

키워드 기반 인과 관계 검색 서비스

NLP Center 검색기술실 내러티브팀 김건

Confidential

2023.02.14 검색기술실 내러티브팀
(인턴십 기간 : 2022-09-05 ~ 2023-02-24)



Contents

01. 서론	3
02. 비지도 문서 키워드 추출	9
03. 문장 인과 관계 분류 및 추출	29
04. 트렌드 키워드 시스템	41

Contents

01. 서론	3
a. 동기	4
b. 데모	5
c. 서비스 구조	8
02. 비지도 문서 키워드 추출	9
03. 문장 인과 관계 분류 및 추출	29
04. 트랜드 키워드 시스템	41

동기

• 키워드 기반 인과 관계 분석

- 키워드 기반 분석

- **키워드들**은 관심 있는 타겟 도메인의 핵심 지식
- 사용자에게 문서들의 중요 정보를 빠르게 파악할 수 있는 효과를 제공
- 정보 검색과 요약 등의 핵심 태스크의 중요한 정보를 제공함

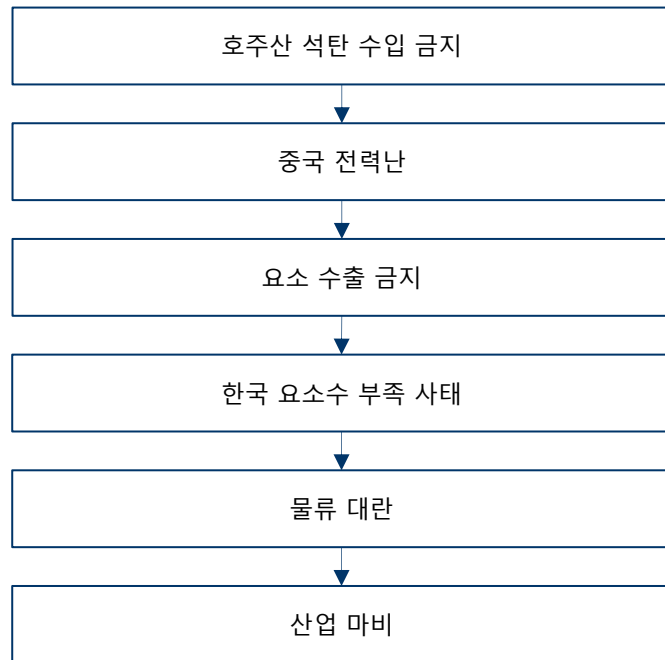
- 인과 관계 분석

- 이벤트 간의 인과 관계 이해는 의학 분야와 금융 분야 등의 다양한 분야에 도움을 주는 핵심 자연어 처리 태스크 중 하나
- 인과 관계 탐지 작업은 실제 환경에서의 문서들에서 나타나는 변화들의 발생 원인을 찾는 것을 목표로 함

- 키워드 기반 인과 관계 검색 시스템

- 키워드 기반 분석과 인과 관계 분석을 결합한 시스템
- 대규모 텍스트 집합에서 사용자가 관심있는 키워드의 인과 관계 구성을 자동으로 탐색하여 간략하게 요약하는 시스템
- **금융 환경 변화 및 각종 사건의 인과 관계를 분석하는 중요한 정보**를 제공

- "2021년 요수수 대란" 인과관계 예제



데모

• 뉴스 도메인 키워드 추출

- http://geon6757-search-web.cloud.ncsoft.com/keyword_extraction

뉴스 키워드 추출 키워드 기반 인과관계 검색 트렌드 키워드

뉴스 키워드 추출

국내 등록된 자동차 2천500만 대 가운데 디젤차 1천만 대에 요소수 불통이 뒤얹었습니다. 매년 전 세계로 수출되는 중국산 요소수 500만t 가운데 절반 가까운 47%가 인도로 유입되고, 한국은 두 번째 많은 14%를 수입합니다. 국내 요소수 전량을 중국에서 수입하던 차인데요. 그런데 중국이 호주와의 '석탄 분쟁'을 겪으면서 사실상 요소수 수출을 중단한 상태이기 때문에 한국이 직격탄을 맞았습니다. 과거에는 국내에서도 요소수를 생산하는 업체들이 있었으나, 중국, 러시아 등과 비교해 가격 경쟁력이 떨어지면서 요소수 생산 업체들이 2013년 전후로 모두 없어졌는데요. 이 때문에 중국이 수출을 재개하는 것이 유일한 방안이지만, 현재 수출을 재개할지는 미지수입니다. 소방 당국은 요소수 사태가 장기화할 경우를 대비해 재고 관리에 힘을 쏟고 있습니다. 전국에서 운영하는 6천748대 소방차 중 80.5%가, 1천675대 구급차량 중 90.0%가 요소수를 사용하는 차량입니다. 중국발 요소수 공급 현상으로 충북지역 제조·판매업체가 문을 닫는 등 피해가 잇따르고 있습니다. 일부 주유소는 요소수 판매 중지해 나섰고, 물류 차량 운전기사는 천정부지로 치솟은 요소수를 '올머저자먹기식'으로 구매해 운행하는 실정인데요.

추출 모델 선택 : ☐ TF-IDF ☐ Pointwise-KLD ☐ UKERank ☒ CorpusRank

추출 키워드 개수 : 15

추출하기

뉴스 키워드 추출 키워드 기반 인과관계 검색 트렌드 키워드

뉴스 키워드 추출 결과

요소수 국내 등록 2천500만 대 수출 1천만 대 디젤차 500만t 중국 올머저자 6천748대 불통 1천675대 지동지 천당

47%

검색 초기화 위 키워드들로 인과관계 검색하기

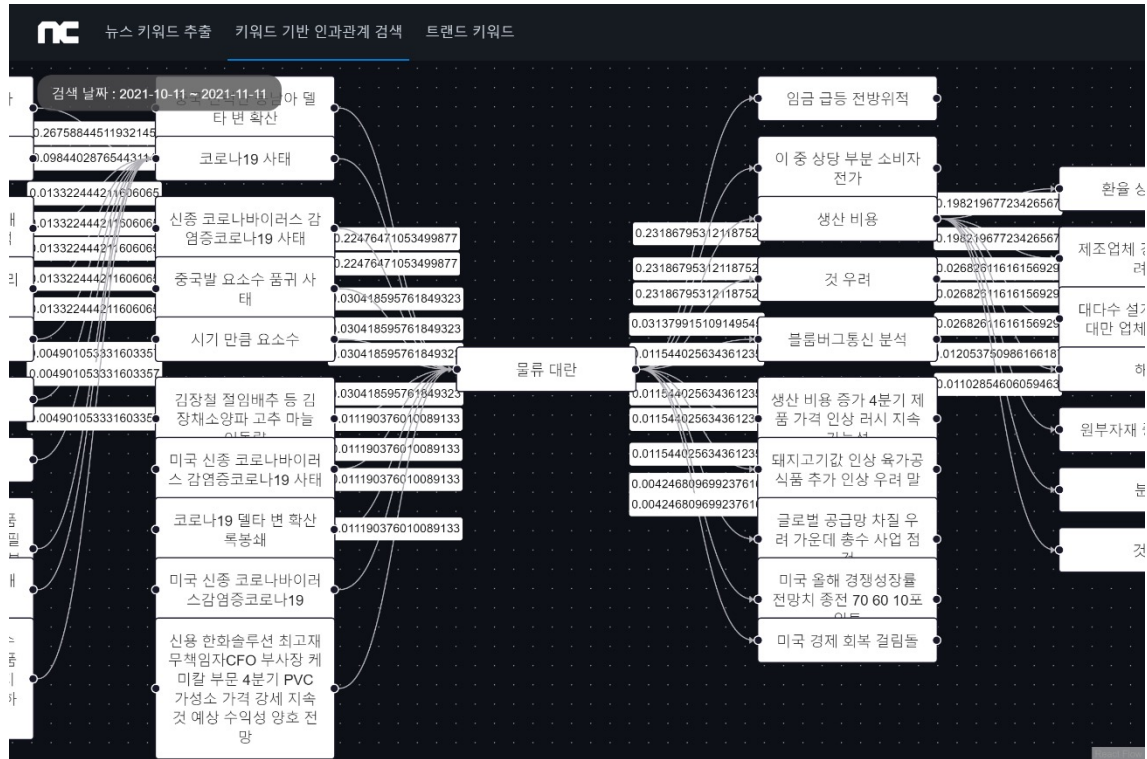
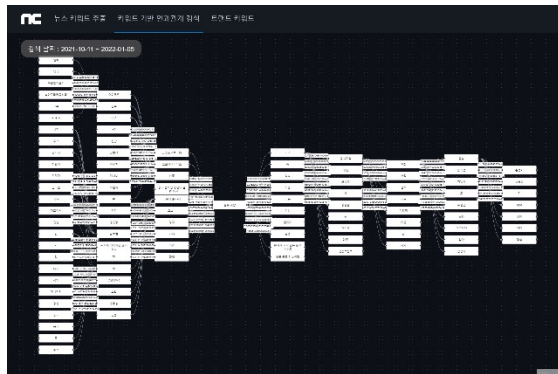
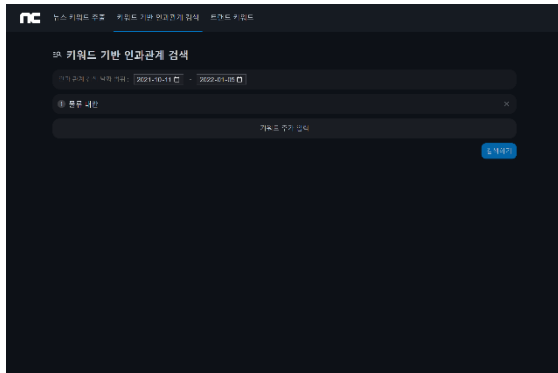
뉴스 청킹 결과

국내 등록 된 자동차 2천500만 대 가운데 디젤차 1천만 대 에 요소수 불통 이 뒤얹었습니다. 매년 전 세계 로 수출 되는 중국산 요소수 500만t 가운데 절반 가까운 47% 가 인도 로 유입 되고, 한국 은 두 번째 많은 14% 를 수입 합니다. 국내 요소수 전량 을 중국 에서 수입 하던 차인데요. 그런데 중국 이 호주 와의 '석탄 분쟁 을 겪으면서 사실상 요소수 수출 을 중단 한 상태 이기 때문에 한국 이 직격탄 을 맞았습니다. 과거 에는 국내 에서도 요소수 를 생산 하는 업체 들이 있었으나, 중국 , 러시아 등과 비교 해 가격 경쟁력 이 떨어지면서 요소수 생산 업체 들이 2013년 전후 로 모두 없어졌는데요. 이 때문에 중국 이 수출 을 재개 하는 것이 유일 한 방안 이지만, 현재 수출 을 재개 할지는 미지수 입니다. 소방 당국 은 요소수 사태 가 상기화 될 경우를 대비 해 재고 관리 에 힘 을 쏟고 있습니다. 전국 에서 운영 하는 6천748대 소방차 중 80.5 %가, 1천675대 구급차량 중 90.0 %가 요소수 를 사용 하는 차량 입니다. 중국발 요소수 공급 현상 으로 충북 지역 제조 · 판매업체 가 문 을 닫는 등 피해 가 잇따르고 있습니다. 일부 주유소 는 요소수 판매 중지 에 나섰고, 물류 차량 운전기사 는 천정부지 로 치솟은 요소수 를 '올머저자먹기식'으로 구매 해 운영 하는 실정 인데요.

데모

• 키워드 인과 관계 검색

- http://geon6757-search-web.cloud.ncsoft.com/keyword_search



데모

트렌드 키워드

- http://geon6757-search-web.cloud.ncsoft.com/trend_keyword

NC 뉴스 키워드 추출 키워드 기반 인과관계 검색 트렌드 키워드

트렌드 키워드

트렌드 키워드 검색 옵션

월별 트렌드 키워드 2023년 2월 6일

검색하기

정치	경제	사회	세계
1 이 장관	1 지난해	1 유가족	1 후보키예
2 후보	2 6일	2 김 전 회장	2 지진
3 6일	3 올해	3 6일	3 시리아
4 외교	4 최근	4 매경	4 발생
5 아이뉴스24	5 이날	5 김성태	5 각수
6 윤석열	6 1분전	6 조 전 장관	6 회소
7 연대	7 증권부	7 불발물	7 건물
8 정치	8 가려한희남	8 서울경찰	8 김진
9 통일	9 이사회	9 실종자	9 이날
10 발의	10 등	10 분향소	10 남부

스포츠	IT과학	생활/문화	기타
1 수원	1 대동항실	1 이태원리 스타인	1 때
2 6일	2 지난해	2 때	2 지난해
3 스포츠 기자	3 당	3 .06	3 난방비
4 아레나	4 여당	4 2023.02	4 WSJ
5 MK	5 이 장관	5 중형	5 이후
6 2022-2023 프로농구	6 대통령	6 베스트	6 최근
7 수원 KT	7 이장민	7 이날	7 난방
8 한진한	8 위법	8 최근	8 피의자
9 KGC	9 행정안전부 장관	9 지난해	9 100일
10 지영은 기자	10 때	10 강만길	10 설치

NC 뉴스 키워드 추출 키워드 기반 인과관계 검색 트렌드 키워드

트렌드 키워드

트렌드 키워드 검색 옵션

월별 트렌드 키워드 2022년 12월

검색하기

정치	경제	사회	세계
1 윤 대통령	1 이 시각	1 협의	1 러시아
2 대동항실	2 기록	2 이태원	2 우크라이나
3 이태원	3 1개월간	3 눈	3 머스크
4 여당	4 체결강도	4 참사	4 부인 대통령
5 여당	5 종매수제결합	5 선정	5 철원수기 대통령
6 이상민	6 종매도제결합	6 구속영장	6 전철
7 예산안	7 가려대금	7 화물연대	7 우크라이나 전쟁
8 참사	8 2023년	8 검찰	8 IRA
9 국정조사	9 화물연대	9 아시아경제	9 이브생티나
10 이재명 대표	10 선전	10 피해자	10 월드컵

스포츠	IT과학	생활/문화	기타
1 OSEN	1 아시아경제	1 OSEN	1 뉴시스
2 월드컵	2 이태원리	2 2022.12	2 화물연대
3 2022.12	3 민주당	3 포츠	3 연합뉴스보oces
4 아르헨티나	4 위믹스	4 osen	4 파업
5 브라질	5 아이뉴스24	5 co	5 허락
6 메시	6 이상민	6 스포츠조선닷컴	6 오진
7 포르투갈	7 여당	7 sportschosun	7 거래
8 프랑스	8 구 대표	8 배우	8 BOJ
9 ESG	9 16강	9 스포츠조선	9 예산안
10 sportschosun	10 선전	10 com	10 승진

NC 뉴스 키워드 추출 키워드 기반 인과관계 검색 트렌드 키워드

트렌드 키워드

트렌드 키워드 검색 옵션

년도별 트렌드 키워드 2023년

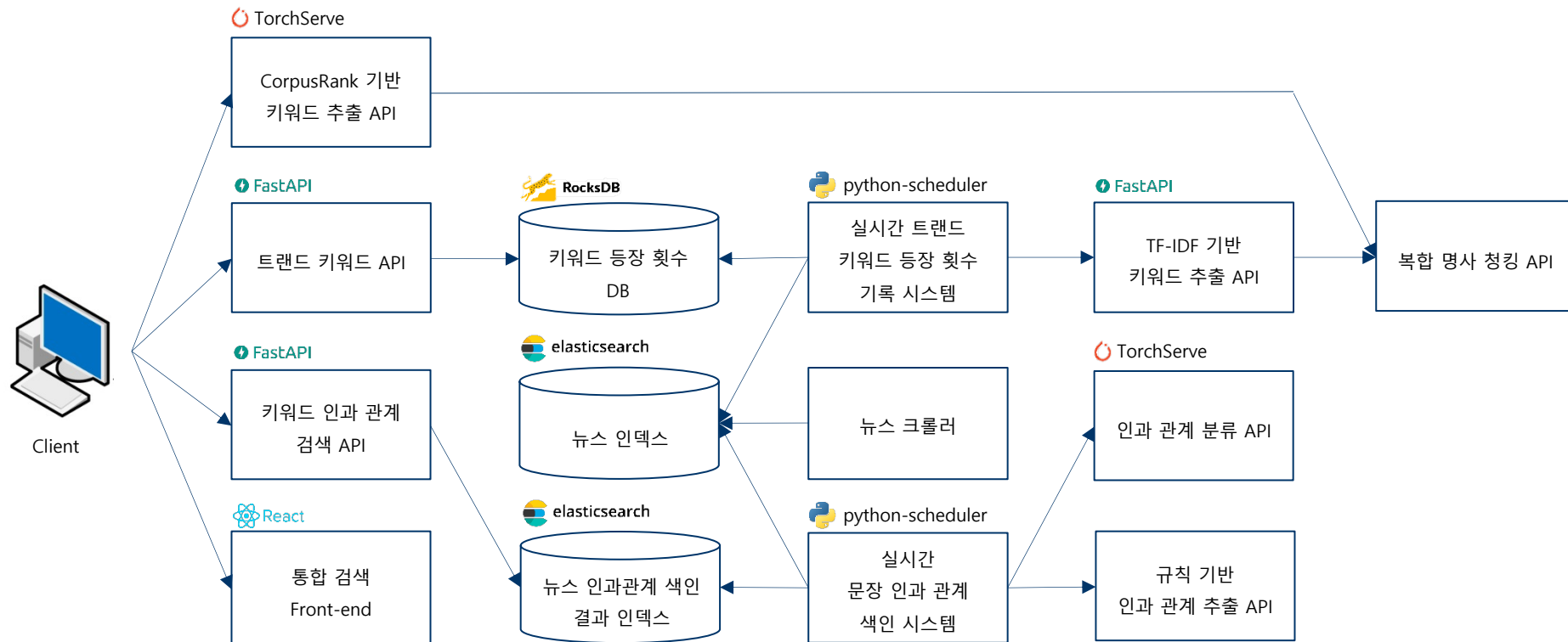
검색하기

정치	경제	사회	세계
1 이 대표	1 오피스 평균 주가수익률	1 이 대표	1
2 윤 대통령	2 상장 주식수	2 김 전 회장	2
3 전 의원	3 종매도제결합	3 김	3
4 김 의원	4 종매수제결합	4 대장동	4
5 UAE	5 1개월간	5 양방울	5
6 북한 무인기	6 체결강도	6 전철연	6
7 대통령	7 가려한희남	7 김성태	7
8 신원대적	8 1분전	8 풀	8
9 무인기	9 지난해	9 중국발 일국자	9
10 김기현 의원	10 CES 2023	10 시위	10

스포츠	IT과학	생활/문화	기타
1 2023.01	1 CES 2023	1 2023.01	1
2 sportschosun	2 CES	2 osen	2
3 중국성명	3 GPT	3 co	3
4 WBC	4 북한 무인기	4 클로리	4
5 osen	5 임유경	5 rumi	5
6 스포츠 기자	6 시난해	6 soul1014	6
7 MK	7 갤럭시S23	7 .kr	7
8 2022-2023 SKT	8 송예리 기자	8 뉴진스	8
9 예미닷	9 올국장	9 올국장	9
10 개로	10 오븐AI	10 육질	10

서비스 구조

• 서비스 구조



[1] 통합 검색 Front-end (Git, Docker) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/search-web.git

Contents

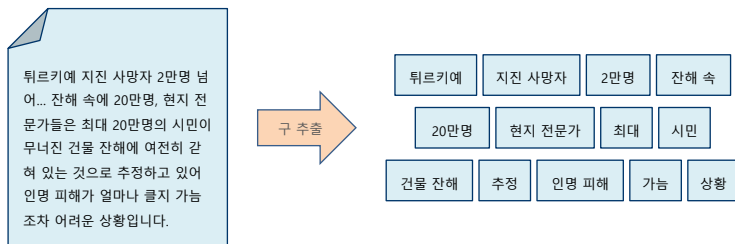
01. 서론	3
02. 비지도 문서 키워드 추출	9
a. 소개	10
b. 방법론	14
c. 환경 및 결과 분석	20
03. 문장 인과 관계 분류 및 추출	29
04. 트렌드 키워드 시스템	41

소개

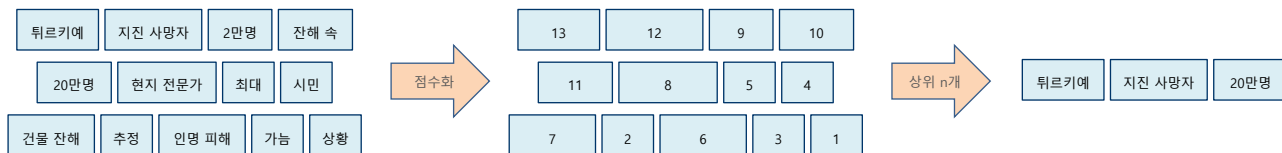
• 문제 정의 | 키워드 추출(Keyphrase Extraction)

- "문서의 중요한 정보를 나타낼 수 있는 단어 또는 구 집합을 문서에서 추출"[1]

- 문서 내 존재하는 모든 구(Phrase)들을 추출



- 모든 구들은 키워드 후보가 되며 후보들을 점수화 한 후, 상위 n개를 문서의 키워드로 지정

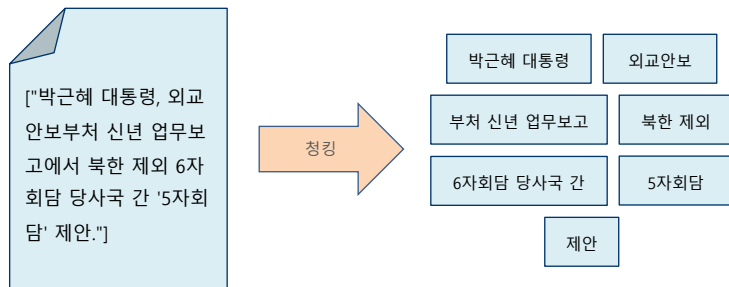


[1] Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.

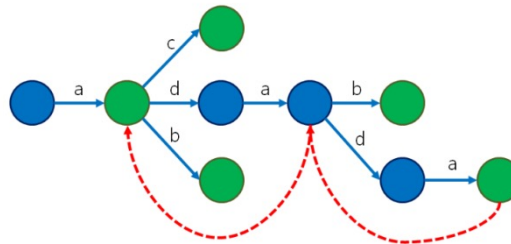
소개

• 문제 접근 | 복합 명사 청킹

- 입력된 문서를 문장 단위로 쪼갬 후, 문장들을 구 분할 사전에 기반하여 복합 명사 청킹
- 청킹 결과인 모든 문장의 각 구(청크)들은 모두 문서의 키워드 후보가 됨
- 청킹 입/출력 예시



- Aho-Corasik^[1]으로 구현된 구 분할 사전에서 매칭되는 모든 복합 명사를 찾음
- $W = \{a, ab, ac, adab, adada\}$



- Weighted Interval Scheduling 알고리즘으로 가중치(길이)가 가장 크고 커버리지가 가장 높은 샘플 추출

신년	맞이						
신년	맞이	고객					
		고객	맞춤형				
			맞춤형	다이어트	식단		
				다이어트	식단		
						단독	판매

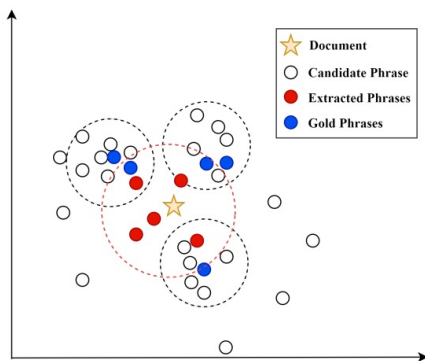
■ 가중치 높음
□ 가중치 낮음

[1] Aho, Alfred V., and Margaret J. Corasick. "Efficient string matching: an aid to bibliographic search." Communications of the ACM 18.6 (1975): 333-340.

소개

• 문제 접근 | UKERank[1]

- 등장 배경 및 개요



- 키워드 후보들을 임베딩 후 추상적 위치를 정점으로, 문서 전체를 임베딩 후 추상적 위치를 별로 표현하였을 때 2차원 좌표계에서 좌측 그래프와 같이 나타낼 수 있음
- 빨간 원 내부 정점들은 문서 전체와 유사한 키워드 후보 집합이나, Local context를 반영하였다고 볼 수 없음
- UKERank에서는 Global context와 Local context를 모두 반영할 수 있도록 각 키워드 후보마다 아래 두 개의 수치를 구한 후 하나의 점수로 합치는 방법론을 제시
 - Global Relevance(Phrase-Document Similarity) : 구가 문서의 Global context와 유사한 정도
 - Local Salience(Boundary-Aware Centrality) : 구가 Local context와 유사한 정도,
 (각 구들을 정점, 구들 사이의 유사도가 간선인 그래프를 생성한 후, 그래프 정보를 사용하여 Local Salience 계산)

- UKERank의 성능 증가 시도 접근 방법

- 키워드 후보의 임베딩 방식 수정
- Bert 외에 다른 임베딩 모델들을 사용
- MLM으로 Domain Adaptation을 하여 임베딩 모델의 뉴스 도메인 입력에 대한 성능 증가

[1] Liang, Xinnian, et al. "Unsupervised keyphrase extraction by jointly modeling local and global context." arXiv preprint arXiv:2109.07293 (2021).

소개

• 문제 접근 | CorpusRank^[1]

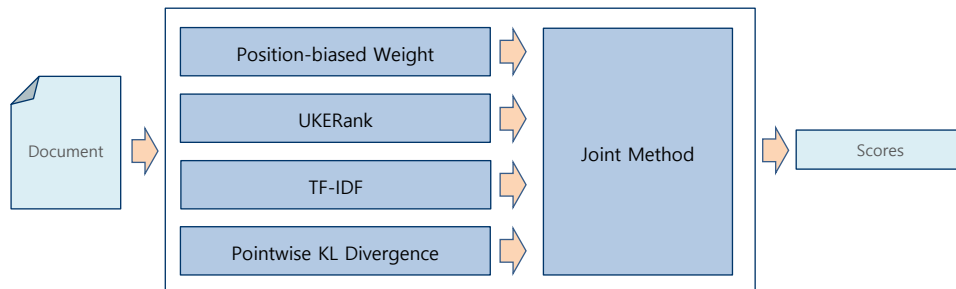
- 위치를 고려한 그래프 기반 추출 방식(UKERank) 외에 통계 기반 추출 방식(TF-IDF), n-gram 언어 모델 기반 추출 방식(Pointwise KL Divergence)과 조합하여 키워드 후보들의 점수를 계산
- CorpusRank 논문에서는 아래 모델들 또한 키워드 후보 점수 계산에 사용하였으나 기존 모델들과의 아이디어가 중복된다는 판단과 여러 모델들을 사용할 경우 전체 모델의 키워드 추출 속도가 매우 느려진다는 이유들로 본 실험에서는 사용하지 않음

- BERTRank : UKERank, PageRank^[2] 혼합 기반 그래프 기반 키워드 추출 모델
- K-document Centrality : UKERank, ExpandRank^[3] 혼합 기반 Local Saliency 방식 수정 모델
- Syntax Score 모델 : <전치사의 명사구>, <문장의 주어 명사구>, <문장의 목적어 명사구> 여부에 따라 차등 점수 부여 모델

- CorpusRank의 성능 증가 시도 접근 방법

- 조합 방식(Joint Method)에서 기존 논문 보다 더 많은 파라미터 조합
- 다른 조합 방식(Joint Method) 시도

- CorpusRank 모델 구조



[1] Williams, Isaiah, and Barry Cheung. "CorpusRank: Corpus Information in Unsupervised Keyphrase Extraction."

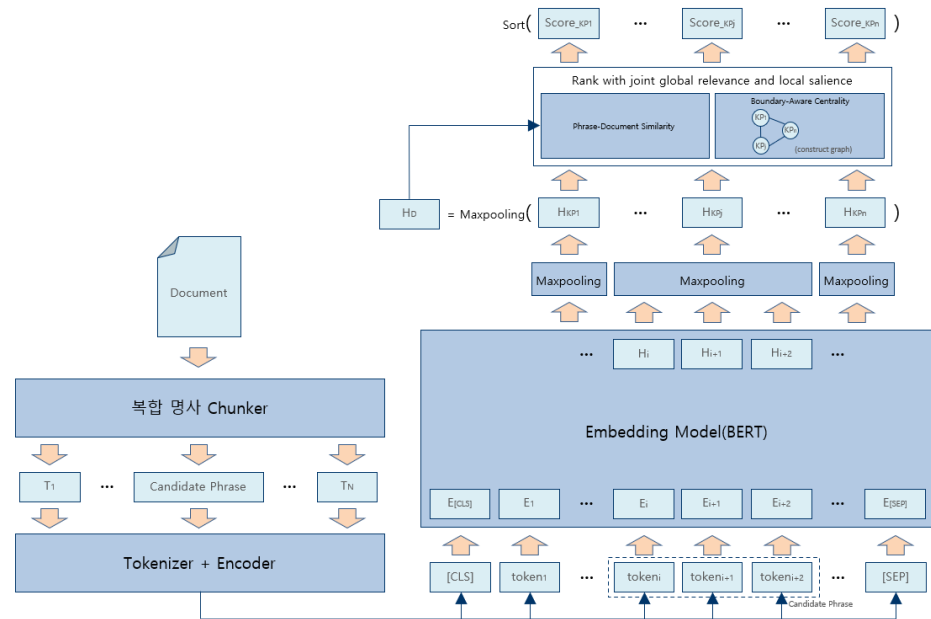
[2] Page, Lawrence, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[3] Wan, Xiaojun, and Jianguo Xiao. "Single document keyphrase extraction using neighborhood knowledge." AAAI. Vol. 8. 2008.

방법론

• UKERank | 모델 작동 원리 및 구조

- 문서를 입력받아 토큰화 및 청킹
 - $\{t_1, t_2, \dots, t_N\}, \{KP_1, KP_2, \dots, KP_n\} = \text{tokenizer_with_chunking}(\text{document})$
 - 청킹된 모든 구(Phrase)들은 키워드 후보
- 각 토큰을 문맥 고려 임베딩
 - $\{H_1, H_2, \dots, H_N\} = \text{BERT}(\{t_1, t_2, \dots, t_N\})$
- 문서와 키워드 후보들을 임베딩
 - $H_D = \text{Maxpooling}(\{H_1, H_2, \dots, H_N\})$
 - $H_{KP_i} = \text{Maxpooling}(\{H_j, \dots, H_{j+KP_size-1}\})$
 - H_{KP_i} 를 UKERank 논문에서는 내부 토큰 벡터들을 평균 내어 정의하였지만, Maxpooling을 시도하였을 때 성능이 더 높게 측정됨
- Rank 알고리즘으로 모든 키워드 후보들의 점수를 구함
 - $\{S_{KP_1}, S_{KP_2}, \dots, S_{KP_n}\} = \text{Rank}(\{H_{KP_1}, \dots, H_{KP_n}\}, H_D)$
- 키워드 후보들의 점수를 정렬하여 키워드 추출



방법론

- UKERank | Global Relevance

- Phrase-Document Similarity

- 맨해튼 거리(Manhattan distance, L1-distance)를 사용하여 구와 문서 전체 간의 유사도를 측정

- $R(H_{KP_i}) = 1 / \|H_D - H_{KP_i}\|_1$

- "구-문서 유사도 측정 방법으로 유클리드 거리, 코사인 유사도, 맨해튼 거리 세 가지를 시도하여 전체 모델의 성능을 측정하였을 때, 맨해튼 거리 사용이 제일 좋은 결과를 얻음"[1]

Similarity Measure	DUC2001			Inspec			SemEval2010		
	F1@5	F1@10	F1@15	F1@5	F1@10	F1@15	F1@5	F1@10	F1@15
Euclidean Distance	23.31	28.04	30.39	28.5	37.01	29.25	10.99	16.37	18.41
Cosine similarity	15.01	17.96	19.44	23.67	30.26	33.35	9.70	12.22	13.29
Manhattan Distance	28.62	35.52	36.29	32.49	40.04	41.05	12.26	19.22	21.42

Table 3: The results of different measure methods for similarity between candidate phrase and the whole document.

[1] Liang, Xinnian, et al. "Unsupervised keyphrase extraction by jointly modeling local and global context." arXiv preprint arXiv:2109.07293 (2021).

방법론

• UKERank | Local Saliency

- Phrase-Phrase Similarity^[1]

- "Phrase-Phrase Similarity으로 코사인 유사도(cosine similarity)보다 내적(dot-product)을 사용하였을 때 모델 성능이 더 높게 측정됨"^[1]

$$- e_{ij} = H_{KP_i}^T \cdot H_{KP_j}$$

- Position-biased Weight^[2]

- "긴 문서나 뉴스 기사들은 작가/저자가 중요 정보를 문서의 앞에 위치시키는 경향이 있음"^[3]

$$- \hat{p}(KP_i) = \frac{\exp(p(KP_i))}{\sum_{k=1}^n \exp(p(KP_k))}, \quad (p(KP_i) = \frac{1}{p_1}, \quad p_1 = \text{키워드 후보 } KP_i \text{가 처음으로 등장한 위치})$$

- Boundary-Aware Centrality

- "일반적으로 문서의 주요 내용은 문서의 시작 혹은 뒷 부분에 존재"^[4]

$$- \text{boundary function } d_b(i) = \min(\alpha \cdot i, n - i)$$

- 정점 i와 j에 대해서 $d_b(i) < d_b(j)$ 이면, 노드 i는 노드 j보다 문서 경계에 더 가까이 있음을 의미

$$- C(H_{KP_i}) = \sum_{d_b(i) < d_b(j)} \max(e_{ij} - \theta, 0) + \lambda \sum_{d_b(i) \geq d_b(j)} \max(e_{ij} - \theta, 0), \quad (\theta = \beta(\max(e_{ij}) - \min(e_{ij})) + \min(e_{ij}))$$

$$- \hat{C}(H_{KP_i}) = \hat{p}(KP_i) C(H_{KP_i})$$

[1] Liang, Xinnian, et al. "Unsupervised keyphrase extraction by jointly modeling local and global context." arXiv preprint arXiv:2109.07293 (2021).

[2] Sun, Yi, et al. "SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model." IEEE Access 8 (2020): 10896-10906.

[3] Florescu, Corina, and Cornelia Caragea. "A position-biased pagerank algorithm for keyphrase extraction." Thirty-first AAAI conference on artificial intelligence. 2017.

[4] Lin, Chin-Yew, and Eduard Hovy. "Identifying topics by position." Fifth Conference on Applied Natural Language Processing. 1997.

방법론

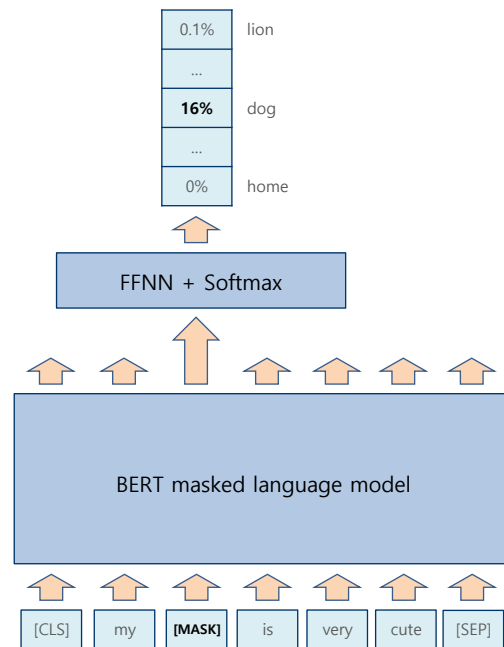
• UKERank | Masked Language Modeling(MLM)

- BERT_[1] 에서 사용된 학습 방법 중 하나
- 입력 텍스트의 일부 토큰을 무작위로 마스킹한 후, 주변 문맥을 바탕으로 마스킹된 토큰을 예측
- 뉴스 관련 데이터셋으로 학습을 진행하여, 뉴스 도메인에 대하여 임베딩 모델의 성능 증가 (Domain Adaptation)

• UKERank | Rank with Global and Local Information

- Joint global context and local context
 - 구의 Phrase-Document Similarity와 Boundary-Aware Centrality를 단순히 곱하여 구의 최종 점수를 구함
 - $S_{UKE}(H_{KP_i}) = R(H_{KP_i}) \cdot \hat{C}(H_{KP_i})$
 - 구들의 최종 점수를 정렬한 후, 상위 n개의 구들을 문서의 키워드들로 지정

- MLM 예시



[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

방법론

- 쿨백-라이블러 발산(KLD, Kullback-Leibler Divergence, Relative entropy)

- 이상적인 분포에 대해, 그 분포에 근사하는 다른 분포를 사용하여 샘플링을 한다면 발생할 수 있는 정보 엔트로피의 차이

- 두 확률분포의 차이를 계산하는 데 사용할 수 있음

- $D_{KL}(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ (p, q = 확률분포 P, Q 의 확률 밀도 함수)

- Pointwise KL Divergence[1]

- N-gram phrase($w = w_1 w_2 \dots w_n$)가 등장할 확률 : $P(w) = \prod_{i=1}^n P(w_i | w_1 w_2 \dots w_{i-1})$ (각 단어가 독립적이라고 가정)

- Pointwise KL Divergence $\delta_x(p||q) = p(x) \log \frac{p(x)}{q(x)}$

- Phraseness of w

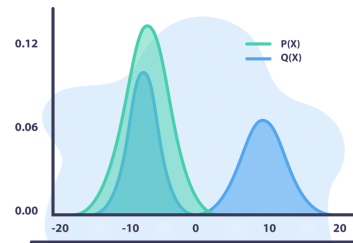
- unigram 모델에서 각 단어가 독립적으로 가정함으로써 잃어버린 정보량 = $\delta_w(LM_{fg}^N \parallel LM_{fg}^1)$

- Informativeness of w

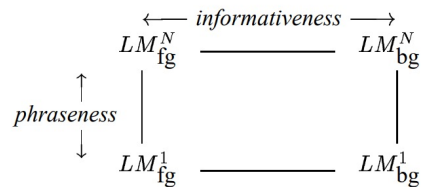
- foreground 모델 대신 background 모델로 샘플링을 함으로써 잃어버린 정보량 = $\delta_w(LM_{fg}^1 \parallel LM_{bg}^1)$

- $S_{ki}(KP_i = w) = (\text{Phraseness of } w) + (\text{Informativeness of } w) = \delta_w(LM_{fg}^N \parallel LM_{fg}^1) + \delta_w(LM_{fg}^1 \parallel LM_{bg}^1)$

- 원본 확률 분포 P 와 근사 확률 분포 Q 예시



- Phraseness와 Informativeness 비교



1 = unigram 모델, N = n-gram 모델
fg = foreground corpus, bg = background corpus

[1] Tomokiyo, Takashi, and Matthew Hurst. "A language model approach to keyphrase extraction." Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment. 2003.

방법론

- TF-IDF(Term Frequency - Inverse Document Frequency)

- 여러 문서로 이루어진 문서군이 있을 때, 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치

- $idf(w, D) = \log\left(\frac{N}{df(w)}\right) + 1$, $tfidf(w, d, D) = tf(w, d) \cdot idf(w, D)$

- (N = 총 문서의 수, $df(w)$ = 단어 w 가 등장한 문서의 수, $tf(w, d)$ = 문서 d 에서 w 의 등장 횟수)

- N-gram phrase($w = w_1 w_2 \dots w_n$)의 키워드 후보가 주어졌을 때,

- $S_{tf(word)}(KP_i = w) = average(tfidf(w_1, D), tfidf(w_2, D), \dots, tfidf(w_n, D))$

- $S_{tf(phrase)}(KP_i = w) = tfidf(w, D)$

- $S_{tf}(KP_i) = S_{tf(phrase)}(KP_i)$ (정량평가 결과, $S_{tf(phrase)}$ 가 $S_{tf(word)}$ 보다 더 우수한 성능을 보임)

- CorpusRank

- Combined Model

- 위치-편향 가중치, KLD, TF-IDF, UKERank의 결과 값들을 파라미터들로 하나의 점수로 변환

- $T_{mul}(H_{KP_i}) = \hat{p}(KP_i) \cdot softmax(S_{kl}(KP_i))^a \cdot softmax(S_{tf}(KP_i))^b \cdot softmax(S_{UKE(\alpha, \beta, \lambda)}(KP_i))^c$

- $T_{sum}(H_{KP_i}) = \hat{p}(KP_i) + a \cdot softmax(S_{kl}(KP_i)) + b \cdot softmax(S_{tf}(KP_i)) + c \cdot softmax(S_{UKE(\alpha, \beta, \lambda)}(KP_i))$

- $a, b, c, \alpha, \beta, \gamma$ = hyper-parameters

환경 및 결과 분석

• 데이터셋

- 한국어 뉴스 말뭉치^[1]
 - 크롤링 된 10만 개의 뉴스 기사 본문들로 구성됨
- 한국어 뉴스 키워드^[2] 학습용 데이터셋
 - 뉴스 기사에서 해시태그 된 복합 명사를 본문의 키워드로 가정



- 크롤링 된 10000개의 뉴스 기사 본문과 키워드들로 구성됨
- 한국어 뉴스 키워드 평가용 데이터셋
 - 학습용 데이터셋과 동일한 방식으로 제작
 - 19239개의 뉴스 기사 본문과 키워드들로 구성됨
 - 본문 : 최소 1개, 최대 411개, 평균 24.13개의 문장
 - 키워드 : 최소 5개, 최대 25개, 평균 8.35개

[1] 한국어 뉴스 말뭉치 전처리 과정 : <http://galadriel:8090/x/wUuFBQ>

[2] 한국어 뉴스 키워드 데이터셋 전처리 과정 : <http://galadriel:8090/x/jNSeBQ>

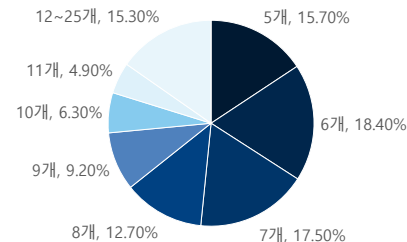
- 한국어 뉴스 말뭉치 예시

document_id	donga-111052329-1
content	"영화 '스파이더맨: 노 웨이 홈'(감독 존 왓츠)이 누적관객수 600만명 돌파를 앞두고 있다. 2일 영화관입장권통합전산망에 따르면 스파이더맨: 노 웨이 홈은 전날 28만9169명을 모았다. 누적관객수는 584만8941명이...(생략)"

- 한국어 뉴스 키워드 평가용 데이터셋 예시

source	중앙일보
content	"文, 현충원 참배로 2022년 첫일정... 선도국가 길, 멈추지 않겠다. 문재인 대통령이 2022년 새해 첫 일정을 서울 동작동 국립서울현충원 참배로 시작했다. 문 대통령은 1일 오전 8시 김부겸 국무총리와 홍남기 경제부총리 겸 기획재정부 장관, 유은혜 사회부총리 겸 교육부 장관, 전해철 행정안전부 장관을 비롯한 국무위원, 유영민 대통령 비서실장 등과 함께 현충...(생략)"
keywords	["선도국가", "현충원", "현충원 참배", "행정안전부 장관", "기획재정부 장관", "문재인", "참배"]

- 한국어 뉴스 키워드 평가용 데이터셋 키워드 개수 분포



환경 및 결과 분석

- UKERank(S_{UKE})

- 아래에 대한 모든 조합에 대하여 "한국어 뉴스 키워드 학습용 데이터셋"으로 최고 성능(F1@10)의 파라미터 조합 찾음
 - Embedding Model = { bert-base-multilingual-uncased, bert-kor-base^[1], kocharelectra-base-discriminator^[2], nc/nlu_electra }
 - $\alpha = \{ 0.1, 0.2, 0.5, 0.7, 1.0, 1.2, 1.5 \}$
 - $\beta = \{ 0.0, 0.1, 0.2, 0.3 \}$
 - $\lambda = \{ 0.2, 0.4, 0.6, 0.8, 1.0 \}$

- CorpusRank(T_{mul})

- 아래에 대한 모든 조합에 대하여 "한국어 뉴스 키워드 학습용 데이터셋"으로 최고 성능(F1@10)의 파라미터 조합 찾음
 - Embedding Model = { bert-base-multilingual-uncased, bert-kor-base^[1] }
 - $a = \{ 0.00, 0.25, 0.50, 0.75, 1.00 \}$
 - $b = \{ 0.00, 0.25, 0.50, 0.75, 1.00 \}$
 - $c = \{ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 \}$
 - $\alpha = \{ 0.0, 0.1, 0.2, 0.3 \}$
 - $\beta = \{ 0.0, 0.1, 0.2, 0.3 \}$
 - $\lambda = \{ 0.2, 0.4, 0.6, 0.8, 1.0 \}$

[1] <https://huggingface.co/kykim/bert-kor-base>

[2] <https://huggingface.co/monologg/kocharelectra-base-discriminator>

[3] 비지도 문서 키워드 추출 실험 : <http://galadriel:8090/x/TMueBQ>

[4] 비지도 문서 키워드 추출 실험 (Git) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/uke.git

환경 및 결과 분석

- UKERank Masked Language Modeling(MLM)

- "한국어 뉴스 말뭉치"로 "bert-base-multilingual-uncased", "kykim/bert-kor-base"을 fine-tuning
- [CLS], [SEP]을 제외한 15%의 입력 토큰에 대하여 [MASK]로 대체
- [MASK] 토큰 맞춤 여부의 accuracy 측정

- 정량 평가 방법

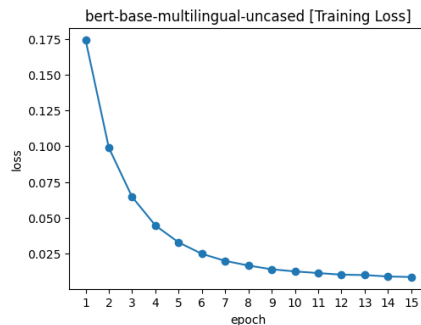
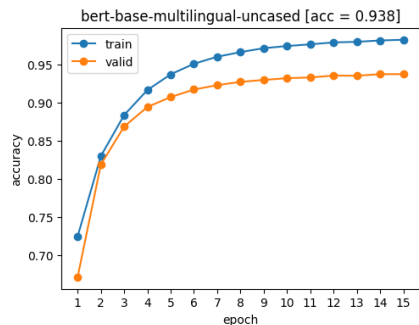
- 모델의 출력 결과의 키워드가 데이터셋의 키워드와 어간 추출(Stemming)하였을 때 동일하면 정답으로 설정
- 모든 모델들이 5개, 10개, 15개의 키워드를 추출하였을 때 각각을 "한국어 뉴스 키워드 평가용 데이터셋"으로 f1-score, precision, recall 측정
- 모델을 torchserve로 GPU서버에 배포 후, REST API의 호출-응답 사이의 평균 시간을 측정

[1] 비지도 문서 키워드 추출 MLM 실험 (Git) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/uke-mlm-bert.git

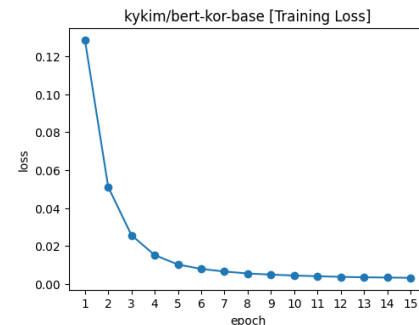
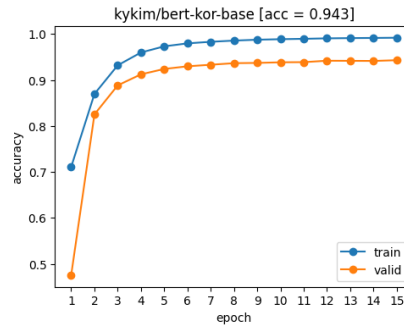
환경 및 결과 분석

- UKERank Masked Language Modeling(MLM)

- bert-base-multilingual-uncased



- kykim/bert-kor-base



- UKERank, CorpusRank 파라미터 설정

model-name	F1@5	F1@10	F1@15	params
CorpusRank(mul)	0.3206	0.3018	0.2697	[bert-kor-base, 0.00, 0.25, 15.0, 0.1, 0.2, 0.2, without-mlm]
CorpusRank(sum)	0.2828	0.2699	0.2434	[bert-kor-base, 0.00, 1.7, 0.00, 0.0, 0.2, 0.2, without-mlm]
UKERank	0.2249	0.2296	0.2172	[bert-kor-base, 0.1, 0.0, 0.2, without-mlm]

(정량 평가 결과, MLM은 오히려 성능 하락)

환경 및 결과 분석

• 정량 평가 결과

model-name	P@5	R@5	F@5	P@10	R@10	F@10	P@15	R@15	F@15	time(sec/doc)
ORACLE	0.8994	0.5887	0.6943	0.5790	0.7092	0.6187	0.3966	0.7189	0.4961	-
LEAD	0.2314	0.1550	0.1811	0.1805	0.2387	0.2002	0.1505	0.2950	0.1943	-
TextRank	0.0802	0.0532	0.0624	0.0834	0.1105	0.0926	0.0816	0.1618	0.1058	
SingleRank	0.0911	0.0607	0.0712	0.0908	0.1201	0.1008	0.0881	0.1745	0.1142	
PositionRank	0.1128	0.0752	0.0881	0.1110	0.1478	0.1235	0.1078	0.2149	0.1402	
TopicRank	0.2334	0.1519	0.1794	0.1762	0.2273	0.1930	0.1445	0.2778	0.1851	
MultipartiteRank	0.2924	0.1894	0.2240	0.2154	0.2765	0.2354	0.1737	0.3319	0.2220	
CorpusRank	0.3317	0.2170	0.2557	0.2402	0.3108	0.2635	0.1904	0.3669	0.2441	0.3066
TF-IDF	0.3203	0.2076	0.2455	0.2359	0.3021	0.2575	0.1879	0.3574	0.2396	0.1791
UKERank	0.2272	0.1521	0.1779	0.1809	0.2391	0.2005	0.1541	0.3022	0.1990	0.3150
pointwise-KLD	0.0274	0.0188	0.0219	0.0309	0.0419	0.0347	0.0327	0.0661	0.0427	0.1809

- Baseline models

- ORACLE : 복합 명사 청킹을 사용하였을 경우 가능한 최대 성능
- LEAD : 상단 n개의 후보 구를 키워드로 추출하는 모델
- TextRank, SingleRank, PositionRank, TopicRank, MultipartiteRank : 그래프 기반 키워드 추출 모델

환경 및 결과 분석

• 정성 평가 | UKERank

Input	첫 여성	법무부 인권국장	...	민변	출신	위은진	변호사	지난해	8월	이후	공석	이었던	법무부 인권국장	에	민주사회	를	위한	변호사 모임	(민변)	출신	위	은진(50)	사법연수원	31기									
)	변호사	가	임용	됐다.	여성	이	인권국장	에	임명	된	건	처음	이다.	세 번째	비	(非)	검사 출신	인권국장	이다.	법무부	는	3일	자	로	위 변호사	를	인권국장	에	임용	한다고	2일					
	밝혔다.	위	신임	국장은	민변 출신	황희석	열린민주당	최고위원	과	이상갑	현 법무부	법무실장	에	이어	세 번째	비검사 출신	인권	국장	이	된다.	문재인 정부	의	'	탈	검찰화	'	기조	에	따라								
	법무부	는	2006년	7월	신설	이후	검사	만	보임	했던	인권국장 자리	를	2017년	부터	비	검사 출신	일반직	공무원	이나	전문가	에게	개방	해	경력경쟁	채용	으로	뽑았다.	위	국장	은							
	이화여대	통계학과	를	졸업	하고	1999년	제41회	사법시험	에	합격	한	후	약	20년간	변호사	로	일	해	왔다.	변호사 시절	여성 폭력	방지	및	피해자	지원	,	이주외국인	,	다문화가족								
	인권	보호	,	시민	인권	침해	구제	활동	등	인권	변호	활동	를	했다.	최근에는	정의기억연대	(정의연)	옛	한국정신대문제대책협의회)	후원금	유용	혐의	등으로	무소속	윤미향	의원	과	함께	재판					
	에	넘겨진	정의연	이사 A	씨를	변호	하다가	지난해	10월	말	인권국장	채용	절차	중에	사임	했다.	국가인권위원회	외국인	인권전문	위원회	전문위원	,	대한변호사협회	인권위원회	부위원장	,											
	민변	여성인권위원장	,	경찰청	인권침해	사건	진상조사위원회	위원	등을	지냈다.	이번에	여성	첫	법무부	인권국장	이	됐다.	인권국장	은	일반직	고위공무원	나	등급	의	직책	으로	정부	의									
	인권	정책	를	총괄	한다.	법무부	는	"	다양	한	현장	활동	를	통해	쌓아	온	풍부한	경험	과	전문성	을	바탕	으로	인권	친화적	법	집행	과	제도	정착	에	역량	을	집중	해	국민	의
	실질적	인	인권보장	수준	향상	에	크게	기여	할	것으로	기대	한다"고	밝혔다.																								
Gold	인권국장	법무부	법무부 인권국장	인권국장	채용	위은진	민변	문재인	윤미향	정의연																											
Output	첫 여성	법무부 인권국장	위은진	민변	황희석	열린민주당	최고위원	민변 출신	변호사 모임	사법연수원	인권	국장	현 법무부																								

Input	인도	의	새해	...	황금사원	에	몰린	수천	명	인파	지난	1일	(현지	시각)	인도	편자브주	암리차르	에	있는	황금	사원	앞	이	새해	첫	기도	를	하러	몰려든	시크교	신자	들로	인산인해	
	를	이루고	있다.	시크교	는	15세기	품	힌두교	에서	갈라져	나온	종교	로,	14억	인도	인구	중	2%	가	시크교	신자	인	것으로	알려져	있다.	이곳	황금사원	은	시크교	의	성전					
	(聖殿)	이다.																																		
Gold	새해	성전	시크교	인도	황금사원																															
Output	편자브주	암리차르	인도	황금사원	수천	명	시크교	신자	인파	새해	현지	시각	시크교	성전																						




gold
output
gold n output

[1] 비지도 문서 키워드 추출 정성평가 : <http://galadriel:8090/x/z6ieBQ>

환경 및 결과 분석

- 정성 평가 | TF-IDF

Input	<p>첫 여성 법무부 인권국장 ... 민변 출신 위은진 변호사 지난해 8월 이후 공석 이었던 법무부 인권국장 에 민주사회 를 위한 변호사 모임 (민변) 출신 위 은진(50 · 사법연수원 31기) 변호사 가 임용 됐다. 여성 이 인권국장 에 임명 된 건 처음 이다. 세 번째 비 (非) 검사 출신 인권국장 이다. 법무부 는 3일 자 로 위 변호사 를 인권국장 에 임용 한다고 2일 밝혔다. 위 신임 국장은 민변 출신 황희석 열린민주당 최고위원 과 이상갑 현 법무부 법무실장 에 이어 세 번째 비검사 출신 인권 국장 이 된다. 문재인 정부 의 ' 탈 검찰화 ' 기조 에 따라 법무부 는 2006년 7월 신설 이후 검사 만 보임 했던 인권국장 자리 를 2017년 부터 비 검사 출신 일반직 공무원 이나 전문가 에게 개방 해 경력경쟁 채용 으로 뽑았다. 위 국장 은 이화여대 통계학과 를 졸업 하고 1999년 제41회 사법시험 에 합격 한 후 약 20년간 변호사 로 일 해 왔다. 변호사 시절 여성 폭력 방지 및 피해자 지원 , 이주외국인 · 다문화가족 인권 보호 , 시민 인권 침해 구제 활동 등 인권 변호 활동 을 했다. 최근에는 정의기억연대 (정의연 · 옛 한국정신대문제대책협의회) 후원금 유용 혐의 등으로 무소속 윤미향 의원 과 함께 재판 에 넘겨진 정의연 이사 A 씨를 변호 하다가 지난해 10월 말 인권국장 채용 절차 중에 사임 했다. 국가인권위원회 외국인 인권전문 위원회 전문위원 , 대한변호사협회 인권위원회 부위원장 , 민변 여성인권위원장 , 경찰청 인권침해 사건 진상조사위원회 위원 등을 지냈다. 이번엔 여성 첫 법무부 인권국장 이 됐다. 인권국장 은 일반직 고위공무원 나 등급 의 직책 으로 정부 의 인권 정책 을 총괄 한다. 법무부 는 " 다양 한 현장 활동 을 통해 쌓아 온 풍부한 경험 과 전문성 을 바탕 으로 인권 친화적 법 집행 과 제도 정착 에 역량 을 집중 해 국민 의 실질적 인 인권보장 수준 향상 에 크게 기여 할 것으로 기대 한다"고 밝혔다.</p>
Gold	<p>인권국장 법무부 법무부 인권국장 인권국장 채용 위은진 민변 문재인 윤미향 정의연</p>
Output	<p>인권국장 민변 법무부 인권국장 정의연 법무부 임용 위은진 인권 국장 탈 검찰화 비 검사 출신</p>
Input	<p>인도 의 새해 ... 황금사원 에 물린 수천 명 인파 지난 1일 (현지 시각) 인도 판자브주 암리차르 에 있는 황금 사원 앞 이 새해 첫 기도를 하려 몰려든 시크교 신자 들로 인산인해를 이루고 있다. 시크교 는 15세기 쯤 힌두교 에서 갈라져 나온 종교 로, 14억 인도 인구 중 2% 가 시크교 신자 인 것으로 알려져 있다. 이곳 황금사원 은 시크교 의 성전 (聖殿)이다.</p>
Gold	<p>새해 성전 시크교 인도 황금사원</p>
Output	<p>황금사원 시크교 신자 시크교 첫 기도 14억 인도 인구 사원 앞 성전 판자브주 암리차르 인산인해 15세기</p>

-  gold
-  output
-  gold n output

[1] 비지도 문서 키워드 추출 정성평가 : <http://galadriel:8090/x/z6ieBQ>

환경 및 결과 분석

• 정성 평가 | CorpusRank

Input	첫 여성	법무부 인권국장	...	민변	출신	위은진	변호사	지난해	8월	이후	공석	이었던	법무부 인권국장	에	민주사회	를	위한	변호사 모임	(민변)	출신	위	은진(50)	사법연수원	31기									
)	변호사	가	임용	됐다.	여성	이	인권국장	에	임명	된	건	처음	이다.	세 번째	비	(非)	검사 출신	인권국장	이다.	법무부	는	3일	자	로	위	변호사	를	인권국장	에	임용	한다고	2일				
	밝혔다.	위	신임	국장은	민변 출신	황희석	열린민주당	최고위원	과	이상갑	현	법무부	법무실장	에	이어	세 번째	비검사 출신	인권	국장	이	된다.	문재인 정부	의	'	탈	검찰화	'	기조	에	따라							
	법무부	는	2006년	7월	신설	이후	검사	만	보임	했던	인권국장 자리	를	2017년	부터	비	검사 출신	일반직	공무원	이나	전문가	에게	개방	해	경력경쟁	채용	으로	뽑았다.	위	국장	은							
	이화여대	통계학과	를	졸업	하고	1999년	제41회	사법시험	에	합격	한	후	약	20년간	변호사	로	일	해	왔다.	변호사 시절	여성	폭력	방지	및	피해자	지원	,	이주외국인		다문화가족							
	인권	보호	,	시민	인권	침해	구제	활동	등	인권	변호	활동	를	했다.	최근에는	정의기억연대	(정의연)	옛	한국정신대문제대책협의회)	후원금	유용	혐의	등으로	무소속	윤미향	의원	과	함께	재판					
	에	넘겨진	정의연	이사	A	씨를	변호	하다가	지난해	10월	말	인권국장	채용	절차	중에	사임	했다.	국가인권위원회	외국인	인권전문	위원회	전문위원	,	대한변호사협회	인권위원회	부위원장	,										
	민변	여성인권위원장	,	경찰청	인권침해	사건	진상조사위원회	위원	등을	지냈다.	이번에	여성	첫	법무부	인권국장	이	됐다.	인권국장	은	일반직	고위공무원	나	등급	의	직책	으로	정부	의									
	인권	정책	를	총괄	한다.	법무부	는	"	다양	한	현장	활동	를	통해	쌓아	온	풍부한	경험	과	전문성	을	바탕	으로	인권	친화적	법	집행	과	제도	정착	에	역량	을	집중	해	국민	의
	실질적	인	인권보장	수준	향상	에	크게	기여	할	것으로	기대	한다"고	밝혔다.																								
Gold	인권국장	법무부	법무부 인권국장	인권국장 채용	위은진	민변	문재인	윤미향	정의연																												
Output	인권국장	법무부 인권국장	민변	첫 여성	위은진	법무부	황희석	열린민주당	최고위원	임용	민변 출신	인권 국장																									

Input	인도	의	새해	...	황금사원	에	몰린	수천 명	인파	지난	1일	(현지 시각)	인도	편자브주	암리차르	에	있는	황금	사원 앞	이	새해	첫	기도	를	하러	몰려든	시크교 신자	들로	인산인해			
	를	이루고	있다.	시크교	는	15세기	품	힌두교	에서	갈라져	나온	종교	로,	14억	인도	인구	중	2%	가	시크교 신자	인	것으로	알려져	있다.	이곳	황금사원	은	시크교	의	성전				
	(聖殿)이다.																																	
Gold	새해	성전	시크교	인도	황금사원																													
Output	황금사원	시크교 신자	편자브주	암리차르	시크교	인도	수천 명	인파	첫 기도	14억 인도 인구	성전																							

gold
output
gold n output

[1] 비지도 문서 키워드 추출 정성평가 : <http://galadriel:8090/x/z6ieBQ>

환경 및 결과 분석

- 정성 평가 오류 분석

- 한국어 뉴스 키워드 데이터셋으로 CorpusRank의 파라미터를 찾은 결과, $a = 0$ (CorpusRank 결과 값에 KLD는 영향을 미치지 않음)
- CorpusRank 결과 값에는 TF-IDF의 결과 값이 가장 크게 반영되어 키워드 후보의 최종 점수가 결정 됨
- "gold"와 "output"이 유사하나 정답으로 인정되지 않은 키워드들이 존재
- 신조어와 같은 복합 명사 청킹 사전에 없는 단어들은 키워드 후보 군에서 제외될 수 있음
- UKERank와 TF-IDF 모델의 결과 값들을 서로 보완하여 CorpusRank의 결과 값으로 추출되는 것을 볼 수 있음
- CorpusRank가 TF-IDF 보다 미묘한 차이로 f1-score가 높지만, 결과 추출 속도가 0.6배 더 느림

Contents

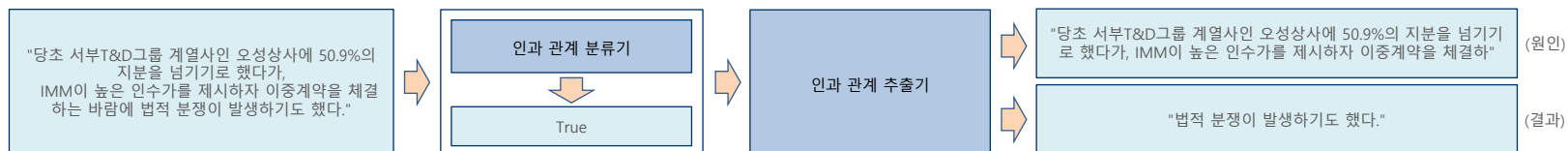
01. 서론	3
02. 비지도 문서 키워드 추출	9
03. 문장 인과 관계 분류 및 추출	29
a. 소개	30
b. 방법론	33
c. 환경 및 결과 분석	35
d. 키워드 인과 관계 검색 서비스	40
04. 트랜드 키워드 시스템	41

소개

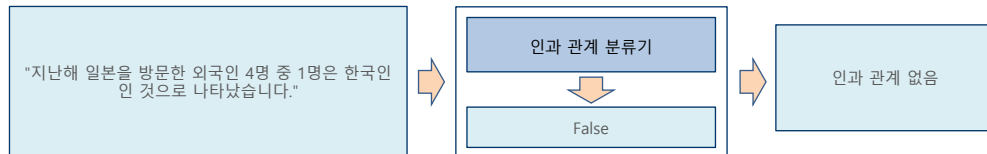
• 문제 정의 | 인과 관계 분류 및 추출

- 인과 관계(Causality) : "텍스트에 표현된 두 객체나 이벤트 간의 원인 결과 관계 (객체 e2가 다른 객체 e1의 결과로써 성립하는 관계)"^[1]
- 문장 인과 관계 분류 : 문장에서 인과 관계가 있는지 여부를 확인
- 문장 인과 관계의 원인과 결과 추출 : 문장에서 인과 관계의 원인과 결과 부분을 추출

- 인과 관계 분류 결과가 "True"일 때, 인과 관계의 원인과 결과 추출



- 인과 관계 분류 결과가 "False"일 때,



[1] 내러티브팀 이호창님 Causality Analysis 발표자료, 2021

소개

• 문장 인과 관계 분류 관련 연구

- FinCausal 2020^[1] Task-1

- 금융 관련 문장이 주어졌을 때, 인과 관계의 여부를 분류하는 Task
- 데이터셋(FinCausal Corpus)이 제공되었으며, 여러 팀들이 참여하고 공통된 기준으로 각 팀들의 모델을 평가

- 1위~6위 팀 모두 fine-tuning 한 Transformer 계열 분류 모델을 사용
- LIORI^[2](1위), UPB^[3](2위), FiNLP^[4](4위) 팀은 Ensemble 전략을 사용
- 상위 팀들이 사용한 Transformer 계열 모델들로는 "BERT-base", "BERT-large", "RoBERTa", "FinBERT"가 있음

- FinCausal 2020 Task1 결과

Rank	Team-name	F1 Score
1	LIORI	97.75
2	UPB	97.55
3	ProsperAMnet	97.23
4	FiNLP	96.99
5	DOMINO	96.12
...
	baseline	95.23

• 문제 접근 | 문장 인과 관계 분류

- FinCausal 2020 Task-1 상위 팀들의 전략을 사용하여 문제를 접근하여, 영어와 한국어 문장의 인과 관계 여부를 분류

[1] Mariko, Dominique, et al. "Financial document causality detection shared task (fincausal 2020)." arXiv preprint arXiv:2012.02505 (2020).

[2] Gordeev, Denis, et al. "LIORI at the FinCausal 2020 Shared task." Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020.

[3] Ionescu, Marius, et al. "Upb at fincausal-2020, tasks 1 & 2: Causality analysis in financial documents using pretrained language models." Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020.

[4] Gupta, Sarthak. "FiNLP at FinCausal 2020 Task 1: Mixture of BERTs for Causal Sentence Identification in Financial Texts." Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020.

소개

• 문제 접근 | 규칙 기반 인과 관계 추출

- 문장의 원인과 결과를 규칙 기반으로 추출
- 규칙은 여러 인과 관계 연구^{[1], [2], [3], [4]}들에서 밝혀진 패턴들을 사용
- 규칙 패턴 예시

규칙	예시문장	원인	결과
~는 탓에	국내외에서 신종 코로나바이러스 감염증(코로나19) 확산속도가 다시 가팔라지 는 탓에 주요 시장에서 이동통신사들의 오프라인 매장 내방객들이 크게 줄어든 것으로 알려졌다.	코로나 감염증 확산속도 가팔라 짐	오프라인 매장 내방객 줄어듦
~ 함으로써	신한금융은 독립·전문 벤처캐피탈 회사를 인수 함으로써 기업 생애주기 전체에 대응하는 투자금융 밸류체인을 완성하게 됐다.	신한금융, 벤처캐피탈 회사 인수	투자금융 밸류 체인 완성
~ 이후 ~ [동사]	종합부동산세와 양도소득세를 대폭 강화하기로 한 정부의 7·10대책 이후 서울 강남에서 아파트 증여가 폭발적으로 늘고 있다 .	7.10 대책	강남 아파트 증여
~원인은 ~ 때문이다[덕분이다]	지난달 저물가의 원인은 우선 사회적 거리 두기 운동으로 외식, 여행 등 서비스 업종 수요가 크게 줄었기 때문이다 .	사회적거리두기운동, 업종 수요 줄어듦	저물가
...

[1] 장두성, and 최기선. "단서 구문과 어휘 쌍 확률을 이용한 인과관계 추출." 한국정보과학회 언어공학연구회 학술발표 논문집 (2003): 163-169.

[2] 최상진, and 임채훈. "인과관계 형성의 인지과정과 연결어미의 상관성: '-아서', '-니까', '-면' 등을 중심으로." 국어학 (國語學) 52 (2008): 127-152.

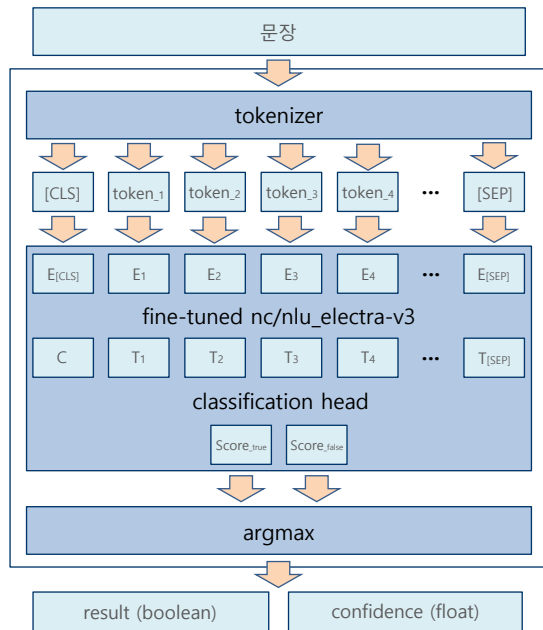
[3] 홍소영. "중국어 학습자의 인과관계 연결어미 습득 양상 연구: '-아서', '-니까', '-느라고', '-므로'를 중심으로." 한국어와 문화 28 (2020): 133-186.

[4] 유로. "한 중 인과관계 표현 형식과 분류 기준 연구." 어문논집 77 (2016): 233-280.

방법론

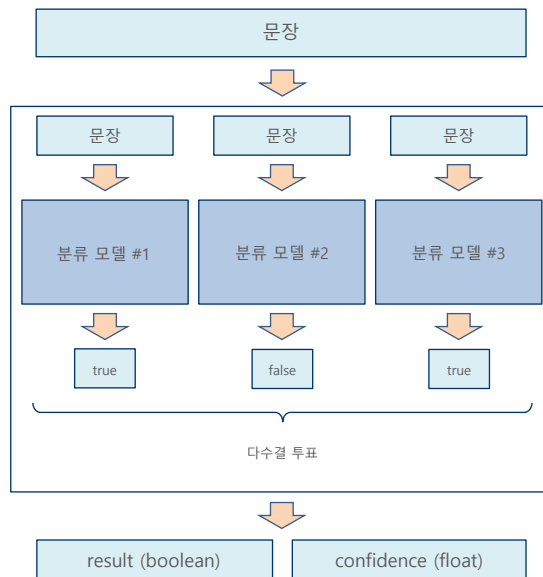
• BERT 분류 모델

- 문장을 입력 받아 인과 관계 여부를 반환
- BERT, ELECTRA 등의 모델을 fine-tuning 하여 성능 향상
- 모델 예시



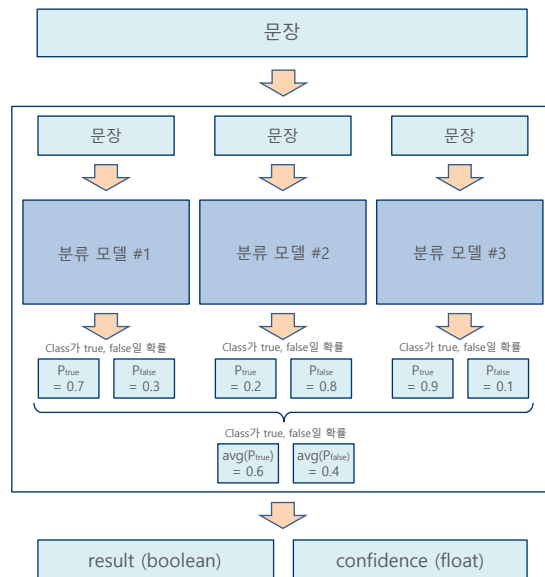
• Hard Voting Ensemble 모델

- 다수 모델의 분류 예측 결과 값을 다수결로 최종 class 결정
- $Ensemble(\hat{y}) = \operatorname{argmax}(\sum_{j=1}^n I(\hat{y}_j = i), i \in (0, 1))$
- 모델 예시



• Soft Voting Ensemble 모델

- 다수 모델의 분류 예측 결과값 간 확률을 평균하여 최종 class 결정
- $Ensemble(\hat{y}) = \operatorname{argmax}(\frac{1}{n} \sum_{j=1}^n P(\hat{y}_j = i), i \in (0, 1))$
- 모델 예시

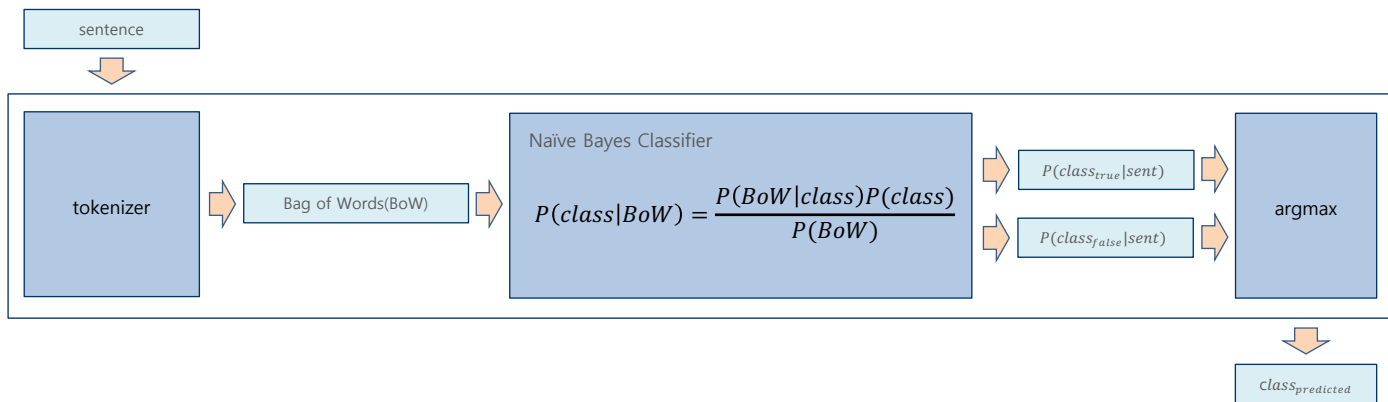


방법론

• Naive Bayes 분류 모델^[1]

- 모든 특성 값들(BoW의 각 빈도 값)은 서로 독립임을 가정한 후, 베이즈 정리를 이용한 확률 분류기
- 베이즈 정리(Bayes' theorem) : 어떤 사건이 서로 배반하는 원인 둘에 의해 일어난다고 할 때 실제 사건이 일어났을 때 이것이 두 원인 중 하나일 확률

- 모델 구조



[1] Raschka, Sebastian. "Naive bayes and text classification i-introduction and theory." arXiv preprint arXiv:1410.5329 (2014).

환경 및 결과 분석

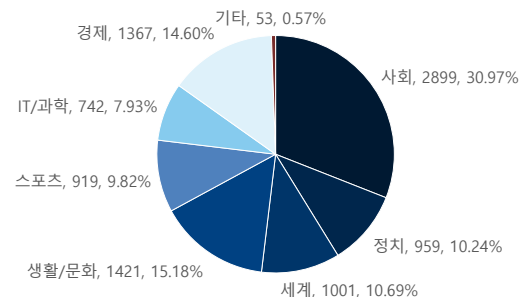
• 데이터셋

- 한국어 뉴스 인과관계 데이터셋

- 뉴스 기사 본문 중 하나의 문장과 인과관계 여부로 구성됨
- 각 section 비율이 동일하게 train, valid, test 데이터셋으로 분리
- train 데이터셋은 6739개, valid 데이터셋은 1685개, test 데이터셋은 937개의 뉴스 데이터를 가짐
- 뉴스 인과관계 데이터 예시

id	content	label	section
056-0010850372	농가 생산비 절감을 돕고, 가축분뇨 자원화로 자연친화적인 축산 환경을 만들겠습니다.	False	사회
079-0003310757	화산재로 인해 이날 오후 6시부터는 마닐라 국 제공항의 항공기 운항이 전면 중단됐다.	True	세계
...

- section 별 개수 통계



환경 및 결과 분석

- BERT 분류 모델

- 아래에 대한 모든 조합에 대하여 "한국어 뉴스" train 데이터셋으로 학습한 모델 생성
 - Pretrained Models = { nc/nlu_electra, bert-kor-base, electra-kor-base, kobert-base-v1, bert-base-multilingual-uncased, KcELECTRA-base, kcbert-large }
 - Learning rate = { 1e-4, 5e-5, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7 }
 - Dropout rate = { 0.00, 0.05, 0.10, 0.15, 0.20, 0.25 }

- Voting Ensemble 모델

- 최고 성능(f1-score)을 가진 top3, top5 모델들을 Hard/Soft Voting Ensemble 한 모델 생성

- Naive Bayes 분류 모델

- 아래에 대한 모든 조합에 대하여 "한국어 뉴스" train 데이터셋으로 학습한 모델 생성
 - tokenizer = { okt, komoran, hannanum, kkma } (KoNLPy[2]의 tokenizer)
 - BoW의 토큰 종류 개수 = { 500, 1000 }
 - pos-tagging 여부 = { true, false }

- 정량 평가 방법

- 모든 모델들에 대하여 "한국어 뉴스" test 데이터셋으로 f1-score, precision, recall 측정

[1] 문장 인과 관계 분류 실험 (Git) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/kor-causality-detection.git

[2] 파이썬 한국어 NLP(KoNLPy) : <https://konlpy.org/ko/latest/index.html>

환경 및 결과 분석

• 정량 평가 결과

BERT 모델, Voting Ensemble 모델 정량 평가 결과 top 10						
rank	model-name	learning-rate	dropout-rate	precision	recall	f1-score
1	voting_ensemble_soft-top3	-	-	0.793132	0.799360	0.791573
2	nc/nlu_electra-v3	0.000050	0.15	0.791064	0.797225	0.791329
3	voting_ensemble_hard-top3	-	-	0.790917	0.797225	0.790784
4	nc/nlu_electra-v1	0.000005	0.10	0.789119	0.795091	0.789934
5	voting_ensemble_hard-top5	-	-	0.789119	0.795091	0.789934
6	voting_ensemble_soft-top5	-	-	0.786604	0.792956	0.787210
7	bert-kor-base	0.000005	0.20	0.784922	0.791889	0.783660
8	electra-kor-base	0.000005	0.05	0.780487	0.785486	0.782041
9	kobert-base-v1	0.000010	0.05	0.764725	0.770544	0.766593
10	bert-base-multilingual-uncased	0.000005	0.15	0.741347	0.751334	0.743265

□ voting ensemble models

□ top-5 models

환경 및 결과 분석

• 정성 평가

- 규칙 기반 인과 관계가 추출되는 20000개의 예시들을 정성평가에 사용
- (모델 결과의 True : False 비율) = 11064개 : 8936개 = 55.32% : 44.68%

text	cause(규칙 기반으로 추출된 원인)	effect(규칙 기반으로 추출된 결과)	model-output
새해 첫날 미세먼지 농도는 원활한 대기 확산의 영향으로 전 권역에서 좋음~보통 수준을 보일 전망입니다.	새해 첫날 미세먼지 농도는 원활한 대기 확산의	전 권역에서 좋음~보통 수준을 보일 전망입니다.	True
석유·가스 등 가격 상승에 따른 에너지 수입 급증도 수입을 늘린 요인으로 꼽혔다.	석유·가스 등 가격 상승	에너지 수입 급증도 수입을 늘린 요인으로 꼽혔다.	True
시는 미분양 증가와 거래량 급감에 따른 시장 침체가 지역 경제 전반에 악영향을 끼칠 것으로 우려해 이같이 건의했다.	시는 미분양 증가와 거래량 급감	시장 침체가 지역 경제 전반에 악영향을 끼칠 것으로 우려해 이같이 건의했다.	True
산업부는 "2020년 12월(12.4%), 코로나19 사태 뒤 처음으로 두 자릿수 증가율을 기록한 것에 연이어 20%에 육박하는 성장세를 기록했다는 점에서 의미가 크다"고 평가했다.	산업부는 "2020년 12월(12.4%), 코로나19	뒤 처음으로 두 자릿수 증가율을 기록한 것에 연이어 20%에 육박하는 성장세를 기록했다는 점에서 의미가 크다"고 평가했다.	False
이에 따라 계도기간 종료 하루 전인 9일까지 접종을 받으려면 2일까지는 예약을 해야 한다.	이	계도기간 종료 하루 전인 9일까지 접종을 받으려면 2일까지는 예약을 해야 한다.	False
이번 담판이 더욱 중요한 이유는 과거 동서 냉전의 주역이었던 미국과 러시아가 유럽에서 세력권을 둘러싸고 정면으로 마주한 자리기 때문이다.	는 과거 동서 냉전의 주역이었던 미국과 러시아가 유럽에서 세력권을 둘러싸고 정면으로 마주한 자리기 때문이다.	이번 담판이 더욱	False
이어 나토가 추가 확장 중단에 관한 법적 보장을 해야한다고 재차 압박하면서 우크라이나에 대한 모든 군사적 지원과 무기 공급을 중단하라고 나토에 촉구했다.	이어 나토가 추가 확장 중단에 관한 법적 보장을 해야한다고 재차 압박하	우크라이나에 대한 모든 군사적 지원과 무기 공급을 중단하라고 나토에 촉구했다.	False

[1] 문장 인과 관계 분류 정성평가(예시 20000개) : <http://galadriel:8090/x/FHBfBQ>

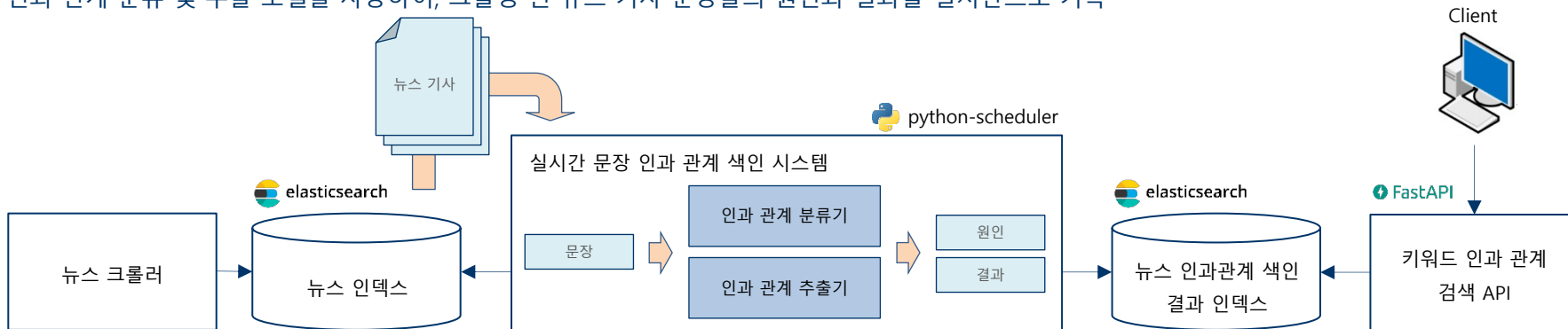
환경 및 결과 분석

- 정성 평가 오류 분석
 - 규칙 기반으로 인과 관계가 추출되더라도, 인과 관계가 없을 수 있음 → 인과 관계 분류는 인과 관계 분석에서 필요한 단계
 - Soft-voting, Hard-voting 외에 Weighted-Voting Ensemble 모델 시도가 없었음
 - 규칙 기반 인과 관계 추출에 대명사가 포함 → 대명사가 가리키는 대상 검색이 필요

키워드 인과 관계 검색 서비스

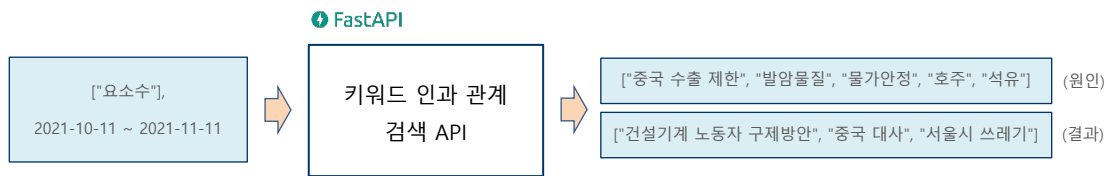
• 실시간 문장 인과 관계 색인 시스템

- 인과 관계 분류 및 추출 모델을 사용하여, 크롤링 된 뉴스 기사 문장들의 원인과 결과를 실시간으로 기록



• 키워드 인과 관계 검색 API

- 날짜 범위와 키워드들을 입력하면, 해당 날짜 범위 내에 키워드들의 원인들과 결과들을 Elastic Search를 사용하여 검색
- 원인들과 결과들은 검색된 개수 순으로 정렬되어 상위 n개만 추출되며, stopwords와 같은 결과값들은 제거



[1] 인과 관계 분류 API (Git, Docker) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/kor-causality-detection-docker.git

[2] 실시간 문장 인과 관계 색인 시스템 및 검색 API (Git, Docker) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/es-casuality.git

Contents

01. 서론	3
02. 비지도 문서 키워드 추출	9
03. 문장 인과 관계 분류 및 추출	29
04. 트렌드 키워드 시스템	41
a. 소개	42
b. 작동 원리	43
c. 결과 분석	45

소개

- 소개

- "도메인을 대표하는 키워드는 관심 대상 도메인의 핵심 정보"^[1]
- 특정 날짜 범위 내의 도메인의 트랜드 키워드를 얻는 것은, 도메인의 중요한 구간 정보에 대한 묘사를 얻는 것과 같음
- 날짜 범위(일별/주별/월별/년별)와 도메인(정치/경제/사회/세계/스포츠/IT-과학/생활-문화)에 대한 트랜드 키워드들을 얻는 것이 목표
- 도메인과 날짜 범위에 따른 트랜드 키워드 예시

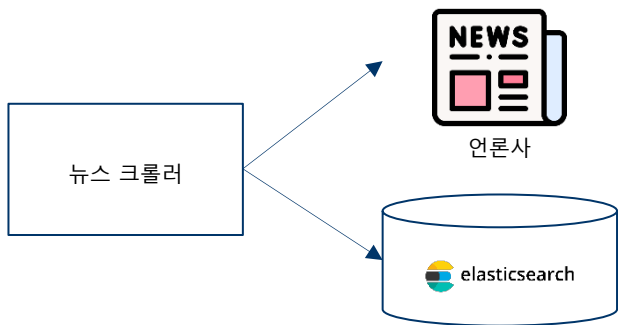
도메인(section)	날짜 범위	트랜드 키워드
세계	일별/2023년 2월 6일	튀르키예, 시리아, 지진
스포츠	월별/2022년 12월	월드컵, 아르헨티나, 메시
사회	주별/2022년 10월 5번째	이태원, 경찰, 참사

[1] Akash, Pritom Saha, et al. "Domain Representative Keywords Selection: A Probabilistic Approach." arXiv preprint arXiv:2203.10365 (2022).

작동 원리

• 실시간 뉴스 크롤러

- 실시간으로 언론사 사이트들을 크롤링하여 Elasticsearch 인덱스에 뉴스 데이터 삽입



- 저장된 뉴스 데이터 document_source 예시

```

{
  "document_id": "0115_202301010228025461",
  "title": "광안리해수욕장에서 새해맞이 대규모 드론 공연",
  "author": "김종호",
  "date": "2022-12-31T17:28:00Z",
  "url": "https://www.ytn.co.kr/_ln/0115_202301010228025461",
  "section": "사회",
  "content": "부산 광안리해수욕장에서 새해맞이 대규모 드론 공연이 열렸습니다. 부산 수영구는 어젯밤 11시 56분부터 10분 동안 해운대해수욕장 상공에 드론 천5백 대를 띄워 새해 첫날 새벽 0시에 맞춘 초원기와 토끼, 달리는 사람 등을 형상화한 공연을 선보였습니다. (...)"
}
  
```

• TF-IDF 기반 키워드 추출 API

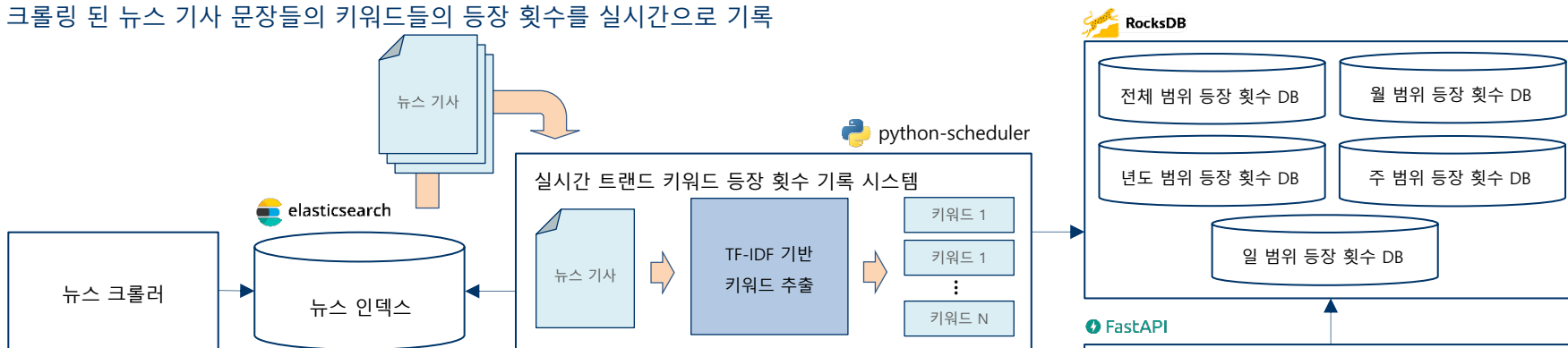
- 비지도 문서 키워드 추출 실험에서 사용한 TF-IDF 기반 키워드 추출 모델을 FastAPI를 사용하여 배포
- 문서를 입력받아 키워드들을 반환

[1] TF-IDF 기반 키워드 추출 API (Git, Docker) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/uke-tf-idf-api-docker.git

작동 원리

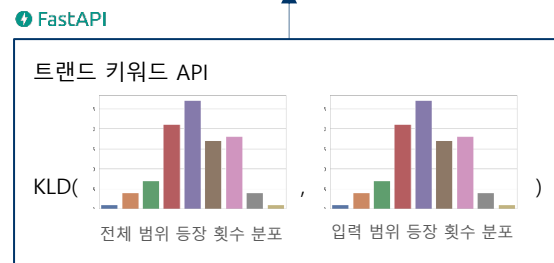
• 실시간 트렌드 키워드 등장 횟수 기록 시스템

- 크롤링 된 뉴스 기사 문장들의 키워드들의 등장 횟수를 실시간으로 기록



• 트렌드 키워드 API

- 년도별, 월별, 주별, 일별 범위 단위의 트렌드 키워드를 추출
- 트렌드 키워드는 전체 범위 등장 횟수 분포와 입력된 범위 등장 횟수 분포 사이의 pointwise KLD 점수가 높은 순으로 추출됨



Client

[1] 트렌드 키워드 (Git, Docker) : http://galadriel02.korea.ncsoft.corp/Geon_Kim/trend-keywords.git

결과 분석

• 정성 평가 및 오류 분석

- 일별 트렌드 키워드 / 2023년 2월 6일 검색 결과

section	트렌드 키워드
정치	이 장관, 후보, 6일, 외교, 아이뉴스24, 윤핵관, 연대, 정치, 통일, 발의
경제	지난해, 6일, 올해, 1분전, 최근, 증권부, 이날, 거래량회전율, 이사회, 이원장
사회	유가족, 김 전 회장, 6일, 해경, 김성태, 조 전 장관, 쌍방울, 서울광장, 실종자, 분향소
세계	튀르키예, 지진, 발생, 시리아, 격추, 최소, 건물, 강진, 남부, 풍선
스포츠	수원, 6일, 스포츠 기자, 아레나, MK, 2022-2023 프로농구, 수원 KT, 천정환, KGC, 볼펜 피칭
IT/과학	대통령실, 지난해, 당, 여당, 이 장관, 대통령, 이상민, 위반, 행정안전부 장관, 때
생활/문화	이데일리 스타in, 때, .06, 2023.02, 충청 베스트, 강민경, 강원영서, SM, 0
기타	때, 난방비, WSJ, 지난해, 이후, 최근, 난방, 피의자, 100일 설치

- 일별 트렌드 키워드 / 2022년 12월 검색 결과

section	트렌드 키워드
정치	윤 대통령, 대통령실, 이태원, 여야, 여당, 이상민, 예산안, 국정조사, 참사, 이재명 대표
경제	이 시각, 기록, 1개월간, 체결강도, 총매수체결량, 총매도체결량, 거래대금, 2023년, 화물연대, 선정
사회	혐의, 이태원, 눈, 참사, 선정, 구속영장, 화물연대, 검찰, 아시아경제, 피해자
세계	러시아, 우크라이나, 머스크, 푸틴 대통령, 젤렌스키 대통령, 전쟁, 우크라이나 전쟁, IRA, 아르헨티나, 월드컵
스포츠	OSEN, 월드컵, 2022.12, 아르헨티나, 브라질, 메시, 포르투갈, 프랑스, 16강, sportschosun
IT/과학	아시아경제, 이데일리, 민주당, 위믹스, 아이뉴스24, 이상민, 여야, 구 대표, ESG, 선정
생활/문화	OSEN, 2022.12, 포즈, osen, co, 스포츠조선닷컴, sportschosun, 스포츠조선, 배우, com
기타	뉴시스, 화물연대, 연합인포맥스, 파업, 하락, 오전, 거래, BOJ, 예산안, 승진

- 날짜, 기자 명, 언론사 명과 같은 단어가 트렌드 키워드로 추출 됨 → 추출 단계에서 stopwords와 같은 키워드 제거 필요

- "윤 대통령", "윤", "대통령 실"이 다른 키워드로 추출 됨 → 유사 키워드 클러스터링 필요

- 신조어와 같은 복합 명사 청킹 사전에 없는 단어들은 키워드 후보 군에서 제외될 수 있음

Q & A

geon6757@ncsoft.com
geon6757@kaist.ac.kr



END OF DOCUMENT