뉴스 기사 키워드 추출을 위한 구묶음 주석 말뭉치 구축

김태영†, 김정아‡, 김보희‡, 오효정○§

국민연금공단†, ㈜엔씨소프트‡, 전북대학교§
fnty127@hanmail.net†, kjeongah@ncsoft.com‡, bohui09@ncsoft.com‡, ohj@jbnu.ac.kr§

Chunking Annotation Corpus Construction for Keyword Extraction in News Domain

Tae-Young Kim†, Jeong Ah Kim‡, Bo Hui Kim‡, Hyo Jung Oh ○§
National Pension Service†, NCSOFT‡, Jeonbuk University§

요 약

빅데이터 시대에서 대용량 문서의 의미를 자동으로 파악하기 위해서는 문서 내에서 주제 및 내용을 포괄하는 핵심 단어가 키워드 단위로 추출되어야 한다. 문서에서 키워드가 될 수 있는 단위는 복합명사를 포함한 단어가 될 수도, 그 이상의 묶음이 될 수도 있다. 한국어는 언어적 특성상 구묶음 개념이 적용되는데, 이를 통해 주요 키워드가 될 수 있는 말덩이 추출이 가능하다. 따라서 본 연구에서는 문서에서 단어뿐만 아니라 다양한 단위의 키워드 묶음을 태깅하는 가이드라인 정의를 비롯해 태깅도구를 활용한 코퍼스 구축 방법론을 고도화하고, 그 방법론을 실제로 뉴스 도메인에 적용하여 주석 말뭉치를 구축함으로써 검증하였다. 본 연구의 결과물은 텍스트 문서의 내용을 파악하고 분석이 필요한 모든 텍스트마이닝 관련 기술의 기초 작업으로 활용 가능하다.

주제어: 구묶음, 주석 말뭉치, 키워드 추출, 뉴스 기사, 언어자원 구축

1. 서론

빅데이터 시대에서 대용량 문서의 의미를 빠르게 특히 자동적으로 파악하는 일은 중요 연구 과제로서, 가장 보편적으로 활용되는 방법은 키워드 자동 추출 및 분석이다[1]. 문서 내용 파악을 위해서는 문서 내에서 키워드를 적절하게 추출하는 것이 중요하다. 키워드 추출(Keyword Extraction)은 각 문서에서 주제와 내용을 포괄하는 핵심 단어를 추출하는 것으로, 뉴스에서 중요한 정보를 추출할 때 매우 중요한 역할을 한다[2]. 문서에서 키워드가 될 수 있는 단위는 단어가 될 수도, 그 이상의 단위가 될 수도 있다. 즉, 단어 이외에 구(phrase) 단위에서도 키워드 추출이 가능하다.

한국어는 언어적 특성상 구묶음(chunking) 개념이 적용되는데, 구묶음은 문장 내에서 말덩이(chunk)라고 하는 구성성분을 찾는 과정이다. 말덩이는 문장 내에서 문법적으로나 의미적으로 같은 역할을 하는 일련의 형태소로서[3], 주요 키워드가 될 수 있다. 이에 본 연구에서는 구문 분석 결과를 기반으로 문서에서 단어뿐만 아니라 다양한 단위의 키워드 묶음을 태깅하는 가이드라인을 정의하였으며, 태깅도구를 활용해 수작업 검증을 최소화하는 말뭉치 구축 방법론을 고도화하였다. 나아가 해당 가이드라인을 실제 뉴스 도메인에 적용하여 통계기반 자동 구묶음 결과를 도구를 통해 보완하는 방식으로 주석 말뭉치를 구축함으로써 제시한 방법론을 검증하였다.

2. 키워드 추출을 위한 구묶음 주석 연구

본 연구의 궁극적인 목적은 뉴스 기사의 내용을 파악하여 대표 키워드를 추출하여 이용자에게 일목요연하게 제공하거나, 뉴스에 자주 나타나는 주요 이슈에 대한 관용표현을 파악하는데 있다. 따라서 본 연구에서는 N사에서 제공한 Y사 뉴스 기사를 주석 대상으로 선별하고, 키워드 추출을 위한 기본적인 구묶음 주석 가이드라인을 작성하였다. 이후 해당 기본 가이드라인에 의거하여 사전에 통계기반으로 자동 구묶음된 후보군에 대한 오류를 태깅도구를 통해 수작업으로 검증하고 보완하였다.

2.1 구묶음 주석 가이드라인 작성

뉴스 기사로부터 키워드를 추출하기 위해 본 연구에서는 가장 먼저 구묶음 주석 대상에 대한 기본 가이드라인을 작성하였다. 가이드라인에서는 각 언어형상에 대한 구체적인 관찰과 유형별 가이드라인을 제시하였다. 기본 가이드라인의 주석 대상에는 체언류, 명사구, 동사/형용사 파생 접사가 붙은 체언 어근, 개체명, 기타 명사처럼 쓰이는 외국어, 외래어, 숫자, 기호 등이 포함되었다. 특히 구의 길이가 너무 길 경우 하나의 단위에 너무 많은 정보가 담기게 될 수 있기 때문에, 작업자 간의 의견 조율 등 다양한 측면을 고려하여 직관적인 기준인 어절 수를 제한하는 방식(본 연구에서는 5어절)을 채택하였다. 1차 가이드라인 수립 후, 주석 결과물에 대한 검증을 진행하면서 구묶음 주석 가이드라인이 보다 상세화되었다.

아래 <표 1>은 상세화된 기본 가이드라인 내용을 요약

한 것이다.

<표 1> 구묶음 주석 가이드라인 요약

1	구묶음 단위는 파싱(parsing)을 통해 문장구조를 반영하여 진행하며, 수정방향은 '최대 범위 묶음'으로 한다.(단, 하나의 chunk가 5어절 미만여야 한다.) - 하나의 구묶음은 명사를 수식하는 어구이다. - 구문분석을 통해 2개의 수식구가 하나의 명사 head를 꾸미는 경우에는 가까운 수식구는 head에 결합하여 묶고, 먼 수식구는 분리한다. - 구문분석을 통해 1개의 수식구가 하나의 명사 head를 꾸미는 경우에는 가장 상위의 수식구를 head에서 분리한다.
2	체언류, 명사구, 동사/형용사 파생접사가 붙은 체언 어근, 개체명, 기타 명사처럼 쓰이는 외국어, 외래어, 숫자, 기호 등 문장을 이루는 전체 요소들을 대상으로 한다.
3	하나의 개체(entity), 행위, 상태 등으로 직관적으로 인식되는 구를 하나의 구묶음으로 묶는다.
4	코퍼스 구축은 기존 통계기반 구묶음 결과(initial chunk)를 기준으로 수정 작업을 진행하는 것으로 한다.

2.2 뉴스 도메인 구묶음 주석 말뭉치 구축

본 연구에서는 N사에서 제공한 Y사 뉴스 기사의 8가지 도메인(경제, 국제, IT, 생활, 논평, 정치, 사회, 스포츠)을 주석 대상으로 선정하였다. 그리고 앞서 정의된 가이드라인에 의거해 통계기반으로 구묶음을 자동 태깅하여 말뭉치를 구축하였으며, 그 결과는 다음 <표 2>와 같다.

<표 2> 뉴스 도메인 구묶음 코퍼스 구축 결과

뉴스 도메인		작업문서 수	문장 수
1	economy	150	2,045
2	international	143	2,273
3	it	141	2,575
4	living	147	2,876
5	opinion	150	3,524
6	politics	142	2,413
7	society	145	2,620
8	sports	144	2,571
총합		1,162	20,897

이때 N사에서 제공한 웹 기반의 구묶음 태깅 도구를 활용하였으며, 이미 구묶음된 상태에서 웹기반 작업도구를 활용하여 작업자가 수작업으로 주석 결과물을 수정 및 보완하였다. 이러한 방법은 수작업을 최소화하고 다수의 작업자로부터 구묶음 주석 결과의 일관성을 유지하는데 효과적이다.

2.3 주석 결과물에 대한 평가

본 연구에서는 앞서 정의된 가이드라인을 토대로 통계기반 구묶음 후보군에 대한 오류를 검증하고 보완하였다. 특히, 기존 모듈의 특성을 분석함으로써 구축된 태깅 코퍼스에 대한 정확도 및 일관성을 평가하였다.

첫 번째로 기존 통계기반 구묶음 결과의 오류 유형을 분석한 결과, 주로 명사 및 부사 수식 관계에서 오류가 많이 발생하였다. 다음 <표 3>, <표 4>와 같이 명사 수식 관계가 기존 구묶음 결과에서 제대로 묶이지 않거나, 동사를 수식하는 부사 및 부사구가 기존 구묶음 결과에서 분리된 경우가 다수였다. 특히, 명사를 중심으로 수식 관계를 분석하여 주요 키워드를 추출하는 것이 본 연구의 주요 목적이므로, 기존 구묶음 결과로 인해 분리된 명사 수식 관계를 하나로 묶어주었다.

<표 3> 기존 구묶음 오류 유형 1 - 명사 수식 관계

원문	검은 터번은 이슬람 예언자 무함마드의 직계임을 뜻한다.
기존 구묶음	[검은 터번은] [이슬람 예언자 무함마드의] [직계임을] [뜻한다.]
수정된 구묶음	[검은 터번은] [이슬람 예언자 무함마드의 직계임을] [뜻한다.]
명사 '직계임'을 앞의 구가 관형격조사 '의'를 중심으로 수식하고 있고, 전체가 5어절 미만이므로 [이슬람 예언자 무함마드의 직계임을] 전체를 하나로 묶음	

<표 4> 기존 구묶음 오류 유형 2 - 부사 수식 관계

원문	뿌듯하게 고개를 끄덕였다.
기존 구묶음	[뿌듯하게] [고개를 끄덕였다.]
수정된 구묶음	[뿌듯하게 고개를 끄덕였다.]
기존 구묶음에서는 '고개를 끄덕였다'가 묶여 있기 때문에 부사 '뿌듯하게'가 뒷부분 전체를 수식한다고 판단하고 하나로 결합함	

두 번째로 작업자별 고빈도 오류 유형을 분석한 결과, 전반적으로 최대 구묶음 단위를 유지하기 위해 5어절 이상의 수식구를 분리하는 과정에서 오류가 가장 많이 나타났다. 그리고 개체명 처리와 관련해 하나의 의미 단위(어절 수)를 어떠한 관점에서 보는가에 따라 이견이 많았다. 이에 작업자별 오류 유형을 다음 <표 5>, <표 6>과 같이 수식 관계 분리와 개체명 처리로 구분하였다.

<표 5> 작업자별 고빈도 오류 유형 1 - 수식 관계 분리

원문	미국이 기밀작전을 통해 죽였다고 발표한 이슬람국가(IS)의 수괴 아부 바크르 알바그다디(48세 추정)는
기존 구묶음	[미국이] [기밀작전을] [통해 죽였다고] [발표한] [이슬람국가(IS)의 수괴] [아부 바크르 알바그다디(48세 추정)는]
수정된 구묶음	[미국이] [기밀작전을 통해 죽였다고 발표한] [이슬람국가(IS)의 수괴 아부 바크르 알바그다디(48세 추정)는]

① [미국이 기밀작전을 통해 죽였다고 발표한 이슬람국가(IS)의 수괴 아부 바크르 알바그다디(48세 추정)는] → '최대 범위 묶음'을 반영하여 전체를 하나로 묶은 뒤, 수식구가 5어절 이상이므로 분리 필요

② [미국이 기밀작전을 통해 죽였다고 발표한] [이슬람국가(IS)의 수괴 아부 바크르 알바그다디(48세 추정)는] → 명사 '아부 바르크 알바그다디'를 중심으로 수식구가 2개이므로 먼 수식구를 분리하고, 여전히 5어절 이상이므로 추가 분리 필요

③ [미국이] [기밀작전을 통해 죽였다고 발표한] [이슬람국가(IS)의 수괴 아부 바크르 알바그다디(48세 추정)는] → 주어인 '미국이'를 분리하고, 전체가 5어절 미만이므로 분리를 멈춤(개체명 중 이름은 전체를 1어절로 취급)

문장 내에서 수식구가 명사와 어떤 관계를 갖는가에 따라 수식 관계 구묶음의 처리 방향이 달라진다. 특히, 5어절 이상인 수식구를 분리하는 과정이 어려워 초기에는 작업자별로 오류가 많이 발생하였다. 이에 본 연구에서는 <표 5>와 같이 수식구가 5어절 이상인 경우에 '수식 관계 분리 → 주어 분리 → 연결어미(의미 단위) 분리' 순으로 검증 및 보완 작업을 수행하였다.

<표 6> 작업자별 고빈도 오류 유형 2 - 개체명 처리

원문	의회 '여성 살해 대책 위원회' 의장인 발레리아 발란테 상원의원도
기존 구묶음	[의회 '여성 살해 대책 위원회'] [의장인] [발레리아] [발란테 상원의원도]
수정된 구묶음	[의회 '여성 살해 대책 위원회' 의장인] [발레리아 발란테 상원의원도]

'여성 살해 대책 위원회'는 단체명, '발레리아 발란테'는 이름으로 각각 1어절로 취급 가능

본 연구에서는 <표 6>과 같이 하나의 개체(entity), 행위, 상태 등 직관적으로 인식되는 구를 하나의 구묶음으로 태깅하였다. 이 중 개체명의 어절 수를 어떻게 처리하는 가에 따라 구묶음의 단위가 달라진다. 따라서 본 연구에서는 인명/단체명, 작품명, 국가명, 지명 등 확실히 하나의 의미 단위인 것만 1어절로 취급하고, 그 외에 일반 명사가 포함된 개체명은 각각의 어절을 개별로 취급하였다. 단, 법률명과 같이 분리하면 의미가 손상되는 경우에는 5어절 이상이어도 하나의 단위로 묶고, 다른 어절과 분리하였다.

작업 전체적으로 뉴스 기사의 복잡한 문장 구조로 인해 수식 관계 파악이 어려웠고, 작업자가 수식 관계를 어떻게 파악하는 가에 따라 개인별 편차가 컸다. 따라서 구문 분석 결과가 제시된 이후에 주석 결과물을 수정 및 보완하는 것이 보다 효율적이고, 품질의 일관성도 담보할 수 있는 방법이다.

3. 결 론

본 연구는 뉴스 도메인 문서집합 분석의 기초 자료로 활용될 수 있는 키워드 추출을 위한 구묶음 주석 방법론 연구 및 구축을 목적으로 수행되었다. 어떠한 문서든 내용 파악을 위해서는 문서 내 대표 키워드를 적절하게 선정하는 것이 중요하다. 문서에서 키워드가 될 수 있는 단위는 단어가 될 수도, 그 이상의 단위가 될 수도 있다. 이에 본 연구에서는 문서에서 단어뿐만 아니라 다양한 단위의 키워드에 주석을 다는 방법론을 고도화하고, 그 방법론을 뉴스 도메인에 실제로 적용하여 2만 문장 이상의 주석 말뭉치를 구축하였다. 본 연구의 결과물은 텍스트 문서의 내용을 파악하고 분석이 필요한 모든 텍스트 마이닝 관련 기술의 기초 작업으로 활용 가능하다.

감사의 글

본 논문은 2020년 ㈜엔씨소프트 연구비 지원에 의해 수행되었음

참고문헌

[1] 김경림, 이다영, 조환규, "문서 요약 및 비교분석을 위한 주제어 네트워크 가시화", 정보과학회논문지, 44권, 2호, pp.139-147, 2017.

[2] 박준현, 조장우, "PositionRank를 이용한 한국어 뉴스 기사 키워드 추출", 한국정보과학회, 학술발표논문집, 431-433, 2020.

[3] 남궁영, 김창현, 천민아, 박호민, 윤호, 최민석, 김재균, 김재훈, "한국어 말덩이 정의와 구묶음: 한국어 말덩이 부착 말뭉치와 Bi-LSTM/CRFs 모델을 활용하여", 정보과학회논문지, 47권, 6호, pp.587-595, 2020.