자연어처리 기반 콘텐츠 범주화 및 재생산에 대한 연구

최준식[1], 이서영[2], 임인호[3], 김영종*
[1,2,3,*]숭실대학교 소프트웨어학부
e-mail: 2318ws@soongsil.ac.kr[1], ww0111@soongsil.ac.kr2,
inho2736@soongsil.ac.kr3, youngjong@ssu.ac.kr*

A Study on Categorization and Reproduction of Content Based on Natural Language Processing

Junsik Choi[1], Seoyoung Lee[2], Inho Lim[3], Youngjong Kim**
[1,2,2,*]School of Software, Soongsil University

요 약

다양한 요약 프로그램이 존재하지만 입력되는 정보의 형태가 텍스트에 한정되어있다. Sumalyze는 그러한 한계점을 보완한 파이썬 언어 기반 라이브러리다. 다양한 형식의 데이터를 입력받으며 컨텐츠의 종류에 맞는 요약과 범주화를 제공한다. 또한, 컨텐츠의 핵심파악을 돕기 위해 요약에 대한 분석을 제공한다...

1. 서론

최근 자연어처리 관련 연구가 활발해지면서 다양한 오픈소스 라이브러리와 패키지가 공개되고 있다. 대표적으로 한국 정보처리를 위한 파이썬 패키지인 'KoNLPy'[1] 가 있다. 이 패키지는 입력된 문장의 형태소를 분석하고 다양한 말뭉치를 제공하고 있다.

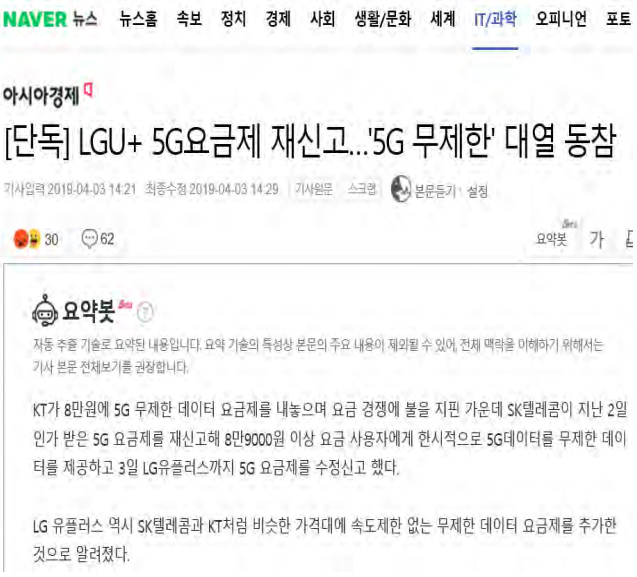
이러한 라이브러리나 패키지를 이용해 정보 요약을 제공하는 대다수의 제품들은 텍스트 형태만을 기반으로 정보를 처리한다. 하지만 최근 웹상에서 공유되는 정보들은 '텍스트' 뿐만 아니라 PDF, 이미지, 음성, 동영상과 같은 다양한 형태와 확장자를 가지고 있어, 이에 대한 요약과 학습을 원하는 사용자가 늘어나고 있다.

본 논문에서는 이런 관점에서 자연어(Natural Language)를 기반으로 정보를 분석해 범주화하고 이를 새로운 콘텐츠로 다시 생산해 내는 방식을 제안한다. 이 방식을 통해 요약의 결과 또한 자신이 원하는 양식으로 변환해서 제공하고 키워드를 제시하며 사용자의 학습이 용이하게 돕는다.

본 논문의 구조는 다음과 같다. 먼저 2장에서는 기존에 있는 다양한 요약관련 연구와 프로그램을 소개한다. 3장에서는 본 논문에서 제시하는 프로그램의 요구사항을 기술하고, 4장에서는 이에 사용된 기술과 동작원리 그리고 소프트웨어 아키텍쳐에 대해 논의한다. 마지막으로 5장에서 요약 및 결론을 내린다.

2. 관련연구 및 프로그램

쉽게 접해볼 수 있는 요약 프로그램으로 네이버 뉴스의 '요약봇'이 있다. 요약봇은 자동요약 기술 '아이리스(IRIS)'를 기반으로 서비스에 구현된다. 텍스트로 작성된 문서(기사 내용)에서 문장의 중요도를 분석해 중요한 문장을 추출해 요약해주는 서비스이다. <그림 1>에서 그 사례를 확인할 수 있다.



<그림 1> 네이버 뉴스 '요약봇' 예시

또한 사용자가 원하는 텍스트를 직접 받아 3줄로 요약해 주는 '서머라이즈3' 웹 프로그램도 주목을 받고 있다.

이 두 프로그램은 텍스트를 요약해 사용자가 긴 글을 보다 빠르게 학습할 수 있도록 도와주는데 의의를 가진다. 하지만 입력되는 정보의 형태가 텍스트에 한정되어 있다는 점에서 한계를 가진다. 또한 도출된 결과가 사용자가 보기 쉽게 정렬된 정보가 아니라 단순히 요약된 문장의 나열이라는 것에서 단점을 가진다.

3. 요구사항

"Sumalyze"는 기존 요약 서비스들의 한계를 보완한 파이썬 언어 기반의 라이브러리이다. Sumalyze는 다음 요구에 의해 개발되었다.

다양한 데이터형식 지원, 기존의 요약 서비스들은 오직 텍스트만 지원하고 자사 프로그램 내에서의 컨텐츠만 요약이 가능하도록 되어있었다. Sumalyze은 사용자 업로드 방식의 라이브러리로서 텍스트뿐만 아니라 영상, 음성, PDF, 이미지의 형식을 지원한다.

컨텐츠 종류에 맞는 요약, 단순히 텍스트의 분량을 줄이는 형식의 요약이 아닌 컨텐츠의 종류에 맞게 요약을 제공한다. 컨텐츠 종류는 소논문, 강의, 연설, 기사, 프로그램요구서, 학습서가 있다.

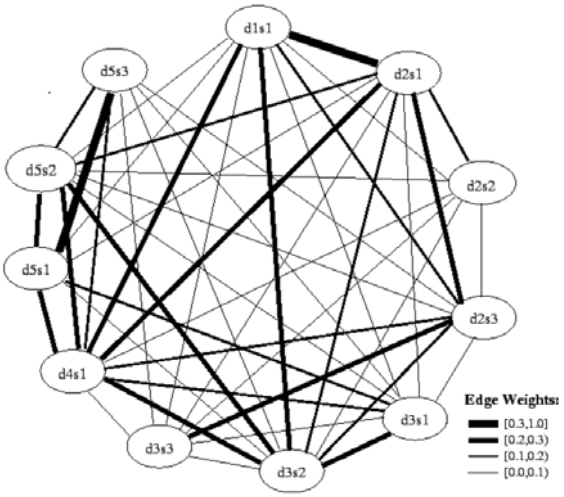
컨텐츠의 명확한 이해를 위한 범주화, 텍스트의 문단에서 핵심 문장을 추출해 컨텐츠 내용을 아우르는 목차를 생성한다. 이로써 사용자는 컨텐츠 내용의 큰 틀을 파악할 수 있다.

요약에 대한 분석 제공, 컨텐츠 요약 후 각 문단 속 문장들의 중요도를 수치화한 데이터, 빈도수가 높은 키워드와 전체 내용을 아우르는 중요개념(핵심주제)를 추출해내 요약된 컨텐츠의 내용의 핵심파악을 쉽도록 만든다.

또한 본 Sumalyze의 소스는 오픈소스로 제공되어 사용자가 원하는 방식으로 수정하여 사용할 수 있게 하였다.

4. 기술 설명 및 사용법

본 연구에서 이루어지는 요약은 LexRank 알고리즘을 한국어에 적합하도록 구현한 LexRankr 파이썬 패키지[2]를 사용한다. LexRank 알고리즘은 구글의 PageRank 알고리즘에서 착안된 방법으로 그래프 클러스터링 기반 랭킹 알고리즘이다.[3] 중요도가 높은 문장은 다른 문장으로부터 참조되는 횟수가 많다는 점에서 착안된 아이디어이다.



<그림2> LexRank의 원리

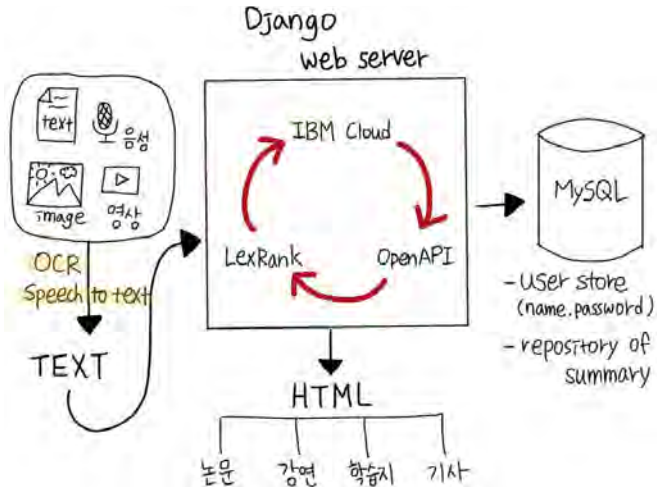
목차 생성 및 키워드 분석을 위해서는 "IBM Cloud"의 API인 자연어 이해 서비스 (Watson Natural Language Understadning)를 사용한다. 자연어 이해 서비스의 텍스트 분석을 통해, 컨셉, 객체, 키워드, 카테고리, 정서(sentiment), 감정(emotion), 관계(relation), 의미역(semantic roles) 과 같은 다양한 입력 텍스트의 의미론적 특징들을 분석할 수 있다.

Sumalyze 사용법은 다음과 같다. 먼저 요약, 분석 하고자 하는 컨텐츠를 업로드한 후 컨텐츠 종류를 선택하면, 그 컨텐츠의 데이터 형식에 맞춰 파이썬 변환 라이브러리와 STT(Speech to Text) 오픈 API를 사용하여 컨텐츠의 모든 내용을 텍스트 형태로 변환한다. 적당한 분량(이하 파트)으로 텍스트를 분할하고 자연어 이해 서비스(IBM Cloud)를 통해 각 파트의 키워드와 중요도, 빈도수 , 개체, 주제를 분석해 목차를 만든다. LexRankr를 이용해 파트 별로 문장들을 요약한 후 사용자가 입력한 콘텐츠 종류에 맞게 정해진 레이아웃에 목차와 요약된 텍스트를 배치한다. (콘텐츠 종류에 따라 상이한 주제가 선택된다. 가령, 학습서 요약의 경우 사용자의 학습을 돕기 위한 간단한 퀴즈가 추가된다.)

추가적으로 Sumalyze 웹 서비스에 가입하면 사용자 프로필에 자신이 요약한 내용을 DB로 저장해 언제든지 다시 열람할 수 있다.

4-1. Software Architecture

아래의 <그림 3>에서 Sumalyze의 소프트웨어 아키텍쳐를 확인할 수 있다.



<그림3> software architecture

5.결론

정보화 시대, 데이터의 공유가 빠르게 이루어지며 더 이상 가지고 있는 데이터의 양은 중요하지 않게 되었다. 이제는 방대한 양의 데이터를 어떻게 효율적으로 학습할 수 있는지가 관건이다. 이러한 이유로 요약 기술에 관심도가 높아지고 있고, 수많은 요약 프로그램이 개발되어 교육 분야에 다양하게 활용되고 있다.

이러한 관점에서, 사용자의 편의에 맞게 요약과 분석을 제공하는 Sumalyze 또한 좋은 교육 프로그램에 활용될 수 있을 것이다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학사업의 연구결과로 수행되었음 (2018-0-00209-001)"

참고문헌

[1] 박은정, 조성준 "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 한글 및 한국어 정보처리 학술대회 논문집, 2014

[2] 설진석, 이상구, "lexrankr - LexRank 기반 한국어 다중문서요약",한국정보과학회 동계학술발표회 , 2016.12

[3] Erkan, Gunes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research 22 (2004): 457-479.