Dense Retrieval 을 이용한 랭킹 보존 기반 키워드 추출

Ranking Preserving Keyphrase Extraction using Dense Retrieval

이종현° • 나승훈 (전북대학교)

Contents

▪ 키워드 추출

문서 내의 중요한 정보를 요약하거나 문서를 대표할 수 있는 워드(word)나 구(phrase)를 예측하는 태스크

국제유가, 미 생산 증가 우려에 약세 •••WTI 0.9%↓ (연합뉴스, 2018.01.20)

국제유가는 19일(현지시간) 하락했다. 이날 뉴욕상업거래소(NYMEX)에서 2월 인도분 서부 텍사스산 원유(WTI)는 배럴당 0.58달러(0.9%) 하락한 63.37달러에 거래를 마쳤다. 이번주 주간 기준으로는 1.5% 하락률을 기록했다. 영국 ICE 선물거래소의 브렌트유 2월물도 같은 시각 배럴당 0.63달러(0.91%) 내린 68.68달러에 거래되고 있다. 미국의 원유 생산량이 사상 최고 수준에 이를 것이라는 전망이 나오면서 투자심리가 위축됐다. 국제에너지기구(IEA)는 올해 미국의 원유 생산량이 사우디아라비아를 넘어설 것으로 전망했다. 국제금값은 강세를 보였다. 뉴욕상품거래소에서 2월물 금 가격은 전날보다 온스당 5.90달러(0.4%) 오른 1,333.10달러에 마감했다. 연방정부의 셧다운(일시적 업무정지) 우려로 안전자산 선호 심리가 강해지면서 금값에 호재로 작용했다.

출처 : SBS 뉴스

원본 링크 : https://news.sbs.co.kr/news/endPage.do?news_id=N1004582660&plink=COPYPASTE&cooper=SBSNEWSEND

- 키워드

  국제유가, 하락, 뉴욕상업거래소, 투자심리가 위축, 국제금값은 강세, ...

▪ 정보 검색 시스템 (IR system (Information Retrieval System))

• Sparse Retrieval (키워드 기반 정보 검색 시스템) : BM25, TF-IDF

- Bag-of-word 를 이용하여 질문과 문서를 고차원의 희소 벡터로 나타내는 방식
- 질문과 문서에 등장하는 단어(word)에 기반하여 질문과 문서 사이의 유사도를 계산하기 때문에 동의어, 유의어의 처리가 까다로움
- 도메인에 맞게 튜닝하기 어려움

• Dense Retrieval (딥러닝 기반의 정보 검색 시스템)

- 질문과 문서를 연속 벡터로 인코딩하여 문서 검색을 신경망 기반의 거리 측정을 통해 수행
- Dense Retrieval 은 자연어에 대한 문맥적 이해에 기반하기 때문에 Sparse Retrieval 에 비해 다양한 질문에 유연하게 대처할 수 있음
- 최근 오픈 도메인 질의 응답에서 Sparse Retrieval 의 성능을 큰 폭으로 뛰어 넘는 모습을 보임

▪ 키워드 추출

문서의 키워드는 해당 문서의 중요한 정보를 요약하거나 특징을 내제 하기도 하므로 후보 키워드를 가지고 정보 검색 시스템(IR system)을 이용해 문서 집합에서 문서를 검색하였을 때, 해당 문서의 검색 점수가 높게 나오는 키워드는 보다 의미 있는 키워드일 가능성이 높음

- 기존의 키워드 추출 모델에 Dense Retrieval 을 적용한 랭킹 보존 기반 키워드 추출 모델을 제안

통계 기반 키워드 추출 모델인 PositionRank[1] + Dense Retrieval 에서는 약 2.94%p 의 성능 개선

딥러닝 기반 키워드 추출 모델인 CatSeq[2] + Dense Retrieval 에서는 약 0.21%p 의 성능 개선

[1] Florescu, Corina, and Cornelia Caragea. "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017.

[2] Yuan, Xingdi, et al. "One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases." *arXiv preprint arXiv:1810.05241* (2018).

- 랭킹 알고리즘을 이용한 키워드 추출

  • Single document keyphrase extraction using neighborhood knowledge [X. Wan and J. Xiao, ‘2008]

  • Textrank: Bringing order into text [R. Mihalcea and P. Tarau, ‘2004]

  • Human-competitive tagging using automatic keyphrase extraction [O. Medelyan, E. Frank, and I. H . Witten, ‘2009]

- 시퀀스 라벨링(Sequence labeling)을 이용한 키워드 추출

  • Keyphrase extraction using deep recurrent neural networks on twitter [Q. Zhang, Y. Wang, Y. Gong, and X.-J. Huang, ‘2016]

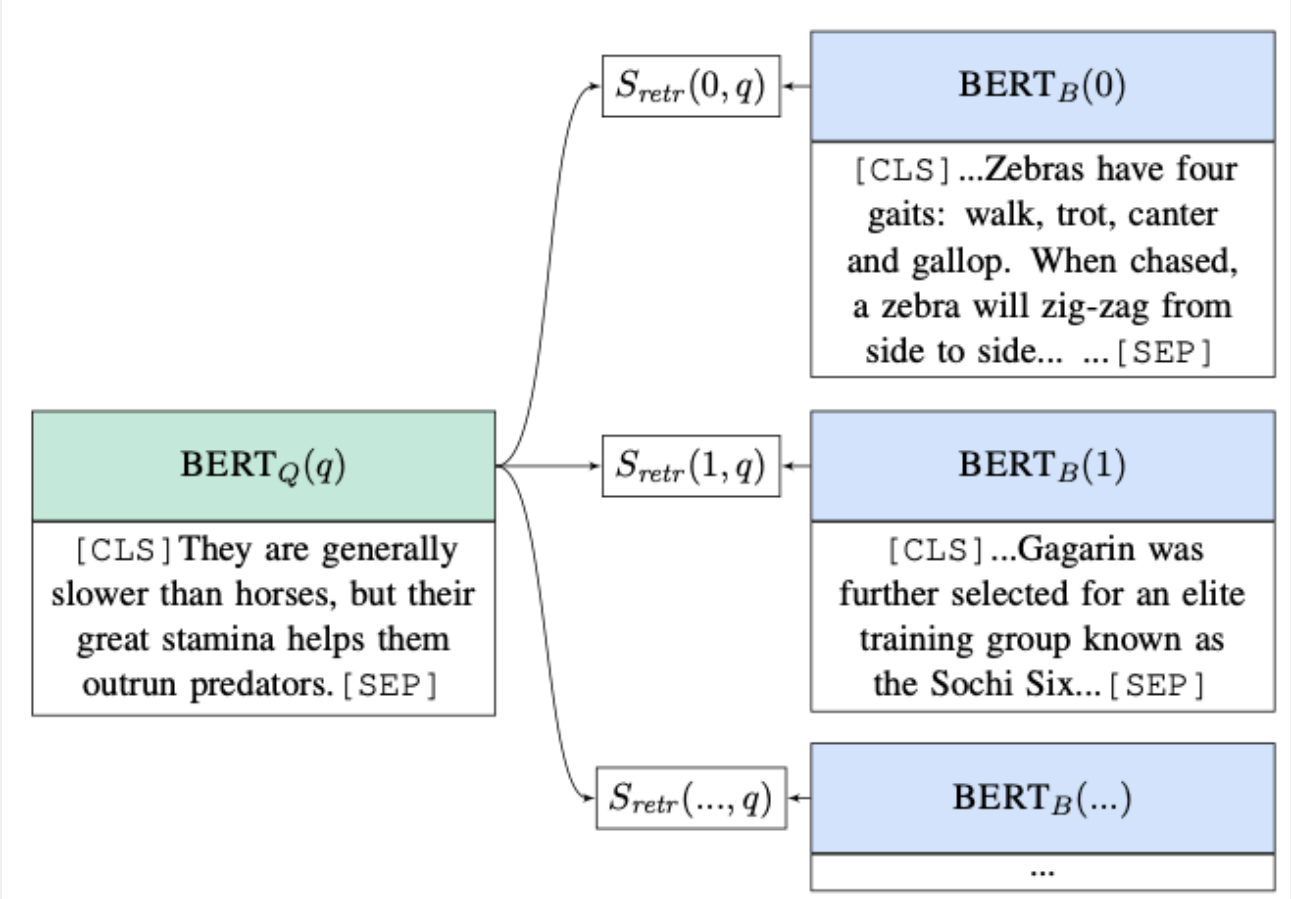
  • Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents [R. Alzaidy, C. Caragea , and C. L. Giles, ‘2019]

- Graph Neural Networks 를 이용한 키워드 추출

  • Divgraphpointer: A graph pointer network for extracting diverse keyphrases [Z. Sun, J. Tang, P. Du, Z.-H . Deng, and J.-Y. Nie, ‘2019]

▪ Dense Retrieval 모델

- 키워드 추출에 적용하기 위한 dense retrieval 모델로 [3] 의 ORQA 모델을 이용



[3] Lee, Kenton, Ming-Wei Chang, and Kristina Toutanova. "Latent retrieval for weakly supervised open domain question answering." *arXiv preprint arXiv:1906.00300* (2019).

▪ Dense Retrieval 모델 사전 학습

- ICT(Inverse cloze task) 사전 학습

ORQA 모델을 한국어 위키피디아 데이터 셋을 이용해 ICT(Inverse cloze task) 사전학습하여 키워드 추출 모델에 적용

블록 인코더 문서에서 임의의 문장이 하나 제거된 문서를 인코딩

$$B = \{[BOS], title, [EOS], [EOS], body, [EOS]\}$$

$$\boldsymbol{H}_B = RoBERTa^B(B)$$

$$\boldsymbol{e}_B = \boldsymbol{W}_B \frac{\sum_i^{L_B} \boldsymbol{H}_B^i}{L_B}$$

쿼리 인코더 제거된 문장을 인코딩

$$Q = \{[BOS], query, [EOS]\}$$

$$\boldsymbol{H}_Q = RoBERTa^Q(Q)$$

$$\boldsymbol{e}_Q = \boldsymbol{W}_Q \frac{\sum_i^{L_Q} \boldsymbol{H}_Q^i}{L_Q}$$

나머지 사전 학습 방식은 [4] 와 동일함

[4] 강동찬, 나승훈, 최윤수, 이혜우, 장두성, "Realm을 이용한 한국어 오픈도메인 질의 응답," *in HCLT 2020*, 2020.

▪ Dense Retrieval Fine-tuning

- 한국어 뉴스 기사로 제작한 총 80,335 개의 문서 셋을 이용
- 전체 80,335 개의 문서 중 일부 문서(7,702 개)에 대해 키워드를 라벨링(labeling)하여 키워드 추출 데이터 셋으로써 사용

• **상위 K 개의 후보 문서 셋을 얻어냄**

정답 키워드 셋 $Y = \{y_1, y_2, \dots, y_N\}$

상위 K 개의 후보 문서 셋 $D = \{D_1, D_2, \dots, D_K\}$

▪ Dense Retrieval Fine-tuning

- 한국어 뉴스 기사로 제작한 총 80,335 개의 문서 셋을 이용
- 전체 80,335 개의 문서 중 일부 문서(7,702 개)에 대해 키워드를 라벨링(labeling)하여 키워드 추출 데이터 셋으로써 사용

• **상위 K 개의 후보 문서 셋을 얻어냄**

정답 키워드 셋 $Y = \{y_1, y_2, \dots, y_N\}$

↓

BM25

↓

상위 K 개의 후보 문서 셋 $D = \{D_1, D_2, \dots, D_K\}$

• **Fine-tuning**

문서 블록 집합 $\bar{B} = \{B_1, B_2, \dots, B_K\}$

쿼리 $Q = \{[BOS],\ y_1, y_2, \dots, y_N, [EOS]\}$

$$e_{\bar{B}} = BlockEncoder(\bar{B})$$

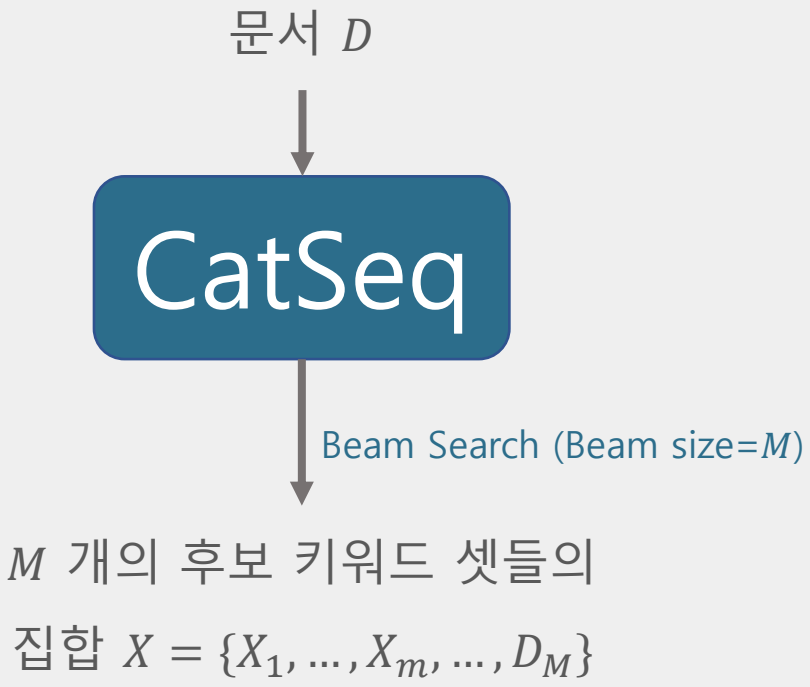
$$e_Q = QueryEncoder(Q)$$

$$P(B_j|Q) = \frac{\exp(e_{B_j} e_Q)}{\sum_{B' \in \bar{B}} \exp(e'_B e_Q)}$$

이후에, CrossEntropy Loss 를 이용하여 Fine-tuning 을 진행

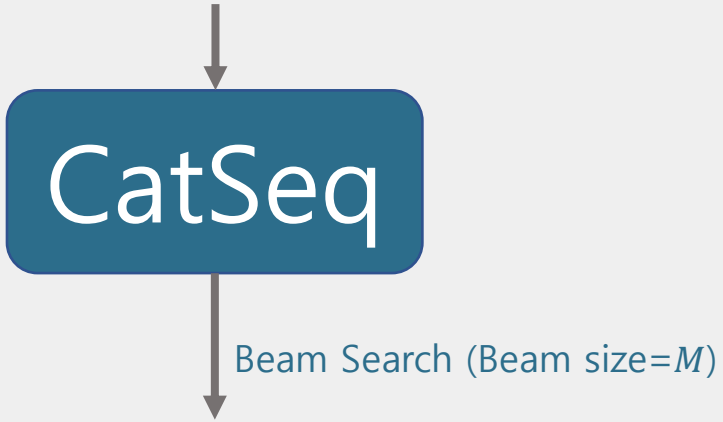
▪ 기존 키워드 추출 모델과 Dense Retrieval 의 결합 및 추론

- 추론 과정에서 fine-tuning 된 Dense Retrieval 은 키워드 추출 모델인 CatSeq의 Beam Search 과정에서 얻어낸 M 개의 후보 키워드 셋들 중 검색 점수가 가장 높은 후보 키워드 셋을 예측하는데 이용됨



▪ 기존 키워드 추출 모델과 Dense Retrieval 의 결합 및 추론

- 추론 과정에서 fine-tuning 된 Dense Retrieval 은 키워드 추출 모델인 CatSeq의 Beam Search 과정에서 얻어낸 M 개의 후보 키워드 셋들 중 검색 점수가 가장 높은 후보 키워드 셋을 예측하는데 이용됨



$Q_X = \{Q_{X_1}, \dots, Q_{X_m}, \dots, Q_{X_M}\}$

B_D = 문서 D 의 블록 인코더 입력

$$e_{B_D} = BlockEncoder(B_D)$$

$$e_{Q_{X_m}} = QueryEncoder(Q_{X_m})$$

$$dense\ score_m = e_{Q_{X_m}} e_{B_D}$$

$$score_m = CatSeq\ score_m + \lambda \cdot dense\ score_m$$

※ $CatSeq\ score_m$ 은 키워드 추출 모델에서 얻어진 후보 키워드 셋 X_m 에 대한 로그 확률(Log probability)를 의미함

▪ 실험 데이터 및 세팅

- 2017 년도의 한국어 뉴스 데이터를 자체 수집하여 데이터 셋을 구축
- 문서 검색을 위한 문서 데이터의 수는 총 80,355 개 이며, 이 중 7,702 개의 문서를 가지고 키워드를 라벨링하여 키워드 추출 데이터로써 사용

	문서 수	평균 키워드 수
학습 셋	5,391	6.09
개발 셋	770	6.15
평가 셋	1,541	6.11
총 합	7,702	-

표 : 키워드 추출 데이터 셋의 분포

Dense Retrieval Pre-training

Projection Layer : $\mathbb{R}^{128}$

Optimizer : AdamW

Learning Rate : 1e-4

Dense Retrieval Fine-tuning

Optimizer : AdamW

Learning Rate : 5e-5

▪ 실험 결과

모 델	λ (lambda)	F1 (%)
PositionRank (F_1@10)	-	24.78
Dense Retrieval + Position Rank (F_1@10)	0.01	25.10
Dense Retrieval + Position Rank (F_1@10)	0.02	**27.72**
CatSeq	-	54.34
Dense Retrieval + CatSeq	0.01	**54.55**
Dense Retrieval + CatSeq	0.02	54.39

▪ 결 론

- 키워드 추출 모델에 Dense Retrieval 을 결합한 랭킹 보존 기반 키워드 추출 모델을 제안

- 문서의 정답 키워드를 가지고 전체 문서 집합에 대해 검색을 하였을 경우에 해당 문서에 대한 검색 점수가 더 높은 키워드가 더 의미 있는 키워드라는 생각에 기반함

- 이를 기존의 키워드 추출 모델에 적용하여 개선된 성능을 얻어냄

감사합니다