주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델

Title Generation Model for which Sequence-to-Sequence RNNs with Attention and Copying Mechanisms are used

저자 (Authors)

이현구, 김학수
Hyeon-gu Lee, Harksoo Kim

출처 (Source)

정보과학회논문지 44(7), 2017.7, 674-679(6 pages)
Journal of KIISE 44(7), 2017.7, 674-679(6 pages)

발행처 (Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07203561

APA Style

이현구, 김학수 (2017). 주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델. 정보과학회논문지, 44(7), 674-679

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 44, No. 7, pp. 674-679, 2017. 7
https://doi.org/10.5626/JOK.2017.44.7.674

주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델

(Title Generation Model for which Sequence-to-Sequence RNNs with Attention and Copying Mechanisms are used)

이 현 구 †　　김 학 수 ††
(Hyeon-gu Lee)　　(Harksoo Kim)

요 약 대용량의 텍스트 문서가 매일 만들어지는 빅데이터 환경에서 제목은 문서의 핵심 아이디어를 빠르게 집어내는데 매우 중요한 단서가 된다. 그러나 블로그 기사나 소셜 미디어 메시지와 같은 많은 종류의 문서들은 제목을 갖고 있지 않다. 본 논문에서는 주의집중 및 복사 작용을 가진 sequence-to-sequence 순환신경망을 사용한 제목 생성 모델을 제안한다. 제안 모델은 양방향 GRU(Gated Recurrent Unit) 네트워크에 기반 하여 입력 문장을 인코딩(encoding)하고, 입력 문장에서 자동 선별된 키워드와 함께 인코딩된 문장을 디코딩함으로써 제목 단어들을 생성한다. 93,631문서의 학습 데이터와 500문서의 평가 데이터를 가진 실험에서 주의집중 작용방법이 복사 작용방법보다 높은 어휘 일치율(ROUGE-1: 0.1935, ROUGE-2: 0.0364, ROUGE-L: 0.1555)을 보였고 사람이 정성평가한 지표는 복사 작용방법이 높은 성능을 보였다.

키워드: sequence-to-sequence 모델, 주의집중 작용, 복사 작용, 제목 생성, 순환신경망

Abstract In big-data environments wherein large amounts of text documents are produced daily, titles are very important clues that enable a prompt catching of the key ideas in documents; however, titles are absent for numerous document types such as blog articles and social-media messages. In this paper, a title-generation model for which sequence-to-sequence RNNs with attention and copying mechanisms are employed is proposed. For the proposed model, input sentences are encoded based on bi-directional GRU (gated recurrent unit) networks, and the title words are generated through a decoding of the encoded sentences with keywords that are automatically selected from the input sentences. Regarding the experiments with 93631 training-data documents and 500 test-data documents, the attention-mechanism performances are more effective (ROUGE-1: 0.1935, ROUGE-2: 0.0364, ROUGE-L: 0.1555) than those of the copying mechanism; in addition, the qualitative-evaluation radiative performance of the former is higher.

Keywords: sequence-to-sequence model, attention mechanism, copying mechanism, title generation, recurrent neural network

· 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.R-20160906-004163, 빅데이터 자동 태깅 및 태그 기반 DaaS 시스템 개발). 또한 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2016R1A2B4007732)
· 이 논문은 제43회 동계학술발표회에서 'Sequence to Sequence 모델과 키워드를 이용한 End-to-End 문서 제목 생성'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 강원대학교 컴퓨터정보통신공학과
nlphglee@kangwon.ac.kr
†† 종신회원 : 강원대학교 컴퓨터정보통신공학과 교수
(Kangwon Nat'l Univ.)
nlpdrkim@kangwon.ac.kr
(Corresponding author임)

논문접수 : 2017년 2월 1일
(Received 1 February 2017)
논문수정 : 2017년 5월 3일
(Revised 3 May 2017)
심사완료 : 2017년 5월 13일
(Accepted 13 May 2017)

정보과학회논문지 제44권 제7호(2017. 7)

1. 서 론

최근 많은 양의 텍스트 데이터가 등장함에 따라 원시 말뭉치를 통해 핵심 정보 요약, 키워드가 표현된 문장 생성 등 사람이 이해하기 쉽도록 정제하는 작업이 필요해지고 있다. 특히 문서의 내용을 잘 표현해 줄 수 있는 제목은 매우 중요한 정보로 데이터 정제에 필수적이다. 그러나 대부분의 SNS, 발췌한 내용 데이터는 키워드와 짧은 양의 텍스트만 존재하여 텍스트 요약을 통한 제목 생성은 적용하기 어려운 단점이 있다. 이러한 이유로 키워드와 요약이나 SNS 내용 같은 짧은 텍스트를 통해 제목을 자동으로 생성하는 문서 제목 생성 모델에 대한 연구가 필요하다. 본 논문에서는 기계번역 분야에서 많이 사용되는 sequence-to-sequence 모델[1]을 사용하여 단어 및 짧은 텍스트를 통해 단어의 추출 및 나열이 아닌 자연스럽고 의미있는 문장으로 제목을 생성하는 end-to-end 문서 제목 생성 모델을 제안한다.

2. 관련연구

기존의 제목 생성 방법은 문장으로 구성된 의미있는 제목 생성이 아닌 문서에 나타나는 중요 키워드를 추출하여 규칙기반으로 나열하는 요약 방식의 연구[2,3]로 진행되었다. sequence-to-sequence 모델은 입력 문장이 주어졌을 때 결과 문장이 출력되도록 하는 Recurrent Neural Network[4] 기반의 end-to-end 생성 모델로 입력의 형식과 출력의 형식이 다르게 사용되는 기계 번역(Machine Translation)분야[5]에서 많이 사용되고 있다. 최근 이러한 sequence-to-sequence 모델을 사용하여 중요 단어를 찾고 문법에 맞게 나열하는 추출, 검색 방법이 아닌 생성 방법으로 전체 문서를 입력하여 문서 추상화 벡터를 생성하고 그 벡터를 이용하여 요약문을 생성하는 문서 요약 시스템[6], 입력에서 나타나는 어휘가 출력에도 나타나는 요약 시스템의 특성을 반영하기 위해 sequence-to-sequence 모델에 복사 방법을 추가한 연구[7]도 진행되었다. 또한 사용자의 질문이나 대화 문장을 인식하여 적절한 답변을 생성하는 질의응답 시스템[8], 대화형 채팅 시스템[9]이 연구되고 있다. 본 논문에서는 sequence-to-sequence 모델을 응용하여 짧은 텍스트와 문서에 나타나는 핵심 키워드를 통해 제목을 자동으로 생성하는 end-to-end 문서 제목 생성 모델을 제안한다.

3. 문장과 키워드를 통한 제목 생성 모델

그림 1은 제안 모델의 전체 구조도를 보여준다. 그림 1에서 보는 것과 같이 제안 모델은 주어진 짧은 문장과 키워드를 sequence-to-sequence 모델을 통해 제목을 생성하는 end-to-end 문서 제목 생성 모델이다. sequence-to-sequence 모델은 요약이나 SNS글처럼 짧은 문장을 입력받아 형태소를 분석하고 인코딩하여 추상화된 정보가 저장되어 있는 문맥벡터(Context vector)를 생성하는 인코더 부분과 생성된 문맥벡터와 입력문장, 키워드를 통해 제목이 될 문장을 형태소 단위로 디코딩하고 형태소를 단어단위의 문장으로 복원하여 제목을 생성하는 디코더부분으로 구성된다.

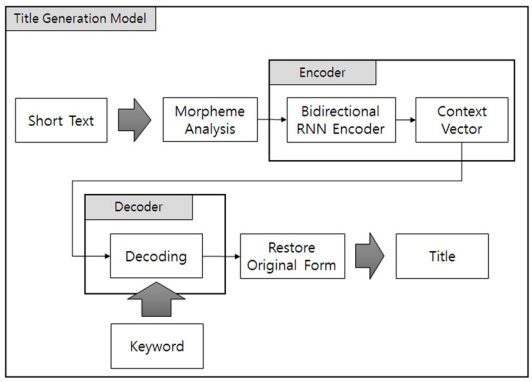

그림 1 제안 모델의 구조도
Fig. 1 Architecture of the proposed model

본 논문에서는 sequence-to-sequence의 응용 모델인 주의집중 작용(Attention mechanism)[10]과 복사작용(Copying mechanism)[11] 방법에 키워드를 반영할 수 있도록 디코더부분을 수정한 모델을 제안한다. 키워드를 디코더에 반영한 이유는 같은 문장이 들어와도 키워드에 따라 문장에서 집중할 내용이 바뀌어 제목이 달라질 수 있도록 하기 위함이다.

3.1 주의집중 작용을 활용한 제목 생성

본 논문에서는 문맥벡터를 전역으로 보기위한 주의집중 작용(Attention mechanism)이 적용된 sequence-to-sequence 모델을 사용한다. 주의집중 작용은 입력된 문장을 하나의 문맥벡터로 인코딩하여 문장 본연의 의미가 추상화로 인해 정보가 손실되는 기존의 sequence-to-sequence 모델의 문제점을 해결하기 위해 제안된 방법으로 인코딩 Recurrent Neural Network의 모든 은닉상태(hidden state)에 가중치를 반영하기 위한 주의집중 벡터를 연결하고 디코딩 때 입력을 전역적으로 참고하며 특정 은닉상태에 높은 가중치를 주기위한 방법이다. 그림 2는 주의집중 작용이 적용된 sequence-to-sequence 모델이다.

그림 2에서 보는 것과 같이 입력된 문장 X를 Bidirectional Recurrent Neural Network[12]의 Cell을 GRU(Gated Recurrent Unit)[13]로 설정한 인코더로 인코딩하여 주의집중 문맥벡터를 생성한다. 문맥 벡터만

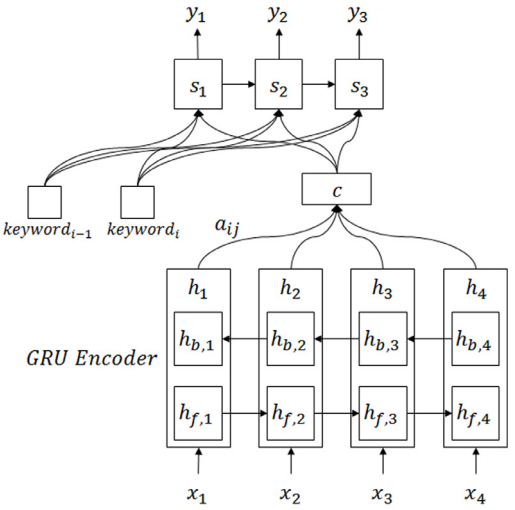

그림 2 주의집중 작용이 적용된 모델

Fig. 2 Attention of the sequence-to-sequence model

을 사용하여 디코딩하는 기존 주의집중 sequence-to-sequence 모델과 달리 생성된 주의집중 벡터에 키워드 벡터를 추가하여 문서 제목을 생성할 때 키워드에 따라 다른 문장이 생성되도록 한다. 이러한 주의집중 작용이 적용된 모델은 문장 전체를 추상화한 벡터를 사용하는 전통적인 모델과 달리 문장의 특정부분을 집중하여 주의집중이 없을 때보다 정보 손실이 적어 좋은 품질의 결과를 얻을 수 있다. 식 (1)은 그림 2의 주의집중 작용이 적용된 sequence-to-sequence 모델을 수식으로 표현한 내용이다.

$$
\begin{aligned}
h_{f,i} &= GRU(Emb_{source}(x_i), h_{f,i-1}) \\
h_{b,i} &= GRU(Emb_{source}(x_i), h_{b,i-1}) \\
h_i &= [h_{f,i}, h_{b,i}] \\
e_{ij} &= f(s_{i-1}, h_j) \\
a_{ij} &= \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \\
c_i &= \sum_{j=1}^{T_x} a_{ij} h_j \\
k &= [Emb(keyword_1) \cdots Emb(keyword_i)] \\
s_i &= f(s_{i-1}, Emb_{target}(y_{i-1}), k, c_i) \\
y_i &= argmax(g(Emb_{target}(y_{i-1}), s_i, c_i))
\end{aligned}
\tag{1}
$$

식 (1)에서 h_i는 i번째 단어의 은닉계층을 나타내며 정방향(forward)과 역방향(backward)의 은닉상태(hidden state)를 연결(concatenate)하여 사용한다. Emb_{source} 와 Emb_{target} 은 입력에서 사용되는 워드임베딩과 출력에서 사용되는 워드임베딩, e_{ij}는 이전 시간의 디코딩 은닉상태와 현재 입력의 은닉상태를 Feedforward Neural Network를 적용하여 출력된 결과값 ,a_{ij}는 i번째 출력을 디코딩할 때 사용되는 j번째 입력의 주의집중 가중치, c_i는 생성된 주의집중 문맥벡터, s_i는 디코딩 은닉상태, $keyword_i$는 입력된 키워드, y_i는 디코딩 결과 생성된 형태소이다.

3.2 복사 작용을 활용한 제목 생성

복사 작용(Copying mechanism)은 주의집중 작용에서 발생하는 문제인 미등록어, 등장확률이 낮은 고유명사 등이 출력에서 확률이 낮아져 출현하지 않는 문제를 해결하기 위해 제안된 방법으로 입력 문장에서 나타나는 어휘의 등장확률을 높여 출력에 복사되도록 하는 방법이다. 복사 작용은 단어를 생성하기 위한 주의집중 작용에 입력 단어의 출력확률을 높이기 위한 복사노드를 추가하여 주의집중 작용 방법으로 계산된 생성확률과 출력으로 나타나야할 단어가 입력에 존재할 시 복사노드를 통해 추가로 계산된 복사확률을 합하여 결과를 생성한다. 그림 3은 복사 작용이 적용된 sequence-to-sequence 모델이다.

복사 작용은 3.1절에서 언급한 주의집중 작용에 복사노드를 추가하고 출력 시 주의집중 작용으로 계산된 생성확률과 복사노드에서 계산된 복사확률을 합하여 생성될 단어를 결정한다. 주의집중 작용 모델에서 변경되거나 추가된 내용은 식 (2)와 같다.

$$
\begin{aligned}
copy_{ij} &= f_{copy}(Emb_{target}(y_{i-1}), s_{i-1}, h_j) \\
generate_i &= g(Emb_{target}(y_{i-1}), s_i, c_i)) \\
p(y_i|X) &= \begin{cases} \frac{1}{Z}\left(\exp(generate_i) + \sum_{j:x_j=y_i} \exp(copy_{ij})\right), & y_i \in X \\ \frac{\exp(generate_i)}{Z}, & otherwise \end{cases}
\end{aligned}
\tag{2}
$$

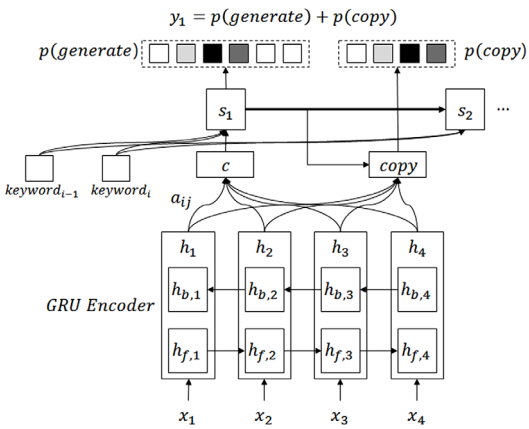

그림 3 복사 작용이 적용된 모델

Fig. 3 Copying of the sequence-to-sequence model

식 (2)에서 $copy_{ij}$는 복사확률을 계산하기 위한 복사 노드로 이전 시간의 출력 단어 임베딩, 이전 시간의 디코딩 은닉상태, 현재의 인코딩 은닉상태를 Feedforward Neural Network에 적용한 값이고, $generate_i$는 주의집중 작용 모델에서 나온 출력 확률, $p(y_i|X)$는 입력 X가 주어졌을 때 단어가 생성될 확률이다. $p(y_i|X)$를 구하는 식의 $y_i \in X$ 조건에서 보는 것과 같이 출력 후보 단어가 입력에 나타나면 복사노드를 통해 계산된 복사확률을 더해 입력에 나타난 단어의 확률을 상승시켜 출력에 반영될 수 있도록 하는 것을 알 수 있다.

4. 실험 및 평가

4.1 실험 준비

본 논문에서는 짧은 텍스트와 키워드를 통한 문서 제목 생성을 실험하기 위해 논문 데이터 94,131개를 사용한다. 논문 데이터는 논문의 주제를 나타내는 키워드, 논문의 요약 그리고 제목으로 구성되어있으며 학습 데이터 93,631개, 평가 데이터 500개를 무작위로 나누어 사용한다. 키워드가 존재하지 않는 데이터는 요약과 제목을 사용하여 $TF \cdot IDF$를 계산한 후 상위 n개의 어휘를 선택하여 사용한다. 본 논문에서 $TF \cdot IDF$를 통한 키워드 어휘 선택은 상위 3개를 사용하였다.

모델에서 사용할 워드 임베딩은 20GB의 뉴스 데이터와 논문 데이터를 Skip-gram Word2Vec[14] 방식으로 생성하였다. 워드 임베딩의 차원수는 50차원이다.

4.2 실험 평가

표 1은 제안 모델의 자동으로 측정한 성능을 보여준다. 성능 지표는 문서 요약 분야에서 많이 사용되는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[15]를 사용한다. ROUGE-N은 정답과 생성된 결과 사이의 N-gram 일치율을 나타내고 ROUGE-L은 LCS(Longest Common Subsequence)[16]를 사용하여 측정한 성능이다. 실험 결과 주의집중 작용방법이 복사 작용방법보다 좋은 성능을 보인다. 그러나 전체적으로 낮은 성능을 보이는데 sequence-to-sequence 모델은 짧은 문장을 생성하고자하는 경향이 있으며 ROUGE 지표는 문장이 짧을수록 불리한 문제가 있고 의미적인 정보는 반영하지 못한다. 즉, 어휘가 많이 일치하지 않아 잘못된 제목이 생성된 것이 아니라는 것을 알 수 있기 위해서는 의미적으로 유사한지를 파악해야한다. 따라서 사람이 직접 정성평가한 자료가 추가로 필요하다.

표 1 제안 방법의 성능 비교

Table 1 Performance comparison of the proposed method

	ROUGE-1	ROUGE-2	ROUGE-L
Attention	0.1935	0.0364	0.1555
Copying	0.1647	0.0332	0.1401

본 논문에서 정성평가는 1점부터 5점까지의 점수를 다음과 같은 기준으로 평가하였다.

1점 : 의미도 문법도 모두 이상한 문장

2점 : 의미는 다르나 문법적으로 이상 없는 문장

3점 : 주제는 비슷하나 세부적 의미가 다른 문장

4점 : 의미가 유사하나 아주 작은 차이를 보이는 문장

5점 : 의미가 정답과 완전히 일치하는 문장

표 2와 그림 4는 생성된 제목의 품질을 사람이 직접 정성평가한 내용을 보여준다.

표 2의 점수는 평가자가 평가 데이터를 통해 생성된 결과를 정답과 비교하여 부여한 점수들의 평균이고, 카파 상관계수(Cohen's Kappa Coefficient)[17]는 평가자들간의 통계적 일치도를 측정한 지표이다. 그림 4는 평가자들의 점수별 빈도의 평균을 나타낸 것이다. 표 2에서 보는 것과 같이 정성평가는 복사 작용이 주의집중 작용보다 좋은 성능을 보이는 것을 알 수 있다. 즉, 표 1에서 주의집중 작용이 복사 작용에 비해 높은 ROUGE 점수를 보여 어휘가 많이 일치되는 것을 알 수 있었지만 실제 의미와 문장 완성도 측면에서는 복사 작용이 더 좋은 것을 알 수 있다. 또한 그림 4의 분포를 보면 '의미도 문법도 모두 이상한 문장'인 1점의 분포가 주의집중 작용보다 복사 작용에서 줄어들고 의미적으로 고득점에 해당하는 '의미가 유사하나 아주 작은 차이를 보

표 2 정성평가의 성능

Table 2 Performance comparison of the qualitative evaluation

	Human 1	Human 2	Human 3
Attention	2.39	2.57	2.11
Copying	2.61	2.6	2.12
Kappa coefficient	0.4519		

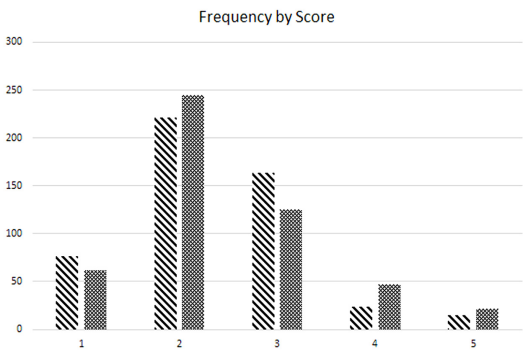

그림 4 점수별 빈도수

Fig. 4 Frequency according to the score

이는 문장', '의미가 정답과 완전히 일치하는 문장'의 빈도가 상승되는 것을 볼 수 있다. 즉 복사 작용이 문법, 의미적으로 모두 좋은 문장이 생성되는 것을 알 수 있었다.

4.3 실험 결과

본 논문에서는 제안한 복사작용을 활용한 제목 생성 모델을 통해 생성된 제목의 예시는 그림 5와 같다. 사용된 예시는 "국내 산성광산배수(AMD)의 수질 특성 분석" (오재일, 심연식 저 , 2003)[18] 논문이다.

그림 5에서 요약문(Abstract)은 모델에 입력으로 사용되는 짧은 문장이며 키워드(Keyword)는 제목 생성 시 사용되는 키워드, 생성된 제목(Generated title)은 모델을 통해 생성된 제목이다. 키워드를 "산성광산배수",

Abstract	
국내에서 배출되는 산성광산배수의 수질 특성을 파악하고자 기존 광산배수 수질 자료를 바탕으로 데이터베이스를 구축한 뒤, 광종별(석탄광산, 금속광산)로 분류하여 비교, 분석하는 작업을 수행하였다. 석탄광산 산성배수와 금속광산 산성배수의 수질 오염도(mg/L 단위)를 비교한 결과, 두 산성광산배수 모두에서 Fe, Al, Pb, Cd가 오염물질 배출허용기준 상의 모든 기준치를 상회하며, Mn, As는 청정지역의 배출허용기준을 상회하고 있어, F^-만이 기준치를 모두 만족하고 있는 것으로 나타났다. 산성광산배수 pH 조건에 따른 발생 이온 수질 특성을 pC/pH inder를 이용하여 분석한 결과, 석탄광산 산성배수에서는 Fe=0.99, Al=0.88, Mn=1.16, Ca=0.79, Mg=0.79, $SO_4{^{2-}}=0.61$ 등으로 나타났으며 금속광산 산성배수에서는 Fe=1.14, Al=0.99, Mn=1.28, $SO_4{^{2-}}=0.71$ 등으로 나타났다. 산성광산배수 내 이온들의 charge balance를 이용한 분석에서 (전체 양이온)/(전체 음이온)의 비율은 석탄광산 산성배수와 금속광산 산성배수에서 각각 94% 75%로 나타나 석탄광산 산성배수에서 상대적으로 많은 수질 항목을 측정하고 있음을 알 수 있었으며, 이러한 주된 차이는 금속광산 산성배수에서의 Ca과 Mg에 대한 자료의 미비가 주된 원인으로 도출되었다(양이온의 eq/L 농도 비중은 석탄광산 산성배수가 Ca>Mg>Al>Fe>H>Mn 순으로 나타났고, 금속광산 산성배수는 Fe>Al>H>Mn 순으로 나타남).	
Correct Title	국내 산성광산배수(AMD)의 수질 특성 분석
Keyword 1	산성광산배수, 수질
Generated Title 1	우리나라 산의 물 성 분석에 관한 연구
Keyword 2	산성광산배수, 석탄광산
Generated Title 2	석탄광산에서 성질 특성 분석

그림 5 생성된 제목의 예
Fig. 5 Example of a generated title

표 3 예제의 성능 평가
Table 3 Performance evaluation of an example

Automatic evaluation			
	ROUGE-1	ROUGE-2	ROUGE-L
Generated Title 1	0.3500	0.2222	0.3921
Generated Title 2	0.4500	0.2963	0.5000
Qualitative evaluation			
	Human 1	Human 2	Human 3
Generated Title 1	5	5	5
Generated Title 2	3	3	3

"수질"을 사용하면 "우리나라 산의 물 성 분석에 관한 연구"라는 제목이 생성되고, "산성광산배수", "석탄광산"을 사용하면 "석탄광산에서 성질 특성 분석"이라는 다른 제목이 생성된다. 즉, 입력된 키워드에 따라 제목이 다르게 생성되는 것을 알 수 있다. 표 3은 그림 5의 예제를 자동평가와 정성평가를 한 결과를 나타낸다.

표 3에서 첫 번째 문장(Generated Title 1)은 두 번째 문장(Generated Title 2)보다 자동평가 성능이 낮게 나왔다. 즉 정답문장과의 어휘 일치도가 두 번째 문장이 더욱 높은 것을 알 수 있다. 그러나 의미적 일치도를 판단하는 정성평가에서 평가자 모두 첫 번째 문장을 고득점에 해당하는 5점을 부여하고 두 번째 문장에는 중간 점수인 3점을 부여하여 첫 번째 문장이 어휘 일치도는 떨어지나 의미적으로 더욱 좋은 문장이라는 것을 알 수 있다. 또한 두 번째 문장은 다른 키워드가 입력되어 결과가 달라진 것이며 키워드는 생성된 문장에 나타나는 것으로 보아 키워드 선택이 중요하다는 것을 알 수 있다.

5. 결 론

본 논문에서는 sequence-to-sequence 모델의 응용인 주의집중 작용방법과 복사 작용방법에 키워드를 반영할 수 있게 수정한 디코더를 사용하여 짧은 텍스트와 키워드를 통한 end-to-end 문서 제목 생성 모델을 제안하였다. 실험 결과 주어진 키워드에 따라 집중되는 어휘가 달라 생성되는 제목이 바뀌는 것을 알 수 있었고 정답과의 어휘 일치도를 보는 ROUGE는 주의집중 작용방법이, 의미적 정보와 문장의 구조적 정보를 확인하기 위한 정성평가는 복사 작용방법이 높은 성능을 보였다.

References

[1] L. Sutskever, O. Vinyals, and Q. V. Le, "sequence-to-sequence learning with neural networks," Advances in neural information processing systems, pp. 3104-3112, 2014.

[2] K. Han and Y. Ahn, "Automatic labeling of Korean document clusters created by LDA," *Proc. of the Korea Computer Congress 2013*, pp. 616-618, 2013.

[3] T. Kim and S. Myaeng, "Automatic Naming of Document Clusters by Using their Hierarchical Structure," *Proc. of the 13th Annual Conference on Human and Cognitive Language Technology*, pp. 163-170, 2001.
[4] F. J. Pineda, "Generalization of back-propagation to recurrent neural networks," *Physical review letters 59.19:* 2229, 1987.
[5] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," arXiv preprint arXiv:1406.1078, 2014.
[6] A. M. Rush, S. Chopra and J. Weston, "A neural attention model for abstractive sentence summarization," arXiv preprint arXiv:1509.00685, 2015.
[7] K. Choi, C. Lee, "End-to-end Document Summarization using Copy Mechanism and Input Feeding," *Proc. of the 28th Annual Conference on Human & Cognitive Language Technology*, pp. 56-61, 2016.
[8] D. Hewlett, A. Lacoste, L. Jones, I. Polosukhin, A. Fandrianto, J. Han, M. Kelcey and D. Berthelot, "WikiReading A Novel Large-scale Language Understanding Task over Wikipedia," arXiv preprint arXiv:1608.03542, 2016.
[9] O. Vinyals, Q. Le, "A neural conversational model," arXiv preprint arXiv:1506.05869, 2015.
[10] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.
[11] J. Gu, Z. Lu, H. Li and V. O. K. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," arXiv preprint arXiv:1603.06393, 2016.
[12] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing 45.11*, pp. 2673-2681, 1997.
[13] K. Cho, B. V. Merriënboer, D. Bahdanau and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv preprint arXiv:1409.1259, 2014.
[14] T. Kikolov, K. Chen, G. Corrado,and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
[15] C. Y. Lin, "ROUGE A Package for Automatic Evaluation of Summaries," *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8, 2004.
[16] M. Paterson and V. Dančík, "Longest common subsequences," *International Symposium on Mathematical Foundations of Computer Science*, 1994.
[17] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, Vol. 70, No. 4, pp. 213-220, 1968.
[18] J. Oh and Y. Shim, "Statisical Analysis of Water Quality of Domestic Acid Mine Drainage(AMD)," *Journal of the Korean Society of Civil Engineers B*, 23.6B, pp. 587-596, 2003.

이 현 구

2016년 강원대학교 컴퓨터정보통신공학과 학석사연계과정 학사. 2016년 강원대학교 컴퓨터정보통신공학과 학석사연계과정 석사. 2016년~현재 강원대학교 컴퓨터정보통신공학과 박사과정. 관심분야는 자연어처리, 정보검색, 정보추출, 질의응답시스템

김 학 수

1996년 건국대학교 전자계산학과 학사 1998년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년 University of Massachusets, Amherst 박사후연구과정. 2005년 한국전자통신연구원 선임연구원. 2006년~현재 강원대학교 컴퓨터정보통신공학전공 교수. 관심분야는 자연어처리, 대화모델링, 정보검색, 정보추출, 질의응답시스템