미등록단어 문제와 데이터 부족 현상을 해결하기 위한 비지도학습 토크나이저와 추출 기반 문서 요약 기법

김 현 중

- "Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling" , **Hyunjoong Kim**, Han Kyul Kim, Sungzoon Cho, Expert Systems with Applications (minor review) **(5장)**
- "Representation learning for unsupervised heterogeneous multivariate time series segmentation and its application", **Hyunjoong Kim**, Han Kyul Kim, Misuk Kim, Joosung Park, Sungzoon Cho, Keyng Bin Im, and Chang Ryeol Ryu, Computers & Industrial Engineering, 130, 272-281. **(6장)**
- "Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation", Han Kyul Kim, **Hyunjoong Kim**, Sungzoon Cho, Neurocomputing. Volume 266, 29 November 2017, Pages 336-352
- "KR-WordRank: A Korean word extraction method based on WordRank and unsupervised learning", **Hyunjoong Kim**, Sungzoon Cho and Pilsung Kang, 대한산업공학회지, 2014, Vol. 40(1): 18-33. **(4장)**
- "추천 시스템 기법 연구동향 분석", 손지은, 김성범, **김현중**, 조성준, 대한산업공학회지, Apr 2015, Vol.41, No.2, pp.185-208.
- "Fast Parameterless Ballistic Launch Point Estimation based on k-NN Search", Soojin Kim, **Hyunjoong Kim**, Sungzoon Cho, Defence Science Journal, Volume 64, No 1, January 2014, Pages 41-47. (SCIE)

Activities

김 현 중 (Hyunjoong Kim)

- 데이터야놀자2018 발표, "문서 군집화를 위한 효율적인 k-means 활용: 빠른 학습, 군집 레이블링, 시각화"
- Naver TeckTalk (2017) 발표, "미등록단어 문제 해결을 위한 비지도학습 기반 한국어자연어처리 방법론 및 응용"
- PyCon2017 발표, "노가다 없는 텍스트 분석을 위한 한국어 NLP"

Projects

- 노이즈가 많은 대화 데이터 분석을 위한 자연어처리 엔진 개발과 일상 대화를 위한 대화형 챗봇 엔진 개발 (2015. 05 - 2016. 04)
- 공정로그 텍스트 데이터 분석을 위한 자연어처리 엔진 개발 및 이를 이용한 클레임 요인 분석 (2014. 12 - 2015. 03)
- Semantic 검색을 위한 topic modeling (2013. 04 - 2013. 11)
- 차량의 시계열 센서데이터의 비지도기반 구간 분리 및 주행 패턴 추출 (2015. 05 - 2016. 04)
- 데이터와 분석 목적에 맞는 데이터마이닝 분석 로드맵 작성 (2013. 12 - 2014. 02)
- 스마트 TV 자동녹화 시스템 패러매터 설정을 위한 데이터 분석 (2012. 04 - 2012. 11)
- Cross selling, Up selling을 위한 고객 데이터 분석 (2011. 05 - 2011. 11)

미등록단어 문제와 데이터 부족 현상을 해결하기 위한 비지도학습 토크나이저와 추출 기반 문서 요약 기법

김 현 중

자연어처리

- 자연어처리는 사람의 언어를 컴퓨터가 이용할 수 있는 형태의 정보로 변환하고, 이를 이용하여 과업들을 수행하는 분야이다.
  - 토크나이징, 문서 요약, 정보 추출 등 다양한 세부 과업으로 나뉘어지며,
  - 자연어처리를 위한 많은 방법들은 머신 러닝 기법을 이용하고 있다.

자연어처리

- 이 논문은 머신러닝 기법을 이용하는 한국어 텍스트 데이터의 두 가지 자연어처리 과업의 문제점을 개선한다.
  - 토크나이징
  - 키워드, 핵심 문장 추출을 통한 문서 요약

토크나이징

- 토크나이징은 주어진 문장을 토큰열로 인식하는 과업이다.
  - 품사 판별은 문장을 단어열, 형태소 분석은 문장을 형태소 열로 인식한다.

- 토크나이징은 다른 자연어처리 과업의 전처리 과정에 해당한다.
  - 문서 요약, 토픽 모델링과 같은 과업의 품질은 토크나이징의 품질에 의존한다.

문서 요약

- 문서 요약의 단위는 키워드 혹은 요약 문장으로 이뤄진다.
- 문서 요약의 접근법은 두 가지로 분류된다.
  - 추출 기반 (extractive) 접근법은 데이터에 존재하는 단어나 문장을 선택한다.
  - 요약 기반 (abstractive) 접근법은 문서의 내용을 표현할 수 있는 새로운 문장을 생성한다.

자연어처리 과업의 어려움

- 자연어처리 과업에서는 다음의 어려움이 존재한다.
  - 미등록단어 : 학습데이터에 존재하지 않은 새로운 단어를 인식하지 못할 수 있다.
  - 학습데이터 구축 : 많은 과업들이 학습데이터를 이용하는 지도학습 기반 머신 러닝 알고리즘을 이용하지만, 이를 위한 학습데이터를 구축하기 어려운 경우가 많다.
  - 오류: 띄어쓰기 및 철자법 오류에 의하여 단어가 제대로 인식되지 않을 수 있다.

논문의 기저

- 언급한 어려움을 해결하기 위하여 비지도학습 기반의 방법으로 접근한다.
  - 비지도학습 기반 방법은 학습데이터의 의존성이 낮기 때문에 다양한 도메인으로의 확장이 용이하다.
  - 지도학습 기반 방법과 보완적으로 이용될 수 있다.
- 한국어의 특징을 이용한다.
  - 교착어인 한국어의 어절 구조를 이용하여 비지도학습 기반 방법이 통계 정보를 효과적으로 이용할 수 있도록 돕는다.

논문의 범위

1장. 개요 및 관련 연구

학습데이터를 이용하지 않으며 토크나이저의 미등록단어 문제를 해결하기 위한 연구

2장. 미등록단어 문제 해결을 위한 단어 추출 기법과 이를 이용한 한국어 토크나이저

3장. 어절 구조를 이용한 통계 기반 명사 추출

4장. 그래프 랭킹 기반 키워드/핵심문장 추출을 이용한 단일주제 문서 집합 요약

5장. 문서 군집화 알고리즘 및 군집화 레이블링을 이용한 다주제 문서 집합 요약

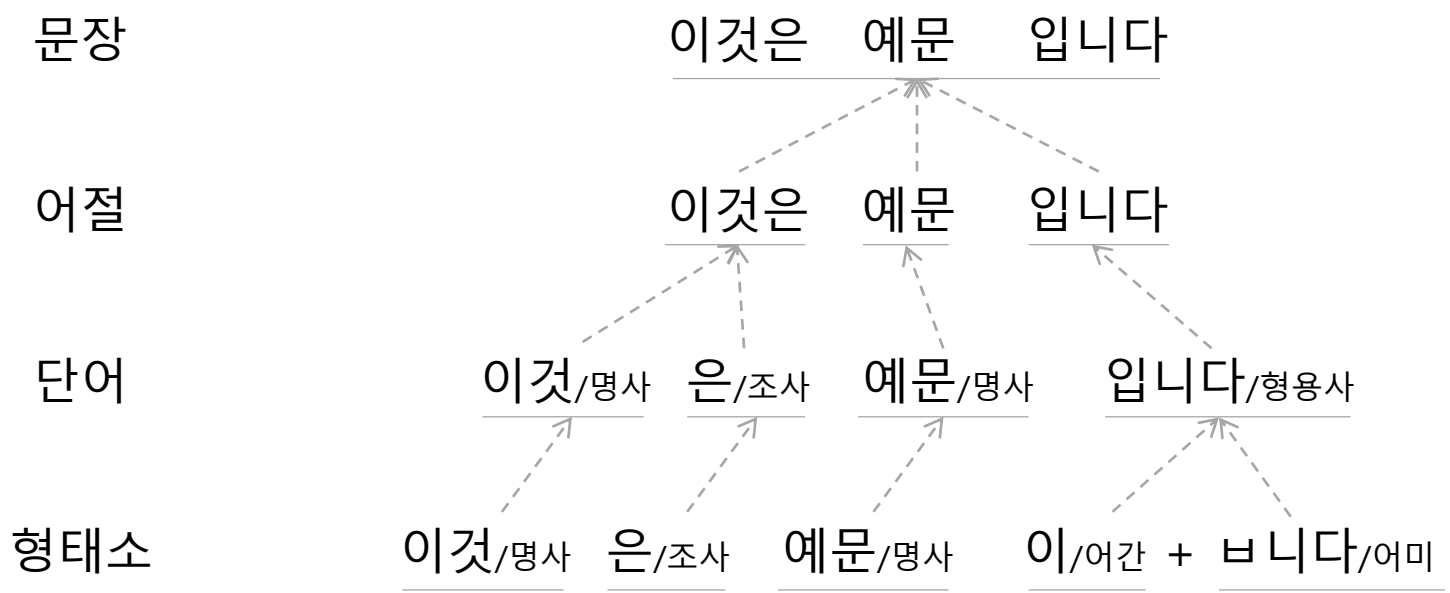
6장. 시계열 형식의 뉴스 문서 집합 요약을 위한 거리 기반 유사 주제 구간 분리

7장. 결론

한국어 어절 구조 제안 : L+[R]

한국어의 구성

- 한국어는 다음의 요소로 구성되어 있다.



제안하는 어절 구조

- 의미를 지니는 단어는 어절의 왼쪽에 등장한다.
  - 명사, 동사, 형용사, 부사, 감탄사 : 의미를 지니는 단어
  - 조사, 어미 : 문법 기능의 단어와 형태소

발표/명사 + 를/조사

하/동사어근 + 면서/어미

< 의미를 지닌 단어, 형태소가 어절의 왼쪽에 등장하는 예시>

제안하는 어절 구조

- 한국어 어절의 형태는 L + [R] 이다.
  - 여러 형태소가 합쳐져 하나의 어절을 구성할 수 있다.
  - 의미 부분과 문법 기능 부분을 각각 L 과 R 의 복합형태소로 표현한다.

국어수업하는데 = 국어/명사 + 수업/명사 + 하/동사파생접미사 + 는데/어미

국어수업하는데 = 국어수업/L + 하는데/R

국어수업 = 국어수업/L

< 제안하는 L + [R] 구조 예시>

1장. 개요 및 관련 연구

2장. 미등록단어 문제 해결을 위한 단어 추출 기법과 이를 이용한 한국어 토크나이저

3장. 어절 구조를 이용한 통계 기반 명사 추출

4장. 그래프 랭킹 기반 키워드/핵심문장 추출을 이용한 단일주제 문서 집합 요약

5장. 문서 군집화 알고리즘 및 군집화 레이블링을 이용한 다주제 문서 집합 요약

6장. 시계열 형식의 뉴스 문서 집합 요약을 위한 거리 기반 유사 주제 구간 분리

7장. 결론

배경

- 토크나이징 (품사 판별)은 단어 사전을 이용하여 단어열 후보를 만들고, 순처적 레이블링 (sequential labeling) 알고리즘을 이용하여 후보를 선별한다.
  - 형태소 분석은 단어 탐색에 이용될 수 있다.

문장 : '집에 간다고 말했다'

Word	Tag	b	e
집	명사	0	1
에	Josa	1	2
간	명사	2	3
가 + ㄴ다고	동사 + 어미	2	5
다	명사	3	4
고	-	4	5
말	명사	5	6
하+았다	동사 + 어미	6	8

품사	단어
명사	이것, 예문, 집, 말, 이, 것, 다, 간
조사	은, 는, 에, 이
형용사	이다,
동사	가다, 하다
어미	-다, -ㄴ다고, -았다

형태소 사전에 어간의 원형과 어미가 있을 경우,
"어간 + 어미 결합 규칙" 을 이용하여 용언을 인식한다.

< 사전을 이용한 단어 후보 탐색 예시 >

배경

- 생성된 단어열 후보는 순차적 레이블링 방법에 의하여 평가된다.
  - 앞, 뒤의 문맥을 이용하여 판별을 위한 변수 (features)를 생성한다.
    - $\widehat{w,t} = \text{argmax}_{w,t \in G(s)} \langle \lambda, F(w,t) \rangle$

단어열 : [집, 에, 가+ㄴ다고, 말, 하+았다]
품사열 : [명사, 조사, 동사+어미, 명사, 동사+어미]

+

$F(w_{i-1}, t_{i-1}, w_{i+1})$

w_{i-1}='에', t_{i-1}='조사', w_{i+1}='말'

< 앞, 뒤 문맥을 이용하여 생성된 변수의 예시 >

배경

- 그러나 사전에 등록되지 않은 단어는 단어열 후보에 포함되지 않는다.
  - 형태소 분석은 "새로운 단어는 알려진 형태소의 결합"이라 가정한다.
  - 한국어는 표의문자 성격을 지니기 때문에 음절을 단어로 인식한다.

문장 : 재공연을 했어요

품사열 : (**재공연**, 명사), (**을**, 조사), (**했어요**, 동사)

형태소열: (**재**, 관형사), (**공연**, 명사), (**을**, 조사), (**하**, 동사), (**었**, 선어말어미), (**어요**, 종결어미)

< 품사 판별과 형태소 분석의 차이 예시 >

너무/MAG, 너무너무/MAG, 는/JX, 아이오/NNG, 아이/NNG, 의/JKG, 노래/NNG, 에/JKM, 요/JX

비/Noun, 선/Verb, 실/PreEomi, 세/PreEomi, 가/Eomi, 드러났/Verb, 다/Eomi

< 형태소 분석에 의한 미등록단어 문제 예시 >

배경

- Word Piece Model (WPM) 은 비지도학습 기반 토크나이저이다 ^[1].
  - Byte Pair Encoding 를 이용하여 자주 등장한 k 개의 부분단어 (subwords) 를 단어 사전으로 학습하고, 문장을 학습된 부분단어로 분할한다.
  - 빈도수가 작은 단어는 잘못된 글자열로 나뉘어지거나, 빈도수가 큰 어절은 단어로 나뉘어지지 않는 문제가 발생한다. (예시) "오늘 + 의"
  - 단어 점수가 부분단어의 빈도수에 의해서만 정의되기 때문이다.

문장 : 오늘의 사건사고 뉴스입니다
부분단어열 : _오늘의 _사건 사고 _뉴스 입 니다

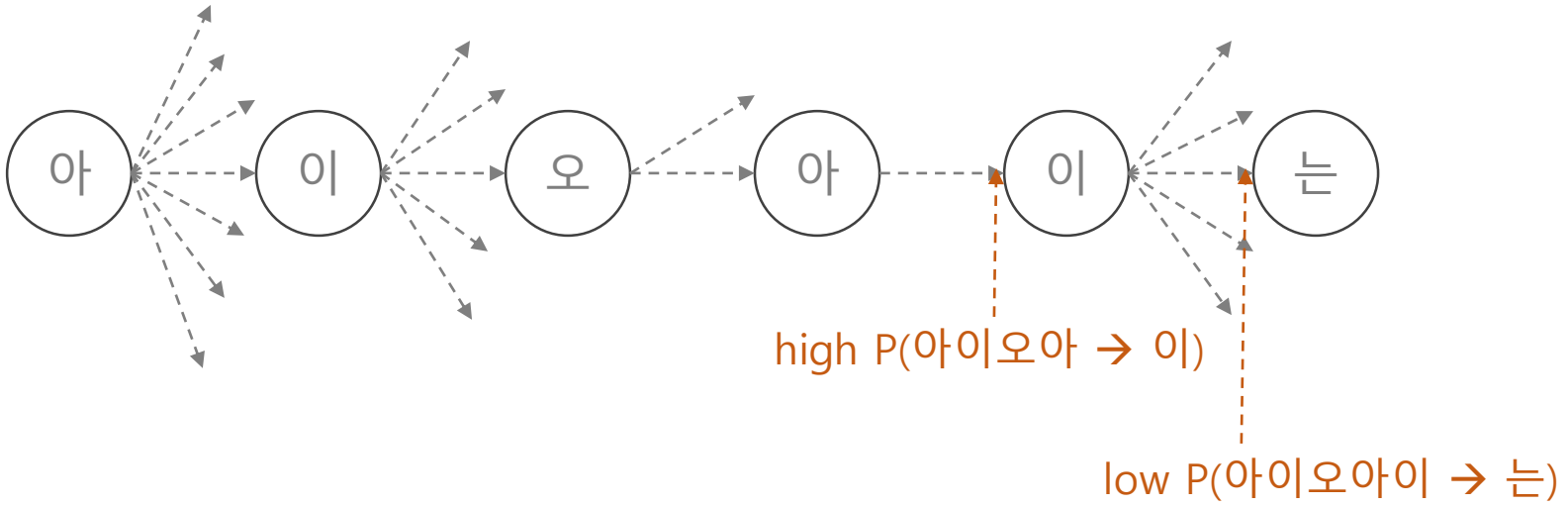
< Word Piece Model 에 의한 토크나이징 결과 예시 >

[1] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

제안하는 방법

- 한국어에 적합한 단어 점수 정의와 이를 이용한 토크나이저를 제안한다.
  - Cohesion 은 글자열 n-gram 을 이용하여 글자열의 단어 점수를 정의한다 ^[1].

$$cohesion(c_{1:n}) = \sqrt[n-1]{\prod_{i=2}^{n} P(c_{1:i}|c_{1:i-1},)}$$

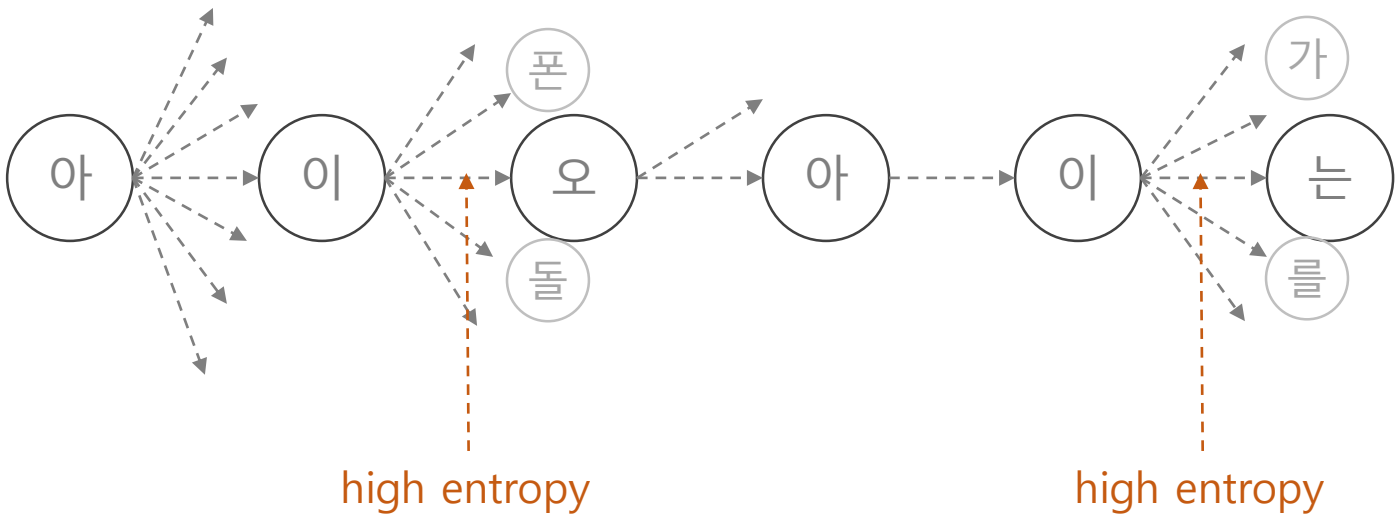


[1] Kim, H. (2013). Cleansing noisy text using corpus extraction and string match. Master's Thesis, Seoul National University

제안하는 방법

- 한국어에 적합한 단어 점수 정의와 이를 이용한 토크나이저를 제안한다.
  - Branching Entropy 는 단어 경계의 글자 분포를 이용하여 단어 점수를 정의한다 ^[1].

$$BE(c_{1:n}) = \sum_{c_{i+1} \in G(c_{1:n})} -P(c_{i+1}|c_{1:n}) \times \log P(c_{i+1}|c_{1:n})$$



[1] Jin, Z. and Tanaka-Ishii, K. (2006). Unsupervised segmentation of chinese text by use of branching entropy. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 428–435. Association for Computational Linguistics.

제안하는 방법

- 두 종류의 단어 점수 정의법은 어절 내 단어 경계를 잘 표현한다.
  - 단어 점수는 글자들의 확률에 의하여 정의되기 때문에 빈도수의 영향을 덜받는다.
  - 어절 내 L 과 R 의 경계를 구분할 수 있다.

subword	frequency	Cohesion score	Branching Entropy
아이	4,910	0.15	3.11
아이오	307	0.10	0.49
아이오아	270	0.20	0
아이오아이	270	0.30	1.72
아이오아이는	40	0.26	0

< 뉴스로부터 학습한 Cohesion 과 Branching Entropy 예시 >

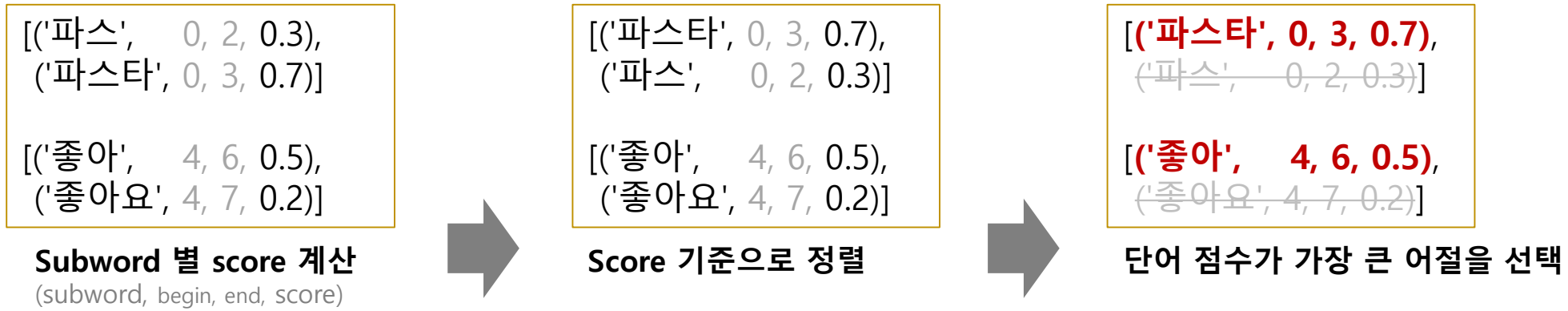
제안하는 방법

- 데이터의 띄어쓰기 정확도에 따라 두 종류의 토크나이저를 선택한다.
  - 띄어쓰기 오류가 적다면 어절을 L + [R] 로 이분한다 (L-Tokenizer).
  - 띄어쓰기 오류가 많다면 문장에서 단어 점수가 높은 L 부분을 우선적으로 선택한다 (Max Score Tokenizer).

제안하는 방법 (L-Tokenizer)

- 띄어쓰기가 잘 되어있다면 어절의 왼쪽에서 점수가 높은 부분을 자른다.

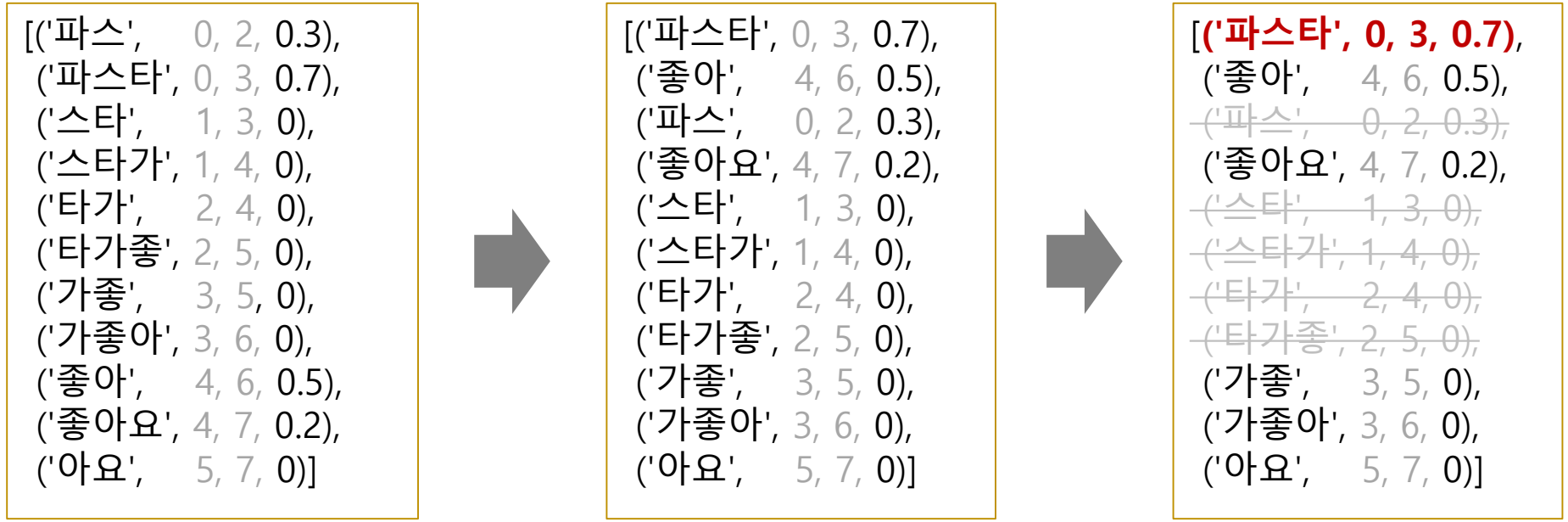
```
단어 점수= {'파스': 0.3, '파스타': 0.7, '좋아요': 0.2, '좋아':0.5}
tokenize('파스타가 좋아요')
```



제안하는 방법 (Max Score Tokenizer)

- 띄어쓰기가 잘 되어있지 않다면 점수가 높은 부분부터 자른다.

```
단어 점수= {'파스': 0.3, '파스타': 0.7, '좋아요': 0.2, '좋아':0.5}
tokenize('파스타가좋아요')
```



Subword 별 score 계산
(subword, begin, end, score)

Score 기준으로 정렬

최고점수의 단어 선택,
위치가 겹치는 단어 제거

제안하는 방법 (Max Score Tokenizer)

- 띄어쓰기가 잘 되어있지 않다면 점수가 높은 부분부터 자른다.

```
단어 점수 = {'파스': 0.3, '파스타': 0.7, '좋아요': 0.2, '좋아':0.5}
```

[파스타]가좋아요

```
[('파스타', 0, 3, 0.7),
 ('좋아',    4, 6, 0.5),
 ('파스',    0, 2, 0.3),
 ('좋아요', 4, 7, 0.2),
 ('스타',    1, 3, 0),
 ('스타가', 1, 4, 0),
 ('타가',    2, 4, 0),
 ('타가좋', 2, 5, 0),
 ('가좋',    3, 5, 0),
 ('가좋아', 3, 6, 0),
 ('아요',    5, 7, 0)]
```

[파스타]가[좋아]요

```
[('파스타', 0, 3, 0.7),
 ('좋아',    4, 6, 0.5),
 ('파스',    0, 2, 0.3),
 ('좋아요', 4, 7, 0.2),
 ('스타',    1, 3, 0),
 ('스타가', 1, 4, 0),
 ('타가',    2, 4, 0),
 ('타가좋', 2, 5, 0),
 ('가좋',    3, 5, 0),
 ('가좋아', 3, 6, 0),
 ('아요',    5, 7, 0)]
```

[파스타, 가, 좋아, 요]

제안하는 방법 (Pseudo code)

```
s : sentence
w : eojeol
D : word – score dictionary

def tokenize(s, D):
    tokens = [ ]
    for w in split(s):
        scores = { }
        for e in (2, |w|):
            sub = w[:e]
        scores[sub] = D.get(sub, 0)
    l ← find sub having maximal score
    r ← w[|l|:]
    tokens += [l, r]
    return tokens
```

< Pseudo code of L-Tokenizer >

```
s : sentence
w : eojeol
D : word – score dictionary

def tokenize(s, D):
    subs ← scan subword score (s, D)
    subs ← sort subs by score in reverse order
    tokens = [ ]
    while subs is not empty:
        t ← pop subs
        tokens += [t]
        subs ← remove overlapped sub with t
    return tokens
```

< Pseudo code of Max Score Tokenizer >

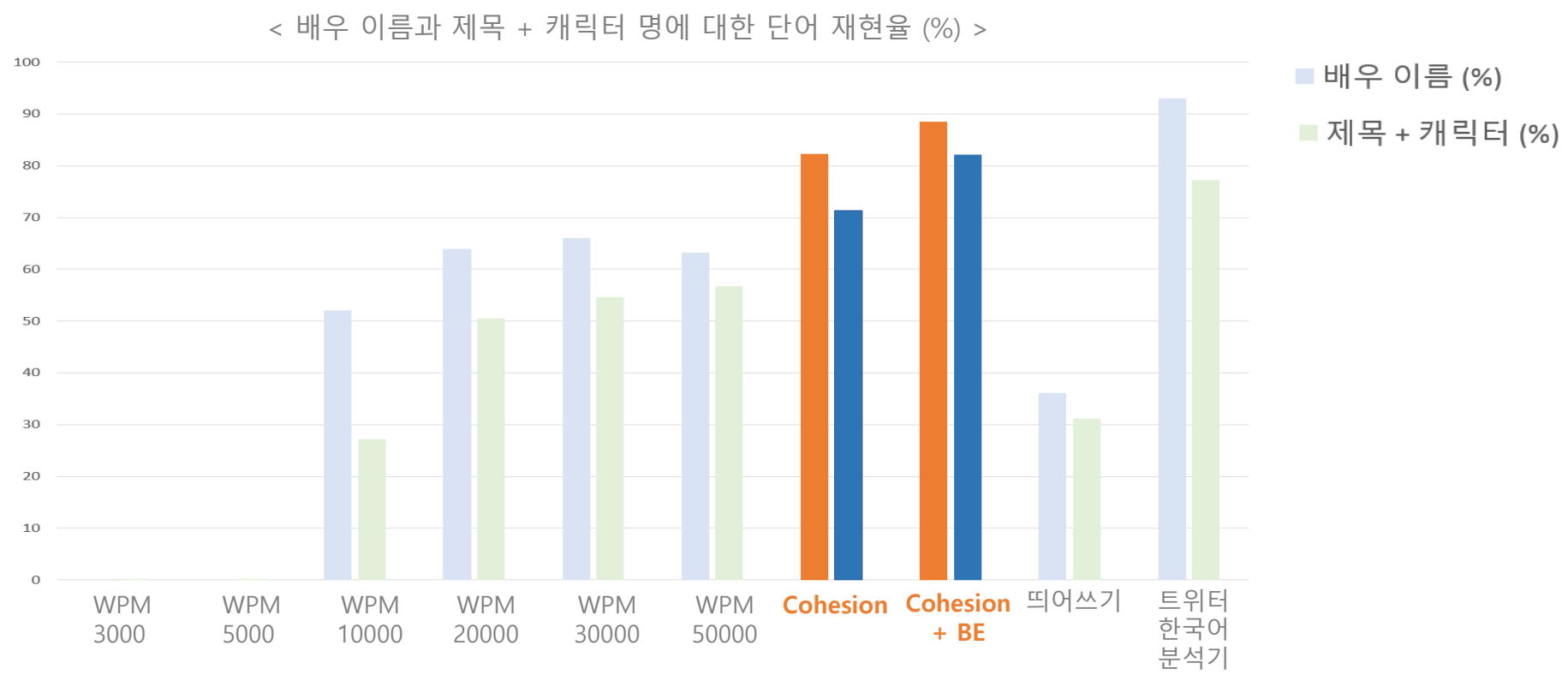
성능 평가

- 영화평 데이터를 이용한 성능 평가를 수행하였다.
  - 리뷰가 5000 개 이상인 영화 152 개, 3,280,685 건을 이용하였다.
  - 단어 추출 방법은 단일 주제 문서 집합에서 잘 작동한다 (세종 말뭉치는 다주제)

- 두 가지 자연어처리 과업을 통하여 토크나이저의 성능을 평가하였다.
  - 배우 / 극중 역할명 / 영화 제목 재현을 통한 미등록단어 인식 능력 평가
  - 평점 분류를 통한 벡터화 능력 평가

성능 평가 1. 단어 추출 재현율

- 토크나이저 별로 특정 단어들이 제대로 인식되었는지 확인하였다.
  - 배우, 캐릭터, 제목을 수집한 뒤, 해당 단어가 등장한 경우 단어가 재현되었다고 정의하였다



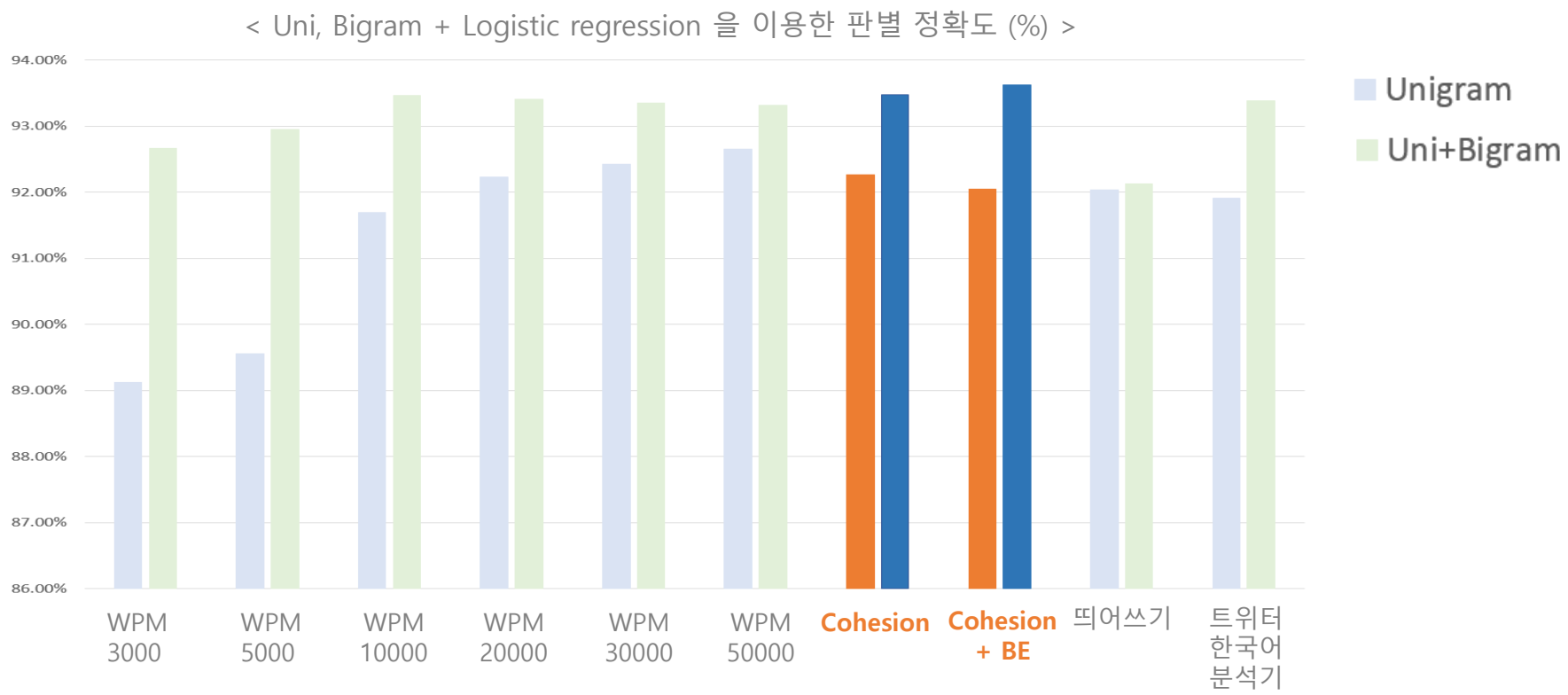
성능 평가 1. 단어 추출 재현율

- 토크나이저 별로 특정 단어들이 제대로 인식되었는지 확인하였다.
  - 배우, 캐릭터, 제목을 수집한 뒤, 해당 단어가 등장한 경우 단어가 재현되었다고 정의하였다

모델	배우 이름 (%)	제목 + 캐릭터 (%)	배우 + 제목 + 캐릭터 (%)
WPM 3000	0	0.13	0.05
WPM 5000	0	0.13	0.05
WPM 10000	52.05	27.21	43.3
WPM 20000	63.93	50.42	59.17
WPM 30000	65.99	54.49	61.94
WPM 50000	63.21	56.62	60.89
Cohesion	**82.23**	**71.29**	**78.38**
Cohesion + BE	**88.4**	**82.08**	**86.17**
띄어쓰기	36.09	31.09	34.33
트위터 한국어 분석기	**92.95**	**77.14**	**87.38**

성능 평가 2. 문서 분류

- 영화평을 이용하여 평점 (sentiment) 를 분류하였다.
  - 평점은 1 ~ 3 점 (부정) / 4 ~ 7 점 (제거) / 8 ~ 10 점 (긍정)으로 사용하였다.



성능 평가 2. 문서 분류

- 영화평을 이용하여 평점 (sentiment) 를 분류하였다.
  - 평점은 1 ~ 3 점 (부정) / 4 ~ 7 점 (제거) / 8 ~ 10 점 (긍정)으로 사용하였다.

모델	unigram		uni + bigram	
	정확도	순위	정확도	순위
WPM 3000	89.12%	10	92.67%	9
WPM 5000	89.56%	9	92.95%	8
WPM 10000	91.69%	8	**93.47%**	**3**
WPM 20000	92.23%	4	93.41%	4
WPM 30000	**92.43%**	**2**	93.35%	6
WPM 50000	**92.65%**	**1**	93.32%	7
Cohesion	**92.27%**	**3**	**93.48%**	**2**
Cohesion + BE	92.05%	5	**93.63%**	**1**
띄어쓰기	92.04%	6	92.13%	10
트위터 한국어 분석기	91.91%	7	93.39%	5

결론

- 제안된 방법은 WPM 와 비교하여 다음의 장점이 있다.
  - 문서의 벡터화 품질과 단어 인식 능력이 좋다.
  - WPM 는 적절한 k 를 탐색해야 하지만, 제안된 방법은 패러매터를 사용하지 않는다.
  - 임의의 단어 점수 정의 방법을 이용할 수 있다.
- 사전을 이용하는 토크나이저에 근접하는 단어 인식 능력을 보인다.

추가 연구

- 제안된 방법은 단어의 품사 추정 능력이 없다.
  - (단어, 품사) 사전을 이용하는 학습 기반 품사 판별기를 보완하기 어렵다.

- 기학습된 토크나이저의 변수로 이용될 수 있다.
  - 비지도학습 기반 토크나이저의 정보를 이용하여 기학습된 토크나이저도 데이터셋에 적합하게 모델을 조절할 수 있다.

배경

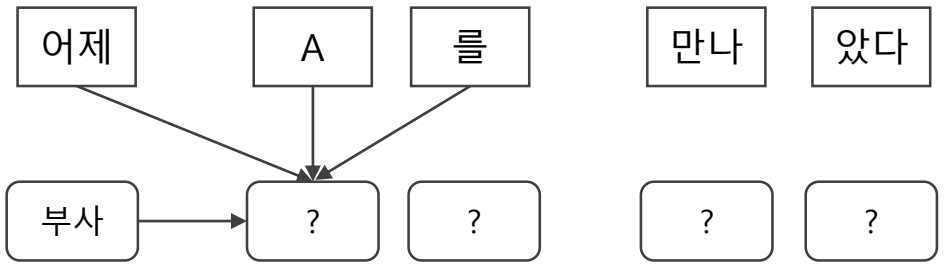
- 새롭게 만들어지는 단어(형태소)는 **명사**와 어미다
  - 명사는 **새로운 개념**을 어미는 다양한 말투를 표현하기 위하여 만들어진다.
    - (예시: **아이오아이** , 하지**말라궁**)
- 미등록단어는 주로 명사에서 발생한다.
  - 명사는 가장 많이 이용되며 다양한 단어로 구성되어 있다.

형태소 품사	출현 빈도수 (비율)	고유 개수	평균 출현 빈도수
명사	7,124,644 (29.13 %)	144,294	49.38
어미	4,688,406 (19.17 %)	3,142	1,492.17
조사	4,184,235 (17.11 %)	289	14,478.32
어간	3,308,743 (13.53 %)	7,989	414.16

< 세종말뭉치에서 계산된 형태소 별 통계 >

관련 연구

- 순차적 레이블링 기반 품사 판별기는 미등록단어를 추정할 수 있다 ^[1,2].



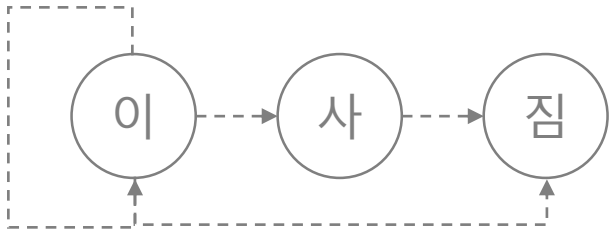
< 순차적 레이블링 알고리즘의 품사 추정 예시
$F(x_{i-1}$ = '어제', y_{i-1}='부사', x_{i+1}='를') >

- 음절 단위의 순차적 레이블링은 문맥을 이용한 단어 추정이 가능하다 ^[3].
  - 학습데이터에 존재하는 문맥만을 이용할 수 있다.

[1] Na, S.-H., Yang, S.-I., Kim, C.-H., Kwon, O.-W., and Kim, Y.-K. (2012). Crfs for korean morpheme segmentation and pos tagging.
In Proc. of 24th Annual Conference on Human and Cognitive Language Technology, pages 12–15
[2] Lee, C., Kim, J., Kim, J., and Kim, H. (2013). Joint models for korean word spacing and pos tagging using structural svm. Journal of KIISE: Software and Applications, 40(12):826–832.
[3] https://github.com/kakao/khaiii

관련 연구

- 템플릿과 규칙 기반을 이용한 명사 추출 방법들이 제안되었다.
  - "[명사]+은" 과 같은 템플릿을 이용하면, "은"이 포함된 명사는 잘못 인식된다 ^[1, 2].
    - (예시) 손나은
  - 오토마타와 규칙을 이용하여 명사를 추출하는 방법이 제안되었다 ^[3].
    - 한글로 표기된 외래어 명사들이 제대로 인식되지 않는다.



< 오토마타 예시 >

[1] Lee, H.-y., Lee, J.-s., Kang, B.-d., and Yang, S.-w. (2016). Functional expansion of morphological analyzer based on longest phrase matching for efficient korean parsing. Journal of Digital Contents Society, 17(3):203–210
[2] Hong, J.-P. and Cha, J.-W. (2008). A new korean morphological analyzer using eojeol pattern dictionary. In Proceedings of the Korean Information Science Society Conference. Korean Institute of Information Scientists and Engineers.
[3] Lee, D.-G., Lee, S.-Z., and Rim, H.-C. (2003). An efficient method for korean noun extraction using noun patterns. Journal of KIISE: Software and Applications, 30(1):173–173.

제안하는 방법

- 단어의 오른쪽에 등장하는 글자 분포를 이용한 명사 추출법을 제안한다.
  - A 의 오른쪽에 등장하는 글자 분포를 이용하면 A 가 명사임을 추정할 수 있다.

어제 A라는 가게에 가봤어

A에서 보자

A로 와줘

< A 가 명사임을 추정할 수 있는 문장 예시 >

제안하는 방법

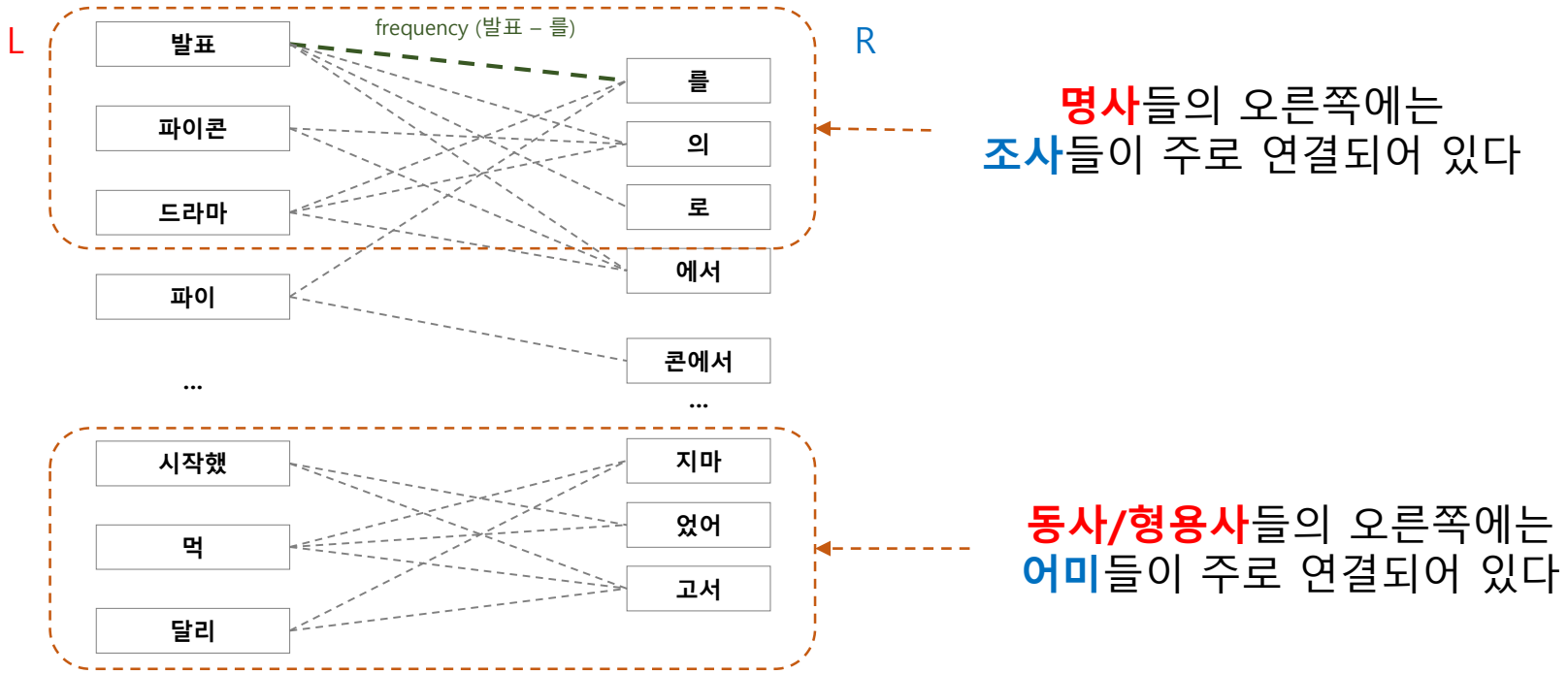
def extract_nouns(D, t, c):
    D : dataset
    t : threshold of noun score
    c : minimum count of word

    $G = ((L, R), E)$ ← construct L-R graph (D, c)
    $N = \{\ \}$
    for l in reverse sort by length(L) :
        s ← noun score(l)
        if $s \geq t$:
            remove (l^*, r^*) from G such as $l \in l^* + r^*$
        $N \leftarrow N \cup \{l\}$
    N ← postprocessing (N)
    return N

< Pseudo code of proposed noun extractor >

제안하는 방법 (1단계: L-R 그래프 구축)

- 모든 어절의 L + [R] 를 이용하여 그래프를 구축한다.
  - 발표를 → [발 + 표를, 발표 + 를, 발표를 + ""]



명사들의 오른쪽에는
조사들이 주로 연결되어 있다

동사/형용사들의 오른쪽에는
어미들이 주로 연결되어 있다

제안하는 방법 (2단계: 판별기 학습)

- 세종 말뭉치의 품사가 태깅된 정보로부터 L – R 행렬을 만든 뒤, 판별기를 이용하여 L 의 명사 유무를 판단하는 모델을 학습한다.

…

예술가의	**예술가/NNG+의/JKG**	**113**
예술가는	예술가/NNG+는/JX	45
예술가가	예술가/NNG+가/JKS	43
예술가들의	예술가/NNG+들/XSN+의/JKG	30

…

단어 품사	단어 / R	- 는	**- 의**	- 고	- 가	-었던
명사	**예술가**	45	**113**	2	43	0
동사	먹	33	0	27	0	27
명사	예술가들	0	30	0	0	0

제안하는 방법 (2단계: 판별기 적용)

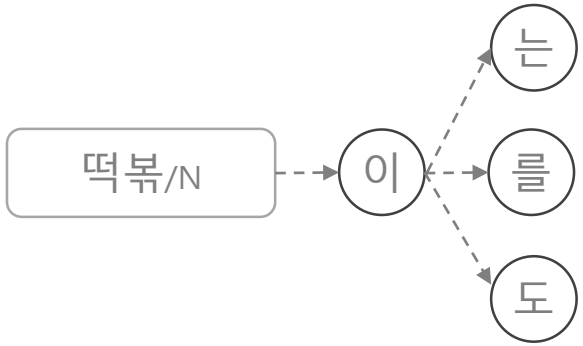
- 길이가 긴 명사 후보 L 부터 명사 유무를 판별한다.
  - 짧은 L 단어에는 잘못된 R 이 변수로 포함되어 있다.
    - '아이디어/명사' → (아이디, 어)

'아이디어' 에 의하여 '아이디'가 명사로 추출되지 못한다

'아이디어' 를 명사로 추출한 뒤, L-R 그래프에서 제거하면 '아이디'가 명사로 추출된다.

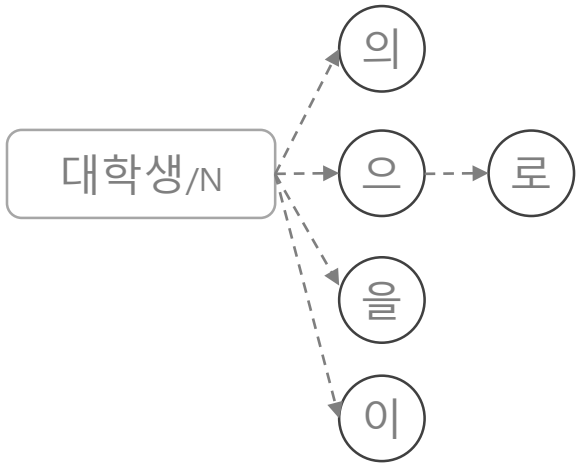
제안하는 방법 (3 단계 후처리)

- $N = N_{sub} + J$ (떡볶 + 이)



- 긴 단어의 Branching Entropy 가 더 크면 짧은 단어를 명사에서 제거

- $N_0 = N + J_{sub}$ (대학생으 + 로)



- 짧은 단어의 Branching Entropy 가 더 크면 긴 단어를 명사에서 제거

성능 평가

- 세종 말뭉치를 이용하여 명사 추출 성능을 평가하였다.
  - 일부 기학습된 모델은 세종말뭉치를 이용하여 학습되었음에도 불구하고 낮은 재현율을 보였다.

모델	정밀도 (precision)	재현율 (recall)	F1 점수
제안하는 방법	**0.96**	0.958	**0.959**
꼬꼬마 형태소 분석기	0.92	**0.963**	0.941
한나눔 형태소 분석기	0.866	0.897	0.881
트위터 한국어 처리기	0.945	0.874	0.908

< 세종말뭉치를 이용한 명사 인식 능력 평가 >

성능 평가

- 뉴스 문서에서의 명사 인식 능력을 측정하였다.
  - 네이버뉴스로부터 수집한 1일의 뉴스 30,091 건으로부터 명사를 추출하였다.
  - 명사 사전으로 신조어가 포함되어 있는 "나무위키"를 이용하였다.
  - 기학습된 모델들은 낮은 재현율을, 제안한 방법은 상대적으로 높은 재현율을 보인다.

모델	정밀도 (precision)	재현율 (recall)	F1 점수
제안하는 방법	0.8672	**0.5484**	**0.6719**
꼬꼬마 형태소 분석기	**0.8968**	0.2941	0.4429
한나눔 형태소 분석기	0.8778	0.3228	0.4721
트위터 한국어 처리기	0.8897	0.3467	0.499

< 나무위키를 이용한 명사 인식능력 평가 >

성능 평가

- 뉴스 문서에서의 명사 인식 능력을 측정하였다.
  - 네이버뉴스로부터 수집한 1일의 뉴스 30,091 건으로부터 명사를 추출하였다.
  - 명사 사전으로 신조어가 포함되어 있는 "네이버 한국어 사전"을 이용하였다.
  - 신조어가 적기 때문에 상대적으로 비슷한 성능을 보인다.

모델	정밀도 (precision)	재현율 (recall)	F1 점수
제안하는 방법	0.9797	**0.4721**	**0.6372**
꼬꼬마 형태소 분석기	**0.9931**	0.3461	0.5133
한나눔 형태소 분석기	0.9858	0.4005	0.5695
트위터 한국어 처리기	0.9843	0.3928	0.5615

< 네이버 한국어 사전을 이용한 명사 인식능력 평가 >

결론

- 한국어 어절의 구조와 문법 기능을 하는 단어집합 (R) 을 이용하여 통계 기반으로 명사를 추출하는 방법을 제안하였다.
  - 제안하는 방법은 세종 말뭉치와 뉴스 기사에서 세종 말뭉치와 외부 단어 사전을 이용하는 모델들과도 비슷하거나 더 좋은 명사 인식 능력을 보였다.

추가 연구

- 문맥상 명사가 아님을 추정할 수 있는 단어가 명사로 추출된다.
  - 그리고 + 는/서/도
- R 에 포함되지 않은 복합형태소는 명사의 일부로 추출된다.
  - 육성한다/개발한다/모색한다 + 고/는
- 추출된 명사 사전 만으로는 기학습된 모델을 보강하지 못한다.
  - 명사 사전은 사용자 사전으로만 이용될 수 있다.

추가 연구

- 추출된 명사 사전 기학습된 모델의 feature 로는 사용되지 못한다.
  - 명사 사전은 사용자 사전으로써 이용될 수 있다.
  - 사용자 사전에 의하여 추가되는 단어는 feature 에는 영향을 주지 못한다.

컴퓨터/명사, 를/조사, (**끄/동사**, 끌/동사), ㅂ니다/어미

'XY[-2] = (컴퓨터, 명사) & XY[0] = (끄, 동사)' : 3.54

'XY[-2] = (컴퓨터, 명사) & XY[0] = (끌, 동사)' : 0.25

< 단어가 문맥을 판단하는 정보로 이용되는 예시 >

성능 평가 (데모)

덴마크	웃돈	너무너무너무	가락동	매뉴얼	지도교수
전망치	강구	언니들	신산업	기뢰전	노스
할리우드	플라자	불법조업	월스트리트저널	2022년	불허
고씨	어플	1987년	불씨	적기	레스
스퀘어	충당금	건축물	뉴질랜드	사각	하나씩
근대	투자주체별	4위	태권	네트웍스	모바일게임
연동	런칭	만성	손질	제작법	현실화
오해영	심사위원들	단점	부장조리	차관급	게시물
인터폰	원화	단기간	편곡	무산	외국인들
세무조사	석유화학	워킹	원피스	서장	공범

< 뉴스 데이터에서의 명사 추출 예시 >

성능 평가 (데모)

변심	인시디어스	코디님	강만후	병원예약	전공필수
이런의미	바깥양반	프레시아	전리품상자	난심일기	근육몬
동피랑	상태이상	흐흐르	수입과자	모래사장	알엔에이
오션윙	옥매미	산넘어산	클럽하우스	물한번	손에다
따른사람	모진말	노들섬	한신아파트	괜차니	핵고생
다음장	아이도루	본사사람	마스크걸	브런치카페	클렌징워터
배아플라	일일미션	버스정거장	리장	안산도착	다리하나
빠른사과	용계	아홉시꺼	걸신	컴푸타	필터빨
성큼성큼	서가엔쿡	경중	스카이워크	인형만	하루시작
퍈의점	신청하기	미모쨜	재밌는시간	운동하규	간사들

< 대화 데이터에서의 명사 추출 예시 >

1장. 개요 및 관련 연구

2장. 미등록단어 문제 해결을 위한 단어 추출 기법과 이를 이용한 한국어 토크나이저

3장. 어절 구조를 이용한 통계 기반 명사 추출

상황 별 문서 집합 요약을 위한 추출 기반 비지도학습 문서 요약 방법론

4장. 그래프 랭킹 기반 키워드/핵심문장 추출을 이용한 단일주제 문서 집합 요약

5장. 문서 군집화 알고리즘 및 군집화 레이블링을 이용한 다주제 문서 집합 요약

6장. 시계열 형식의 뉴스 문서 집합 요약을 위한 거리 기반 유사 주제 구간 분리

7장. 결론

배경

- 추출 기반 문서 요약 방법은 데이터에 존재하는 단어 혹은 문장을 키워드나 요약문으로 선택하여 문서 집합을 요약하는 방법이다.
- 문서 집합의 종류에 따라 접근 방법이 달라야 한다.
  - 문서 집합이 하나의 주제로 구성되어 있을 때
  - 문서 집합이 여러 개의 주제로 구성되어 있을 때
  - 문서에 생성 시간의 정보가 포함되어 있을 때

1장. 개요 및 관련 연구

2장. 미등록단어 문제 해결을 위한 단어 추출 기법과 이를 이용한 한국어 토크나이저

3장. 어절 구조를 이용한 통계 기반 명사 추출

4장. 그래프 랭킹 기반 키워드/핵심문장 추출을 이용한 단일주제 문서 집합 요약

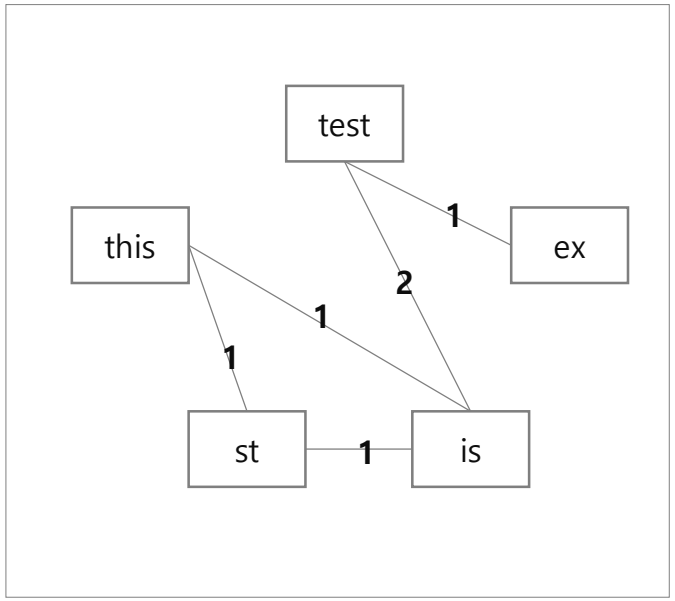
5장. 문서 군집화 알고리즘 및 군집화 레이블링을 이용한 다주제 문서 집합 요약

6장. 시계열 형식의 뉴스 문서 집합 요약을 위한 거리 기반 유사 주제 구간 분리

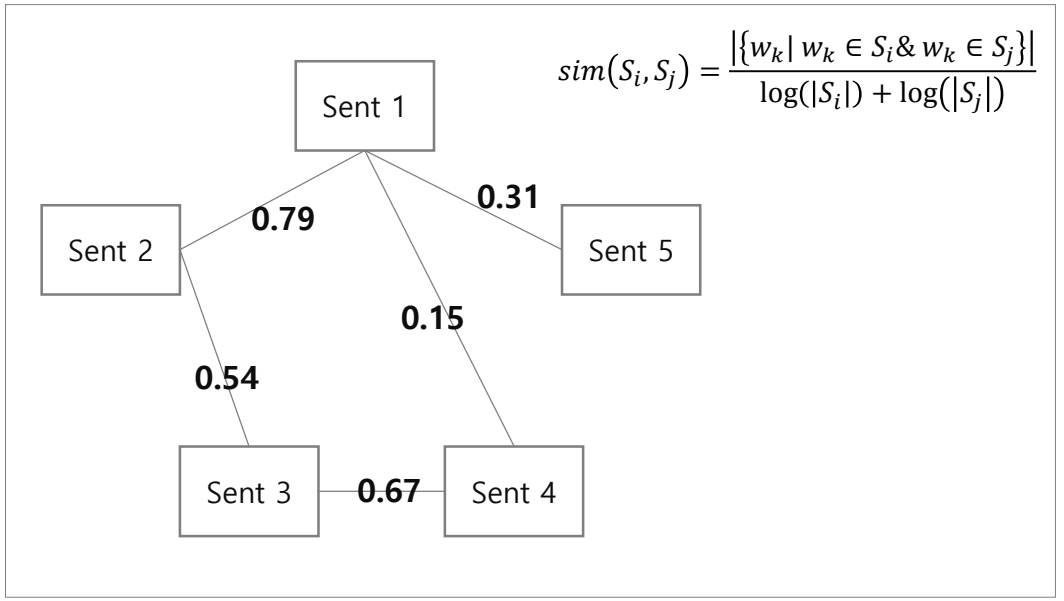
7장. 결론

배경

- 문서 집합의 주제가 단일할 경우에는 TextRank 가 이용될 수 있다 ^[1].
  - 토크나이저를 이용하여 단어, 문장 그래프를 구축한 뒤, PageRank 를 학습한다.
  - 랭크가 높은 마디 (단어, 문장)을 키워드, 핵심문장으로 선택한다.



< 단어 그래프 예시 >



< 문장 그래프 예시 >

[1] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

TextRank 의 한계점

- TextRank 의 키워드는 토크나이저의 성능에 의존한다.
  - 키워드가 단어로 인식되지 못하면 그 단어는 키워드로 선택되지 못한다.

- TextRank 의 핵심 문장은 다양성을 보장하지 않는다.
  - TextRank 는 문장 그래프의 각 마디(문장) 마다 독립적으로 랭크를 계산한다.
  - 한 문장이 높은 랭크를 얻으면 이와 비슷한 다른 문장도 높은 랭크를 얻는다.

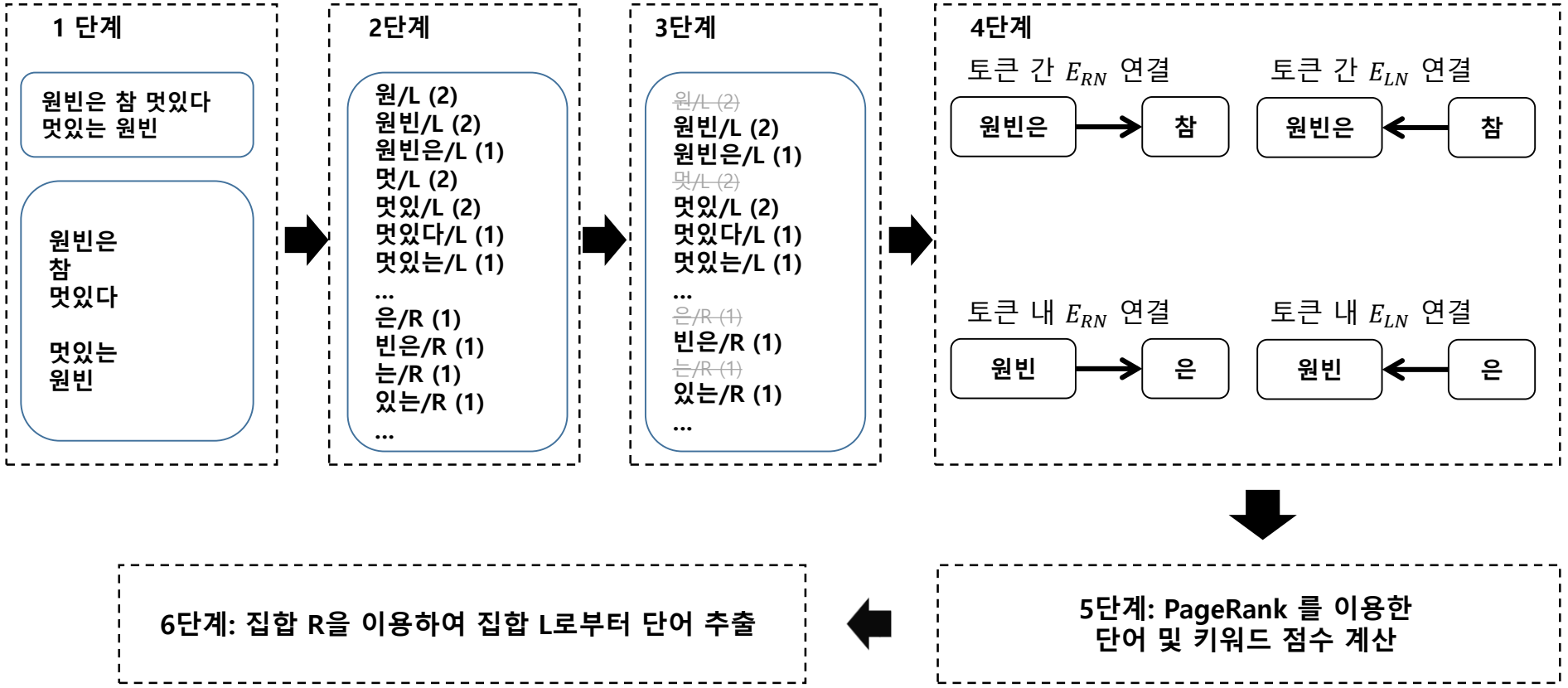
제안하는 방법

- 부분단어 (subword) 그래프를 이용하여 키워드를 추출한다.
  - 토크나이저에 의존하지 않기 때문에 미등록단어 문제를 우회할 수 있다.

- 문장과 키워드 간의 벡터 유사성을 이용하여 핵심 문장을 선택한다.
  - 토크나이저의 의존하지 않는 문장 그래프 생성이 가능하다.
  - 핵심 문장 선택 시, 핵심 문장들이 서로 이질적이도록 정규화 (regularize) 한다.

제안하는 방법 (키워드)

- 어절의 L + [R] 구조를 이용한 부분단어 그래프를 이용하여 키워드를 추출화는 과정



제안하는 방법 (키워드)

- 6단계: 집합 R을 이용하여 집합 L로부터 단어 추출

부분단어 L (랭크)	부분단어 R (랭크) (top k)
영화/L (355.3)	은/R (571.9)
~~영화가/L (127.3)~~	는/R (423.9)
원빈/L (283.3)	이/R (396.1)
~~영화다/L (55.1)~~	가/R (277.1)
~~원빈이/L (38.1)~~	다/R (158.3)
~~원빈은/L (35.3)~~	이다/R (29.5)

def select_keywords(L, R_k):

$KW = \{\}$ # keyword set

For c in L:

if c is not form of $(kw_i + r_j)$:

$KW \leftarrow KW \cup \{c\}$

return KW

$L = [l_1, l_2, \ldots l_L]$ # 모든 부분단어 L

$R_k = [r_1, r_2, \ldots r_k]$ # 랭크 기준 상위 k 개의 부분단어 R

< 6단계 pseudo code>

제안하는 방법 (핵심 문장)

1 단계 : 부분단어 그래프를 이용한 키워드 추출 및 이를 이용한 키워드 벡터 생성

키워드 (index)	영화 (0)	원빈 (1)	아저씨 (2)	감동 (3)	마지막 (4)	액션 (5)	멋있 (6)
랭크	55.3	25.1	19.5	5.5	3.2	3.1	1.5
벡터값	0.489	0.222	0.172	0.049	0.028	0.027	0.013

2 단계 : 키워드 포함 유무를 이용한 문장의 벡터화

문장	벡터
s_1 : 영화 아저씨에서 원빈 정말 멋있더라	[(0, 1), (1, 1), (2, 1), (6,1)]
s_2 : 마지막장면에서 좀 감동적이었어	[(4, 1), (3,1)]
s_3 : 아저씨에서 원빈 액션 멋있더라	[(2, 1), (1, 1), (5, 1), (6, 1)]

3 단계 : 선택된 문장 간 유사도를 이용한 핵심 문장 선택

제안하는 방법 (핵심 문장)

3 단계 : 선택된 문장 간 유사도를 이용한 핵심 문장 선택 ($\sigma = 0.5$)

키워드 (index)	영화 (0)	원빈 (1)	아저씨 (2)	감동 (3)	마지막 (4)	액션 (5)	멋있 (6)
랭크	55.3	25.1	19.5	5.5	3.2	3.1	1.5
벡터값	0.489	0.222	0.172	0.049	0.028	0.027	0.013

	문장	벡터	핵심 문장 비용
(1) 핵심 문장 선택	s_1 : 영화 아저씨에서 원빈 정말 멋있더라	[(0, 1), (1, 1), (2, 1), (6,1)]	0.1042
(3) 핵심 문장 선택	s_2 : 마지막장면에서 좀 감동적이었어	[(4, 1), (3,1)]	0.9231 + $2 \cdot I(\cos(s_1, s_2) > \sigma)$
	s_3 : 아저씨에서 원빈 액션 멋있더라	[(2, 1), (1, 1), (5, 1), (6, 1)]	0.5654 + $2 \cdot I(\cos(s_1, s_3) > \sigma)$

(2) 핵심 문장 비용증가

제안하는 방법 (핵심 문장)

3 단계 : 선택된 문장 간 유사도를 이용한 핵심 문장 선택 ($\sigma = 0.5$)

def select_keysentence (S, KV, σ):

σ # user-specified parameter

$KS = [\,]$ # key-sentence list

$C = dist(KV, S)$ # initial cost

while $|KS| < k$:

$i \leftarrow \arg\min C$

$KS \leftarrow KS + [s_i]$

$C \leftarrow C + I(dist(s, s_i) < \sigma)$

return KS

KV : # keyword vector

S : # vectorized sentences

< 3단계 pseudo code>

성능 평가

- 제안된 방법의 성능을 평가하기 위하여 온라인에서 수집한 영화평 데이터를 이용하였다.

데이터	목적
영화 별 영화평	• 기학습된 토크나이저에서 미등록단어 문제가 자주 발생 • 중복된 문장이 존재하는 긴 문서 • ROUGE-1 을 이용한 키워드 및 핵심 문장 추출 능력의 정성적, 정량적 평가

성능 평가

TextRank		KR-WordRank	
영화/NNG (173.00)	번/NNB (20.26)	영화 (201.34)	현실 (15.21)
보/VV (128.93)	거/NNB (19.67)	너무 (81.80)	생각 (14.94)
좋/VA (65.55)	최고/NNG (19.18)	정말 (40.62)	지루 (13.81)
하/VV (52.02)	때/NNG (19.15)	음악 (40.52)	다시 (13.62)
것/NNB (47.43)	사람/NNG (19.04)	마지막 (38.73)	감동 (13.61)
같/VA (45.37)	여운/NNP (17.55)	뮤지컬 (23.24)	보는 (12.49)
영화/NNP (43.89)	뮤지컬/NNP (16.94)	최고 (21.85)	좋아 (12.01)
음악/NNG (43.59)	나오/VV (16.54)	사랑 (20.69)	재밌 (11.91)
꿈/NNG (41.43)	듯/NNB (16.11)	꿈을 (20.47)	재미 (11.41)
있/VV (40.79)	영상미/NNG (15.95)	아름 (20.36)	좋고 (11.39)
없/VA (35.94)	지루/XR (15.66)	영상 (20.33)	계속 (11.16)
마지막/NNG (31.92)	처음/NNG (15.25)	여운이 (19.51)	조금 (10.95)
수/NNB (30.08)	장면/NNG (15.15)	진짜 (19.10)	느낌 (10.94)
사랑/NNG (28.25)	감동/NNG (15.14)	노래 (18.77)	처음 (10.76)
아름답/VA (26.47)	가/VV (15.03)	보고 (18.60)	결말 (10.60)
현실/NNG (24.82)	만들/VV (13.50)	좋았 (17.66)	연기 (10.52)
되/VV (23.91)	들/VV (13.24)	그냥 (16.60)	장면 (10.38)
노래/NNG (23.39)	남/VV (13.21)	스토리 (16.27)	그리고 (10.36)
생각/NNG (23.19)	느낌/NNG (13.13)	좋은 (15.67)	하는 (10.28)
스토리/NNP (21.35)	말/NNG (13.13)	인생 (15.41)	있는 (10.17)

< 라라랜드 영화 리뷰에서 추출된 키워드. TextRank 는 코모란을 사용>

성능 평가

추출 방법	핵심 문장
KR-WordRank	여운이 크게남는영화 엠마스톤 너무 사랑스럽고 라이언고슬링 남자가봐도 정말 매력적인 배우인듯 영상미 음악 연기 구성 전부 좋았고 마지막 엔딩까지 신선하면서 애틋하구요 30중반에 감정이 많이 메말라있었는데 오랜만에 가슴이 촉촉해지네요
	영상미도 너무 아름답고 신나는 음악도 좋았다 마지막 세바스찬과 미아의 눈빛교환은 정말 마음 아팠음 영화관에 고딩들이 엄청 많던데 고딩들은 영화 내용 이해를 못하더라——사랑을 깊게 해본 사람이라면 누구나 느껴볼수있는 먹먹함이 있다
	정말 영상미랑 음악은 최고였다 그리고 신선했다 음악이 너무 멋있어서 연기를 봐야 할지 노래를 들어야 할지 모를 정도로 그리고 보고 나서 생각 좀 많아진 영화 정말 이 연말에 보기 좋은 영화 인 것 같다
	무언의 마지막 피아노연주 완전 슬픔ㅠ보는이들에게 꿈을 상기시켜줄듯 또 보고 싶은 내생에 최고의 뮤지컬영화였음 단순할수 있는 내용에 뮤지컬을 가미시켜째즈음악과 춤으로 지루할틈없이 빠져서봄 ost너무좋았음
TextRank	처음엔 초딩들 보는 그냥 그런영화인줄 알았는데 정말로 눈과 귀가 즐거운 영화였습니다 어찌보면 뻔한 스토리일지 몰라도 그냥 보고 듣는게 즐거운 그러다가 정말 마지막엔 너무 아름답고 슬픈 음악이 되어버린
	시사회 보고 왔어요 꿈과 사랑에 관한 이야기인데 뭔가 진한 여운이 남는 영화예요
	시사회 갔다왔어요 제가 라이언고슬링팬이라서 하는말이아니고 너무 재밌어요 꿈과 현실이 잘 보여지는영화 사랑스런 영화 전 개봉하면 또 볼생각입니당
	황홀하고 따뜻한 꿈이었어요 imax로 또 보려합니다 좋은 영화 시사해주셔서 감사해요
TextRank (Cosine similarity)	시사회에서 보고왔는데 여운쩔었다 엠마스톤과 라이언 고슬링의 케미가 도입부의 강렬한음악좋았고 예고편에 나왓던 오디션 노래 감동적이어서 눈물나왔다ㅠ 이영화는 위플래쉬처럼 꼭 영화관에봐야함 색감 노래 배우 환상적인 영화
	방금 시사회로 봤는데 인생영화 하나 또 탄생했네 롱테이크 촬영이 예술 영상이 넘나 아름답고 라이언고슬링의 멋진 피아노 연주 엠마스톤과의 춤과 노래 눈과 귀가 호강한다 재미를 기대하면 약간 실망할수도 있지만 충분히 훌륭한 영화
	좋다 좋다 정말 너무 좋다 그 말 밖엔 인생영화 등극 ㅠㅠ
	음악도 좋고 다 좋고 좋고좋고 다 좋고 씁쓸한 결말 뭔가 아쉽다

< 라라랜드 영화 리뷰에서 추출된 핵심 문장들. TextRank 는 코모란을 사용>

성능 평가

- 각 데이터셋 별로 정답 요약 문장을 구축하는 것은 주관이 개입되며, 다양한 데이터에 확장 가능한 평가 방법이 아니다.
- 좋은 핵심 문장들은 다수의 키워드를 포함해야 한다.
  - 핵심 문장으로 선택된 문장 내의 키워드 재현율을 확인한다 (ROUGE-1).
  - 키워드를 정답 요약 문장으로 이용한다.

성능 평가

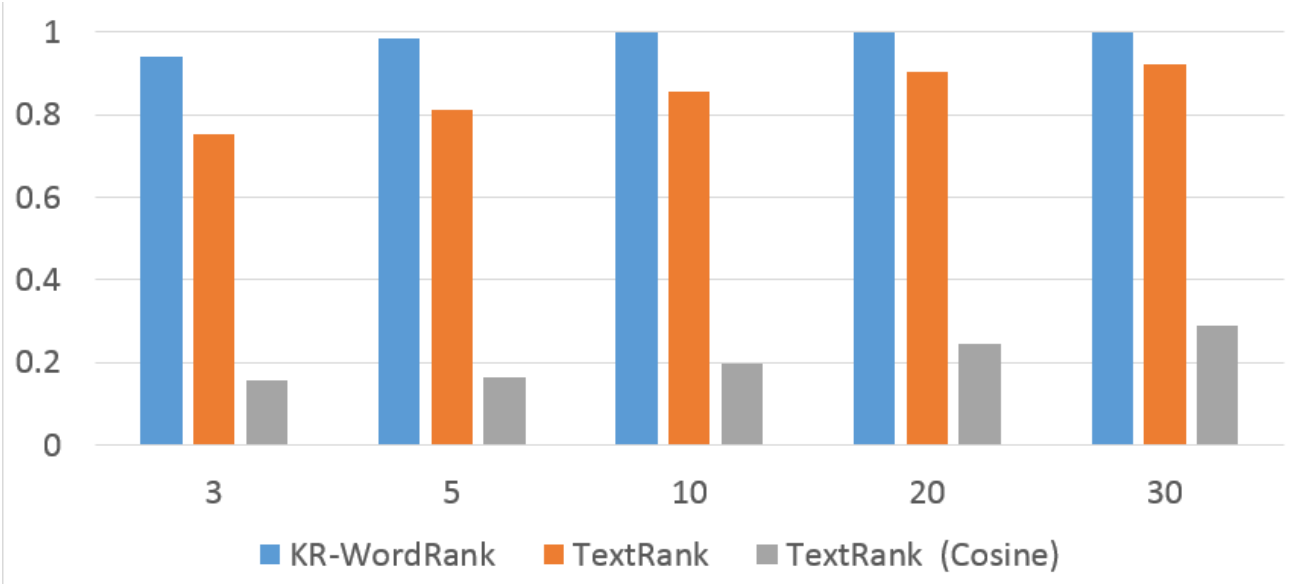
- ROUGE-N 은 문서 요약 과업의 정량평가 척도로 이용된다.
  - 정답 요약 문장이 주어졌을 때, 추출된 k 개의 핵심 문장에 포함된 정답 문장의 N gram 비율 (recall)

	Sentence	Bigrams
Reference	the cat was under the bed	the cat, cat was, was under, under the, the bed
Key-sentence	the cat was found under the bed	the cat, cat was, was found, found under, under the, the bed

< ROUGE-2 = 4/5 인 경우 예시 >

성능 평가

- 제안하는 방법 (KR-WordRank) 은 TextRank 보다 더 많은 종류의 키워드를 포함하는 문장을 핵심 문장으로 선택한다.
  - 네이버영화에서 수집한 영화평이 5000 개 이상인 152 개 영화



< 10 개의 키워드를 정답 문장으로 설정한 경우 핵심 문장 개수 별 ROUGE-1 점수
키워드 개수 별 실험 결과는 appendix 에 기술 >

성능 평가

- 제안하는 방법 (KR-WordRank) 은 TextRank 보다 더 많은 종류의 키워드를 포함하는 문장을 핵심 문장으로 선택한다.

# keywords	# keysents	KR-WordRank	TextRank	TextRank (Cosine)
10	3	0.942	0.752	0.156
10	5	0.983	0.810	0.166
10	10	0.998	0.855	0.197
10	20	1.000	0.904	0.244
10	30	1.000	0.920	0.291
20	3	0.758	0.603	0.078
20	5	0.873	0.690	0.084
20	10	0.966	0.767	0.101
20	20	0.994	0.839	0.135
20	30	0.998	0.869	0.170
30	3	0.631	0.496	0.053
30	5	0.771	0.587	0.056
30	10	0.911	0.688	0.069
30	20	0.980	0.778	0.093
30	30	0.994	0.820	0.120
50	3	0.480	0.363	0.032
50	5	0.621	0.452	0.034
50	10	0.804	0.565	0.042
50	20	0.929	0.674	0.058
50	30	0.968	0.731	0.078

< 네이버 영화에서 수집한 영화평을 이용한 성능 평가 결과 >

결론

- 토크나이저를 이용하지 않는 키워드, 핵심문장 추출 방법을 제안하였다.
  - 부분단어 그래프를 이용하여 문서집합에서 직접 단어를 학습함으로써 키워드가 미등록단어로 인식되는 문제를 해결하였다.
- 다양한 키워드들을 포함하는 핵심 문장의 추출을 유도할 수 있다.

1장. 개요 및 관련 연구

2장. 미등록단어 문제 해결을 위한 단어 추출 기법과 이를 이용한 한국어 토크나이저

3장. 어절 구조를 이용한 통계 기반 명사 추출

4장. 그래프 랭킹 기반 키워드/핵심문장 추출을 이용한 단일주제 문서 집합 요약

5장. 문서 군집화 알고리즘 및 군집화 레이블링을 이용한 다주제 문서 집합 요약

6장. 시계열 형식의 뉴스 문서 집합 요약을 위한 거리 기반 유사 주제 구간 분리

7장. 결론

배경

- 문서 집합의 주제가 다양할 경우 문서 요약이 어렵다 ^[1].
- 단일한 주제의 부분 집합으로 문서 집합을 나누어야 한다 ^[2].
- Spherical k-means 는 대량의 문서 집합을 효율적으로 군집화 한다.
  - *k*-means 는 학습 비용이 작은 군집화 방법이다 ^[3].
  - Cosine similarity 는 문서 간 거리를 잘 정의한다.

[1] Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 322–330. Association for Computational Linguistics.
[2] Twinandilla, S., Adhy, S., Surarso, B., and Kusumaningrum, R. (2018). Multi-document summarization using k-means and latent dirichlet allocation (lda)–significance sentences. Procedia Computer Science, 135:663–670
[3] Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. Journal of Statistical Software, 50(10):1–22.

문서 군집화를 위한 Spherical *k*-means 의 한계점

- 안정적인 initial centroids 를 선택하기 위한 *k*-means++ 알고리즘은 고차원 데이터에서 비효율적으로 작동한다.
- 각 군집의 문서를 요약하기 위해서는 별도의 모델을 학습해야 한다.

제안하는 방법

- 이 장에서는 다양한 주제의 문서들로 이뤄진 문서 집합을 효율적으로 요약하기 위해 “개선된 Spherical *k*-means” 알고리즘을 제안한다.
  - 고차원 벡터 공간의 특징을 이용한 효율적인 initializer
  - Centroids 를 이용한 군집 레이블링

제안하는 방법 (initializer)

- *k*-means++ [1] 은 안정적인 initial centroids 를 선택하는데 이용된다.

1. Select a point c_0 randomly
2. Select next point c_t with prob. $\frac{\mathrm{d}(x)^2}{\sum_{x \in D} \mathrm{d}(x)^2}$

   where $\mathrm{d}(x)^2$ is min distance between x and centroids already chosen
3. Repeat step 2 until choosing k points

[1] Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

제안하는 방법 (initializer)

- 고차원 벡터 공간에서는 대부분의 거리가 비슷한 값을 지니기 때문에 *k*-means++ 의 $\frac{\mathrm{d}(x)^2}{\sum_{x \in D} \mathrm{d}(x)^2}$ 는 uniform 에 가까워진다.

Data set	<= 0.7	0.7 - 0.8	0.8 - 0.9	0.9 - 1.0
A6 blogs	0.249	0.378	1.628	97.745
Tucson blogs	0.59	1.121	3.89	94.399
Sonata blogs	0.323	0.455	1.984	97.239
IMDb reviews	0.272	7.751	57.271	34.706
Reuter RCV1	0.045	0.067	0.316	99.573
MovieLens 20M	1.456	2.386	12.458	83.7
Yelp reviews	0.01	0.286	11.073	88.632

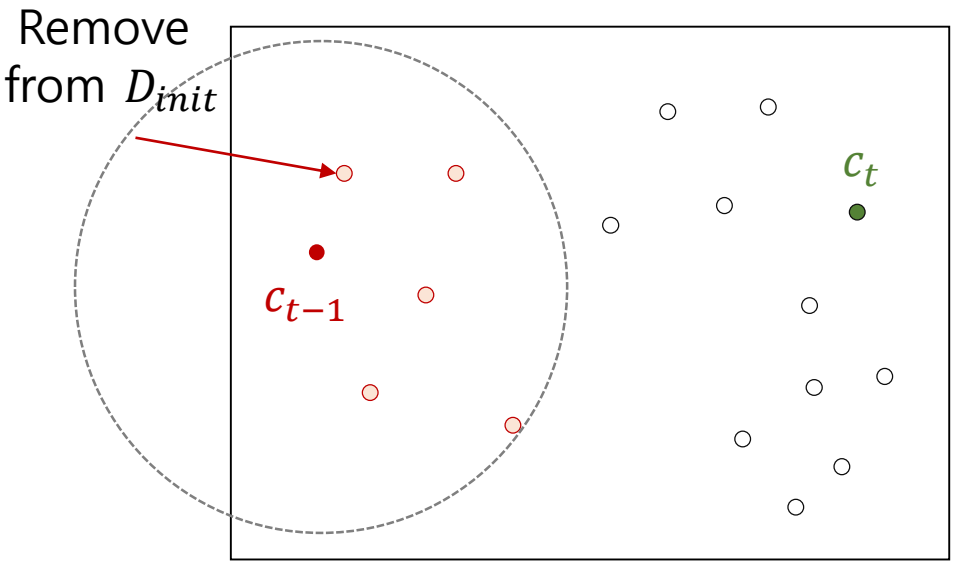
< 7 종류의 데이터에서의 문서 간 Cosine 거리 분포>

제안하는 방법 (initializer)

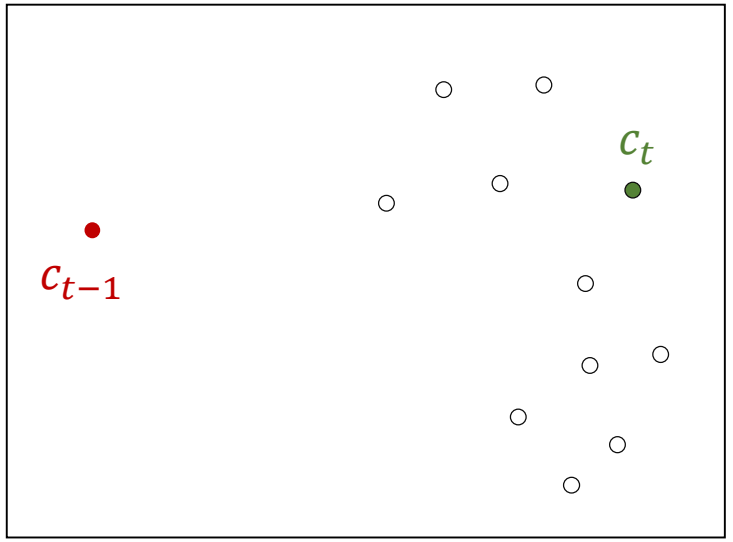
1. $\alpha * k$ 개의 후보를 random sampling $D_{init} \subset D$
2. D_{init} 에서 한 개의 점 c_i 을 임의로 선택.
3. D_{init} 에서 c_i 와 Cosine similarity 가 t_{init} 보다 큰 점을 D_{init} 에서 삭제
4. D_{init} 이 공집합이 아니면 k 개의 점을 뽑을 때까지 step 2 – 3 반복
5. D_{init} 이 공집합이며, 현재까지 선택한 점의 개수, k_0 가 $k_0 < k$ 이면
   $k - k_0$ 개의 점을 $D - D_{init}$ 에서 임의로 추출

제안하는 방법 (initializer)

- D_{init} 에서 c_{t-1} 주변 점들을 제거함으로 c_t 는 c_{t-1} 와 떨어져 있다.



c_{t-1} 에서 먼 점 중 하나는 c_t



$c_{t-1} \cong c_{t+1}$ 인 점은 선택되지 않음

제안하는 방법 (Centroid 기반 군집 레이블링)

- 군집 C_i 에서의 단어 w_{ij} 의 키워드 점수, $s(w, c_i)$
  - 다른 군집보다 상대적으로 자주 등장한 단어를 키워드로 정의

$$s(w, c_i) = \frac{p_{i\,(w)}}{p_i(w) + p_{-i}(w)}$$

$p_i(w)$: C_i 에서 단어 w 의 비율

$p_{-i}(w)$: C_i 외에서 단어 w 의 비율

제안하는 방법 (Pseudo code)

D: input dataset
k: a number of clusters
α: a factor for determining the search space of initial points
t_{init}: minimum distance between initial points
β: sparsity threshold factor
k_0: a number of keyword candidates
k_1: a number of selected keywords

def improved_spherical_kmeans $(D, k, \alpha, t_{init}, \beta, k_0, k_1)$:
  $C \leftarrow$ Initialize k centroids with sparse initializer (D, k, α, t_{init})
  $L \leftarrow$ Assign every points to its closest centroid
  while (not converged):
    $C \leftarrow$ Update centroid. Averaging their comprising data points
    $C \leftarrow$ Project centroids to sparse vector (C, L, β)
    $L \leftarrow$ Re-assign all the points to its closest centroid
  $Z \leftarrow$ cluster labeling (C, L, k_0, k_1)
  return C, L, Z

성능 평가 (initializer 의 정량적 평가)

- 7 종류의 데이터에 대하여 initializer 의 성능향상 정도를 측정하였다.

Data	A6 blogs	Tucson blogs	Sonata blogs	IMDb reviews	Reuter RCV1	MovieLens 20M	Yelp reviews
Number of documents	63,153	105,755	229,253	1,228,348	804,414	138,493	5,261,669
Total number of words	91,302	81,497	85,129	68,049	47,236	131,262	27,247
Number of nonzero elements	18,051,341	29,192,999	60,861,803	181,411,713	60,915,113	20,000,263	365,341,887
Sparsity	0.9969	0.9966	0.9969	0.9978	0.9984	0.9989	0.9974

< 데이터셋 통계 >

성능 평가 (initializer 의 정량적 평가)

- Initializer 의 속도가 개선되었으며, 데이터의 크기가 클수록 개선폭이 크다.

Data	α	k			
		10	20	50	100
A6 blogs	1	x 288	x 268	x 260	x 233
	3	x 265	x 257	x 213	x 150
	5	x 248	x 226	x 166	x 99
	10	x 217	x 159	x 100	x 54
	k-means++	5 s	10 s	25 s	52 s
Tucson blogs	1	x 464	x 487	x 397	x 367
	3	x 388	x 440	x 306	x 244
	5	x 358	x 376	x 261	x 172
	10	x 312	x 279	x 160	x 102
	k-means++	7 s	16 s	41 s	82 s
Sonata blogs	1	x 777	x 941	x 860	x 777
	3	x 785	x 841	x 614	x 495
	5	x 707	x 770	x 534	x 376
	10	x 600	x 615	x 330	x 208
	k-means++	16 s	33 s	86 s	175 s
IMDb review	1	x 1165	x 1257	x 2137	x 2253
	3	x 803	x 714	x 1988	x 1787
	5	x 1180	x 1172	x 1715	x 1381
	10	x 815	x 1062	x 1301	x 866
	k-means++	41 s	84 s	214 s	431 s

Data	α	k			
		10	20	50	100
Reuters RCV1	1	x 511	x 686	x 819	x 892
	3	x 439	x 713	x 850	x 772
	5	x 520	x 685	x 672	x 678
	10	x 518	x 639	x 622	x 425
	k-means++	13 s	28 s	71 s	146 s
MovieLens 20M	1	x 193	x 215	x 218	x 210
	3	x 202	x 213	x 214	x 184
	5	x 202	x 204	x 186	x 145
	10	x 189	x 172	x 144	x 103
	k-means++	4 s	9 s	24 s	49 s
Yelp reviews	1	x 484	x 876	x 1535	x 3092
	3	x 368	x 908	x 1508	x 2917
	5	x 362	x 903	x 1877	x 1595
	10	x 351	x 598	x 1143	x 2120
	k-means++	81 s	164 s	421 s	848 s

성능 평가 (initializer 의 정량적 평가)

- Initializer 가 바뀌었음에도 군집화 결과의 성능은 유지되는가?
  - k-means++ 을 이용한 경우와 비슷한 품질의 군집화 결과를 얻었다.

data	Intra distance	Inter distance	Silhouette score
A6 blogs	0.986	1.001	1.124
Tucson blogs	1.006	1.000	0.998
Sonata blogs	1.005	1.000	1.038
IMDb reviews	0.999	1.000	0.984
Reuters RCV1	0.999	1.001	1.019
MovieLens 20M	1.002	0.996	1.102
Yelp reviews	1.001	0.991	0.980

< 제안한 initializer 를 이용한 군집화 결과의 품질 값과
k-means++ 를 이용한 군집화 결과의 품질 값의 비율 >

성능 평가 (군집 레이블링의 정성적 평가)

- IMDB reviews (k=1,000)
  - 2,514 개의 영화 리뷰를 k=1,000 의 군집으로 학습
  - 5 개의 예시 군집들의 의미가 잘 파악된다.

영화 "타이타닉"	iceberg, zane, sinking, titanic, rose, winslet, camerons, 1997, leonardo, leo, ship, cameron, dicaprio, kate, tragedy, jack, disaster, james, romance, love, effects, special, story, people, best, ever, made
Marvle comics 의 heros (Avengers)	zemo, chadwick, boseman, bucky, panther, holland, cap, infinity, mcu, russo, civil, bvs, antman, winter, ultron, airport, avengers, marvel, captain, superheroes, soldier, stark, evans, america, iron, spiderman, downey, tony, superhero, heroes
Cover-field, District 9 등 외계인 관련 영화	skyline, jarrod, balfour, strause, invasion, independence, cloverfield, angeles, district, los, worlds, aliens, alien, la, budget, scifi, battle, cgi, day, effects, war, special, ending, bad, better, why, they, characters, their, people
살인자가 출연하는 공포 영화	gayheart, loretta, candyman, legends, urban, witt, campus, tara, reid, legend, alicia, englund, leto, rebecca, jared, scream, murders, slasher, helen, killer, student, college, students, teen, summer, cut, horror, final, sequel, scary
영화 "매트릭스 "	neo, morpheus, neos, oracle, trinity, zion, architect, hacker, reloaded, revolutions, wachowski, fishburne, machines, agents, matrix, keanu, smith, reeves, agent, jesus, machine, computer, humans, fighting, fight, world, cool, real, special, effects

< IMDB 에서의 군집 레이블링 결과 예시 >

성능 평가 (군집 레이블링의 정성적 평가)

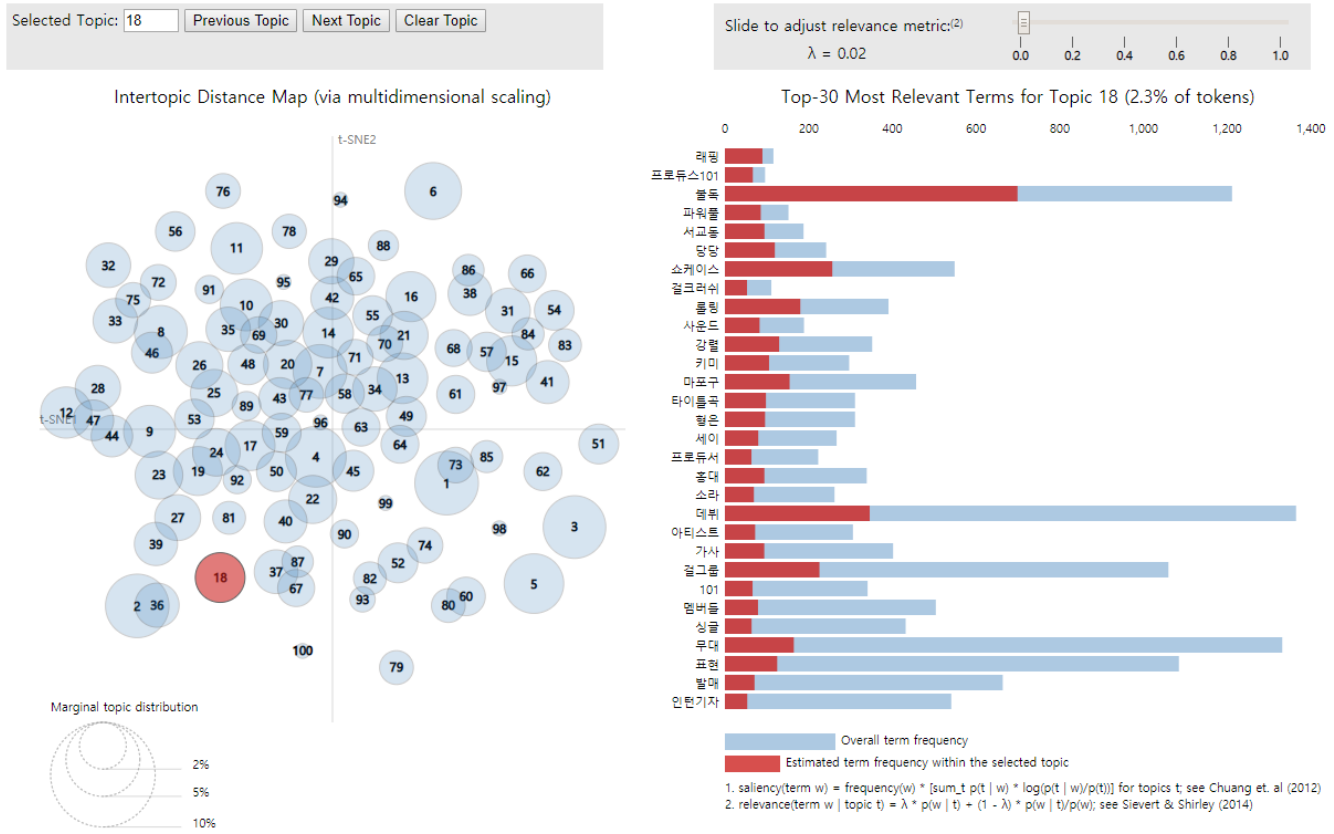
- "소나타"가 포함된 네이버 블로그 (k=500)
  - 소나타는 다의어이기 때문에 다양한 문맥에서 이용된다.
  - 비지도학습기반 단어 추출을 수행하는 토크나이저를 이용하였다.

렌트카 광고	제주렌트카, 부산출발제주도, 제주신, 이끌림, 제주올레, 왕복항공, 불포함, 제주도렌트카, 064, 롯데호텔, 자유여행, 객실, 제주여행, 특가, 해비치, 제주시, 제주항, 티몬, 2박3일, 올레, 유류, 항공권, 조식, 제주도여행, 제주공항, 2인
중고차 매매	최고급형중고, 최고급, 프리미어, 프라임, 2011년식, YF소나타TOP, 2010년식, 풀옵션, 2011년, YF소나타PR, 1인, Y20, 2010년, 완전무사고, 판매완료, 군포, 검정색, YF쏘나타, 2011, 하이패스, 2010, 무사고, 등급, 파노라마, 허위매물
클래식 음악	금관악기, 아이엠, Tru, 트럼펫, 트럼, 나팔, 금관, 텔레만, Eb, 호른, 오보에, Tr, Concerto, 하이든, 협주곡, Ha, 악기, 연주하는, 오케, 오케스트라, 독주, 악장, 작곡가, 곡
아이비 "유혹의 소나타 "	Song, 공부할, 부른, 노래, 가사, 부르는, 가수, 보컬, 목소리, 발라드, 명곡, 신나, 들으면, 듣기, 유혹의, 앨범,아이비, 제목
광염 소나타 및 일제강점기 소설들	백성수, 발가락, 현진, 이광수, 김유, 자연주의, 친일, 평양, 운수, 유미, 저지르, 야성, 탐미, 김동인, 복녀, 광염, 닮았다, 사실주의, 광기, 저지, 1920, 단편소설, 범죄, 감자, 동인, 한국문학

< "소나타"가 포함된 블로그의 군집 레이블링 예시 >

성능 평가 (군집 레이블링의 정성적 평가)

- PCA 를 이용하여 2 차원 벡터로 표현, 군집 레이블을 키워드로 이용하면 군집화 결과를 시각적으로 표현할 수 있다.



< pyLDAvis 와 군집 레이블링 결과를 이용한 군집화 시각화 예시 >

결론

- 여러 주제를 포함하는 문서 군집을 요약하기 위한 문서 군집화 및 군집 레이블링 방법을 제안하였다.
  - 제안한 방법은 적은 학습 시간으로 *k*-means++ 과 비슷한 품질의 군집화 결과를 학습할 수 있다.
  - Centroid 를 이용한 군집 레이블링 방법으로도 군집의 의미를 해석할 수 있다.

추가 연구

- k-means 계열의 군집화 방법은 적절한 k 를 사용자가 직접 설정해야 한다.
  - 군집 레이블링 방법은 각 군집이 서로 다른 주제로 나뉘어졌다고 가정한다.

- Outlier, imbalanced class distribution 상황에서 잘 작동하지 않는다.
  - Outlier 에서 키워드가 선택될 경우 군집 해석이 어려워진다.

배경

- 시계열 형식으로 생성되는 문서 집합의 시간 정보는 주제의 변화를 설명하는 정보로 이용될 수 있다.
- Mixed type 은 군집화 방법이 이용되기 어렵다.
  - 단어 빈도와 시간으로 이뤄진 벡터는 점들 간 거리를 정의하기 어렵다.

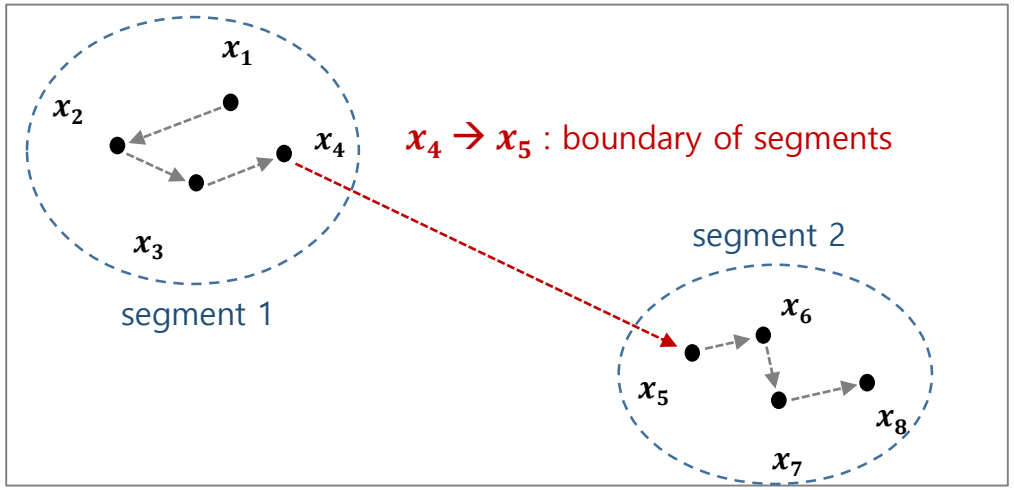
관련 연구

- 시간 구간에 따라 여러 개의 LDA 모델을 학습하는 방법 ^[1]
  - 시점 t 의 토픽 – 단어 행렬 β_t 이 β_{t-1} 와 비슷하도록 제약, $\beta_t \sim N(\beta_{t-1}, \sigma^2 I)$
  - 매 구간 별 토픽 개수가 모두 동일하며, 사용자가 임의로 구간을 정의해야 함
- 군집 간 거리를 이용한 새로운 토픽 탐색 ^[2,3]
  - 이전 시점의 군집과의 거리가 임계값 이상인 문서는 새로운 군집으로 구성
  - 여러 개의 세부 토픽이 발생하여 큰 트렌드를 요약하기 어려움

[1] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120. ACM, 2006
[2] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 370–378. ACM, 2012.
[3] Dominik Wurzer, Victor Lavrenko, and Miles Osborne. Tracking unbounded topic streams. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 1765– 1773, 2015.

제안하는 방법

- 시계열로 생성되는 x_t 의 거리가 멀어지는 지점을 구간의 경계로 추출
  - 시간 별 문서 집합을 벡터 x_t 로 표현
  - 각 구간은 문서 군집으로 해석할 수 있으며, 앞 장의 키워드, 핵심 문장 추출 방법을 이용하여 각 구간의 문서 집합을 요약



concept of time-series segmentation in vector space

제안하는 방법

D: documents stream
w: window length
m_l: minimum segment length
m_b: minimum boundary score
k_0: number of keyword candidates
k_1: number of selected keywords
k_2: number of keysentences
σ: minimum distance between selected key-sentences

def summarize_documents_stream $(D, m_l, m_b, k_0, k_1, k_2)$:
  $KW = [\,]$: keyword list
  $KS = [\,]$: keysentence list
  $X \leftarrow \text{encode}(D)$
  $B \leftarrow$ compute change point score(X, w)
  $S \leftarrow$ segment documents stream(X, B, m_l, m_b)
  for s in S:
    $KW_s, KV_s \leftarrow$ segment labeling (S, s, k_0, k_1)
    $KS_s \leftarrow$ seleft keysentences (s, KV_s, σ, k_2)
    $KW += KW_s$
    $KS += KS_s$
  return KW, KS

성능 평가

- 질의어 별로 수집된 뉴스 데이터의 문서 집합 요약을 통한 정성 평가
  - 2013-01-01 ~ 2019-03-10 (2,260 일) 간의 정치인 이름이 포함된 뉴스 수집
  - 각 질의어 (김무성, 박근혜, 유시민) 별 트렌드가 변하는 구간을 분리한 뒤,
    6 장의 키워드 추출 방법과 5 장의 핵심 문장 추출 방법을 적용

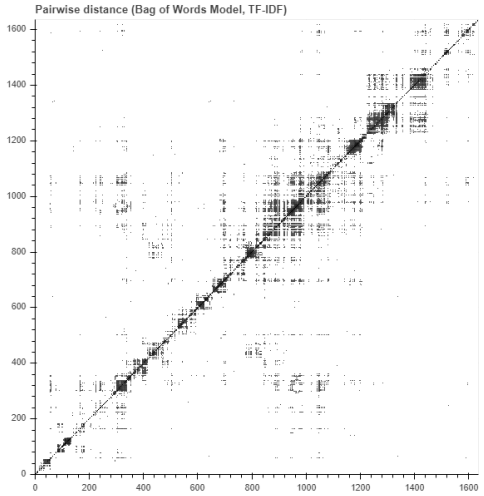
성능 평가

- 명사 추출기와 L-Tokenizer 를 이용하여 전처리를 수행하였다
  - 질의어가 포함된 문서 집합마다 각 일별로 생성된 뉴스를 병합한 뒤
  - 불용어를 제거하기 위하여 TF-IDF 로 벡터화를 수행하였다

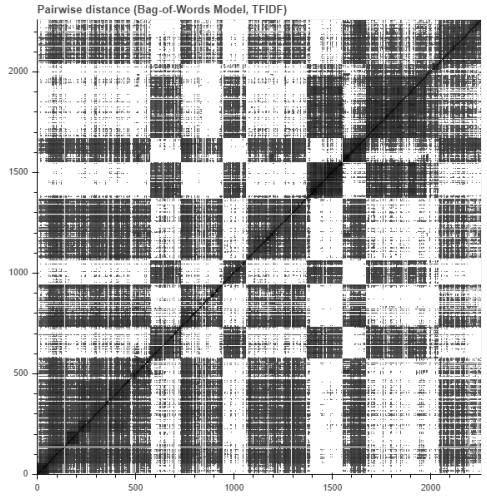
질의어	고유 날짜 수	뉴스 개수	고유 단어 개수	일평균 뉴스 개수
김무성	1,636	223,590	69,597	136.67
박근혜	2,260	1,339,266	249,152	592.60
유시민	447	18,239	13,619	40.90

성능 평가

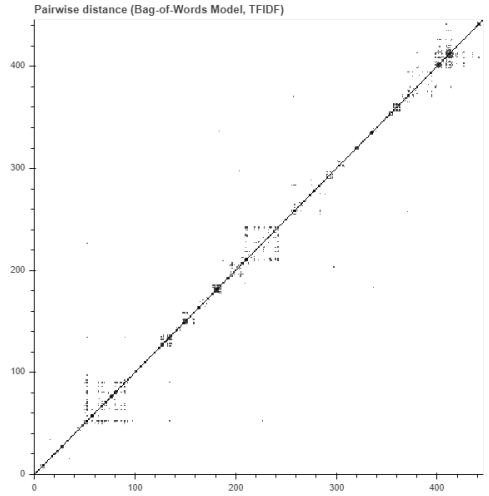
- 데이터 별 평균 구간의 길이와 주제의 세분성 (granularity) 이 다르다
  - 한 시점 여러 주제의 문서가 발생할수록 상위 주제 기준으로 구간이 나뉜다
  - 일 간 거리 행렬 (pairwise distance of date) 로부터 전체 구조를 파악할 수 있다



'김무성'의 일 간 거리 행렬
구간 개수 : 99 개
구간 평균 길이 : 8.74 일



'박근혜'의 일 간 거리 행렬
구간 개수 : 68 개
구간 평균 길이 : 32.5 일



'유시민'의 일 간 거리 행렬
구간 개수 : 91 개
구간 평균 길이 : 1.5 일

성능 평가 (김무성)

2016-12-10 ~ 2016-12-30	**[키워드]** 개혁보수신당, 보수신당, 가칭, 신당, 창당, 개혁, 중도, 분당, 비대위원장, 보수, 탈당, 유승민, 유, 주호영, 13일, 당을, 원내, 비상시국회, 분열, 친박계, 모임, 황영철, 나경원, 정책, 친박, 선출, 고민, 의원, 의원회관, 비박, 선언
	- 김무성 유승민 등 새누리당 비박 비박근혜 계 의원 34명이 탈당을 선언하고 신당의 가칭을 보수신당 이라고 정했다 - 28일 국회 의원회관에서 주호영 원내대표 김무성 유승민 의원 등이 참석한 가운데 가 개혁보수신당 정강정책 토론회가 열리고 있다 - 김무성 전 대표가 친박 지도부로는 당 재건이 어렵다며 탈당을 시사해 비대위원장 선출이 분당 분수령이 될 거란 관측이 나옵니다 - 아시아경제 윤동주 기자 김무성 새누리당 전 대표가 26일 국회에서 열린 가칭 개혁보수신당 창당 준비위원회의에 참석하고 있다 - 김무성 전 새누리당 대표가 친박근혜계를 향해 대통령의 정치적 노예들 이라며 강하게 비판하며 중도 보수 창당을 고민하고 있다고 밝혔다
2017-05-25 ~ 2017-05-26	**[키워드]** 세비, 유병재, 패러디, 반납, 방송인, 서명, 광고, 노룩, 노룩패스, 공항, 네티즌들, 웃음, 캐리어, 화제, 제목, 공개, KBS, 커뮤니티
	- 서울경제 방송인 유병재가 김무성 의원의 캐리어 노 룩 패스 를 패러디해 눈길을 끌었다 - 그룹 투포케이 역시 26일 오전 KBS 2TV 뮤직뱅크 리허설을 가는 길에 김무성의 노룩패스 패러디를 해 웃음을 자아냈다 - 앞서 김무성 의원이 공항에서 수행원을 보지 않고 캐리어를 굴려 전달하는 모습이 포착돼 큰 화제를 모은 바 있다 - 김무성 의원 등 전 새누리당 소속 의원 27명이 오는 5월31일까지 1년치 세비를 국가에 반납할 지 관심이 쏠리고 있다 - 김무성 의원의 행동은 온라인에서 지탄을 받았고 해외 온라인 커뮤니티 레딧 등에서 한국인의 스웨그 라는 제목으로 공개되기까지 했다
2017-11-02 ~ 2017-11-15	**[키워드]** 탈당계, 재입당, 복당파, 9명, 9일, 탈당파, 복당, 홍철호, 8명, 입당, 강길부, 지위, 정양석, 의총, 제출, 간담회, 탈당, 선언, 이종구, 황영철, 김용태, 집단, 당사, 기자회견, 김영우, 교섭단체, 의원총회, 친박, 창당, 예정, 홍준표, 오전
	재입당 국회의원 간담회에서 바른정당을 탈당한 김무성 강길부 김영우 김용태 이종구 황영철 정양석 홍철호 의원과 손을 잡고 기념촬영을 하고 있다 - 바른정당은 지난 8일 김무성 의원 등 8명이 일제히 탈당계를 제출한 뒤 자유한국당에 복당 교섭단체 지위를 잃었다 - 홍준표 자유한국당 대표가 9일 오전 서울 여의도 당사에서 열린 바른정당 탈당파 의원들의 재입당 간담회에서 김무성 의원과 대화를 나누고 있다 - 바른정당의 김무성 의원 등 8명이 6일 탈당을 선언하기로 했다 이로써 바른정당은 창당 9개월여 만에 절반으로 쪼개지게 됐다 - 실제 강성 친박계는 김무성 의원 등 8명의 탈당파 복귀만으로도 의총 소집을 요구하는 등 조직적 반발 기류를 드러내고 있다

성능 평가 (박근혜)

2013-01-02 ~ 2013-02-24	**[키워드]** 내정자, 후보자, 인사청문, 의혹, 삼청동, 지명, 취임식, 핵실험, 김용준, 장관, 대통령직인수위원회, 업무, 청와대, 오후, 부처, 인수위원회, 대통령직, 위원장, 논의, 검토, 방안
	- 정 총리 후보자는 8일 삼청동 인수위에서 기자회견을 갖고 박근혜 정부의 초대 국무총리로 지명된 소감에 대해 이같이 말했다 - 박근혜 대통령 당선인과 김용준 대통령직인수위원회 위원장이 30일 오후 서울 종로구 삼청동 대통령직 인수위원회에서 열린 정무분과 업무보고에 참석해 있다 - 20일 정홍원 국무총리 후보자의 국회 인사청문회를 시작으로 박근혜 정부의 17개 부처 초대 장관 내정자들의 청문회가 줄줄이 이어진다 - 박근혜 대통령 당선인은 이번주 초 총리 후보자와 청와대 인선을 함께 발표하는 방안을 검토하고 있는 것으로 전해졌습니다 - 박근혜 정부 초대 국무총리로 지명된 김용준 내정자 75 의 검증 과정에서 재산 문제가 가장 먼저 도마에 오를 것으로 보인다
2016-10-02 ~ 2016-12-02	**[키워드]** 하야, 국정농단, 최순실, 촛불, 퇴진, 시국, 탄핵, 게이트, 비선, 특검, 광화문, 행진, 검찰, 수사, 집회, 대국민, 시민, 의혹, 분노, 재단, 혐의, 헌법, 사태, 조사, 개입, 사건, 촉구, 요구, 사과, 임명, 연설
	- 비선 실세 최순실 씨의 국정농단 사태의 책임을 물어 박근혜 대통령의 퇴진을 촉구하는 5차 주말 촛불집회가 26일 열린다 - 774억 원을 모금해 세웠는데 재단에 박근혜 대통령 측근인 최순실 차은택 씨가 개입했다는 의혹은 현재 검찰 수사가 진행 중입니다 - 5일 오후 서울 광화문광장에서 박근혜 정권 퇴진 집회에 가족 단위로 참여한 시민들이 촛불을 들고 행진하고 있다 - 박영수 변호사가 박근혜 게이트 수사의 특별검사로 임명되었다 대통령은 특검 수사에 적극적으로 협조하고 직접 조사에도 응하겠다고 한다 - 국정농단 으로 박근혜 대통령의 하야와 탄핵을 요구하는 대학가의 시국선언이 번지고 있다
2018-02-18 ~ 2018-04-27	**[키워드]** 1심, 구형, 세월호, 징역, 이명박, 유죄, 피고인, 선고, 한국당, 참사, 4월, 대변인, 국정, 혐의, 법원, 역사, 재판, 원장, 검찰, 인정, 사건, 조사, 뇌물, 소환, 지난해, 회장, 구속, 받은, 판단
	- 국정농단 사건으로 징역 30년을 구형받은 박근혜 전 대통령의 1심 선고가 오는 4월 6일 내려집니다 - 또 박근혜 전 대통령이 지난해 3월 21일 검찰 조사를 받은 지 358일만에 소환된 전직 대통령이 됐습니다 - 법원은 박근혜 전 대통령의 삼성그룹 뇌물수수 혐의와 관련해 앞서 일부 유죄가 인정된 최순실씨와 동일한 판단을 내렸다 - 법원이 최순실 박근혜 1심 선고를 근거로 이 부회장 항소심 재판부의 판단이 잘못됐다고 본다면 뇌물액수가 커질 가능성이 크다 - 그는 이명박 박근혜 전직 대통령이 구속된 데 대해 한국당을 향해 쓴 목소리도 냈다

성능 평가 (유시민)

2016-03-04 ~ 2016-03-04	**[키워드]** 자본, 무전취식, 삼청각, 시술, 20만, 고급, 세종문화회관, 직권상정, 국회의장, 직원, 테러방지법, 필리버스터, 홍보, 식사, 문구, 가득, 몰려, 방청객, 국가, 외모, 원인, 넥타이
	- 이어 유시민은 고급 시술전문 성형외과나 피부과에서 조용히 한다 며 요즘은 외모도 신체 자본이다 이라고 말했다 - 이날 유시민 작가는 국회의장이 국가비상사태 라고 테러방지법을 직권상정했다 고 말문을 열었다 - 유시민은 방청객만 가득하고 의원석은 텅텅 비어있다 면서 편의점에 알바생 하나 있고 손님들만 몰려 있는 것과 같은 것 이라고 답했다 - 한편 이날 전원책과 유시민은 필리버스터 삼청각 무전취식 논란 등 다양한 사안을 놓고 토론을 펼쳤다 - 유시민은 세종문화회관 직원인 정팀장의 행각을 지적하며 20만 9천 원 짜리 밥 구경도 못해본 사람이 99.9% 다 라고 말했다
2017-07-29 ~ 2017-07-29	**[키워드]** 지식인, 보성, 시즌2, 순천, 통영, 출연진, 제작, 양정우, 도시, 28일, 기획, 지식
	- 이후 편집 영상에서 유시민은 여행지와 관련한 풍부한 지식을 드러내 최고의 지식인다운 면모를 보였다 - 제가 유시민 선생님의 이름을 많이 팔고 다녔죠 라고 말한 양정우 PD는 섭외 비하인드에 대해 털어놓기도 했다 - 통영 순천 보성 강릉 경주 공주 부여 세종 춘천 전주 등 전국 10개 도시를 돌아다녔다 - tvN 알쓸신잡 은 28일 마지막 방송에서 출연진이 한 공간에 모여 그간 방송에서 전하지 못한 뒷얘기를 나눴다 - 한편 알쓸신잡 후속으로는 삼시세끼 바다목장편 이 방송되며 시즌2는 현재 기획 중이다
2018-05-25 ~ 2018-05-25	**[키워드]** 관전, 군사적, 불안감, 멸균, 남북고위급회담, 취소, 후보자, 제재, 포인트, 마감, 네거티브, 단일화, 발제, 북한, 무소속, 비핵화, 한반도, 불안
	- 24일 방송된 JTBC 썰전 에서 유시민 작가가 남북고위급회담 취소에 대한 자신의 의견을 밝혔다 - 그리고 북한이 원하는 게 있다 면서 군사적 안전 보장 과 국제 무대에서의 제재 철회 를 꼽았다 - 이어 이런 것은 보건학적으로 설명이 가능하다 북한은 오랜 시간 주체사상 외 모든 다양한 의견을 멸균했다 고 덧붙였다 - 6.13 지방선거의 두 번째 관전 포인트 네거티브입니다 선거 시즌만 되면 고질적으로 되풀이 되는 네거티브 논란 이번에도 시동이 걸렸습니다 - 이어 북한은 아직도 비핵화를 한반도 전체 비핵화로 이해하고 있다 그것은 미국 전략자산 배제하는 것을 포함한다 고 덧붙였다

결론

- 벡터 거리 기반의 시계열 분리 알고리즘을 적용하여 문서 집합의 주제가 변하는 시점을 기준으로 구간을 나누는 방법을 제안하였다.
- 각 구간은 문서 군집으로 해석되어 4, 5 장의 문서 요약 방법이 적용된다.
  - 구간 내 주제가 다양할 경우 상위 주제에 의하여 구간이 구분되며, 주제가 단일한 경우 세분화된 주제에 의하여 구간이 구분된다.
  - 구간 내 키워드와 핵심 문장은 각 기간의 문서 집합을 요약할 수 있다.

추가 연구

- 구간 내 주제가 다양한 경우 큰 트렌드와 다른 세부 주제가 존재하며, 밀접한 주제 내의 계층 구조가 존재한다.
  - 계층 구조를 기반으로 한 문서 집합을 요약하는 방법이 연구되어야 한다.
  - 큰 트렌드와 다른 세부 주제의 문서는 노이즈로 제거되어야 한다.

- 한국어 문서의 토크나이징과 문서 요약 과업을 위한 비지도기반 자연어처리 방법을 제안하였다.
  - 토크나이저와 명사 추출 방법은 학습 말뭉치를 이용하는 모델들과 비슷하거나 더 높은 단어 인식 능력을 보였다.
  - 단일 주제로 구성된 문서 집합에서 미등록단어 문제 해결과 핵심 문장의 다양성 보장을 위한 키워드/핵심 문장 추출 방법이 제안되었다.
  - 다양한 주제로 구성된 문서 집합을 요약하기 위한 효율적인 문서 군집화 방법 및 군집 별 키워드 추출 방법은 기존 군집화 방법보다 효율적으로 학습하며, 좋은 품질의 키워드로 각 군집을 요약하였다.
  - 시계열 형식의 문서 집합을 요약하기 위하여 트렌드가 변하는 시점을 분리한 뒤, 각 구간에서 키워드와 핵심 문장을 추출하는 방법을 제안하였다.

- 토크나이저는 다른 과업에 이용되는 데이터를 처리하는 기초 과업으로 알려진 문장에 대해서는 지도기반 접근법이 더 정확히 작동하나, 알려지지 않은 문장에서는 오작동할 위험이 있다.
- 두 접근 방법을 상호 보완적으로 이용하는 방법에 대한 후속 연구가 필요하다.
  - 비지도기반 접근법은 분석할 문서 집합에 대한 특징을 잘 추출할 수 있다.
  - 지도기반 접근법은 문맥적 모호성 해결과 용언의 원형복원 같은 과업에 좋은 성능을 보인다.
  - 두 접근법을 보완한다면 정확하면서도 데이터에 적합하게 작동하는 토크나이저를 얻을 수 있을 것으로 기대한다.