TextRank 알고리즘과 주의 집중 순환 신경망을 이용한 하이브리드 문서 요약

정석원○, 이현구, 김학수
강원대학교
nlpsw@kangwon.ac.kr, nlphglee@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Hybrid Document Summarization using a TextRank Algorithm and an Attentive Recurrent Neural Networks

Seok-won Jeong○, Hyeon-gu Lee, Harksoo Kim
Kangwon National University

요 약

문서 요약은 입력 문서가 가진 주제를 유지하면서 크기가 축약된 새로운 문서를 생성하는 것이다. 문서 요약의 방법론은 크게 추출 요약과 추상 요약으로 구분된다. 추출 요약의 경우 결과가 문서 전체를 충분히 대표하지 못하거나 문장들 간의 호응이 떨어지는 문제점이 있다. 최근에는 순환 신경망 구조의 모델을 이용한 추상 요약이 활발히 연구되고 있으나, 이러한 방법은 입력이 길어지는 경우 정보가 누락된다는 문제점을 가지고 있다. 본 논문에서는 이러한 단점들을 해소하기 위해 추출 요약으로 입력 문서의 중요한 일부 문장들을 선별하고 이를 추상 요약의 입력으로 사용했을 때의 성능 변화를 관찰한다. 추출 요약을 통해 원문 대비 30%까지 문서를 요약한 후 요약을 생성했을 때, ROUGE-1 0.2802, ROUGE-2 0.1294, ROUGE-L 0.3254의 성능을 보였다.

주제어: 문서 요약, TextRank, 순환 신경망

1. 서론

문서 요약(Document summarization)은 입력 문서가 가진 주제를 유지하면서 크기가 축약된 새로운 문서를 생성하는 과정이다. 문서 요약의 방법론은 크게 추출 요약(extractive summarization)과 추상 요약(abstractive summarization)으로 구분된다. 추출 요약은 원본 문서가 가진 문장들을 그대로 활용하여 요약하는 것으로, 문장들의 상대적 중요도에 따라 가장 중요한 일부 문장을 선별함으로써 문서를 요약한다. 추출 요약은 과거부터 활발히 연구되었으며[1, 2], 비교적 단순한 방법으로도 그럴듯한 결과들을 보였다. 그러나 추출 요약의 결과로 선별된 문장들이 문서 전체를 충분히 대표하지 못하거나, 선별된 문장들 간의 호응이 떨어지는 문제점들이 존재한다. 최근에는 순환 신경망(Recurrent Neural Network) 구조의 모델을 이용해 원본 문서에 없는 새로운 문장들을 생성해내는 추상 요약이 활발히 연구되고 있다[3, 4, 5]. 그러나 추상 요약은 특정 출력을 반복하거나 입력이 길어지는 경우 일부 정보들이 누락되는 등의 문제점들이 있다. 본 논문에서는 이러한 문제들을 해소하기 위해 추출 요약을 통해 중요 문장들을 선별하고, 이를 추상 요약의 입력으로 사용하여 원문에 없는 새로운 문장을 생성하는 하이브리드 문서 요약을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들에 대해서 살펴보고, 3장에서는 제안 시스템의 추출 요약과 추상 요약 방법에 대해서 설명한다. 4장에서는 실험에 대해서 살펴보고 5장에서 끝을 맺는다.

2. 관련 연구

PageRank[6]는 하이퍼링크를 통해 연결된 웹 문서 간의 상대적 중요도를 계산하는 그래프 기반 순위화 알고리즘으로, 중요한 페이지일수록 더 많은 인용을 받는다는 것을 기초로 한다. TextRank[7]는 PageRank의 개념을 자연어 처리에 응용한 것으로 문장, 단어와 같은 특정 단위들 간의 중요도를 계산하는 알고리즘이다. 문서 내의 각 문장을 그래프의 정점(vertex)으로 가정하는 경우 중요한 문장들을 선별할 수 있으며, 이를 통해 문서 요약이 가능하다. 같은 원리로 각 단어를 정점으로 가정할 경우 중요 키워드를 선별할 수 있다. 본 논문에서는 TextRank를 이용한 추출 요약으로 입력 문서의 중요한 문장들을 선별하여 추상 요약의 입력으로 사용한다.

최경호 외[8]는 주의 집중 순환 신경망(Attentive RNN)[9]을 비롯한 다양한 모델을 이용해 한국어 문서 요약을 수행하였으며 음절, 형태소, 음절+형태소 혼합 등 입력 형태에 따른 추상 요약의 성능을 비교하였다. 이 결과를 통해 본 논문에서는 형태소 단위의 입력으로 추상 요약을 수행한다.

3. 제안 시스템

제안 시스템은 추출 요약을 통해 입력 문서를 먼저 요약한 뒤, 그 결과를 추상 요약의 입력으로 사용하는 파이프라인 형태의 구조이다. 구조도는 [그림 1]과 같다.

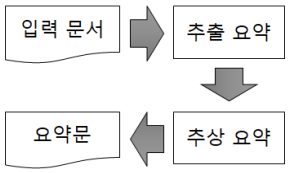

그림 1 시스템 구조도

3.1 TextRank를 이용한 추출 요약

본 논문에서는 TextRank를 이용한 추출 요약으로 입력 문서의 문장들을 선별한다. 이를 위해 입력 문서의 각 문장들에 대해 형태소 분석을 수행하고, 체언류와 용언류의 $TF \cdot IDF$를 계산하여 문장-단어 행렬을 생성한다. 그 뒤 생성된 문장-단어 행렬의 전치 행렬을 구하여 서로 곱해주면 문장 간의 상관관계(correlation)를 나타내는 행렬을 얻을 수 있다. 상관행렬을 구하는 예시는 아래 [그림 2] 같다.

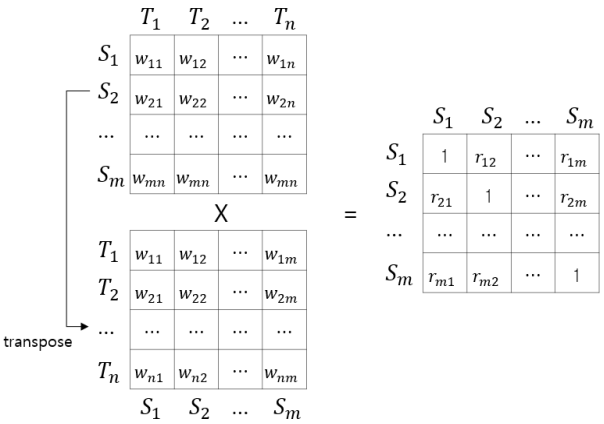

그림 2. 문장 간 상관행렬 예

문장 간 상관행렬은 문장 간의 가중치 그래프로 나타낼 수 있으며, TextRank 알고리즘을 통해 각 문장의 중요도를 구할 수 있다. TextRank의 수식은 아래 식 (1)과 같다.

$$TR(V_i) = (1-d) + d^* \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TR(V_j) \quad (2)$$

TextRank를 통해 구한 중요도 순으로 문장들을 정렬한 뒤, 상위 n개 문장 외의 나머지를 제거하고 남은 문장들을 출현 순서대로 재배치하면 요약 결과를 얻을 수 있다. TextRank를 이용한 추출 요약의 결과 예시는 [표 1]과 같다.

(기사 출처 : http://www.insight.co.kr/news/87744)

표 1. TextRank를 이용한 추출 요약 예

원본 문서	22일 서울 여의도 국회에서 진행중인 최순실 국정농단의혹사건 진상규명을 위한 국정조사특별위원회 제5차 청문회에서 위원들은 우병우 전 수석을 향해 끊임없이 "최순실을 아느냐"고 질문했다. 이에 한결같이 "모른다"는 답변만 늘어놓은 우 전 수석에게 손혜원 더불어민주당 의원은 일침을 가했다. 손 의원은 "우병우 증인은 거짓말을 할 때 눈을 깜빡 깜빡 3번한다"면서 "지금도 역시 눈을 깜빡 거리고 있다"고 소리쳤다. 이 말을 들은 우 전 수석은 웃음이 나는지 이를 참기 위해 애쓰는 모습을 보이기도 했다. 손 의원이 질문을 바꿔 "차은택도 모르냐"고 호통했고, 우 전 수석은 역시 "모른다"고 답변했다. 그러자 손 의원은 "차은택은 평소에도 우병우 수석이 봐준다고 했다"며 참고인으로 출석한 K스포츠 재단 노승일 부장을 향해 "우병우가 정말 차은택을 모르는 것 같냐"고 질의했다. 이에 노 전 부장은 "차은택의 법조 조력자가 김기동인데, 우병우가 김기동을 소개시켜줬다는 이야기를 고영태에게 들었다"고 증언했다. 이처럼 참고인 노 전 부장을 비롯한 다른 증인들과 우 전 수석이 전혀 다른 주장을 펼치면서 누군가는 위증을 하고 있다는 의혹이 커지고 있는 상황이다. 한편 이날 청문회에서 우 전 수석은 질의를 받는 동안 불량한 태도로 김성태 국조특위 위원장에게 지적을 받기도 했다.
요약 결과	이에 한결같이 "모른다"는 답변만 늘어놓은 우 전 수석에게 손혜원 더불어민주당 의원은 일침을 가했다. 이처럼 참고인 노 전 부장을 비롯한 다른 증인들과 우 전 수석이 전혀 다른 주장을 펼치면서 누군가는 위증을 하고 있다는 의혹이 커지고 있는 상황이다.

3.2 주의 집중 순환 신경망

주의 집중 순환 신경망은 기본적인 형태의 순환 신경망 인코더-디코더(RNN encoder-decoder)에 주의 집중(attention mechanism)을 추가한 구조이다. 인코더-디코더 구조의 신경망은 입력을 길이에 상관없이 고정된 크기의 벡터로 인코딩하며, 이로 인해 누락되는 정보가 생길 수 있다. 주의 집중 순환 신경망은 고정 길이 벡터의 사용이 인코더-디코더 구조의 성능을 향상시키는 데 있어 병목 현상이 되는 것을 방지하기 위해, 모델이 자동으로 입력의 적절한 부분을 찾도록 확장한다[9]. 이를 통해, 주의 집중 순환 신경망은 기계 번역 분야에서 좋은 성능을 보였으며, 다른 자연어처리 분야에도 효과적으로 활용된다. 주의 집중 순환 신경망의 구조도는 아래 [그림 3]과 같다.

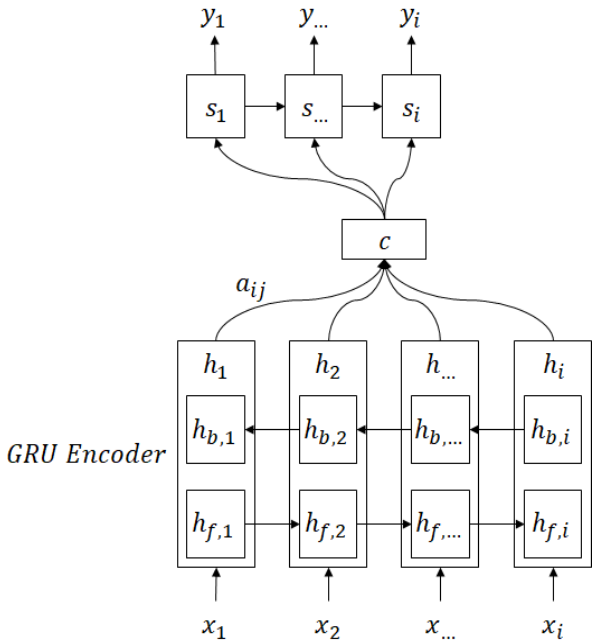

그림 3. 주의 집중 순환 신경망 구조

주의 집중 순환 신경망 모델의 인코더로는 양방향 GRU(Gated Recurrent Unit)를 사용한다. 모델을 수식으로 표현하면 아래 식 (2)과 같다.

$$
\begin{aligned}
h_{f,i} &= GRU(x_i, h_{f,i-1}) \\
h_{b,i} &= GRU(x_i, h_{b,i+1}) \\
h_i &= [h_{f,i}; h_{b,i}] \\
e_{ij} &= f(s_{i-1}, h_j) \\
\alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \\
c_i &= \sum_{j=1}^{T_x} \alpha_{ij} h_j \\
s_i &= f(s_{i-1}, c_i)
\end{aligned}
\qquad (3)
$$

식 (2)에서 입력 열 $X = \{x_1, x_2, \cdots, x_i\}$는 GRU를 통해 인코딩되며 $h_{f,i}$와 $h_{b,i}$는 각각 정방향, 역방향의 은닉 계층을 나타내고, h_i는 두 은닉 계층의 결합을 나타낸다. a_{ij}는 주의 집중 가중치를 나타내며 c_i는 주의 집중 가중치 및 입력을 통해 생성된 문맥 벡터, s_i는 디코더의 은닉 계층을 나타낸다.

제안 시스템은 위에서 설명한 주의 집중 순환 신경망을 이용하여 추상 요약을 수행한다. 신경망의 입력은 TextRank를 통해 일부 요약된 결과이며 출력으로 요약된 문장을 출력한다. 주의 집중 순환 신경망을 이용한 추상 요약의 예제는 아래 [표 2]와 같다.

표 2. 주의 집중 순환 신경망을 이용한 추상 요약 예

입력	이에 한결같이 "모른다"는 답변만 늘어놓은 우전 수석에게 손혜원 더불어민주당 의원은 일침을 가했다. 이처럼 참고인 노 전 부장을 비롯한 다른 증인들과 우 전 수석이 전혀 다른 주장을 펼치면서 누군가는 위증을 하고 있다는 의혹이 커지고 있는 상황이다.
출력	우병우 청와대 민정수석이 '최순실 게이트' 진상규명을 위해 자신의 집 앞에 섰다 .
정답 요약	최순실 국정농단 사태 진실규명을 위한 청문회에 증인으로 출석한 우병우 청와대 전 민정수석은 '최순실을 아느냐'는 질문에 "모른다"로 일관했다.

[표 2]에서 출력은 추상 요약의 결과이다. 정답 요약은 모델 학습 시에 정답으로 사용한 문장으로, 기자가 해당 기사에 작성한 코멘트이다. 실제 입력 및 출력은 형태소 단위로 이루어지며, [표 2]에서는 가독성을 위해 일반 문장으로 표시하였다.

4. 실험

4.1 실험 데이터

본 논문에서는 실험을 위해 인사이트 뉴스 기사 21,072 문서를 수집하였다. 전체 데이터 중 19,999 문서를 학습에 사용하였으며, 1,073 문서를 평가에 사용하였다. 학습을 위한 정답으로는 기사를 작성한 기자가 남긴 코멘트를 사용하였다. 요약 결과를 평가하기 위한 지표로는 ROUGE[10]를 사용하였으며, 그 중에서도 ROUGE-1, ROUGE-2, ROUGE-L을 사용하였다.

4.2 성능

제안 시스템의 성능은 아래 [표 3]과 같다.

표 3. 입력에 따른 요약 성능 비교

입력 길이	ROUGE-1	ROUGE-2	ROUGE-L
100%	0.2961	0.1453	0.3348
70%	0.2901	0.1395	0.3317
30%	0.2802	0.1294	0.3254

[표 3]에서 성능은 각 ROUGE의 F1-점수이며, 입력 길이는 추상 요약에 사용한 입력의 길이를 나타낸다. 입력 길이 100%는 추출 요약을 거치지 않은 원본 문서를 나타내며, 입력 길이 70%의 경우 TextRank를 통해 원본 문서 대비 70% 수준으로 문서를 요약한 것을 나타낸다.

성능 평가 결과, 중요한 일부 문장을 선별하여 입력할

경우 성능이 향상될 것이라는 예상과 달리 입력의 길이를 줄이더라도 추상 요약의 성능이 향상되지는 않았다. 그러나 원본 문서 대비 30% 수준까지 문서를 축약하여 입력하였음에도 불구하고 성능 감소가 크지 않은 것을 확인할 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 추출 요약과 추상 요약의 결합을 통해 두 가지 방법론이 가진 단점을 해소함으로써 문서 요약 성능의 개선을 시도하였다. 그러나 원본 문서를 의미 있는 방법으로 축약하더라도 추상 요약의 성능이 향상되지는 않았으며, 출력된 문장이 요약으로서 충분하지 않은 경우도 많았다. 분석 결과, 이는 문서 요약 분야에서 현재의 신경망 모델 및 데이터가 가진 한계로 보인다. 그러나 입력되는 문서의 크기를 크게 줄이더라도 요약 성능이 크게 하락되지는 않았다는 점으로 미루어 볼 때, 충분한 데이터를 확보하고 모델 구조를 개선한다면 의미 있는 결과를 보일 수 있을 것으로 생각된다.

향후 연구로 TextRank외에 다양한 방법을 통해 추출 요약 성능을 향상시킬 것이며, 더 많은 데이터를 수집하고, 개선된 형태의 신경망을 활용하는 것으로 추상 요약을 성능을 향상시킬 예정이다.

감사의 글

이 논문은 2016년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.R-20160906-004163, 빅데이터 자동 태깅 및 태그 기반 DaaS 시스템 개발)

참고문헌

[1] K. Knight and D. Marcu, Summarization beyond sentence extraction: A probabilistic approach to sentence compression, *Artificial Intelligence*, 139(1), pp. 91-107, 2002.

[2] R. Mihalcea, Language independent extractive summarization, *In Proceedings of the ACL 2005 on Interactive poster and demonstration sessions,* pp. 49-52, 2005.

[3] J. Tan, X. Wan and J. Xiao, Abstractive Document Summarization with a Graph-Based Attentional Neural Model. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* Vol. 1, pp. 1171-1181, 2017.

[4] P. Nema, M. Khapra, A. Laha and B. Ravindran, Diversity driven Attention Model for Query-based Abstractive Summarization. *arXiv preprint arXiv:1704.08300,* 2017.

[5] Q. Zhou, N. Yang, F. Wei and M. Zhou, Selective Encoding for Abstractive Sentence Summarization. *arXiv preprint arXiv:1704.07073,* 2017.

[6] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank citation ranking: Bringing order to the web. Stanford InfoLab. 1999.

[7] R. Mihalcea and P. Tarau, TextRank: Bringing Order into Text. *In EMNLP* Vol. 4, pp. 404-411, 2004.

[8] 최경호, 이창기. 복사 방법론과 입력 추가 구조를 이용한 End-to-End 한국어 문서요약. 정보과학회논문지, 44(5), pp. 503-509, 2017.

[9] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473,* 2014.

[10] Lin, Chin-Yew, "ROUGE: a Package for Automatic Evaluation of Summaries," *Proceedings of the Workshop on Text Summarization Branches Out*, 2004.