테크니컬 리포트

2020년 1차

인공지능

학습용

데이터 구축

자연어 영역

문서요약 텍스트

개요: 문서요약 텍스트 데이터셋이란?

텍스트 요약 기술은 자연어처리(NLP, Natural Language Processing) 분야의 주요 연구 분야 중 하나로서, 언어를 이해하고 지식을 추출하여 새로운 가치를 창출하는 과정에 있어 중요한 역할을 한다. 최근 인터넷 매체의 발달로 다양한 컨텐츠를 소비할 수 있는 환경에 처한 컨텐츠 소비자들을 대상으로 AI를 이용한 요약기술은 예컨대 법률문서 요약, 뉴스기사 요약 등을 통해 핵심 내용을 신속하고 정확하게 파악하는 과정에 도움을 줌으로써 다양한 수요를 파생시킨다.

AI가 해당 텍스트를 이해하고 핵심 내용을 요약적으로 전달하기 위해서는 AI가 해당 텍스트의 주요 내용이 무엇인지를 이해할 수 있는 형태로 가공된, 다양한 유형의 대규모 요약 텍스트 데이터셋이 필요하나, 이러한 데이터셋이 대규모로 구축되어 있는 영어와 달리 한국어로 된 요약 데이터셋은 현재까지 미진한 상황이었다.

이에 비플라이소프트㈜는 그 동안 부족했던 한국어로 된 문서요약 텍스트 데이터의 축적 및 활용을 위하여 총 40만 건의 한국어 데이터셋을 구축하였으며 해당 데이터셋은 뉴스기사, 기고문, 법률 판결문, 잡지 등 다양한 도메인으로 구성되어 요약 AI 알고리즘화 성능 향상을 추구하였다. 또한 데이터의 지속적인 이용과 AI 알고리즘의 응용에 제한이 없도록 저작권 문제를 완전히 해결하였다.

데이터셋의 구성

문서요약 텍스트 데이터셋은 원문데이터에 대한 추출요약과 생성요약으로 구성되며, 즉 하나의 데이터셋에 총 3건의 데이터 (원문, 추출, 생성)가 포함된다. 원문데이터의 수량은 뉴스기사 30만 건, 기고문 6만 건, 잡지 1만 건, 법원 판결문 3만 건으로 총 40만 건이며, 이에 대한 추출요약과 생성요약도 각각 40만 건 씩 구축되었다.

데이터 종류	데이터 형태	데이터 수량	제공 방식
신문기사	뉴스 텍스트	30만 세트	JSON 포맷 파일
기고문	오피니언 텍스트	6만 세트	
잡지	웹진 기사 텍스트	1만 세트	
법률	법원 판결문 텍스트	3만 세트	
합계		40만 세트	

각각의 데이터셋에는 출처와 카테고리, 데이터 크기, 데이터 생성일자 등의 메타정보가 포함되며, 그 형태에 따라 뉴스기사, 기고문, 잡지, 법률 데이터들은 각각의 속성에 맞게 수정된 메타정보를 포함하고 있다.

구체적인 데이터의 구성은 다음과 같다.

Key	Description	Type	Child Type
name	파일명	String	
delivery_date	생성시간 yyyy-MM-dd hh:mm:ss	String	
documents	문서 배열	JsonArray	JsonObject
[	문서	JsonObject	
id	문서 아이디	String	
category	카테고리	String	
media_type	미디어 유형 (ex: online)	String	
media_sub_type	비디어 유형 (ex: 중앙지)	String	
media_name	미디어 명 (ex: 국민일보)	String	
size	길이 (ex: small)	String	
char_count	본문길이	String	
publish_date	게시시간 yyyy-MM-dd hh:mm:ss	String	
title	제목	String	
text	본문(문단/문장)	JsonArray	JsonArray
[	문단	JsonArray	JsonObject
[	문장		
index	순번	String	
sentence	문장	String	
highlight_indices	불용어 위치 정보	String	
]			
]			
]			

데이터셋의 설계 기준과 분포

특정 채널에 편향되지 않은 요약 AI 알고리즘 개발을 위하여 채널별로 균형 있게 데이터셋을 설계하는 데에 가장 중점을 두었다.

- 뉴스기사 : 뉴스기사는 요약 AI 알고리즘의 핵심 데이터로서 10개 언론사로부터 30만 건의 원문데이터를 확보하였으며, 이 중 종합면 30%, 정치 20%, 경제 20%, 사회 20%, 문화 및 스포츠 기타 10%의 비율로 구성되어 있다.
- 기고문 : 기고문은 사실관계의 전달에 중점을 둔 일반적인 뉴스기사와 달리 개인적인 주장을 담고 있는 형태의 문서로서 신문의 오피니언 면을 통해 확보하였다. 특정 매체에 대한 집중도를 줄이고 정치/경제/사회/문화/과학 등 다양한 주제를 균등하게 배분하여 학습데이터를 구축하였다.
- 잡지 : 잡지는 전문성이 뚜렷하고 텍스트의 길이가 긴 편이므로 1만 건 수준으로 제한하였으며, 시사/경제, 공학/기술, 문화/라이프. 예술/엔터테인먼트, 요리/건강, 취미/레포츠, 컴퓨터/인터넷 등 7개의 카테고리로 구분하여 구성하였다.
- 법률 : 법률은 공공데이터 포털 (open.law.go.kr)을 통하여 판례 원문의 오픈 API를 제공 받아 원문데이터를 확보하였으며, 민사, 형사 등 다양한 사건 판례로 구성하였다.

구분		원문 건수	비율
뉴스기사	종합	9만 건	22.5%
	정치	3만 7,500건	9.3%
	경제	3만 7,500건	9.3%
	사회	3만 7,500건	9.3%
	문화	3만 7,500건	9.3%
	스포츠	3만 건	7.5%
	IT/과학	3만 건	7.5%
기고문	주장	6만 건	15%
잡지	시사	5,000건	1.2%
	문화예술	3,000건	0.7%
	기타	2,000건	0.5%
법률	판결문	3만 건	7.5%
합계		40만 건	100%

[표 1] 문서요약 텍스트 데이터의 주제별 분포

데이터 구조

문서요약 텍스트 데이터의 구조는 아래 표와 같다.

No		항목		길이	타입	필수여부	비고
		한글명	영문명				
1		데이터셋 정보			JsonObject	Y	
	1-1	데이터셋 명	name		String	Y	
	1-2	데이터셋 전달일자	delivery_date		String	Y	yyyy-mm-dd hh:mm:ss
2		문서 정보			JsonObject	Y	
	2-1	문서 번호	id		String	Y	
	2-2	카테고리	category		String	Y	예) 문서번호
	2-3	매체 유형	media_type		String	Y	예) 온라인

	2-4	매체 구분	media_sub_type		String	Y	예) 중앙지
	2-5	미디어 명	media_name		String	Y	예) 국민일보
	2-6	본문길이	size		String	Y	
	2-7	본문글자수	char_count		String	Y	
	2-8	발행일시	publish_date		String	Y	yyyy-mm-dd hh:mm:ss
	2-9	제목	title		String	Y	
3		본문(문단/문장) 정보	text		Array	Y	
	3-1	문단	[		Array	Y	
	3-2	문장	{		JsonObject	Y	
	3-3	순번	index		Integer	Y	
	3-4	문장	sentence		String	Y	
	3-5	불용어 위치 정보	highlight_indices		String	Y	
			}				
			]				
4		원문 평가 정보			JsonObject	Y	
	4-1	가독성	readable		Integer	Y	
	4-2	정확성	accurate		Integer	Y	
	4-3	정보성	informative		Integer	Y	
	4-4	신뢰성	trustworthy		Integer	Y	
5		추출요약문 정보	extractive		Array	Y	
6		생성요약문 정보	abstractive		Array	Y	

데이터 예시

아래 데이터는 요약 AI를 학습시키기 위한 문서요약 텍스트 데이터의 실제 예시로서, 광주매일신문의 2019년 5월 2일자 정치면 온라인 기사이다. 총 글자수는 814자로 '소' 데이터로 분류되었으며 기사의 품질은 각 5점 만점에 가독성 4점, 정확성 4점, 정보성 3점, 신뢰성 4점으로 평가되었다. 기사 내의 모든 문장은 0번부터 순차적으로 index가 붙으며, 해당 index는 추출요약에 사용된다. 즉 아래 데이터에서 전체 기사 내용을 요약하는 3개 문장은 2, 3, 4번 문장으로 추출되었으며, 이렇게 추출된 3개 문장을 이용하여 "광주시에 따르면 국방부·전남도와 협의해 후보지를 직접 찾아가 주민들에게 사업의 필요성을 설명할 사업 소개와 함께 후보지로

선정되면 지원되는 다양한 사업이 상세하게 담긴 군 공항 이전사업 관련 홍보 영상물을 제작 중이다"라는 1개의 생성요약문이 최종적으로 도출되었다.


```
{
"name": "문서요약 프로젝트",
"delivery_date": "2020-09-08 11:36:41",
"documents": [
{
"id": "343753195",
"category": "정치",
"media_type": "online",
"media_sub_type": "지역지",
"media_name": "광주매일신문",
"size": "small",
"char_count": "814",
"publish_date": "2019-05-02 18:32:00",
"title": "군 공항 이전사업 홍보 영상물 만든다",
"text": [
[
{
"index": 0,
"sentence": "市, 국방부·전남도 협의 통해 이전 후보지서 필요성 설명",
"highlight_indices": ""
},
],
...
],
"document_quality_scores": {
"readable": 4,
"accurate": 4,
"informative": 3,
"trustworthy": 4
},
"extractive": [
2,
```

3,

4

],

"abstractive": [

" 광주시에 따르면 국방부·전남도와 협의해 후보지를 직접 찾아가 주민들에게 사업의 필요성을 설명할 사업 소개와 함께 후보지로 선정되면 지원되는 다양한 사업이 상세하게 담긴 군 공항 이전사업 관련 홍보 영상물을 제작 중이다."

]

},

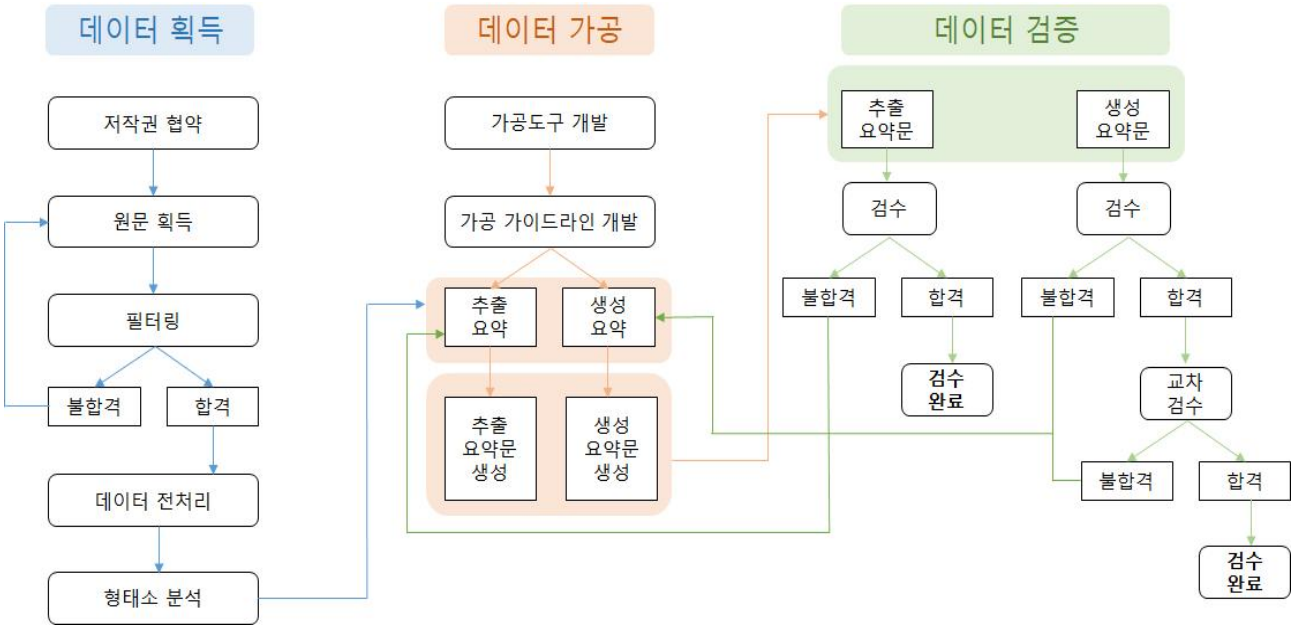
.....

]

}

데이터 구축 과정

문서요약 텍스트 데이터의 전체적인 구축 프로세스는 아래 그림과 같다.



[그림] 문서요약 텍스트 데이터의 구축 과정

- 데이터 획득 : 언론사 및 공공데이터 포털을 통해 확보된 원문데이터에는 메타정보를 부여하여 문장구분 오류 등의 1차 필터링을 거치고, 검수를 통해 합격된 데이터만 형태소 분석을 통해 요약작업용 텍스트 데이터로서 서버에 구축된다.

- 데이터 가공 : 추출요약은 원문의 형태적 구조를 바탕으로 내용을 잘 표현하는 문장 3개 (리드문 포함)를 추출하여 우선순위에 따라 차례로 작성하며, 생성요약은 추출요약으로 형성된 문장, 제목의 키워드, 동의어 등을 활용하여 기사의 내용을 한 문장으로 요약한다.
- 데이터 검증 : 가공시 사용한 가이드라인 준수 여부를 중심으로 검수를 진행하되, 추출요약은 중요 문장에 대한 추출여부를 5점 척도를 활용하여 4점 이상을 합격 기준으로 하고 생성요약은 추상적인 내용이 포함되어 있으므로 질적 요소를 포함하는 5점 척도를 활용하여 4점 이상을 합격 기준으로 한다. 검증에서 불합격 처리된 요약문은 재가공한다.

검수와 품질 확보

데이터에 대한 검수는 총 3차례에 걸쳐 진행되며, 1차 검수에서는 추출요약문과 생성요약문을 모두 검수하고, 2차 검수 (교차검수)에서는 1차 검수가 완료된 생성요약문을 검수한다. 생성요약의 경우 추상성이 높기 때문에 교차검수를 통해 검수의 객관성을 확보하고자 하였다. 각각의 요약문에 대한 검수 기준은 아래의 표와 같다.

검수 기준	추출요약	생성요약
키워드의 활용	제목에서 제시된 주요 단어가 포함된 문장 선택 여부	제목에서 제시된 주요 단어의 활용 여부
리드문의 활용	가장 중요하게 선택된 문장의 구성 형태 확인	가장 중요하게 선택된 문장 활용 여부
6하원칙 내용의 포함	6하원칙 내용이 포괄된 문장의 선택 여부	생성된 문장에서 6하원칙 내용의 포함 정도
주관적 문장의 포함	주관적 문장 수정 불가	-
의미의 중복	선택된 3개 문장 중 중복된 내용의 수준 - 극단적 중복을 회피	-
구체적 문장의 활용	3개 문장 중 1문장은 구체적 문장이 포함되어야 함	구체적 문장의 수정 여부
문장의 추출	문장 변형 금지	문장 원형 추출 금지
추상화, 집단화, 동의어의 선택	-	문장 수정시 추상화, 집단화, 동의 선택의 정도를 질적으로 판단
비문, 미완성 문장	-	비문, 미완성 문장은 반드시 2점 이하로 입력
문장의 길이	-	전체 문장의 10% 내외로 요약 - 극단적 축약, 극단적인 복문은 요약 실패로 판정

검수를 통해 불합격된 데이터의 경우, 검수자가 가공자에게 검수의견을 제시하며 가공자는 이를 참고하여 재가공을 진행한다. 이에 대한 재검수 후에도 불합격 판정을 받은 경우 3회까지 재가공을 요청할 수 있으며, 3차 검수 결과도 불합격인 경우에는 해당 데이터를 별도 저장하여 다른 작업자를 통해 재가공을 요청한다.

데이터 구축 담당자

수행기관(주관) : 비플라이소프트㈜ (전화: 070-7091-8562), 이메일: rice127@bflysoft.com