용하였다. SummaQA (Scialom et al., 2019)에서는 요약문의 본문에 대한 Factual consistency 평가지표로 TLDR(Cachola et al., 2020)에서 학습 한 모델을 CNNDM에서의 성능(F1-Score)와 가능도(confidence)를 측정하였다.

QuestEval(Scialom et al., 2021)은 기존의 질의생성-질의응답 기반 Factual consistency 평가모델의 수동평가와의 상관도가 ROUGE(Lin, 2004) 대비 못 미치는 점을 보완하기 위해 두 가지 평가모델의 조화평균을 측정하도록 제안하였다. QuestEval(Scialom et al., 2021)에서는 SummaQA(Scialom et al., 2019), SummaC(Laban et al., 2022) 등은 Factual consistency(precision-based)와 관련이 있고, QAGS(Wang et al., 2020)는 질의응답 모델의 관련도(recall-based)와 관련이 있다고 주장하여, 둘의 조화 평균인 QuestEval을 새로운 평가지표로 제안하였다.

단일 Factual Consistency 평가지표의 한계점

최근의 생성요약 모델에서는 요약문에서의 Hallucination 현상을 개선하기 위한 방법들 이 제안되었다(S. Cao & Wang, 2021). 대부분의 관련연구에서는 BERTScore(Zhang et al., 2020), FactCC(Kryscinski et al., 2020) 등의 의미적 유사도 기반 평가지표로 Factual consistency 개선 정도를 평가하였다. 하지만 이후의 연구에서 단일 종류의 평가방식을 통한 Factual consistency 평가에 대한 문제를 제기하였다. FASum(Zhu et al., 2021)은 NLI 평가지표를 downstream 데이터에 finetuning할 경우 사전학습된 평가 모델의 사실 연속성 평가성능 저하 발생을 주장하였다.

또한 (Amplayo et al., 2022, Sun et al., 2022)에서는 평가모델을 검증 데이터에 finetuning하거나 요약문 생성모델과 같은 종류의 encoder-decoder 구조의 사전학습 모델을 사용하면, 해당 방법에

유리한 방향으로 성별, 인종 등의 social bias가 생긴다고 주장하였다. 따라서 언어 생성모델의 출력에 Factual consistency 평가모델의 능력을 종합적으로 평가하기 위해, downstream 과제에서 필요한 언어모델의 능력을 정의하고 종합적으로 비교해 평가하는 Meta-Evaluation에 대한 연구가 진행되었다.

생성요약 과업의 메타평가(Meta-Evaluation)

FFCI(Koto et al., 2020)와 GoFigure(Gabriel et al., 2021)에서는 생성요약 모델에서 필요한 특성에 대해 서로 다른 정의를 내리고, 각각의 특성 별로 평가지표와 수동평가 점수 간의 상관분석을 진행하였다.

FFCI(Koto et al., 2020)에서는 생성요약 모델의 요약능력 평가 기준으로 4가지 특성(Faithfulness, Focus, Coverage, Inter-Sentence Coherence)을 제안하였다. Faithfulness는 Factual consistency의 정도를 측정한 것이고, 나머지 3개의 특성은 정답요약을 활용하는 기존의 수동평가 방식을 모방하여 Focus와 Coverage는 재현율과 정밀도를, Inter-Sentential Coherence는 생성한 요약문의 일관성을 측정하였다. GoFigure(Gabriel et al., 2021)는 요약문의 Factual consistency를 종합적으로 평가하기 위한 특성을 평가점수의 일관성과 민감도 등에 따라 5가지로 정의하였다.

여러 평가지표를 활용한 Meta-Evaluation을 진행했음에도, Factual consistency 자동 평가를 위한 평가지표는 인간의 수동평가 점수와 매우 낮은 상관도를 보였고(Koto et al., 2020; Gabriel et al., 2021), 여러 종류의 평가모델을 앙상블 하여 요약모델 평가에 이용하는 연구(Honovich et al., 2022)가 진행되었다.

AggreFACT 데이터셋

AggreFACT(Tang et al., 2022)는 기존 추상요약 모델에서의 Factual consistency 평가에 관한 연구결과를 통합하여 Factual Inconsistency가 발생하는 오류 유형과 분석에 사용하는 요약모델 종류를 재분류하여 Factual consistency 평가를 위한 통일된 생성요약 벤치마크를 제안했다. AggreFACT (Tang et al., 2022)는 관련된 9개 선행연구의 서로 다른 오류 유형을 발생 원인과 문법적 특성에 따라, 공통된 4가지로 오류 라벨로 재분류하였다.

AggreFACT(Tang et al., 2022)에서는 이 중 4개의 선행연구 결과를 활용해 오류 유형 표지를 재부착했고, 요약모델의 제안 시기에 따라 3 종류로 구분하였다. AggreFACT(Tang et al., 2022)에서는 요약문에서 Hallucination이 발생한 원인과 문법 요소에 따라 오류유형을 재분류하였다. 오류발생의 원인을 본문 내의 요소 중 확인할 수 있는 경우를 'Intrinsic' 오류라 하고, 그 외에는 'Extrinsic' 오류라 했다. 문법 요소에 따른 분류로는 요약문의 오류가 객체, 개수 등의 명사로부

터 기인한 경우를 'Noun-Phrase(NP)' 오류라 하고, 관계연결이나 부정 등의 오류로부터 기인한 경우를 'Predicate' 오류라고 하였다. 이외에도 선행연구의 결과 중, 문장 전체에 걸쳐서 오류 유형을 파악해야 하는 경우를 'Entire-Sentence'로 분류하였다. AggreFACT(Tang et al., 2022)에서 분류한 4가지 유형의 본문-요약문에서의 오류 유형 별 예시는 부록의 <표 C>와 같다.

AggreFACT(Tang et al., 2022)는 생성요약 모델의 제안시기에 따라 'OLD', 'XFormer', 'SOTA'의 세 가지로 나눴다. Transformer(Vaswani et al., 2017) 이전의 LSTM 등의 Seq2Seq 구조를 자연어 생성모델에 적용한 전통적인 생성요약 모델유형을 'OLD'라고 하였고, Transformer(Vaswani et al., 2017) 구조를 생성요약 모델에 활용한 경우를 'XFormer'라 하였다. 또한 Transformer(Vaswani et al., 2017) 구조에 대규모 데이터를 사전학습한 BERT(Devlin et al., 2018) 이후에 추가적인 사전 학습 방법 등을 적용해 자연어처리의 다양한 하위 분야에서 높은 성능을 보이는 경우를 'SOTA'라고 분류했다.

<표 1>은 AggreFACT(Tang et al., 2022)이 재분류한 4개의 선행연구에서 수동평가를 진행한 방법과 사용한 데이터와 요약모델을 분류한 것이다. 검증 데이터셋으로 사용한 본문 출처와 요약모델, 정답사용 여부와 각각의 데이터 수가 표시하였고, 참고데이터에서 수동평가 시에 숙련된 작업자가 참여한 경우는 굵은 글씨로 표기하였다.

〈표 1〉 AggreFACT에서 정리한 데이터 정보

		참고 데이터셋			
		CLIFF	FRANK	Goyal	XSumFaith
Document 출처	XSum	O	O	O	O
	CNNDM	O	O	O	X
요약모델 타입	OLD	-	PtGen,TConvS2S	-	PtGen,TConvS2S
	XFormers	-	BERTS2S,TransS2S	-	BERTS2S,TransS2S
	SOTA	BART,PEGASUS	-	DAE	-
정답요약 사용		X	X	O	X
개수 정보	#데이터	600	1246	125	1853
	#작업자	2	3	2	3

CLIFF(S. Cao & Wang, 2021)는 요약문에서 자주 발생하는 오류유형에 따라 인위적인 오류를 더한 문장을 negative sample로 사용해 finetuning 하였다. CLIFF(S. Cao & Wang, 2021)에서는 숙련된 2명의 작업자가 수동평가를 하여, 평가지표로 사용한 FactCC(Kryscinski et al., 2020)와 BertScore(Zhang et al., 2020)에 대한 상관분석을 통해 요약모델에서의 Factual Consistency 개선정도를 Seq2Seq 구조의 사전학습 모델인 BART(Lewis et al., 2020), PEGASUS(Zhang et al., 2020)와 비교하였다.

FRANK(Pagnoni et al., 2021)는 본문의 주제정보를 활용하는 TConvS2S,와 Transformer(Vaswani

et al., 2017)의 레이어를 랜덤 값과 BERT(Devlin et al., 2018)의 레이어 값으로 초기화한 Trans-S2S, BERTS2S(Rothe et al., 2020), Pointer-Generator 네트워크를 활용한 PtGen(See et al., 2017)와 에 대해 요약모델의 요약문에 대해 발생 원인과 문법 요인 등 다양하게 분류하여 임의의 작업자 3명의 수동평가 결과와 상관분석을 하였다.

CLIFF(S. Cao & Wang, 2021)는 요약문에서 자주 발생하는 오류유형에 따라 인위적인 오류를 더한 문장을 negative sample로 사용해 finetuning 하였다. CLIFF(S. Cao & Wang, 2021)에서는 숙련된 2명의 작업자가 수동평가를 하여, 평가지표로 사용한 FactCC(Kryscinski et al., 2020)와 BERTScore(Zhang et al., 2020)에 대한 상관분석을 통해 요약모델에서의 Factual consistency 개선정도를 Seq2Seq 구조의 사전학습 모델 인 BART(Lewis et al., 2020), PEGASUS(Zhang et al., 2020)와 비교하였다.

FRANK(Pagnoni et al., 2021)는 본문의 주제정보를 활용하는 TConvS2S(Rothe et al., 2020), Pointer-Generator 네트워크를 활용한 PtGen(See et al., 2017)과 Transformer(Vaswani et al., 2017)의 레이어를 랜덤 값과 BERT(Devlin et al., 2018)의 레이어 값으로 초기화한 TransS2S, BERTS2S(Rothe et al., 2020) 모델의 요약문에 대해 발생 원인과 문법 요인 등 다양하게 분류하여 임의의 작업자 3명의 수동평가 결과와 상관분석을 하였다.

XSumFaith(Maynez et al., 2020)는 Hallucination 발생 여부에 집중하여, 숙련된 작업자 3명이 본문-요약문 쌍에서 span 위치마다 Hallucination 발생 여부를 표시하고, 수동평가 위치와의 IoU (Intersection over Union)을 통해 상관분석을 진행하였다.

본 논문에서는 AggreFACT(Tang et al., 2022)에서 분류한 요약문의 오류 유형에 따라 Factual consistency 평가모델 점수와 요약문에서의 오류 여부 간의 상관관계에 대해 알아본다. 평가지표의 공정한 비교를 위하여, 본 연구에서는 본문-요약문 쌍에 대한 Factual consistency 평가지표 중, 검증 요약 데이터로 활용할 XSum(Narayan et al., 2018)과 CNNDM(Hermann et al., 2015) 데이터에 평가모델을 finetuning 하지 않는 방법만 사용한다.

제안 방법

본 연구에서는 FFCI(Koto et al., 2020)에서 제시한 생성요약 모델의 평가기준을 차용하여 정답 요약을 참조하지 않는 형태로 수정한 요약문의 자동방식을 제안한다. 제안한 방식에서는 좋은 요약문에 해당하는 특성을 정의하고, 본문-요약문 간의 각 특성 별 평가점수를 측정한다. 제안한 평가기준을 AggreFACT(Tang et al., 2022)의 오류유형 분류에 따라 적용하고, 제안한 평가방식의 특성 별 점수와 수동평가 간의 상관관계를 분석한다.

본 연구에서는 요약모델의 자동평가를 위한 특성 중, faithfulness, focus와 coverage에 대한 본문

과 요약문 간의 비지도평가 점수를 측정한다. FFCI(Koto et al., 2020)는 요약문의 평가점수 측정을 위해 문서 요약 데이터셋의 정답요약을 사용했다. 하지만 본문-요약문 간의 성능측정을 위한 평가지표로 활용한 모델을 특정 데이터에 finetuning하면 2.5에서 언급한 정답요약으로 활용한 문장의 희소성과 편향성 등의 문제가 발생하기 쉽다. 따라서 본 연구에서는 평가지표 선정 과정에서 생성요약 검증 데이터(Narayan et al., 2018; Hermann et al., 2015)에 finetuning이 필요한 경우는 제외하였다.

본 실험에서는 요약문 평가를 위해 각각의 특성마다 대치한 평가점수를 평균 또는 가중합하여, AggreFACT(Tang et al., 2022)의 수동평가와 비교한다. 각 특성마다 선정한 본문-요약문 쌍의 평가지표는 다음과 같다.

요약문의 평가특성: Faithfulness

요약문 자동평가를 위한 첫 번째 특성인 faithfulness는 요약문과 본문 간의 Factual consistency의 정도를 측정하기 위한 특성으로, 요약문의 정보 중 본문에 없거나 상충되는 내용이 얼마나 있는지를 확인한다. FFCI(Koto et al., 2020)는 요약모델의 faithfulness 평가점수로 정답요약을 생성요약 결과와 함께 사용하는 단일 평가지표 (Lin, 2004; Papineni et al., 2002; Lavie & Agarwal, 2007)와 요약문 만을 사용하는 단일 평가지표(Zhang et al., 2020; Kryscinski et al., 2020; Maynez et al., 2020)를 각각 사용해서 실험을 진행했다.

본 연구에서는 faithfulness 특성의 평가점수를 본문과 요약문 간의 의미적 유사도로 정의한다. 본문-요약문 쌍의 문장 단위 (sentence-level) 의미적 유사도를 측정하기 위해 문장 쌍을 분류하는 NLI(Natural Language Inference) 데이터 분류를 위해 학습된 모델을 사용한다. 본문-요약문 쌍의 단어 수준(word-level) 의미적 유사도를 측정하기 위해 PLM의 토큰 출현확률을 평가 점수로 활용하는 Cloze-Task 방식의 평가모델을 사용한다. Cloze-Task 방식의 평가모델은 사전학습 단계에서 일부 토큰을 마스킹(masking)하고 채우는 방식(Masked Language Modeling, MLM)으로 사전학습을 진행한 PLM(Devlin et al., 2018; Lewis et al., 2020)을 활용해 요약문 중 일부를 마스킹하여 예측값과 정답을 비교한다.

$$S_{faithfulness} = \lambda S_{NLI} + (1-\lambda) S_{MLM} \quad (0<\lambda<1) \quad (1)$$

본 실험에서 사용한 faithfulness의 평가점수는 식 (1)과 같다. 본문-요약문 쌍에 대한 문장과 단어 수준의 faithfulness 평가모델로 각각 NLI와 Cloze-Task 기반의 평가지표를 사용했을 때, 각각의 점수를 S_{NLI}와 S_{MLM}라고 한다. 최종적으로 두 점수를 가중합하여 faithfulness 특성의 평가점수 $S_{faithfulness}$를 계산한다. S_{NLI}와 S_{MLM}를 위해 선정한 각각의 평가모델은 다음과 같다.

Faithfulness를 위한 의미 유사성 기반 평가지표

본문-요약문 간 문장 수준 의미적 유사성 기반 평가지표로는 BERTScore (Zhang et al., 2020)를 사용한다. FFCI(Koto et al., 2020)는 Factual Consistency를 측정하기 위한 NLI 기반 평가지표로 FactCC(Kryscinski et al., 2020) 등도 사용했지만, FactCC(Kryscinski et al., 2020)는 downstream task로 사용할 요약 데이터셋에 대한 별도의 finetuning 과정이 필요하다. 따라서, 본 연구에서는 별도의 finetuning 과정 없이 PLM(Devlin et al., 2018)를 그대로 사용하는 방식인 BERTScore(Zhang et al., 2020)를 사용한다.

BERTScore(Zhang et al., 2020)에서는 사전학습 모델인 BERT(Devlin et al., 2018)의 출력인 contextual embedding을 사용한다. BERTScore(Zhang et al., 2020)는 정답요약과 요약문을 입력으로 받아 두 문장에서의 각 토큰 쌍에 대해 pairwise 코사인 유사도를 계산한다.

본 연구에서는 정답 요약 문장이 주어지지 않는 상황을 가정하여, BERTScore의 입력으로 정답요약 대신 본문을 사용해 본문-요약문 간의 의미 유사도 점수 S_{NLI}를 계산한다.

Faithfulness를 위한 Cloze-Task 기반 평가지표

본문-요약문 간 단어 수준의 Cloze-Task 기반 평가지표로는 BLANC(Vasilyev et al., 2020)를 사용한다. BLANC(Vasilyev et al., 2020)는 본문-요약문 쌍의 자동평가를 위해 PLM의 본문을 활용한 마스킹 된 요약문 복원 정확도를 요약문의 본문에 대한 faithfulness 평가지표로 제안했다. 요약문의 객체 중 랜덤하게 마스킹 된 토큰을 복원 하는 PLM의 사전학습 과제인 MLM을 수행하였고, 평가모델의 입력방식에 따라 본문과 요약문을 접합(concat)하는 $BLANC_{help}$와 본문과 요약문을 각각 입력받는 형태로 모델을 학습하는 $BLANC_{tune}$을 제안했다. 본 실험에서는 평가모델을 특정 데이터에 finetuning할 경우 생기는 평가지표의 편향성 문제(Sun et al., 2022)를 피하기 위해, $BLANC_{help}$를 식 (1)의 본문-요약문의 단어 간의 관련도 S_{MLM}로 활용한다.

요약문의 평가특성: Focus and Coverage

두 번째 평가요소인 focus와 coverage는 본문-요약문 쌍에 대해 서로 정보를 얼마나 잘 포함하고 있는지를 확인한다. FFCI(Koto et al., 2020)는 focus는 본문의 정보 중 요약문이 포함하는 정도로, coverage는 요약문의 정보 중 본문에 없는 내용이 포함된 정도로 정의했다.

FFCI(Koto et al., 2020)에서는 본문-요약문 간의 정보의 양을 측정하기 위해 질의생성-질의응답 기반 모델을 활용했다. FFCI(Koto et al., 2020)는 질의생성-질의응답 기반 Factual consistency 평가지표로 QAGS(Wang et al., 2020)를 사용했다. QAGS(Wang et al., 2020)는 요약문이 포함하는 본문의 정보를 측정하기 위해 요약문을 기반으로 질의 문장과 정답 스팬을 생성하고, 생성한 질의에 대해 본문을 기반으로 답변한 질의응답 모델의 스팬 예측 정확도를 요약문이 포함하는 본문

의 정보량으로 사용한다. FFCI(Koto et al., 2020)는 QAGS(Wang et al., 2020)에서 생성결과 요약문 기반의 질의 를 생성했을 때, 정답요약 기반으로 답변한 질의응답 모델의 정확도를 focus 특성의 평가점수로 활용했다. 반대로, 정답요약 기반으로 생성된 질의에 대한 생성결과 요약문 기반의 질의응답 정확도를 coverage 평가점수로 활용했다.

본 연구에서는 focus 특성을 본문의 정보 중 요약문 기반의 응답할 수 있는 정도로 정의한다. coverage 특성은 요약문의 정보 중 본문 기반의 질의응답을 할 수 있는 정도로 정의한다. FFCI (Koto et al., 2020)에서는 각각의 평가과정에서 정답요약을 사용한 것과는 달리, 본 실험에서는 정답요약을 활용하지 않고 본문과 요약문을 질의생성-질의응답 기반의 Factual consistency 평가모델 입력으로 사용한다.

각 특성에 대한 평가점수를 구하기 위해 사용한 본문-요약문 간 Factual consistency 평가지표는 다음과 같다.

Focus를 위한 의미 질의생성-질의응답 기반 평가지표

본문-요약문 간 Focus 특성의 평가지표로는 SummaQA(Scialom et al., 2019)를 사용한다. SummaQA(Scialom et al., 2019)는 강화학습 기반의 생성요약 모델 학습을 위해 질의응답 모델 기반의 요약문 평가지표를 제안했다. 질의생성-질의응답 모델은 SQuAD(Rajpurkar et al., 2016)로 학습한 질의생성 모델과 질의응답 모델을 활용한다. 질의생성 모델은 본문 기반으로 선택한 객체를 정답으로 하는 질의들을 생성하고, 질의응답 모델은 생성한 질의들에 대해 요약문을 context로 활용했을 때의 가능도(confidence)와 질문들에 대한 F1-score를 생성요약 모델에 보상으로 제공한다.

본 실험에서는 본문에 대한 focus 특성의 평가지표로 식 (2)와 같이 본문-요약문 쌍에 대한 SummaQA의 가능도 $SummaQA_{confidence}$와 F1-score $SummaQA_{f1}$를 가중합 하여 focus 특성의 평가점수 S_{focus}로 사용한다.

$$S_{focus} = \lambda SummaQA_{confidence} + (1-\lambda)SummaQA_{f1} \ (0<\lambda<1) \qquad (2)$$

Coverage를 위한 의미 질의생성-질의응답 기반 평가지표

본문-요약문 간 Coverage 특성의 평가지표로는 QuestEval(Scialom et al., 2021)을 사용한다. QuestEval(Scialom et al., 2021)은 좋은 요약 문장은 본문의 내용과 Factual Consistency을 유지하면서 중요한 내용을 담고 있어야 한다고 주장했다. 따라서 QuestEval(Scialom et al., 2021)은 요약문의 본문 내용에 대한 Factual Consistency을 평가하는 Factual Consistency(precision-based) 지표인 SummaQA(Scialom et al., 2019)와 본문의 요약문 내용에 대한 관련도(recall-based) 지표인 QAGS (Wang et al., 2020) 사이의 조화평균을 요약 문장에 대한 평가지표로 사용한다.

실험 및 결과

실험은 앞서 정의한 생성요약 결과의 평가를 위한 각 특성 별 평균 평가점수와 AggreFACT (Tang et al., 2022) 데이터의 수동평가 간의 상관분석을 진행하고, 요약문 분류 별로 상관관계를 관찰한다. 상관분석을 위한 요약문의 분류는 Hallucination이 발생한 이유와 요약문 생성에 활용한 본문의 출처 및 생성모델에 따라 나눴으며, 그 분류는 AggreFACT(Tang et al., 2022)에서 정의한 오류 유형과 요약문 생성에 활용한 생성요약 데이터 및 모델 분류를 따랐다.

실험에 사용한 AggreFACT 데이터 개수 및 평가지표 사용법은 다음과 같다.

데이터셋 상세

실험에는 오류 유형 별 평가지표를 적용하기 위해 2.6의 AggreFACT(Tang et al., 2022) 데이터를 활용한다. AggreFACT(Tang et al., 2022)은 뉴스요약 데이터셋인 XSum(Narayan et al., 2018)과 CNNDM(Hermann et al., 2015)에 대한 다양한 생성요약 모델의 결과에 대해 발생한 오류의 유형을 분류하였다. AggreFACT(Tang et al., 2022)은 오류의 유형을 <표 C>의 예시와 같이 발생원인과 문법적 특성에 따라 분류하였다. AggreFACT(Tang et al., 2022) 데이터를 활용하기 위해 전처리는 다음과 같이 진행 했다. AggreFACT(Tang et al., 2022)에서 언급된 에러 유형 외에 표시된 라벨은 오류없음을 의미하는 “correct” 라벨과 같음을 확인해 데이터에 따라 통합시키거나 실험에서 제외하였다. AggreFACT(Tang et al., 2022)의 FRANK(Pagnoni et al., 2021) 데이터 중 일부는 실험에 사용한 평가모델 중 SummaQA(Scialom et al., 2019)의 허용 입력크기를 초과하여 제외하였다. 전처리 과정에서 Goyal'21(GoyalDurrett, 2021)의 데이터는 전처리 결과 사용할 수 없게 돼서 실험에서 제외하였다.

〈표 2〉 AggreFACT의 오류 유형 별 XSum의 본문에 대한 요약문 개수

Data (#num)	correct	Total Error	Extrinsic-Error				Intrinsic-Error			
			Total	NP	Pred	EntSent	Total	NP	Pred	EntSent
CLIFF	126	**174**	**147**	125	28	7	48	33	10	6
FRANK	54	**938**	**843**	735	209	-	171	152	19	-
XSumFaith	202	**2298**	**1943**	1101	516	553	**685**	443	262	36
Total	256	**3536**	**2933**	1961	753	560	**904**	628	291	42

〈표 3〉 AggreFACT의 오류 유형 별 CNNDM의 본문에 대한 요약문 개수

Data (#num)	correct	Total Error	Extrinsic-Error				Intrinsic-Error			
			Total	NP	Pred	EntSent	Total	NP	Pred	EntSent
CLIFF	247	**52**	27	25	2	-	**27**	20	8	-
FRANK	698	**552**	**419**	313	188	-	**288**	230	77	-
Total	698	**851**	**446**	338	190	-	**315**	250	85	-

최종적으로 실험에서 사용한 데이터 중 XSum(Narayan et al., 2018)과 CNNDM(Hermann et al., 2015)의 본문을 활용한 AggreFACT(Tang et al., 2022) 데이터의 오류 유형 별 개수는 <표 2>, <표 3>과 같다. 본 실험에서는 AggreFACT(Tang et al., 2022)에서 발생 이유에 따라 묶은 Ext-Error와 In-Error, 그리고 전체 오류 개수를 Total-Error로 표시하여 집계하였다. 하나의 문장에서 여러 종류의 하위 오류가 동시에 발생할 수 있기에 중복을 제외하고 집계하였으며, 하위 유형에서의 중복을 제거한 경우는 볼드체로 표시하였다.

평가모델 상세

실험에는 BERTScore[5], BLANC[6], QuestEval[7], SummaQA[8]에서 공개한 버전의 평가모델을 사용한다. 요약문 평가를 위한 특성 faithfulness와 focus의 평균 평가점수를 구하기 위한 식 (1)과 식 (2)의 하이퍼파라미터 lambda는 모두 0.5로 설정하여 실험을 진행하였다. BERTScore는 BERT-large 모델의 최대길이인 512 토큰을 평가모델의 사전학습에 사용하였다. 따라서 본문의 길이가 이를 초과하는 경우는 본문의 앞에서부터 512 토큰에 맞춰서 사용한다. BLANC는 본문과 요약문장을 접합한 길이가 모델의 최대입력 길이인 512 토큰을 초과하지 않도록 한다. SummaQA(Scialom et al., 2019)과 QuestEval(Scialom et al., 2021)은 해당 논문의 실험과정을 따라, 본문의 길이를 400 토큰으로, 요약문과 생성되는 질의의 길이를 100토큰으로 제한했으며, 본문과 질의생성 단계에서 생성한 질의문장을 접합해 질의응답 모델의 입력으로 사용한다.

실험 결과

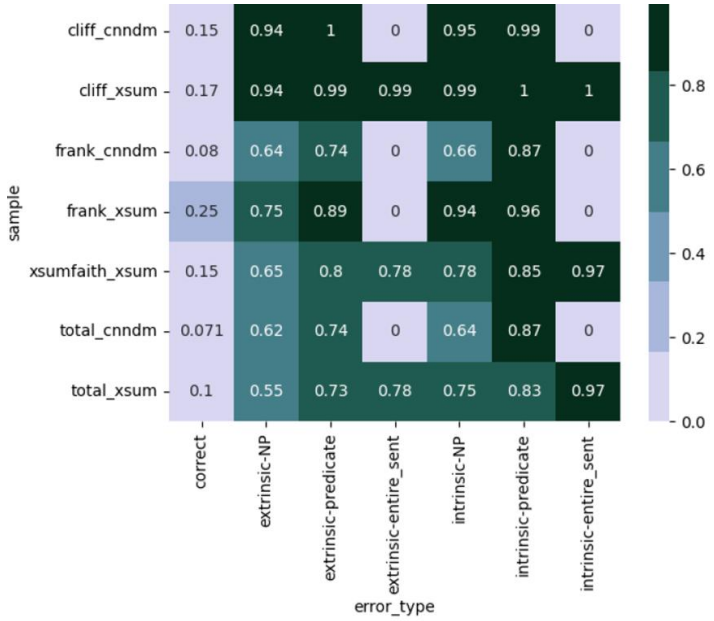
<표 2>와 <표 3>의 요약문 샘플을 생성요약 결과와 요약모델에 따라 나누고, 각각은 평가특성마다 요약문 분류에 따른 각각의 요약문 평가특성 별 평균점수 사이의 상관도를 계산했다.

5) BERTScore (github): https://github.com/Tiiiger/bert_score

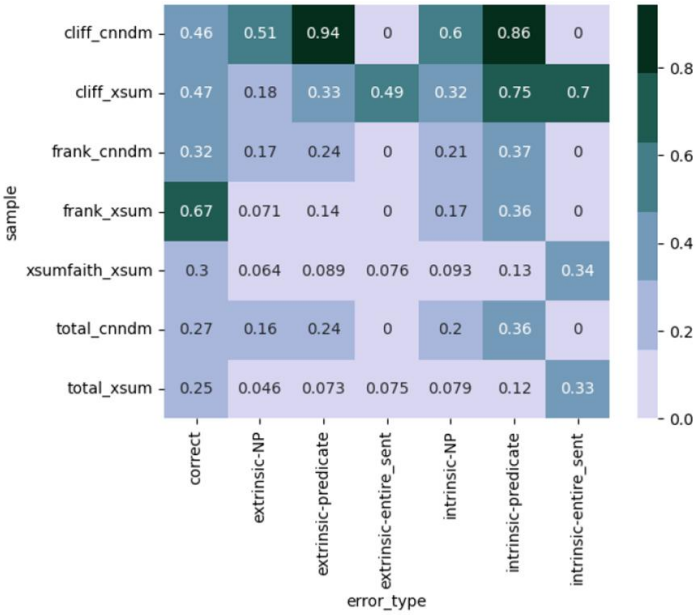
6) BLANC (github): https://github.com/PrimerAI/blanc

7) QuestEval (github): https://github.com/ThomasScialom/summa-qa

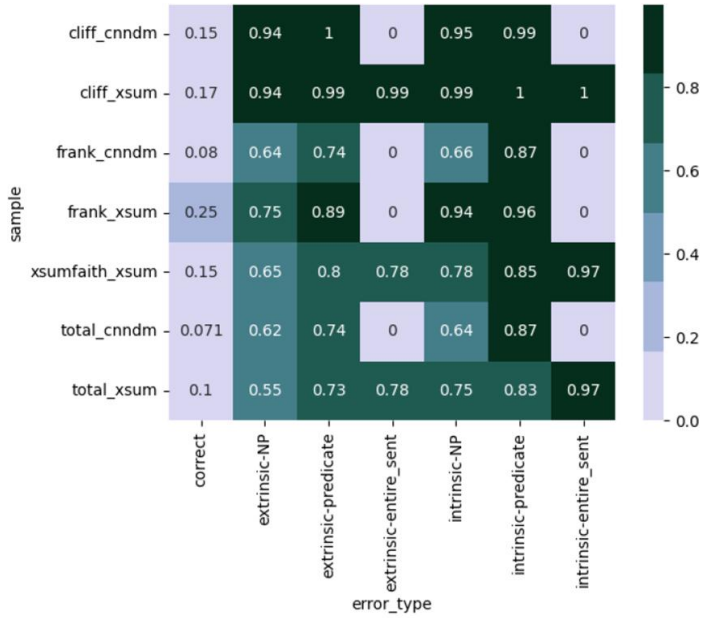
8) SummaQA (github): https://github.com/ThomasScialom/QuestEval



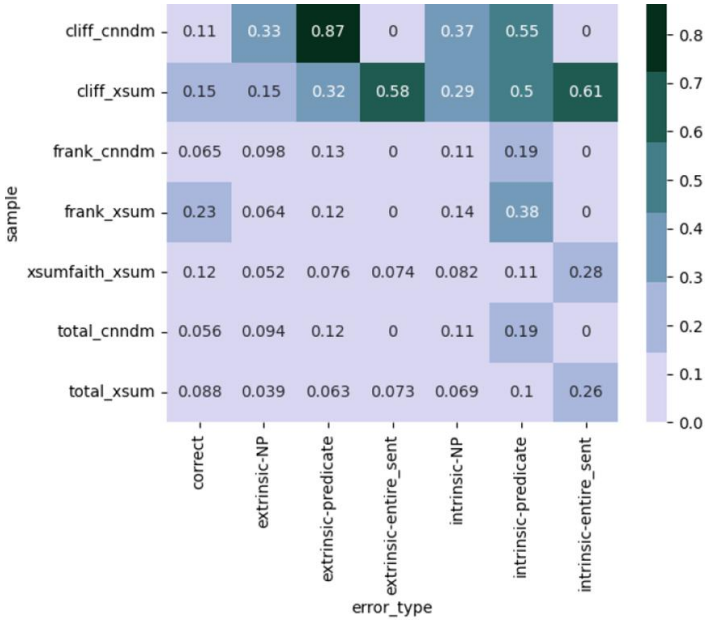
(그림 2) 평균 Faithfulness 평가점수의 Pearson 상관도



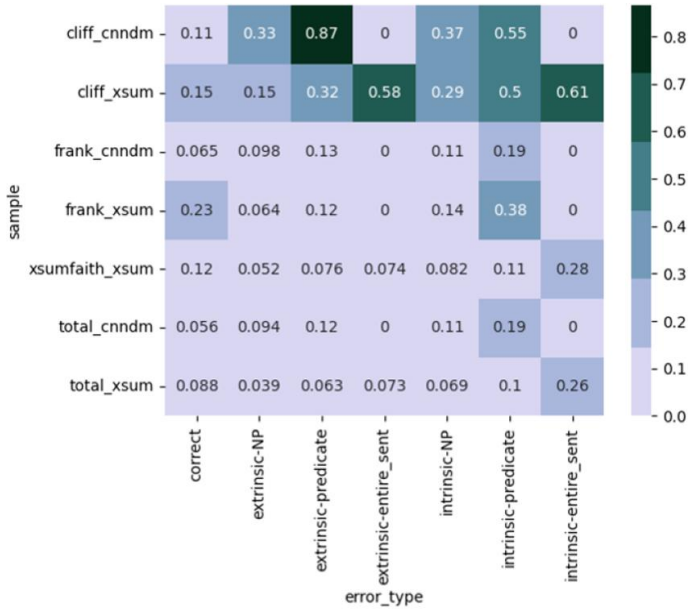
(그림 3) 평균 Focus 평가점수의 Pearson 상관도



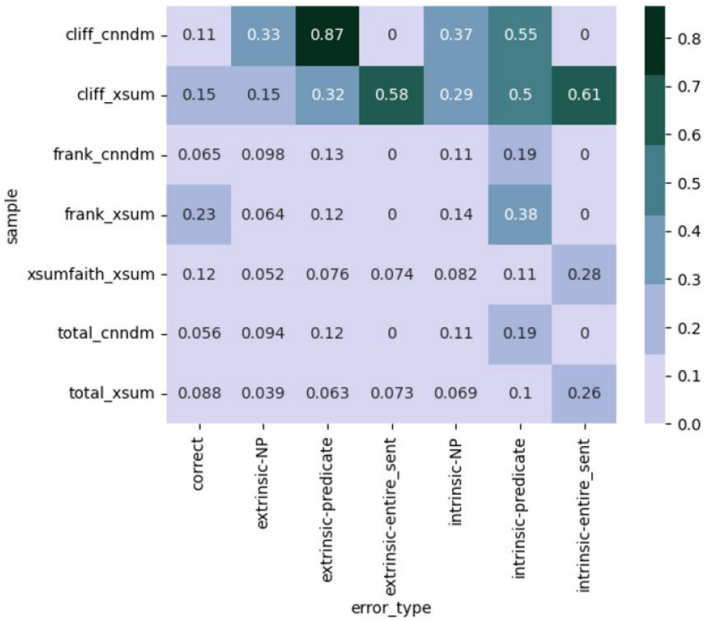
(그림 4) 평균 Coverage 평가점수의 Pearson 상관도



(그림 5) 평균 Faithfulness 평가점수의 Spearman 상관도



(그림 6) 평균 Focus 평가점수의 Spearman 상관도



(그림 7) 평균 Coverage 평가점수의 Spearman 상관도

(그림 2, 3, 4)와 (그림 5, 6, 7)은 각각 <표 2>와 <표 3>의 분류에 따라 Pearson과 Spearman 상관도를 시각화한 결과이다. (그림 2, 3, 4, 5, 6, 7)에서 공통적으로 술어 부분에 대한 오류가 발생한 'predicate'과 명사구와 술어 부분에서 모두 오류가 발생한 'entire-sentence' 오류 유형에서 높은 상관도를 보였다. 또한 오류발생의 이유에 따라서는 원인을 본문과 요약문 내부에서 찾을 수 없는 'extrinsic'에 비해 'intrinsic' 오류 유형에서 상관도가 높게 나타났다.

각 평가모델 별로 Pearson 상관계수를 표시한 (그림 2, 3, 4)에서는 본 연구의 3장에서 제안한 faithfulness와 coverage 평가점수에서 높은 상관도를 보였다. 본문과 요약문 간 의미적 유사도를 측정한 faithfulness 평가점수가 높은 상관도를 보임에 따라, 실험에서 사용한 본문-요약문 간에는 단어 수준(word-level)으로 유사한 의미를 가진다. 질의응답 모델을 활용해 문장 수준(sentence-level)의 의미적 유사도를 확인하기 위한 focus와 coverage 평가점수를 나타낸 (그림 3,4) 중 (그림 4)의 coverage 평가점수만 높은 상관도를 보임에 따라, 실험에서 사용한 요약모델은 본문의 내용을 잘 포함한 생성요약 결과를 만든다고 해석된다. 하지만 함축성(abstractiveness)이 높은 생성요약 과제의 특성 상, 본문의 내용 중 요약문에서 새로 등장한 요소를 확인하기 위한 focus 평가점수는 전반적으로 낮게 측정되었다. 생성요약에서 축약 및 함축이 빈번한 요약과제의 특성 상, 요약문이 지나치게 함축적인 요약문을 생성한 경우, focus 평가 점수가 높은 상관도를 보이기 힘들다고 생각된다.

각 평가모델 별로 Spearman 상관계수를 표시한 (그림 5, 6, 7)에서는 전반적으로 (그림 2, 3, 4) 대비 낮은 상관도를 보였다. 다만, 오류발생 원인을 본문과 요약문 쌍 내부 에서 찾을 수 있는 'intrinsic'유형 중 술어와 문장 전체에 관련된 'intrinsic predicate'과 'intrinsic entire-sentence' 오류 유형에 대해서는 비교적 높은 상관도를 보였다.

(그림 2, 3, 4, 5, 6, 7)에서 요약문을 생성한 요약모델 별로 살펴봤을 때, 다음의 경우에서 비교적 높은 상관관계를 보였다. CLIFF(S. Cao & Wang, 2021)를 사용한 경우, 전 반적으로 다른 모델 대비 높은 상관도를 보였다. (그림 2, 4)에서는 0.94~1로 매우 높은 상관도를 보였고, (그림 5,6,7)에서는 다른 요약결과는 대부분 0.2 미만의 상관도로 매우 떨어짐에 반해 비교적 높은 상관도를 유지했다. CLIFF에서 요약결과에 대해 반례를 생성해 요약모델의 대조학습에 사용한 방법이 타 방법론 대비 견고한 요약모델을 만드는 데 효과적이었다고 생각된다.

(그림 2, 4)의 FRANK(Pagnoni et al., 2021)의 경우도 높은 상관도를 보였다. FRANK 모델을 사용해 xsum에서 추출한 본문을 요약한 경우, 주어나 목적어에 해당하는 명사구에 관련된 'extrinsic Noun-Phrase' 유형을 제외한 모든 유형의 오류에 대해 높은 Pearson 상관도를 보였다.

(그림 2, 3, 4, 5, 6, 7)에서 본문-요약문 간 Hallucination이 발생하지 않은 'correct' 유형에 대해서는 공통적으로 매우 낮은 상관도를 보였다. <표 2>와 <표 3>에서 확인할 수 있듯이, 본 연구에서는 오류 유형 별 부착된 라벨의 부족으로 사용한 데이터 수에 불균형이 있었다. 실험에 사용한 두 가지 데이터 모두 'correct' 유형의 표본 수가 Hallucination이 발생한 다른 오류 유형

대비 적은 표본 수를 사용했다. 따라서 오류 유형 간 표본 수를 정규화하여 추가적인 확인이 필요하다.

<표 A>와 <표 B>는 각각의 평가특성 별 오류유형 마다의 샘플 데이터의 평균평가모델 점수로 부록에 표기하였다.

결론

본 연구에서는 FFCI(Koto et al., 2020)의 요약문 자동평가 방법론을 사람의 작업이 필요한 정답요약을 사용하지 않는 방법으로 적용될 수 있도록 수정하였다. AggreFACT(Tang et al., 2022)에서 분류한 요약문 분류와 오류 유형을 따라 방법론을 실험하였고, 실험결과를 정리하면 다음과 같다.

제안한 생성요약 결과의 평가특성에 따른 실험 결과, 요약문 생성에 사용한 요약모델과 요약문에서 Factual Inconsistency가 발생한 특정 경우에 제안한 방법이 높은 상관도를 보였다. 요약문 생성 시에, BART(Lewis et al., 2020)와 PEGASUS(Zhang et al., 2020) 등의 Transfomer(Vaswani et al., 2017) 구조로 대규모 사전학습을 진행한 'SOTA' 유형의 생성요약 모델을 사용한 경우 AggreFACT(Tang et al., 2022)의 모든 경우 높은 상관관계를 보였다. 본문-요약문 간에 전체 문장에 걸쳐 Factual Inconsisteny 오류가 발생한 'entire-sentence' 유형의 경우, 요약문을 생성한 본문의 출처와 모델에 상관없이 높은 상관계수를 보였다. 반면에 본문-요약문 간 내용의 불일치가 없는 경우는 수동평가 결과와 낮은 상관도를 보여, 제안한 방법의 평가점수가 높은 경우 Factual Consisteny 오류가 없다는 결론을 내릴 수는 없다고 보였다.

본 연구에서는 본문-요약문 간의 평가지표를 활용하여, 요약문 평가모델의 입력을 최대 512 토큰까지 제한하여 실험했다. 하지만 본문의 길이가 긴 요약 과업의 특성 상, 더 긴 본문을 활용할 경우에 대해 적용 가능성을 확인할 필요가 있다. 향후 연구 에서는 더 긴 길이의 입력을 활용할 수 있도록 제안된 Transformer 계열의 PLM(Beltagy et al., 2020; Zaheer et al., 2020)을 활용한 연구가 필요하다. 또한 모델 기반의 평가지표는 어휘 중첩도 기반의 평가지표(Lin, 2004; Lavie & Agarwal, 2007; Papineni et al., 2002) 대비 의미론적 요소를 더 잘 포착한다는 장점이 있지만, 요약문을 평가하는 과정이 매우 무겁다는 단점이 있다. 따라서 지식 증류 등을 이용해 더 가벼운 평가모델(Sanh et al., 2019) 등을 활용한 평가지표의 경량화 또 한 필요하다.

참고문헌

Amplayo, R. K., Liu, P. J., Zhao, Y., & Narayan, S. (2022). Smart: Sentences as basic units for text evaluation. arXiv preprint arXiv:2208.01030.

Arumae, K., & Liu, F. (2019, June). Guiding extractive summarization with question-answering rewards. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 2566-2577). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1264 doi: 10.18653/v1/N19-1264

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv: 2004.05150.

Berry, M. W., Dumais, S. T., & O' Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM Review, 37(4), 573-595. Retrieved from https://doi.org/10.1137/1037127 doi: 10.1137/1037127

Cachola, I., Lo, K., Cohan, A., & Weld, D. (2020, November). TLDR: Extreme summarization of scientific documents. In Findings of the association for computational linguistics: Emnlp 2020 (pp. 4766-4777). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.428 doi: 10.18653/v1/2020 .findings-emnlp.428

Cao, S., & Wang, L. (2021). Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2109.09209.

Cao, Z., Wei, F., Li, W., & Li, S. (2018). Faithful to the original: Fact-aware neural abstractive summarization. In Proceedings of the thirty-second aaai conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth aaai symposium on educational advances in artificial intelligence. AAAI Press.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018, June). A discourse-aware attention model for abstractive summarization of long documents. 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. 615-621). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2097 doi: 10.18653/v1/N18-2097

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Durmus, E., He, H., & Diab, M. (2020, July). FEQA: A question answering evaluation framework for

faithfulness assessment in abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5055-5070). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/ 2020.acl-main.454
doi: 10.18653/v1/2020.acl-main.454

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.

Fabbri, A., Wu, C.-S., Liu, W., & Xiong, C. (2022, July). QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 2587-2601). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main .187 doi: 10.18653/v1/2022.naacl-main.187

Falke, T., Ribeiro, L. F. R., Utama, P. A., Dagan, I., & Gurevych, I. (2019, July). Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 2214-2220). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1213 doi: 10.18653/v1/P19-1213

Gabriel, S., Celikyilmaz, A., Jha, R., Choi, Y., & Gao, J. (2021, August). GO FIGURE: A meta evaluation of factuality in summarization. In Findings of the association for computational linguistics: Acl-ijcnlp 2021 (pp. 478-487). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.42 doi: 10.18653/v1/ 2021.findings-acl.42

Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019, November). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Proceedings of the 2nd workshop on new frontiers in summarization (pp. 70-79). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5409 doi:10.18653/v1/D19-5409

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT Press.

Goyal, T., & Durrett, G. (2021). Annotating and modeling fine-grained factuality in summarization. In Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies.

Gudivada, V. N. (2018). Chapter 12 - natural language core tasks and applications. In V. N. Gudivada & C. Rao (Eds.), Computational analysis and understanding of natural languages: Principles, methods and applications (Vol. 38, p. 403-428). Elsevier. Retrieved from https://www .sciencedirect.com/science/article/pii/S0169716118300257
doi: https://doi.org/10.1016/bs.host.2018.07.010

Gupta, P., Wu, C.-S., Liu, W., & Xiong, C. (2021). Dialfact: A benchmark for fact-checking in dialogue. arXiv preprint arXiv:2110.08222.

Hardy, Narayan, S., & Vlachos, A. (2019). Highres: Highlight-based referenceless evaluation of summarization.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in neural information processing systems, 28.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., ⋯ Matias, Y. (2022). True: Re-evaluating factual consistency evaluation. In Workshop on document-grounded dialogue and conversational question answering.

Huang, L., Cao, S., Parulian, N., Ji, H., & Wang, L. (2021, June). Efficient attentions for long document summarization. In Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 1419-1436). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.112 doi: 10.18653/v1/ 2021.naacl-main.112

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ⋯ Fung, P. (2023, mar). Survey of hallucination in natural language generation. ACM Comput. Surv., 55(12). Retrieved from https://doi.org/10.1145/3571730 doi: 10.1145/3571730

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2014). Extractive summarization using continuous vector space models. In Proceedings of the 2nd workshop on continuous vector space models and their compositionality (cvsc) (pp. 31-39).

Kim, B., Kim, H., & Kim, G. (2019, June). Abstractive summarization of Reddit posts with multi-level memory networks. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 2519-2531). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1260 doi: 10.18653/v1/N19-1260

Kim, D.-H., Lee, S.-W., & Lee, G. G.-B. (2002). Query-based document summarization using important sentence selection heuristics and mmr. In Annual conference on human and language technology (pp. 285-291).

Koto, F., Lau, J. H., & Baldwin, T. (2020). Ffci: A framework for interpretable automatic evaluation of summarization. J. Artif. Intell. Res., 73.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020, November). Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 conference on empirical

methods in natural language processing (emnlp) (pp. 9332-9346). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/ 2020.emnlp-main.750 doi: 10.18653/v1/2020.emnlp-main.750

Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Revisiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics, 10, 163-177. Retrieved from https://aclanthology.org/2022.tacl-1.10 doi: 10.1162/tacl_a_00453

Ladhak, F., Durmus, E., Cardie, C., & McKeown, K. (2020, November). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In Findings of the association for computational linguistics: Emnlp 2020 (pp. 4034-4048). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.360 doi: 10.18653/v1/2020.findings-emnlp.360

Lavie, A., & Agarwal, A. (2007, June). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the second workshop on statistical machine translation (pp. 228-231). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W07-0734

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2020, July). 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). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/ 2020.acl-main.703 doi: 10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https:// aclanthology.org/W04-1013

Liu, Y. (2019). Fine-tune bert for extractive summarization. arXiv preprint arXiv:1903.10318.

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. arXiv preprint arXiv: 1908.08345.

Liu, Y., & Liu, P. (2021). Simcls: A simple framework for contrastive learning of abstractive summarization. arXiv preprint arXiv:2106.01890.

Liu, Y., Liu, P., Radev, D., & Neubig, G. (2022). Brio: Bringing order to abstractive summarization. arXiv preprint arXiv:2203.16804.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th annual meeting of the association for

computational linguistics (pp. 1906-1919). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.173 doi: 10.18653/v1/2020.acl-main.173

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv: 1808.08745.

Pagnoni, A., Balachandran, V., & Tsvetkov, Y. (2021, June). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 4812-4829). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/ 2021.naacl-main.383 doi: 10.18653/v1/2021.naacl-main.383

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). 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). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P02-1040 doi: 10.3115/1073083.1073135

Pasunuru, R., & Bansal, M. (2018, June). Multi-reward reinforced summarization with saliency and entailment. 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. 646-653). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2102 doi: 10.18653/v1/N18-2102

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. ArXiv, abs/1705.04304.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified textto- text transformer. Journal of Machine Learning Research, 21(140), 1-67. Retrieved from http://jmlr.org/papers/v21/20-074.html

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 2383-2392). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1264 doi: 10.18653/v1/D16-1264

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. Transactions of the Association for Computational Linguistics, 8, 264-280. Retrieved from

https:// aclanthology.org/2020.tacl-1.18 doi: 10.1162/tacl_a_00313

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/ 1910.01108. Retrieved from http://arxiv.org/abs/1910.01108

Schuster, T., Fisch, A., & Barzilay, R. (2021, June). Get your vitamin C! robust fact verification with contrastive evidence. In Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 624-643). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.52 doi: 10.18653/v1/ 2021.naacl-main.52

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., & Staiano, J. (2020, November). MLSUM: The multilingual summarization corpus. In Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp) (pp. 8051-8067). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020 .emnlp-main.647
doi: 10.18653/v1/2020.emnlp-main.647

Scialom, T., Dray, P.-A., Patrick, G., Sylvain, L., Benjamin, P., Jacopo, S., & Alex, W. (2021). Questeval: Summarization asks for fact-based evaluation. arXiv preprint arXiv:2103.12693.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. (2019, November). Answers unite! unsupervised metrics for reinforced summarization models. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp) (pp. 3246-3256). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1320
doi: 10.18653/v1/D19-1320

See, A., Liu, P. J., & Manning, C. D. (2017, July). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 1073-1083). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-1099
doi: 10.18653/v1/P17-1099

Sizov, G. (2010). Extraction-based automatic summarization: Theoretical and empirical investigation of summarization techniques..

Sun, T., He, J., Qiu, X., & Huang, X. (2022, December). BERTScore is unfair: On social bias in language model-based metrics for text generation. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 3726-3739). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.245

Tang, L., Goyal, T., Fabbri, A. R., Laban, P., Xu, J., Yahvuz, S., ⋯ Durrett, G. (2022). Understanding

factual errors in summarization: Errors, summarizers, datasets, error detectors. arXiv preprint arXiv:2205.12854.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers) (pp. 809-819). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology .org/N18-1074 doi: 10.18653/v1/N18-1074

Vasilyev, O., Dharnidharka, V., & Bohannon, J. (2020). Fill in the blanc: Human-free quality estimation of document summaries. arXiv preprint arXiv:2002.09836.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ⋯ Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, A., Cho, K., & Lewis, M. (2020, July). Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5008-5020). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.450 doi: 10.18653/ v1/2020.acl-main.450

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 353-355). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-5446 doi: 10.18653/v1/W18-5446

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. ACM Computing Surveys, 54(11s), 1-38.

Yuan, W., Neubig, G., & Liu, P. (2021). Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), Advances in neural information processing systems (Vol. 34, pp. 27263-27277). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/file/ e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ⋯ others (2020). Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th international conference on machine learning. JMLR.org.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In International conference on learning representations. Retrieved from

https://openreview .net/forum?id=SkeHuCVFDr

Zhu, C. (2021). Chapter 8 - applications and future of machine reading comprehension. In C. Zhu (Ed.), Machine reading comprehension (p. 185-207). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780323901185000084 doi: https://doi.org/10.1016/B978-0-323-90118-5.00008-4

Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., & Jiang, M. (2021, June). Enhancing factual consistency of abstractive summarization. In Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 718-733). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.58 doi: 10.18653/v1/2021.naacl-main.58

김탁영, 김지나, 강형원, 김수빈, & 강필성 (2022). 한국어 문서요약 및 음성합성 통합 프레임워크 구축. **대한산업공학회지, 48**(1), 80-90.

박은환, 나승훈, 신동욱, 김선훈, & 강인호 (2021). Summary-to-document 를 이용한 텍스트 생성 요약. **한국정보과학회 학술발표논문집**, 308-310.

최경호, & 이창기 (2016). Copy mechanism과 input feeding을 이용한 end-to-end 한국어 문서요약. **한국어정보학회 학술대회**, 56-61.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop

1차 원고 접수: 2023. 04. 17
1차 심사 완료: 2023. 06. 07
2차 원고 접수: 2023. 07. 08
2차 심사 완료: 2023. 07. 22
최종 게재 확정: 2023. 07. 28

(*Abstract*)

Empirical Study for Automatic Evaluation of Abstractive Summarization by Error-Types

Seungsoo Lee Sangwoo Kang

School of Computing, Gachon University

Generative Text Summarization is one of the Natural Language Processing tasks. It generates a short abbreviated summary while preserving the content of the long text. ROUGE is a widely used lexical-overlap based metric for text summarization models in generative summarization benchmarks. Although it shows very high performance, the studies report that 30% of the generated summary and the text are still inconsistent. This paper proposes a methodology for evaluating the performance of the summary model without using the correct summary. AggreFACT is a human-annotated dataset that classifies the types of errors in neural text summarization models. Among all the test candidates, the two cases, generation summary, and when errors occurred throughout the summary showed the highest correlation results. We observed that the proposed evaluation score showed a high correlation with models finetuned with BART and PEGASUS, which is pretrained with a large-scale Transformer structure.

Key words : Natural Language Processing, Generative Text Summarization, Quality Estimation, Meta-Evaluation

부록

<표 A,B>는 AggreFACT의 오류 유형 별로 측정한 평균 평가점수이다. 오류가 없는 'correct'와 전체 오류를 합한 'Total Error', 그리고 각각의 오류 유형 별로 표기했고, 해당 오류 유형의 데이터가 없는 경우는 '-'로 표기했다.

〈표 A〉 AggreFACT$_{XSum}$에서의 오류 유형 별 본문-요약문 쌍의 평균 평가점수

Data (#num)	correct	Total Error	Extrinsic-Error				Intrinsic-Error			
			Total	NP	Pred	EntSent	Total	NP	Pred	EntSent
Table A.1 - Average faithfulness score : (BERTScore,BLANC)										
CLIFF	0.4621	0.4608	0.4626	0.4622	0.4691	0.4531	0.4558	0.4629	0.4539	0.4216
FRANK	0.4314	0.4217	0.4211	0.4209	0.4216	-	0.4248	0.4247	0.4257	-
XSumFaith	0.4480	0.4430	0.4426	0.4460	0.4422	0.4366	0.4444	0.4455	0.4439	0.4358
Total	0.4445	0.4389	0.4374	0.4376	0.4375	0.4369	0.4413	0.4414	0.4430	0.4338
Table A.2.1 - Average focus score : SummaQA(avg$_{prob}$, $fscore$)										
CLIFF	0.0810	0.0660	0.0653	0.0653	0.0549	0.0787	0.0659	0.0680	0.0622	0.0563
FRANK	0.0613	0.0612	0.0607	0.0595	0.0642	-	0.0637	0.0615	0.0810	-
XSumFaith	0.0670	0.0630	0.0620	0.0655	0.0620	0.0551	0.0662	0.0662	0.0653	0.0632
Total	0.0658	0.0633	0.0618	0.0632	0.0624	0.0554	0.0657	0.0652	0.0662	0.0622
Table A.2.2 - Average coverage score : QuestEval										
CLIFF	0.7755	0.8008	0.8112	0.8053	0.8012	0.8938	0.7734	0.7963	0.7605	0.6785
FRANK	0.7610	0.7299	0.7286	0.7330	0.7029	-	0.7388	0.7458	0.6831	-
XSumFaith	0.7163	0.7290	0.7269	0.7378	0.7342	0.7003	0.7408	0.7506	0.7323	0.6663
Total	0.7258	0.7345	0.7316	0.0000	0.7280	0.7027	0.7421	0.7519	0.7301	0.6680

〈표 B〉 AggreFACT$_{CNNDM}$에서의 오류 유형 별 본문-요약문 쌍의 평균 평가점수

Data (#num)	correct	Total Error	Extrinsic-Error				Intrinsic-Error			
			Total	NP	Pred	EntSent	Total	NP	Pred	EntSent
Table B.1 - Average M$_{faithfulness}$: $(BERTScore, BLANC)$										
CLIFF	0.4996	0.4855	0.4810	0.4865	0.4226	-	0.4896	0.4955	0.4781	-
FRANK	0.4929	0.4697	0.4666	0.4651	0.4656	-	0.4679	0.4661	0.4723	-
Total	0.4929	0.4793	0.4674	0.4664	0.4657	-	0.4699	0.4685	0.4729	-
Table B.2.1 - Average M$_{focus}$: $SummaQA(avg_{prob}, fscore)$										
CLIFF	0.1553	0.1447	0.1455	0.14979	0.1160	-	0.1430	0.1501	0.1154	-
FRANK	0.1494	0.1197	0.1143	0.1123	0.1150	-	0.1190	0.1174	0.1189	-
Total	0.1494	0.1315	0.1162	0.1149	0.1151	-	0.1211	0.1201	0.1195	-
Table B.2.2 - Average M$_{coverage}$: $QuestEval$										
CLIFF	0.7323	0.7250	0.7363	0.7384	0.7103	-	0.7060	0.6984	0.7165	-
FRANK	0.7103	0.6856	0.6811	0.6822	0.6701	-	0.6894	0.6864	0.6902	-
Total	0.7103	0.7015	0.6844	0.6863	0.6706	-	0.6908	0.6873	0.6927	-

<표 C>는 AggreFACT(Tang et al., 2022)에서 정의한 요약문에서 발생하는 오류 유형 4가지의 예시이다. 각각의 유형 별로 요약문에서 오류가 발생한 부분에는 붉은 글씨로 표기하였다.

〈표 C〉 AggreFACT의 오류 유형 별 예시

	예시
본문	The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.
에러 유형	요약문 예시
Intrinsic-NounPhrase	The first Ebola vaccine was approved in 2021.
Intrinsic-Predicate	The lengthy process of FDA approval for the Ebola vaccine has led doubt to a COVID-19 vaccine could be developed safely.
Extrinsic-NounPhrase	China has already started clinical trials of the COVID-19 vaccine.
Extrinsic-Predicate	The first Ebola vaccine is approved in the US, five years after the clinical trials have been started.