딥러닝과 Maximal Marginal Relevance를 이용한 2단계 문서 요약

전재원○, 황현선, 이창기

강원대학교 컴퓨터과학과
{jwj, hhs4322, leeck}@kangwon.ac.kr

Two-step Document Summarization using Deep Learning and Maximal Marginal Relevance

Jaewon Jeon○, Hyunsun Hwang, Changki Lee
Kangwon National University Dept. of Computer Science

요 약

문서 요약은 길이가 긴 원본 문서의 의미는 유지한 채 원본보다 짧은 문서나 문장을 생성하는 자연어 처리 태스크이다. 본 논문에서는 Maximal Marginal Relevance(MMR)를 이용한 sequence-to-sequence 문장 추출 모델을 이용하여 의미가 중복되는 문장을 최소화하는 문장을 추출하고 추출된 문장을 sequence-to-sequence 모델을 통해 요약문을 생성하는 2단계 문서 요약 모델을 제안한다. 실험 결과 MMR을 활용하지 않았던 기존의 방법론보다 Rouge 성능이 향상되었다.

주제어: Maximal Marginal Relevance (MMR), 추출 요약, 생성 요약

1. 서론

문서 요약이란 원본 문서의 의미는 그대로 유지 하면서 주요한 정보가 담긴 원본 문서보다 짧은 문서나 문장으로 출력하는 것을 말한다[1].

문서 요약은 크게 추출 요약(Extractive Summarization)과 생성 요약(Abstractive Summarization) 두 가지 방법으로 나뉜다. 추출 요약[2]은 원본 문서에 있는 문장들 중에서 문서의 주제를 담고 있는 주요한 문장 몇개를 선택해서 출력해 요약문으로 사용하는 방법이고, 생성 요약[3]은 문서 전체의 의미를 이해하고 이에 맞는 문장을 생성하여 요약문으로 사용하는 방법이다. 추출 요약 방법보다 생성 요약 방법이 난이도가 더 높은 작업에 속하기 때문에 추출 요약 방법이 주로 연구되어 왔다. 그러나 최근 심층 신경망을 이용한 학습(Deep Learning)의 연구가 활발하게 진행되면서 품질이 높은 문장을 생성하는 방법에 대한 연구가 진행되면서 생성 요약 방법이 많이 연구되고 있다.

추출 요약은 원본 문서의 모든 문장들 중에서 중요하지 않은 문장을 제외하고 문서의 주제가 담긴 주요한 문장들만 선택하여 추출하므로 생성 요약을 할 때 기존 문서 전체가 아닌 추출 요약을 통해 추출된 문장들을 입력으로 사용하면 더 좋은 요약문을 생성할 수 있다[4].

본 논문에서는 [4]의 추출 요약과 생성 요약을 결합한 2단계 문서 요약 모델에서 사용된 추출 모델에 Maximal Marginal Relevance(MMR)[5]를 적용하여 중복된 문장을 최소화하여 추출된 문장들에 원본 문서의 다양한 내용이 담길 수 있도록 하고 이를 생성 요약의 입력으로 사용하는 2단계 문서 요약 모델을 제안한다.

2. 관련 연구

2.1 추출 요약

추출 요약은 원본 문서에 있는 문장들 중에서 문서의 주제를 담고 있는 주요한 문장 몇 개를 선택해서 출력해 요약문으로 사용하는 방법이다.

문장을 추출하는 방법은 단어의 출현 빈도수를 기반으로 추출하는 방법, 그래프를 이용한 추출 방법, 문장 간의 유사도를 기반으로 추출하는 방법이 있다.

단어의 출현 빈도수를 기반으로 추출하는 방법[6]은 문장 내에서 명사 등의 주요한 단어가 얼마나 들어 있는지 계산해서 수치를 기반으로 추출할 문장을 결정한다.

그래프를 이용한 추출 방법[7]은 문장들 간의 연결 그래프를 만든 후에 각 그래프들을 그룹화해서 그룹별로 주요 문장이 무엇인지 결정하여 주요 문장만 추출하는 방법이다.

문장 간의 유사도를 기반으로 추출하는 방법[8]은 각 문장의 자질들을 코사인 유사도(Cosine Similarity) 등의 방법으로 문장 간의 유사도를 계산하여 추출할 문장을 결정하는 방법이다.

2.2 생성 요약

생성 요약은 원본 문서의 전체 내용을 이해하고 이를 바탕으로 요약문을 새로 생성해 내는 요약 방법이다. 주로 sequence-to-sequence 모델을 통하여 요약문을 생성하는 방법을 사용한다. 그러나 문서 요약은 입력 문장의 길이가 매우 길다는 특징이 있고, 또한 출력 문장에 등장하는 다수의 단어가 입력 문장에 존재한다는 특징이 있다. 이러한 특징 때문에 생성 요약에서는 주로 주의 집

중 방법(Attention Mechanism)과 복사 방법론(Copy Mechanism) 등이 함께 사용된다.

2.3 Maximal Marginal Relevance

[5]에서는 제안한 Maximal Marginal Relevance(MMR)는 문서 주제와 관련성이 높은 문장을 선택하면서도 이전에 선택된 문장들에 대한 유사도가 낮은 문장을 선택하여 선택된 문장들간의 정보 중복 문제를 해결할 수 있는 기법으로 다음과 같이 정의된다.

$$\text{Arg}\max_{S_i \epsilon D}\left[\alpha Sim_1(S_i, Q) - (1-\alpha)\max_{S_j \epsilon R} Sim_2(S_i, S_j)\right] \quad (1)$$

D 는 문서를 이루고 있는 문장들의 전체 집합을 의미하고 S는 문장, R은 D 의 부분집합으로 이미 선택된 문장들의 집합을 의미한다. Sim_1과 Sim_2는 유사도 함수를 의미하며 서로 같은 함수를 사용하거나 다른 함수를 사용할 수 있다.

3. 2단계 문서 요약 모델

본 논문에서는 기존 순환 신경망 모델이나 sequence-to-sequence 모델을 활용하여 문장을 추출하는 추출 모델에 MMR기법을 적용하여 추출된 문장들의 중복성을 최소화하고 추출된 문장을 이용하여 copy mechanism 모델 또는 transformer 모델을 통해 요약문을 생성하는 2단계 파이프라인 방식의 문서 요약 모델을 제안한다.

3.1 MMR을 활용한 문장 추출

추출 요약은 문서를 이루고 있는 여러 개의 문장들 중에서 주요한 문장 몇 개를 선택해서 선택된 문장들을 요약문으로 사용하는 것이 핵심 목표이다. 본 논문에서는 [9]의 방법론을 따라, 대상 문서의 각 문장들에 추출할 문장이면 1, 추출하지 않아도 될 문장이면 0으로 레이블을 달아서 추출할 문장을 선택하는 방법을 사용한다.

본 논문에서 사용된 추출 모델은 각 문장의 추출 스코어를 다음과 같이 계산한다. 입력 문장을 3가지 인코더 중에 하나를 선택해 인코딩하고 문장 추출 모델을 거쳐 1차 스코어를 계산하고, 이 스코어에 MMR을 적용하여 최종 추출 스코어를 계산한다.

[9]의 문장 추출 모델에서 사용한 인코더는 3가지로 평균 인코더(Average Embedding), 순환 신경망(Recurrent Neural Network; RNN) 인코더, 합성곱 신경망(Convolution Neural Network; CNN) 인코더가 사용되었다.

평균 인코더는 단순히 입력 문장의 단어 임베딩들의 평균을 계산해서 나온 결과를 문장 임베딩으로 사용하는 방식이다. 문장 임베딩 h는 문장 s의 각 단어 임베딩 w_i의 평균으로 다음과 같은 수식으로 계산한다.

$$\mathrm{h} = \frac{1}{|s|}\sum_{i=1}^{|s|} w_i \quad (2)$$

순환 신경망 인코더는 GRU(Gated Recurrent Unit)를 사용한 양방향 순환 신경망(Bi-directional RNN)에 입력 문장을 넣고 마지막 hidden state를 연결(Concatenation)하여 문장 임베딩으로 사용한다.

합성곱 신경망 인코더는 입력 문장을 넣고 max pooling을 거친 모든 합성곱 필터를 통해 나온 최종 출력을 연결(Concatenation)하여 문장 임베딩으로 사용한다.

3가지 인코더 중에 선택된 인코더를 통해 나온 문장 임베딩들은 문장 추출 모델에 입력으로 사용된다. 문장 추출 모델은 순환 신경망 추출 모델, sequence-to-sequence 추출 모델을 사용한다.

순환 신경망 추출 모델은 GRU를 사용한 양방향 순환 신경망을 사용한다. 문장 임베딩을 입력 받아서 나온 각 forward 및 backward 출력을 다층 퍼셉트론에 통과시켜 1차 스코어를 계산한다.

Sequence-to-sequence 추출 모델은 순환 신경망으로 GRU를 사용하였고, 장거리 종속성 문제를 해결하기 위해 주의집중 방법론(Attention Mechanism)을 적용한 모델을 사용했다. 이 모델은 입력된 문장 임베딩을 쿼리 벡터로 변환하여 attention의 출력과 쿼리 벡터를 연결(Concatenation)하여 다층 퍼셉트론에 통과시켜 1차 스코어를 계산한다.

문장에 대한 최종 스코어는 1차 스코어에 MMR을 적용해서 계산한다. MMR은 아래 수식과 같이 적용하였다.

$$\max_{S_i \epsilon D}\left[\boldsymbol{\alpha} \cdot score(S_i) - (1-\alpha)\max_{E_k \epsilon R}[sim(S_i, E_k)]\right] \quad (3)$$

D 는 원본 문서의 문장들 중 아직 추출되지 않은 문장들의 집합이고, R 은 이미 추출된 문장들의 집합이다. score(S) 는 문장 S 에 대한 1차 스코어를 나타내고 sim(S, E)는 문장 S와 문장 E의 유사도를 계산하는 함수이다. 유사도 함수는 코사인 유사도(Cosine Similarity)와 내적 유사도(Dot Product Similarity)를 사용했다.

코사인 유사도는 두 벡터 간의 각도를 이용하여 유사도를 계산한다. 계산 결과는 -1이상 1이하의 값이 나오며 1에 가까울수록 두 벡터는 유사하다.

$$\boldsymbol{sim_{cos}(S_1, S_2)} = \frac{\boldsymbol{S_1 \cdot S_2}}{\|\boldsymbol{S_1}\|\,\|\boldsymbol{S_3}\|} \quad (4)$$

내적 유사도는 두 벡터의 내적 값을 유사도로 사용한 방법이다.

$$\boldsymbol{sim_{dot}(S_1, S_2) = S_1 \cdot S_2} \quad (5)$$

3.2 요약문 생성

Sequence-to-sequence 모델만을 이용한 문서 요약은 원본 문서 전체를 입력으로 받아 해당 문서의 내용을 이해하고 요약문을 생성한다. 그러나 입력 문장이 길어질수록 장거리 종속성 문제가 발생하여 입력 문장의 품질이 떨어지는 문제가 생긴다.

본 논문에서는 이 문제를 해결하기 위해 sequence-to-sequence 요약 생성 모델에 입력 문장을 문서 전체가 아닌 3.1절에서 추출한 문장들을 사용한다.

Sequence-to-sequence 모델은 copy mechanism을 적용한 RNN모델과 transformer 모델을 사용한다.

4. 실험 및 성능

본 논문에서는 한국어 문서 요약 데이터로 인터넷 사

이트 인사이트[1]의 뉴스 기사를 활용하여 수집한 뉴스 데이터를 사용했다. 인사이트 뉴스 기사의 경우 제목, 한 줄 요약문, 본문의 형태로 구성되어 있다. 원본 문서로는 뉴스 기사 본문을 활용했고 한 줄 요약문이 있는 경우 정답 요약 문장으로 사용하였고, 없는 경우 제목을 정답 요약 문장으로 사용하였다. 수집한 문서들 중 109,558개의 문서를 학습 셋, 3,786개의 문서를 개발 셋, 3,705개의 문서를 평가 셋으로 사용하여 실험하였다.

뉴스 데이터들은 모두 형태소 분석기[10]를 사용하여 형태소 단위로 이루어진 문장으로 만들어 사용하였다(표 1).

표 1 데이터 형식

원본 기사	2일 서울 왕십리 CGV에서 영화 '널 기다리며'의 언론 배급 시사회에는 모홍진 감독과 김성오, 심은경, 윤제문이 참석했다.
변형 기사	2/SN 일/NNB 서울/NNP 왕십리/NNP CGV/SL 에서/JKB 영화/NNG '/SS 너/NP 르/JKO 기다리/VV 며/EC '/SS 의/JKG 언론/NNG 배급/NNG 시사회/NNG 에/JKB 는/JX 모홍진/NNP 감독/NNG 과/JC 김성오/NNP ,/SP 심은경/NNP ,/SP 윤제문/NNP 이/JKS 참석/NNG 하/XSV 았/EP 다/EF ./SF

문장을 추출할 때 1차 스코어를 결정하는 모델은 추출 문장의 rouge 스코어를 기준으로 가장 성능이 좋았던 평균 임베딩 인코더와 sequence-to-sequence 문장 추출 모델을 사용하였다.

요약문을 생성하는 생성 요약 모델은 단어 임베딩으로 200차원을 사용하여 학습을 진행하였고 OpenNMT[2]에 구현된 모델을 사용했다. Copy mechanism을 사용한 RNN 모델의 경우, 순환 신경망의 hidden state 크기를 256으로 하였고 bi-directional LSTM을 사용하였다. Transformer의 레이어는 8개를 사용하였다.

MMR에서 유사도 함수의 경우 코사인 유사도와 내적 유사도를 모두 실험했으며 α값을 0.4부터 0.7까지 조절하며 실험하였다. 코사인 유사도의 경우 α값이 0.5일 때가 성능이 제일 좋았고 내적 유사도의 경우 α값이 0.7일 일 때 성능에 제일 좋았다.

Baseline 모델은 추출 모델을 통해 문장을 추출하지 않고 문서 전체를 입력 문장으로 사용하여 Copy Mechanism을 사용한 RNN모델이나 transformer를 통해 생성한 요약문으로 성능을 측정하였다.

각 모델 별 rouge score 성능은 아래 표2에서 확인할 수 있다. 표 2를 통해, 기존 생성 요약 모델보다 MMR을 활용하여 중복을 최소화한 2단계 요약 모델이 더 높은 성능을 보임을 알 수 있다.

표 2 모델에 따른 Rouge 성능표

모델	Rouge - 1	Rouge - 2	Rouge - L
LSTM(Attn. + Copy)	38.11	15.42	32.31
추출 요약 + LSTM(Attn. + Copy)	38.69	15.59	32.74
추출 요약(Cos 유사도 MMR) + LSTM(Attn. + Copy)	**40.83**	**16.38**	**34.22**
추출 요약(내적 유사도 MMR) + LSTM(Attn. + Copy)	40.68	16.27	33.99
Transformer	37.22	14.03	31.04
추출 요약 +Transformer	37.40	14.44	31.62
추출 요약(Cos 유사도 MMR) +Transformer	39.26	14.98	33.18
추출 요약(내적 유사도 MMR) +Transformer	39.76	14.86	32.88

5. 결론 및 연구 방향

본 논문에서는 문장 추출 모델에 MMR을 적용하여 중복된 문장을 최소화하여 주요한 문장을 추출하고 이 추출된 문장을 copy mechanism을 적용한 RNN모델과 transformer를 이용하여 요약문을 생성하는 2단계 문서 요약 모델을 제안하였다. 기존 생성 요약 모델보다 MMR을 활용하여 중복을 최소화한 2단계 요약 모델이 더 높은 성능을 보임을 알 수 있다.

감사의 글

이 논문은 SK주식회사C&C의 지원을 받아 연구되었음.

참고문헌

[1] 장동현, 맹성현, “자동 요약 시스템” 정보과학회지. 제 15권. 제10호 pp. 42-49. 1997.

[1] https://insight.co.kr/

[2] https://github.com/OpenNMT/OpenNMT-py

문서	지난달 31일 한 유튜브 채널에 올라온 '수박 껍질만 벗기는 방법'이라는 영상이 큰 인기를 끌고 있다. '수박 껍질만 벗기는 방법'은 의외로 간단하다. 먼저 비슷한 크기의 수박 두 통을 준비한 뒤 한 통의 껍질을 칼로 슥슥 발라낸다. 연두색 부분이 안 보일 정도도 꼼꼼하게 발라낸 후 수세미로 살살 밀어주면서 매끄럽게 만든다. 그리고 남은 수박 한 통을 반으로 잘라 과육을 파낸 후, 앞의 수박과 원래 하나였던 것(?)처럼 합쳐 주기만 하면 끝이다. 그러면 앞의 과정을 보지 못한 사람들은 수박의 알맹이와 껍질을 쉽게 분리할 수 있게 된다. 비록 반전이 숨어있긴 하지만 누리꾼들은 "공룡알 모양의 수박이다"라면서 재밌어 하는 반응이다.	
추출 모델 추출 문장	지난달 31일 한 유튜브 채널에 올라온 '수박 껍질만 벗기는 방법'이라는 영상이 큰 인기를 끌고 있다. 비록 반전이 숨어있긴 하지만 누리꾼들은 "공룡알 모양의 수박이다"라면서 재미있어 하는 반응이다. '수박 껍질만 벗기는 방법'은 의외로 간단하다. 그러면 앞의 과정을 보지 못한 사람들은 수박의 알맹이와 껍질을 쉽게 분리할 수 있게 된다.	
추출 모델 + 내적 추출 문장	지난달 31일 한 유튜브 채널에 올라온 '수박 껍질만 벗기는 방법'이라는 영상이 큰 인기를 끌고 있다. '수박 껍질만 벗기는 방법'은 의외로 간단하다. 먼저 비슷한 크기의 수박 두 통을 준비한 뒤 한 통의 껍질을 칼로 슥슥 발라낸다. 그러면 앞의 과정을 보지 못한 사람들은 수박의 알맹이와 껍질을 쉽게 분리할 수 있게 된다.	
추출 모델 + Cos 추출 문장	지난달 31일 한 유튜브 채널에 올라온 '수박 껍질만 벗기는 방법'이라는 영상이 큰 인기를 끌고 있다. '수박 껍질만 벗기는 방법'은 의외로 간단하다. 연두색 부분이 안 보일 정도도 꼼꼼하게 발라낸 후 수세미로 살살 밀어주면서 매끄럽게 만든다. 그러면 앞의 과정을 보지 못한 사람들은 수박의 알맹이와 껍질을 쉽게 분리할 수 있게 된다.	
정답 문장	"수박의 '알맹이'만 그대로 분리하는 방법이 있을까?"	
예측 문장	LSTM (Attn. + Copy)	수박 껍질만 벗길 수 있는 방법이 등장해 누리꾼들의 관심을 끌었던 것으로 알려졌다.
	LSTM (Attn. + Copy) + 내적	'수박 껍질만 벗기는 방법' 이 누리꾼들 사이에서 화제를 모으고 있는 가운데, 많은 이들을 하게한다.
	LSTM (Attn. + Copy) + Cos	'수박 껍질만 벗기는 방법이 공개되어 누리꾼들의 이목을 집중시키고 있다.

표 3 추출된 문장 및 예측한 문장 결과

[2] Jae-Min Yoon, You-Jin Chung, Jong-Hyeok Lee, "Automatic Extractive Summarization of Newspaper Articles using Activation Degree of 5W1H," Journal of KIISE, SA, Vol. 31, No. 4, pp. 505-515, 2004.

[3] Sumit Chopra, Michael Auli, Alexander M Rush, and SEAS Harvard. Abstractive sentence summarization with attentive recurrent neural networks. Proceedings of NAACL-HLT16, pp. 93-98, 2016.

[4] 전재원, 황현선, 이창기. 추출 요약과 생성 요약을 결합한 2단계 문서 요약.

[5] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the 21st ACM-SIGIR InternationalConference on Research and Development in Information Retrieval.

[6] K.McKeown, J.Robin, and K.Kukich, “Generating Concise Natural Language Summaries”, Advances in Automatic Text Summarization, MIT press, pp.233-264, 1999.

[7] G.Erkan, and D.R. Radev, “LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization”, Journal of Artificial Intelligence Research, vol.22, no.2004, pp.457-479, 2004.

[8] Takaharu Takeda, Atsuhiro Takasu, “UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing”, International Conference on Digital Libraries, pp.438-439, 2007.

[9] Chris Kedzie, Kathleen McKeown and Hal Daume III, "Content Selection in Deep Learning Models of Summarization", 2018

[10] 이창기. Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델. 정보과학회논문지: 소프트웨어 및 응용, 제40권 제12호, 2013.12, 826~832