패러프레이즈 추출을 위한 키프레이즈 데이터셋 구축 방법론 연구

강혜린[0], 강예지, 박서윤, 장연지, 김한샘†
연세대학교 언어정보학협동과정, 언어정보연구원†
{hyerink, yjkang5009, seoyoon.park, yeonji3547, khss†}@yonsei.ac.kr

A Study on the Construction of keyphrase dataset for paraphrase extraction

Hyerin Kang[O], Yejee Kang, Seoyoon park, Yeonji Jang, Hansaem Kim†
Interdisciplinary Graduate Program of Linguistics and Informatics, Yonsei University

요 약

자연어 처리 응용 시스템이 패러프레이즈 표현을 얼마나 정확하게 포착하는가에 따라 응용 시스템의 성능 측면에서 차이가 난다. 따라서 자연어 처리의 응용 분야 전반에서 패러프레이즈 표현에 대한 중요성이 커지고 있다. 시스템의 성능 향상을 위해서는 모델을 학습시킬 충분한 말뭉치가 필요하다. 특히 이러한 패러프레이즈 말뭉치를 구축하기 위해서는 정확한 패러프레이즈 추출이 필수적이다. 따라서 본 연구에서는 패러프레이즈를 추출을 위한 언어 자원으로 키프레이즈 데이터셋을 제안하고 이를 기반으로 유사한 의미를 전달하는 패러프레이즈 관계의 문장을 추출하였다. 구축한 키프레이즈 데이터셋을 패러프레이즈 추출에 활용한다면 본 연구에서 수행한 것과 같은 간단한 방법으로 패러프레이즈 관계에 있는 문장을 찾을 수 있다는 것을 보였다.

주제어: 패러프레이즈, 패러프레이즈 추출, 키프레이즈, 키프레이즈 데이터셋

1. 서론

하나의 사건을 표현할 때에 이를 표현할 수 있는 방법은 매우 다양하다. 서로 다른 표현을 사용하여 유사한 의미를 전달하는 두 개 이상의 문장을 패러프레이즈(paraphrase)라 한다. 패러프레이즈 표현은 언어 전반에서 다양하게 나타나는 현상이기 때문에 자연어처리 분야에서도 주목을 받고 있다. 기계가 문장을 이해하고 패러프레이즈 문장들을 같은 문장으로 인식해야 하기 때문이다. 패러프레이즈 표현을 기계가 얼마나 정확하게 포착하는가에 따라 응용 시스템의 성능에 큰 영향을 미칠 수 있기 때문에 문장 간 관계가 패러프레이즈인지를 판별하기 위해서는 모델을 학습시킬 양질의 말뭉치가 필요하다. 최근에는 모델 학습을 위한 일정 수준의 품질을 보장하는 말뭉치를 빠르게 확보할 필요성이 있는데[1] 현재 공개된 한국어 패러프레이즈 말뭉치의 수는 필요성에 비해 충분하지 않다. 또한 공개된 말뭉치라고 해도 새로운 패러프레이즈 표현 정보가 새로운 패러프레이즈 표현 정보가 말뭉치에 실시간으로 반영되는 것은 현실적으로 어렵다. 이러한 패러프레이즈 말뭉치를 구축하기 위해서는 정확한 패러프레이즈 추출이 필수적이다. 패러프레이즈 추출에서 문장 유사도가 높은 문장 쌍들이 추출된다면 이러한 문장들을 모아 양질의 패러프레이즈 말뭉치를 구축할 수 있다. 따라서 질 높은 패러프레이즈 말뭉치를 위해서는 패러프레이즈 추출의 방법 또한 중요하다.

본 연구에서는 패러프레이즈 추출의 중요성을 인식하고 한국어 패러프레이즈 말뭉치 구축을 위해 패러프레이즈를 추출하는 새로운 방법을 제안하고자 한다. 본 연구에서 제안하는 것은 키프레이즈 데이터셋을 활용하는 방법론이다. 키프레이즈(keyphrase)란 해당 문장이나 문서의 핵심을 나타내는 구(phrase)를 말한다. 키프레이즈 데이터셋은 패러프레이즈 관계에 있는 문장들은 키프레이즈를 공유할 것이라는 가정과 개체명 기반의 패러프레이즈 추출 접근법을 접목시켜 본 연구에서 새롭게 제안하는 개념이다. 패러프레이즈 추출에 정성적 방법으로 구축된 키프레이즈 데이터셋을 활용한다면 추출이 훨씬 쉽고 간편할 것이며 서로 유사도 높은 문장을 찾을 수 있을 것이다. 따라서 본 연구에서는 개체명을 기반으로 키프레이즈 데이터셋을 구축하고 이를 활용하여 패러프레이즈를 추출하고자 한다.

현재 많은 연구에서 키프레이즈와 키워드라는 용어를 혼용하고 있다. 키프레이즈란 구를 전제로 하는 용어이지만 한 단어인 키워드를 표현할 때에도 키프레이즈라는 용어를 빈번하게 사용하고 있다. 키프레이즈 추출(keyphrase extraction)과 키워드 추출(keyword extraction)의 용어 사용이 혼용되고 있다는 것이다. [2]에서는 키프레이즈와 키워드의 차이에 대해 서술하면서 키프레이즈는 multi-word lexeme, keyword는 single word term이라고 정의하였다. 구 단위로 이해해야 하는 단어를 single word 단위에서 해석하려고 하면 오류가

발생할 수 있다고 하면서 키프레이즈와 키워드를 분리하여 용어 사용을 해야 한다고 하였다. 본 연구에서는 mult-word lexeme를 토대로 하여 개체명과 결합된 형태를 패러프레이즈 추출의 기준으로 삼고자 하기 때문에 키프레이즈라는 용어를 사용한다.

2. 관련 연구

개체명에 기반하여 패러프레이즈 추출에 접근한 연구들은 특정한 날짜에 발생한 사건에 대한 기사는 같은 사건을 표현한 것이지만 다른 언어로 표현된 것이 많을 것이라는 가정을 전제로 한다. [3]에서는 개체명이 나타난 신문 기사문을 대상으로 하여 개체명, 구문 주석 작업을 진행하였다. 이후 구문 분석된 문장들을 각각의 노드들과 연결하여 문장 내에서 사용 가능한 패턴만을 추출하였다. 추출한 패턴들 내에는 개체명이 포함되어 있는데 일반화된 패러프레이즈 패턴을 추출하기 위해 PERSON, LOCATION과 같은 개체명 대분류로 각 논항을 교체하는 작업을 거쳤다. 이 과정을 통해 구축된 후보 패러프레이즈 패턴들을 TF-IDF에 기반한 문장 유사도를 사용하여 최종 패러프레이즈 패턴 선별 작업을 진행하였다. 개체명 기반의 패러프레이즈 접근법의 기초를 마련했다는 것에서 의의가 있다. [4]는 최신 영어 신문 기사에서 개체명이 같은 문장은 추출한 후 후보 패러프레이즈 문장에서 개체명을 제외한 나머지 명사, 부사, 동사, 형용사 간의 Jaccard Coefficient 값의 조화평균으로 두 문장 사이의 유사도를 구하여 최종 패러프레이즈로 판별하였다.

[5]에서는 질문과 요구의 목적을 가진 지시성 발화를 대상으로 하여 핵심 내용인 키프레이즈를 추출하고 추출한 키프레이즈를 통해 데이터를 증강하는 방법론을 제안하였다. 지시 발화의 키프레이즈를 추출하는 방법으로 담화 성분을 이용하여 연구에 적합한 원리를 목록화하여 제시하였다. 구조화된 방식으로 추출한 키프레이즈들은 하나의 키프레이즈 집합을 이루고 이를 데이터 증강에 활용하였다. 주석자들은 키프레이즈 하나당 열 개의 발화를 생성하였고 다양한 발화가 생성될 수 있도록 구체적인 가이드라인도 마련하였다. 이러한 과정을 통해 키프레이즈와 질문/요구 문장 쌍을 구축하였다. 이러한 데이터셋을 구축함으로써 같은 담화 성분에서 생성된 문장들이 모두 같은 핵심 내용(키프레이즈)를 공유하기 때문에 각각의 문장 쌍들은 모두 패러프레이즈 관계에 있다는 것을 보였다. 이는 키프레이즈 연구가 패러프레이즈 연구로까지 확장될 수 있다는 것을 보여줬다는 점에서 의미가 있다.

개체명이 일치한다고 모든 문장이 패러프레이즈 관계에 있는 것은 아니다. [3],[4]에서와 같이 문장 간의 유사도를 확인하여 패러프레이즈인지를 확인하는 과정을 거쳐야 한다. 이는 개체명 기반으로 패러프레이즈를 추출할 경우 항상 그 문장이 유사하거나 같은 의미를 전달하지는 않는다는 것이다. 본 연구에서는 개체명 기반의 패러프레이즈 추출 접근법이 가진 한계와 패러프레이즈 추출 과정의 불완전성을 보완하고자 [5]에서 보인 문장 간의 높은 유사도를 나타내는 기준이 키프레이즈라는 증명을 활용하고자 한다.

3. 키프레이즈 데이터셋의 구성

키프레이즈 데이터셋 구축의 구체적인 방법에 관해 살펴보기 전에 키프레이즈 데이터셋의 개념과 기본 구성, 표현 방식에 대해 살펴본다. 키프레이즈 데이터셋은 담화 성분을 활용하여 지시 발화의 키프레이즈를 추출한 [5]와 개체명을 기반으로 하여 패러프레이즈를 추출한 [3], [4]의 방법을 접목시킨 개념이다.

키프레이즈 데이터셋은 1차, 2차, 3차 분류로 구성된 계층 구조를 가진다. 1차, 2차는 텍스트에 나타난 개체명이며 3차는 본 연구에서 제안하는 방법론으로 구성된 키프레이즈이다. 이 키프레이즈는 텍스트의 핵심 내용을 담고 있다. 그림 1은 키프레이즈 데이터셋의 계층 구조를 나타낸 것이다.

그림 1 키프레이즈 데이터셋의 구조

표 1 1차, 2차 분류 개체명의 분류와 설정 방법

1차 분류	국가명, 도시명 등 기관, 회사명 등 소속 관계로 정의할 수 없는 둘 이상의 사람이 등장할 경우 그 중 한 사람의 이름 가능 예) 키프레이즈의 주체가 소속된 기관, 소속 관계가 아닌 대등한 관계일 경우 기사문에서 빈도수가 더 높은 개체명, 사건이 일어나는 장소
2차 분류	사람 이름, 사건 이름, 정책 등의 명칭, 질병, 약품명 등

1차, 2차 분류 개체명은 키프레이즈와의 문장 내의 연관 관계를 통해 설정해야 한다. 개체명과 키프레이즈 간의 관계가 파악되었다면 1차, 2차의 관계성을 따져 각각 분류해야 한다. 1차, 2차의 구체적인 분류는 다음 표 1과 같다. 이를 통해 1차, 2차, 3차 분류를 설정한 예시는 다음 그림 2와 같다.

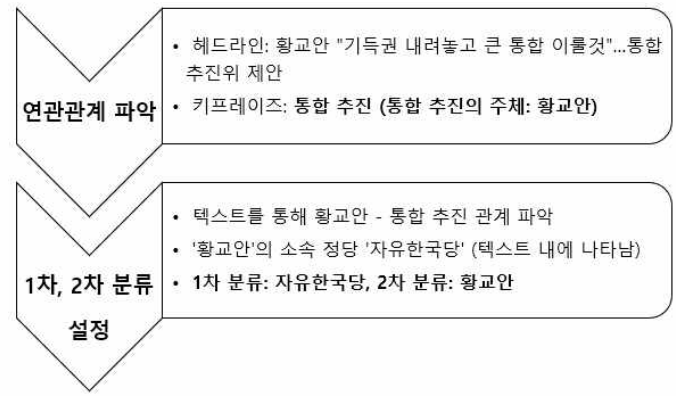

그림 2 1차, 2차 분류 개체명의 분류와 설정 방법의 예시

1차, 2차 분류 개체명은 패러프레이즈 관계에 있는 문장을 찾을 때에 대상이 되는 텍스트의 범위를 좁혀주는 역할을 한다. 마지막으로 3차 분류는 키프레이즈이다. 패러프레이즈 관계에 있는 문장은 유사한 의미를 가진 서로 다른 언어 표현을 사용하여 표현되기 때문에 키프레이즈와 유사한 의미를 가지는 표현도 추가해야 한다. 구체적인 방법은 4장에서 다룬다. 키프레이즈와 유사한 의미를 가지는 단어를 찾아 3차 분류 키프레이즈에 추가한다.

1차, 2차, 3차 분류로 구성된 키프레이즈 데이터셋은 키프레이즈 추출의 대상이 된 텍스트의 핵심 문장과 일치하는 형태를 보인다. 계층 구조를 표현한 그림 1에서도 알 수 있듯이 키프레이즈 데이터셋을 활용할 때에는 3차 분류인 키프레이즈를 공유하는 문장을 찾기 위해서는 상위 개체명도 동일하게 나타나야 한다. 인명, 기관명 등의 경우 줄임말을 사용하여 표현하는 경우가 있는데, 이런 경우 모두 데이터셋에 입력하여 동일 개체명 추출에 이용할 수 있도록 한다. 키프레이즈 데이터셋을 통해 패러프레이즈 관계에 있는 문장들을 찾기 위해서는 1차, 2차, 3차 분류가 함께 결합된 형태로 문장을 찾아야 한다.

패러프레이즈를 추출할 때에는 키프레이즈 내의 단어 중 키프레이즈 내에서 '-하다', '-되다'가 붙어 서술성을 가지는 것으로 해석되는 것은 제외하고 활용한다. 예를 들어, 키프레이즈가 '시민 참여'일 경우 '시민이 참여하다'로 해석되고, '일본 탈출'의 경우에는 '일본을 탈출하다'로 해석된다. '시민, 일본'의 경우 명사이기 때문에 서술성 명사를 쉽게 변별할 수 있지만, '가동 중단, 통합 추진'과 같은 경우는 모두 서술성 명사인 구조인데 이러한 경우 'A 이/가 혹은 을/를 B하다/되다'의 키프레이즈 구조를 고려하여 해당 키프레이즈에서 서술성을 가지지 않는 것만을 가려내 활용한다. 이러한 과정을 거치는 이유는 패러프레이즈 관계에 있는 문장들의 서술어의 경우 형태가 변하기 때문에 키프레이즈로 모두 찾아낼 수 없고 키프레이즈에 나타난 서술성 명사로만 패러프레이즈를 찾을 경우 다양한 패러프레이즈 동사 패턴을 포착할 수 없기 때문이다. 따라서 본 연구에서는 개체명과 3차 분류 키프레이즈 내에서 서술성을 가지지 않는 것만을 활용한다. 패러프레이즈 추출 이후 문장 쌍을 구성할 때에는 키프레이즈 데이터셋 내의 모든 단어를 사용하여 기준으로 삼는다.

또한 2차 분류 개체명이 1차 분류 개체명의 의미를 포함하고 있는 경우나 1차 분류 개체명이 전제되는 경우에는 2차 분류와 키프레이즈만의 조합으로도 패러프레이즈 추출이 가능할 것이다.

4. 키프레이즈 데이터셋 구축 방법

4.1. 데이터

패러프레이즈 추출을 위해 본 연구에서 제안하는 키프레이즈 데이터셋은 개체명 기반의 패러프레이즈 추출 접근법을 따른다. 개체명 기반의 패러프레이즈 추출을 위한 키프레이즈 데이터셋 구축의 대상이 되는 텍스트는 개체명이 많이 쓰이고 잘 드러나는 뉴스 기사문이 가장 적합하다고 판단하여 네이버 연합뉴스 속보 페이지에서 대상 기사문을 선정하였다. 키프레이즈 데이터셋에 최대한 다양한 주제를 포함하기 위하여 매일 주요 뉴스들을 편집하여 구성하는 네이버 뉴스 연합뉴스 속보 페이지를 대상으로 삼았다. 해당 페이지는 그날에 크게 이슈가 되었던 내용이나 주요 내용들을 고르게 담고 있다. 키프레이즈 데이터셋 구축의 대상이 되는 기사의 출처는 기사 주제 및 구체화된 개체명의 중복을 지양하기 위해 연합뉴스 한 곳으로 제한하였다.

4.2. 개체명 태그셋에 따른 기사문 선정

본 연구는 개체명 기반의 패러프레이즈 추출을 중심으로 하기 때문에 개체명의 개념과 분류 체계가 중요하다. 본 연구는 TTA 개체명 태그셋을 따르되 본 연구의 목적에 맞게 이를 재구성하였다. '네이버 뉴스- 연합뉴스 속보' 페이지 내에 배치된 기사문의 특성을 파악하여 주요하게 나타나는 개체명을 대상으로 하여 태그셋을 재구성하였다. 본 연구의 목적에 맞게 재구성한 TTA 개체명 태그셋은 표 2와 같다.

기사 헤드라인에 표 2에 해당하는 개체명이 있다면 키프레이즈 데이터셋 구축 대상 기사문으로 선정하여 작업을 진행하였다.

표 2 TTA 개체명 태그셋을 본 연구의 목적에 맞게 재구성한 개체명 목록

대분류	정의
PERSON (PS)	사람 이름
LOCATION (LC)	국가명, 지역명, 도시명, 수도명 등
ORGANIZATION (OG)	경제, 교육, 군사, 스포츠, 법률 등 관련 기관/ 단체
CIVILIZATION (CV)	조세 / 제도 / 정책 / 직업 등 명칭
EVENT (EV)	사회 운동 및 선언, 전쟁/ 혁명, 축제 명칭, 스포츠/ 레저 관련 행사
TERM (TM)	증세/증상/질병, 약/약품 명 등

4.3. 키프레이즈 데이터셋 구축 방법론

선정한 기사문을 대상으로 하여 키프레이즈 데이터셋을 구축하기 위하여 키프레이즈 추출 알고리즘과 키워드 알고리즘을 이용하였다. 키프레이즈 추출 알고리즘은 TextRank이며 키워드 추출 알고리즘은 LDA와 KR-WordRank이다. KR-WordRank는 [6]에서 제안한 방법으로 일본어와 중국어의 단어 분할을 위해 제안된 방법인 WordRank를 한국어의 언어적 특성에 맞게 개선한 것이다. 한 문서에 나타난 단어들 중 명사, 동사, 형용사, 부사 중에서 빈도수가 높거나 주요 단어들과의 사용이 빈번한 단어를 키워드로 추출하는 것이 주요 개념이다. 후보 단어를 기반으로 단어 그래프를 생성한 뒤 하이퍼링크를 가지는 웹 문서에 상대적 중요도에 따라 가중치를 부여하는 PageRank 알고리즘 내 랭킹 순 학습 방식을 이용하여 키워드를 추출한다. 본 연구에서 제안하는 키프레이즈 데이터셋 중 3차 분류에 해당하는 키프레이즈를 표현하기 위한 방법론은 1) TextRank의 키프레이즈 추출 결과만 활용 2) TextRank의 키프레이즈 추출 결과와 LDA, KR-WordRank의 키워드 추출 결과 조합으로 나눌 수 있다. 1)에서 해당 기사문의 내용을 모두 표현할 수 있는 키프레이즈 결과가 나왔다면 TextRank의 결과만으로 키프레이즈를 표현한다. 이러한 판단의 기준은 TextRank에서 제공하는 핵심 문장 추출 기능에 있다. TextRank는 문장 중요도 순서로 결과를 출력하는데, 이 중 상위 5개의 문장을 TextRank의 결과가 포괄할 수 있다면 1)의 방법만을 사용하였다. TextRank의 키프레이즈 추출 결과 중 핵심 내용과 상위 5개 문장에 가장 적합한 것을 선택하였다. 1)의 방법의 예는 다음 표 3과 같다.

표 3 1) TextRank의 키프레이즈 추출 결과만을 활용할 수 있는 예

기사의 헤드라인	<코로나 여파에…대한항공 객실승무원 최대 1년 무급휴직한다> (연합뉴스, 2020.06.11.)[1]
TextRank 결과	**무급 휴직**

표 3을 토대로 키프레이즈 데이터셋을 구성한다면, 키프레이즈 '무급 휴직'은 '객실 승무원'이 하는 것이고 객실 승무원의 소속은 '대한항공'이다. 따라서 1차 분류는 '대한항공'이 되고 2차 분류는 '객실 승무원'이 된다. 이렇게 키프레이즈 데이터셋을 구성한다면 기사의 핵심 내용을 모두 나타낼 수 있다.

TextRank의 키프레이즈 추출만으로 해당 기사문의 핵심 내용을 나타낼 수 있는 키프레이즈가 추출되지 않았다면 2)로 넘어가게 된다. 2)는 TextRank를 통해 계산된 주요 문장 상위 5개와 키프레이즈와 LDA, KR-WordRank 키워드 추출 결과를 조합하는 방법이다. TextRank에서 추출된 키프레이즈에 부가적으로 추가되어야만 기사문의 핵심 내용을 나타낼 수 있을 때에 이 방법을 따른다. 2)의 방법의 예시는 다음 표 4와 같다.

표 4 2) TextRank의 키프레이즈 추출 결과와 LDA, KR-WordRank의 키워드 추출 결과를 모두 활용해야 하는 예

기사의 헤드라인	<트럼프, '백악관 속살 폭로' 볼턴 회고록에 출판금지 소송> (연합뉴스, 2020.06.17.)[2]
TextRank 결과	누설 금지, 기밀 누설, 안보 보좌관, 국가 안보, **회고록 출간**, 책 출간
LDA 결과	**볼턴**, 회고록, 백악관, 검토, 출판, 안보, 법무부, 기밀, **금지**, 국가, **소송**
KR-WordRank 결과	회고록, **볼턴**, **제기**, 법무부, 백악관, 출간, 출판

1) https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=001&aid=0011670652&isYeonhapFlash=Y&rc=N

표 4를 보면, TextRank의 결과만으로 기사의 핵심 내용을 모두 담을 수 없다. 따라서 2)의 방법에 따라 LDA 결과와 KR-WordRank의 결과를 모두 참고하여 키프레이즈를 조합하여야 한다. 키프레이즈를 조합할 때에는 TextRank 핵심 문장 상위 5개 문장에 나타난 기사의 핵심 내용이 모두 표현될 수 있도록 해야 한다. TextRank의 결과와 LDA, KR-WordRank의 결과를 조합한 키프레이즈와 텍스트 내의 개체명 간의 관계를 파악하여 1차, 2차 분류를 설정해야 한다. '볼턴의 회고록이 출간되는 것의 금지 소송을 제기'한 것은 트럼프이기 때문에 '트럼프'가 키프레이즈 데이터셋의 개체명으로 추가되어야 한다. '트럼프'는 미국과의 소속 관계를 가진다. 따라서 1차 분류는 '미국'으로 설정한다. 이러한 기준에 따라 키프레이즈 데이터셋을 구성하면 다음 표 5와 같다.

표 5 2)의 방법으로 구축된 키프레이즈 데이터셋

1차 분류	2차 분류	3차 분류
미국	트럼프	볼턴 회고록 출간 금지 소송 제기

구축된 키프레이즈 데이터셋은 패러프레이즈 추출 활용에 그 목적이 있기 때문에 의미가 유사한 단어를 데이터셋에 추가해야 한다. 이를 위해 의미에 기반하여 키프레이즈를 구성하는 어휘와 유의 관계에 있는 어휘를 교체한 키프레이즈를 추가하였다. 이를 위해 우리말샘 사전의 어휘 지도와 한국어 어휘 의미망(KorLex)를 사용하여 연관 어휘를 추가하였다.

표 5 키프레이즈 데이터셋을 예로 들어 이 과정을 설명하면 3차 분류 키프레이즈에 해당하는 '회고록 출간 금지 소송' 중 유의어 쌍을 가지면서 기사에서 패러프레이즈 표현으로 사용되는 단어를 파악해야 한다. '출간'의 어휘 관계를 파악하여 교체 가능한 유사한 의미의 단어를 추가해야 한다. '우리말샘 어휘지도'에 '출간'을 검색하면 비슷한 말로 '간행, 출판, 간출'이 어휘 지도에 나타나 있다. 이 세 단어를 모두 키프레이즈와의 유의 관계에 볼 수는 없는데, 그 이유는 '출간'이 '금지 소송'과 함께 쓰였기 때문이다. '출판 금지 소송'으로는 빈번하게 사용되지만 '간행, 간출 금지 소송'으로는 쓰이지 않기 때문에 '출판'만을 해당 키프레이즈의 '출간'과 교체 가능한 것으로 하여 최종 키프레이즈 데이터셋 3차 분류에 '회고록 출판 금지 소송'을 추가하여 준다. 이러한 과정을 통해 구축된 최종 키프레이즈 데이터셋은 표 6과 같다.

키프레이즈 유의어 추가 과정을 거치면 1차, 2차, 3차 분류의 최종 키프레이즈 데이터셋 구축이 완료된다.

2) https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=004&oid=001&aid=0011683507

표 6 키프레이즈 유의어가 추가된 키프레이즈 데이터셋

1차 분류	2차 분류	3차 분류
미국	트럼프	볼턴 회고록 출간 금지 소송 제기 볼턴 회고록 출판 금지 소송 제기

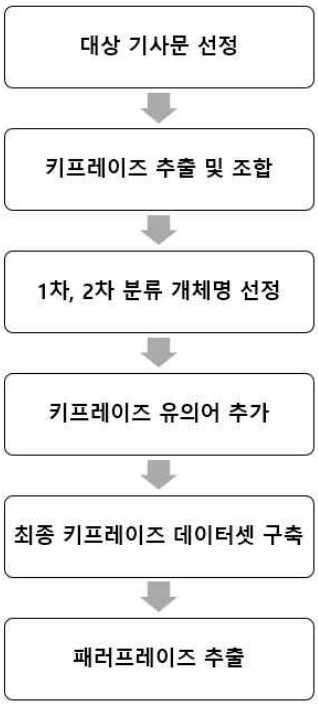

그림 3 연구 방법 도식화

4장의 전체적인 연구 방법의 흐름을 도식화한 것은 그림 3과 같다.

5. 키프레이즈 데이터셋 기반 패러프레이즈 추출

최종적으로 구축한 키프레이즈 데이터셋을 이용하여 패러프레이즈를 추출한다. 패러프레이즈를 추출하기 위해 키프레이즈 데이터셋 구축 대상 기사문의 기간에 해당하는 네이버 뉴스에 서비스를 제공하는 모든 언론사의 기사문을 크롤링하였다. 크롤링된 결과 중 기사문만을 추출하여 텍스트로 저장하였다. 이를 정규 표현식을 이용하여 키프레이즈 데이터셋 1차, 2차 개체명과 3차 분류인 키프레이즈 내에서 서술성이 없는 것으로 해석되는 것만을 대상으로 이 단어들이 모두 포함된 문장을 검색

하고 해당 문장들만을 추출하였다. 1차 개체명이 2차 개체명을 전제하고 있는 경우나 2차 개체명이 1차 개체명을 전제하고 있는 경우에는 하나의 개체명만 선택하였다. 다음 표 7은 최종 키프레이즈 데이터셋으로 구축된 표 6의 개체명과 키프레이즈를 결합한 형태로 검색하였을 때 추출되는 패러프레이즈 문장들이다. 표 5의 경우, 2차 분류인 '트럼프'는 '미국'이라는 의미가 전제되기 때문에 개체명은 '트럼프'만을 선택하였고 3차 분류 키프레이즈의 경우 '볼턴의 회고록을 출간(출판)하는 것을 금지하는 소송을 제기하다'의 의미가 되기 때문에 서술성이 있는 '출간, 금지, 제기'는 제외한 '볼턴, 회고록, 소송'만을 패러프레이즈 추출 시에 사용하였다. 이를 통해 추출된 문장 중 표 6의 키프레이즈 데이터셋을 기반으로 하여 문장을 선별하였다.

표 7 키프레이즈 데이터셋 기반을 활용하여 패러프레이즈를 추출하였을 때의 예

존 볼턴 전 미국 백악관 국가안보회의 NSC 보좌관의 회고록을 두고 트럼프 대통령이 출판금지 소송을 제기했다.
도널드 트럼프 미국 대통령이 존 볼턴 전 전 백악관 국가안보보좌관의 회고록 출판을 금지해달라는 소송을 제기했다.
도널드 트럼프 미국 대통령이 결국 존 볼턴 전 백악관 국가안보보좌관의 회고록 출판을 막기 위해 소송을 제기했다.

표 7에 나타난 문장들은 표 6의 키프레이즈 데이터셋을 공유하고 있는 것을 확인할 수 있다. 이는 패러프레이즈 관계에 있는 문장은 서로 유사하거나 같은 키프레이즈를 공유한다는 것을 보여주는 결과이다.

기존 개체명 기반의 패러프레이즈 추출 접근법에서의 방법론은 같은 개체명을 가진 여러 문장에서 패러프레이즈 관계에 있는 문장을 찾기 위해 정제 과정을 거쳐야 했다. 하지만 본 연구는 패러프레이즈 관계에 있는 문장들이 서로 공유하고 있는 키프레이즈를 찾고 개체명과의 연관 관계를 통해 키프레이즈 데이터셋을 구축하였다. 이를 토대로 문장을 추출하여 별도의 정제 과정을 생략하고도 표 7과 같이 패러프레이즈 문장을 추출될 수 있음을 보였다.

6. 결론

본 연구에서는 패러프레이즈를 추출하기 위하여 키프레이즈 데이터셋을 제안하고 이를 기반으로 패러프레이즈를 추출하였다. 패러프레이즈 말뭉치를 구축하기 위해서는 패러프레이즈 관계에 있는 문장을 추출하는 과정이 필수적이다. 기존 패러프레이즈 말뭉치 구축 과정에서는 추출된 문장의 유사도 정도에 따라 점수를 매기고 계속해서 정제하는 과정을 거쳤다. 그러나 이것은 말뭉치 구축 과정에서 추출된 문장이 모두 패러프레이즈 관계에 있다는 것이 아니라는 것을 반증한다. 따라서 본 연구에서는 이러한 과정을 최소화하고 문장을 처음 추출할 때부터 패러프레이즈 관계에 있는 문장을 추출하고자 키프레이즈 데이터셋 방법론을 제안하였다. 키프레이즈 데이터셋은 텍스트에 개체명 기반으로 접근하여 대상 텍스트의 범위를 좁히고 패러프레이즈 관계에 있는 문장은 서로 유사하거나 같은 키프레이즈를 공유한다는 언어학적인 가정을 바탕으로 한 것이다. 구축한 키프레이즈 데이터셋을 활용한다면 본 연구에서 수행한 것과 같은 간편한 방법으로 패러프레이즈 관계에 있는 문장을 추출할 수 있다는 것을 보였다. 이후의 연구에서는 본 연구에서 구축한 방법론을 바탕으로 양질의 키프레이즈 데이터셋을 구축하고 패러프레이즈 추출에 활용하여 패러프레이즈 말뭉치 구축까지 나아가고자 한다.

참고문헌

[1] 오교중, 김현민, 고보원, 남제현, 최호진, "문장 유사성 분석, 문장 유사성 분석을 위한 한국어 패러프레이즈 말뭉치 및 구축 가이드라인", 제 31회 한글 및 한국어 정보처리 학술대회 논문집, 527-530, 2019.

[2] Siddiq i, Sifatullah, and Aditi Sharan, "Keyword and keyphrase extraction techniques: a literature review", International Journal of Computer Applications 109.2 ,2015.

[3] Shinyama, Yusuke, et al, "Automatic paraphrase acquisition from news articles", Proceedings of HLT, Vol. 2. San Diego, US, 2002.

[4] 김광준, "최신 신문 코퍼스 기반 자동 패러프레이즈 문장 생성 기법", 서강대학교 대학원 석사 학위 논문, 2015.

[5] 조원익, 문영기, 김종인, 김남수, "담화 성분을 활용한 지시 발화의 키프레이즈 추출: 한국어 병렬 코퍼스 구축 및 데이터 증강 방법론", 제 31회 한글 및 한국어 정보처리 학술대회 논문집, 241-245, 2019.

[6] 김현중, 조성준, 강필성, "KR-WordRank : WordRank를 개선한 비지도학습 기반 한국어 단어 추출 방법", 대한산업공학회지, 40(1), 18-33, 2014.