



비정형 텍스트 데이터의 토픽 분석의 정확도 향상을 위한 형태소 연결 기법 연구

김혜린, 박건호, 김성주



HANKUK UNIVERSITY OF FOREIGN STUDIES

1. 소개
2. 방법론
3. 결과
4. 결론

1. 소개

(1)	대학은	창의적	인재를	양성하여야	한다.
	<i>tayhakun</i>	<i>changuycek</i>	<i>inceylul</i>	<i>yangsenghayeya</i>	<i>hanta</i>
	university.TOP	creative	talented.OM	foster.OBLG	AUX.DCL
	'University must foster creative and talented human resources.'				
(1')	대학-은	창의적	인재-를	양성하-여야	하-는다.
	<i>tayhak-un</i>	<i>changuycek</i>	<i>incey-lul</i>	<i>yangsengha-yeya</i>	<i>ha-nta</i>
	university-TOP	creative	talented-OM	foster-OBLG	AUX-DCL
	'University must foster creative and talented human resources.'				

한국어 어휘 묶음의 형태소 기반 분석 예시

- 한글 텍스트 데이터로부터 키워드를 추출하는 방법 중 하나는 단어의 빈도를 확인하는 방법이 있다.
- 이를 위해서는 형태소 분석이 정확하게 선행되어야 한다.

해결하고자 하는 문제 정의

남자웃코디 :: 형스타일 양털반집업후리스 오버핏 남자맨투맨추천 너무 너무 예쁘고, 지금 입기 딱 좋은 옷이라 남자맨투맨추천 하는 양털반집업후리스 오버핏 맨투맨 입니
당 아이보리, 카키, 블루 컬러가 있는데 제 동생은 블루로...

나이키 배색반집업후리스 봉봉빈티지 ♥제품명: 나이키 배색반집업후리스 ♥가격: 38000 ♥사이즈: 총길이67센치 어깨단면45센치 가슴단면56센치 ♥제품설명:
컨디션8/10 세탁완료 95-100...

어커버 하프집업후리스 :: HALF ZIP UP SILKY FLEECE HALF ZIP UP SILKY FLEECE 원래 양털아우터를 살려고 했는데 어쩌다가 어커버반집업을 사버려따 21900원이던가?
할인하는데 거기에 11% 쿠폰할인이랑 친구할인이랑...

[18DAY] 남미배낭여행 (볼리비아 라파즈 데스로드/킬리킬리 야경/라파즈... 시간반의 이동 후 라파즈 도착 도둑 아입니다 새벽에 숙소 앞으로 픽업을 오기때문에 아침먹고
숙소앞에서 기다리는 중 한명씩 다 픽업하고 이동한다 이러시면...

Made U] 스윗서브 후리스반집업후드-뽀글이 양털 후드티] 털날리는 것... 상품명 - [Made U] 스윗서브 후리스반집업후드-뽀글이 양털 후드티 상품가격 - 19,900원 [상품
바로가기] [계시물 바로가기] 털날리는 것만 빼면 괜찮은거 같아요

/choiroom/ 터틀 울 골지 원피스 니트 세트 뒷면 초이룸의 베스트상품 함께보기 [무배]반폴라 기모 배색 프릴 롱 원피스 초이룸...원피스 초이룸 shopping.naver.com [무배
]집업폴라 기모 배색 롱 원피스 초이룸...

남자웃코디 :: 형스타일 양털반집업후리스 오버핏 남자맨투맨추천 너무 너무 예쁘고, 지금 입기 딱 좋은 옷이라 남자맨투맨추천 하는 양털반집업후리스 오버핏 맨투맨 입니
당 아이보리, 카키, 블루 컬러가 있는데 제 동생은 블루로...

나이키 배색반집업후리스 봉봉빈티지 ♥제품명: 나이키 배색반집업후리스 ♥가격: 38000 ♥사이즈: 총길이67센치 어깨단면45센치 가슴단면56센치 ♥제품설명:
컨디션8/10 세탁완료 95-100...

‘집업’이라는 키워드로 수집한 정제되어 있지 않은 텍스트 데이터 일부

- 위의 예시에서도 확인할 수 있듯이, 인터넷상의 비정형 데이터를 정제하기 위해서는 브랜드명, 신조어, 합성어 등 기존의 형태소 분석기만으로는 결과를 도출해 내기 어려운 점이 있다.

1. 소개

선행 연구

문장 : 내년도 최저임금을 기존 방식으로 전체 업종에 동일하게 적용하기로 결정했다.

한나눔		KKMA		komoran		okt		mecab		khaiii	
내년	N	내년도	NNG	내년	NNG	내년	Noun	내년도	NNG	내년	NNG
도	J	최저	NNG	도	JX	도	Josa	최저	NNG	도	JX
최저임금	N	임금	NNG	최저임금	NNP	최저임금	Noun	임금	NNG	최저	NNG
을	J	을	JKO	을	JKO	을	Josa	을	JKO	임금	NNG
기존	N	기존	NNG	기존	NNG	기존	Noun	기존	NNG	을	JKO
방식	N	방식	NNG	방식	NNG	방식	Noun	방식	NNG	기존	NNG
대로	J	대로	JX	대로	JX	대로	Josa	대로	JX	방식	NNG

- 이미 다양한 형태소 분석기들이 오픈소스로 공개되어 있지만, 각각의 형태소 분석기 마다 도출해내는 결과와 성능에 차이가 있다.

앞선 예시와 같이 정형화 되어 있지 않은 텍스트를 처리하기 위해
본 연구에서 제시하는 방법론의 Diagram

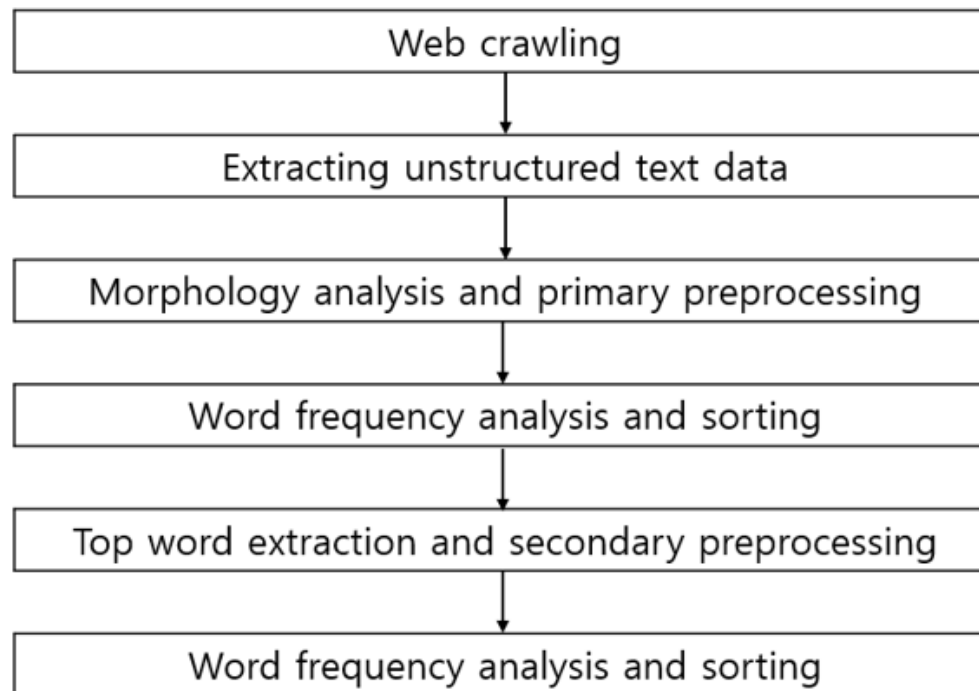


Figure 1. Block diagram of showing processes of proposed method

2. 방법론

Hannanum

[('미사용', 'N')]
[('오버핏', 'N')]
[('후리스', 'N')]
[('반집업', 'N')]
[('나이키', 'N')]
[('디스커버', 'N'), ('이', 'J'), ('리', 'E')]
[('아디다스', 'N')]

Kkma

[('미사', 'NNG'), ('용', 'XSN')]
[('오버', 'NNG'), ('핏', 'NNG')]
[('후', 'NNG'), ('리스', 'NNG')]
[('반집', 'NNG'), ('업', 'NNG')]
[('나이키', 'NNG')]
[('디스', 'UN'), ('커버', 'NNG'), ('리', 'XSN')]
[('아디다스', 'NNG')]

Komoran

[('미', 'NNP'), ('사용', 'NNG')]
[('오버핏', 'NA')]
[('후', 'NNG'), ('리스', 'NNP')]
[('반', 'XPN'), ('집', 'NNG'), ('업', 'NNG')]
[('나이키', 'NNP')]
[('디스커버리', 'NNP')]
[('아디다스', 'NNP')]

Mecab

[('미', 'NNP'), ('사용', 'NNG')]
[('오버', 'NNG'), ('핏', 'NNG')]
[('후', 'NNG'), ('리스', 'NNG')]
[('반집', 'NNG'), ('업', 'NNG')]
[('나이키', 'NNG')]
[('디스커버리', 'NNP')]
[('아디', 'NNG'), ('다스', 'NNG')]

Okt

[('미사', 'Noun'), ('용', 'Noun')]
[('오', 'Modifier'), ('버핏', 'Noun')]
[('후', 'Noun'), ('리스', 'Noun')]
[('반', 'Modifier'), ('집업', 'Noun')]
[('나이키', 'Noun')]
[('디스커버리', 'Noun')]
[('아디다스', 'Noun')]

본 연구에서 텍스트 분석을 위해 사용한 5가지의 형태소 분석기

[('미사용', 'N')]
[('오버핏', 'N')]
[('후리스', 'N')]
[('반집업', 'N')]

Figure 2. Compound words analyzed by Hannanum

[('나이키', 'NNP')]
[('디스커버리', 'NNP')]
[('아디다스', 'NNP')]

Figure 3. Compound words analyzed by Komoran and Okt

Figure2. '한나눔'의 경우, 합성어를 의도에 맞게 출력하는 것을 확인할 수 있다.
Figure3. 'Komoran'과 'Okt'는 브랜드명을 누락없이 출력한다.

각각의 형태소 분석기는 위와 같이 장단점을 가지고 있다.

[('미사용', 'N')]
[('오버핏', 'N')]
[('후리스', 'N')]
[('반집업', 'N')]

Figure 2. Compound words analyzed by Hannanum

[('나이키', 'NNP')]
[('디스커버리', 'NNP')]
[('아디다스', 'NNP')]

Figure 3. Compound words analyzed by Komoran and Okt

본 연구에서는 **각각의 형태소 분석기로부터 올바르게 도출된 결과만을 최종 단어로 채택**하게 되는데,
예를 들어 형태소 분석 결과에 어느 하나의 형태소 분석 결과에 '아디'가 있고,
다른 형태소 분석 결과에 '아디다스'가 있다면,
'아디'는 탈락되고, '아디다스'가 최종 결과로 도출된다.

다양한 형태소 분석기의 장점만을 결합하는 방식

1. 동일한 비정형 한글 텍스트 데이터에 대해 5가지의 형태소 분석기를 사용하여 **명사만을 추출**한다.
2. 각 단어의 빈도를 확인하고 각각의 형태소 분석기 결과 중 **상위 100개의 단어만 추출**한다.
3. 나머지 형태소 분석기에 적절한 값을 곱하여 **빈도수 값을 균일하게 조정**한다.
4. 하나의 형태소 분석기에서 도출된 결과를 다른 형태소 분석기의 결과와 **비교**한다.
5. 하나의 형태소 분석기로부터 얻어진 결과는 제거한다.
6. 하나의 단어가 다른 단어에 포함되면, 보다 짧은 단어는 제거된다.

제안방식의 효율성

Table 1. Comparison of results of used words (previous approach vs. our proposed approach)

Incomplete words (previous method)	corrected words (proposed method)
아디, 다스	아디다스
직접, 개인, 안전, 거래	직접거래, 개인거래, 안전거래
즉시, 송금	즉시송금
랄프, 로렌	랄프로렌
안전, 거래	안전거래
보물, 하우스	보물하우스
파우더, 룸	파우더룸
뉴, 발란스	뉴발란스
칼, 하트	칼하트
타미, 힐, 피거	타미힐피거
공식, 공구	공식공구

이전에 하나의 형태소 분석 결과로 '아디' 와 '다스' 로 분리되었지만, 본 연구의 방법론을 적용한 결과 '아디다스' 로 결과를 얻을 수 있었으며, '직접', '개인', '거래' 로 각기 분리되었던 결과들이 '직접거래', '개인거래', '안전거래' 로 본문에서 자주 사용되었던 의도를 파악할 수 있게 되었다.

3. 결과

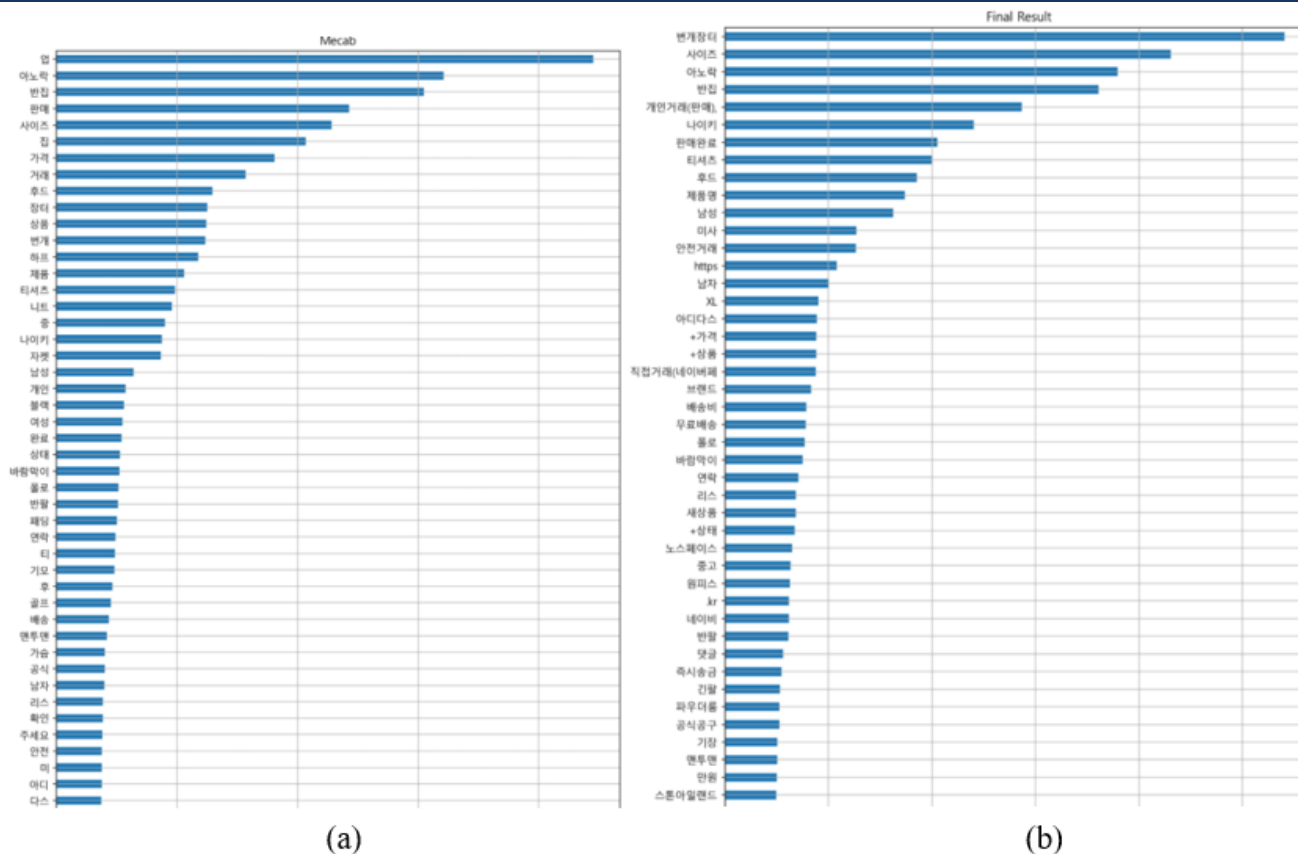


Figure 3. Comparison of the results generated using the existing method and our proposed method. (a) previous method, (b) proposed method

(a) 이전의 하나의 형태소 분석기로 분석된 결과

(b) 본 연구의 방법론을 적용한 결과

- 특정 형태소 분석기 하나만을 사용하여 비정형 한글 텍스트 데이터의 핵심 키워드를 추출하는 데에는 한계가 있다. 특히, 브랜드명이나 신조어, 합성어 등은 분해되거나 제거되는 **문제점**이 있다.
- 본 연구에서는 5가지의 형태소 분석기를 적절히 **조합**하여 브랜드명과 신조어 합성어들을 분해하지 않고 그대로 출력하는 방법을 제시한다.
- 이를 통해 비정형 한글 텍스트 데이터에서 핵심이 되는 키워드들을 **보다 정확하게 추출**해 내는 데에 다방면으로 활용할 수 있을 것으로 기대한다.

감사합니다.