국립국어원 2019-01-15

발 간 등 록 번 호
11-1371028-000774-01

말뭉치 분석 연구 및 시범 구축

사업 책임자
김 한 샘

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '말뭉치 분석 연구 및 시범 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 7월 ~ 2019년 12월

2019년 12월 4일

사업 책임자: 김한샘(연세대학교)

사업 수행 기관 연세대학교 산학협력단,
㈜슬로워크

사업 책임자 김한샘

사업 참여자 이공주, 김학수, 주민재

〈사업 수행자〉 연세대학교 산학협력단, ㈜슬로워크

사업 책임자	김한샘(연세대학교, 책임연구원)
사업 참여자	김학수(강원대학교, 공동연구원)
	이공주(충남대학교, 공동연구원)
	주민재(명지대학교, 공동연구원)
	강범일(연세대학교, 연구보조원)
	한지윤(연세대학교, 연구보조원)
	오태환(연세대학교, 연구보조원)
	강예지(연세대학교, 연구보조원)
	최현수(연세대학교, 연구보조원)
	박석원(연세대학교, 연구보조원)
	정석원(강원대학교, 연구보조원)
	장영진(강원대학교, 연구보조원)
	최기현(강원대학교, 연구보조원)
	이경호(충남대학교, 연구보조원)
	김동준(한국신문윤리위원회, 연구보조원)
	여규병(전적 동아일보, 연구보조원)
	이세강(우송대학교, 연구보조원)
	이용원(동국대학교, 연구보조원)
	이창호(전적 스포츠조선, 연구보조원)
	김민지(연세대학교, 보조원)
	김아영(연세대학교, 보조원)
	손영랑(연세대학교, 보조원)
	윤선영(연세대학교, 보조원)
	이종혁(연세대학교, 보조원)
	정지원(연세대학교, 보조원)
	정혜진(연세대학교, 보조원)
	조연수(연세대학교, 보조원)
	허나연(연세대학교, 보조원)
	김담린(강원대학교, 보조원)
	김보은(강원대학교, 보조원)
	박주일(슬로워크, 개발인력)
	김형우(슬로워크, 개발인력)
	임동근(슬로워크, 개발인력)
	양정화(슬로워크, 개발인력)

<국문 초록>

말뭉치 분석 연구 및 시범 구축

본 사업의 목적은 신문기사 요약 말뭉치를 구축하여 4차 산업 혁명의 핵심인 빅데이터 분야의 발전으로 대두된 자동 요약 기술 개발에 이바지하는 데에 있다. 자동 요약은 정보 소스와 채널의 다각화 및 대량화, 전송의 신속화, 정보 소비 패턴의 변화 등으로 인해 매우 중요한 서비스로 등장하였다. 방대한 양의 텍스트에서 사용자에게 필요한 정보를 정확하게 추출하여 제공하기 위해서는 요약 기술의 발전이 필수적이며, 이러한 기술 개발에 반드시 필요한 것이 바로 요약 말뭉치이다. 본 사업의 주요 과업과 연구 성과는 다음과 같다.

신문기사 말뭉치 대상 요약 말뭉치 구축 : 추출 요약을 위해 주석된 13,167개의 주제문과 추상 요약을 위해 도메인 전문가가 검수하여 완성한 13,167개의 요약문을 포함한 총 4,389건의 기사로 구성된 신문기사 요약 말뭉치를 구축하였다.

한국어의 특성을 살린 고품질 추상 요약문 작성을 위한 지침 개발 : 요약문 작성을 위하여 한국어의 특성과 실용성을 반영하여 작성자가 고품질의 요약문을 생성할 수 있는 지침을 개발하였다. 지침에는 구축 작업자와 도메인 전문가, 작문 전공 연구진의 의견을 고루 반영하여 요약 말뭉치를 구축하고자 하는 누구나에게 참고가 될 수 있도록 작성하였다.

요약 말뭉치 구축 및 평가 방법론 개발 : 추출 요약과 추상 요약 방식의 요약 시스템에 모두 적용할 수 있는 요약 말뭉치를 유기적이고 효율성 높은 방식으로 구축할 수 있는 구축 기준과 체계를 수립하였다. 또한 구축된 요약 말뭉치를 기존에 널리 이용되는 요약 말뭉치 평가 방식을 통해 검증하고 정성적으로 평가할 수 있는 평가 기준을 제시하였다.

요약 말뭉치 활용 방안 모색 : 요약 기술을 실제 개발하고 활용해야 하는 산업체의 자문을 받아 실제 기술 개발에 사용될 수 있도록 말뭉치를 설계하고 산업계의 실수요를 조사하였다. 이어 실제 자동 요약 모델에 구축된 요약 말뭉치를 적용하여 활용 가능성을 검토하였다.

주요어: 요약 말뭉치, 기사 요약 말뭉치, 신문기사 요약, 자동 요약, 신문 빅데이터

차 례

차 례

제 6장 결론

제 1 장

서 론

1. 사업의 목적 및 개요

본 사업의 목적은 신문기사 요약 말뭉치를 구축하여 4차 산업 혁명의 핵심인 빅데이터 분야의 발전으로 대두된 자동 요약 기술 개발에 이바지하는 데에 있다. 자동 요약은 정보 소스와 채널의 다각화 및 대량화, 전송의 신속화, 정보 소비 패턴의 변화 등으로 인해 매우 중요한 서비스로 등장하였다. 방대한 양의 텍스트에서 사용자에게 필요한 정보를 정확하게 추출하여 제공하기 위해서는 요약 기술의 발전이 필수적이며, 이러한 기술 개발에 반드시 필요한 것이 바로 요약 말뭉치이다. 이에 따라 본 사업에서는 4차 산업혁명 대비 기반 기술 개발 및 인공지능 기술 개발, 활용을 위하여 대규모 말뭉치 구축이 가장 시급하게 필요한 분야로 요약 분야를 선정하였다. 요약 기술 중 가장 보편적인 신문기사 텍스트를 기반으로 추출 요약 모델과 추상 요약 모델에 모두 적용할 수 있는 요약 말뭉치를 구축하고 실제 적용 방안을 검토하였다. 이러한 요약 말뭉치를 구축하기 위해 설정하였던 하위 목표는 아래와 같다.

▶ **신문기사 말뭉치 대상 요약 말뭉치 구축**

추출 요약을 위한 주제문 주석 말뭉치와 추상 요약을 위해 도메인 전문가가 작성한 요약문으로 구성된 말뭉치 구축

▶ **한국어의 특성을 살린 고품질 추상 요약문 작성을 위한 지침 개발**

요약문 작성을 위한 기존 사례를 토대로 한국어의 특성과 실용성을 반영하여 작성자가 고품질의 요약문을 생성할 수 있는 지침을 개발

▶ **산업계 실수요를 반영한 말뭉치 구축 계획 수립 및 활용**

요약 기술을 실제 개발하고 활용해야 하는 산업체의 자문을 받아 실제 기술 개발에 사용될 수 있도록 말뭉치를 설계하고 실제 활용 가능성을 검증하여 산업계의 실수요를 해소할 수 있는 자원 개발

▶ **요약 기술 시범 개발**

본 사업을 통해 구축되는 말뭉치를 활용하여 실제 요약 모델에 적용

2. 사업 범위

본 과제에서 수행한 과업의 범위는 아래와 같다.

과업의 범위	세부 과업 내용
요약 말뭉치 구축 방법론 수립	추출 및 추상 요약 말뭉치 구축 체계 수립 추출 및 추상 요약 말뭉치 구축 시스템 활용 추출 및 추상 요약 말뭉치 구축 기준 수립
요약 말뭉치 구축 지침	요약 말뭉치 구축 지침 개발 말뭉치 구축 및 검수 인력 실무 교육 운영
요약 말뭉치 시범 구축	‘2018 국어 말뭉치 분석 및 구축 연구’ 사업 구축 말뭉치 중 신문기사 말뭉치 7,265건의 기사를 대상으로 요약 대상 4,451건의 기사 추출 후 정제 과정을 거쳐 최종적으로 4,389건의 기사에 주제문에 주석하고 요약문을 생성
요약 말뭉치 활용 방안 모색	자동 요약 모델 적용 요약 말뭉치의 산업적 활용 가치 검증

표 1 과업 범위 및 세부 과업 내용

3. 사업의 필요성 및 기대 효과

본 사업의 필요성과 기대효과는 다음과 같다.

▶ 인공지능을 위한 한국어 요약 말뭉치 수요 발생

텍스트 자동 요약은 인공지능을 활용한 고차원의 자연어 처리 기술이다. 정보의 소스와 채널 다각화와 대량화, 전송의 신속화, 정보 소비 패턴의 변화 등으로 인해 요약이 매우 중요한 서비스로 등장하였다. 구글 어시스턴트나 알렉사와 같은 AI 비서에서 뉴스 브리핑은 매우 중요한 컨텐츠로 자리 잡았다. 미국의 오토메이티드 인사이츠(Automated Insights), 아골로(Agolo) 등의 스타트업은 다양한 소스의 정보를 다양한 형태의 요약 서비스로 제공하는 기술을 개발/제공 중이며, 마이크로소프트 등으로부터 기술적 가치를 인정받아 투자를 유치한 바 있다. 그러나 공개된 국내 자동 요약 시스템은 네이버 기사 요약, 다음 기사 요약, NCSOFT PAIGE 등이다. 공개된 시스템들은 모두 추출 요약 방법을 사용하고 있다. 현재 국내 자동 요약의 수준은 단순히 주제를 잘 나타내는 문장을 뽑는 추출 요약의 정확도도 확보하지 못하는 수준이다.

영어의 경우 CNN/daily 말뭉치 등 대규모의 자원이 공개되어 있으나, 한국어 경우 네이버와 다음 등 포털사이트에서 제공하는 요약봇 서비스의 기반이 되는 말뭉치는 공개하지 않고 있다. 이에 따라 산업계에서 대규모의 고품질 한국어 요약 말뭉치를 확보하는 데에 어려움을 겪고 있다. 또한 개발된 요약 기술의 성능을 객관적으로 평가할 수 있는 평가용 데이터셋도 부재한 실정이다.

▶ 사업 기대 효과

(1) **자동 요약 기술 성능의 객관적 검증 데이터 마련**: 향후에 개발되는 자동 요약 기술 성능을 검증하기 위해서는 기준으로 삼을 만한 신뢰할 수 있는 평가용 데이터가 필요하다. 본 과제에서는 한국어 요약 말뭉치 구축을 통해 요약 모델의 성능을 평가할 수 있는 객관적인 근거를 마련하였다.

(2) **추출 요약 기술 및 추상 요약 기술 개선 및 성능 향상**: 기계학습 기반의 자동 요약 기술을 개발하기 위해서는 모델의 훈련을 위한 대량의 고품질 데이터가 필수적이

다. 본 과제에서는 자동 요약 모델 개발을 위한 고품질의 훈련 데이터를 공급하여 기술을 개선하고 성능 향상에 기여하고자 한다.

(3) **데이터 부재로 기술 개발에 난항을 겪는 기업체 수요 충족**: 자연어 처리 기술을 개발하기 위해서는 목적에 맞게 구축된 대량의 고품질 데이터가 갖춰져야 하지만, 투입되는 시간과 인력, 비용 등의 문제로 이러한 자원을 확보하는 것이 쉽지 않다. 본 과제에서는 공익의 목적으로 데이터를 공개하여 자체적으로 자동 요약 기술 개발을 위한 데이터 구축에 어려움을 겪는 기업의 수요를 충족시키고자 한다.

(4) **신문 빅데이터의 활용도 제고**: 본 과제는 2019년 국어 빅데이터 구축 사업의 일환으로 진행되는 ‘신문기사 원문자료 수집 및 정제’ 사업의 결과물인 12억 어절 신문 말뭉치의 산업적 활용 방안을 제시한다. 본 과제에서 제안하는 신문 요약 말뭉치 구축 방법론에 따라 신문 빅데이터를 재가공하여 요약 개술 개발의 밑거름으로 삼을 수 있다.

제 2 장

신문기사
말뭉치 구축

1. 신문기사 말뭉치 구축 개요

1.1. 말뭉치 구축 목표

뉴스 기사 및 요약 관련 프로젝트 수행 경험을 바탕으로 본 사업에서 자체 기술력과 노하우를 활용한 구축 시스템의 사용성을 확보하고, 전산적 기술 지원 체계를 통하여 구축 시스템의 성능을 최적화하였다. 말뭉치 구축 목표는 다음과 같다.

목표	개요
요약 기술 개발을 위한 말뭉치 구축	기사 원문에 주제문이 주석된 추출 요약 말뭉치 기사 원문의 내용을 바탕으로 사람이 직접 생성한 추상 요약 말뭉치
요약 말뭉치 평가 지표 개발	기존 평가 알고리즘을 활용한 정량적 평가 지표 개발 작문 전문가를 활용하여 정성적 평가 지표 개발
요약 말뭉치 구축 지침 개발	한국어에 특화된 요약 말뭉치 구축 지침 개발 신문기사 요약문의 정략적, 정성적 평가 기준 설계
요약문 구축 시스템 개선	사용자 편의성 증대에 초점을 맞춘 요약문 구축 시스템 개선 대규모 요약 말뭉치 구축을 위한 다중 사용자 관리 기능 안정화
자동 요약 모델 검증	구축된 말뭉치를 기반으로 요약 모델 적용 및 성능 평가 최신 추출 및 추상 요약 방법론 적용
산업적 활용 가치 검증	국내에서 산업적으로 활용 중인 요약 서비스 검토

표 2 말뭉치 구축 목표 및 개요

1.2. 말뭉치 구축 방안

1.2.1. 구축 대상 말뭉치 선정

국립국어원에서 제공한 7,265 기사(2,000,215 어절) 중 아래 기준에 따라 선정된 기사를 우선 작업 대상으로 하여 4,451 기사(1,323,076 어절)를 대상으로 주석 및 요약 작업을 하였다. 그 중 작업 과정에서 납품 제외 대상을 제외하고 최종 납품 대상 기사로 4,389 기사(1,305,427 어절)를 선정하였다.

▶ 객관적인 정보 전달을 위하여 작성된 기사

신문기사의 주제는 '경제, 과학, 국제, 문화, 사람들, 사회, 스포츠, 정치, 지역, 기획, 사설, 오피니언' 등으로 분류할 수 있다. 이중 객관적인 정보만을 담고 있는 기사들을 자동 요약을 위한 말뭉치의 대상으로 선정하였다. 이에 따라 '기획, 사설, 오피니언'은 말뭉치 구축 대상에서 제외하였다. 이 세 분류로 구분되어 있지 않은 기사 중 이 세 분류에 해당하는 기사 또한 추후 정제 작업을 통하여 제외하였다.

▶ 적정 분량의 기사

분량이 너무 짧아서 요약이 불필요하거나 반대로 너무 길어서 세 문장 이내로 요약을 할 수 없는 기사는 말뭉치 구축 대상에서 제외하였다. 기사문에서 한 문장은 평균적으로 14개의 어절 정도로 이루어지기 때문에, 부제목(subhead)을 제외한 본문을 기준으로 200개 어절 이상 600개 어절 미만의 기사만을 말뭉치 구축 대상으로 선정하였다.

1.2.2. 납품 말뭉치 형식

납품 말뭉치는 기본적으로 국립국어원에서 제공한 파일과 동일한 형식인 XML 형식으로 구성하였다. 그림과 같이 메타 정보로 text date에 작성일, id에 원본 데이터의 기사 id, subclass에 기사 분류, subclass2는 요약 시 유의해야하는 기사 유형을 나타내어 주석하였다. 주제문은 <s type="key">를 통해 나타냈으며, 요약문은 <summarization> 태그를 이용해 나타냈다.

<text date="20110628" id="NWRW1800000030-0403" subclass="지역" subclass2="">

<s type="head">[서울] "태풍 시속 70㎞로 북상 중"(26일 오전 4시)… "2단계 비상근무령"(오전 7시 20분)… "상황 끝"(오후 4시 30분)</s>

<s type="subhead">●태풍 '메아리'로 긴박했던 재난안전대책본부 72시간</s>

<s type="key">지난 26일 오전 4시 서울 중구 종합방재센터 지하 2층 서울시 재난안전대책본부 종합상황실.</s>
<s>조용하던 50평 남짓한 상황실 내부가 술렁이기 시작했다.</s>
<s>직원들이 "어어…" 하면서 상황실에 설치된 가로세로 1.5m 스크린에서 눈을 떼지 못했다.</s>
<s type="key">시속 30㎞의 속도로 한반도를 향해 북상하던 태풍 '메아리'가 제주도 서남쪽 270㎞ 해상에서 갑자기 시속 70㎞로 빨라진 것이다.</s>

<s>사흘째 집에 못가고 24시간 비상 대기 중이던 재난안전대책본부 직원들의 움직임이 분주해졌다.</s>
<s>오전 7시 20분.</s>
<s>상황실을 지휘하던 이인근 서울시 도시안전본부장은 2단계 비상 근무령을 내렸다.</s>
<s>보통 비상 근무령은 재난안전대책본부장인 서울시장이 내리지만 이날은 통제관인 이인근 본부장이 직접 내린 후 오세훈 서울시장에게 보고했다.</s>
<s>1단계를 거치지 않고 바로 2단계로 격상했다.</s>
<s>그만큼 상황이 급박하다고 판단했다.</s>
<s>보고를 받은 오세훈 시장은 상황실을 찾아 피해 상황을 묻고 지시했다.</s>

<s>2단계 비상 근무가 시작되자 직원 50여명은 시내 태풍 피해 상황을 확인했다.</s>
<s>상황실에는 비가 내리고 있음을 표시하는 파란색 비상등에 계속 불이 들어오고 있었다.</s>
<s>오후 1시쯤 상황실에는 "태풍 메아리가 신의주 방향으로 북상하면서 별 영향 없이 서울 지역을 벗어날 것"이라는 기상업체의 예보가 들어왔다.</s>
<s>서울시는 기상 상황에 따라 기민하게 대처하기 위해 1시간마다 강수량을 예보하는 케이웨더라는 민간 기상업체와 계약을 맺고 있다.</s>

(중략)

<summarization>
<s>지난 26일 오전 태풍 '메아리'의 속도가 갑자기 빨라져 서울시 재난안전대책본부가 2단계 비상근무령을 내렸으나 다행히 태풍 진로가 바뀌어 큰 피해는 없었다.</s>
<s>수방 시스템을 자신해 왔던 서울시는 작년 '광화문 기습 폭우' 로 교훈을 얻어 장마 초반부터 비상 근무에 돌입했다.</s>
<s>예전의 수방 대책은 사후 대책이었으나 이제는 예방 시스템을 촘촘하게 갖춰 피해를 최소화하는 대책을 마련했다. </s>
</summarization>
<byline>최인준</byline>
</text>

표 3 요약 말뭉치 납품 xml 형식 예시

1.2.3. 요약 말뭉치 구축 지침 개발

요약 말뭉치 구축 지침은 구축할 말뭉치의 목적에 따라 두 종류로 나눌 수 있다. 첫 번째는 추출 요약이고 두 번째는 추상 요약이다. 해외 구축 요약 말뭉치의 구축 가이드라인, 국내 요약 작문 기존 사례, 실제 산업계 활용 연계 방안 등을 검토하여 요약 말뭉치 구축을 위한 지침을 설계하였다. 상세한 내용은 3장에 기술하였다. 추출 요약 말뭉치와 추상 요약 말뭉치 각각을 위하여 추가적으로 검토한 기준은 아래와 같다.

▶ 추출 요약 말뭉치

추출 요약은 기사 원문 중에서 전체 텍스트를 나타낼 수 있는 주제문을 선정하는 것이다. 주제문은 가장 상위의 정보를 담고 있어야 하며 중복된 정보를 가지고 있어서는 안 된다. 따라서 선정된 문장이 실제로 전체 텍스트의 내용을 잘 함축하고 있는지에 대한 기술적 품질 평가 기준이 필요하다. 또한 선정된 주제문이 잘못된 구문 구조나 맞춤법 오류 등을 가지고 있지 않은지에 대한 언어학적 품질 평가 기준이 필요하다.

▶ 추상 요약 말뭉치

추상 요약은 기사 원문의 내용을 바탕으로 전체 텍스트의 내용을 잘 나타낼 수 있는 새로운 문장을 생성하는 것이다. 요약문은 사소하거나 반복되는 내용을 삭제하고, 제시되는 항목과 행동을 보다 상위의 개념으로 일반화한 뒤, 중요한 내용들을 선정하여 구성된다. 따라서 생성된 요약문이 실제로 전체 텍스트의 내용을 잘 함축하고 있는지에 대한 품질 평가 방법을 제시하였다.

1.2.4. 작업자 교육 및 전문가 워크숍

사업 착수 보고회 이후 구축 작업자 교육 워크숍을 개최하였다. 워크숍은 주석 및 요약 말뭉치 구축 작업자를 대상으로 이루어졌으며 구축 작업 지침, 구축 시스템 튜토리얼 등을 교육하였다. 이를 통해 작업자들의 업무 역량 및 시스템 이용 역량을 강화하였다. 또한 구축 및 검수 과정의 이슈 산정과 지침 보완을 위하여 전문가 워크숍을 진행하였다.

교육 업무	교육을 위한 강사 선정 1) 요약 말뭉치 구축 시스템 구축 및 관리 담당자 2) 요약 말뭉치 구축 지침 개발자 교육 대상자에게 요약과 시스템 사용에 대한 교육 실시
반영/절차	사용자 교육 계획 수립: 사용자 교육 계획서 집합 교육, 현장 교육 등 다양한 교육 방법 지원 교육 계획, 준비, 실시, 평가 및 결과 분석 실시
대상/내용	전체 공통 교육 사용자 맞춤 교육 매뉴얼 등 교육 교재 제공 오류 보고 방안 교육
일정/조직	교육 훈련 인력 배치 교육 요구 사항을 파악하여 수행 단계별 교육 계획 수립

표 4 작업자 교육 프로세스

▶ 구축 작업자 대상 교육

- 1차

일시: 2019년 8월 8일 17:00 ~ 18:00

장소: 연세대학교 위당관 514호 언어정보연구원

참석자: 김동준, 여규병, 이세강, 이용원, 이창원

교육자: 주민재

교육 내용: 요약문 구축 지침 초안 공유 및 실무 교육

- 자동 요약 기술 개요 및 요약 데이터 구축 방법론 교육
    - 전체 데이터 산출 과정 개요
    - 자동 요약 기술 관련 주요 개념 설명
    - 요약문 데이터 구축 시 유의사항 공유

- 요약문 구축 도구 사용법 교육
    - 구축 작업 데모 시연
    - 보고 대상 이슈에 대해 설명
- 작업 절차 및 추후 일정 사항 공유
    - 작업량 할당 및 마감 일정 공유
    - 보고 체계 공유

- 2차

일시: 2019년 8월 30일 16:00 ~ 17:00

장소: 연세대학교 위당관 514호 언어정보연구원

참석자: 김동준, 여규병, 이세강, 이용원, 이창원

교육자: 주민재

교육 내용: 구축 데이터 대상 1차 검수 작업 안내

- 개정 지침 관련 (2019.8.26) 변경 사항 안내
    - 서브헤드라인 배제 이슈 → 시스템에도 관련 사항 반영하여 예방 조치
    - 길이, 내용 등 요약문 품질 관련 사항 안내
- 1차 검수 관련 작업 할당 및 일정 공유
    - 추후 작업자 충원 관련 계획 안내
    - 1차 검수 작업 절차 안내 및 일정 공유
- 요약문 생성 및 검수 시 이슈 피드백 진행
    - 시스템에 보고된 이슈 사항 관련 피드백 진행 및 지침 반영

▶ 학부생 작업자 대상 교육 내용

- 1차

일시: 2019년 9월 4일 15:00 ~ 14:00

장소: 연세대학교 위당관 514호 언어정보연구원

참석자: 강예지, 김민지, 손영랑, 윤선영, 조연수, 허나연

교육자: 오태환, 박석원

교육 내용: 요약문 구축 지침 공유 및 실무 교육

- 자동 요약 기술 개요 및 요약 데이터 구축 방법론 교육
    - 전체 데이터 산출 과정 개요
    - 자동 요약 기술 관련 주요 개념 설명

        - 요약문 데이터 구축 시 유의사항 공유
    - 요약문 구축 도구 사용법 교육
        - 구축 작업 데모 시연
        - 보고 대상 이슈에 대해 설명
    - 작업 절차 및 추후 일정 사항 공유
        - 작업량 할당 및 마감 일정 공유
        - 보고 체계 공유

- 2차

일시: 2019년 9월 5일 15:00 ~ 14:00

장소: 연세대학교 위당관 514호 언어정보연구원

참석자: 김아영, 이종혁, 정지원, 정혜진

교육자: 오태환, 박석원

교육 내용: (9/4 미참여자 대상) 요약문 구축 지침 공유 및 실무 교육

- 자동 요약 기술 개요 및 요약 데이터 구축 방법론 교육
    - 전체 데이터 산출 과정 개요
    - 자동 요약 기술 관련 주요 개념 설명
    - 요약문 데이터 구축 시 유의사항 공유
- 요약문 구축 도구 사용법 교육
    - 구축 작업 데모 시연
    - 보고 대상 이슈에 대해 설명
- 작업 절차 및 추후 일정 사항 공유
    - 작업량 할당 및 마감 일정 공유
    - 보고 체계 공유

▶ **전문가 워크숍**

일시: 2019년 11월 15일 13:00 ~ 16:00

장소: 백양누리 104호

참석자: 김한샘, 주민재, 김동준, 여규병, 이세강, 이용원, 이창호

내용: 요약 말뭉치 구축 과정 및 지침에 대한 최종 피드백 및 구축 방법론 점검

아래 7가지 사항에 대한 주요 피드백 청취 및 의견서 수합

1) 주제문 선정 지침에 대한 수정 사항

2) 요약문 생성 지침에 대한 수정 사항 (생성 기준 추가)

3) 서브클래스 부여 기준에 대한 사항

4) 요약 대상 기사 유형에 대한 사항

5) 구축 및 작업 과정에 대한 의견 사항

6) 구축 및 작업 도구에 대한 의견 사항

7) 그 외 의견

2019년 말뭉치 분석 연구 및 시범 구축 사업 전문가 주요 의견 요약

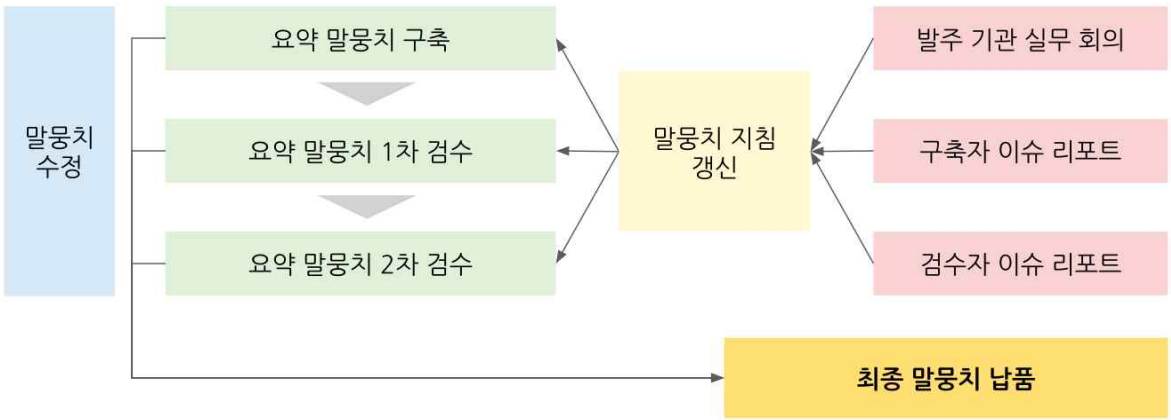
*전체 내용은 부록 참고

구분	내용
주제문 선정 지침에 대한 수정 사항	뉴스 기사의 주제문을 선택할 때 시간적 순서에 대한 기준이 필요함. (과거-현재-미래 순의 사건시 기준보다 기사 발행 시점을 우선적으로 고려할 필요가 있음.) 가급적 구체적 정보가 많이 포함된 문장을 선정하는 것이 추출 요약에 더 적합한 기준으로 사료됨. 선정된 주제문 어디에도 기사의 의도를 최대한 포괄하는 문장이 없는 경우도 빈번. 이 경우에는 오히려 주요 인물이나 사건의 이름이 분명히 드러난 문장을 선택하는 등의 대안적 지침이 필요.
요약문 생성 지침에 대한 수정 사항(생성 기준 추가)	요약문을 생성할 때 반드시 기사에서 다루고 있는 인물의 이름과 소속을 명기해야 한다는 지침도 추가할 필요가 있음. 일반적인 기사로 보이나 광고성 기사인 경우에 대한 처리 기준 필요. 요약문 생성이 기사 전체를 포괄할 수 있는 문장을 작성하는 데 초점이 있으므로 팩트 위주보다는 상위어 위주로 작성해야 함에도, 실제 작업에서는 이를 적용하기 어려운 경우가 있음: 요약문 작성의 확실한 방향성 정립 필요. 시간적, 논리적 순서에 입각하여 요약문을 재배열하는 것은 충실한 기사 요약에 방해가 될 소지가 있음. 따라서 기사의 의도를 충분히 반영할 수 있도록 요약 방향을 개정하는 것이 바람직.
서브클래스 부여 기준에 대한 사항	‘리뷰’ 기사의 경우 보도 시점과 사건시가 일치하지 않는 경우가 있는데 이에 대해 각 상황에 맞는 서브클래스 기준 마련 필요. 대부분의 기사가 ‘전언’에 해당될 수 있으므로 폐지 검토 필요. ‘인터뷰’ 기사는 ‘일문일답형’과 ‘풀어쓰기형’으로 구분할 필요가 있고, 외신이나 타 매체 인터뷰 인용 유형 역시 다룰 필요가 있음. ‘리뷰’를 세분하고 ‘서평’은 별도로 취급할 필요가 있음.
요약 대상 기사 유형에 대한 사항	기사 형식을 차용한 기획 광고는 요약 대상에서 제외하는 것이 바람직함. 탐사(르포), 묘사(스케치) 유형 기사에 대한 지침 필요.
구축 및 작업 과정에 대한 의견 사항	작업자의 경험에 따라 구축 품질에서 큰 편차를 보이는 것을 조정할 수 있어야 함. 기사를 써 본 작업자와 그렇지 못한 사람의 차이는 능률을 떨어뜨릴 뿐 아니라 큰 실수를 할 수 있는 원인으로 작용하기 때문. 검수A 직후 작업자에게 피드백을 제공하는 것이 바람직. 검수B 작업을 하면서 구축문과 검수A의 결과물을 비교할 수 있는 기회가 주어졌는데, 이 과정이 이후 작업에 도움이 되었음. 비전문가 그룹에게도 이러한 기회가 주어진다면 전체 결과물의 질적 향상이 도움이 될 것으로 기대. 현행 구축-검수A-검수B 절차에는 구축 초기 단계부터 전문 인력이 투입되는 것이 말뭉치 품질 제고에 바람직.
구축 및 작업 도구에 대한 의견 사항	검수A 이후 검수B 단계에서도 수정 작업이 가능한 도구가 필요함.
그 외 의견	언론사별로 들여쓰기, 구두점, 영어 철자, 외래어 표기 등의 내부 기준이 동일하지 않아 일관된 요약문 작성에 어려움이 있으며, 이에 대해 별도로 참고 기준을 마련할 필요가 있음. 최초 구축 작업이 추후 작업 품질에 큰 영향을 미침. 작업 할당 분량 및 주기에 대해 최적화 필요. 검수B 이후에도 추가적인 교열 과정이 필요할 수 있음. 기사 원문에 사실관계가 부정확한 경우 이에 대한 처리 기준 필요.

표 5 전문가 주요 의견

2. 말뭉치 구축 과정

산정된 작업량과 작업 속도를 검증하기 위한 시범 구축을 시행하고 국립국어원과의 협의를 통하여 1차, 2차 구축 시기를 조율하여 총 3차에 걸쳐 말뭉치를 구축하였다. 각 구축 단계마다 단계별 목표를 설정하고 피드백을 통한 개선 작업을 진행하였다. 반복적으로 기존 작업의 성과가 누적되어 반영되는 작업 파이프라인에 따라 작업이 반복될수록 고품질의 요약 말뭉치를 얻을 수 있었다.



<그림 1> 말뭉치 구축 파이프라인

▶ 시범 구축

(1) 시범 구축 대상: 전체 구축 대상 기사 중 약 10% 400건의 기사

- 최종 납품본에서 1건의 기사 제외

(2) 시범 구축 목표

- 구축 시스템 점검
- 주제문 주석 및 요약문 생성 방안 타당성 검토
- 작업 일정 점검
- 작업 품질 검증
- 작업자 특성 파악
- 구축 작업 필요 자원 산정

▶ 1차 구축

(1) 구축 대상: 전체 구축 대상 기사 중 약 30% 1,335건의 기사

- 최종 납품본에서 13건의 기사 제외

(2) 구축 목표

- 시범 구축 피드백 내용 반영
- 표본 추출 검수 및 자동 검수를 통한 품질 확보
- 오류에 대한 일괄 처리
- 처리 결과에 대한 재검토
- 고빈도 오류에 대한 재교육

▶ 2차 구축

(1) 구축 대상: 전체 구축 대상 기사 중 약 60% 2,716건의 기사
- 최종 납품본에서 2건의 기사 제외

(2) 구축 목표
- 1차 구축 피드백 내용 반영
- 표본 추출 검수 및 자동 검수를 통한 품질 확보
- 오류에 대한 일괄 처리
- 처리 결과에 대한 재검토

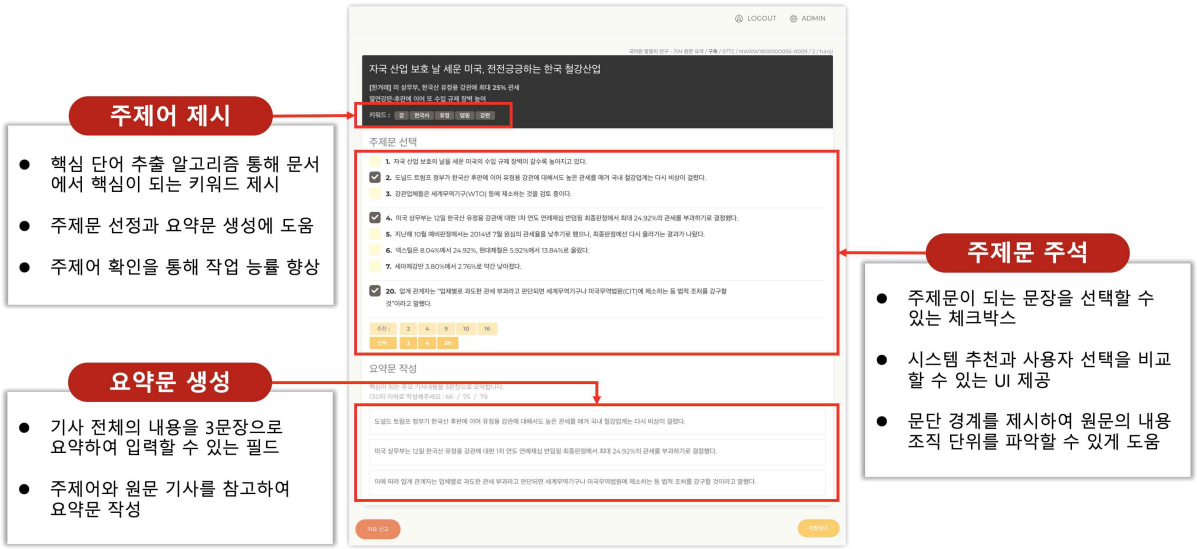
▶ 최종 말뭉치

(1) 구축 대상: 시범, 1차, 2차 구축 말뭉치 전체 검토 최종 4,389건의 기사(1,305,427어절)

(2) 구축 목표
- 개정된 지침과 기준에 따른 납품 대상 기사 확정
- 협의를 통한 수정 사항 반영

3. 말뭉치 구축 도구

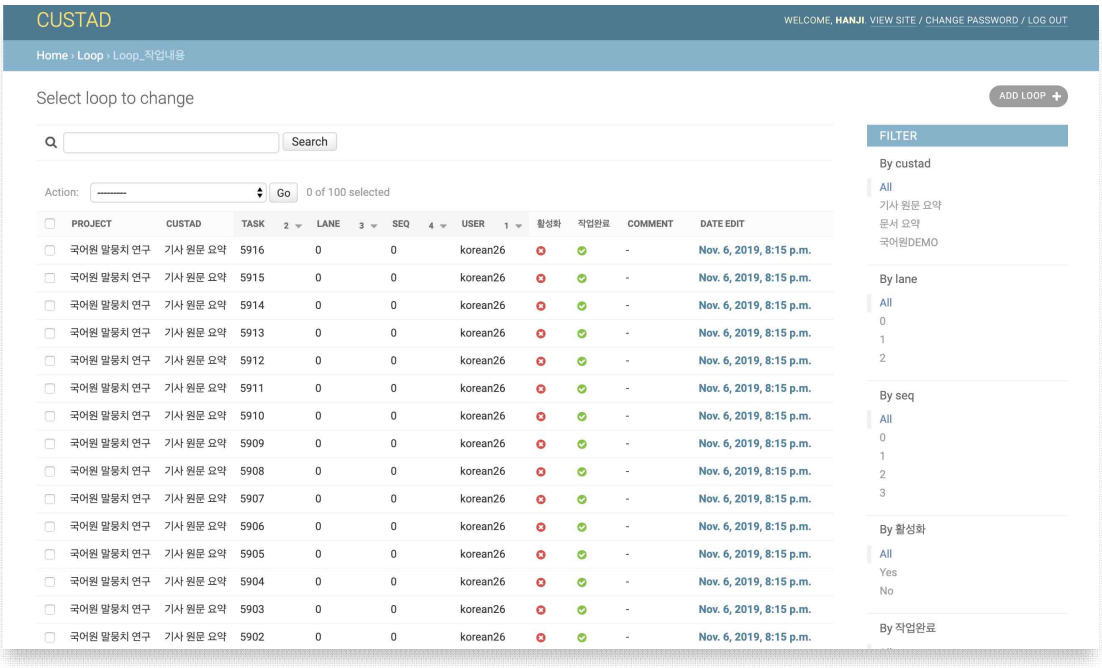
기존 말뭉치 주석 시스템을 활용하여 추출 요약과 추상 요약 말뭉치를 한 화면에서 동시 구축 가능한 시스템을 구현했다. 동일 원문 기사에 대하여 세 문장의 주제문을 선택하고, 원문과 다른 어휘를 사용한 요약문을 생성할 수 있다. 이러한 요약문 구축 도구를 토대로 말뭉치 구축을 진행하면서 작업자의 피드백을 받아 작업 편의를 도모하는 방식으로 개선하였다.



<그림 2> 작업 화면

3.1 관리 도구

프로젝트 구축 도구에서 최고관리자 권한으로 다음과 같은 기능을 제공, 사용 가능



<그림 3> 전처리 기사 업로드 관리자 도구 예시

(a) 프로젝트 생성/수정

- 새로운 프로젝트를 생성
- 기존에 생성된 프로젝트의 정보 수정

(b) 전처리 데이터 업로드 및 작업 할당

- 프로젝트에 할당을 위한 전처리 데이터 업로드 지원
- 등록된 데이터를 작업자에 중복 없이 할당

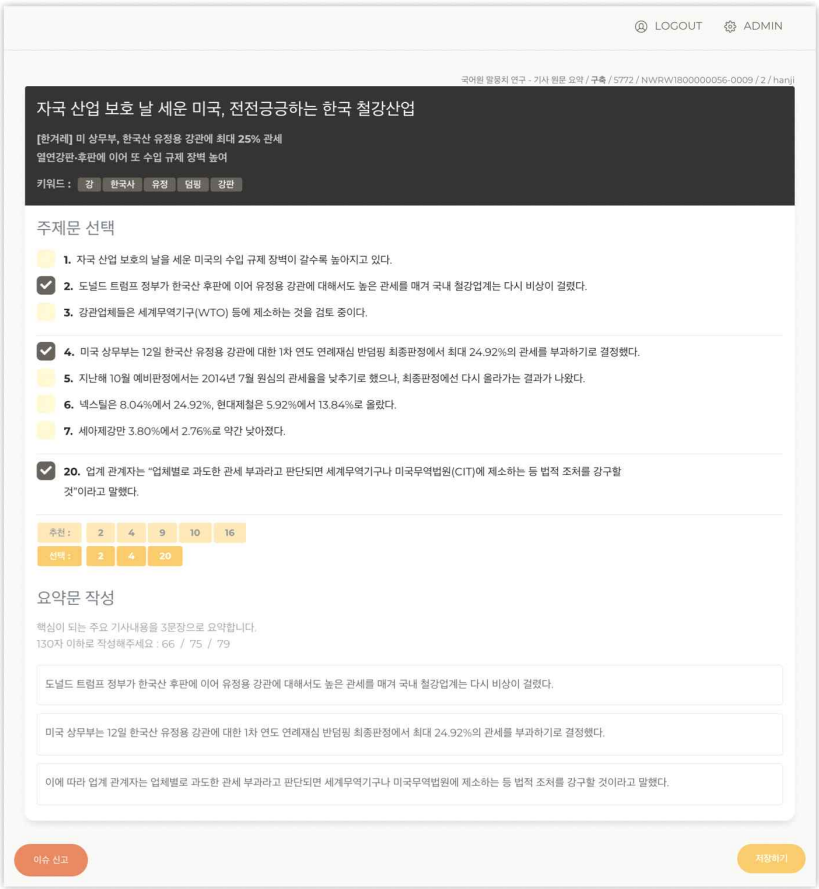
(c) 결과물(산출물) 확인

- 작업자의 완료된 작업물 조회/수정
- 1차, 2차 검수된 검수물 조회/수정

(d) 결과물 추출

- 최종 검수 완료 된 최종 검수물의 조회/수정
- 최종 구축 말뭉치의 추출

3.2. 작업 도구



<그림 4> 기사문 요약 도구 예시

(a) 할당된 프로젝트 주제문 선정 및 요약문 작성 작업 진행

- 프로젝트별로 업로드된 전처리 기사 내용을 작업자에 할당
- 할당된 기사의 내용 본문을 보고 요약문 세 문장 작성 후 제출
- 주제문의 경우 문장별 분리된 내용 중 5개의 주제문 선 제시 후, 그중 3개의 요약문을 선택하고 2개를 해제하는 형태의 주제문 선정 작업 진행

(b) 작업 진행 내용 확인/수정

- 작업 완료한 내용에 대한 조회 가능
- 기존 작업한 내용에 대한 수정 가능 (단, 검수 할당 후부터 수정 불가)

(c) 이슈 작업 부분 신고

- 기사나 작업에 문제가 있는 경우 이슈 신고 가능

3.3. 검수 도구

<그림 5> 검수 도구 예시

(a) 1차 : 구축 작업물에 대한 검수 진행

- 구축 작업을 진행한 작업자를 제외, 검수 진행
- 1차 검수 내용 완료 혹은 통과 시 2차 검수 대상으로 이관

(b) 2차 : 1차 검수물과 구축 작업물의 내용 상호 비교 검수

- 검수 품질을 높이기 위한 이중 검수 작업 진행
- 1차 검수에 대한 이슈 발견 시 검토 의견 첨부해 최종으로 이관

(c) 최종 : 2차 검수물에 대한 최종 검수 후 승인

- 구축 작업물, 1차 검수물, 2차 검수물에 대한 총체적 검수 진행
- 최종 검수 완료 시 산출물 정리 및 추출

4. 말뭉치 평가 방법론

4.1 정량적 평가 방법론

요약문 자동 평가 방법과 사람의 수동 평가 사이의 상관관계를 측정하여 로 어떤 평가 방법이 한국어 문서 요약에 가장 적합한지 실험을 통해 살펴보았다.

(1) ROUGE 점수 : ROUGE 점수는 추출 요약과 추상 요약 모두에서 공통으로 사용되는 평가 방법이다. ROUGE-1, 2와 ROUGE-L이 가장 많이 사용된다.

(2) ROUGE-BERT 점수 : ROUGE의 가장 큰 단점이 어휘 일치도만을 측정하고 의미적 유사성을 반영하지 못한다는 점이다. BERTScore는 의미 유사성을 고려한 문장 단위의 평가 방법이다. 본 과제에서는 한국어 문서 요약의 자동 평가를 위해 ROUGE와 BERTScore를 결합한 ROUGE-BERT를 사용한다. ROUGE-BERT는 수식 (1)과 (2)를 이용하여 계산한다.

$$ROUGE_{BERT}-N=\frac{\sum_{S\in Ref}\sum_{x_n\in S}\max_{y_n} f(x_n,y_n)}{\sum_{S\in Ref}\sum_{x_n\in S} count(x_n)} \qquad (1)$$

$$f(w_x,w_y)=\max\left(\frac{\mathrm{v_x}\cdot\mathrm{v_y}}{\|\mathrm{v_x}\|\times\|\mathrm{v_y}\|},\ 0\right) \qquad (2)$$

한국어 요약 자동 평가를 어떻게 수행하는 것이 인간 평가자와의 일치도가 가장 높은지 살펴보기 위해 다음과 같은 실험을 수행하였다. 실험 결과를 통해 한국어 요약 자동 평가의 기준을 마련할 수 있다. 실험 과정은 다음과 같으며, <그림 6>에 이를 도식화 하였다.

① 구축한 요약 말뭉치 중 48개의 신문기사를 선택한다. 해당 기사에 대해 주석자가 작성한 추상 요약을 함께 준비한다. 선택된 문장은 추상 요약에 추출 요약문이 하나도 포함되어 있지 않으며, 비교를 위해 Title과 Subtitle이 세 문장 이상인 기사들만 선택하였다. (시범 말뭉치와 1차 말뭉치 집합에서만 추출하였기 때문에 추상 요약이 추출 요약문보다 긴 것들이 선택되었다.)

② 단계 ①에서 준비한 신문기사에 대해 시스템의 요약 출력 결과를 얻는다. 본 과제에서는 Lead-3, SummaRuNNer, SummaRuNNer+Compression, 제목+부제, 사람의 추출요약의 5개 요약을 비교대상으로 삼는다.

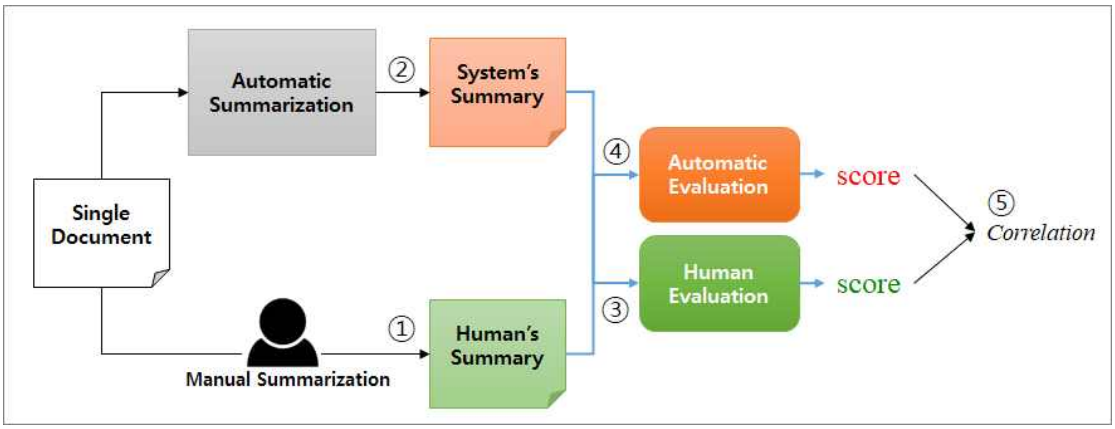
summary-1	Title + Subtitle (제목 및 부제)
summary-2	Human Extraction (수동 추출 요약)
summary-3	Lead-3 (기사 첫 세 문장)
summary-4	SummaRuNNer (추출 요약)
summary-5	SummaRuNNer + Compression (추출+축약 요약)

표 6 요약별 구성

③ 2명의 평가자가 ②에서 수집한 시스템 요약에 대해 수동으로 평가한다. 평가 방법은 1~5의 총괄 점수(holistic score)를 부여한다.

④ 단계 ②에서 수집한 요약들에 대해 평가 선택 사항을 바꿔가며 자동 평가를 수행한다. 자동 평가에는 ROUGE와 ROUGE-BERT를 사용한다.

⑤ 단계 ③과 ④의 평가 점수 사이의 상관관계를 측정한다.



<그림 6> 자동 평가와 인간 평가의 상관관계 측정 과정

우선 각 요약별 사람이 작성한 요약(reference summary)과의 평가 결과는 다음과 같다. 비교 요약의 길이가 모두 다르기 때문에 모든 평가는 ROUGE 재현율과 정확률을 함께 고려한 F-1 score이다.

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-BERT
summary-1 (Human Extraction)	71.56	58.58	60.81	72.35
summary-2 (Title + Subtitle)	22.35	6.39	14.14	42.20
summary-3 (Lead-3)	44.60	26.01	30.19	50.85
summary-4 (SummaRuNNer)	48.82	31.56	34.38	55.55
summary-5 (SummaRuNNer + Comp)	43.37	26.90	31.89	52.36

표 7 형태소 단위의 ROUGE (F1)

표 7의 Title과 Subtitle로 구성되는 summary-2의 경우 exact-matching만을 사용하는 ROUGE-1, 2, L의 경우에는 매우 낮은 점수를 받았다. 그러나 의미를 평가하

고자 도입된 ROUGE-BERT의 경우 그 점수 차이가 다른 평가 방법에 비해 줄어 든 것을 볼 수 있다. Human Extraction으로 구축된 summary-1은 매우 높은 점수로 평가되었는데 본 실험에서 사용된 reference summary가 추출 요약(summary-1)을 기반으로 작성되어 summary-1과 유사한 부분이 많은 것도 한 이유라고 하겠다.

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-BERT
summary-1 (Human Extraction)	56.26	44.95	51.17	71.48
summary-2 (Title + Subtitle)	6.05	0.74	4.97	40.57
summary-3 (Lead-3)	24.39	16.32	20.19	49.39
summary-4 (SummaRuNNer)	28.66	19.33	23.18	53.71
summary-5 (SummaRuNNer + Comp)	22.72	11.71	18.56	50.43

표 8 어절 단위의 ROUGE (F1)

표 8 어절 단위의 ROUGE 평가도 형태소 단위의 평가와 유사한 결과가 관찰되었다. 어절 단위이다 보니 summary-2의 결과가 다른 요약문보다 더 많이 낮아지는 것을 확인할 수 있었다. 또한 문장 축약까지 수행한 summary-5의 결과가 첫 세 문장을 추출한 summary-3보다 낮은데 reference 요약에 사용된 표현들이 원문과는 조사나 어미 등에서 약간의 차이를 보이기 때문으로 생각된다.

	ROUGE-1	ROUGE-2	ROUGE-L
summary-1 (Human Extraction)	69.15	56.65	60.48
summary-2 (Title + Subtitle)	25.14	8.67	17.41
summary-3 (Lead-3)	37.29	23.92	28.64
summary-4 (SummaRuNNer)	43.37	29.44	33.40
summary-5 (SummaRuNNer + Comp)	37.92	22.34	29.43

표 9 내용어만 고려 ROUGE (F1)

	평가자-1	평가자-2	평가자-평균
summary-1 (Human Extraction)	4.188	3.875	4.031
summary-2 (Title + Subtitle)	3.542	2.958	3.250
summary-3 (Lead-3)	2.083	2.542	2.313
summary-4 (SummaRuNNer)	2.889	2.844	2.867
summary-5 (SummaRuNNer + Comp)	2.844	2.622	2.733

표 10 Human Holistic Evaluation (1~5)

표 10은 2명의 인간 평가자가 각 요약문에 대해 1~5 점 사이의 총괄 점수를 부여한 결과이다. 인간 평가자의 가장 뚜렷한 평가 차이는 Title+Subtitle로 구성되는 summary-2에 대한 평가이다. summary-2는 기존의 자동 평가 방법 (ROUGE)에서는 항상 최하위의 성능을 보이던 반면 인간 평가자의 평가에서는 summary-1 다음의 평가 점수를 받았다. 이는 reference 요약과 전혀 다른 단어와 형식으로 작성되어 있는 요약문의 경우 자동 평가 방법으로는 좋은 점수를 얻을 수 없지만 인간 평

가자의 이해를 바탕으로는 좋은 점수를 얻을 수 있음을 보여준다. 또한, 자동 평가 방법에서 summary-3의 성능에 못 미치던 summary-5도 인간 평가자의 평가 결과에서는 더 좋은 평가 점수를 받았다.

다음 표 11은 전체 성능결과를 통합한 표이다. 같은 요약문에 대해 각 평가 방법에 따라 어느 정도의 점수 차이가 나는지를 확인해 볼 수 있다.

		summary-1	summary-2	summary-3	summary-4	summary-5
ROUGE-1	어절	56.26	6.05	24.39	28.66	22.72
	형태소	71.56	22.35	44.60	48.82	44.75
	내용어	69.15	25.14	37.29	43.37	37.92
ROUGE-2	어절	44.95	0.74	16.32	19.33	11.71
	형태소	58.58	6.39	26.01	31.56	26.90
	내용어	56.65	8.67	23.92	29.44	22.34
ROUGE-L	어절	51.17	4.97	20.19	23.18	18.56
	형태소	60.81	14.14	30.19	34.48	31.89
	내용어	60.48	17.41	28.64	33.40	29.43
ROUGE-BERT	어절	71.48	40.57	49.39	53.71	50.43
	형태소	72.35	42.20	50.85	55.55	52.36
평가자-1		4.188	3.542	2.083	2.889	2.844
평가자-2		3.875	2.958	2.542	2.844	2.622
평가자-평균		4.031	3.250	2.313	2.867	2.733

표 11 전체 평가 방법 비교

다음은 어떤 자동 평가 방법이 인간 평가자와 가장 유사한 평가 결과를 도출하는지 살펴보기 위하여 인간 평가자와의 피어슨 상관계수를 구해 보았다.

2명의 인간 평가자와 자동 평가 간의 피어슨 상관계수는 표 12, 표 13과 표 14에 제시하였다.

		summary-1	summary-2	summary-3	summary-4	summary-5
ROUGE-1	어절	0.154530	0.273302	0.703823	0.438845	0.498494
	형태소	0.183271	0.273936	0.685459	0.484793	0.549751
	내용어	0.164880	0.262711	0.695547	0.481010	0.544899
ROUGE-2	어절	0.144571	0.340693	0.682784	0.370900	0.426264
	형태소	0.150265	0.276562	0.719429	0.461542	0.508948
	내용어	0.131846	0.185143	0.700901	0.432865	0.519399
ROUGE-L	어절	0.142580	0.322707	0.688256	0.393127	0.428780
	형태소	0.145761	0.337197	0.727356	0.431730	0.458688
	내용어	0.145803	0.273234	0.680056	0.417615	0.464527
ROUGE-BERT	어절	0.173393	0.270005	0.692474	0.505259	0.498709
	형태소	0.217423	0.254284	0.697888	0.531966	0.522321
평가자-2		0.261636	0.440331	0.726647	0.626290	0.608915

표 12 인간평가자-1과 자동 평가 방법 간의 피어슨 상관계수

우선 전체적인 상관계수가 낮게 나옴을 볼 수 있다. 피어슨 상관계수는 0.7~1.0의 경우 강한 양적 상관관계, 0.3~0.7의 경우 뚜렷한 양적 상관관계, 0.1~0.3의 경우 약한 양적 상관관계가 있다고 해석한다.

summary-1에 대한 상관계수가 제일 낮은데 그 이유는 summary-1은 사람이 추출한 추출요약으로 인간평가자에게는 4.0(1~5) 이상의 점수를 받은 반면 자동 평가 방법에서는 ROUGE-1-어절의 경우 13.48~91.93까지 다양한 점수를 받았다. 자동 평가로 받은 점수가 13.48인 요약문에 인간평가자는 4점을 부여한 경우가 있는 반면, 자동 평가 점수가 91.93인 요약문에 인간평가자가 5점을 부여하였다. 그러다보니 인간평가자와의 상관관계가 매우 낮게 나오고 있다. summary-3의 경우 기사의 첫 세 문장을 추출하여 만든 요약이다 보니 요약문의 성능에 뚜렷한 차이가 보이기 때문에 인간평가자 사이의 상관계수도 높고 자동 평가 방법과의 상관계수도 높게 나타났다.

summary-3, 4, 5에서 ROUGE-1-형태소를 제외하고는 어절보다는 형태소 단위로 평가하는 것이 인간평가자와의 상관계수가 높았다.

		summary-1	summary-2	summary-3	summary-4	summary-5
ROUGE-1	어절	0.137184	0.209903	0.756780	0.596608	0.588579
	형태소	0.192444	0.165499	0.741774	0.646307	0.641482
	내용어	0.167294	0.283095	0.756342	0.609340	0.628056
ROUGE-2	어절	0.122608	0.143678	0.715418	0.518699	0.516784
	형태소	0.173418	0.128358	0.774525	0.607020	0.585007
	내용어	0.127345	0.092941	0.740236	0.549001	0.554771
ROUGE-L	어절	0.132147	0.255156	0.703942	0.598816	0.609312
	형태소	0.209738	0.249896	0.735632	0.621574	0.586190
	내용어	0.176408	0.304134	0.713569	0.582197	0.562472
ROUGE-BERT	어절	0.177489	0.298135	0.752734	0.591563	0.556974
	형태소	0.181052	0.258265	0.746973	0.605146	0.564424
평가자-1		0.261636	0.440331	0.726647	0.626290	0.608915

표 13 인간평가자-2와 자동 평가 방법 간의 피어슨 상관계수

평가자-2의 경우에도 평가자-1과 유사한 경향을 보인다. 평가자-2와의 상관계수는 전반적으로 ROUGE-1에서 가장 높게 나왔다.

		summary-1	summary-2	summary-3	summary-4	summary-5
ROUGE-1	어절	0.179801	0.277178	0.787411	0.578809	0.607741
	형태소	0.235229	0.247216	0.769562	0.631940	0.665832
	내용어	0.207308	0.321882	0.782964	0.608327	0.655469
ROUGE-2	어절	0.163973	0.265089	0.753216	0.497656	0.527474

	형태소	0.204421	0.223183	0.805419	0.596781	0.611292
	내용어	0.161019	0.154189	0.776521	0.547820	0.599436
ROUGE-L	어절	0.170015	0.332266	0.749465	0.556173	0.582282
	형태소	0.228983	0.335789	0.787321	0.589694	0.584985
	내용어	0.204367	0.341388	0.750781	0.559264	0.574418
ROUGE-BERT	어절	0.219163	0.335714	0.779353	0.610616	0.589598
	형태소	0.244133	0.300820	0.778823	0.632543	0.606583
평가자-1		0.700706	0.790773	0.920512	0.889875	0.888811
평가자-2		0.871928	0.897776	0.937312	0.912972	0.904730

표 14 2명의 인간평가자-평균과 자동 평가 방법 간의 피어슨 상관계수

다음 표 15는 BERT-ROUGE-형태소와 다른 자동 평가 방법 간의 상관계수를 측정해 보았다. BERT-ROUGE-형태소의 경우 ROUGE-1-내용어 평가와 가장 상관계수가 높았다.

		summary-1	summary-2	summary-3	summary-4	summary-5
ROUGE-1	어절	0.919854	0.460935	0.917319	0.865961	0.825220
	형태소	0.928383	0.747927	0.933779	0.928039	0.912767
	내용어	0.946613	0.801949	0.961091	0.940050	0.932119
ROUGE-2	어절	0.892470	0.359888	0.853832	0.788590	0.753394
	형태소	0.926574	0.583234	0.928541	0.896991	0.887799
	내용어	0.931644	0.586422	0.924386	0.883181	0.888674
ROUGE-L	어절	0.879132	0.466876	0.896164	0.862526	0.837041
	형태소	0.862940	0.669125	0.912126	0.901934	0.895610
	내용어	0.884728	0.673477	0.921009	0.895098	0.895699

표 15 BERT-ROUGE-형태소와 다른 자동 평가 방법간의 상관계수

4.2 정성적 평가 방법론

4.2.1. 총체적 평가

요약문에 대한 정성적 검증 방안을 수립하기 위해 구축된 요약 데이터의 일부를 무선표집하여 요약문 전반에서 확인되는 특성들을 확인하였다. 분석적 평가에 필요한 항목은 1)주제의 명확성, 2)주제문 주석의 정확성, 3)요약문 내용의 완결성, 4)요약문의 간결성, 5)접속어・지시어의 활용이 추출되었다. 이를 정성 평가의 주요 점검 기준으로 삼아 5점 척도 평가를 진행하였다. 도출된 요약문 품질 점검 기준안은 다음과 같다.

점검 목록	세부 점검 기준 체크리스트	5점 척도 평가
생성된 요약문이 주제문 주석 결과와 질적으로 차이가 있는가?	☑ 선택된 주제문을 그대로 연결하는 방식으로 요약문을 생성하지는 않았는가? ☑ 생성된 요약문이 원 기사의 핵심 내용을 충분히 포괄할 수 있는가? ☑ 전체 기사 중 생성된 요약문 각각이 내용적으로 대표하지 못하고 있는 단락은 없는가?	그렇다 ⑤ ④ ③ ② ① 그렇지 않다
요약문에 직접 인용을 하지는 않았는가?	☑ 생성된 요약문에 직접 인용이 그대로 포함되지는 않았는가? ☑ 기사 내 (여러) 직접 인용문의 내용을 충분히 포괄할 수 있도록 재기술되었는가? ☑ 직접 인용의 내용을 옮기는 과정에서 단순히 인용 부호만을 누락하는 방식으로 요약하지는 않았는가?	그렇다 ⑤ ④ ③ ② ① 그렇지 않다
요약문에 적절한 연결어를 사용하였는가?	☑ 문장과 문장을 연결하는 연결어의 사용이 적절한가? ☑ 연결어가 필요한 경우에 이를 생략하지는 않았는가?	그렇다 ⑤ ④ ③ ② ① 그렇지 않다
기사 내용이 요약문에 충분히 압축되었는가?	☑ 지엽적인 내용을 무분별하게 포함하여 문장이 길어지지는 않았는가? ☑ 문장 내에서 상위어나 더 일반적인 기술로 대체할 수 있는 부분은 없는가? ☑ 요약문이 주제문보다 길이가 더 길면서 주제문보다 더 적은 내용을 담고 있지는 않은가?	그렇다 ⑤ ④ ③ ② ① 그렇지 않다

표 16 요약문 품질 점검 기준안

4.2.2. 총체적 평가에 따른 요약문 분류

무선표집된 50개의 요약문을 총체적으로 평가한 결과, 모두 5개의 CASE로 분류할 수 있었다. 이번 작업은 주제문 주석과 요약문 생성이 연결되어 있는 바, 주제문 주석의 정확성과 요약문의 질을 종합적으로 판단해야 했다. 5개 CASE의 성격은 다음과 같다.

CASE	양상
CASE 1	주제문 주석과 생성된 요약문이 모두 양호한 경우
CASE 2	주제문 주석은 양호하나 생성된 요약문 내용이 주제문과 거의 차이가 없는 경우
CASE 3	주제문 주석은 양호하나 생성된 요약문 내용이 부족한 경우
CASE 4	주제문 주석 일부가 적절하지 않으나 생성된 요약문은 양호한 경우
CASE 5	주제문 주석 일부가 적절하지 않고 생성된 요약문도 적절하지 않은 경우

표 17 총체적 평가에 따른 최종 요약문의 양상 분류

4.2.3. 분석적 평가 항목에 따른 요약문 평가

총체적 평가를 통해 추출한 5개 항목에 따른 평가는 CASE 1이 다른 CASE에 비해 모든 항목에서 높은 점수를, CASE 5는 모든 항목에서 전반적으로 낮은 점수를 기록했다. 특히 CASE 5는 '주제문 주석의 정확성'이 다른 CASE에 비해 매우 낮은 편이었다(자세한 표는 별도 액셀 파일 참조). CASE별 특성에 대해서는 4.2.4. 주제문 주석과 요약문 질의 관계 참조.

4.2.4. 주제문 주석과 요약문 질의 관계

이번 작업에서는 '주제문 주석하기 → 요약문 생성하기'라는 단계를 설정했다. '주제문 주석하기'는 원문에서 요약문을 보다 정확하면서도 효과적으로 생성하기 위해 거쳐야 하는 중간 단계로서의 역할을 담당한다. 원문에서 요약문 생성 단계로 바로 넘어가는 경우, 요약문이 생성되는 실질적인 근거가 미약하고 요약문의 질이 작업자의 관

점에 따라 좌우될 가능성이 적지 않다. 이러한 한계를 극복하기 위해 설정한 중간 단계가 ‘주제문 주석하기’라고 할 수 있다. 주제문 주석을 통해 요약문 생성에 최소한의 객관적인 근거를 확보될 수 있다고 볼 수 있다. 객관적으로 원문에서 가장 핵심적인 문장들을 주제문으로 주석함으로써 작업자의 관점이나 성향에 따라 요약문의 질이 좌우되는 가능성을 최소화하는 것은 물론, 요약문의 질을 제고하는데 효과적이라고 볼 수 있다. 하지만 이렇게 주제문의 역할을 기대하기 위해서는 하나의 전제가 필요하다. 즉, 주제문(들)이 정확하게 주석되어야 한다. 주제문이 정확하게 주석되지 않으면 주제문을 기반으로 생성되는 요약문의 질을 보장하기 어렵다.

정성적으로 요약문을 검증한 결과, 일부 요약문에서 이러한 문제가 확인되었다. 주제문이 적절하게 주석되지 않은 경우, 발생할 수 있는 문제는 요약문의 질 저하였지만 세부적으로는 두 가지 부정적 결과가 유발될 수 있다. 첫째, 원문의 핵심 내용이 정확하게 파악되지 않음으로써 주석된 주제문을 기반으로 생성된 요약문이 원문의 핵심 내용의 일부를 포함하지 못할 수 있다. 둘째, 잘못 주석된 주제문이 이후 주제문 주석 과정 및 결과에 부정적인 영향을 미칠 수 있다.

1, 2차 검수를 거쳐서 최종적인 요약문들을 대상으로 무선표집된 50개의 요약문을 정성적으로 검증한 결과, ‘원문 → 주제문 주석 → 요약문 생성’이라는 3단계를 통해 주석・생성되는 주제문과 요약문의 결과는 아래와 같이 확인되었다.

CASE	상태	개수	백분율(%)
CASE 1	주제문 주석과 생성된 요약문이 모두 양호한 경우	27	54
CASE 2	주제문 주석은 양호하나 생성된 요약문 내용이 주제문과 거의 차이가 없는 경우	2	4
CASE 3	주제문 주석은 양호하나 생성된 요약문 내용이 부족한 경우	2	4
CASE 4	주제문 주석 일부가 적절하지 않으나 생성된 요약문은 양호한 경우	8	16
CASE 5	주제문 주석 일부가 적절하지 않고 생성된 요약문도 적절하지 않은 경우	11	22

표 18 최종 요약문의 주제문과 요약문의 양상(CASE 별)

50개의 최종 요약문 중에서 25개 요약문은 주제문 주석과 생성된 요약문이 모두 양호(CASE 1)했다. CASE 1이 최종 요약문 중에서 54%를 차지한다는 결과는 주제문 주석 결과가 최종 요약문의 질에 미치는 영향이 매우 높다는 것을 보여준다. 주제문 주석은 양호하나 생성된 요약문 내용이 주제문과 거의 차이가 없었던 CASE 2는 전체에서 매우 적은 수를 차지했는데 이는 이번 작업 중에서 배포된 지침(2019.8.19./8.26 업데이트 지침- 주석된 주제문과 (최종) 요약문 생성의 관계, 주제문과 최종 요약문 차이 두기)이 일정한 영향을 미친 것으로 보인다.

주목해야 할 부분은 CASE 3, 4, 5이다. 주제문 주석은 양호하나 생성된 요약문 내용이 부족한 경우인 CASE 3은 최종 요약문에서 차지하는 비율이 4%로 높지는 않다. 구체적으로 살펴보면 주제문에 비해 요약문의 길이가 너무 길거나 주제문의 내용을 요약문이 충분히 포괄하지 못한 경우이다. 전자의 경우 요약문에 최대한 많은 내용을 기술하려는 결과로, 후자의 경우는 반대로 주제문 내용을 모두 반영할 경우 요약문이 지나치게 길어질 수 있다고 생각하고 원문의 일부만 기술한 것으로 보인다.

주제문	요약문	비고
- 북한이 지난해 12월 1일 단행한 군사분계선 육로 통행 및 개성공단 등 체류 제한 조치를 21일자로 해제한다. - 이번 조치로 개성공단과 금강산을 오가는 인력 및 물자의 이동과 경협 사무를 위한 당국 간 의사소통 채널이 지난해 12월 1일 이전으로 대부분 원상회복될 것으로 보인다. - 북한은 또 김대중 전 대통령의 서거에 조의를 표하기 위해 조문단 6명을 1박 2일 일정으로 서울에 보내기로 했다.	- 북한이 지난해 12월 1일 단행한 군사분계선 육로 통행 및 개성공단 등 체류 제한 조치를 21일 자로 해제하고 판문역과 파주역을 오가는 남북 화물열차 운행을 재개하며 개성공단 내 남북 경제협력협의사무소를 다시 열기로 했다. - 이번 조치로 개성공단과 금강산을 오가는 인력 및 물자의 이동과 경협 사무를 위한 당국 간 의사소통 채널이 지난해 12월 1일 이전으로 대부분 원상회복될 것으로 보인다. - 북한은 또 김기남 노동당 중앙위원회 비서 등 김대중 전 대통령 조문단 6명을 1박 2일 일정으로 서울에 보내기로 했으며, 조문단은 북한 특별기편으로 서해 직항로로 21일 오후 김포공항에 도착해 조문할 예정이다.	주제문에 비해 요약문의 길이가 김.
- 국회의원 총선거와 대통령 선거를 앞두고 대기업 정책을 둘러싼 논쟁이 뜨거운 가운데 한국경제연구원이 29일 서울 중구 웨스틴조선호텔에서 ‘대기업 정책의 쟁점과 바람직한 방향’을 주제로 심포지엄을 열었다. - 참석자들은 정치권의 출총제 부활 주장 및 정부의 동반성장 정책에 대해 집중 토론했다. - 동반성장 정책에 대해서도 찬반이 엇갈렸다.	- 국회의원 총선거와 대통령 선거를 앞두고 대기업 정책을 둘러싼 논쟁이 뜨거운 가운데 한국경제연구원이 29일 서울 중구 웨스틴조선호텔에서 ‘대기업 정책의 쟁점과 바람직한 방향’을 주제로 심포지엄을 열었다. - 출자총액제한제도에 대해 황인학 한경연 기업정책연구실장은 “30대 그룹의 매출 집중도는 점차 낮아지는 추세”라며 “재벌 규제 방안은 설득력이 없다”고, 전성인 홍익대 경제학과 교수는 “재벌의 몸집을 줄이기 위한 노력이 시급하다”고 말했다. - 동반성장 정책의 효과에 대해서도 김경묵 덕성여대 교수는	요약문에는 3번째 주제문이 구체화된 내용이 부족함. 원문과 대조했을 때 동반성장정책에 대한 찬반 중 한 쪽 입장만 기술되어 있음.

	“정부 기대와 달리 소득 양극화 해소에 도움이 되지 않고 있다”고, 김세종 중소기업연구원 선임연구위원은 “중소기업 발전을 위한 대기업의 역할을 더 고민해야 한다”고 조언했다.	

표 19 CASE 3의 구체적 양상

CASE 4는 주석된 주제문 중의 일부가 적절하지 않으나 생성된 요약문은 양호한 경우이다. 일부 주제문이 원문 전체 내용을 포괄하지 못하거나 지엽적인 내용을 기술하고 있다. 원문의 내용을 보다 포괄적으로 기술하고 있는 문장이 있으므로 대체되어야 한다. 최종 요약문에서는 원문의 다른 문장(포괄적으로 기술하고 있는 문장)을 활용하여 적절하게 요약된 것으로 판단된다. 이 경우는 요약문이 1 ,2차 검수과정에서 수정되어 최종 요약문의 질이 높아졌을 것으로 추정된다.

주제문	요약문	비고
- 이날 정조국의 선제 결승골로 1-0으로 제주를 꺾은 서울은 스플릿시스템 A그룹에서 승점 90으로 울산과 3-3으로 비긴 전북(승점 78)을 따돌리고 2010년에 이어 2년 만에 정상에 복귀했다. - 30골로 역대 한 시즌 최다골을 터뜨린 데얀과 17골, 18도움을 한 몰리나의 파괴력은 그야말로 핵폭탄 같다. - 서울은 우승상금 5억 원과 2013년 아시아축구연맹(AFC) 챔피언스리그 출전권을 획득했다.	- 서울은 정조국의 선제 결승골로 제주를 1-0으로 꺾어 스플릿시스템 A그룹에서 2년 만에 정상에 복귀했다. - 서울은 30골로 역대 한 시즌 최다골을 터뜨린 데얀과 17골 18도움을 기록한 몰리나의 핵폭탄같은 파괴력으로 가장 안정적인 경기력을 유지했다. - 서울은 우승상금 5억 원과 2013년 아시아축구연맹(AFC) 챔피언스리그 출전권을 획득했다.	두 번째 주제문은 원문에서 ‘서울은 가장 안정적인 경기력을 보유하고 있다’는 내용의 지엽적인 부분으로 주제문으로 적절하지 않으며 ‘박 감독의 지적처럼 서울은 가장 안정적인 경기력을 보유하고 있다’로 대체되어야 함. 하지만 요약문에서는 ‘가장 안정적인 경기력을 유지했다’는 내용이 포함되어 있어 양호한 것으로 판단됨.
- 법관 재임용제도 개선방안을 논의하기 위한 판사회의가 17일 서울중앙지법 등 서울지역 3개 법원에서 열렸다. - 둘째, 현행 근무평정제도를 근간으로 하는 연임심사제도는 객관성 투명성이 담보되고 방어권이 보장되도록 개선돼야 한다”는 내용을 담았다. - 이번 단독판사회의를 시작으로 판사회의는 당분간 확대될 것으로 보인다.	- 법관 재임용제도 개선방안을 논의하는 판사회의가 17일 서울중앙지법 등 서울지역 3개 법원에서 열렸다. - 이날 판사회의는 “이번 연임심사과정에서 나타난 문제점이 재판의 독립을 해칠 우려가 있으며, 현행 근무평정제도를 근간으로 하는 연임심사제도는 객관성 투명성과 방어권이 보장되어야 한다”는 결의문을 발표했다. - 이를 시작으로 20일에는 의정부지법에서, 21일에는 수원지법과 광주지법에서 각각 단독판사	두 번째 주제문은 원문에서 ‘이날 의제는 ‘연임심사제도의 제반 문제점 및 개선방안 논의’ ‘근무평정제도의 문제점 및 개선방안 논의’ 등 두 가지다.‘로 대체되어야 한다.’ 내용의 지엽적인 부분으로 주제문으로 적절하지 않으며 앞의 문장으로 대체되어야 함. 하지만 요약문에서는 이 내용이 전반적으로 기술되고 두 번째 주제문 내용이 생략되어 양호한 것으로 판단됨

	회의가 열릴 예정이어서 판사회의가 당분간 확대될 것으로 보인다.	

표 20 CASE 4의 구체적 양상

문제가 되는 것은 CASE 5이다. 주제문 주석의 일부가 적절하지 않고 생성된 요약문도 적절하지 않은 경우가 전체의 22%(11개)를 차지하는 것으로 확인되었다. CASE 5는 CASE 1 다음으로 차지하는 비율이 높다. CASE 5의 경우 주제문 주석 과정부터 문제가 발생했을 가능성이 있다. 주제문을 잘못 주석하는 경우 이후 주제문(들) 주석에도 부정적인 영향을 미칠 수 있다. 주석자 입장에서는 이전 정보(이미 주석된 주제문)와의 연관되는 정보를 처리하는 것이 자연스러운 것으로 이해하기 때문이다.

표 21은 CASE 5의 요약문의 질이 상대적으로 낮은 원인이 잘못 주석된 주제문에 있다는 점을 잘 보여준다. 지엽적인 내용을 다루는 문장이 주제문으로 주석됨으로써 이를 기반으로 생성되는 요약문의 질에도 부정적인 영향을 미치는 것이다.

주제문	요약문	비고
- 코피는 여자보다 남자가 많이 흘리고, 특히 6세 남자 어린이가 코피 때문에 가장 많이 병원을 찾는 것으로 나타났다. - 6세 어린이가 코피를 가장 많이 흘리는 이유는 코를 후비거나 코를 세게 푸는 습관 때문으로 보인다. - 또 고혈압과 간질환이 있는 사람이 코피를 흘릴 가능성이 높다.	- 국민건강보험공단이 2006년부터 2010년까지 코피로 진료를 받은 환자를 분석한 결과 코피는 여자보다 남자가, 특히 6세 남자 어린이가 많이 흘리는 것으로 나타났다. - 6세 어린이가 코피를 가장 많이 흘리는 이유는 코를 후비거나 코를 세게 푸는 습관 때문으로 보인다. - 또 고혈압과 간질환이 있는 사람이 코피를 흘릴 가능성이 높고 혈우병, 백혈병, 혈소판 감소증 등 혈액질환으로 코피가 나기도 한다.	‘코피는 남자가 더 많이 흘리고 특히 6세 어린이들이 그 증상이 더 심한 이유 그리고 코피를 예방하는 방법’ 이 원문의 중심 내용이다. 따라서 세 번째 주제문은 ‘코피를 예방하려면 코를 건드리는 습관부터 버려야 한다’는 문장으로 대체되어야 한다. 요약문은 잘못 주석된 주제문들을 기반으로 생성됨으로써 코피를 예방하는 방법에 대한 내용이 전혀 기술되지 않았으며 상대적으로 지엽적인 고혈압, 간질환 환자들의 코피 가능성에 대해 기술되었다.
- 명품창출포럼은 이날 서울 서초구 양재동 엘타워에서 창립총회를 열었다. - 명품포럼의 초대 회장으로 선출된 박성철 신원 대표는 “경제가 성장하고 국민소득이 증가하면 소비재든 자본재든 최고 상품이 세계 시장을 리드한다”며 “한국 기업도 이제 세계 최고의 명품이 되지 않으면 경쟁에서 뒤처질 수밖에 없다”고 말했다. - 허경 지경부 기술표준원 원장은 “원가나 품질 등 고전적인	- 대기업 30곳과 중견·중소기업 70곳으로 구성된 명품창출포럼이 1일 서울 서초구 양재동 엘타워에서 창립총회를 열었다. - 명품포럼의 초대 회장으로 선출된 박성철 신원 대표는 “경제가 성장하고 국민소득이 증가하면 소비재든 자본재든 최고 상품이 세계 시장을 리드한다”며 “한국 기업도 이제 세계 최고의 명품이 되지 않으면 경쟁에서 뒤처질 수밖에 없다”고 말했다. - 허경 지경부 기술표준원 원장	두 번째와 세 번째 주제문이 원문에서 상대적으로 지엽적인 내용을 기술하고 있는 것으로 모두 잘못 주석된 것으로 판단된다. 원문의 전체적인 내용을 감안할 때 두 번째 주제문은 ‘명품포럼은 회원 수를 100개 기업으로 한정해 결속력을 유지하면서도 기존 회원이 세계 시장점유율 3위 이내의 명품을 만들면 명예회원이 되는 대신에 신규회원 1개사를 새로 가입시키는 방식으로 운영된다’로, 세 번째

경쟁 요소로는 세계 시장에서 경쟁력 확보에 한계가 있다”며 “한국 기업들이 소비자의 마음을 움직이고 감성을 자극하는 명품을 만들 수 있도록 정부도 적극 지원하겠다”고 말했다.	은 “원가나 품질 등 고전적인 경쟁 요소로는 세계 시장에서 경쟁력 확보에 한계가 있다”며 “ 소비자의 마음을 움직이고 감성을 자극하는 명품을 만들 수 있도록 지원하겠다”고 말했다.	주제문은 ‘정부 역시 명품 제품을 만들기 위해 월간 명품잡지를 발행하고 ‘대한민국 명품 발굴 콘테스트’를 개최하기로 했다.‘로 대체되어야 한다. 두 개의 주제문이 잘못 주석됨으로써 요약문 역시 원문의 지엽적인 내용을 중심으로 생성된 것으로 판단된다.
- 특히 서울은 2010년 5월 5일 성남과의 홈경기에서 6만747명의 역대 K리그 최다 관중을 기록하는 등 역대 관중 ‘톱 10’ 중 9번을 기록할 정도로 인기를 독차지하고 있다. - 프로 스포츠의 성공은 인구밀도가 중요하다. - 그래서 요즘 서울의 큰 시장을 감안해 기존 1부 팀을 입성시키자는 얘기도 나온다.	- 서울은 2010년 5월 5일 성남과의 홈경기에서 6만747명의 역대 K리그 최다 관중을 기록하는 등 역대 관중 ‘톱 10’ 중 9번을 기록할 정도로 인기를 독차지하고 있다. - 팬을 확보하지 못해 흥행이 되지 않으면 프로 스포츠의 존재 가치는 떨어지므로 프로 스포츠의 성공은 인구밀도가 중요하다. - 그래서 요즘 서울의 큰 시장을 감안해 지역 지지 기반이 약한 기존 1부 팀을 입성시켜 잠실종합운동장을 사용하게 해 강남 팬들을 확보하자는 이야기가 나오고 있다.	원문에서는 최근 프로축구 서울팀의 인기는 경기력과 함께 관중이 많은 서울이 연고인 것도 이유이다, 프로 스포츠의 성공은 인구밀도가 중요하다, 따라서 서울의 시장성을 감안하여 기존 1부 팀을 서울로 연고를 바꾸자는 내용이 중심이다. 그런데 첫 번째 주제문은 서울팀의 최근 경기력만을 언급하고 있어 원문의 핵심 내용을 포괄하지 못한다. 특히 두 번째, 세 번째 주제문이 서울 연고의 중요성에 대해 다루고 있다는 점에서도 첫 번째 주석된 주제문과 연결성이 높지 않은 것으로 판단할 수 있다. 또한 첫 번째 주제문이 적절하게 주석되지 못하여 요약문 역시 원문의 핵심 내용을 제대로 반영하지 못하고 있다.

표 21 CASE 5의 구체적 양상

CASE 1과 CASE 5가 전체의 76%를 차지하는 것으로 볼 때, 주제문의 주석 결과가 요약문의 질에 미치는 영향이 매우 높다는 것을 알 수 있다. 따라서 최종 요약문의 질을 높이기 위해서는 주제문 주석 과정이 보다 정밀하게 수행되어야 한다. 특히 첫 번째 주제문 주석이 적절하게 수행되어야 할 것으로 보인다. 위의 지침에서 언급한 것처럼 이전의 주제문이 다음 주제문 주석에 일정한 영향을 미치기 때문이다. 따라서 주제문 주석 과정이 다른 문장들을 최대한 포괄하는 문장들을 중심으로 수행될 수 있도록 관련 지침이 보다 명확하고 면밀하게 구성되는 것은 물론, 필요한 경우 수행자(작업자)를 대상으로 일정 수준의 학습이 필요한 것으로 판단된다.

제 3 장

요약 말뭉치 구축 지침

1. 요약 말뭉치 구축 대상 기사 선정

1.1. 신문기사 선정

▶ 객관적인 정보 전달을 위하여 작성된 기사

신문기사의 주제는 ‘경제, 과학, 국제, 문화, 사람들, 사회, 스포츠, 정치, 지역, 기획, 시설, 오피니언’ 등으로 분류할 수 있다. 이 분류는 국립국어원의 2018년 국어 말뭉치 연구 및 구축 사업에서 구축된 신문기사 말뭉치의 분류를 따른다. 이중 객관적인 정보만을 담고 있는 기사들을 자동 요약을 위한 말뭉치의 대상으로 선정한다. 이에 따라 ‘기획, 사설, 오피니언’은 말뭉치 구축 대상에서 제외한다. 실제 분류가 경제, 과학, 국제, 문화, 사람들, 사회, 스포츠, 정치, 지역, 기획, 시설, 오피니언’에 해당하는 기사더라도, 특정인의 시각으로 쓰인 논설 또는 칼럼에 해당할 경우 구축 대상에서 제외한다. 이러한 제외 기사의 발굴은 작업자의 이슈 보고 및 검수 과정을 통해 이루어진다. 제외 기사 목록에서 논설 또는 칼럼, 기획 기사에 해당하는 키워드를 발굴하고 목록화하여 전체 작업 기사에서 최종적으로 제외될 기사를 확정한다. 검토 대상 키워드는 “데스크 진단”, “문화好통”, “기자수첩”, “삼시세평”, “탐사기획”, “기자의 눈”, “커버스토리” 등 실제 칼럼 명 및 실제 인명으로 구성되었다.

▶ 적정 분량의 기사

분량이 너무 짧아서 요약이 불필요하거나 반대로 너무 길어서 세 문장 이내로 요약을 할 수 없는 기사는 말뭉치 구축 대상에서 제외한다. 기사문에서 한 문장은 평균적으로 14개의 어절 정도로 이루어지기 때문에, 본문 기준 200개 어절 이상 600개 어절 미만의 기사만을 말뭉치 구축 대상으로 선정한다. 주제문과 요약문은 일반적인 문장으로 구성하는 것을 원칙으로 하고 있기 때문에 명사구로 종결되는 부주제문은 본문에서 제외한다.

1.2. 요약 난이도를 높이는 기사문 유형

위에서 제시한 도메인 분류와 관계없이, 기사의 문장 형식 또는 목적이 일반적인 기사와 다른 것들이 있다. 특정한 제품이나 행사를 홍보하기 위한 목적을 가진 기사, 인터뷰이의 전체 발화를 그대로 옮겨놓거나 질의응답 형식으로 정리한 기사, 3개 이상

의 정보가 열거된 기사 등이 이에 속하며, 이러한 하위분류를 서브 클래스라고 지칭하였다. 서브 클래스를 가지는 기사들은 주제문을 주석하거나 요약문을 생성할 때 일반적인 기사와는 다른 별도의 처리가 필요하다.

1.2.1. 서브 클래스 분류 및 예시

본 과제에서 분류한 서브 클래스와 각 서브 클래스의 정의는 아래와 같다.

분류	정의
인터뷰	질문과 답변으로 되어 있는 기사 구성
리뷰	문화 행사, 전시, 서평, 방송, 시승기 등 리뷰 및 안내
열거	유사한 성격의 정보가 3개 이상 나열된 기사
광고	네이티브 광고로 추정되는 기사
부가정보제시	일반적인 형식의 기사 뒤에 개념 설명 기사, 연혁 등이 부가적으로 제시되는 기사
전언	기자간담회, 연설 등 긴 발화를 전체적으로 옮긴 기사

표 22 서브 클래스 분류 및 정의

서브 클래스는 기사의 특징에 따라서 6개로 분류되었다. 위 서브 클래스들은 한 기사에 여러 개가 해당할 수 있는데, 예를 들어 동일한 질문에 대한 여러 사람과의 인터뷰를 다루는 기사는 ‘인터뷰’와 ‘열거’의 두 가지 서브 클래스를 할당 받는다.

기사 원문
“한국식 거리응원 도입, 남아공 화합 축제로” 축구는 스포츠 이상이다. 국민을 하나로 묶고 정치가 해결할 수 없는 일을 해낸다. 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵도 축구 이상의 큰 의미를 담고 있다. 대니 조든 남아공월드컵조직위원회(SALOC) 사무총장(57)은 “내년 월드컵은 남아공을 하나로 묶어 인종 차별의 잔재를 없앨 것이며 세계 평화에도 기여할 것”이라고 말했다. SALOC의 최고경영자로 월드컵의 모든 것을 관장하는 그를 29일 남아공 요하네스버그 샌턴의 미켈란젤로호텔에서 만났다. 그동안 일부 국내 언론과 짧은 소감 인터뷰는 있었지만 장시간에 걸쳐 월드컵 청사진을 제시한 것은 이번이 처음이다. —1년 남았는데 월드컵 준비는 잘되고 있는가. “한국의 본선 진출을 축하한다. 아직도 2002년 한일 월드컵 기억이 생생하다. 정말 환상적이었다. 수백만 한국 팬들이 붉은색 옷을 입고 경기장 밖 광장에 모여 응원한 것은 우리도 본받고 싶다. 내년 남아공에서도 2002년 한국의 응원 방식을 도입해 국민들을 하나로 뭉치게 할 것이다. 이번 2009 컨페더레이션스컵 때 16경기를 치렀는데 전혀 문제가 없었다. 현재 경기장은 제대로 건설되고 있다. 5개는 이미 완성됐다. 내년 초까진 10개 모두 완성될 것이다.” —남아공이 월드컵 개최를 통해 얻는 효과는 무엇인가.

“먼저 각종 인프라 구축이 남아공을 발전시킬 것이다. 공항을 증개축하고 길을 넓히고 25개 고급 호텔을 새로 짓고 있다. 둘째는 관광객 유치다. 현재 950만 명이 매년 남아공을 찾고 있다. 내년엔 1050만 명 이상의 관광객 유치가 목표다. 셋째는 국가 이미지 제고다. 월드컵을 통해 지구촌 사람들이 남아공을 다시 보게 될 것이다. 넷째는 흑과 백이 둘이 아니라는 것을 지구촌에 보여줄 것이다. 이번 컨페드컵 때 경기장에는 흑인과 백인이 모두 스탠드를 채우고 ‘바파나 바파나(Bafana Bafana · 줄루어로 소년들, 남아공 축구 대표팀을 의미)’를 응원했다.”

—남북한이 본선에 동반 진출했다.

“정말 대단하다. 남한과 북한은 월드컵 본선 무대에서 처음 만나게 됐다. 그 역사가 남아공 월드컵에서 이뤄진 것이 너무 기쁘다.”

... 후략

표 23 인터뷰 기사 예시

위의 표 23은 서브 클래스가 인터뷰인 기사의 예시이다. 기사의 도입부는 일반적인 기사들과 크게 차이나지 않지만 그 이후부터는 ‘-’로 표시되는 기자의 질문과 큰따옴표로 표시되는 인터뷰이의 대답으로 기사가 구성되어 있다. 이러한 인터뷰 기사들은 일반적인 기사에 비하여 정보가 함축적으로 전달되지 않으며 기사의 길이가 길기 때문에 주제문을 주석하기에 어려운 측면이 있다.

기사 원문

극단 신기루만화경의 ‘설공찬전’

1997년 한글소설본이 발견돼 ‘홍길동전’을 밀어내고 최초의 한글소설 자리를 꿰찬 ‘설공찬전’은 조선 중종조에 이미 금서(禁書)가 된 작품이다. 명분은 혹세무민하는 혼백의 세계를 다뤘다는 것이었다. 저승세계를 소개하면서 현실 권력을 매섭게 풍자한 대목이 연산군을 몰아내고 권력을 잡은 중종시대 권력층의 역린을 건드렸기 때문이다.
극단 ‘신기루만화경’의 연극 ‘설공찬전’(이해제 작, 연출)은 이 고전소설에서 저승세계에 대한 소개는 빼고 권력의 생리를 비판한 내용을 전면에 부각했다. 요절한 수재 설공찬(황도연)은 죽은 지 3년 뒤 아버지 설충란(임진순)에게 못다 한 효도를 하려고 망나니 사촌동생 설공침(정재성)의 몸에 강림한다. 여기서 작품은 공침의 몸을 빌려 뮤지컬 ‘지킬 앤드 하이드’식의 선악 대결을 한국적 해학으로 녹여낸다.
정재성씨는 점잖은 공찬과 패악스러운 공침이 한 몸을 두고 벌이는 입씨름과 몸씨름을 능청스럽게 풀어낸다. 당대 실세인 정익로(이장원) 대감을 구워삶아 아들에게 관직의 길을 열어주려는 공침의 아비 설충수(최재섭)는 죽은 공찬이 아들의 몸에 들어왔다는 사실을 알고도 이를 묵인한다. 공침의 몸을 빌린 공찬이 정 대감 앞에서 감춰뒀던 자신의 경륜을 펼치려는 순간, 정 대감은 다음과 같은 질문으로 그의 입을 막는다.

“세상엔 많은 문답이 있다. …그 물음이 어디에서 흐르느냐에 따라 그 대답의 방법이 엄연히 달라지는 것. 그것이 세상 살아가는 법이다. 자, 중천에 해가 떠 있다. 내 눈엔 저 해가 네 개의 모가 있는 바둑판으로 보이는데 자네의 눈엔 어떻게 보이는가.”

한때 세상을 내려다보긴 같지만 스스로 몸을 드러내는 해와 몸을 감추는 달은 서로 다른 족속이라고 일갈했던 공찬은 이 문답을 통해 권력의 본질을 깨닫고 이 몸 저 몸으로 옮겨 다니며 한바탕 놀이를 펼친다. 작품은 정권교체기 ‘영혼 없는 공무원들’의 행태를 수없이 목도하는 요즘의 현실에 따끔

한 일침을 놓는다. 중국어 발음으로 한자성어를 남발하는 정 대감은 어색한 영어 발음으로 세간의 비웃음을 사던 지식인들을 연상시킨다. 요즘 세태에 대해 직접적이면서도 매서운 비판의 수위를 높인다면 더욱 호응이 뜨거울 작품이다. 서울 대학로 정보소극장에서 2월 8일까지. 2만 원. 02-764-7462

표 24 리뷰 기사 예시

위의 표 24는 서브 클래스가 리뷰인 기사의 예시이다. ‘설공찬전’이라는 한 편의 연극에 대한 감상이 전체 기사의 주된 내용을 이루고 있다. 연극의 줄거리, 주연 배우, 대사에 대한 정보가 담겨 있으며 기사의 후반부에는 연극을 볼 수 있는 장소와 가격대, 문의를 위한 전화번호까지 표기되어 있다. 리뷰는 다른 기사들과 달리 객관적인 내용이 아닌 작성자의 주관적인 감상을 드러내고 있으며, 기사 후반부에 나타나는 광고와 같은 측면은 기사의 객관성을 더욱 저하시킨다. 따라서 이러한 기사들은 객관적인 정보만을 전달하는 일반적인 기사들과는 따로 분류하여 요약을 하여야 한다.

기사 원문
“안되는 거 빼고 다 된다” 홈케어 서비스 ‘후끈’ 매트리스·후드·배관 등 집안 관리 한샘, 에어컨 청소 추가 ‘승부수’ 전자랜드·카카오 등 잇따라 가세 홈케어 업체들의 ‘안방 쟁탈전’이 달아오르고 있다. 홈케어 시장 진출 기업들이 줄을 잇고, 서비스 종류도 침대 매트리스나 에어컨 등 가정용품 청소에서 수도와 난방 배관 청소, 새집증후군 해결에 이르기까지 다양해지고 있다. 홈케어 시장 선두주자인 코웨이는 이 분야의 1분기 매출액이 전년 동기 대비 113% 증가했다고 16일 밝혔다. 매트리스 렌털·관리가 주축인 코웨이 홈케어 사업의 이런 성장 속도는 지난해 분기 평균 성장률(21%)의 5배가 넘는다. 한샘은 주방의 후드와 세탁기·매트리스 청소를 전문적으로 해주는 ‘한샘 홈케어 서비스’에 최근 에어컨 서비스를 추가했다. 여름을 앞두고 전문 기사들이 방문해 송풍팬·냉각핀·필터를 세척하고, 고압 스팀·유브이(UV) 자외선·피톤치드 살균 서비스를 해준다. 한샘 김영태 상무는 “매트리스나 주방가구 등 제품을 판매하는 데서 끝나지 않고 소비자들이 제품을 사용하면서 겪는 어려움을 해결해주는 관리 서비스까지 제공함으로써 차별화된 경쟁력을 가질 수 있다”고 말했다. 건강생활용품 기업인 한경희생활과학도 최근 홈클리닝 서비스 종류를 대폭 늘렸다. 기존의 에어컨·세탁기·냉장고·비데 등 11개 품목 살균 클리닝과 이사·입주·대청소 홈클리닝에 곰팡이 제거, 수도·보일러 배관 청소, 바닥재 제거와 코팅 시공, 방충망·외풍 차단 시공 서비스를 추가했다. 서비스 지역도 서울·수도권에서 전국으로 확대했다. 이 회사 홈케어 서비스는 지(G)마켓을 통해서도 이용할 수 있다. 지난 1월 홈클리닝 브랜드 ‘마이크로 케어’로 침대 매트리스 렌탈·관리 서비스를 시작한 청호나이스는 올해 안에 에어컨·세탁기·후드 등으로 서비스 품목을 늘릴 예정이다. 또 가정과 사무실의 가전제품을 통합 관리해주는 서비스를 제공한다는 계획도 갖고 있다. 전자랜드도 가전제품 배송과 설치 노하우를 바탕으로 올 1월부터 홈클리닝 서비스인 ‘전자랜드 클린킹’ 사업을 시작했다. 에어컨·냉장고·공기청정기 세척·살균에서 새집증후군 해결, 유리창·배관 청소, 일반 청소까지 해준다. 카카오톡이란 막강한 플랫폼을 보유한 카카오는 하반기 중 가사도우미를 중개하는 홈클리닝 서비스

를 출시할 예정이다. 한경희생활과학 김윤채 팀장은 “미세먼지와 집먼지 진드기, 새집증후군 등으로 인해 집안을 청결하게 유지하려는 욕구가 높다. 하지만 맞벌이 등으로 청소할 시간은 부족해 홈케어 서비스를 이용하는 가정이 늘고 있다”고 말했다.

표 25 열거 기사 예시

위의 표 25는 서브 클래스가 열거인 기사의 예시이다. ‘홈케어 서비스’라는 하나의 큰 주제를 가지고 여러 기업들의 현황에 대해서 다루고 있다. 일반적인 기사들과 마찬가지로 하나의 큰 주제를 가지고 있으며 객관적인 정보만을 다루고 있지만, 비교적 길지 않은 두세 문장의 정보들이 나열되어 있기 때문에 작업자가 주제문을 선정하거나 요약문을 생성하기에 어려운 측면이 있다.

기사 원문
[아파트 미리보기]김포 한강신도시 ‘우미 린’ 《‘김포 한강신도시에서도 분양 훈풍이 이어질까.’ 우미건설이 12일 경기 김포 한강신도시 ‘우미 린’의 모델하우스를 개관하자 부동산 업계에서는 인천 청라지구와 송도지구에서 불었던 분양 훈풍이 이곳에도 불지 눈여겨보고 있다. 최근 ‘경기가 바닥을 쳤다’ ‘수도권 분양시장이 살아났다’는 주장이 꾸준히 제기되면서 분양시장의 회복세를 가늠하는 잣대로 우미 린의 분양률에 관심이 집중되고 있는 것이다.》 ○3.3㎡당 분양가 1041만원 우미 린은 가격 경쟁력이 높다는 평가가 많다. 우미건설에 따르면 우미 린의 3.3m² (1평)당 분양가는 1041만 원 선으로 지난해 분양됐던 ‘우남 퍼스트빌’보다 약 30만 원 싸다. 인근 지역인 김포시 고촌면과 걸포동, 고양시 일산 지역에 비해서는 3.3m² 당 분양가가 200만~500만 원 저렴하다. 이 아파트는 모두 4개의 크기로 구성돼 있다. 전용면적 △105m² A 331채 △105m² B 331채 △128m² 278채 △130m² 118채 등 총 1058채. 지하 2층에 지상 20~26층 총 14개 동이다. 특히 128m² 와 130m² 는 중도금의 60%를 전액 무이자로 빌려주는 게 특징으로 꼽힌다. 우미건설 관계자는 “128m² 와 130m² 는 중도금의 60%에 무이자를, 105m² 는 중도금의 30%에 무이자를 각각 적용했다”며 “무이자 혜택을 감안하면 128, 130m² 와 105m² 의 3.3m² 당 분양가 차이가 37만 원에서 14만 원 정도로 줄어든다”고 말했다. 내년 2월 11일까지 계약하면 양도소득세가 5년간 100% 면제되며 전매제한 기한은 1년이다. 단지가 11만 m² 크기의 호수공원에서 걸어서 10분 거리이고 인근에 경전철역이 신설되며 단지 바로 옆에 초등학교와 중학교가 들어설 예정이라는 것도 이 아파트의 장점이다. ○벽 떼내 구조 바꿀 수 있어 우미 린의 전체적인 인테리어는 은은한 그레이 톤을 적용해 모던하고 담백한 이미지를 연출하는 데 초점을 맞췄다. 특히 실용성을 강조하기 위해 수납공간에 신경을 많이 쓴 모습이 보였다. 수납공간이 넉넉하고 효과적으로 배치돼 있기 때문이다. 주방, 현관, 안방의 멀티룸 등은 전체 면적의 약 25%가 수납공간으로 이뤄져 벽 전체가 크고 작은 수납공간으로 채워졌다는 느낌이 들 정도였다. 우미건설 관계자는 “오랫동안 새 집 같은 분위기를 유지하려면 거주 기간이 늘어날수록 쌓이는 각종 잡동사니를 최대한 넣어둘 수 있는 공간이 필요하다는 주부들의 의견을 반영한 결과”라고 설명했다. 입주자들이 자유롭게 실내 구조를 바꿀 수 있도록 한 것도 특징이다. 거실과 방, 방과 방 사이의 벽 중 상당수를 떼어낼 수 있는 무량판 구조를 적용했다. 바닥도 폴리싱 타일과 온돌마루 중 선택이 가능하다. 주방에는 주부들이 일을 하면서 편안하게 TV를 시청할 수 있도록 10인치 크기의 고화질(HD)

주방액정TV를 무료로 설치했다.
김포 한강신도시는 김포고속화도로가 곧 개통될 예정이고 김포공항까지 이어지는 경전철도 2012년 완공 예정이어서 교통 여건이 계속 좋아질 것으로 보인다. 우미 린의 청약은 16일 특별공급을 시작으로 17일부터 19일까지 실시된다. 당첨자 발표는 25일이며 다음 달 1일부터 3일까지 계약을 한다. 입주는 2011년 10월 예정이다.
수도권 분양시장이 살아났다는 주장이 나오는 가운데 우미건설이 12일 경기 김포 한강신도시에서 '우미 린' 모델하우스를 개관하자 부동산 업계에서는 이곳에서도 분양 훈풍이 불지 눈여겨보고 있다. 호수공원에서 걸어서 10분 거리이고 인근에 경전철역, 초등학교, 중학교가 들어설 예정인 이 아파트는 내년 2월 11일까지 계약하면 양도소득세가 5년간 100% 면제되며 16일부터 청약이 실시된다.
김포 한강신도시는 김포고속화도로가 곧 개통되고 김포공항까지 이어지는 경전철도 2012년 완공 예정이어서 교통 여건이 계속 좋아질 것으로 보인다.

표 26 광고 기사 예시

위의 표 26은 서브 클래스가 광고인 기사의 예시이다. 곧 완공되는 아파트의 분양 시기, 입주 일정, 가격, 평수, 장점 등에 대한 내용이 주를 이루고 있다. 이처럼 광고 서브 클래스 기사들은 형식적인 측면에서는 일반적인 기사들과 큰 차이가 없으나 한 제품을 소비자들에게 홍보하기 위한 독특한 목적을 가지고 있으며, 따라서 해당 기사에 나타난 내용들이 얼마나 객관적인지에 대한 평가를 내리기가 어렵다.

기사 원문

[프로축구] "포항은 진화중… 한국의 바르사 꿈꾼다"
■ K리그 1위 질주, 포항 황선홍 감독

감독과 팀 사이에도 궁합이 있다. 축구인들은 요즘 잘나가는 포항 스틸러스와 '황새' 황선홍 감독(43)을 두고 찰떡궁합이라고 말한다. 포항은 선수단 구성이 크게 달라진 것이 없는데 지난해 K리그 9위에서 올 시즌 1위(5승 3무)의 고공비행을 하고 있다. 황 감독의 역할이 컸다는 평가다.

○ 애정
1991년 건국대를 졸업하고 독일 2부 리그에서 활약했던 황 감독은 1993년 포항에 둥지를 틀었다. 국내 첫 프로팀이 포항이었다. 지난해 말 포항 사령탑으로 컴백한 황 감독은 "포항은 내 고향 같은 팀"이라고 말했다. 2008년 프로 첫 사령탑의 인연을 맺은 부산의 재계약 제의를 뿌리치고 포항을 선택한 것은 이런 개인적인 사정과 1973년 팀을 창단해 40년 가까이 한국 축구 발전을 위해 노력한 포항에 대한 각별한 애정 때문이었다.

○ 자부심
황 감독은 취임 후 포항이 아시아 챔피언스리그 챔피언과 K리그 2위를 한 2009년과 9위로 곤두박질친 지난해를 분석한 결과 동기 결여가 문제라는 결론을 내렸다. 지난해에는 동아시아선수권 대표팀 차출로 인한 조직력 와해와 브라질 출신 레모스 감독의 중도 사퇴 등 우여곡절이 있었다. 황 감독은 선수들에게 포항의 역사와 자부심에 대해 설명하고 경기 전 라커룸 보드에 '우리는 포항이다'는 문구를 적어두면서 선수들의 마음을 움직였다. '승리에 집착하기보다는 5분 더 뛰고, 깨끗한 매너를 지킨다'는 등 2009년 만들어 K리그에 큰 반향을 일으켰던 '스틸러스 웨이'도 부활시켰다. 포항이 변하자 팬들도 반응했다. K리그 평균 홈 관중 1만4757명으로 지난해(1만1174명)에 비해 32%가 증가했다.

○ 소통
요즘 포항은 프런트와 감독, 선수, 팬이 사위일체가 됐다. 이기는 경기보다는 팬들이 즐거운 경기를

해야 한다는 김태만 사장의 뜻에 황 감독도 전적으로 따르고 있다. 이런 구단의 모토를 황 감독이 선수들에게 잘 설명해 플레이를 바꿨다. 프런트는 다시 활기를 띤 지역 팬들을 흥겹게 하기 위해 매 홈 경기에 자동차를 경품으로 내놓는 등 다양한 이벤트를 마련했다. 팬들은 스탠드를 뜨겁게 달구고 있다.

○ 공부
황 감독은 요즘 비디오를 분석하며 밤을 새우는 재미에 빠져 있다. 소속팀과 상대팀은 물론이고 잉글랜드와 스페인 등 해외 경기도 분석해 응용하고 있다. 세계 최고의 팀으로 불리는 스페인 바르셀로나같이 패싱플레이로 효율적인 축구를 하는 팀을 만드는 게 목표다. 황 감독은 2003년과 2007년 잉글랜드에 축구 유학을 하는 등 끊임없이 노력하고 있다.
8일 친정팀 부산과 첫 대결을 하는 황 감독은 “별다른 감정은 없다. 우리 플레이를 잘하느냐가 중요하다. 우리는 더 발전해야 하고 계속 진화하는 중”이라고 말했다.

■ 황선홍 감독은?
△생년월일: 1968년 7월 14일
△출신교: 서울 용문중-용문고-건국대
△선수 시절 소속팀: 독일 레버쿠젠Ⅱ(아마추어・1991년), 독일 부퍼탈SV(1992년), 포항(1993～1998년), 세레소 오사카(1998～1999년), 수원(2000년), 가시와 레이솔(2000～2002년), 전남(2002～2003년), 프로 통산 134경기 73득점
△국가대표: 1988년 12월 아시안컵 일본전 데뷔, 2002년 한일 월드컵까지 통산 103경기 50골
△지도자 경력: 전남 코치(2003～2006년), 부산 감독(2008～2010년)

표 27 부가정보제시 기사 예시

위의 표 27은 서브 클래스가 부가정보제시인 기사의 예시이다. 기사의 형식과 내용은 일반적인 기사들과 유사하지만 기사의 후반부에 특정 인물에 대한 정보들이 제시되고 있다. 이처럼 부가정보제시 서브 클래스 기사들은 기사의 내용과 관련된 정보들을 따로 제시하고 있는 기사들이다.

2. 한국어 쓰기 특성을 고려한 요약문 생성

이번 사업에서 주목해야 할 것은 영미나 유럽의 경우 단락 쓰기는 topic sentence (주제문장) + supporting sentence(뒷받침 문장)의 형식이 굳어진 반면, 한국어는 아직 이러한 형식이 일반화되어 있지 않다는 점이었다. 최근에는 중등·고등교육의 쓰기 교육에서 이러한 형식을 교육하고 있으나 아직 일반화되었다고 보기 어렵다. 이러한 한국어 쓰기의 특성은 요약문 생성 자체가 기술적으로 어려울 수 있다는 점과 연결된다는 점에서 주의를 요한다. 구체적으로는 요약문 생성 기준 중 주제문의 선별이 영미나 유럽의 텍스트에 비해 상대적으로 용이하지 않을 수 있다. 또한 원문의 길이가 짧을 경우 상대적으로 요약용 말뭉치 추출이 용이하지 않을 수 있으므로 이에 대한 충분한 기술적 고려가 필요하였다. 이러한 측면은 주제문을 생성하고 중심 내용을 재구성하는 '구성' 단계는 물론, 최종 요약문 생성에도 동일하게 적용되었다. 즉, 상위어를 중심으로 구성된 주제 문장과 이를 뒷받침하는 문장으로 최종 요약문을 생성함으로써 독자에게 원문의 핵심적인 내용을 정확하게 전달할 수 있는 요약문의 질을 높일 수 있도록 구성 단계에서 요약문 생성의 세부 규칙을 보다 정교하게 설정해야 할 필요가 있다. 또한 이러한 과정은 선택과 삭제에 의해 추출된 요약 말뭉치를 기반으로 상위어 생성이 원활하게 이루어져야만 효과를 거둘 수 있으므로 아울러 상위어 생성에 대한 기술적 고려가 필수적이다.

주석된 주제문	올해 프로야구에선 2명의 '일본 U턴파'가 뛴다. 이혜천은 야쿠르트와의 2년 계약이 끝난 뒤 방출됐고 이범호는 계약기간이 남았지만 복귀를 택했다. 일본 U턴파 환대에 대한 우려의 목소리도 크다.
요약문 A	두산 이혜천(32)와 KIA 이범호(30)는 '일본 U턴파'이다. 이해천은 일본에 가기 전에 소속된 구단인 두산으로 돌아왔고, 이범호는 친정인 한화 대신 KIA로 돌아왔다. 국내 구단 소속으로 일본 진출 후 돌아온 선수들이 대부분 좋은 성적을 거두기는 했으나 이를 환대하는 데에 대한 우려의 목소리도 있다.
요약문 B	올해 프로야구에서 '일본 U턴파'는 두산 이혜천(32)과 KIA 이범호(30)이다. 이해천은 일본에 가기 전 소속된 구단인 두산으로, 이범호는 KIA로 돌아왔다. 국내 구단 소속으로 일본 진출 후 돌아온 선수들이 대부분 좋은 성적을 거두기는 했으나 이들을 무조건 환대하는 데에 대한 우려의 목소리도 있다.

표 28 주제문 및 요약문 예시

요약문 A에는 '프로야구'라는 상위어가 빠져 있다. 원 기사에서 전체 내용을 포괄하는 상위어는 '프로야구'이다. 기사의 모든 내용은 프로야구라는 상위어에 입각할 때 의미를 갖기 때문에 요약문의 질을 높이기 위해서는 요약문 B와 같이 상위어가 반드시

포함되어야 한다. 또한 요약문 A의 두 번째 문장에는 '돌아오다'가 두 번 중복되어 삭제가 필요하다. 세 번째 문장에서 환대에 대한 우려의 목소리는 원 기사의 어조를 볼 때 '무조건 환대'에 대한 것으로 이해된다. 그러므로 요약문 B와 같이 '무조건 환대'와 같이 원 기사의 의도를 보다 정확히 전달할 수 있는 기술적 고려가 필요하다.

위의 언급의 연장선상에서 단락별로 첫 문장이 영미 유럽과 달리 세부 내용을 다루는 경우도 많아 첫 문장을 중심으로 요약이 이루어지는 경우 기술적인 주의가 필요하다. 원문의 단락별로 첫 번째 문장이 단락의 내용을 포괄하는 역할을 하는 경우가 일반적이기는 하지만 일부 단락은 부연, 상술, 예시를 담당하는 경우가 있으므로 이 경우에는 첫 문장이라도 최종 요약문 생성에 그대로 활용하는 경우 문제가 발생할 가능성이 있다.

원문	① 이선균과 정려원이 JTBC 새 월화드라마 '검사내전'의 주연으로 캐스팅을 확정지었다. ② 올해 말 방송되는 '검사내전'(연출 이태곤/크리에이터 박연선/극본 이현 서자연)은 생활형 검사들의 오피스드라마로, 현직 검사 김웅이 저술한 동명의 베스트셀러가 원작이다. 미디어 속 화려한 법조인이 아닌 지방도시 진영에서 하루하루 살아가는 평범한 '직장인 검사'들의 이야기를 그린다. ③ 최근 영화 '기생충'을 통해 믿고 보는 명품 배우임을 다시금 입증한 이선균이 이번 '검사내전'에서 생활밀착형 검사 이선웅 역으로 출연을 확정지었다. 이선웅은 진영지청 형사 2부 소속으로 선한 인상에 출세욕 없이 느긋해 보이지만 보기와 달리 만만치 않은 '한 방'을 지닌 인물. 이선균 특유의 따뜻하고 묵직한 연기 톤이 배역과 높은 싱크로율을 자랑하는 만큼 이선균 표 이선웅의 모습이 벌써부터 기대감을 불러일으킨다. ④ 중앙지검에서 승승장구하다 하루아침에 머나먼 진영지청으로 발령받은 엘리트검사 차명주 역으로는 배우 정려원이 캐스팅됐다. '기름진 멜로', '마녀의 법정' 등 다양한 작품에서 장르를 넘나드는 팔색조 매력을 선보인 정려원은 이번 '검사내전'에서 빈틈없이 완벽한 검사 차명주로 변신, 걸 크러시 면모를 뽐낼 예정이다. ⑤ 이로써 '검사내전'은 탄탄한 베스트셀러 원작을 바탕으로 JTBC '청춘시대' 시리즈로 섬세한 연출력과 필력을 인정받은 이태곤 감독-크리에이터 박연선 작가의 재결합, 그리고 주연배우 이선균, 정려원의 합류까지 삼박자를 완성하며 올 하반기 최고의 기대작으로 떠올랐다. 2019년 말 방송 예정.

표 29 기사 예시

위의 예시는 연예 기사이다. 이 기사는 모두 4개의 단락으로 구성되어 있는데 첫 번째 단락은 1개 문장으로 구성되었다. '단락의 첫 번째 문장이 원문의 전체적인 내용을 포괄한다'는 관점에서 보면 기사의 전체 내용을 포괄하는 핵심 문장은 ①이 된다. ②-⑤문장 중에서 ②, ⑤는 원문에서 중요한 내용을 포함하고 있으며 최종 요약문을 생성하는데 주된 역할을 한다고 간주할 수 있지만 ③, ④의 경우 ①의 문장을 **상술**하는 기능을 담당하고 있다. 따라서 ③, ④의 문장을 최종 요약문 생성에 활용하는 경우 요약문이 지나치게 길어질 수 있으므로 ③, ④의 문장은 최종 요약문 생성에서 배제하

는 기술적 고려가 필요하다.

한국어 문장은 연결어 등의 활용도가 높아 문장이 길어지는 경우가 적지 않다. 문장이 길어지는 경우 전체적인 내용 파악이 어려워지는 등의 가독성이 떨어질 수 있다. 따라서 요약 과정 및 최종 요약문 산출에서 이에 대한 기술적 고려가 필요하다. 이번 사업에서 요약문의 1차 검수 결과, 일부 요약문의 경우 문장이 너무 길어서 내용 파악에 장애가 될 수 있는 경우들이 발견되었다. 이러한 현상의 원인은 크게 두 가지로 추측할 수 있다. 첫째, 원문의 긴 문장을 그대로 요약에 활용한 경우이다. 원문의 문장들이 정제되지 않고 장황하게 쓰인 경우가 많을 때 요약에 그대로 활용될 수 있다. 둘째, 작업자들이 원문의 주요 내용을 최대한 많이 넣기 위해 요약한 문장을 길게 생성하는 경우이다. 특히 이번 작업의 경우 최종 요약문을 3개 문장으로 한정하였으므로 작업자가 원문의 내용을 최대한 반영하려고 시도하는 경우 문장이 길어질 수 있다. 요약문의 각 문장이 너무 길어지면 이후 데이터 처리 등의 작업에서 기술적 문제가 발생할 수도 있다. 이번 사업에서는 이러한 문제를 최소화하기 위해 아래와 같이 사례를 제시, 활용하였다.

기 구축된 요약문	제안 요약문
‘천국의 계단’에서 최지우와 신현준이 함께 죽음을 맞고, ‘발리에서 생긴 일’의 조인성이 권총자살하는 결말은 자극적이지만 시청자들은 비현실적 상황을 계속 보면서 인과관계나 개연성을 따지기보다 오히려 묘한 판타지를 느낀다는 분석이 나온다.	두 드라마의 등장인물들의 죽음이나 결말이 매우 자극적이고 비현실적이다. 하지만 시청자들은 인과관계나 개연성을 따지지 않고 오히려 묘한 판타지를 느낀다는 것이다.

표 30 요약문 제안 예시

이와 함께 최종 요약문 생성에서 가독성을 높일 수 있는 명확한 방향 제시가 필요하다. 원문에서 주석한 주제문들을 그대로 결합하는 것이 아니라 요약문을 시간적 또는 논리적 순서에 따라 재구성하는데 필요한 구체적인 지침을 제공할 필요가 있다. 이번 사업에서 주제문들이 명시적으로 드러나는 원문을 요약하는 경우, 주석한 주제문들을 그대로 결합하는 경우들이 일부 발견되었다.

아래 주제문은 원문에서 주석한 것이다. 원문이 길지 않고 문장이 곧 단락인 경우이므로 주제문 주석은 비교적 용이하지만 이를 그대로 요약문으로 활용하면 가독성을 확보하기 어렵다. 주제문과 요약문을 비교해보면 요약문의 첫 문장은 주제문의 마지막 문장을 기반으로 구성된 것임을 알 수 있다. 근거는 주제문의 [3] 문장에서 ‘앞서’라는

단어가 있기 때문이다. 따라서 시간적 순서에 따르면 [3]의 문장이 최종 요약문의 맨 앞에 위치하는 것이 자연스럽다. 하지만 독자의 내용 이해도는 무의식중에 시간적 순서에 따라 사건을 배치하거나 논리적으로 정합적이 갖추어질 때 높아진다. 따라서 최종 요약문은 원문에서 추출된 주제문(장)의 순서에 구애받지 말고 요약문 자체가 논리적일 수 있도록 생성되어야 한다.

이번 사업에서는 이와 관련하여 별도의 지침을 제시하였고 검수를 거쳐 관련된 문제들의 발생을 최소화하기 위해 노력하였다. 이에 대해서도 보다 면밀한 기술적 검토와 관련 지침이 마련되어야 할 것으로 생각한다.

원문	배우 송혜교 측이 송중기와의 이혼 소식 후 확산한 관련 악성 댓글과 루머 유포자들을 일괄 고소했다. 25일 경찰에 따르면 송혜교 측은 이날 분당경찰서에 허위사실 적시에 따른 명예훼손 및 모욕에 대한 내용으로 다수를 상대로 한 고소장을 냈다. 송혜교 측은 이번 고소와 관련, 어떠한 선처나 합의 없이 강경하게 대응하겠다는 입장을 경찰에 피력한 것으로 알려졌다. 앞서 송혜교와 송중기는 결혼 1년 8개월 만에 이혼 조정에 나선 사실이 보도되며 다양한 악성 댓글과 루머에 시달려왔다. 송중기 소속사 역시 루머 유포에 대해 법적 대응하겠다는 입장을 밝힌 바 있다. 두 사람의 이혼 조정은 지난 22일 성립돼 두 사람은 결혼 1년 9개월 만에 완전히 남남이 됐다.
주석된 주제문	[1] 배우 송혜교 측이 송중기와의 이혼 소식 후 확산한 관련 악성 댓글과 루머 유포자들을 일괄 고소했다. [2] 송혜교 측은 이번 고소와 관련, 어떠한 선처나 합의 없이 강경하게 대응하겠다는 입장을 경찰에 피력한 것으로 알려졌다. [3] 앞서 송혜교와 송중기는 결혼 1년 8개월 만에 이혼 조정에 나선 사실이 보도되며 다양한 악성 댓글과 루머에 시달려왔다.
생성된 요약문	송혜교와 송중기가 결혼 1년 8개월 만에 이혼 조정에 나섰다. 이들은 이혼 소식이 전해진 후 다양한 악성 댓글과 루머에 시달렸다. 이에 송혜교 측은 악성 댓글과 루머 유포자들을 모두 고소했으며, 어떤 선처나 합의 없이 강경하게 대응하겠다는 입장을 밝혔다.

표 31 주제문 주석 및 요약문 생성 예시

2.1. 읽기, 쓰기와 ‘읽기-쓰기’의 차이

‘읽기-쓰기’란 새로운 지식을 획득하였음을 확인하고 표현하는 과정 자체이며 가장 일반적으로 활용되는 학습 활동이자 지식의 측정 방식이다. 따라서 ‘읽기-쓰기’는 지식의 습득과 활용에서 가장 중요한 기능이지만 실제 수행에는 많은 어려움이 따른다. 한

연구에서는 '읽기-쓰기'의 어려움을 다음과 같이 기술하고 있다. 본 연구의 범위는 다음과 같다.

(1) 텍스트를 읽고 쓰는 것을 배우기란 모국어를 사용하는 학생들에게도 완전히 습득하기 어려울 정도로 어려운 과제이다. 요약, 정보의 통합, 텍스트에 대한 비평, 학술논문의 작성 등과 같은 읽기-쓰기를 통합한 과제들을 해 내기 위해서는 많은 연습이 필요하다 (Grabe & Zhang, 2013:9).

전통적으로 읽기 영역에서 '읽기-쓰기'는 독자가 텍스트에서 내용을 선택, 평가, 활용하는 능력 향상에 필요한 학습방법으로 인식되었다(Trites & McGroarty, 2005). 그러나 읽기 능력과 '읽기-쓰기' 능력의 상관관계에 대한 연구들은 다양한 결과들을 산출하면서 현재까지는 뚜렷한 합의점을 찾지 못한 것으로 보인다. Watanabe(2001)는 학생들의 '읽기-쓰기' 수행과 읽기와 쓰기의 평가 점수 사이에 상관관계를 분석했는데 읽기 점수는 '읽기-쓰기' 수행 점수와 상관관계가 확인되지 않으나 쓰기 점수와는 상당히 높은 상관관계를 보인다는 점을 확인했다. Delaney(2008)의 연구 역시 읽기 점수와 '읽기-쓰기' 수행 점수에서 매우 낮은 수준의 상관관계만을 확인했다는 점에서 같은 맥락에 놓인다. 반면, Trites & McGroarty(2005)는 측정한 모든 읽기 해석 점수가 '읽기-쓰기'와 일정한 상관관계를 보인다고 주장했다.

쓰기와 '읽기-쓰기'의 관계에 대한 이론적 관심 1980년대 미국에서 쓰기 과정을 단계별로 구분하려는 시도에서 시작되었다. 소위 '과정 중심 글쓰기'에서 읽기와 쓰기의 통합을 쓰기 과정의 일부로 이해하면서 '읽기-쓰기'에 대한 관심이 높아진 것이다 (Silva, 1990:15). 읽기 능력과 '읽기-쓰기' 능력의 상관관계는 선행 연구들에서 입장이 엇갈리지만 쓰기 능력과는 비교적 일관되게 높은 상관관계를 확인할 수 있다 (Lewkowicz, 1994; Watanabe, 2001)

2.2. '읽기-쓰기'의 연구 동향

전통적으로 읽기와 쓰기는 주체의 사고를 언어로 표현하는 과정에서 서로 연결되어 있는 것으로 간주되었다. 그러나 실제로는 별개의 영역으로 인식되는 관습으로 인해 읽기와 쓰기는 각각 다른 학문적 배경에서 독자적으로 연구되었다. 읽기와 쓰기의 관계에 대해 주목하기 시작한 것은 1970년대 '학습의 인지 모형(cognitive model of learning)'이 등장하고 인지구성주의 관점에서 읽기와 쓰기를 연구하는 경향이 뚜렷해지면서부터였다. 인지구성주의 관점에서 쓰기와 읽기는 공통적으로 텍스트의 의미를 적

극적으로 구성(active construction of meaning)하는 행위로 이해되었다. 특히 읽기는 지각, 언어, 개념적 작용이 복합된 정신적 과정으로 이해되었다. 즉, 독자가 특정 텍스트를 이해하는 과정은 텍스트가 제공하는 정보와 이전부터 습득하고 있는 관련 정보들과의 관련성을 파악하고 둘 사이의 영향관계를 형성되는 것으로 파악한 것이다.

스키마 이론은 읽기와 쓰기 관계에 대한 연구에 자극을 가하면서 방향성 가설(directional hypothesis)과 비방향성 가설(non-directional hypothesis) 형성에 일정한 영향을 끼쳤다. 방향성 가설은 읽기와 쓰기가 구조적 요소를 공유하고 있으며 한쪽에서 습득한 인지구조를 다른 쪽에 적용할 수 있다고 보는 견해로서 읽기 학습을 통해 형성된 인지구조가 쓰기에 영향을 주며 쓰기 지식을 학습하는 것은 중요하지 않다는 읽기 중심 견해와 쓰기 학습이 독해 능력과 정보기억능력을 향상시킬 수 있다는 쓰기중심 견해로 분류된다. 비방향성 가설은 읽기나 쓰기 어느 한 쪽이 다른 쪽에 영향을 미치는 것이 아니라 어느 방향이든 인지구조의 전이가 가능하다는 점에서 방향성 가설과 다르다. 비방향성 가설은 텍스트를 인지하는 능력(읽기 능력)과 생성하는 능력(쓰기 능력) 모두 텍스트 구조에 대한 동일한 지식에 기반을 두고 있다고 가정한다. 따라서 비방향성 가설에서 읽기와 쓰기는 상호작용이 역동적으로 발생하는 관계로 파악한다. 즉, 서로의 인지구조가 상호작용을 통해 영향을 받고 정교화될 수 있다고 보는 것이다. 이러한 견해에서 볼 때 교실 수업에서 '쓰기'를 '읽기 전 활동'이나 '읽기 후 활동'으로 활용하는 것은 인지구조의 전이를 활성화시키려는 의도로 이해할 수 있다.

1980년대 들어서면서 방향성 가설과 비방향성 가설은 Rosenblatt의 교류이론(transactional theory)으로 연결된다. 교류는 '상호 구성적 상황에서 각 요소들이 서로 조건화하거나 조건화되는 관계'로 교류 이론은 기본적으로 양방향적 모형(bidirectional model)이라고 할 수 있다. 방향성·비방향적 가설은 어느 한 쪽의 인지구조가 다른 쪽 인지구조에 영향을 미치는 것이라면 교류이론은 양쪽 방향으로 모두 영향을 미친다는 점에서 구별된다. 교류이론에 따르면 읽기는 '독자의 마음과 텍스트의 언어 사이에 발생하는 교류 과정'으로 읽기를 통해 독자가 형성하는 글의 의미는 텍스트 정보와 텍스트를 작성한 필자가 전달하려는 메시지(구조) 그리고 독자의 독서 경험 및 이전에 습득한 지식들이 상호교류하는 과정에서 구성된 것으로 파악한다. 또한 쓰기는 필자의 지식과 필자의 개인적·문화적·사회적 환경 사이의 상호교류 과정이 된다.

교류이론에서는 읽기와 쓰기가 서로 조건으로 작용하는 상호조건화(mutually condition) 개념을 도입한다. 즉 읽기 과정에서는 쓰기의 인지구조, 쓰기에는 읽기의 인지구조가 조건화되면서 상호영향관계를 형성한다는 것이다. 이는 쓰기 과정에 관한 연구를 통해서도 일부 확인이 되었다. 필자가 글을 수정하는 과정에서 다시 읽기

(re-reading)는 현재 작성한 자신의 글을 읽으면서 자신의 의도가 제대로 전달되지 못하는 부분을 파악하는 행위로서 읽기와 쓰기가 역동적으로 상호영향관계를 형성하는 행위라는 점에서 읽기와 쓰기의 상호영향관계를 확인할 수 있다.

1970년대 이후 읽기와 쓰기 연구에 인지구성주의가 도입되면서 읽기와 쓰기가 각각 다른 영역이며 상호 간에 영향관계를 상정하지 않는 전통적 견해는 더 이상 설득력을 얻지 못하게 되었다. 또한 방향성·비방향성 가설과 교류이론으로 관련 연구가 발전하면서 읽기와 쓰기는 어느 한 쪽이 우위에서 다른 쪽에 영향을 미치는 고정적인 관계가 아니라 의미구성의 맥락에 따라 영향관계는 양방향으로 그리고 유동적으로 형성된다는 견해가 힘을 얻게 되었다.

80-90년대 연구는 읽기와 쓰기의 동시 작용에서 발생하는 상승효과에 초점을 맞춘 것으로 볼 수 있는데 대부분의 연구는 단일한 텍스트의 읽기와 쓰기 관계를 다루었다. 그러나 실제 읽기와 쓰기는 필자와 독자와 속한 사회 내의 문식적 환경과 맥락에 영향을 받는다는 사실에서 볼 때, 좀 더 다양한 맥락과 의미구성의 조건들을 고려한 연구가 진행되어야 한다. 이와 함께 특정한 맥락에서 읽기와 쓰기 관계는 어떤 방식으로 구체화되는지 그리고 읽기가 쓰기를, 또는 쓰기가 읽기를 간섭하는 시점과 지점에 대한 연구들이 폭넓게 진행될 필요가 있다.

외국 연구에 비추어 볼 때, 국내의 '읽기-쓰기' 연구는 매우 미진한 수준이다. 초등교육에서 교수학습 방법의 일부로 다루거나(송지언, 2015), 한국어교육 분야에서 한국어 학습자의 요약문 쓰기 양상 분석(이은경, 2012, 2013) 정도를 들 수 있다. 대학 수준의 의사소통 교육에서는 유사한 연구들이 수행되기는 했지만 본격적인 '읽기-쓰기' 연구는 거의 없는 실정이다.

2.3. '요약문 쓰기'의 연구 동향

요약문 쓰기는 학술적 영역에서 읽기와 쓰기가 동시에 관여하는 학습 활동이기 때문에 국외에서는 이에 관한 다양한 연구가 진행되어 왔다. 특히 요약문 쓰기가 진행되는 과정은 상대적으로 상세히 밝혀진 바 있다. Ferris & Hedgcock(2005:106)에 의하면 요약문 쓰기는 다음의 다섯 단계의 과정을 거친다.

(1) 요약문 쓰기의 진행 과정

1. 원자료에 대해 이해한다.
2. 텍스트 중에 가장 중요한 정보를 선정한다.

3. 텍스트 중에 부가적 정보를 삭제한다.
4. 선택한 정보에 관해 이해하고 이를 종합한다.
5. 원자료의 수사 구조를 반영하여 선택한 정보를 배열한다.

요약하기는 원텍스트(제시문)에 대한 이해, 가장 핵심적인 정보의 선택, 핵심을 벗어나는 정보의 삭제, 선택한 정보의 이해와 통합, 원텍스트의 수사적 구조를 반영하는 방식으로 재구조화의 과정을 거친다. 특히 핵심적인 정보들을 선택하고 이를 재구조화하는 과정이 중요하며 상대적으로 난이도가 높다.

위의 5단계를 본 연구의 요약문 생성기준(추상 요약용,van Dijk & Kintsch(1977, 1978)와 Browan & Day(1983)의 규칙 결합)에 대입하면 다음과 같다.

단계	Ferris & Hedcock	요약문 생성기준	구체적인 활동양상
1	원자료에 대해 이해한다		
2	텍스트 중에 가장 중요한 정보를 선정한다	선택	주제문의 선별, 중심 내용의 선정
3	텍스트 중에 부가적 정보를 삭제한다	삭제	반복적이고 사소한 내용 삭제
4	선택한 정보에 관해 이해하고 이를 종합한다	일반화	사우이어로 대체
5	원자료의 수사 구조를 반영하여 선택한 정보를 배열한다	구성	주제문 생성, 중심 내용의 재구성

표 32 요약문 생성기준 및 구체적인 활동양상

요약문 쓰기는 원텍스트(제시문)에 대한 이해, 핵심적인 정보의 선택, 핵심을 벗어나는 정보의 삭제, 선택한 정보의 이해와 통합, 원텍스트의 수사적 구조 반영 등의 재구조화 과정을 기반으로 수행된다. 특히 핵심적인 정보들을 선택하여 재구조화하는 과정이 쓰기 능력과 상당히 높은 관계를 갖고 있다고 볼 수 있다. 이러한 측면에서 볼 때, 요약문 쓰기는 읽기와 쓰기 능력 모두에 관여된다는 잠정적 결론이 가능하다.

따라서 학술 영역에서 요약문 쓰기는 일반적으로 읽기와 쓰기와 동시에 관련을 맺는 활동으로 “학문적 읽기와 쓰기의 능력을 발달시킬 수 있다는 도구적” 역할을 담당(이은경, 2013a:3-4)하는 것으로 이해된다. 또한 요약문 쓰기는 요약이라는 ‘과제’ 수행을 전제되어 있으므로 보다 적극적인 읽기를 유도함으로써 읽기 능력을 향상시키며(Friend, 2001), 특히 읽기 전, 읽기 중, 읽기 후에 수행되는 ‘모든 종류의 쓰기’4)가 내용 이해에 큰 도움을 준다(Leki, 1993). 즉, 요약문 쓰기는 독자로 하여금 텍스트에 대한 이해도를 높이는 것은 물론, 보다 심화된 이해를 할 수 있는 전략을 적극적으로 활

용하도록 유도하는 기능을 갖고 있다.

텍스트에 대해 심화된 이해를 추구하는 성격으로 인해 요약문 쓰기는 학술적 글쓰기를 구성하는 내용적 역할을 담당한다(이은경, 2013a:4). 다시 말해 요약문 쓰기는 읽고 쓰는 능력 향상에만 영향을 미치는 것에 그치지 않고 학술적 글쓰기를 실질적으로 구성하는 역할을 수행하고 있는 것이다. 요약문 쓰기는 교재의 요약, 논문에서 주요 아이디어 정리, 자료로부터 아이디어 요약, 수업의 필기 등 학술적 글쓰기를 수행하는데 필수적인 활동들에 해당된다고 볼 수 있다(Kennedy & Smith, 2006:55).

3. 주제문 주석 및 요약문 생성 지침

본 장은 실제 요약 말뭉치 구축을 위해 활용한 주제문 주석 및 요약문 생성 지침을 수록하고 있다. 요약 말뭉치는 기사 원문 중에서 세 문장의 주제문을 선택하고, 기사 내용을 요약한 세 문장의 요약문을 생성하는 방식으로 구축된다. 3.1절에서는 요약 말뭉치 구축의 기본 원칙을 소개하고, 3.2절에서는 요약 말뭉치 구축 작업 절차, 3.3절에서는 각 절차별 세부 지침을 예시와 함께 제시하고 있다.

3.1. 기본 원칙

가. 주제문 주석

주제문 주석은 기사 원문의 문장을 수정 없이 단순 선택하는 것이다. 주제문은 기사당 세 문장씩 주석한다.

나. 요약문 생성

요약문 생성은 기사 전체의 내용을 포괄하는 요약 문장을 새로 작성하는 것이다. 요약문은 기사당 세 문장씩 생성한다.

3.2. 주제문 주석 및 요약문 생성 절차

요약하기는 원텍스트(제시문)에 대한 이해, 가장 핵심적인 정보의 선택, 핵심을 벗어나는 정보의 삭제, 선택한 정보의 이해와 통합, 원텍스트의 수사적 구조를 반영하는 방식으로 재구조화의 과정을 거치면서 이루어진다. 특히 핵심적인 정보들을 선택하고 이를 재구조화하는 과정이 중요하다. Ferris & Hedgcock(2005)에 따르면 요약하기는 아래의 5단계를 거친다.

1. 원자료에 대해 이해한다.
2. 텍스트 중에 가장 중요한 정보를 선정한다
3. 텍스트 중에 부가적 정보를 삭제한다
4. 선택한 정보에 관해 이해하고 이를 종합한다
5. 원자료의 수사 구조를 반영하여 선택한 정보를 배열한다

위의 5단계를 단순화하고 각 단계에 대하여 작업자들이 실제로 수행할 내용을 정리하면 아래의 표 33과 같다.

단계	작업 단계	실제 작업 방식
1	원문 읽기	기사 원문을 읽는다
2	핵심 문장 찾기	최상위어/상위어(구)를 선정한다
		단락별 핵심 문장 찾기
3	주제문 추출하기	주제문을 참고하여 중복, 지엽적 내용을 생략하여 불필요한 문장 성분 등을 삭제하기
4	요약문 생성하기	선택된 정보를 포괄할 수 있도록 일반화된 서술을 통해 요약문 작성하기
		시간적/논리적 순서에 입각하여 요약문 재배열하기

표 33 주제문 주석 및 요약문 생성 절차

3.3. 각 절차별 세부 지침

3.3.1. 주제문 주석

주제문 주석은 기사 전반의 내용을 대표하는 3개의 문장을 선택하는 작업이다. 주석된 주제문은 추출 요약의 데이터가 되는 동시에 요약문 생성의 참고 자료가 된다. 표 33의 주제문 주석 및 요약문 생성 절차 중에서 2단계와 3단계가 주제문 주석에 해당된다.

1. 주제문으로 주석된 문장은 원문 그대로 유지하며 작업자 임의로 수정할 수 없다.
2. 주석된 세 문장은 가급적 중복되는 내용을 담고 있지 않도록 한다.
3. 원문의 전체 내용을 포괄할 수 있도록 상위어 또는 키워드를 포함하는 문장을 주제문으로 주석하도록 한다.

1) 핵심 문장 찾기

주제문 주석 및 요약문 생성의 두 번째 절차인 '핵심 문장 찾기'는 원문의 내용을 모두 포괄하는 핵심어인 '상위어'를 찾고 상위어에 입각하여 핵심 문장들을 찾아내는 단계이다.

가) 최상위어/상위어(구) 선정

상위어는 원문의 전체 내용을 모두 포괄하는 단어나 어구를 뜻한다. 선정된 상위어는 주제문을 주석할 때뿐만이 아니라 요약문을 생성할 때도 중요한 기준이 된다. 작업자는 선정된 상위어에 입각하여 불필요하거나 거리가 먼 문장이나 내용을 배제해야 한다. 상위어를 제대로 파악하지 않으면 중요한 내용을 누락하거나 지엽적인 내용이 중심적인 위치를 차지할 수 있으니 주의해야 한다.

원문	헤드라인: [광주세계수영]'쑨양 이슈'에 中매체·팬들 "불편"·"담담 " '도핑 스캔들'이 '쑨양 패싱'으로 이어지면서 자국의 '수영스타' 쑨양(28)을 바라보는 중국인들의 심정이 불쾌하면서도 복잡하다. 담담한 팬심도 드러난다. 일각에선 국제무대에서 쑨양을 무시한 경쟁자들의 행동을 문제 삼았으나 다른 한편에선 '신경쓰지 않는다'며 담담한 태도를 보였다. 도핑 의혹에 대해선 대다수가 쑨양에 대한 강한 신뢰를 드러냈다. 신민이브닝뉴스의 루 웨이쉰 기자는 "아무나 설 수 없는 공식적이고 중요한 시상대에서 보인 맥 호튼(23·호주)과 던컨 스콧(22·영국)의 행동은 무례했다. 두 선수 모두 본인에게 좋을 것이 없다"고 말했다. 이어 "쑨양의 격앙된 반응도 물론 불필요한 행동이다. 시상식에서는 승리 세레머니에만 집중해야 한다"고 거듭 강조했다. 그러면서 "국제수영연맹이 시상식에서 보인 두 선수의 태도에 대해 경고한 것은 공정하고 필요한 조치였다고 본다"고 조심스레 평가했다. 그는 논란을 바라보는 중국인들의 격앙된 반응과 불쾌감을 전하기도 했다. 논란에 크게 개의치 않는다는 입장도 나왔다. 경영 경기장에서 만난 한 중국인 남성(29)은 "경기는 모든 선수에게 공정하다. 호튼과 스콧의 행동은 스스로의 선택이며, 존중해야 한다"고 밝혔다. 도핑 회피 의혹에 대해서는 "많은 중국인들은 쑨양과 그의 말을 변함없이 믿는다"고 답했다. 어머니와 경기장을 찾은 첸첸(34·여)은 "경기에서 이긴 것이 중요하다. 다른 선수들이 보이는 태도에 크게 신경쓰지 않는다"며 "쑨양을 둘러싼 의혹도 언젠가는 진실이 밝혀질 거라 믿는다"고 전했다. 쑨양이 지난해 도핑테스트를 고의로 회피했다는 의혹이 일면서 쑨양의 경쟁자들이 불편한 감정을 잇따라 드러내고 있다. 한편 400m 자유형에서 은메달을 목에 건 호튼은 도핑 의혹에 항의하는 의미에서 시상대에 나란히 서지 않으며, 금메달리스트 쑨양을 외면했다. 200m 자유형에서도 동메달리스트 던컨 스콧(22·영국)이 금메달을 따낸 쑨양과 악수도, 기념촬영도 모두 거절하면서 '쑨양 패싱' 논란이 가열됐다.
상위어	쑨양, 쑨양 패싱, 도핑 의혹, 도핑 스캔들

표 34 최상위어/상위어(구) 선정 예시

위 표 34의 기사는 2019년 광주세계수영대회에서 일어난 해프닝에 대하여 다루고 있다. 도핑 의혹을 받고 있는 중국 선수 '쑨양'에 대한 항의의 표시로 다른 나라 선수들이 쑨양을 무시했다는 내용이다. 따라서 위 예시에서는 '도핑 스캔들'과 '쑨양 패싱', '도핑 의혹' 등이 전체 기사의 내용을 포괄하는 상위어가 될 수 있다. 이중 '도핑 스캔

들', '도핑 의혹'은 기사의 주 내용인 '쑨양 패싱'이 일어난 이유이며 위 기사는 '쑨양 패싱'으로 인해 발생한 사건을 구체적으로 다루고 있기 때문에, 이들 중 최상위어는 '쑨양 패싱'이 된다.

나) 단락별 핵심 문장 찾기

기사의 주제문은 원문의 맨 앞에 위치하는 경우가 일반적이지만 항상 그런 것은 아니다. 오히려 이후 문장이 전체 내용을 포괄하는 핵심 문장인 경우도 적지 않기 때문에, 항상 작업자 스스로 자신이 선택한 문장보다 전체 내용을 더 잘 포괄하는 다른 문장이 있는가에 대한 점검을 수행해야 한다.

원문	**일본프로야구 히로시마 구단이 선수에게 손찌검한 오가타 고이치(51) 감독에게 엄중주의 조치했다고 24일 발표했다.** 일본 '데일리스포츠' 등 주요 언론에 따르면 오가타 감독은 지난 6월30일 요코하마전 당시 연장 11회 전력 질주를 게을리했다는 이유로 외야수 노마 타카요시에게 몇차례 손찌검한 것으로 알려졌다. 이날 9회 대주자로 나섰던 노마는 2-2로 맞선 연장 11회 초 1사에서 타석에 들어섰다. 투수 플라이성 타구를 때렸는데, 1루로 전력 질주하지 않았다. 문제는 상대 투수가 포구에 실패하면서 공이 떨어졌는데, 노마가 앞서 달리지 않았다. 투수가 재빠르게 공을 잡아 1루로 던져 아웃카운트를 잡았다. 오가타 감독으로서는 화가 날 수밖에 없는 상황이었다. 하지만 히로시마 구단은 "(어떠한 상황에도)손을 올려서는 안 된다. 과한 행동은 무슨 상황이 벌어졌다고 해도 용납할 수 없는 일"이라고 했다. 오가타 감독 역시 "두 번 다시 이같은 일이 발생하지 않도록 하겠다"고 사죄했다. '데일리스포츠'는 '오가타 감독은 지난 15일 요코하마전을 앞두고도 선수단과 프런트 앞에서 심려를 끼쳤다면서 사죄했다'고 보도했다.
원문	① 대한민국 축구 대표팀을 지휘했던 두 감독이 중국 FA컵 8강에서 격돌했다. **② 최강희 감독이 이끄는 상하이 선화가 울리 슈틸리케 감독의 톈진 터다를 꺾고 4강에 올랐다.** 상하이는 24일 밤 중국 톈진 올림픽 센터 스타디움에서 열린 2019 중국 FA컵 8강전에서 3-1 완승을 거뒀다. 핵심 공격수 김신욱을 쉬게 한 최 감독은 전반 39분 가오디, 후반 29분 빈진하오, 후반 38분 차오윈딩의 골로 톈진 터다를 완파했다. 톈진 터다는 경기 종료 직전 수원 삼성 출신 브라질 공격수 조나탄이 한 골을 만회해 영패를 면했다. 최강희 감독은 중국 무대 입성 첫 시즌에 FA컵 4강 진출의 쾌거를 이뤘다. 같은 시간 상하이 상강은 헐크와 오스카의 득점으로 중국 슈퍼리그 선두를 달리는 광저우 헝다를 2-0으로 꺾고 4강에 올랐다. 상하이의 두 팀이 나란히 4강

	진출에 성공했다. 박충균 감독이 이끄는 톈진 톈하이는 라파 베니테스 감독의 다롄 이팡과 홈 경기에서 0-4로 크게 졌다. 박 감독은 이 경기에 외국인 선수는 물론 주전급 중국 선수들도 주말 리그 경기에 대비해 모두 제외했다. 박 감독은 “홈 팬들에게 죄송하다. 하지만 지금은 잔류 경쟁이 더 중요하다. 톈진와 더비 전을 잘 준비하겠다“고 했다. 톈진 톈하이는 오는 28일 슈틸리케 감독의 톈진 터다와 더비전을 앞두고 있다. 톈진의 두 팀은 모두 탈락했다. 김민재가 출전한 베이징 궈안도 산둥 루넝과 원정 경기에서 연장전 끝에 1-2로 패해 4강 진출에 실패했다. 산둥은 브라질 공격수 게디스와 이탈리아 공격수 그라치아노 펠레의 골로 승리했다. 베이징은 박성의 코너킥을 세드릭 바캄부가 헤더로 연결해 승부를 연장전으로 끌고 갔으나 조나탄 비에라가 부상으로 이탈한 것에 이어 헤나투 아우구스투까지 부상자 명단에 올라 힘을 쓰지 못했다. 중국 FA컵 4강전은 상하이 선화의 최강희 감독이 전 소속팀 다롄 이팡과 격돌하고, 상하이 상강은 산둥 루넝과 맞붙는다.

표 35 단락별 핵심 문장 찾기 예시

위 표 35의 예시 중 첫 번째 기사에서는 밑줄 친 첫 번째 문장이 해당 기사의 핵심 문장이다. 그러나 두 번째 기사에서는 첫 번째 문장이 전체 내용을 포괄하는 것으로 보이지만 실제 원문의 내용을 살펴보면 두 번째 문장이 기사의 전체 내용을 보다 잘 포괄하고 있다는 것을 알 수 있다.

2) 주제문 추출하기

주제문 주석 및 요약문 생성의 세 번째 절차인 ‘주제문 추출하기’는 중복적이거나 지엽적인 내용을 생략하고 불필요한 문장을 삭제하는 단계이다.

가) 단락별 첫 번째 문장 중에서 세부 내용을 다룬 문장 생략하기

원문의 단락별 첫 번째 문장은 대부분 단락의 전체 내용을 포괄하는 역할을 한다. 그러나 일부 단락은 전체 기사의 부연, 상술, 예시만을 담당하는 경우가 있으므로 이러한 문장들은 주제문으로 선택하지 않아야 한다.

원문	① 이선균과 정려원이 JTBC 새 월화드라마 ’검사내전’의 주연으로 캐스팅을 확정지었다. ② 올해 말 방송되는 ’검사내전’(연출 이태곤/크리에이터 박연선/극본 이현 서자연)은 생활형 검사들의 오피스드라마로, 현직 검사 김웅이 저술한 동명의 베스트셀러가 원작이다. 미디어 속 화려한 법조인이 아닌 지방도시 진영에서 하루하루

	살아가는 평범한 ‘직장인 검사’들의 이야기를 그린다. ③ 최근 영화 '기생충'을 통해 믿고 보는 명품 배우임을 다시금 입증한 이선균이 이번 '검사내전'에서 생활밀착형 검사 이선웅 역으로 출연을 확정지었다. 이선웅은 진영지청 형사 2부 소속으로 선한 인상에 출세욕 없이 느긋해 보이지만 보기와 달리 만만치 않은 '한 방'을 지닌 인물. 이선균 특유의 따뜻하고 묵직한 연기 톤이 배역과 높은 싱크로율을 자랑하는 만큼 이선균 표 이선웅의 모습이 벌써부터 기대감을 불러일으킨다. ④ 중앙지검에서 승승장구하다 하루아침에 머나먼 진영지청으로 발령받은 엘리트검사 차명주 역으로는 배우 정려원이 캐스팅됐다. '기름진 멜로', '마녀의 법정' 등 다양한 작품에서 장르를 넘나드는 팔색조 매력을 선보인 정려원은 이번 '검사내전'에서 빈틈없이 완벽한 검사 차명주로 변신, 걸 크러시 면모를 뽐낼 예정이다. ⑤ 이로써 '검사내전'은 탄탄한 베스트셀러 원작을 바탕으로 JTBC '청춘시대' 시리즈로 섬세한 연출력과 필력을 인정받은 이태곤 감독-크리에이터 박연선 작가의 재결합, 그리고 주연배우 이선균, 정려원의 합류까지 삼박자를 완성하며 올 하반기 최고의 기대작으로 떠올랐다. 2019년 말 방송 예정.

표 36 세부 내용 포함 문장 생략 예시

표 36의 기사는 모두 5개의 단락으로 구성되어 있으며 기사의 전체 내용을 포괄하는 핵심 문장은 ①이다. ②-⑤문장 중에서 ②, ⑤는 원문에서 중요한 내용을 포함하고 있으며 ①과 함께 주제문으로 주석될 수 있다. 하지만 ③, ④의 경우 ①의 문장을 상술하는 기능만을 담당하고 있기 때문에 ③, ④의 문장은 최종적으로 주제문 주석에서 배제해야 한다.

나) 주제문 추출 시 지엽적인 내용이 있는지에 대한 점검

주제문 추출은 최상위어와 상위어(구)에 입각하면서도 지엽적이거나 불필요한 부분은 최대한 배제해야 한다.

주제문	[1] 이선균과 정려원이 JTBC 새 월화드라마 '검사내전'의 주연으로 캐스팅을 확정지었다. [2] 올해 말 방송되는 '검사내전'(연출 이태곤/크리에이터 박연선/극본 이현 서자연)은 생활형 검사들의 오피스드라마로, 현직 검사 김웅이 저술한 동명의 베스트셀러가 원작이다. [3] 이로써 '검사내전'은 탄탄한 베스트셀러 원작을 바탕으로 JTBC '청춘시대' 시리즈로 섬세한 연출력과 필력을 인정받은 이태곤 감독-크리에이터 박연선 작가의 재결합, 그리고 주연배우 이선균, 정려원의 합류까지 삼박자를 완성하며 올 하반기 최고의 기대작으로 떠올랐다.

표 37 지엽적 내용 점검 예시

표 37은 원문인 표 36의 기사를 기반으로 추출한 주제문(장)들이다. 표 36의 최상위어는 '검사내전'이고 상위어(구)는 '이선균과 정려원이 '검사내전'의 주연으로 캐스팅'이다. 표 37은 앞선 지침들에 의거하여 선정한 주제문들이나 다소 지엽적인 내용이 포함되어 있다. 예를 들어 표 37의 [2]번 문장 <올해 말 방송되는 '검사내전'(연출 이태곤/크리에이터 박연선/극본 이현 서자연)은 생활형 검사들의 오피스드라마로, 현직 검사 김웅이 저술한 동명의 베스트셀러가 원작이다>에서 "올해 말 방송되는 '검사내전'(연출 이태곤/크리에이터 박연선/극본 이현 서자연)은 생활형 검사들의 오피스드라마"까지는 최상위어('검사내전')와 연관되지만 이후 문장 "생활형 검사들의 오피스드라마로 현직 검사 김웅이 저술한 동명의 베스트셀러가 원작이다."는 최상위어와 상위어(구)와 거리가 있어 지엽적인 내용이라고 할 수 있다. 만약 추출한 주제문에 이렇게 지엽적인 내용이 있는 경우에는 첫째, 비슷하게 전체적인 내용을 포괄하지만 지엽적인 부분은 담고 있지 않은 다른 문장을 주제문으로 주석하거나 둘째, 그러한 문장이 존재하지 않는 경우에는 해당 문장을 그대로 주제문으로 주석하되 요약문을 생성할 때 지엽적인 정보를 삭제하여야 한다.

다) 기사의 의도를 최대한 포괄하는 문장으로 주제문 주석하기

주제문을 주석하기 위해서는 기사가 전달하려는 의도를 명확히 파악해야 한다. 기사의 핵심 의도를 잘못 파악하는 경우 주제문을 주석하는 방향이 왜곡될 수 있기 때문이다.

원문	(이전 생략) [5] WEST 프로그램은 지난해 8월 한미 정상회담에서 합의된 것으로 지난해 8월 한미 정상회담에서 합의된 것으로 대학생 및 최근 졸업생들이 미국에서 최장 18개월 동안 체류하면 어학연구(5개월), 인터취업(12개월), 관광(1개월)을 할 수 있도록 한 제도. [6] 미국 국무부가 관리하는 스폰서 업체가 연수기관과 일자를 알선해 준다. [7] 정부는 WEST 프로그램이 적은 비용으로도 어학연수를 대체할 것으로 예상했지만 참여자들의 반응은 회의적이다. [8] 이 프로그램에 참여한 장모(중앙대 4년)씨는 "신청할 때는 대한민국을 대표해 나간다는 자부심이 있었는데 진행 상황을 보니 실망이 크다"면서 "외교통상부에 문의해도 처음이라 시행착오가 많다는 말뿐이고 아직 오리엔테이션 한 번 열지 않았다"고 불만을 드러냈다. [9] 학생들은 어학연구 기간 체류비로 1만 2500달러를 먼저 마련해야 하지만 별도의 지원 체계가 마련돼 있지 않다. [10] 학생들은 "이 비용을 전액 부담하려면 유학원을 이용해 어학연수를 다녀오는 것과 큰 차이가 없다"고 입을 모은다.

	[11] 유모(서울대 4학년)씨는 “다른 어학연수 프로그램에 비해 인턴을 병행할 수 있다는 게 장점인데 인턴 취업에 대한 확신이 없어 불안하다”면서 “정부에서는 인턴 임금으로 체류비를 충당할 수 있을 것이란 말만 되풀이하고 있는데 취업이 안 되면 결국 시간낭비만 하는 것 아니냐”고 지적했다. [12] 이에 대해 외교부 관계자는 “인턴 배정 가능 인원을 생각해 325명을 선발한 것”이라며 “어학연수 5개월 중 취업박람회 등을 통해 일자리를 확정할 예정”이라고 해명했다.

표 38 주제문 주석 예시

위 표 38의 기사 주요 내용은 ‘정부가 미국과 합의하며 마련한 WEST 프로그램의 지원이 미비하여 대학생들의 반응의 회의적’이다. 주제문을 주석한다면 먼저 WEST 프로그램을 설명하는 [5]를 선택할 수 있다. 다음으로는 [7]을 선택할 수 있는데, [7]은 [8]을 포괄하는 문장으로 볼 수 있기 때문이다. 그 다음 주제문으로는 [9]를 선택할 수 있다. [10]-[12]는 [9]를 상술하거나 보완하는 내용이기에 주제문이 될 수 없다. 즉, 위 기사에서 [7]과 [9]는 각각 뒤의 문장(들)의 정보를 포괄하는 추상적 문장으로 볼 수 있다.

기사가 비용의 문제, 인프라 부족, 취업 불확실성 등을 한꺼번에 다루기 때문에, 어떤 문장을 주제문으로 선택하는가가 다음 주제문 주석에도 강한 영향을 미친다. 주제문 주석도 작업자의 인지 처리 과정이므로 작업자 입장에서는 이전 정보와의 연관되는 정보를 처리하는 것이 자연스러운 것으로 이해하기 때문이다. 예를 들어 작업자에 따라 [5] 이후에 [8]과 [11] 또는 [9]과 [11]을 주제문으로 주석할 수 있다. 프로그램에 대한 불만을 표시하는 내용인 [8]을 두 번째 주제문으로 주석하는 경우 [11]을 이후 주제문으로 주석하는 것이 자연스럽다. [11]은 역시 [8]처럼 인터뷰의 구체적인 내용으로 두 문장의 정보 순위가 등위적이기 때문이다. 만약 두 번째 주제문으로 [9]를 선택하는 경우에는 이를 구체화하는 [11]을 선택하는 것이 자연스럽다. 이처럼 기사의 핵심 내용이나 의도를 제대로 파악하지 못한다면 주석된 주제문의 방향이 왜곡될 수 있다.

라) 서브 헤드라인 배제하기

주제문 주석의 대상이 되는 기사들 중에서는 서브 헤드라인을 가지고 있는 것들이 있다. 서브 헤드라인은 헤드라인을 보조해서 설명하는 역할을 가지고 있으며 가사의 전체 내용을 포괄하고 있지만, 기사의 본문이 아니기 때문에 주제문 추출의 대상에 포함되지 않는다. 따라서 본문의 첫 부분이나 중간에 나타나는 서브 헤드라인은 주제문으로 주석하지 않아야 한다.

3) 서브 클래스 기사 주제문 주석

일반적인 기사와 그 목적이나 형식이 상이한 기사들은 서브 클래스로 분류되며, 서브 클래스로는 인터뷰, 열거, 리뷰, 광고, 부가정보제시가 있다. 이 중에서 리뷰와 광고는 다른 기사들과 달리 특정한 목적을 지니고 있거나 객관적이지 않은 정보를 다루고 있지만 기사의 형식면에서는 큰 차이가 없다. 따라서 일반적인 기사와 마찬가지의 방법으로 주제문을 주석할 수 있다. 그러나 인터뷰, 열거, 부가정보제시의 경우에는 일반적인 기사들과 형식적인 면에서 차이를 보이고 있기 때문에 주제문 주석에 있어 주의할 부분이 따로 존재한다.

가) 인터뷰 서브 클래스 기사 주제문 주석

인터뷰 기사의 경우 인터뷰가 수행된 시공간, 인터뷰이(interviewee)의 신상 등을 중심으로 요약을 하는 방식은 되도록 지양해야 하며, 인터뷰에서 다루는 내용(사건)을 중심으로 요약을 해야 한다. 이를 성공적으로 수행하기 위해서는 인터뷰 내용 중에서 전체를 포괄할 수 있는 문장들을 주제문을 주석해야 한다.

원문	[1] 축구는 스포츠 이상이다. [2] 국민을 하나로 묶고 정치가 해결할 수 없는 일을 해낸다. [3] 사상 처음 아프리카 대학에서 열리는 2010년 남아프리카 공화국 월드컵도 축구 이상의 큰 의미를 담고 있다. [4] 대니 조든 남아공월드컵 조직위원회(SALOC) 사무총잠은(57)은 “내년 월드컵은 남아공을 하나로 묶어 인정 차별의 잔재를 없앨 것이며 세계 평화에도 기여할 것”이라고 말했다. [5] SALOC의 최고경영자로 월드컵의 모든 것을 관장하는 그를 29일 남아공 요하네스버그 샌런의 미켈란젤로 호텔에서 만났다. [6] 그동안 일부 국내 언론과 짧은 소감 인터뷰는 있었지만 장시간에 걸쳐 월드컵 청사진을 제시한 것은 이번이 처음이다. [7] - 1년 남았는데 월드컵 준비는 잘되고 있는가. [8] “한국의 본선 진출을 축하한다. [9] 아직도 2002년 한일 월드컵 기억이 생생하다. [10] 정말 환상적이었다. [11] 수백만 한국 팬들이 붉은색 옷을 입고 경기장 밖 과장에 모여 응원한 것은 우리도 본받고 싶다. [12] 내년 남아공에서도 2002년 한국의 응원 방식을 도입해 국민들을 하나로 뭉치게 할 것이다. [13] 이번 2009 컨페더레이션스컵 때 16경기를 치렀는지 문제가 없었다. [14] 현재 경기장은 재대로 건설되고 있다. [15] 5개는 이미 완성됐다. [16] 내년 초까진 10개 모두 완성될 것이다 “ (이하 생략)

표 39 인터뷰 서브 클래스 기사 주제문 주석 예시

위의 인터뷰 기사 주제문 주석 예시에서는 전체 내용을 포괄할 수 있는 문장은 [3]이다. 이후 [4]와 [12]가 주제문으로 주석될 수 있다. [3]은 기사의 전체 내용을 포괄하고, [4]는 인터뷰이의 정보와 인터뷰이가 타인에게 전달하려고 하는 핵심 메시지이며, [12]는 [4]를 보완하고 있기 때문이다. [5]나 [6]처럼 인터뷰가 진행된 시공간이나 인터뷰이에 대한 지나치게 상세한 내용은 주제문으로 주석하지 않아야 한다.

나) 열거 서브 클래스 기사 주제문 주석

기사 중에는 전시회나 공연 등 다수의 이벤트가 나열식으로 서술된 것이 있다. 이렇듯 정보들이 등위적으로 구성되어 있는 기사에서는 각각의 이벤트의 성격을 포괄하여 설명하는 문장을 주제문으로 주석해야 한다. 그리고 이벤트가 많이 제시되는 경우 '대표적인 이벤트'를 설명하는 문장을 주제문으로 주석한다. 만약 모든 이벤트들이 비슷한 분량으로 다루어진 경우에는 기사의 맨 앞에 언급된 이벤트를 중심으로 주제문 주석을 해야 한다.

원문	[1] 듣고, 보고, 만져보고. [2] 세계 각국의 전통 악기를 한자리에서 체험할 수 있는 공연과 전시회가 잇따라 찾아온다. [3] 음악을 통해 그 나라의 문화를 이해하는 교육 효과도 기대할 수 있다. [4] ▽국내외 전통 악기가 만드는 화음 [5] 7~10일 서울 종로구 원서동 북촌창우극장에서는 '세계전통악기축제'가 열린다. [6] 일본 베트남 우즈베키스탄 몽골 인도 등 5개국의 전통 악기를 연주하고 해설 및 연주자 인터뷰를 통해 해당 국가의 전통 악기와 문화를 소개하는 자리다. [7] 1998년 프랑스 월드컵과 2002년 한일 월드컵의 폐막공연에 참가한 일본의 전통 타악기 다이코 연주자 스이치 히다노 씨, 벨기에 나이지리아 등 세계를 돌며 연주를 펼쳐온 인도 전통 현악기 시타르 연주자 자그딥 싱 베디 씨 등이 출연한다. [8] 특히 각국 악기가 한국 전통 악기와 함께하는 특색 있는 협연 무대가 펼쳐져 기대를 모은다. [9] 다이코는 한국의 장구와, 베트남 현악기 단짠은 해금과, 몽골의 '후미 창법'은 한국의 정가와 호흡을 맞춘다. [10] 거문고 연주자이자 북촌창우극장 대표인 허윤정 씨는 "100여 석의 작은 소극장에서 이국의 악기 명인과 친밀한 시간을 가질 수 있을 것"이라고 말했다. [11] 2만 원.

	[12] 02-747-3809 [13] ▽세계 악기 2000개를 한자리에= [14] 10일부터 내년 2월 13일까지 경기 고양시 일산 킨텍스 1A홀에서 열리는 ‘시끌벅적 악기궁전’은 세계에서 온 악기 2000여 개를 만나는 악기 체험전. [15] 스코틀랜드의 ‘백파이프’, 그리스의 ‘팬플루트’처럼 다소 익숙한 악기부터 가나의 ‘크판로고 드럼’, 브라질의 ‘탐발’, 인도의 ‘탄푸라’ 등 이름조차 생소한 악기까지 총출동했다. [16] ‘바람의 소리’(관악기), ‘손가락 소리’(건반 악기), ‘두드림 소리’(타악기), ‘줄의 소리’(현악기) 등으로 전시관을 나눴으며 일부 악기는 직접 연주해 볼 수 있다. [17] 자연과 흡사한 소리를 내는 악기들도 눈길을 모은다. [18] 바람 소리가 나는 ‘윈드머신’, 빗소리가 나는 ‘레인스틱’, 파도 소리가 나는 ‘오션드럼’ 등이 아이들의 호기심을 자극한다. [19] 이 악기들은 그림자극 ‘우리 집이 최고야’를 통해 공연으로도 만나볼 수 있다. [20] 1만2000원. [21] 1544-1555 [22] ▽세계 악기 명인들의 무료 공연= [23] 여수시와 월드마스터조직위원회는 2010년 여수세계박람회 성공 개최를 기원하는 ‘월드마스터 페스티벌’을 3~5일 전남 여수시 오림동 여수진남체육관에서 연다. [24] 세르비아의 전통 백파이프 연주자 에디 타임 씨를 비롯해 전통 공연과 미술 전시에 30여 개 나라의 장인 60여 명이 참여한다. [25] 무료. [26] 070-8228-0990

표 40 열거 서브 클래스 기사 주제문 주석 예시

위의 열거 기사는 여러 공연과 전시회에 대한 정보가 나열되어 있다. 우선 모든 정보를 포괄적으로 나타내는 [2]가 주제문으로 주석될 수 있다. 전체 기사에서 중점적으로 다루어지고 있는 공연이나 전시회가 없이 3개의 전시회가 비슷한 분량을 가지고 있으므로, 나머지 두 개의 주제문은 가장 첫 번째로 제시되는 전시회에 대한 문장으로 주석한다. 전시회를 소개하는 [5]와 부연 설명을 하는 [6]이 주제문이 될 수 있다.

다) 부가정보제시 서브 클래스 주제문 주석

부가정보제시 서브 클래스 기사는 일반적인 기사와 형식이 유사하지만 기사의 중간이나 후반부에 기사의 등장인물 혹은 소재에 대한 부가정보가 따로 제시된다. 운동선수의 연혁 등이 이에 포함된다. 이렇게 제시되는 정보들은 기사 전체의 내용과는 큰 관련이 없기 때문에, 부가정보가 아닌 부분에서 주제문을 주석하여야 한다.

원문	[1] 미국 연방준비제도이사회(FRB)가 조만간 경기 부양을 위한 2차 '양적 완화'에 나설 것이 확실시되고 있다. [2] 양적 완화(Quantitative Easing)는 중앙은행이 국채 매입 등을 통해 시중에 유동성(流動性・자금)을 공급하는 것을 말한다. [3] 12일(현지시각) 미국 FRB가 공개한 9월 연방공개시장위원회(FOMC) 회의록에 따르면, 통화 정책을 결정하는 FOMC 위원들은 조만간 2차 양적 완화 조치에 나서기로 의견을 모은 것으로 확인됐다. [4] 상당수 FOMC 위원은 지난달 회의에서 "현재 미국 경제의 물가 상승률이 지나치게 낮고 실업률이 너무 높다"면서 "경기 회복을 위한 새로운 조치가 필요한 상황"이라고 말했다. [5] 미국 월가(街)에선 FRB가 다음 달 2~3일 열리는 FOMC에서 최대 1조달러(약 1100조원) 규모의 양적 완화 조치를 발표할 가능성이 큰 것으로 보고 있다. [6] FRB가 이처럼 양적 완화에 나서는 것은 최근 미국 경제에 디플레이션(물가가 지속적으로 내리는 현상) 위험이 고조되고 있기 때문이다. [7] FRB가 정한 장기 인플레이션 관리 목표는 1.7~2%인데, 지난 8월 근원 소비자물가지수는 지난해 같은 달에 비해 1.4% 상승하는 데 그쳤다. [8] 미국 실업률도 9% 중반대에서 좀처럼 떨어지지 않는 상황이다. [9] 이에 따라 FRB는 올 하반기와 내년 경제 성장률 전망치를 하향 조정했으며, 인플레를 유발할 특단의 대책도 검토하고 있다. [10] 특단의 대책이란 FRB가 물가 수준 목표제를 도입한 뒤 끊임없이 자금을 공급해 시장에 인플레 기대 심리를 불러일으키는 것을 말한다. [11] 하지만 추가 양적 완화가 역효과를 가져올 것이라는 시각도 없지 않다. [12] 리처드 피셔(Fisher) 댈러스 연방준비은행 총재는 지난 7일 "양적 완화가 고용을 촉진시킬 것이라는 주장에 회의적"이라고 밝혔다. [13] 토머스 호니그(Hoenig) 캔자스시티 연방준비은행 총재도 12일 "양적 완화는 금융시장에 불확실성만 더해줄 뿐 이득은 별로 없는 매우 위험한 전략"이라고 말했다. [14] ☞ 양적 완화(量的緩和・Quantitative Easing) [15] 중앙은행이 고유의 기능인 발권력을 동원해 화폐를 찍어낸 뒤 국채나 회사채를 매입하는 방식으로 시중에 유동성(流動性・자금)을 공급하는 정책을 말한다. [16] 기준금리가 제로에 근접한 상황에선 더 이상 금리 인하를 통해 자금을 풀기 어렵기 때문에 경기 부양을 위한 각종 양적 완화정책이 동원된다.

표 41 부가정보제시 서브 클래스 기사 주제문 주석 예시

위의 부가정보제시 기사의 후반부에는 '양적 완화'라는 경제적 개념에 대한 설명이 부가적으로 제시되어 있다. 그러나 기사에서 중요하게 다루고 있는 내용은 미국 FRB가 양적 완화에 나선다는 사실과 그 이유, 그리고 그에 대한 의견이기 때문에 양적 완화에 대한 개념 설명이 주제문으로 주석될 필요는 없다. 따라서 양적 완화에 나선다는 사실을 밝히고 있는 [1]과 그 이유를 설명해 주는 [6], 그에 대한 반응을 보여주는 [11]이 주

제문으로 주석될 수 있다.

3.3.2. 요약문 생성

요약문 생성은 기사 원문의 내용을 포괄할 수 있는 요약 문장을 새로 작성하는 것이다. 요약문을 생성할 때는 미리 주석해 놓은 주제문을 활용할 수도 있으며 기사 전체에서 중요하다고 판단되는 내용이 3개의 주제문에 누락되어 있거나 주제문이 기사 전체의 내용을 완전하게 포괄하고 있지 않은 경우에는 주제문과는 별개로 생성할 수도 있다. 즉, 요약문 생성은 주제문을 바탕으로 하되, 주제문이 포괄하지 못하는 텍스트의 핵심 정보들을 최대한 반영하는 방향으로 수행되어야 한다.

1) 요약문 생성하기

가) 주석한 주제문을 가독성을 높이는 방향으로 재구성하여 요약문 생성하기

주석된 주제문들은 전체 기사의 내용을 가장 많이 담아내는 문장들을 기사에 등장하는 순서대로 뽑아놓은 것이기 때문에 3개의 문장만 보았을 때 시간적/논리적 순서가 뒤틀어져 있거나 문장 간의 연결이 어색할 수 있다. 따라서 주제문을 활용하여 요약문을 생성하는 경우에는 주제문들을 시간적/논리적 순서에 따라 재구성하는 과정이 필요하다.

원문	배우 송혜교 측이 송중기와의 이혼 소식 후 확산한 관련 악성 댓글과 루머 유포자들을 일괄 고소했다. 25일 경찰에 따르면 송혜교 측은 이날 분당경찰서에 허위사실 적시에 따른 명예훼손 및 모욕에 대한 내용으로 다수를 상대로 한 고소장을 냈다. 송혜교 측은 이번 고소와 관련, 어떠한 선처나 합의 없이 강경하게 대응하겠다는 입장을 경찰에 피력한 것으로 알려졌다. 앞서 송혜교와 송중기는 결혼 1년 8개월 만에 이혼 조정에 나선 사실이 보도되며 다양한 악성 댓글과 루머에 시달려왔다. 송중기 소속사 역시 루머 유포에 대해 법적 대응 하겠다는 입장을 밝힌 바 있다. 두 사람의 이혼 조정은 지난 22일 성립돼 두 사람은 결혼 1년 9개월 만에 완전히 남남이 됐다.
주제문	[1] 배우 송혜교 측이 송중기와의 이혼 소식 후 확산한 관련 악성 댓글과 루머 유포자들을 일괄 고소했다. [2] 송혜교 측은 이번 고소와 관련, 어떠한 선처나 합의 없이 강경하게 대응하겠다는 입장을 경찰에 피력한 것으로 알려졌다.

	[3] 앞서 송혜교와 송중기는 결혼 1년 8개월 만에 이혼 조정에 나선 사실이 보도되며 다양한 악성 댓글과 루머에 시달려왔다.
요약문	송혜교와 송중기가 결혼 1년 8개월 만에 이혼 조정에 나섰다. 이들은 이혼 소식이 전해진 후 다양한 악성 댓글과 루머에 시달렸다. 이에 송혜교 측은 악성 댓글과 루머 유포자들을 모두 고소했으며, 어떤 선처나 합의 없이 강경하게 대응하겠다는 입장을 밝혔다.

표 42 주제문을 재구성하여 요약문 생성하기 예시

표 42는 기사 원문과 그에 대한 주제문과 요약문의 예시이다. 위 기사는 원문이 길지 않고 문장이 곧 단락인 경우이므로 주제문(장) 주석이 비교적 쉬운 편이다. 주목해야 할 것은 생성한 요약문이다. 요약문의 첫 문장은 주제문의 마지막 문장을 기반으로 구성된 것이다. 이렇게 구성한 근거는 주제문 예시의 [3] 문장에서 '앞서'라는 단어가 있기 때문이다. 따라서 시간적 순서에 따르면 [3]의 문장이 생성한 요약문의 맨 앞에 위치하는 것이 자연스럽다. 독자의 내용 이해도는 무의식중에 시간적 순서에 따라 사건을 배치하거나 논리적으로 정합적일 때 높아지기 때문에 요약문을 생성할 때는 주제문(장)의 순서에 구애받지 말고 요약문 자체가 논리적일 수 있도록 생성해야 한다.

나) 주제문에서 중복되는 부분을 생략하여 요약문 생성하기

주제문을 참고하여 요약문을 생성하는 경우, 가독성을 높이기 위하여 주제문에서 반복되는 단어와 어구들을 생략하고 간결하게 문장을 구성해야 한다. 주제문을 추출할 때에도 최대한 중복되는 내용이 담기지 않도록 하는 것이 원칙이지만, 가시 원문을 수정하지 않고 그대로 추출하는 주제문 주석의 특성상 정보가 겹쳐서 나타날 수 있다. 따라서 주제문으로 주석된 문장들을 참고하여 요약문을 생성하는 경우에는 이러한 정보들을 모두 생략해 주어야 한다.

원문	올해 프로야구에선 2명의 '일본 U턴파'가 뛴다. 두산 이혜천(32)과 KIA 이범호(30)다. 2008년 두산에서 연봉 1억5000만 원을 받았던 이혜천은 일본에서 두 시즌을 보낸 뒤 돌아와 계약금 8억 원과 옵션 등 최대 11억 원에 사인했다. 2009년 연봉 3억3000만 원을 받았던 이범호는 친정 한화 대신 KIA 유니폼을 입으며 계약금 8억 원 등 총 12억 원을 받는다. 대개 프로선수의 몸값은 성적이 좌우하지만 둘은 특별하다. 이혜천은 지난해 야쿠르트에서 19경기에 출전해 1패, 평균자책 5.09에 그쳤다. 2009년에도 1승 1패 1세이브에 평균자책 3.65로 좋지 않았다. 이범호는 48경기에 나가 타율 0.224, 4홈런, 11타점의 초라한 성적을 올렸다. 둘 다 2군에 있던 시간이 많았다. [2] 이혜천은 야쿠르트와의 2년 계약이 끝난 뒤 방출됐고 이범호는 계약기간이 남았지만 복귀를 택했다. 국내 구단 소속으로 일본에 진출했다 돌아온 사례는 KIA 이종범이 원조다. 1997년 연봉 1억1000만 원을 받았던 그는 2001년 시즌 도중 복귀하며 3억5000만 원에 계약했다. 활동 기간이 짧아 실제 받은 액수는 훨씬 적었지만 3억5000만 원은 그해 리그

	최고 연봉이었다. 이후 U턴한 정민철과 정민태 역시 가기 전보다 훨씬 많은 몸값을 받으며 금의환향했다. 2006년 LG에서 연봉 5억 원을 받았던 이병규는 지난해 컴백하면서 연봉이 줄었지만 국내에 있을 때 고액 연봉자였던 데다 나이와 일본에서의 성적 등을 고려할 때 결코 적은 몸값은 아니었다. 일본에서는 체면을 구겼지만 돌아와서는 대부분 이름값은 했다. 특히 요미우리에서 2년 동안 2승 1패, 평균자책 6.28에 그쳤던 정민태는 2003년 복귀하자마자 17승(2패)을 올리며 다승왕과 골든글러브를 거머쥐었다. 두산 김경문 감독은 일찌감치 이혜천을 왼손 선발로 낙점했다. 이혜천은 13일 삼성과의 시범경기에서 삼진 7개를 솎아내며 3안타 무실점으로 막았다. 이범호는 15일 복귀 무대에서 3번 3루수로 출전해 LG를 상대로 3타수 2안타 1타점(결승타)을 기록하며 녹슬지 않은 기량을 과시했다. 일본 U턴파 환대에 대한 우려의 목소리도 크다. 무엇보다 '일본에서 실패해도 돌아오면 반겨준다'는 인식이 확산되는 게 문제다. 이혜천과 이범호는 구단의 환대에 어떻게 보답할까.
주제문	[1] 올해 프로야구에선 2명의 '일본 U턴파'가 뛴다. [2] 이혜천은 야쿠르트와의 2년 계약이 끝난 뒤 방출됐고 이범호는 계약기간이 남았지만 복귀를 택했다. [3] 일본 U턴파 환대에 대한 우려의 목소리도 크다
요약문(1)	두산 이혜천(32)와 KIA 이범호(30)는 '일본 U턴파'이다. 이해천은 일본에 가기 전에 소속된 구단인 두산으로 돌아왔고, 이범호는 친정인 한화 대신 KIA로 돌아왔다. 국내 구단 소속으로 일본 진출 후 돌아온 선수들이 대부분 좋은 성적을 거두기는 했으나 이를 환대하는 데에 대한 우려의 목소리도 있다.
요약문(2)	올해 프로야구에서 '일본 U턴파'는 두산 이혜천(32)과 KIA 이범호(30)이다. 이해천은 일본에 가기 전 소속된 구단인 두산으로, 이범호는 KIA로 돌아왔다. 국내 구단 소속으로 일본 진출 후 돌아온 선수들이 대부분 좋은 성적을 거두기는 했으나 이들을 무조건 환대하는 데에 대한 우려의 목소리도 있다.

표 43 주제문에서 중복되는 내용을 생략하여 요약문 생성하기

표 43은 기사 원문에 대한 주제문과 그를 참고하여 생성한 요약문의 예시이다. 먼저 생성한 요약문(1)의 경우, 두 번째 문장에 '돌아오다'가 두 번 중복되어 나타나고 있다. 따라서 요약문(2)의 두 번째 문장과 같이 보다 간결한 요약문으로 수정되어야 한다.

다) 요약문의 길이

요약문은 새로운 문장을 생성하는 것이다. 따라서 최대한 많은 정보를 담으려는 작업자의 의도에 따라 필요 이상으로 긴 문장이 완성될 수 있다. 그러나 요약문의 문장이 너무 길어지면 데이터 처리 작업 등의 난도가 높아질 수 있으며, 신문 기사를 최대한 간략하게 요약한다는 본래의 취지에도 부합하지 않게 된다. 따라서 요약문은 되도록 압축적으로 생성하여야 한다. 이를 위해서는 정확한 수가 아니거나 필수적이지 않은 날짜 정보와 '등, 따위'와 같이 삭제해도 문제되지 않는 단어들은 삭제하여 요약한다. '~할 것이다, ~으로 보인다'와 같은 완화 표현 역시 간결한 종결어미로 바꾸어 요약한다.

또한 접속어가 많이 쓰이는 경우 문장이 복잡해지고 의미가 불분명해질 수 있다. 따라서 요약문 안에서의 접속은 되도록 2개까지로 한정하며, 요약문의 총 글자 수는 130자로 제한하도록 한다. 연결어미는 고빈도로 사용되는 '-고, -아서/어서, -며' 등의 보편적인 것들로 사용한다.

주제문 예시	이에 앞서 북상초교 학교운영위원과 조선제 군의원 등 30여 명은 3일 도교육청에서 기자회견을 열어 "교장공모를 위한 심사과정에 아무런 하자가 없는데도 지역 언론 보도를 이유로 이를 취소한 것은 독선적인 행정"이라고 비난했다. 전교생이 43명인 이 학교는 올 6월 학부모들이 "폐교를 막고 학교를 살리려면 교장공모제가 필요하다"는 데 의견을 모았다. 그러나 도교육청은 일부 후보의 문제 제기와 지역 언론 보도 등을 이유로 현지 조사를 벌인 뒤 지난달 31일 시범학교 지정을 전격 취소했다.
요약문 예시 (1)	[1] 경북 거창군 북상초교 학교운영위원 등 30여 명은 교장 공모 과정에 아무런 하자가 없는데도 지역 언론 보도를 이유로 이를 취소한 것은 독선적인 행정이라며 도교육청에 철회를 요구했다. [2] 전교생이 43명인 이 학교는 올 6월 학부모들이 폐교를 막고 학교를 살리기 위해 교장공모제를 신청해 다른 5개 초중학교와 함께 시범학교로 지정됐으며 공모를 진행해 1, 2위 후보 2명을 도교육감에게 추천했다. [3] 그러나 도교육청은 일부 후보의 문제 제기와 지역 언론 보도 등을 이유로 현지 조사를 벌인 뒤 지난달 31일 시범학교 지정을 취소했다.
요약문 예시 (2)	[1'] 경북 거창군 북상초교 학교운영위원 30여 명은 교장 공모 취소는 독선적인 행정이라며 도교육청에 철회를 요구했다. [2'] 전교생이 43명인 이 학교는 올 6월 학부모들이 폐교를 막기 위해 교장공모제를 신청해 시범학교로 지정됐으며 공모를 진행해 후보 2명을 도교육감에게 추천했다. [3'] 그러나 도교육청은 일부 후보의 문제 제기와 지역 언론 보도를 이유로 지난달 31일 시범학교 지정을 취소했다.

표 44 요약문 생성 예시 - 요약문의 길이

위 표 44는 요약문 생성의 예시이다. 요약문 예시(1)의 [1] 중 '교장 공모 과정에 아무런 하자가 없는데도 지역 언론 보도를 이유로 이를 취소한' 부분은 [3]에 '지역 언론 보도'라는 정보가 주어지기 때문에 요약문 예시(2)의 [1']에서는 '교장 공모 취소'로 축약하였다. 또한 요약문 예시(1)의 [2] 중 '다른 5개 초중학교와 함께'는 필수적이지 않은 부분이기에 [2']에서 삭제되었다. [1]과 [3]의 '등'은 [1']과 [3']에서 삭제되었다. 또한 [2]의 '폐교를 막고 학교를 살리기' 부분은 동일하거나 거의 유사한 의미의 구조가 반복되고 있기 때문에 [2']에서 압축적으로 서술하였다.

라) 주제문과 차이 두기

요약문은 주제문을 참고하여 만들게 되지만, 주제문은 기사 전체의 정보를 담고 있지는 못하다. 따라서 요약문을 생성할 때에는 되도록 주제문과 완전히 동일한 문장이

아닌 더 많은 정보를 포괄하고 있는 문장을 생성하여야 한다. 또한 주제문 주석은 기사 원문의 문장을 그대로 사용하기 때문에 따옴표나 특수기호 등이 포함되어 있지만 요약문을 생성할 때는 이를 수정하도록 한다.

주제문 예시	[1] 순항하던 미국의 소셜 네트워킹 서비스(SNS · 사람들이 자신의 생각과 의견, 경험 등을 서로 공유하기 위해 사용하는 쌍방향 온라인 서비스)가 역풍을 맞고 있다. [2] 미 오클라호마 주 에너지회사인 원오크는 15일 누군가가 트위터에서 회사명과 로고를 도용해 글을 올렸다며 트위터를 상대로 상표권 침해 소송을 냈다. [3] ‘미 연방 통신품위법’ 등에 따라 사이트에서 발생한 이름, 상표 도용을 해당 사이트 측이 책임질 필요는 없다는 주장도 있지만 트위터 측은 이를 계기로 불거진 취약점을 수정하겠다는 의지를 보였다.
요약문 예시	[1’] 트위터와 페이스북이 사용자 이름 도용과 사생활 침해로 연이어 법정에 서는 등 SNS가 역풍을 맞고 있다. [2’] 트위터는 에너지회사 원오크로부터 상표권 침해 소송을 당했으며, 페이스북은 사생활 침해 논란이 인 비컨 프로그램을 없애기로 합의했다. [3’] ’ 미 연방 통신품위법 ‘에 따라 사이트에서 발생한 이름, 상표 도용을 해당 사이트 측이 책임질 필요는 없다는 주장도 있지만 트위터는 취약점을 수정하겠다고 하였다.

표 45 요약문 생성 예시 - 주제문과 차이 두기

위 표 45는 요약문 생성의 예시이다. 주제문 [1]에서의 ‘소셜 네트워킹 서비스(SNS· 사람들이 자신의 생각과 의견, 경험 등을 서로 공유하기 위해 사용하는 쌍방향 온라인 서비스)가’는 괄호 속의 정보가 괄호 밖의 정보를 부연하고 있다. 요약문 [1’]에서는 괄호를 풀고 동일한 정보 중 짧은 것을 남기고 부연 정보를 삭제하여 ’SNS가‘로 요약하였다. 이처럼 요약문을 생성할 때는 원어 표시나 부연 정보 제시를 위한 괄호, 강조를 위한 따옴표를 삭제하여 요약한다. 다만 [3]과 [3’]에서와 같이 고유명을 표현하기 위해 사용된 따옴표는 남겨두는 것을 원칙으로 한다.

또한 주제문에서는 나타나지 않던 페이스북에 대한 정보가 요약문 [2’]에 추가되었다.

2) 서브 클래스 기사 요약문 생성

가) 인터뷰 서브 클래스 기사 요약문 생성

주제문을 주석할 때와 마찬가지로 인터뷰가 수행된 시공간, 인터뷰이(interviewee)의 신상 등을 중심으로 요약을 하는 방식은 되도록 지양해야 하며, 인터뷰에서 다루는 내용(사건)을 중심으로 요약을 해야 한다. 또한, 인터뷰 기사는 인터뷰이의 목소리를 왜곡 없이 전달하기 위해 직접 인용을 하는 경우가 많다. 그러나 요약문 생성에서는 직접

인용된 문장은 모두 따옴표를 풀어서 요약한다.

주제문	[1] 사상 처음 아프리카 대학에서 열리는 2010년 남아프리카 공화국 월드컵도 축구 이상의 큰 의미를 담고 있다. [2] 대니 조든 남아공월드컵 조직위원회(SALOC) 사무총잠은(57)은 “내년 월드컵은 남아공을 하나로 묶어 인정 차별의 잔재를 없앨 것이며 세계 평화에도 기여할 것”이라고 말했다. [3] “먼저 각종 인프라 구축이 남아공을 발전시킬 것이다.
요약문	[1’] 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵은 국민을 하나로 묶고 정치가 해결할 수 없는 일을 해낼 것으로 기대된다. [2’] 남아공은 월드컵 개최로 각종 인프라를 구축해 국가 발전을 꾀하고, 많은 관광객을 유치하고, 국가 이미지를 제고하고, 흑인과 백인이 함께 어우러지는 효과를 얻을 것으로 보고 있다. [3’] 대니 조든 남아공월드컵조직위원회 사무총장은 월드컵은 남아공을 하나로 묶어 인종 차별의 잔재를 없앨 것이며 세계 평화에도 기여할 것이라고 말했다.

표 46 인터뷰 서브 클래스 기사 요약문 생성 예시

위 표의 주제문 예시를 보면, 원문을 그대로 가져왔기 때문에 [2]와 [3]에서 큰따옴표를 발견할 수 있다. 특히 [3]은 인터뷰이의 긴 발화의 시작 부분이기 때문에 닫는 큰따옴표 없이 여는 큰따옴표만 나타난 것을 확인할 수 있다. [2]는 요약문에서는 큰따옴표 없이 간접 인용방식으로 수정하여 [3‘]으로 요약되었다. 또한 주제문에서는 담을 수 없던 ‘각종 인프라 구축’에 대한 내용을 기사 전체에서 간추려 [2’]로 요약하였다.

나) 열거 서브 클래스 기사 요약문 생성

기사 중에는 전시회나 공연 등 다수의 이벤트가 나열식으로 서술된 것이 있다. 이렇듯 정보들이 등위적으로 구성되어 있는 기사에서는 각각의 이벤트의 성격을 포괄하여 설명할 수 있도록 요약문을 생성해야 한다. 이벤트가 많이 제시되는 경우에는 기사에서 가장 분량이 많은 ‘대표적인 이벤트’를 중심으로 요약문을 생성한다. 만약 모든 이벤트들이 비슷한 분량으로 다루어진 경우에는 기사의 맨 앞에 각각의 이벤트들을 최대한 간추려서 요약문을 생성한다.

주제문	[1] 세계 각국의 전통 악기를 한자리에서 체험할 수 있는 공연과 전시회가 잇따라 찾아온다. [2] 7~10일 서울 종로구 원서동 북촌창우극장에서는 ‘세계전통악기축제’가 열린다. [3] 일본 베트남 우즈베키스탄 몽골 인도 등 5개국의 전통 악기를 연주하고 해설 및 연주자 인터뷰를 통해 해당 국가의 전통 악기와 문화를 소개하는 자리다.
요약문	[1’] 세계 각국의 전통 악기를 한자리에서 체험할 수 있는 공연과 전시회가 잇따라 찾아온다. [2’] 7~10일 서울 종로구 원서동 북촌창우극장에서 열리는 ‘세계전통악기축제’는 일본 베트남 우즈베키스탄 몽골 인도 등 5개국의 전통 악기를 연주하고, 해설 및 연주자 인터뷰를 통해 해당 국가의 전통 악기와 문화를 소개한다. [3’] 경기 고양시 킨텍스 1A홀에서는 10일부터 내년 2월 13일까지 세계 악기 2000개를 한자리에

	모은 ‘시끌벅적 악기궁전’이, 전남 여수시 진남체육관에서는 3~5일 ‘월드마스터 페스티벌’이 각각 열린다.

표 47 열거 서브 클래스 기사 요약문 생성 예시

주제문 주석의 [1]과 동일하게, 모든 정보를 포괄적으로 나타내는 문장이 [1‘]로 요약되어 있다. 중점적으로 다루어지는 이벤트가 없었기에 주제문 주석에서는 첫 번째 공연에 대한 설명에서 문장을 추출하였지만, 기사를 간추려 새로운 문장을 생성할 수 있는 요약문 생성에서는 ’세계전통악기축제‘ 외 다른 2개의 공연을 [3’]으로 요약하였다.

다) 부가정보제시 서브 클래스 요약문 생성

부가정보제시 서브 클래스 기사는 일반적인 기사와 형식이 유사하지만 기사의 중간이나 후반부에 기사의 등장인물 혹은 소재에 대한 부가정보가 따로 제시된다. 운동선수의 연혁 등이 이에 포함된다. 이렇게 제시되는 정보들은 기사 전체의 내용과는 큰 관련이 없기 때문에, 부가정보가 아닌 부분에서 주제문을 주석하여야 한다.

주제문	[1] 미국 연방준비제도이사회(FRB)가 조만간 경기 부양을 위한 2차 ’양적 완화’에 나설 것이 확실시되고 있다. [2] FRB가 이처럼 양적 완화에 나서는 것은 최근 미국 경제에 디플레이션(물가가 지속적으로 내리는 현상) 위험이 고조되고 있기 때문이다. [3] 하지만 추가 양적 완화가 역효과를 가져올 것이라는 시각도 없지 않다.
요약문	[1 ‘] 연방공개시장위원회 위원들이 조만간 2차 양적 완화 조치에 나서기로 의견을 모아 미국 연방준비제도이사회가 경기 부양을 위한 최대 1조달러 규모의 양적 완화 조치에 나설 것이 확실시되고 있다. [2’] 이처럼 양적 완화에 나서는 것은 최근 미국 경제에 디플레이션 현상 위험과 실업률 정체가 고조되고 있으며 인플레를 유발할 특단의 대책을 검토하고 있기 때문이다. [3 ‘] 한편 연방준비은행 총재들은 회의적인 태도를 보이며 추가 양적 완화가 역효과를 가져올 것이라는 시각을 내비쳤다.

표 48 부가정보제시 서브 클래스 기사 주제문 주석 예시

주제문 주석에서 확인한 것과 같이, 위 기사의 후반부에는 ‘양적 완화’라는 경제적 개념에 대한 설명이 부가적으로 제시되어 있다. 그러나 기사에서 중요하게 다루고 있는 내용은 미국 FRB가 양적 완화에 나선다는 사실과 그 이유, 그리고 그에 대한 의견이기 때문에 양적 완화에 대한 부가정보제시 부분은 주제문으로 주석되지 않았다. 요약문 생성도 이와 마찬가지이기 때문에 양적 완화의 규모와 보다 구체적인 이유, 그에 대한 의견의 주체 등을 추가하여 문장을 요약하였다.

제 4 장

자동 요약 기술 적용

1. 자동 요약 기술

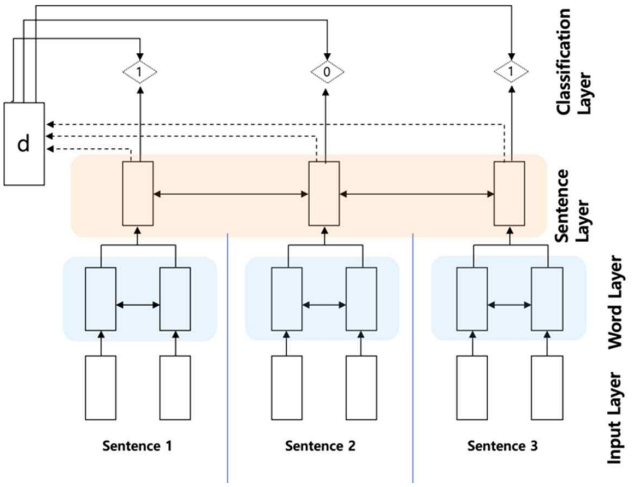
문서 요약은 불필요한 문장을 삭제하거나 주요 문장을 추출 또는 생성하여 문서의 길이를 줄이고 핵심 내용을 남기는 과정을 말한다. 문서 요약은 요약하는 방법에 따라 크게 추출 요약(extractive summarization)과 추상 요약(abstractive summarization)으로 나누어진다. 추출 요약은 원본 문서로부터 핵심 내용을 담고 있는 문장을 선별하여 제시하는 것으로써 문장 분류(sentence classification) 문제에 속한다. 즉, 원본 문서의 각 문장을 요약에 사용할지 말지를 결정하는 이진 분류(binary classification) 또는 문장 집합에 대한 순차 표지 부착(sequence labeling) 문제라고 할 수 있다. 추출 요약을 통해 생성된 요약은 원본 문서에 존재하는 문장이기 때문에 문법적, 의미적 오류를 포함하지 않는다는 장점이 있다. 그러나 원본의 일부만을 추출한 것이기 때문에 부정확한 참조를 사용하거나 일관성이 떨어지는 등 요약된 문장들의 연결이 매끄럽지 못하다는 단점이 있다. 또한 원본 문서가 가진 표현력을 뛰어넘지 못하며 저작권 문제로부터 자유롭지 못하다는 한계가 있다. 추상 요약은 추출 요약과 달리 원본 문서에 없는 새로운 단어나 문장을 생성하는 방법으로 최근 다양한 형태의 심층 신경망(deep neural network)을 통해 가능성을 보여주고 있다. 이와 같은 두 기술이 모두 실제 요약 서비스를 제공하는데 실제적으로 사용되고 있기 때문에 본 과제에서는 추출 요약과 추상 요약 기술을 개발하고 평가하는데 활용할 수 있는 말뭉치를 구축하고자 하였다.

이에 따라 한국어 요약 말뭉치를 활용한 문서 요약 모델 적용을 위해 기존에 연구된 문서 요약 모델을 조사하였고 이 중 효과적인 모델을 선정하여 본 과제에서 구축한 한국어 요약 말뭉치를 적용하였다. 모델의 성능을 높이기 위한 실험 및 다양한 입출력 단위에 대한 실험을 진행하였다. 본 장에서는 이에 대한 실험 내용 및 결과를 서술하고 한국어 요약 말뭉치의 효용을 검토한다.

2. 추출 요약 기술 적용

2.1. 한국어 추출 요약 기술

한국어 문서 추출 요약 시스템으로는 영어권 문서 요약에서 비교 모델로 널리 사용되고 있는 SummaRuNNer를 사용하였다. 이 모델은 RNN 2계층으로 구성된 신경망 모델로 기본적인 구조는 <그림 7>과 같다.



<그림 7> SummaRuNNer의 모델 구조

이 모델은 n개의 문장으로 구성된 문서 $X=[S_1, S_2, ... S_n]$에 대해 $Y=[y_1, y_2, ... y_n]$을 순차적으로 부여하여 추출 요약을 수행한다. 이 때, $y_k \in [0,1]$는 k번째 문장이 요약에 포함되어야 할지(1) 아닌지(0)을 나타내는 것으로 수식 (3)을 통해 결정된다.

수식 (3)에서 h_k는 k번째 입력 문장에 대한 벡터 표현이며, d는 입력 문서 전체를 표현하는 벡터이다. g_k는 k번째 이전까지 만들어진 부분 요약을 표현하는 벡터이다. 이들 정보와 각 항목에 대한 가중치를 이용하여 문장 자체의 중요성(content), 전체 문서에 대한 해당 문장의 대표성(salience), 이전까지 생성된 요약과 해당 문장과의 중복도(novelty), 문장의 절대, 상대 위치에 따른 중요성을 수치화하고 이를 합하여 해당 문장이 요약으로 선택될 확률을 계산한다.

$$
\begin{aligned}
p(y_k = 1 \mid h_k, g_k, d) = \sigma(&W_c h_k \quad \#\ \text{content} \\
&+ h_k^T W_k d \quad \#\ \text{salience} \\
&- h_k^T W_r \tanh(g_k) \quad \#\ \text{novelty} \\
&+ W_{ap} P_k^a \quad \#\ \text{absolute position} \\
&+ W_{rp} P_k^r \quad \#\ \text{relate position} \\
&+ b) \quad \#\ \text{bias term}
\end{aligned} \tag{3}
$$

입력 문장은 2 계층의 RNN을 통해 벡터로 표현된다. 첫 번째 층에서는 개별 문장을 구성하는 단어들의 양방향 GRU 출력의 차원별 평균(element-wise average)을 취해 문장을 표현하는 벡터를 구성한다. 각 문장을 표현하는 벡터는 두 번째 GRU 층에 입력되어 주변 문장들과의 문맥 정보가 반영된 문장 벡터 h를 생성한다.

본 과제에서는 SummaRuNNer 모델에 단어의 표층 자질뿐만 아니라 기본적인 언어 분석 자질을 함께 사용하여 모델을 구축하였다. 영어 데이터를 기반으로 한 실험에서 단어의 표층 정보만 활용한 경우보다 언어 분석 자질을 함께 사용했을 때 더 나은 추출 요약 결과를 얻을 수 있었다. 본 과제에서 사용한 언어 분석 자질은 표 49와 같다.

feature	description	category
Surface	단어 표층 정보	number of vocab
TF	단어의 문서 내 빈도	0~50
POS	단어의 품사	number of POS tags
NE	단어의 named entity 여부	boolean
STOP	단어의 stop word 여부	boolean

표 49 추출 분석에 사용한 언어 분석 자질

단어의 품사 정보(POS)는 의미를 담고 있는 단어들과 문법적 역할을 수행하는 단어들을 구분 지을 수 있도록 도와준다. 단어의 빈도(TF)는 추출 문서 요약과 키워드 추출 관련 연구에서 중요한 자질로 사용되어 왔다. 문서 요약이 '누가', '무엇을', '언제' 등과 같이 개체명 정보를 중심으로 이루어지는 영향이 있기 때문에 개체명 여부를 자질로 추가하였다. 이와 같은 언어 분석 자질을 단어의 표층 정보와 함께 사용한다.

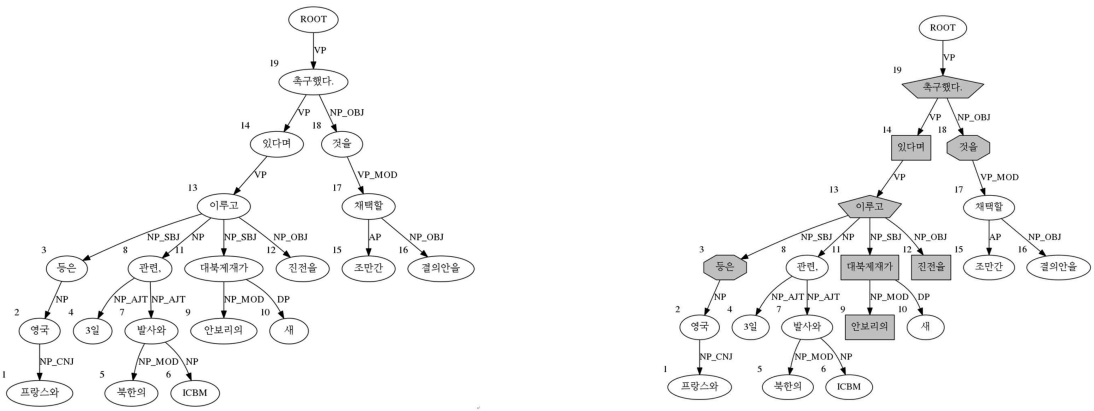
BERT(Bidirectional Encoder Representations from Transformer)는 대량의 말뭉치로부터 미리 학습된(pre-trained) 범용 언어 표현(general language representation) 모델이다. BERT는 기존의 word2vec과 같이 정적인 단어 표현이 아닌 양방향 주변 문맥 정보가 반영된 임베딩을 제공한다. 본 과제에서는 BERT를 문서 축약에 도입하였다. 문서를 표현하기 위한 단어 벡터를 BERT의 출력으로 대치하였다. 본 연구에서는 korBERT[1] 모델을 사용하였다.

2.2. 한국어 문장 축약 모델

본 과제에서는 한국어 문서 요약은 두 단계를 거쳐 수행된다. 첫 번째 단계는 문서로부터 중요한 3 문장을 추출하는 것이고, 두 번째 단계는 추출된 문장을 요약문처럼 축약하는 것이다. 이렇게 두 단계를 거쳐 요약문을 만드는 이유는 문서로부터 추출되는 2~3 문장이 요약문이라고 보기에는 지나치게 자세한 내용을 담고 있는 긴 문장인 경우가 많이 때문이다. 문장을 핵심 내용으로만 축약함으로써 간결한 요약을 생성할 수 있다.

문장 축약은 삭제 기반(deletion-based) 방식을 취한다. 즉, 원문의 핵심적인 부분만을 남기고 나머지 부분은 삭제함으로써 문장을 축약하는 방법이다. 한국어 문장 축약 모델을 구축하기 위해서는 원문-축약문으로 구성된 학습 말뭉치가 필요하다. 영어권에서는 이와 같은 용도로 만들어진 Google dataset이 존재하는 반면 한국어에는 이와 같은 말뭉치가 존재하지 않는다. 본 연구에서는 한국어 신문기사의 제목과 첫 문장을 이용하여 원문-축약문의 학습 말뭉치를 반자동으로 구축하였다. 구축 과정을 간략히 설명하면 다음과 같다.

(1) 신문기사의 제목과 첫 문장을 수집한다.
(2) 첫 문장을 구문분석하여 의존구문트리를 얻는다.
(3) 의존구문트리의 노드 중, 제목에 나타난 단어를 마킹한다.
(4) 마킹된 노드에서 ROOT 사이의 노드 중 마킹되지 않은 노드가 있다면 마킹한다.
(5) 마킹된 노드의 자식노드 중 필수격, 부정어가 마킹되지 않았다면 마킹한다.
(6) 마킹되지 않은 노드를 삭제하고 남은 노드로 축약문을 구성한다.



1) http://aiopen.etri.re.kr/service_dataset.php

(a) (b)

<그림 8> 문장 축약 과정

<그림 8>의 (a)는 원문에 대한 의존 구문 트리이다. (b)의 네모 노드가 제목에 나타난 단어를 마킹한 것이다. 축약문의 구조적 완결성을 위하여 네모 노드에서 ROOT 노드 중 마킹되지 않은 노드를 마킹한 것이 오각형 모양의 노드이다. 마킹된 노드의 자식 노드 중 필수격으로 마킹한 것이 육각형 모양의 노드이다.

1	**제목**	유엔 안보리의 대북제재가 진전을 보이고 있다.
	원문	프랑스와 영국 등은 3일 북한의 IBCM 발사와 관련, 안보리의 새 대북제재가 진전을 이루고 있다며 조만간 결의안을 채택할 것을 촉구했다.
	축약문	프랑스와 영국 등은 안보리의 새 대북제재가 진전을 이루고 있다며 결의안을 채택할 것을 촉구했다.
2	**제목**	공무원 한명 음주운전에 4급부터 9급까지 90명 사흘간 교육
	원문	전남 여수시 경제해양수산국 4급 국장부터 9급 말단 공무원까지 총 90여 명 전원이 사흘간 음주운전 예방 교육을 받게 됐다.
	축약문	4급 국장부터 9급 공무원까지 총 90여 명 전원이 사흘간 음주운전 예방 교육을 받게 됐다.
3	**제목**	檢 vs 양승태, '직권남용' 놓고 치열한 법리 공방
	원문	양승태 전 대법원장과 임종헌 전 법원행정처 차장이 사법농단과 관련한 직권남용 권리행사방해 혐의에 대해 "죄가 되지 않는다"라는 취지의 주장을 펼치며 치열한 법리 공방을 예고했다.
	축약문	양승태 전 대법원장과 임종헌 전 차장이 직권남용 권리행사방해 혐의에 대해 주장을 펼치며 치열한 법리 공방을 예고했다.
4	**제목**	서민 교수 "문재인 지지자분들께 사과드립니다" 사과문 게재
	원문	'기생충 전문가'로 알려진 서민 교수가 24일 자신의 블로그에 문재인 지지자들에 대한 사과문을 올렸다.
	축약문	서민 교수가 블로그에 문재인 지지자들에 대한 사과문을 올렸다.

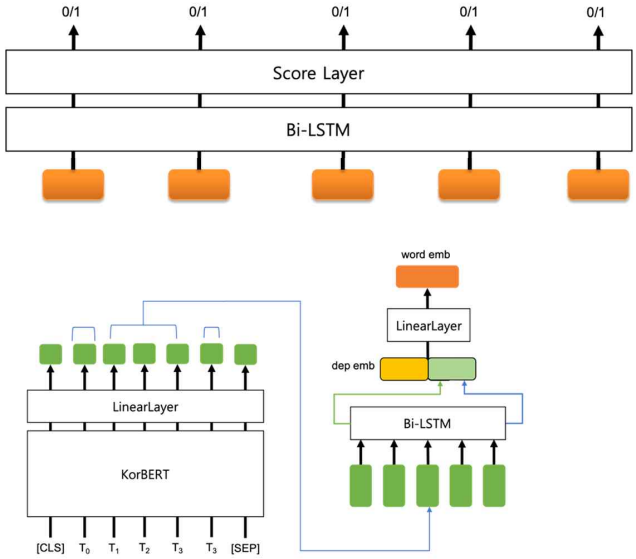
표 50 규칙을 이용하여 구축한 원문-축약문 예제

위의 단계를 거쳐 만든 축약문의 예제를 표 50에 제시하였다. 이와 같이 반자동으로 구축된 원문-요약문 150,000쌍의 학습 말뭉치를 이용하여 문장 축약 모델을 구축하였다. 앞서 구축한 한국어 축약 데이터를 이용한 문장 축약 모델은 다음과 같다. 축약 모델의 입력은 원문장의 언어 분석 결과인 구문트리정보이며 출력은 각 어절을 삭제할지(0) 유지할지(1)의 확률을 출력한다.

$$Y_i^p = Model(G_i) \qquad (4)$$

수식 (4)에서 G_i는 입력문장 S_i의 언어분석 결과인 구문트리 정보이며 $Y^p = \{y_1^p, y_2^p, ...y_k^p, ...y_n^p\}$는 n개 어절에 대한 축약 여부의 표지이다. 앞서 구축한 원문-요약

문 쌍으로부터 정답(oracle) 표지 $Y^o = \{y_1^o, y_2^o, ...y_k^o, ...y_n^o\}$을 알 수 있기 때문에 Y^o와 Y^p 차이를 통해 모델을 학습한다.



<그림 9> 문장 축약 모델

<그림 9>에 문장 축약 모델을 도식화하였다. BERT는 주변 문맥 정보를 고려한 언어 표현을 제공한다. BERT를 이용하여 입력 문장에 대한 임베딩 정보를 구한다. 구문 분석 트리 정보는 입력 어절을 단위로 어절의 부모 어절, 자식 어절, 관계 레이블로 표현할 수 있다. 본 연구에서는 이러한 어절 단위의 구문 정보를 모델의 입력으로 만들기 위하여 <그림 9>의 아래쪽 네트워크를 구축하였다. BERT의 입/출력 단위가 wordpiece이기 때문에 이를 하나의 어절로 묶어주고 동시에 해당 어절의 의존구문분석 트리 정보를 함께 추가하여 어절 입력을 표현한다. <그림 9>의 위쪽 네트워크는 어절 벡터로 표시된 문장을 입력으로 받고 문장의 문맥 정보를 반영하기 위해 Bi-LSTM layer를 거친 후 마지막 Score Layer를 거쳐 최종적으로 각 어절의 선택 확률을 계산한다.

2.3. 한국어 추출 문서 요약 및 축약 모델 실험 결과

실험에 사용한 데이터는 본 과제에서 구축한 요약 데이터로 그 특성은 다음 표 51과 같다.

	시범 데이터 집합	1차 데이터 집합	2차 데이터 집합
문서 수	400	1,335	2,671
본문 평균 문장 수	21.963	22.064	19.956
본문 문장 당 평균 어절 수	14.989	14.622	15.713
추출 요약의 문장 당 평균 어절 수	17.748	17.206	18.042
추상 요약의 문장 당 평균 어절 수	21.447	20.066	17.873
head+subhead 문장 수	1.52	1.732	2.841
head+subhead 문장 당 평균 어절 수	6.219	8.953	6.377

표 51 본 과제에서 구축한 요약 데이터 특성 (평가데이터)

첫째, 추출과 추상 요약 모두 문장 당 평균 어절수가 본문의 평균보다 상당히 많음을 볼 수 있다. 둘째, 초반에 구축된 시범 데이터와 1차 데이터의 경우 추상 요약이 추출 요약보다 더 긴 것을 볼 수 있다.

시범 데이터 집합과 1차 데이터 집합을 하나로 묶어 실험을 수행하고 2차 데이터 집합을 따로 실험을 진행하였다. 본 실험에서는 시범 데이터, 1~2차 데이터를 이용해 모델을 학습하지 않았고 평가 데이터로만 사용하였다.

학습 데이터로는 자체적으로 구축한 신문기사 말뭉치를 이용하였다. 신문기사의 제목, 부제, 그리고 소셜 미디어에서 제공하는 요약문을 이용하여 기사 내용의 요약문으로 사용하였다. 학습 데이터에 대한 특성은 표 52와 같다.

학습데이터	
기사 수	344,198 건
평균 문장 수	28.6문장
평균 어절 수	385.5 어절

표 52 학습 데이터 특성

실험 결과는 다음 표 53과 같다. 'Extraction'은 SummaRuNNer를 이용한 추출 요약 결과이며 'Extra+Comp'는 추출된 문서를 축약한 결과이다.

(형태소 단위)		ROUGE-1			ROUGE-2			ROUGE-L		
		F1	P	R	F1	P	R	F1	P	R
시범+1차 데이터 집합	Extraction	48.73	64.21	41.32	31.66	41.18	26.99	35.33	46.44	30.00
	Extra+Comp	44.48	67.06	34.72	26.76	40.06	20.92	32.18	48.51	25.11
2차 데이터 집합	Extraction	52.44	60.61	48.72	34.62	39.71	32.26	39.12	45.09	36,4
	Extra+Comp	48.86	63.94	41.33	29.97	38.98	25.39	36.2	47.29	30.63

표 53 추출 요약 + 문장 축약 실험 결과

2.4.한국어 문서 요약 말뭉치의 유용성에 대한 고찰

본 과제에서 수동 구축한 문서 요약 말뭉치의 유용성을 살펴보기 위하여 다음과 같은 실험을 수행하였다. 일반적으로 추출 문서 요약은 문서를 이루는 모든 문장에 대하여 각각의 문장을 요약에 포함시킬 것인지(레이블:1) 아닌지(레이블:0)를 결정하는 이진 분류(binary classification) 문제이다. 이와 같이 이진 분류 문제로 다루기 위해서는 정답 레이블이 주석되어 있는 말뭉치가 필요하다. 그러나 이러한 말뭉치가 없기 때문에 대부분의 경우 제목이나 부제를 이용하여 이와 가장 유사한 내용을 담고 있는 문장을 greedy 방식으로 추출하여 사용한다. 추출 과정은 다음과 같다.

1	summary = { }
2	(summary에 이미 추가된 문장을 제외하고) 문서의 각 문장을 summary에 추가해 보고 제목/부제와 ROUGE score를 계산
3	ROUGE score가 가장 높아지는 문장을 실제로 summary에 추가
4	ROUGE score가 떨어지지 않을 때까지 (2, 3) 단계를 반복

표 54 greedy 방식 추출 과정

다음 표에서는 본 과제에서 구축한 (1) 추출 정답 3 문장과 (2) 제목과 부제를 이용하여 greedy 방식으로 추출한 세 문장, (3) 기사 첫 세 문장을 사람이 작성한 요약과 비교하여 ROUGE 값을 측정해 보았다. 표 55에서 보듯이 본 과제에서 사람이 직접 태깅한 추출 정답-3 문장의 ROUGE 값이 가장 높게 나왔다. 즉, 추출 문서 요약의 정답 집합을 추출할 때, 기존에 하던 제목이나 부제를 이용한 것보다 사람이 직접 추출한 세 문장이 더 유사하였다. 그렇기 때문에 본 과제에서 구축한 추출 정답-3 요약문은 차후 추출 요약 시스템 개발이나 성능 평가에 유용한 데이터 집합으로 활용될 수 있다.

		(1) 추출 정답-3	(2) 제목/부제 활용 Greedy-3 추출	(3) Lead-3 추출
형태소	ROUGE-1	0.741558	0.364271	0.492308
	ROUGE-2	0.625889	0.227549	0.309258
	ROUGE-L	0.663689	0.277900	0.355039
내용어	ROUGE-1	0.718696	0.308413	0.410803
	ROUGE-2	0.602634	0.201509	0.279303
	ROUGE-L	0.663288	0.256995	0.332720

표 55 추출 문서 요약 시스템의 정답 레이블에 대한 성능 비교

또한, 제목/부제를 활용한 Greedy-3과 Lead-3이 사람이 직접 추출한 정답-3 문장과 얼마나 유사한지 살펴보기 위하여 추출 정답-3과의 문장 레이블에 대한 F-1 score값을 계산해 보았다. 표 56에서 보는 바와 같이 Greedy 방식으로 추출한 추출 문장이 약 23%의 정확도를 보였다. Greedy-3은 Lead-3보다도 낮은 성능을 보였는데 이는 추출 문서 요약 시스템 구축에서 제목이나 부제보다는 다른 정보를 활용하여 정답을 간접적으로 추정하는 것이 필요함을 보여준다.

	제목/부제를 활용한 Greedy-3 추출	Lead-3 추출
F-1 score	0.232	0.311

표 56 제목/부제를 활용한 추출 문서 요약의 레이블 성능

기존의 추출 문서 요약 시스템은 제목이나 부제를 활용하여 greedy 방식으로 추출한 세 문장을 정답으로 사용하여 학습하였다. 본 과제에서 구축한 추출 정답-세 문장의 유용성을 살펴보기 위하여 greedy 방식으로 추출한 약 344,000 문서의 학습 말뭉치에 추가로 본 과제에서 구축한 3,524 문서를 더하여 학습을 진행하였다. 추가한 3,524 문서의 정답 레이블을 추출하는 방식에 따라 실험을 나누어 수행하고 비교해 보았다.

	기본 학습 말뭉치	제목/부제를 활용한 Greedy-3 추출 약 344,000 문서		
	추가 학습 말뭉치 (3,524 문서)	(1) 추출 정답-3	(2) 제목/부제를 활용한 Greedy-3 추출	(3) Lead-3 추출
형태소	ROUGE-1	0.508217	0.505567	0.498410
	ROUGE-2	0.329134	0.324037	0.318725
	ROUGE-L	0.374887	0.371537	0.364018
내용어	ROUGE-1	0.435940	0.429446	0.417062
	ROUGE-2	0.298915	0.290725	0.286332
	ROUGE-L	0.356378	0.350263	0.339520

표 57 추출 문서 요약 시스템의 학습 레이블에 대한 요약 시스템의 성능 비교

평가는 학습에 사용하지 않은 882 문서에 대해 수행하고 그 ROUGE 값을 표 57에 서술하였다. 추가 학습 말뭉치가 기본 학습 말뭉치에 비해 약 1% 미만의 양이지만 추출 정답-3 문장으로 추출 문서 요약 시스템을 구축하였을 때 성능 향상을 살펴볼 수 있었다. 즉, 추출 문서 요약 시스템 개발 시 본 과제에서 구축한 정답 문장이 주석된 말뭉치를 학습데이터로 추가하였을 때 요약 시스템의 성능 향상을 기대할 수 있다.

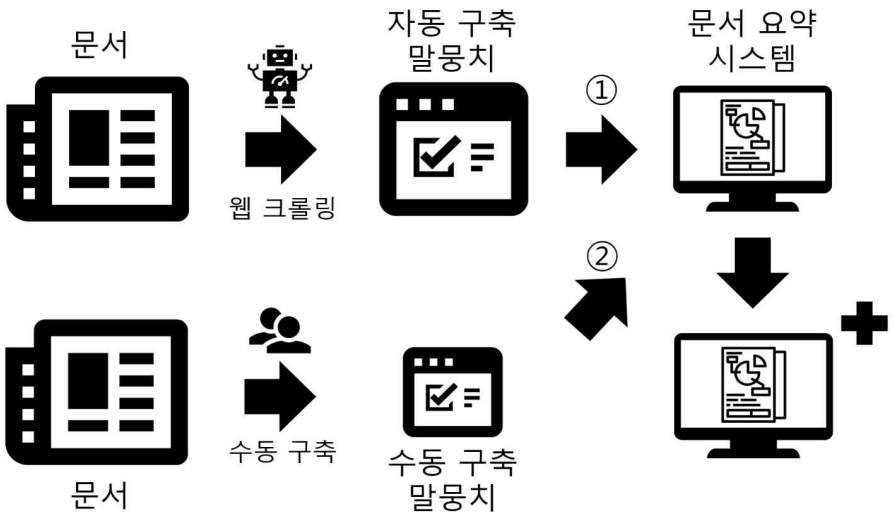
3. 추상 요약 기술 적용

3.1. 웹 크롤링을 통한 문서 요약 말뭉치 구축

심층 신경망 기반 시스템에서는 처리하고자 하는 문제에 대한 대량의 데이터를 확보하는 것이 필수적이다. 그러나 문서 요약을 비롯한 고수준의 자연어처리 응용 영역에서 신뢰도 높은 정답을 포함한 데이터를 대량으로 구축하는 것은 매우 어려운 일이다. 본 과제를 통해 구축된 한국어 요약 말뭉치(수동 구축 말뭉치)의 경우 고품질의 데이터이지만 요약 모델 학습에 활용하기 위해서는 그 양이 충분하지 않다고 볼 수 있다. 따라서 이를 보완하기 위해 웹에서 뉴스 기사를 크롤링하여 대량의 문서 요약 말뭉치(자동 구축 말뭉치)를 구축하였다. <그림 10>은 웹 크롤링 대상인 뉴스 기사의 예시이다.

<그림 10> 웹 크롤링 대상 뉴스 기사의 예

웹에 있는 일부 뉴스 기사들은 <그림 10>과 같이 제목 및 부제목을 포함한다. 제목, 부제목은 비교적 짧고 완전한 문장과는 다른 헤드라인 형태이지만 이것 또한 기사의 본문을 요약한다고 볼 수 있다. 따라서 제목과 부제목을 요약으로 하여 본문-요약 쌍으로 이루어진 말뭉치를 구축하였다. 자동 구축 말뭉치는 약 14만 쌍을 구축하였으며 <그림 11>과 같은 방법으로 활용하였다.



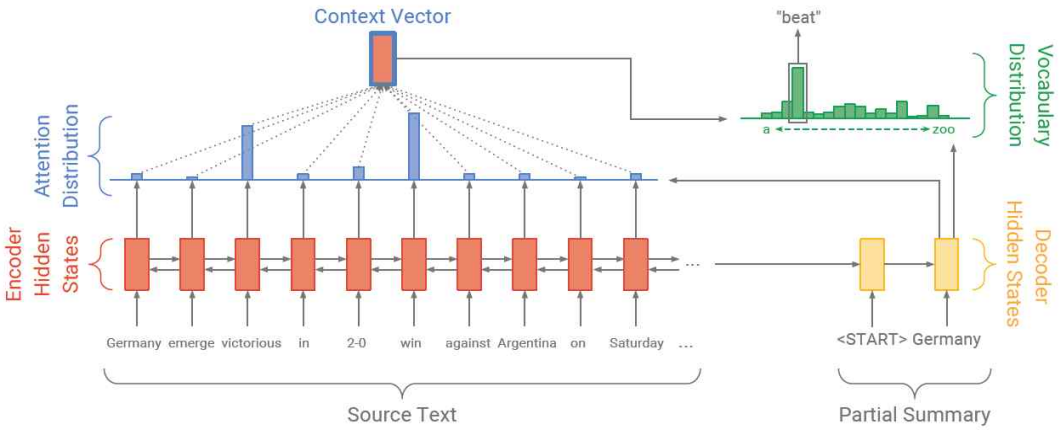
<그림 11> 문서 요약 시스템 학습 과정

<그림 11>과 같이 먼저 대량의 자동 구축 말뭉치를 이용해 문서 요약 시스템을 학습시킨 뒤 소량이지만 양질의 데이터인 수동 구축 말뭉치를 이용해 시스템을 추가 학습한다.

3.2. 추상 요약 모델 적용

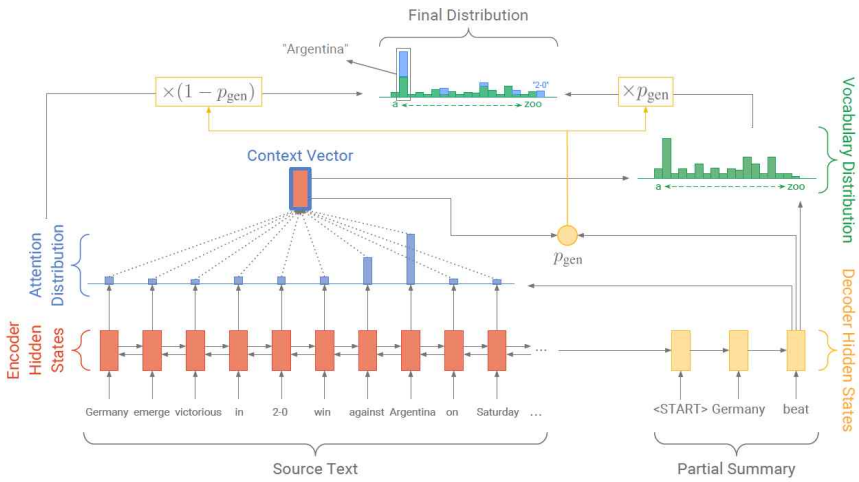
3.2.1. 포인터-제너레이터

문서 요약을 비롯한 자연어처리의 다양한 문야에서는 입력의 주요 어휘가 출력에 그대로 나타나야 하는 경우가 있다. 예를 들어, 채팅 시스템에서 “너 마마무 좋아해?” 라는 문장에 대한 응답으로 “응, 나 트와이스 좋아해.” 라는 답변은 잘못된 답변이 된다. 그러나 심층 신경망 기반 시스템에서 추상화된 어휘 정보들을 학습한 뒤 개념적으로 유사한 다른 어휘를 출력하는 문제는 꽤 빈번하게 일어난다. 문서 요약에서 고유명사와 같은 중요 어휘가 개념적으로 유사하지만 전혀 다른 개체를 의미하는 어휘로 바뀐다면 성능 하락뿐만 아니라 신뢰성을 떨어뜨리는 심각한 문제가 된다. 포인터-제너레이터(See 외, 2017)는 추상 요약 과정에서 빈번히 발생하는 유사 의미 단어 오생성 문제를 해결하기 위해 제안된 모델로 주의 집중 시퀀스-투-시퀀스(Sequence-to-sequence with attentions)를 기반으로 한다. 주의 집중 시퀀스-투-시퀀스의 구조는 <그림 12>와 같다.



<그림 12> 주의 집중 시퀀스-투-시퀀스 구조도

시퀀스-투-시퀀스는 두 개의 순환 신경망(recurrent neural network)을 각각 인코더와 디코더로 활용하는 모델로 한 시퀀스를 입력받아 다른 시퀀스를 출력해야 하는 다양한 분야에 사용된다. 시퀀스-투-시퀀스는 인코더를 통해 입력된 시퀀스를 압축된 형태로 표현한 문맥 벡터(context vector)를 만든다. 그리고 디코더를 통해 문맥 벡터를 순차적으로 디코딩하여 다른 시퀀스를 출력하는 방식으로 동작한다. 주의 집중 기법(attention mechanism)은 기계 번역 분야에서 처음 제안된 것으로 디코더가 출력 단어를 예측하는 매 시점마다 입력 시퀀스 중 가장 연관 있는 부분을 더 집중하여 반영하도록 하는 기법이다. 주의 집중 시퀀스-투-시퀀스는 기계 번역, 문서 요약뿐 아니라 다양한 분야에서 보편적으로 활용된다. 포인터-제너레이터는 이러한 주의 집중 시퀀스-투-시퀀스를 개선한 모델이며 모델 구조는 <그림 13>과 같다.



<그림 13> 포인터-제너레이터 구조도

포인터-제너레이터가 주의 집중 시퀀스-투-시퀀스와 다른 점은 <그림 13>과 같이 주의 집중 기법을 통해 계산된 입력 단어들(원본 문서에 포함된 단어들)의 분포와 디코딩 과정에서 생성되는 단어들의 분포를 결합하여 최종 요약에 포함될 단어를 생성한다는 것이다. 이를 수식으로 표현하면 아래 수식 (5)로 나타낼 수 있다.

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$
$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t \quad (5)$$

즉, 주의 집중 분포를 바탕으로 계산된 문맥 벡터 h_t^*, 디코더 상태 s_t, 디코더 입력 x_t를 바탕으로 게이트 함수(gate function) p_{gen}을 계산하고, 주의 집중을 통해 계산된 입력 단어 분포와 디코더의 단어 생성 분포를 p_{gen}에 따라 가중합하는 방식으로 요약문에 포함될 최종 단어 단어를 생성한다. CNN/Daily Mail 말뭉치를 이용한 실험에서 포인터-제너레이터의 성능은 아래 표 58과 같다.

Models	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	40.34	17.70	36.57
Seq2Seq+Attention (150k vocab)	30.49	11.17	28.08
Seq2Seq+Attention (50k vocab)	31.33	11.81	28.83
Pointer-generator	36.44	15.66	33.42
Pointer-generator + Coverage	39.53	17.28	36.38

표 58 포인터 제너레이터의 성능 비교

표 58과 같이 포인터-제너레이터는 주의 집중 시퀀스-투-시퀀스 모델과 비교하여 월등히 향상된 성능을 보였다. 또한, 동일 단어 반복 출력을 제어하기 위한 커버리지 기법(coverage mechanism)을 적용하여 추가적인 성능 향상을 보였다. 이러한 결과를 통해 포인터-제너레이터는 추상 요약을 대표하는 모델로 자리매김했으며, 최근의 문서 요약 연구에서 성능 비교의 대상이 되고 있다.

3.2.2. 한국어 요약 말뭉치를 활용한 포인터-제너레이터 실험

위의 내용을 바탕으로 한국어 요약 말뭉치를 이용해 포인터-제너레이터를 학습 및

평가하였다. 모델 학습에는 가장 보편적인 입출력 형태인 형태소 단위를 사용하였고 약 11만 8천여 개의 자동 구축 말뭉치를 사용하여 학습하였다. 그 뒤 1,335개의 수동 구축 말뭉치를 사용하여 추가 학습하였고, 논문과 동일하게 커버리지 기법을 적용하였다. 평가에는 학습에 사용되지 않은 400개의 수동 구축 말뭉치를 사용하였다. 평가 결과는 아래 표 59와 같다.

	ROUGE-1	ROUGE-2	ROUGE-L
자동 구축 말뭉치	0.2764	0.1137	0.1669
+ 수동 구축 말뭉치	0.4586	0.2701	0.3572
+ 커버리지 기법	0.5184	0.3154	0.4356

표 59 형태소 단위 포인터-제너레이터 평가 결과

평가는 요약 모델 평가에 주로 사용되고 있는 ROUGE를 이용하였다. ROUGE는 모델 출력이 얼마나 정답에 가깝게 재현되었는지를 나타내는 지표이며, ROUGE-1은 유니그램, ROUGE-2는 바이그램, ROUGE-L은 최장 공통 부분열(LCS:Longest Common Subsequence)의 재현율을 나타낸다. 평가 결과, 자동 구축 말뭉치만으로 학습한 뒤 평가했을 때보다 수동 구축 말뭉치로 추가 학습을 진행하였을 때 모델의 성능이 향상된 것을 확인할 수 있다. 자세한 출력은 아래 표 60, 표 61과 같다.

자동 구축 말뭉치	+ 커버리지 기법
“ 한국 주식 투자 늘리 ㄹ 것 ” 버핏 *버크셔해서웨이 회장 “ 한국 은 위기 에 잘 대응 ” 3 일 한국 기업 주식 투자 늘리 기 위하 아 “ 위기 대응 ” “ 한국 은 위기 에 잘 대응 … 몇 년 간 매우 잘 하 아 나가 ㄹ 것 ”	투자 가 이루어지 ㄴ 신흥 시장 이 ㄴ 나라 , 한국 은 위기 에 잘 대응 하 고 있 고 한국 은 위기 에 잘 대응 하 고 있 기 위하 아 바쁘 ㄹ 것 이라고 말 하 았 다 . 그 는 3 월 포스코 가 발행 하 ㄴ 채권 을 추가 로 매입 하 ㄴ 것 도 하 았 다 . 그 는 또 적당 하 ㄴ 가격 (*right *price) 에 살 ㄴ다면 많 은 수익 을 올리 ㄹ 것 으로 알리 어 지 었 다 .
‘ 경북 관광 ’ 알리 는 42 인치 TV 설치 하 시 어요 일본 , 경주 · 안동 , 영주 등 10 분 짜리 홍보 영상물 ‘ 경북 관광 ’ 알리 는 42 인치 TV 설치	일본 규슈 지역 의 대표 적 항구 이 ㄴ 하 카 타 (*博多) 항의 국제 여객선 터미널 에 최근 ‘ 경북 관광 ’ 을 알리 는 42 인치 TV 가 설치 되 었 다 . 경북도 는 지난해 11 월 후쿠오카 시 를 방문 하 아 오 는 일본인 이 좋아하 ㄹ 수 있 도록 관광 정보 를 알리 는 전광판 을 설치 하 고 있 다 .
파키스탄 정부 시위 앞두 고 파키스탄 정국 이틀째 반 정부 시위 앞두 고 파키스탄 무 슬림 하 ㄴ	파키스탄 정부 가 16 일 *이슬라마바드 의회 앞 에서 열리 ㄹ 예정 이 ㄴ 대 규모 반 정부 시위

가택 연금 명령 법관 복직 시키 ㄹ ‘ 재 기소 ’ 개입 하 았 다는 파키스탄 정부 “ 공직 선거 출마 금지 판결 ”	를 앞두 고 *나와즈 샤리프 와 *거국 미 의 고위 당직자 를 비롯 하 아 야당 지도자 을 을 가택 연금 하 아 파키스탄 정국 이 혼미 하 아 지 고 있 다 . 파키스탄 정부 는 야당 인사 400 여 명 을 검거 하 ㄹ 등 대 규모 반 정부 시위 에 나서 었 다 .

표 60 포인터-제너레이터 출력 결과 (* : 미등록어)

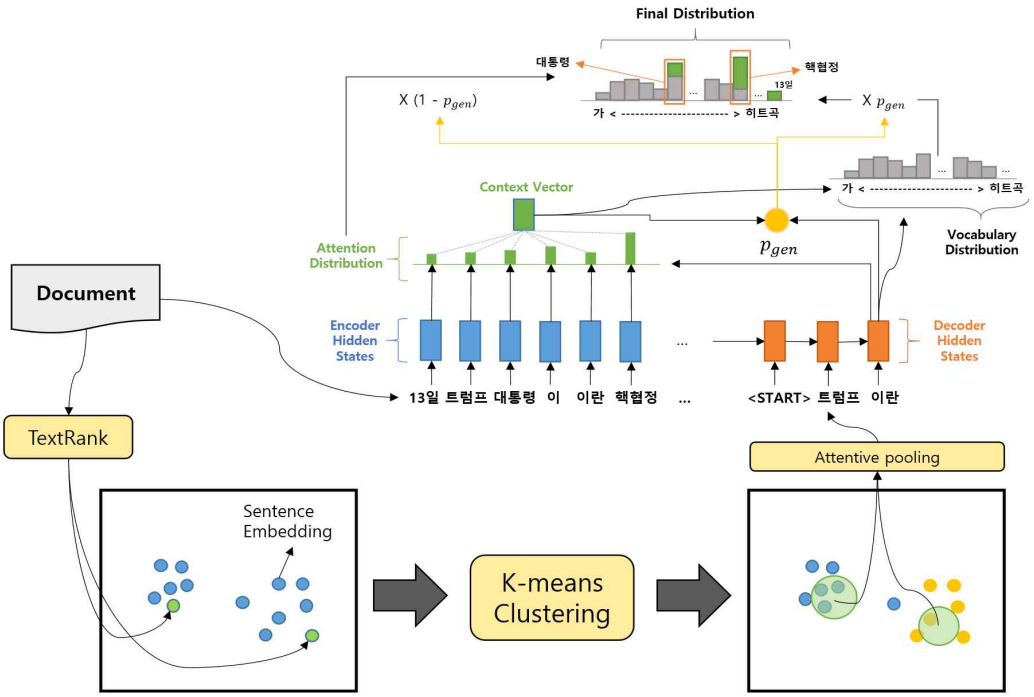
+ 수동 구축 말뭉치	+ 커버리지 기법
현대 미술 의 전당 *한가람미술관 에서 ‘ 동화책 속 세계 여행 ’ (02 - 960 - 585 - *9991) 6 월 23 일 까지 , 6 월 23 일 까지 ‘ 동화책 속 세계 여행 ’ (02 - 960 - 585 - *9991) 6 월 23 일 까지 , 6 월 23 일 까지 ‘ 동화책 속 세계 여행 ’ (02 - 960 - 585 - *9991) 6 월 23 일 까지 , 6 월 23 일 까지 ‘ 동화책 속 세계 여행 ’ (02 - 960 - 585 - *9991) 에서	6 월 23 일 까지 서울 서초구 예술 의 전당 *한가람미술관 에서 ‘ 동화책 속 세계 여행 (02 - 960 - 585) ’ 에서 ㄴ 인기 작가 앤서니 브라운 의 ‘ 돼지 책 ’ 원화 를 만나 ㄹ 수 있 는 세계 여행 (02 - *9991) 에서 ㄴ 인기 작가 65 명 의 원화 및 폐품 을 활용 하 ㄴ 설치 미술 과 *손뜨개를 등 다양 하 ㄴ 작업 을 선보이 ㄴ다 .
수도 권 에서 대형 건설사 들 의 경쟁 이 다시 불붙 고 있 다 . 한국건설산업연구원 은 서울 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ 결과 주택 경기 실사 지수 (*BSI) 를 조사 하 ㄴ	부동산 시장 에 회복 조짐 을 보이 는 가운데 수도 권 에 재 개발 , 재 건축 사업 물량 을 따 아내 려는 대형 건설사 들 의 경쟁 이 다시 불붙 고 있 다 . 이 는 시장 침체 로 사업 을 부진 하 아 쟁점 을 알 고 있 다 .

표 61 커버리지 기법 적용 결과 (* : 미등록어)

표 60의 결과에서 자동 구축 말뭉치만으로 학습한 모델은 뉴스 기사의 제목과 같이 짧고 축약된 형태의 결과를 출력하는 것을 확인할 수 있다. 이는 자동 구축 말뭉치가 뉴스 기사의 제목 및 부제목을 요약으로 하여 구축되었기 때문이다. 반면 수동 구축 말뭉치를 이용해 추가적인 학습을 진행한 후의 모델은 보다 긴 문장 형태로 결과를 출력하였다. 수동 구축 말뭉치를 이용하여 추가 학습을 진행한 뒤 성능이 크게 향상된 것은 이처럼 말뭉치 간의 형태 차이가 큰 영향을 미친 것으로 파악된다. 표 61은 출력에서 나타난 반복 문제와 커버리지 기법을 통해 반복 문제를 해결했을 때의 결과 비교이다. 커버리지 기법 적용 전 출력에서는 동일한 시퀀스가 반복해서 나타나는 문제가 있었지만 커버리지 기법 적용 후에 이러한 부분들이 일부 해결되었다. 또한, 전체 결과를 통해 포인터-제너레이터는 기존의 시퀀스-투-시퀀스 모델이 출력할 수 없는 미등록어들을 잘 출력해내는 것을 볼 수 있다.

3.2.3. 추출 요약과 군집화를 이용한 문맥 추가 실험

포인터-제너레이터가 문서 전체에서 문장의 구분 없이 중요한 정보를 학습하는 반면 추출 요약은 문서 전체에서 문장 단위로 중요한 부분을 학습한다. 따라서 추출 요약을 통해 중요 문장을 선별한 뒤 포인터-제너레이터에 추가한다면 모델이 문장 단위의 중요 정보를 기반으로 요약을 생성하는 것을 기대할 수 있다. 이에 따른 변화를 확인하기 위해 아래의 <그림 14>와 같은 실험을 진행하였다.



<그림 14> 추출 요약과 군집화를 활용한 문맥 벡터 추가

먼저, 문서를 문장 단위로 분할한 뒤 텍스트랭크(TextRank)를 이용하여 중요 문장을 추출한다. 이 때 각 문장의 중요 단어(체언류, 용언류)의 빈도수를 통해 중요도를 계산하며, 가장 중요도가 높은 것으로 계산된 세 문장을 추출한다. 그 뒤, 텍스트랭크를 통해 추출된 문장들의 벡터를 초기 군집의 중심으로 하여 문장 단위로 K-평균 군집화(K-means clustering)를 수행한다. 아래 표 62는 K-평균 군집화를 수행한 뒤의 출력 예시이다.

	문서
군집 1	- 제조사별로 보면 삼성전자의 국내 단말기 판매가격은 평균 508달러로, 해외 평균 223달러보다 2.3배 높았다. - 엘지전자는 국내 판매가격은 평균 361달러인 반면, 해외는 평균 176달러로 2.1배 높았다. - 제품이 대부분 프리미엄폰인 애플은 2배 차이를 보이지는 않았지만, 국내 판매가가 45달러 높은 것으로 나타났다.
군집 2	- 이에 따라 2015년부터 올해 2분기까지의 국내 단말기 평균 판매가격(ASP・전체 단말기 매출을 출하량으로 나눈 수치)은 514달러(약 58만5천원)로 해외 단말기 평균가격(197달러, 약 22만4천원)보다 2.6배 높은 수준인 것으로 나타났다. - 삼성과 애플의 단말기를 사용하는 소비자는 전체 평균보다 높은 수준의 단말기 할부금을 지출하고 있었다. - 변재일 의원은 “제조사들이 해외에서는 유틸리티폰 등 저가폰을 많이 판매하는 반면 국내에서는 프리미엄폰 위주의 단말기 판매 전략을 펴고 있어 국내의 평균 단말 판매 가격이 높은 것으로 분석된다”며 “높은 단말기 가격이 가계 통신비 부담을 키우고 있는 만큼, 저가의 단말기 보급을 확대해 국민의 단말기 선택권을 확대시키고 저렴한 단말기를 사용할 수 있는 환경을 마련해야 한다”고 말했다.
군집 3	- 변재일 더불어민주당 의원이 10일 발표한 시장조사기관 가트너의 보고서를 보면, 지난해 4분기 기준 해외의 프리미엄폰 시장 비중은 약 32%인 반면, 국내는 87.9%에 달했다. - 가트너는 단말기를 프리미엄폰, 베이직폰, 유틸리티폰으로 구분했는데, 전세계적으로는 각각 시장의 3분의 1 정도씩 차지한 반면, 우리나라는 프리미엄폰이 월등히 높고, 유틸리티폰(3~4만원대) 시장은 아예 없었다. - 변재일 의원이 녹색소비자연대와 함께 지난달 12~22일 성인 1천명을 대상으로 온라인 인식조사를 진행한 결과, 응답자의 87.4%가 엘티이(LTE) 스마트폰을 이용하고 있다고 대답했다.

표 62 K-평균 군집화 결과 예시

표 62의 문장들은 군집의 중심 벡터와 문장 임베딩 벡터의 코사인 유사도를 계산하여 군집별로 가장 가까운 세 문장을 출력한 것이다. 군집 1은 주로 제조사별 단말기 가격 비교에 대한 내용이며 군집 2는 국내 평균 단말 가격에 대한 분석이다. 군집 3의 문장들은 스마트폰 사용자들이 사용하는 스마트폰의 종류에 대한 내용이다. 이처럼 각 군집에는 일관성 있는 내용들끼리 모이게 되므로 군집의 중심 벡터를 문서 내 특정 주제에 대한 문맥으로 활용할 수 있다. 문맥은 각각의 군집 중심 벡터들을 attentive pooling한 뒤 이를 디코더 입력에 추가하는 방식으로 활용한다. 문맥을 추가하였을 때의 성능은 아래 표 63과 같다.

	ROUGE-1	ROUGE-2	ROUGE-L
기존 모델	0.3230	0.1304	0.2506
+ 문맥 벡터	0.3206	0.1276	0.2483

표 63 문맥 벡터 추가에 따른 성능

표 63에서는 예상과 달리 군집화를 이용한 문맥을 모델에 추가하는 경우 성능이 감소한 것을 확인할 수 있다. 출력된 결과에서 크게 달라진 부분을 확인할 수는 없었으나 군집화 결과에 크게 문제가 없음을 감안하였을 때, 군집의 중심 벡터를 이용해 문맥 벡터를 생성하는 방법과 생성된 문맥 벡터를 모델에 입력하는 방법에서 더 좋은 방법을 구상해야 할 것으로 보인다.

3.2.4. 품사 게이트 자질 추가 실험

포인터 제너레이터는 주의 집중 분포를 통해 디코더에서 출력한 사전 분포를 보정하여 입력에 나타난 중요 단어가 출력에 복사되도록 하는 모델이다. 그러므로 중요한 단어가 출력에 잘 복사되도록 하려면 모델이 주의 집중 분포를 잘 생성할 수 있어야 한다. 중요한 단어는 형태소 분석 시 일반명사(NNG), 고유명사(NNP), 분석불가(NA) 등과 같이 특정 품사로 나타날 것이라는 가정 하에 입력의 각 토큰마다 품사 정보를 자질로 추가하여 주의 집중 분포를 생성한다면 모델이 출력에 복사되어야 할 중요한 단어를 더 잘 학습할 것으로 기대하였다. 이에 따른 실험은 아래 수식 (6), (7)과 같이 진행하였다. 수식 (6)은 기존의 포인터-제너레이터가 주의 집중 분포를 계산하는 방법이며, 수식 (7)은 품사 정보를 추가하여 주의 집중 분포를 계산하는 방법이다.

$$
\begin{aligned}
h_i &= BiRNN(x_i^{word}) \\
e_i^t &= \mathrm{v}^{\mathrm{T}}\tanh(\mathrm{W_h h_i} + \mathrm{W_s s_t} + \mathrm{b_{attn}}) \\
a^t &= \mathrm{softmax}(\mathrm{e^t})
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
h_i &= BiRNN(x_i^{word}) \\
g_i &= \sigma\,(W_j[h_i; feature_i]) \\
z_i &= g_i * feature_i \\
h'_i &= FNN([h_i; z_i]) \\
e_i^t &= \mathrm{v}^{\mathrm{T}}\tanh(\mathrm{W_h h'_i} + \mathrm{W_s s_t} + \mathrm{b_{attn}}) \\
a^t &= \mathrm{softmax}(\mathrm{e^t})
\end{aligned}
\tag{7}
$$

(7)의 $feature_i$는 단어 x_i^{word}에 대한 품사 벡터이다. 인코더 상태 h_i는 품사 게이트 자질 z_i와 합쳐진 뒤 FNN에 입력되어 새로운 인코더 상태 h'_i를 만들고, 이를 이용해 품사 정보가 반영된 주의 집중 분포를 생성한다. 품사 정보 추가에 따른 모델의 성능은 아래 표 64와 같다.

	ROUGE-1	ROUGE-2	ROUGE-L
기존 모델	0.3230	0.1304	0.2506
+ 품사 정보	0.3123	0.1194	0.2400

표 64 품사 정보 추가에 따른 성능

표 64의 결과에서는 품사 정보를 추가했음에도 불구하고 요약 결과의 성능이 하락한 것을 확인할 수 있다. 이 결과는 현재 사용하고 있는 입출력 단위가 형태소/품사의 형태이기 때문에 입력 자체가 품사 정보를 어느 정도 반영하고 있어서 추가된 정보가 자질로서 잘 활용되지 못한 것으로 보인다.

3.2.5. 입출력 단위 변경 실험

이전까지의 실험에서는 한국어에서 가장 보편적인 입출력인 형태소 단위로 실험을 진행하였다. 그러나 형태소의 경우 모델에서 사용하는 단어 사전의 크기 제한으로 인해 미등록어 문제가 발생할 수 있고 형태소 분석을 위한 형태소 분석 모듈 및 모델의 출력 결과를 복원하기 위한 형태소 복원 모듈이 필요하다는 등의 단점이 있다. 따라서 형태소 단위 외에 다른 입출력 단위를 사용한 실험을 진행하였다. 형태소 외에 가장 먼저 사용한 것은 음절 단위의 입출력이다. 음절은 형태소에 비해 단어 사전의 크기가 작아서 미등록어 문제에서 자유롭고 추가적인 모듈이 불필요하다는 장점이 있다. 아래의 표 65는 음절 단위로 모델을 학습하고 출력한 결과이다.

정답 요약	출력 결과
- 김 위원장, 문 대통령에 “하나도 숨차 안 하십니다” - 문 대통령 “네, 뭐, 이 정도는” 태연한 반응에 웃음꽃 - 리 여사 농담에 김정숙 여사 “정말 얄미우십니다” 맞장구	국무위원장 부인 리설주 여사 웃으며 건넨 말 “정말 얄밉다”

표 65 음절 단위로 출력한 결과 예

표 65는 실제 문서 크기를 고려하지 않고 모델 크기를 제한적으로 적용했을 때의 출력 결과이다. 포인터-제너레이터가 음절 단위에서도 괜찮은 출력을 내는 것을 보여주지만 음절 단위의 경우 인코더와 디코더의 크기를 형태소 단위에 비해 아주 크게 설정해야 하며, 문서 전체를 고려할 경우에는 학습이 거의 불가능하다. 따라서 음절 단위의 입출력은 사용하지 않는 것으로 결정하였다. 그 외의 입출력으로 워드피스

(wordpiece) 단위를 사용하였다. 워드피스는 모든 단어를 빈도수에 따른 부분 단어로 치환하여 표현하는 것이다. 워드피스도 음절과 마찬가지로 형태소에 비해 더 작은 크기의 단어 사전을 사용하므로 미등록어 문제에서 비교적 자유롭다는 장점이 있으며 복원에도 추가적인 모듈이 필요한 형태소와 달리 워드피스 단위는 출력 결과를 문장의 형태로 복원하는 것이 훨씬 간편하다는 장점이 있다. 또한 음절보다는 토큰의 크기가 크기 때문에 문서 전체를 고려하여 학습을 진행하기 어려운 음절과 달리 워드피스는 문서 전체를 고려하여도 적당한 인코더와 디코더 크기로 학습을 진행할 수 있다. 평가를 위해 약 11만 8천여 개의 자동 구축 말뭉치를 사용하여 워드피스 단위로 학습한 뒤 4,006개의 수동 구축 말뭉치를 사용하여 추가 학습하였고, 그 결과는 아래 표 66과 같다.

	자동 구축 평가 말뭉치			수동 구축 평가 말뭉치		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
자동 구축 말뭉치	0.3275	0.1734	0.2523	0.2397	0.1273	0.1452
+ 수동 구축 말뭉치	0.2598	0.1158	0.1886	0.4531	0.3378	0.3923
+ 커버리지 기법	0.2537	0.1103	0.1752	0.4692	0.3429	0.4124

표 66 워드피스 단위 실험 성능

수동 구축 평가 말뭉치를 이용한 성능에서는 형태소 단위와 마찬가지로 성능 향상이 이루어지는 것을 확인할 수 있다. 자동 구축 평가 말뭉치 상에서 성능이 하락하는 것으로 나타나는 것은 자동 구축 말뭉치와 수동 구축 말뭉치 간의 정답 형태의 차이 때문인 것으로 보인다. 자동 구축 말뭉치의 정답이 제목과 같은 짧은 형태인 반면 수동 구축 말뭉치의 정답은 문장의 형태로 이루어졌기 때문에 수동 구축 말뭉치를 통한 학습 후에는 모델이 문장의 형태로 더 긴 요약을 출력하기 때문이다. 출력된 요약 결과는 아래 표 67에서 확인할 수 있다.

	수동 구축 평가 말뭉치
원본 문서	1997년 한글소설본이 발견돼 ‘홍길동전’을 밀어내고 최초의 한글소설 자리를 꿰찬 ‘설공찬전’은 조선 중종조에 이미 금서(禁書)가 된 작품이다. 명분은 혹세무민하는 혼백의 세계를 다뤘다는 것이었다. 저승세계를 소개하면서 현실 권력을 매섭게 풍자한 대목이 연산군을 몰아내고 권력을 잡은 중종 시대 권력층의 역린을 건드렸기 때문이다. 극단 ‘신기루만화경’의 연극 ‘설공찬전’(이해제 작, 연출)은 이 고전소설에서 저승세계에 대한 소개는 빼고 권력의 생리를 비판한 내용을 전면에 부각했다. 요절한 수재 설공찬(황도연)은 죽은 지 3년 뒤 아버지 설충란(임진순)에게 못다 한 효도를 하려고 망나니 사촌동생 설공침(정재성)의 몸에 강림한다. 여기서 작품은 공침의 몸을 빌려 뮤지컬 ‘지킬 앤드 하이드’식의 선악 대결을 한국적 해학으로 녹여낸다. 정재성씨는 점잖은 공찬과 패악스러운 공침이 한 몸을 두고 벌이는 입씨름과 몸씨름을 능청스럽게 풀어낸다. 당대 실세인 정익로(이장원) 대감을 구워삶아 아들에게 관직의 길을 열어주려는 공침의 아비 설충수(최재섭)는 죽은 공찬이 아들의 몸에 들어왔다는 사실을 알고도 이를 묵인한다. 공침의 몸을 빌린 공찬이 정 대감 앞에서 감춰뒀던 자신의 경륜을 펼치려는 순간, 정 대감은 다음과 같은 질문으로 그의 입을 막는다. “세상엔 많은 문답이 있다. …그 물음이 어디에서 흐르느냐에 따라 그 대답의 방법이 엄연히 달라지는 것. 그것이 세상 살아가는 법이다. 자, 중천에 해가 떠 있다. 내 눈엔 저 해가 네 개의 모가 있는 바둑판으로 보이는데 자네의 눈엔 어떻게 보이는가.” 한때 세상을 내려다보긴 같지만 스스로 몸을 드러내는 해와 몸을 감추는 달은 서로 다른 족속이라고 일갈했던 공찬은 이 문답을 통해 권력의 본질을 깨닫고 이 몸 저 몸으로 옮겨 다니며 한바탕 놀이를 펼친다. 작품은 정권교체기 ‘영혼 없는 공무원들’의 행태를 수없이 목도하는 요즘의 현실에 따끔한 일침을 놓는다. 중국어 발음으로 한자성어를 남발하는 정 대감은 어색한 영어 발음으로 세간의 비웃음을 사던 지식인들을 연상시킨다. 요즘 세태에 대해 직접적이면서도 매서운 비판의 수위를 높인다면 더욱 호응이 뜨거울 작품이다. 서울 대학로 정보소극장에서 2월 8일까지. 2만 원. 02-764-7462
정답 요약	서울 대학로 정보소극장에서 2월 8일까지 공연되는 극단 ‘신기루만화경’의 연극 ‘설공찬전’(이해제 작, 연출)은 권력의 생리를 비판한 내용을 부각했다. 중국어로 한자성어를 남발하는 정익로 대감은 어색한 영어 발음으로 세간의 비웃음을 사는 지식인들을 연상시키고, 정권교체기 ‘영혼 없는 공무원들’의 행태를 수없이 목도하는 요즘의 현실에 따끔한 일침을 놓는 연극이다. 1997년 한글소설본이 발견돼 ‘홍길동전’을 밀어내고 최초의 한글소설로 인정받은 ‘설공찬전’은 혹세무민하는 혼백의 세계를 다뤘다는 이유로 조선 중종조에 이미 금서(禁書)가 된 작품이다.
자동 구축 말뭉치	’홍길동전’ 밀어내고 한글소설 설공찬·황도연 ’설공찬전’ 이해제 작·역린…’신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극 ‘신기루만화경’ 연극
+ 수동 구축 말뭉치	‘홍길동전’을 밀어내고 최초의 한글소설 자리를 꿰찬 ‘설공찬전’은 ‘홍길동전’을 밀어내고 최초의 한글소설 자리를 꿰찬 ‘설공찬전’은 조선 중종조에 이미 금서(禁書)가 된 작품이다. 극단 ‘신기루만화경’의 연극 ‘설공찬전’(이해제 작, 연출)은 이 고전소설에서 저승세계에 대한 소개는 빼고 권력의 생리를 비판한 내용을 전면에 부각했다. 극단 ‘신기루만화경’의 연극 ‘설공찬전’(이해제 작, 연출)은 이 고전소설에서 저승세계에 대한 소개는 빼고 권력의 생리를 비판한 내용을 전면에 부각했다.
+ 커버리지 기법	홍길동전’을 밀어내고 최초의 한글소설 자리를 꿰찬 ‘설공찬전’은 조선 중종조에 이미 금서(禁書)가 됐다. 극단 ‘신기루만화경’의 연극 ‘설공찬전’은 이 고전소설에서 저승세계에 대한 소개는 빼고 권력의 생리를 비판한 내용을 전면에 부각했다. 작품은 요즘 세태에 대해 직접적이면서도 매서운 비판의 수위를 높인다면 더욱 호응이 뜨거울 작품이다.

표 67 워드피스 단위 요약 출력 결과

3.2.6. 워드피스 단위 품사 게이트 자질 추가 실험

워드피스 단위 포인터-제너레이터에 품사 게이트 자질을 추가하여 성능을 비교하였다. 형태소 단위와 달리 워드피스의 경우 토큰 자체가 품사를 가지지 않으므로, 각 워드피스의 첫 음절과 끝 음절 품사를 통합하여 해당 워드피스의 품사로 사용하였다. 품사 게이트 자질을 추가한 성능은 아래 표 68 과 같다.

	수동 구축 평가 말뭉치					
	품사 자질 X			품사 자질 O		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
자동 구축 말뭉치	0.2397	0.1273	0.1452	0.2333	0.1209	0.1387
+ 수동 구축 말뭉치	0.4531	0.3378	0.3923	0.4649	0.3448	0.4081
+ 커버리지 기법	0.4692	0.3429	0.4124	0.4949	0.3662	0.4498

표 68 품사 게이트 자질 추가에 따른 성능 (ROUGE)

표 68 을 통해 워드피스 단위의 모델에서 품사 게이트 자질을 추가했을 때 성능이 향상되는 것을 확인할 수 있다. 또한, 어휘 재현율을 평가하는 ROUGE 점수 외에 정답과의 유사도를 고려한 성능을 확인하기 위해 정답 요약과 모델 출력의 코사인 유사도를 비교하였고 결과는 표 69 와 같다.

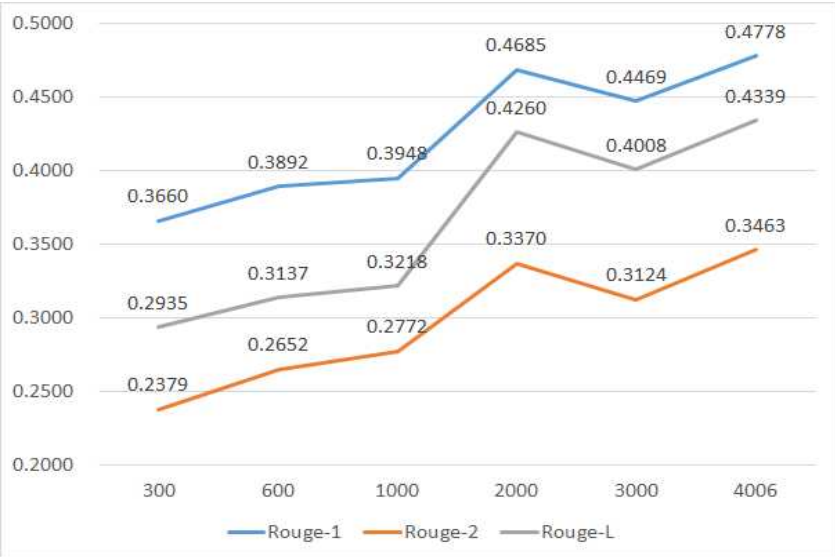
	수동 구축 평가 말뭉치					
	품사 자질 X			품사 자질 O		
	Greedy	Average	Extrema	Greedy	Average	Extrema
자동 구축 말뭉치	0.6767	0.6977	0.2942	0.6741	0.6786	0.2761
+ 수동 구축 말뭉치	0.7715	0.8371	0.4568	0.7798	0.8348	0.4642
+ 커버리지 기법	0.7788	0.8629	0.4658	0.7919	0.8752	0.4747

표 69 품사 게이트 자질 추가에 따른 성능 (임베딩)

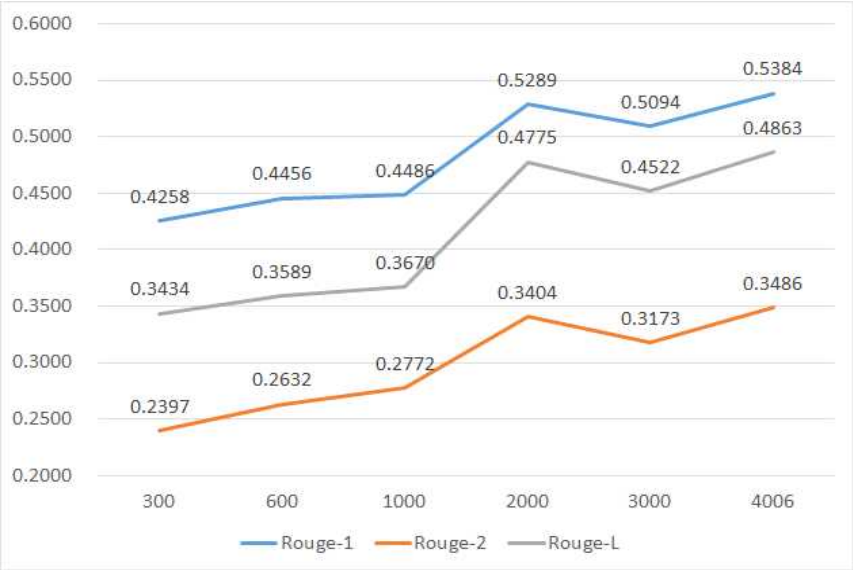
표 69 의 ‘Greedy’ 는 출력과 정답의 각 토큰 중 가장 유사한 토큰들끼리 코사인 유사도를 계산하여 평균한 값이다. ‘Average’ 는 출력과 정답의 모든 토큰의 벡터를 평균한 뒤 코사인 유사도를 계산한 값이다. ‘Extrema’ 는 토큰 임베딩의 극값들로 벡터를 만들고 이것을 코사인 유사도로 비교한 결과이다. 임베딩의 코사인 유사도 평가에서도 품사 게이트 자질을 추가했을 경우가 그렇지 않은 경우보다 높게 측정되는 것을 확인할 수 있다.

3.3. 한국어 요약 말뭉치 효용성 분석

본 과제를 통해 구축한 한국어 요약 말뭉치(수동 구축 말뭉치)가 요약 모델 학습에 얼마나 유효한지 효용성을 분석하기 위해 학습 말뭉치의 크기(문서 수)에 따른 요약 성능을 측정하였다. 포인터-제너레이터에 커버리지 기법을 적용한 모델을 사용하였고 워드피스 단위의 입출력을 사용하였다. 학습은 최소 300 개부터 최대 4,006 개로 말뭉치 크기에 차등을 주었고 평가는 400 개를 사용하였다. 아래 <그림 15>, <그림 16>은 학습 말뭉치의 크기에 따른 요약 성능이다.



<그림 15> 학습 말뭉치의 크기에 따른 워드피스 단위 요약 성능



<그림 16> 학습 말뭉치의 크기에 따른 형태소 단위 요약 성능

<그림 15>, <그림 16>에서 x 축은 각각 학습 말뭉치의 크기(문서 수)를 의미하고, y 축은 ROUGE 성능을 의미한다. <그림 15>은 워드피스 단위에 대한 성능이고 <그림 16>은 워드피스 단위의 결과를 문장으로 복원한 뒤 다시 형태소 분석하여 형태소 단위로 측정한 성능이다. 워드피스 단위와 형태소 단위 모두에서 300 개 문서만을 학습했을 때에 비해 많은 문서를 학습할수록 성능이 향상되는 것을 확인할 수 있다. 즉, 구축한 한국어 요약 말뭉치가 현재 평가 말뭉치 상에서 충분한 효용성을 가진다는 것을 의미한다.

제 5 장

요약 기술 현황 및 요약 말뭉치 활용 방안

1. 요약 기술 현황

1.1 한국어 신문기사 요약봇의 현황

현재 공개된 국내 주요 자동 요약 시스템은 네이버 요약봇, 카카오의 다음 뉴스 요약봇, 엔씨소프트의 페이지(PAIGE) 등이다. 이들 서비스는 공통적으로 추출 요약 기반의 서비스다. 현재 국내 자동 요약의 수준은 단순히 주제를 잘 나타내는 문장을 뽑는 추출 요약에 그치고 있다. 이는 자동 요약 기술의 수준을 높이기 위한 완성도 높은 요약 말뭉치 구축의 필요성을 시사하는 대목이기도 하다.

1.1.1. 네이버 요약봇

▶ 개요

기사를 모두 읽지 않아도 인공지능(AI) 기술을 활용해 기사의 핵심 내용을 요약해 알려주는 '뉴스 자동 요약' 서비스이다. 요약봇은 AI 기반 자동 추출 기술로 기사 본문의 키워드, 문장 중요도 등을 판단하여 뉴스 기사를 최대 세 문장 이내에서 요약하여 보여준다. 이용자는 PC와 모바일 네이버 뉴스화면에서 기사 상단의 요약봇 버튼을 눌러 볼 수 있다.

<그림 17> 네이버 뉴스 요약봇

▶ 자동 추출 기술

요약봇은 자동 요약 기술 '아이리스'(IRIS)를 기반으로 작동한다. 텍스트로 작성된 문서에서 중요한 문장을 추출하는 요약 기술이 서비스의 핵심이다. 요약봇은 문장의 중요도를 분석하고 문서가 작성된 형식(두괄식, 미괄식)에 따라 적절한 요약 결과를 추출한다.

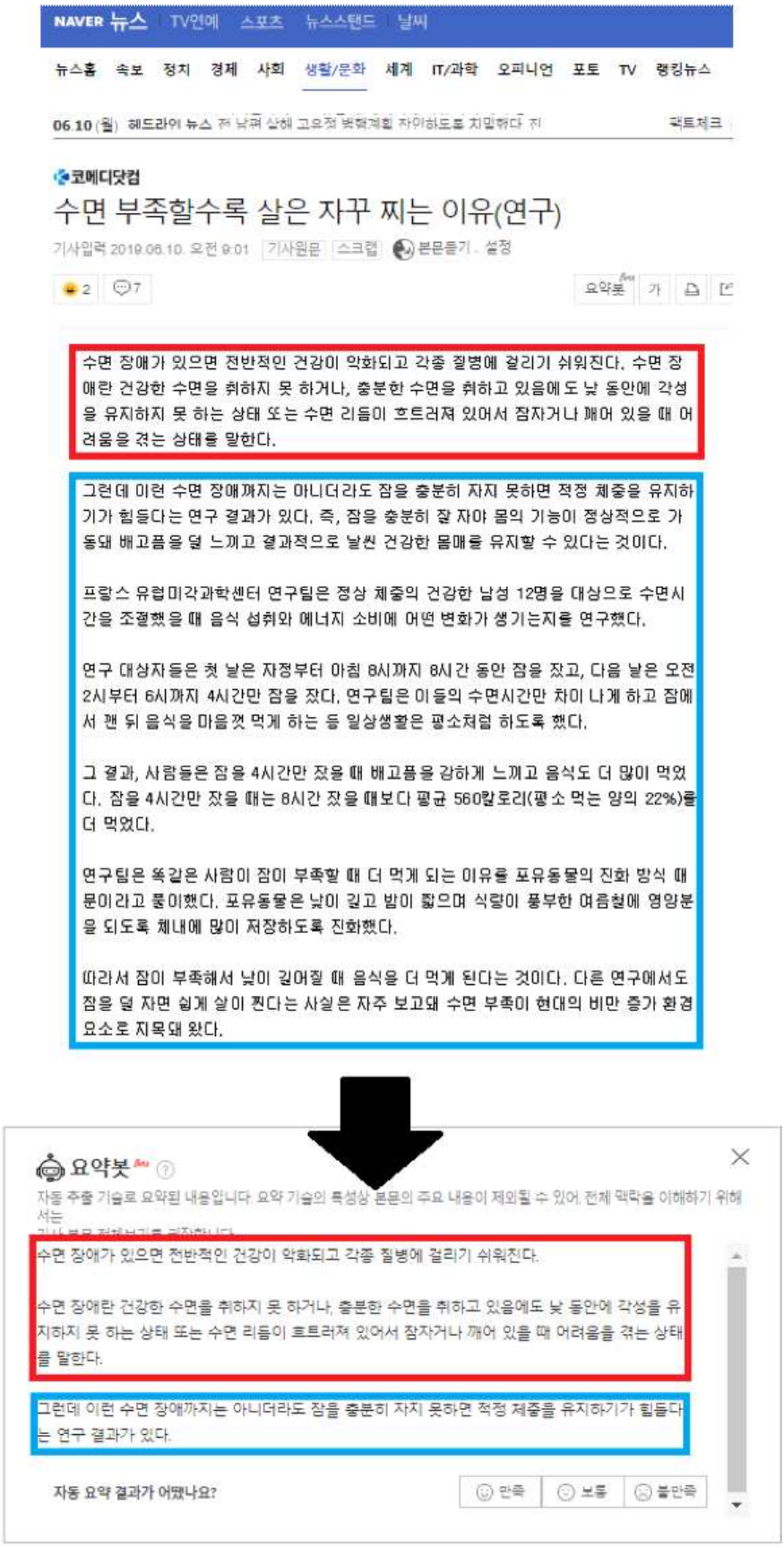
뉴스 기사의 첫 번째 문장(리드문)은 가중치를 높게 부여되므로 주로 첫 번째 문장이 추출되고, 중요도가 높은 나머지 두 문장을 추가로 판단하여 모두 세 문장으로 요약해 보여준다. 이 과정에서 문장이나 단어의 순서를 변경하거나 새롭게 추가하지 않는 등 원래 의미가 훼손되지 않도록 하고 있다. 자동 요약 모듈 개발 파트에서 제공한 3가지 모델 중 뉴스에 가장 적합하다고 판단된 모델이 적용되었다.

▶ 요약봇 제공 기사 기준

요약봇은 현재 네이버 뉴스의 정치・경제・사회・IT・생활・세계・랭킹 분야에 적용됐다. 단, 칼럼이나 사설 등 오피니언 기사, 동영상 기사, 본문이 외국어로 구성된 기사, 본문 기사가 300자 이하인 경우 또는 세 문장 이하인 기사 등에 대해서는 자동 요약 기능을 지원하지 않는다.

▶ 네이버 요약봇 예시

예시 기사를 보면, 첫 번째 단락의 2개의 문장이 요약으로 추출됐다. 그리고 두 번째 단락의 첫 문장이 추출되어 총 세 문장이 추출됐다. 네이버 뉴스봇의 기능에 따라 추출 요약된 것을 알 수 있다.



<그림 18> 네이버 뉴스 요약봇 실제 예시

1.1.2 다음 뉴스 요약봇

<그림 19> 카카오 다음 뉴스 요약봇

▶ 소개

다음은 1,800자 이하 기사에만 뉴스 요약 기능을 제공하고 있는데, 이는 뉴스의 목적과 성격에 따라 다른 접근 방법을 취하는 것이라 볼 수 있다. 1,800자 이내의 스트레이트성 기사는 이용자에게 더 쉽고 빠르게 전달하는 것을 목적으로 하는 경우가 많은데, 그 목적을 보조하는 차원에서 요약 기능을 제공하는 것이다. 1,800자 이상의 기획 기사, 인터뷰 등 장문의 기사는 오히려 독자가 시간을 들여 꼼꼼히 읽도록 유도하는 것이 바람직하다. 따라서 이러한 심층 기사는 원문을 요약하지 않고, 더 많은 이용자들에게 노출하여 전문을 읽도록 유도하고 있다.

▶ 자동 추출 기술

다음 뉴스 요약봇에는 특허 등록된 '기사 요약 서비스 서버 및 방법' 알고리즘이 적용됐다. 자동 요약 기술은 '사회 연결망 분석'을 활용한 알고리즘을 이용하는 '연결 중심성 분석'에 따라 이뤄진다.

해당 알고리듬은 기사의 제목 및 리드(lead) 문장을 구분한 뒤, 제목 및 리드 문장에서 적어도 하나 이상의 핵심 키워드를 선정한다. 이후 핵심 키워드를 포함하고 있는 문장들 사이의 연결 중심성을 평가하여 연결 중심성이 높은 순서대로 부가적인 문장들을 추출한다. 이렇게 추출된 문장들을 리드 문장과 함께 요약문을 구성하는 보충 문장으로 이용하는 것이 해당 알고리즘의 원리다.

이러한 방식은 단순히 단어들의 빈도가 높게 나온 문장을 우선적으로 보여주는 방

식이 아니라, 문장들끼리의 참조 관계를 바탕으로 상대적인 중요성을 산출하여 요약 문장을 결정 및 추출하는 방식이다.

▶ 다음 요약봇 예시

다음 뉴스의 요약봇도 원문 자체는 변경하지 않고 첫 번째 문단과 중간 문단의 문장을 이용해서 요약을 했다. 요약 구성은 세 문장이다.

연합뉴스

코레일, KTX 121회 부정승차자 적발..'징수 부가운임 1천여만원'

입력 2019.07.11. 10:40 댓글 985개

'출발 후 승차권 반환서비스' 악용, 서울~광명 구간 8개월간 부정 이용

KTX 열차와 코레일 사옥 [코레일 제공. 재판매 및 DB 금지]

(대전=연합뉴스) 유의주 기자 = 코레일이 운영하는 열차 승차권 '출발 후 반환 서비스'를 악용해 상습적으로 부정 승차를 하던 사람이 적발돼 거액의 부가운임을 물게 됐다.

코레일은 11일 광명역에서 서울역까지 KTX를 상습적으로 부정 이용하던 승차자 A씨를 적발해 1천여만원의 부가운임을 징수했다고 밝혔다.

지난해 10월 도입된 출발 후 반환 서비스는 열차 출발 이후 10분 이내에는 역을 방문할 필요 없이 스마트폰 앱 '코레일톡'에서 곧바로 구매한 승차권을 반환할 수 있는 고객 서비스다.

부정 이용을 막기 위해 스마트폰 GPS를 활용한다. 해당 열차에 탑승하면 반환할 수 없도록 조치한다.

적발된 A씨는 열차 내에서는 반환이 되지 않지만 열차가 아닌 곳에서는 취소가 가능한 점에 착안해, 자인 B씨를 이용해 승차권을 구매하도록 하고 본인은 사진으로 전송받은 승차권으로 열차를 이용하는 수법을 사용했다.

승차권을 구매하고 10분이 지나기 전에 B씨가 승차권을 반환하고 A씨는 도착역에서 자연스럽게 내리는 방식으로 부정 승차를 했다.

적발되더라도 한 번의 부정 승차에 대한 부가운임만 지불하면 된다는 생각으로 지난해 12월부터 지난 7월까지 8개월에 거쳐 모두 121번의 부정 승차를 해왔다.

코레일은 승차권 발매현황에 대한 빅데이터를 분석하던 중 A씨의 이용 패턴을 수상히 여기고 수차례 확인과 추적을 거쳐 부정 승차자인 A씨를 현장에서 적발했다.

코레일은 철도사업법에 따라 A씨로부터 부정 승차 121회의 원래 운임 10만6천400원과 10배에 해당하는 부가운임 1천16만4천원을 징수했다.

이선권 코레일 고객마케팅단장은 "다수 선의의 고객을 위한 판의 서비스를 악용해 부정 승차를 하는 것은 엄연한 범죄행위"라며 "지속적인 모니터링으로 부정 승차를 단속해 올바른 철도 이용 문화가 정착되도록 하겠다"고 말했다.

↓

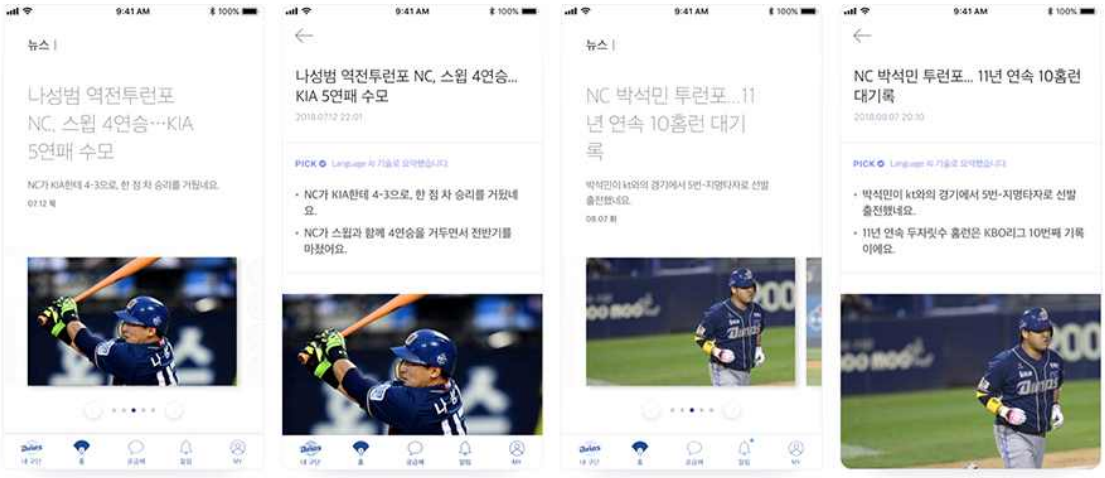
코레일이 운영하는 열차 승차권 '출발 후 반환 서비스'를 악용해 상습적으로 부정 승차를 하던 사람이 적발돼 거액의 부가운임을 물게 됐다.
코레일은 11일 광명역에서 서울역까지 KTX를 상습적으로 부정 이용하던 승차자 A씨를 적발해 1천여만원의 부가운임을 징수했다고 밝혔다.

코레일은 승차권 발매현황에 대한 빅데이터를 분석하던 중 A씨의 이용 패턴을 수상히 여기고 수차례 확인과 추적을 거쳐 부정 승차자인 A씨를 현장에서 적발했다.

ⓘ 기사 제목과 주요 문장을 기반으로 자동요약한 결과입니다. 전체 맥락을 이해하기 위해서는 본문 보기를 권장합니다.

<그림 20> 다음 뉴스 요약봇 실제 예시

1.1.3. 엔씨소프트 페이지(PAIGE)



<그림 21> 엔씨소프트 페이지 예시

▶ **소개**

엔씨소프트의 인공지능 기반 야구 정보 서비스 페이지(PAIGE) 모바일 앱에 탑재된 뉴스 기사 자동 요약 서비스이며, 서비스 도메인은 야구에 한정된다. 기사별로 요약이 제공되며, 문어체인 원문을 구어체 요약으로 문체 변환을 하여 요약문을 생성한다.

▶ **자동 요약 기술**

기술 현황

페이지(PAIGE)의 요약봇은 최소한의 정보량으로 야구 분야 핵심 뉴스를 빠르게 파악하고, 모바일 플랫폼, AI 봇 페르소나에 적절한 요약 표현으로 전달하는 것을 목표로 한다. 그러나, 현재 아래와 같은 전형적인 비지도 추출 요약(Unsupervised Extractive Summarization)의 기술 구조를 사용하고 있어, 품질에 한계가 있다.

Content Selection: 문장의 중요도에 따라 추출할 문장 선택, 중복 제거

Information Ordering: 사건의 흐름, 레토리컬 구조, Coherence 반영하여 추출된 문장을 재배열

Sentence Realization: 문맥에 맞도록 문장을 변경, 플랫폼에 따라 문장을 단순화하거나 자연스러운 문체로 변환하는 과정

전통적인 비지도(Unsupervised) 추출 요약 방법을 사용함에 있어 한계점

현재 한국어 뉴스 요약을 위해 어떤 문장이 중요한지에 대한 학습데이터가 없고,

이에 대한 가이드도 연구된 바가 없어, 대부분의 요약봇이 위와 같은 비지도(Unsupervised) 알고리즘을 사용하여 문장을 추출하고 있다. 다음과 같은 경우에는 성능의 한계로 인해서 오류가 발생하는데 선별된 문장들의 정보량이 부족한 경우, 문장간 Coherency가 부족한 경우, 선행사가 없는 지시어, 대용어가 등장하는 경우가 있다. Surface Realization 단계에서도 요약문에 적합하도록 생략(Simplification)이나, 상호 참조 생성(Co-reference generation), 혹은 상호 참조 복원이 되어야함에도 불구하고 이에 대한 학습 데이터 부재로 인해 정보 왜곡이나 부자연스러운 문서가 생성되는 현상이 발생한다.

또한 전통적인 파이프라인(Content Selection -> Information Ordering -> Sentence Realization) 방식의 추출 요약은 각 단계의 오류 전파로 인한 성능 저하 문제 등이 발생하고 있다.

▶ 엔씨소프트 페이지(PAIGE) 서비스 예시

예시에서는 첫 번째 문장과 두 번째를 추출한 것을 확인할 수 있다. 이 과정에서 문어체의 원문을 구어체로 수정한 것이 특징이다. 원문은 ‘낙찰됐다’ 인데 요약은 ‘낙찰됐어요.’로 문체가 변환된 것을 확인할 수 있다.

NCsoft PAIGE 서비스 앱

본문 :

메이저리그(MLB) 명예의 전당 입회자 11명의 사인이 담긴 야구공이 경매가 약 2억8천만원에 팔렸다.

AP 통신에 따르면, 미국 릴랜드 스프링 클래식 옥션에 나온 이 야구공은 23만6천389달러에 낙찰됐다.

전 메이저리그 투수 에디 로멜의 가족이 내놓은 이 공은 1939년에 받은 명예의 전당 입회자 11명의 사인을 담고 있다.

로멜은 1920~1930년대 필라델피아 애슬레틱스 투수였다가 아메리칸 리그 심판으로도 활동했다. 로멜의 가족은 이 공을 포함해 총 42점을 경매에 내놓았다.

이번 경매에서는 스포츠 사진 최고가 기록이 나왔다.

1914년 베이브 루스가 포함된 볼티모어 오리올스의 팀 사진 원본이 19만373달러(2억2천600만원)에 팔린 것이다. 루스가 볼티모어 유니폼을 입고 있는 사진 원본은 릴랜드 옥션에 처음 등장했을 정도로 희귀하다.

1914년 베이브 루스가 포함된 볼티모어 오리올스의 팀 사진 원본이 19만373달러(2억2천600만원)에 팔린 것이다. 루스가 볼티모어 유니폼을 입고 있는 사진 원본은 릴랜드 옥션에 처음 등장했을 정도로 희귀하다.

기존 스포츠 사진 최고가는 77만98달러였다. 이 사진은 1910년 사진가 찰스 콘런이 뉴욕 힐탑파크에서 찍은 타이 콥의 슬라이딩 사진으로, 2015년 경매에서 팔렸다.

이 밖에 아이스하키 전설 보비 오어가 1872년께 보스턴 브루인스 시절 입은 실착 유니폼이 11만3천924달러에, 미국프로농구(NBA) 뉴욕 닉스의 스타였던 월트 프레이저가 1970년 파이널 7차전에서 입은 저지가 10만81달러에 낙찰됐다.

또 1924년 신인이던 루 게릭을 포함해 뉴욕 양키스 선수들의 사인이 담긴 야구공이 7만1천508달러에, 1924년 워싱턴 세너터스의 월드시리즈 챔피언 팀 사인 야구공이 5만4천301달러에 팔렸다.

기사 상단 요약 제공 :

MLB 명예의 전당 입회자 사인 야구공, 2억8천만원에 낙찰

2019.06.10 (월) 18:00

PAIGE SUMMARY

메이저리그(MLB) 명예의 전당 입회자 11명의 사인이 담긴 야구공이 경매가 약 2억8천만원에 팔렸습니다.

AP 통신에 따르면, 미국 릴랜드 스프링 클래식 옥션에 나온 이 야구공은 23만6천389달러에 낙찰됐어요.

Sports Collector @SportsCollector · 5월 8일
Former Philadelphia A's pitcher Eddie Rommel had quite a collection and it's now at @Leilandsdotcom including a '39 HOF induction ball and a '24 Yankees with young Lou Gehrig: sportscollectorsdaily.com/1939-hof-signe...

<그림 22> 페이지 요약봇 실제 예시

2. 해외 요약 말뭉치 현황

영어권에서 요약 기술 개발을 위해서 공개된 요약 말뭉치는 다음과 같다. 누구나 활용 가능하도록 공개되어 있어 각 모델의 개발과 성능평가에 활용되고 있다. 아래 표는 해외 요약 말뭉치의 현황을 보여주며, 소개된 말뭉치 중 널리 이용되는 말뭉치에 대한 설명을 각 절에서 기술한다.

말뭉치	소개	규모
CNN / Daily Mail	1. 구성 온라인 뉴스 기사(원본) + 요약 문서(사람 작성) = 1쌍 2. 구축 시기 - 개체명 익명 버전(2016) - 개체명 공개 버전(2017)	1. 평균 토큰 - 원본 기사: 평균 781토큰 - 요약 문서: 평균 56토큰, 3.75 문장 2. data set(pairs) - training: 287,226 - validation: 13,368 - test pairs: 11,490
Gigaword	1. 구성 뉴스 기사(원본) + 헤드라인 1문장 = 1쌍 2. 구축 시기 Rush et al.(2015) (전 버전 Gigaword(2012)를 전처리 후 training set 구축)	1. 평균 토큰 - 원본 문서: 31.4토큰 - 요약 문장: 8.3토큰 2. data set(pairs) - training: 3,800,000 - development: 189,000 - test instances: 1,951
DUC 2004 TASK1	1. 구성 뉴스 기사(원본) +요약(사람 작성, 4개 요약) = 1쌍(문서) 2. 구축 시기 DUC 2004 (전 버전 DUC 2003)	1. 평균 토큰 - 원본 문서: 35.6토큰 - 요약 문장: 10.4 토큰 2. data set(pairs) 500개 문서
Webis-TLDR-17Corpus	1. 구성 소셜미디어의 content(원본) +요약(원본 작성자의 요약) =1쌍	1. data set(pairs) 4,000,000
Sentence compression -Google dataset	1. 구성 인터넷 뉴스 기사(원본) + 축약 문장 = 1쌍 2. 구축 시기 traing portion 추가(2017) (전 버전 Fillippova et al. 2013)	1. data set(pairs) - total: 210,000 - training: 200,000 - evaluation: 1,000 sentences

표 70 해외 요약 말뭉치 현황

2.1. CNN/Daily Mail Corpus

▶ 소개

Nallapati et al. (2016)에 의해 구축된 데이터셋으로 요약 평가용으로 사용되었다.

▶ 구성

CNN, Daily Mail 웹사이트의 new-stories에서 사람이 작성한 추상 요약을 질문으로 사용하고, 추가로 빈 칸 채우기 질문의 답으로 예상되는 구절로 stories를 사용한다. 즉, 웹사이트에서 단락과 질문을 추출하여 쌍을 생성하였다.

▶ 특징

1) 추상 요약 말뭉치 Gigaword와 DUC은 요약이 1문장이지만, CNN/Daily Mail 말뭉치는 여러 문장으로 구성되어 있다.

2) anonymized 버전(2016)과 non-anonymized 버전(2017)으로 2가지 버전 구축되었다.

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

Table 3: Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

<그림 23> CNN/Daily Mail Corpus 예시

2.2. Gigaword

▶ 소개

Rush et al. (2015)에서 처음 사용된 데이터 셋이다.

▶ 구성

Gigaword는 지난 20년간의 다양한 국내와 국제 뉴스 서비스의 950만 뉴스 기사를 포함하고 있다. Rush et al. (2015) 학습 셋에 필요한 입력으로 요약 쌍을 만들기 위해 기사의 헤드라인과 헤드라인의 첫 번째 문장을 쌍으로 묶었다.

▶ 특징

1) Rush(2015)는 Gigaword data set(Graff et al., 2003; Napoles et al.,2012)를 스탠포드 Core NLP tool(Manning et al. 2014)로 전처리한 standard Gigaword로 구성된 데이터 셋을 사용했다.

2) 일부 베이스라인들은 파싱과 태깅도 사용했지만, Rush(2015) 모델은 오직 토믄화와 문장 분리를 위한 전처리 모듈만 사용한다.

I(1): a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .
G: iranian-american academic held in tehran released on bail
A: detained iranian-american academic released from jail after posting bail
A+: detained iranian-american academic released from prison after hefty bail

I(2): ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .
G: european mediterranean ministers gather for landmark conference by julie bradford
A: mediterranean neighbors gather for unprecedented conference on heavy security
A+: mediterranean neighbors gather under heavy security for unprecedented conference

<그림 24> Gigaword Corpus 예시

2.3. DUC 2004 TASK1

▶ 소개

DUC 2004는 총 5개 TASK로 진행하였으며, 그중에서 task1에서 제공한 데이터이다. DUC 2003 data를 training으로 사용하기 위해서 DUC 2004의 Task 1, 2는 DUC 2003과 기본적으로 같다.

▶ 구성

문서 출처는 AP newswire, New York Times이며 사람이 작성한 헤드라인이 아닌 4개의 참조 요약문으로 구성되어 있다. TASK1의 경우, 75bytes의 제한을 두어 문서 당 짧은 요약으로 작성되었다.

▶ 특징

1) 요약 연구에서 Gigaword, DUC 2003과 함께 평가 데이터셋으로 쓰였다.
2) DUC 2004는 500 pairs (DUC 2003은 문서와 요약으로 624쌍)로 구성되어 있다.
3) ROUGE n-gram 매칭으로 평가되었다.

2.4. Webis-TLDR-17 Corpus

▶ 소개

소셜미디어 ‘Reddit’의 콘텐츠와 self-written 요약으로 구성되어 있다. 소셜미디어 도메인으로 된 요약 데이터셋 중에서 첫 번째 대규모 데이터셋이다. Reddit은 social news 집합, 웹 콘텐츠 평가, 토론 등이 이루어지는 인터넷 커뮤니티로 submissions과 comments로 구성되어 있다. submissions은 유저들이 업로드하는 top-level posts로 제목을 포함하고 있고 내용은 링크만 있거나 유저가 작성한 body text도 포함되어 있다. comments는 body text로 구성되어 있다.

▶ 구성

Reddit의 submissions와 comments를 필터해서 최종 pairs를 구성하였다. sub reddits 32,778, submissions 1,667,129, comments 2,377,372로 구성되어 있다.

Table 3: Examples of content-summary pairs.

Example Submission
Title: Ultimate travel kit **Body:** Doing some traveling this year and I am looking to build the ultimate travel kit ... So far I have a Bonavita 0.5L travel kettle and AeroPress. Looking for a grinder that would maybe fit into the AeroPress. This way I can stack them in each other and have a compact travel kit. **TL;DR:** What grinder would you recommend that fits in AeroPress?
Example Comment (to a different submission)
Body: Oh man this brings back memories. When I was little, around five, we were putting in a new shower system in the bathroom and had to open up the wall. The plumber opened up the wall first, then put in the shower system, and then left it there while he took a lunch break. After his break he patched up the wall and left, having completed the job. Then we couldn’t find our cat. But we heard the cat. Before long we realized it was stuck in the wall, and could not get out. We called up the plumber again and he came back the next day and opened the wall. Out came our black cat, Socrates, covered in dust and filth. **TL;DR:** plumber opens wall, cat climbs in, plumber closes wall, fucking meows everywhere until plumber returns the next day

<그림 25> Webis-TLDR-17Corpus 예시

3. 요약 말뭉치 활용 방안

본 과제에서 구축한 추출 요약 및 추상 요약에 대한 주석 지침과 말뭉치는 요약 모델의 성능 향상과 문서 생성 기술 개발에 있어 중요한 자원으로 활용될 수 있다. 요약 말뭉치의 구축과 개발이 요약 모델의 성능 향상의 기반이 되고, 이 기술을 응용한 다양한 분야의 발전을 견인하는 역할을 하게 될 것이다.

1) 요약 모델의 요약 성능 고도화

최근의 딥러닝 네트워크는 메모리 네트워크의 형태로 다양한 문맥 정보를 동시에 고려하는 구조로 발전하고 있다. 신규로 구축하는 추출 요약에 대한 정답 말뭉치는 요약의 대상이 되는 원시 말뭉치의 다양한 특징(문서 카테고리, 추출 문장의 문서 내 위치 등)을 반영한 메타 정보가 주석 과정에서 함께 저장된다. 따라서 이러한 자질들을 하나의 딥러닝 네트워크로 입력하여 문장 선별 및 최종 요약문에 대한 지도 학습이 효율적으로 가능해진다. 더불어 기존에 존재하는 문서 내의 문장에서 추출하는 추상 말뭉치 이외에 추상 요약 말뭉치는 실제 작업자가 새로운 요약문을 생성하는 것으로 구축 난이도가 높다. 본 과제를 통해 구축된 추상 요약 말뭉치를 요약 모델에 적용해본 결과 기계적인 방법을 통해 구축한 대규모의 말뭉치에 상대적으로 소규모의 잘 설계된 말뭉치를 더하여 품질 향상을 이끌어낼 수 있음을 확인하였다.

2) 요약 말뭉치 지침 공개

기업에서는 기술적인 난이도가 높은 추상 요약을 위해 도메인 전문가를 초빙하여 주석 가이드라인을 개발하는 것도 쉽지 않은 상황이다. 따라서 이에 대한 가이드라인만으로도 향후 기업이 학습 말뭉치를 확장하여 기술 개발하는 데 도움을 줄 수 있다. 언어학적 말뭉치 구축 가이드라인에 대한 개발은 이후 기업에서 개별적으로 말뭉치를 확장할 때 중요한 지침으로, 매우 활용도가 높다.

3) 요약 모델 성능 평가용 데이터

영어의 경우 CNN/daily 말뭉치 등 대규모의 자원이 공개되어 있으나, 한국어 경우 공개된 한국어 요약 말뭉치가 부재한 상황이다. 요약 모델의 성능 평가와 비교를 위해

서는 개발된 요약 기술의 성능을 객관적으로 평가할 수 있는 평가용 데이터셋이 필요하다. 본 과제를 통하여 구축한 말뭉치는 추출 요약을 위한 주제문과 추상 요약을 위한 요약문으로 구성되어 있어 한국어 요약 모델의 성능을 평가하는 평가용 데이터셋으로 활용될 수 있다.

4) 로봇 저널리즘 등 응용 기술 개발을 위한 기반 마련

로봇 저널리즘은 로봇과 저널리즘이 결합된 합성어로 AI 기술을 바탕으로 취재, 보도, 논평 등 저널리즘 영역 전반에 걸친 모든 과정에서 도움을 주거나 대신해 주는 광범위한 기술 분야라고 할 수 있다. 기계가 저널리즘 활동을 보조하거나 대신하기 위해서는 광범위하고 방대한 분량의 정보를 수집하여 사건을 분석하고, 정보를 효과적으로 전달하기 위한 의견이나 예측을 생성하는 기술들이 뒷받침되어야 한다.

수집된 정보를 분석하는 과정은 중요한 정보와 그렇지 않은 정보를 선별하고, 필요한 정보만을 압축적으로 제시하는 기술이 동반될 때 그 효율성이 증대될 수 있다. 기계가 대용량 문서 집합으로부터 주요 사건을 추출하고 이를 적절한 형태로 제시하려면 다중 문서 요약에 관한 기반 기술이 필수적이다. 자동 요약 기술이란 무수한 양의 비정형 데이터로부터 주요 사건을 추출하고, 불필요한 중복을 제거하고, 선후 관계에 따라 올바른 순서로 정보를 제시할 수 있는 기술을 의미하기 때문이다. 영어권에서는 자동 요약 기술을 위한 여러 학습데이터가 공개되어 있고, 이를 통해 연관된 응용 기술들의 발전이 급격하게 이루어지고 있다. 한국어 요약 말뭉치 역시 로봇 저널리즘과 같은 고도의 응용 기술 개발에 있어서 중대한 기여를 할 수 있다.

제 6 장

결론

1. 연구 요약

본 사업의 목적은 신문기사 요약 말뭉치를 구축하여 4차 산업 혁명의 핵심인 빅데이터 분야의 발전으로 대두된 자동 요약 기술 개발에 이바지하는 데에 있다. 자동 요약은 정보 소스와 채널의 다각화 및 대량화, 전송의 신속화, 정보 소비 패턴의 변화 등으로 인해 매우 중요한 서비스로 등장하였다. 방대한 양의 텍스트에서 사용자에게 필요한 정보를 정확하게 추출하여 제공하기 위해서는 요약 기술의 발전이 필수적이며, 이러한 기술 개발에 반드시 필요한 것이 바로 요약 말뭉치이다. 국내에서 공개된 문서 요약 시스템은 포털 사이트의 신문 기사를 대상으로 한 추출 요약의 베타 서비스 수준에 그치고 있다. 한국어 요약 기술이 앞으로 나아가기 위해 필요한 완성도 높은 한국어 추출, 추상 요약 말뭉치를 시범 구축하는 것을 목표로 하여 진행된 본 사업의 주요 과업과 연구 성과는 다음과 같다.

○ **신문기사 말뭉치 대상 요약 말뭉치 구축** : 추출 요약을 위해 주석된 13,167 문장의 주제문과 추상 요약을 위해 도메인 전문가가 검수하여 완성한 13,167 문장의 요약문을 포함한 총 4,389 기사로 구성된 신문기사 요약 말뭉치를 구축하였다. 이를 위해 요약 말뭉치 구축을 위한 워크벤치를 커스터마이징하여 구축과 검수 과정을 관리하였다. 추출 요약을 위해서 키워드와 추출 후보 문장을 제시하였으며 추출 요약, 추상 요약 모두 세 문장으로 요약하고 XML 형식으로 정제하였다.

○ **한국어의 특성을 살린 고품질 추상 요약문 작성을 위한 지침 개발** : 요약문 작성을 위하여 한국어의 특성과 실용성을 반영하여 작성자가 고품질의 요약문을 생성할 수 있는 지침을 개발하였다. 지침에는 구축 작업자와 도메인 전문가, 작문 전공 연구진의 의견을 고루 반영하여 요약 말뭉치를 구축하고자 하는 누구나에게 참고가 될 수 있도록 작성하였다. 지침은 기본 원칙, 주제문 주석 및 요약문 생성 절차, 각 절차별 세부 지침 등으로 구성되었다.

○ **요약 말뭉치 구축 및 평가 방법론 개발** : 추출 요약과 추상 요약 방식의 요약 시스템에 모두 적용할 수 있는 요약 말뭉치를 유기적이고 효율성 높은 방식으로 구축할 수 있는 구축 기준과 체계를 수립하였다. 또한 구축된 요약 말뭉치를 기존에 널리 이용되는 요약 말뭉치 평가 방식을 통해 검증하고 정성적으로 평가할 수 있는 평가 기준을 제시하였다. 정량정 평가 방법론 연구에서는 요약문 자동 평가 방법과 사람의 수

동 평가 사이의 상관관계를 측정하여 어떤 평가 방법이 한국어 문서 요약에 가장 적합한지 실험을 통해 밝혔다. 무선표집된 50개의 요약문을 정성적으로 평가한 결과, 모두 5개의 CASE로 분류할 수 있었다. 주제문 주석의 정확성과 요약문의 질을 종합적으로 판단하였다.

○ **자동 요약 기술 적용 및 요약 말뭉치 활용 방안 모색** : 요약 기술을 실제 개발하고 활용해야 하는 산업체의 자문을 받아 기술 개발에 사용될 수 있도록 말뭉치를 설계하고 산업계의 실수요를 조사하였다. 또한 실제 자동 요약 모델에 구축된 요약 말뭉치를 적용하여 활용 가능성을 검토하였다. 본 과제에서 구축한 추출 요약 말뭉치의 유용성을 살펴보기 위하여 greedy 방식으로 추출한 학습 코퍼스에 추가로 본 과제에서 구축한 요약 말뭉치를 더하여 학습을 진행한 결과 성능 향상을 살펴볼 수 있었다. 포인터-제너레이터 적용, 추출 요약과 군집화를 이용한 문맥 추가 실험, 품사 게이트 자질 추가 실험, 워드피스 단위 품사 게이트 자질 추가 실험 등을 통해 구축한 추상 요약 말뭉치가 평가 말뭉치로서 충분한 효용성을 가짐을 증명하였다.

2. 정책 제언

○ 장르별 요약 말뭉치 구축 방법론 연구 필요

본 과제를 수행하는 과정에서 원시 말뭉치의 메타데이터에 기재된 정치, 사회, 스포츠 등의 주제 영역보다 세부적인 기사의 형식적 유형이 요약 말뭉치를 구축하는 데에 영향을 크게 미친다는 것이 밝혀졌다. 질문과 답변으로 되어 있는 인터뷰, 문화 행사나 전시, 도서, 방송에 대한 리뷰, 유사한 성격의 정보가 여러 개 나열된 기사의 형식으로 제시된 광고 등은 전형적인 정보적 기사에서 벗어나 있어 현재의 기술 수준과 말뭉치 구축 방식으로는 효율적으로 접근하기 힘들며 활용도가 떨어진다. 이러한 정보를 바탕으로 신문 장르 내부적으로 혹은 신문 외 장르에 대해 장르별 요약 말뭉치 구축 방법론을 연구할 필요가 있다.

○ 고차원 자동 요약 기술을 위한 다중 문서 말뭉치 구축

현재 한국어 자동 요약 기술은 어플리케이션을 개발하여 경제적 가치를 창출할 만한 수준에 미치지 못하고 있다. 신문 기사의 리드 문장과 관련 문장 몇 개를 추출하여 제시한 자동 요약 결과에 대한 인간의 만족도가 높지 않아 베타 버전을 서비스하고 있는 실정이다. 그러나 메가 데이터를 대상으로 주요 사건을 추출하고, 불필요한 중복을 제거하여, 선후 관계에 따라 정보를 제시하는 자동 요약은 정보가 홍수처럼 쏟아지는 현대사회에서 반드시 지향해야 할 기술이다. 현 시점에서 한국어에 관한 한 현장에 적용되는 기술은 추출 요약이 대부분이고 추상 요약은 시작 단계일 뿐이지만 이러한 단일 문서 요약에서 한걸음 더 나아가 여러 문서를 대상으로 필요한 정보만을 압축적으로 제시하는 다중 문서 요약에도 도전해야 한다. 이를 위해서는 다중 문서 요약을 위한 대상과 결과를 평가셋으로 구축하여 고차원 자동 요약 기술의 개발을 위한 말뭉치를 확보할 필요가 있다.

○ 요약 기술 완성도 제고를 위한 관련 말뭉치 구축 및 적용 모색

주요 언어를 중심으로 발전하고 있는 요약 기술의 완성도를 높이기 위해서 최근에는 직접적인 요약 말뭉치뿐만 아니라 관련된 다른 차원의 말뭉치를 적용하기도 한다. SNS 자료나 상품 리뷰 등의 실용적이고 대중적인 텍스트를 대상으로 의존 관계를 분석하고 통합한 후 이를 그래프화하여 정보를 추출하고 요약을 할 수 있다. 또한 최근에는 명제적 의미에 대한 주석인 추상 의미 표상(AMR)을 활용하여 요약의 성능을 높일 가능성에 대해 탐구한 논문이 발표되기도 하였다. 그러나 한국어 언어 자원 중에는

실용성이 높은 텍스트에 대한 의존 관계 분석 말뭉치나 추상 의미 표상(AMR) 말뭉치가 제대로 구축되어 있지 않기 때문에 요약 기술 등 난도가 높은 언어 처리 기술에 적용할 수가 없는 실정이다. 이러한 상위 레이어의 주석 말뭉치 구축을 장려하여 자동 요약을 비롯한 응용 기술 발전에 활용할 필요가 있다.

○ 국제학술대회 수준의 요약 기술 개발을 위한 평가 말뭉치 구축

자동 요약은 자연 언어 처리 및 전산 언어학 분야에서 ACL(Association for Computational Linguistics), EMNLP(Empirical Methods in Natural Language Processing), NAACL(North American Chapter of the Association for Computational Linguistics) 등의 상위 그룹 국제학술대회에서 따로 세션을 만들거나 워크숍을 운영할 만큼 중요한 언어 처리 관련 연구 분야이다. 2016년부터 2018년까지 최근 3년간의 국제 학술대회의 세부 연구 분야별 통계를 내어 보면 늘 10위 내외에 위치할 만큼 중요도가 높다. 그러나 한국어와 관련해서는 공인된 평가 데이터가 없어서 국제 개방 경진대회(Shared Task) 등에서 다루어지지 못하고 있다. 시범 구축 결과를 바탕으로 요약 말뭉치를 본격 구축하여 상위 그룹 국제학술대회 수준의 요약 기술 개발을 시도할 필요가 있다.

○ 국내 경진대회를 통한 한국어 자동 요약 연구 활성화

국립국어원에서는 2013년부터 형태소 분석, 질의응답, 개체명 인식, 의존구문분석 등의 주제로 꾸준히 국어 정보 처리 시스템 경진 대회를 개최해 오고 있는데 이러한 지정 분야를 통해 공식적인 평가 말뭉치를 제공하고 후진 양성에 기여했다는 평가를 받고 있다. 다만 지정 분야가 주로 자연 언어 이해 등의 기초에 초점이 있어 이제 자연 언어 생성으로 확장될 필요가 있다. 본 과제의 주제인 자동 요약은 일종의 자연 언어 생성 분야로서 응용 분야가 넓기 때문에 수요는 많으나 국내에서 공식적인 평가 말뭉치가 제공된 바 없으므로 향후 본 과제 결과물을 확장한 한국어 자동 요약 평가 말뭉치를 제공하고 관련 분야 연구를 활성화하는 것이 바람직하다.

참고문헌

이은경, & 배현숙. (2013). 학문 목적 중국인 학습자의 수준별 요약문 쓰기 전략 분석. 한국언어문화교육학회 학술대회, 261-282.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K. (2017). Text Summarization Techniques: A Brief Survey, arXiv:1707.02268.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. Journal of verbal learning and verbal behavior, 22(1), 1-14.

Celikyilmaz, A., Bosselut, A., He, X. and Choi, Y. (2018). Deep Communicating Agents for Abstractive Summarization, In Proceedings of NAACL-HLT 2018, pages 1662-1675.

Chen, Y. and Bansal, M. (2018). Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting, In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics , pages 675-686.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. Journal of English for academic purposes, 7(3), 140-150.

Denkowski, M and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language, In Proceedings of the Ninth Workshop on Statistical Machine Translation.

Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proceedings of NAACL-HLT 2018.

Di Vesta, J. G., & Gray, S. G. (1972). Listening and notetaking. Journal of Educational Psychology, 63, 8-14.

Einstein, G. O., Morris, J., & Smith, S. (1985). Note-taking, individual differences, and memory for lecture information. Journal of Educational Psychology, 77, 522-532.

Fan, L., Yu, D. and Wang, L. (2018). Robust Neural Abstractive Summarization Systems and Evaluation against Adversarial Information, In Proceedings of 32nd Conference on Neural Information Processing Systems (NIPS 2018).

Filippova, K. and Y. Altun, (2013). Overcoming the lack of parallel data in sentence compression, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. Contemporary Educational Psychology, 26(1), 3-24.

Ganesan, Kavita (2015), ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks, arXiv:1803.01937.

Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English for academic purposes teaching and learning. TESOL Journal, 4(1), 9-24.

Graham, Yvette (2015). Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 128-137.

Kiewra, K. A., DuBois, N., Christensen, M., Kim, S.I., & Lindberg, N. (1989). A more equitable account of the notetaking functions in learning from lecture and from text. Instructional Science, 18 (3), 217-232.

Kryściński, W., Paulus, R., Xiong, C. and Socher, R. (2018). Improving Abstraction in Text Summarization, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1808-1817.

Lewkowicz, J. (1994). Writing from Sources: Does Source Material Help or Hinder Students' Performance?.

Li, W., Xiao, X., Lyu, Y. and Wang, Y. (2018). Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1787-1796.

Lin, C. (2004), ROUGE: A Package for Automatic Evaluation of Summaries, In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004.

Nallapati, R., F. Zhai, and B. Zhou. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Thirty-First AAAI Conference on Artificial Intelligence.

Ng, J. and Abrecht, V. (2015). Better Summarization Evaluation with Word Embeddings for ROUGE, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1925-1930.

Paulus, R., Xiong, C. and Socher, R. (2018). A DEEP REINFORCED MODEL FOR

ABSTRACTIVE SUMMARIZATION, In Proceedings of ICLR 2018.

See, A., Liu, P. J. and Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks, In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1073-1083.

Silva, T. (1990). Second language composition instruction: Developments, issues, and directions in ESL. Second language writing: Research insights for the classroom, 11-23.

Tan, J., Wan, X. and Xiao, J. (2017). Abstractive Document Summarization with a Graph-Based Attentional Neural Model, In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1171-1181.

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests?. Language Testing, 22(2), 174-210.

Van Dijk, T. A., & van Dijk, T. A. (1977). Text and context: Explorations in the semantics and pragmatics of discourse.

Watanabe, Y. (2001). Read-to-write tasks for the assessment of second language academic writing skills: investigating text features and rater reactions. Unpublished doctoral dissertation, University of Hawaii.

Wu, Y., and Hu, B. (2018). Learning to Extract Coherent Summary via Deep Reinforcement Learning, In Proceedings of AAAI.

Yang, Q, Passonneau, R. J. and Gerard de Melo (2016). PEAK: pyramid evaluation via automated knowledge extraction, In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence Pages 2673-2679.

Zhang, J., Zhou, Y,, and Zong, C. (2016). Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 24, NO. 10, OCTOBER 2016.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2019). BERTSCORE: Evaluating Text Generation with BERT, arXiv:1904.09675.

부록. 요약 구축 작업에 대한 전문가 의견

전문가 의견 1 : 김동준
- 추출요약을 위한 주제문 선정이라면, 구체적 정보가 많이 포함된 문장을 선정하여 원문을 읽지 않아도 전체 정보를 파악할 수 있도록 유도함이 바람직할 것으로 사료됨. - 요약문이 기사전체를 포괄할 수 있는 새로운 문장 생성이라면, 팩트 위주보다는 상위어 위주로 작성해야함에도 실제 작업에서는 한정된 문장으로 원문이 전달하려는 정보를 충실하게 전달하려는 방향으로 문장을 작성하는 경향이 큼. - 요약문 작성의 확실한 방향성 정립이 절실함(이는 오히려 제목을 활용하는 게 효율적인 듯함) - 전형적 기사문장이 요약에 용이하나, 향후 문장을 이야기 하듯 쉽게 쓰거나, 탐사(르포) 묘사(스케치) 유형 기사에 대한 요약 지침 필요. - 들여쓰기, 구두점, 영어철자 포함한 고유명사 표기법 등 통일된 표현법 확정필요. - 최초 구축문 작성이 가장 중요함. 작업시간이 충분해야 양질의 구축문 생성됨. 양질의 구축문에 따른 이후 작업은 오히려 수월함.

전문가 의견 2 : 여규병
-기사의 의도를 최대한 포괄하는 문장 중심으로 주제문을 선정하는 것이 기본 지침이지만 그에 따르다 보면 선정된 주제문 3개 어디에도 해당 기사가 전하고자 하는 인물이나 사건의 이름 등이 나타나지 않는 사례가 적지 않습니다. 예컨대 전시회 리뷰 기사인 ID1018의 선정된 주제문에는 작가의 이름이 전혀 나타나지 않습니다. 기사 요약이 인터넷상에서 제공되는 서비스이며 독자가 이미 제목을 보고 기사 요약을 선택한 것으로 가정해 제목에 인물 사건 이름이 확실하게 드러나면 인물 사건 이름이 드러나지 않는 문장의 선택을 용인하되, 그렇지 않으면 인물 사건 이름이 드러나는 문장을 주제문으로 선정토록 지침을 마련할 필요가 있다고 봅니다. -기자의 관점에 따라 기사 작성 방식이 달라지는바 시간적 논리적 순서에 입각하여 요약문을 재배열하는 것은 충실한 기사 요약에 방해가 될 소지가 있습니다. 따라서 이 지침은 기사의 의도를 충분히 반영할 수 있도록 요약하는 방향으로 개정하는 것이 좋을 듯합니다. -검수A 작업이 끝나면 구축 작업자에게 피드백을 제공하는 것이 좋을 듯합니다. 검수B 작업을 하면서 비록 다른 작업자들이 작업한 것이지만 구축문과 검수A의 결과물을 비교할 수 있었던 것이 다음 작업을 하는 데 참고가 되었습니다. 비전문가 그룹에는 더욱 큰 도움이 될 것이며 전체 결과물의 질적 향상에도 기여할 것으로 생각합니다. -기사 요약의 목적은 두 가지로 상정할 수 있을 듯합니다. 독자가 기사를 읽지 않아도 될 정도로 충실한 내용을 구축하는 것, 독자가 요약문을 보고 호기심을 느껴 전체 기사를 보도록 유인하는 것이라고 하겠습니다. 독자는 전자를 선호할 것이고, 포털이나 언론사는 후자를 선호할 것입니다. 이 프로젝트에서도 두 가지 관점을 염두에 두고 각각의 방향을 설정하여 서로 다른 결과물을 도출해 볼 필요가 있지 않을까 생각합니다.

전문가 의견 3 : 이창호
- 주제문을 선택하는 지침은 현행대로 유지하면 될 듯 하다. 반드시 3개를 선택해야 하는 과정이니 현재 지침이 최선이다. 단 모든 것이 기사(뉴스) 선택과 요약인 만큼 흐름을 과거, 현재, 미래 순이 아니라 뉴스가 발생한 시점을 가장 먼저 처리해야 한다. - 말뭉치 작업은 기본적으로 기사(뉴스)를 요약하고 함축적인 정보를 제공하는 것이 가장 중요한 목적인만큼 위의 사례처럼 지나친 일반화를 금지하는 기준을 마련해야 한다. - 기사 형식을 차용한 기획 광고는 요약 대상에서 제외해야 한다. 아파트나 오피스텔 분양, 각종 카드의 신규 상품, 여행 상품 등이 자주 기획성 광고 기사로 다뤄지고 있다. 신문의 경우 ‘광고 페이지’라는 문패를 통해 독자들에게 기사와 광고를 분리할 수 있도록 하고 있지만 온라인 기사의 경우 기사와 광고를 구분하는 것이 쉽지 않아 독자들이 헷갈릴 수 있기 때문이다. - 작업자의 경험에 따라 큰 편차를 보이는 점을 조정해야 한다. 기사를 써 본 작업자와 그렇지 못한 사람의 차이는 능률을 떨어뜨릴 뿐 아니라 큰 실수를 할 수 있는 원인으로 작용하기 때문이다. - 외래어 표기의 통일이 필요하다. 한겨레와 다른 언론사의 차이를 통일하는 작업을 해야 한다.

전문가 의견 4 : 이세강
- 요약문에도 제목이 필요하다. 요약기사 수요자에게 요약된 기사와 함께 기사 원문을 제공한다고 한다. 이 때에도 기사 제목없이 요약된 기사를 준다고 하면 한눈에 볼 수 없기 때문에 흡인력이 떨어질 것으로 판단된다. 최소한 간략하게 정리된 목록을 주는 것이 요약기사나 기사 원문을 보도록 유인하게 될 것이다. - 요약기사를 완성형 기사로 만들어야 한다. 요약기사가 세줄짜리여서 완성형 기사로 만드는데 어려움이 많은 것은 사실이다. 하지만 매번 기사원문을 찾아보도록 한다면 요약기사로 만드는 것이 무의미해진다. 따라서 되도록 완성형으로 만드는 것이 바람직하다. 이를 위해서는 기사 틀에 대해 일관된 규칙이 필요하다. - 약어의 사용 : 예를 들어 북대서양조약기구(NATO, 나토) 등의 표기 통일 - 기사에 등장하는 인물에 대한 정보(나이,소속,주소,성별 등) 제공 - 맥락 해석적 기사문 또는 리드의 억제 1: 팩트를 우선시하지 않고 자극적인 맥락해석이 앞서면 기사의 투명성이 훼손될 우려가 크고 주관성이 크게 늘어날 것이다. 기사가 전하는 정보의 양이 충분하지 않는 상태에서 맥락해석적 기사의 양이 늘면 그만큼 주관에 치우칠 우려가 크다. 따라서 팩트 전달에 중점을 두어야 한다. - 맥락해석적 기사의 억제 필요성 2: 예를 들어 후보자에 대한 지지도 여론조사를 보도할 때, A후보가 51%, B후보가 49%의 지지도가 나왔고 오차범위가 +- 5%라고 할 때 여론조사 보도에 대한 선관위 지침은 오차범위 내에서는 우열을 가리지 말도록 하고 있다. 이럴 경우 기사는 오차범위내 팽팽한 접전, 혼전 등으로 기사화한다. 과연 오차범위내 지지도 차이라고 의미가 없는 것일까? 요약문에서는 51:49로 명시하고 오차범위에 대한 해설을 전하면 된다. 독자의 해석 여지를 기자 또는 언론사가 빼앗을 권리는 없다고 본다. - 최근 시민저널리즘 또는 참여 저널리즘이 언론계를 강타하고 있다. 저마다 객관성에 봉사하는 전통적인 저널리즘에서 언론사의 목소리나 기자의 주장을 전하는 기사가 넘쳐나고 있다. 기사의 내용이 충분하지 않은 상태에서 주장이 많아지면 의도적 오보나 왜곡, 과장이 심해질 우려가 크다. - 기사의 투명성 제고를 위해 기사의 출처, 취재원 명시가 요청된다. - 구축요약문 작성에 전문가적 능력이 요구된다. 구축요약문이 부실하면 검수A와 검수B 작업이 더 어려워진다. 기사는 정답이 없다고 할 정도로 개인차가 크고 다양한 형태로 작성된다. 그러나 기본적으로 기사의 가치나 요건을 갖추지 못할 경우 단순히 개인차로 치부하기에 부적절한 경우가 많 다고 본다. 요약기사의 질을 유지하기 위해서는 구축요약문 단계에서 전문가의 참여가 필요하다.

전문가 의견 5 : 이용원
- 구축-검수A-검수B 3단계로 작업하려면 구축 단계부터 전문가 집단만이 맡아야 함. 구축이 제대로 되지 않으면 검수A 단계가 힘들어지고, 검수B에도 영향 미쳐 최종 요약문이 부실해지기 마련임. - 기사 원문에 사실관계 부정확, 비문 등의 문제가 있을 때 이를 구축자가 고쳐서 할 수 있는지, 그냥 원문대로 요약할지를 정해야 함. - 일부 언론사가 지향하는 로마자 사용 배제, ‘아무개’ 등의 표기를 그대로 인정할지, 언론계에서 통용하는 표기로 바꿀지도 원칙 정해야 함.<예시>김 아무개->김 모, 엠비시->MBC

<Abstract>

Analysis research and trial establishment of corpus

The purpose of this research project is to contribute to the development of automatic summarization technology that has emerged as a development in the big data sector which is the core of the 4th industrial revolution through establishing news articles summarization corpus. Automatic summarization has emerged as a very important service which is caused by diversification and maximization of information sources and channels, quickness of transmission, and changes in information consumption patterns. The development of summary technology is essential to correctly extract and provide information needed by users from a vast amount of text. For developing these technologies, the summary corpus is positively necessary. The tasks and research results of the project are as follows.

Establishment of newspaper articles-oriented summarization corpus : We made up the newspaper articles summary corpus made of total 4,389 newspaper articles including 13,167 topic sentences annotated for extraction-based summarization and 13,167 summarized sentences that has completed inspection by domain experts are for abstractive summarization.

Development of guidelines for making high-quality abstractive summarization based on the characteristics of Korean : For making summary we built the guideline to allow preparers to generate high quality summary, reflecting the characteristics of Korean language and practicality. The opinions of preparers, domain experts and researchers who majored in writing were reflected in the guideline, so it was designed to serve as a reference to anyone who wants to make a summarization corpus.

Development of summarization corpus construction and evaluation methodology :

We established construction criteria and system to construct summarization corpus that can be applied to summarization system of extractive and abstractive summarization method in an organic and high efficiency method. Also we validated the established summary corpus through evaluation method of summarization corpus which has widely used and we proposed the evaluation criteria that can be qualitatively assessed.

The seek for method of utilizing the summarization corpus : We surveyed actual demand of industry and established the corpus to be used in the development of real technology by taking counsel from industry company that needs to actually develop and utilize the summarization technology. And we reviewed usability by applying the summarization corpus built on the actual automatic summarization model.

Keywords: summarization dataset, summarization corpus, newswire summarization corpus, automatic summarization, news big data

Project Director: Kim Hansaem(Yonsei University)

사업 책임자 김한샘(연세대학교)

사업 참여자 김학수(강원대학교)

이공주(충남대학교)

주민재(명지대학교)

강범일(연세대학교)

한지윤(연세대학교)

오태환(연세대학교)

강예지(연세대학교)

최현수(연세대학교)

박석원(연세대학교)

정석원(강원대학교)

장영진(강원대학교)

최기현(강원대학교)

이경호(충남대학교)

김동준(한국신문윤리위원회)

여규병(전적 동아일보)

이세강(우송대학교)

이용원(동국대학교)

이창호(전적 스포츠조선)

김민지(연세대학교)

김아영(연세대학교)

손영랑(연세대학교)

윤선영(연세대학교)

이종혁(연세대학교)

정지원(연세대학교)

정혜진(연세대학교)

조연수(연세대학교)

허나연(연세대학교)
김담린(강원대학교)
김보은(강원대학교)
박주일(슬로워크)
김형우(슬로워크)
임동근(슬로워크)
양정화(슬로워크)
담당 연구원 이승재(국립국어원 언어정보과장)
이수미(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9718
전송 02-2669-9727
인쇄일: 2019년 12월 4일
발행일: 2019년 12월 4일
인 쇄: 연세대학교 POD센터

※ 이 책은 국립국어원의 용역비로 수행한 '말뭉치 분석 연구 및 시범 구축' 사업의 결과물을 발간한 것입니다.