KIPS Trans. Softw. and Data Eng.
Vol.8, No.8 pp.343~348 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.8.343

Single Document Extractive Summarization Based on Deep Neural Networks Using Linguistic Analysis Features

Gyoung Ho Lee† · Kong Joo Lee††

ABSTRACT

In recent years, extractive summarization systems based on end-to-end deep learning models have become popular. These systems do not require human-crafted features and adopt data-driven approaches. However, previous related studies have shown that linguistic analysis features such as part-of-speeches, named entities and word's frequencies are useful for extracting important sentences from a document to generate a summary. In this paper, we propose an extractive summarization system based on deep neural networks using conventional linguistic analysis features. In order to prove the usefulness of the linguistic analysis features, we compare the models with and without those features. The experimental results show that the model with the linguistic analysis features improves the Rouge-2 F1 score by 0.5 points compared to the model without those features.

Keywords : Single Document Summarization, Extractive Summarization, Linguistic Analysis Features, Deep Neural Networks

언어 분석 자질을 활용한 인공신경망 기반의 단일 문서 추출 요약

이 경 호† · 이 공 주††

요 약

최근의 문서요약 시스템은 인공신경망을 이용한 End-to-End 방식이 주류를 이루고 있다. 이러한 시스템은 인간의 자질 추출 과정이 필요 없으며 데이터 중심의 접근 방법을 채택한다. 그러나 기존의 관련 연구들은 품사 정보, 개체명 정보, 단어의 빈도 정보와 같은 언어 분석 자질이 중요 문장을 선택하여 요약을 작성하는데 유용함을 보여왔다. 본 연구에서는 기존의 언어 분석 자질을 활용하여 인공신경망을 기반으로 한 단일 문서의 추출 요약 시스템을 제안한다. 언어 분석 자질의 유용성을 보이기 위해 자질을 사용하는 모델과 사용하지 않는 모델을 비교하였다. 실험 결과 자질을 사용하는 모델이 그렇지 않은 모델에 비해 약 0.5점의 Rouge-2 F1점수 향상을 보였다.

키워드 : 단일 문서 요약, 추출 요약, 언어 분석 자질, 인공신경망

1. 서 론

문서요약(Text Summarization)은 주어진 문서의 주요 내용을 담은 짧고 간결한 글을 생성하는 것이다. 정보매체의 다양화로 개인이 소화해야 하는 정보의 양이 증가하고, 이를 돕기 위한 문서요약 기술도 계속 발달 되어왔다.

요약을 만드는 방법은 크게 생성 요약(Abstractive Summarization)과 추출 요약(Extractive Summarization)으로 나눌 수 있다. 생성 요약은 문서의 핵심 내용을 담은 새로운 문장을 생성하여 요약을 만드는 방법이다[1]. 생성 요약과 함께, 문서 요약의 또다른 방법은 추출 요약이다. 추출 요약은 문서의 구성 문장들 중, 문서를 대표하는 일부 문장을 선택하여 이를 요약으로 사용하는 방법이다[2]. 사람이 작성한 문장을 이용해 요약을 만들기 때문에 문법적으로 온전한 문장들로 요약을 생성할 수 있다. 그렇기 때문에 생성 요약에 비해 상대적으로 적은 노력으로 높은 수준의 요약을 생성할 수 있다. 본 연구는 단일 문서에 대한 추출 요약을 다룬다.

최근 자연언어처리 연구의 큰 흐름은 인공신경망을 이용한 End-to-End 방식의 학습이다. 이는 자질들의 복잡한 조합을 사람이 고안하고 이를 이용하여 문제를 해결하던 기존의 연구 방법과 달리, 문제 해결에 적합한 인공신경망의 입력과 출력을 설계하고 대량의 학습데이터를 입력으로 사용하여 모델을 학습시키는 방법이다[3]. 이러한 흐름은 추출 문서요약에도 적용되고 있다. 추출 요약의 최근 연구 성과들은 자질 설계의 과정 없이 문서를 구성하는 단어의 표층정보만을 이용하여 문장을 표현하고, 각 문장과 전체 문서의 관계를 이용하여 중요 문장을 선택하는 방식으로 진행되었다[2, 4]. 하지만 이러한 흐름 이전에도 문서요약에 대한 연구들은 진행되어 왔다[5, 6]. 이들 연구에서는 단어의 표층정보 뿐만 아니라 품사 정보, 문서 군집에서의 단어빈도수와 역문서빈도, 문장 안에서의 의존구문관계 등 언어 분석을 통한 다양한 자질을 추출하고 이들을 조합하여 문서 요약을 수행하였다. 이러한 자질들은 전처리에 대한 비용이 필요하고 이때 발생하는 오류가 다음 단계로 전파되는 위험이 있지만, 이러한 단점에도 불구하고 기존의 데이터와 연구에서 의미 있는 결과를 내어왔다. 본 연구에서는 이러한 기존의 언어 분석 자질을 최신의

† 준 회 원 : 충남대학교 전자전파정보통신공학과 박사과정
†† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : April 15, 2019
Accepted : May 19, 2019
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

End-to-End 방식의 인공신경망에 함께 결합하여 단일 문서 추출 요약 시스템을 구축해 보고자 한다.

일반적으로 추출 문서요약은 문서를 구성하는 주요 단어를 기반으로 특징적인(salience) 문장을 찾는 문제로 다루어진다. 이는 전통적인 문서요약 방법[7, 8]에서 뿐만 아니라 인공신경망을 이용한 최근의 추출 문서요약 방법[4]에도 적용되고 있는 개념이다. 최신의 인공신경망 기반 추출 문서요약 모델은 입력 문서를 구성하는 단어 정보만을 자질로 사용하여 모델을 설계하고 대량의 학습데이터를 이용하여 인공신경망 모델의 연결 구조 안에서 문서의 중요 단어와 중요 문장이 자동적으로 학습되도록 한다.

간단한 언어 분석을 통하면 문서 내에서 단어의 역할이나 문장 내부에서 단어 사이의 구조적 관계, 문서 내에서 단어의 발생 빈도 등 단어의 언어 분석 자질을 추출할 수 있다. 본 연구에서는 언어 분석 자질을 최신의 인공신경망 방식과 결합하여 문서 요약에 적용하는 방안을 제안한다. 구체적으로는 단어의 품사정보, 개체명 정보, stop word 여부와 단어 빈도 정보를 기존의 단어 표층정보와 함께 결합하여 문장과 문서를 표현하고 이를 기반으로 추출 요약을 수행할 수 있는 인공신경망 모델을 제안한다. 또한 이러한 언어 분석 자질이 문서요약에 유용함을 보이기 위해 기존의 인공신경망 모델과 언어 분석 자질을 결합한 모델의 추출 요약 성능을 비교하고 이를 통해 모델의 유효성을 입증하였다.

2. 관련 연구

문장의 관계를 그래프로 표현하고 이를 통해 중요 문장을 추출하는 그래프-기반 방법[7]이나 ILP(Integer Linear Programming), Knapsack 알고리즘 등을 이용하여 문장의 최적 조합을 찾는 연구들이 추출 요약을 위한 전통적인 방법으로 연구되어왔다[9]. 또한 문장이 추출 될지 여부를 SVM, CRF와 같은 분류기가 판단하여 요약을 생성하는 분류기-기반 요약 연구도 진행되었다[10]. 이러한 분류기에서 문장을 표현하기 위해서는 자질이 필요하다. 이러한 자질로는 문장의 위치[11] 역문서빈도[7], 단어 빈도[5] 등 문장과 문장을 구성하는 단어들의 중요도를 파악하기 위한 정보들이 자질로서 사용되었다.

인공신경망 연구가 활성화된 이후로 이를 기반으로 문서요약을 수행하고자 하는 연구들이 진행되었다. 그 중 [12]의 연구에서 추출 문서요약 학습에 필요한 대량의 말뭉치를 확보하는 방안을 제안하였고 이후 많은 연구들이 이 말뭉치를 기반으로 문서 요약을 수행하였다. 이 말뭉치에서 문장의 선택 여부 레이블은 문장의 위치, 요약문과 문장의 n-gram 겹침 정도, 개체명 개수 등을 조합한 규칙을 통해 정해진다. [2]의 연구에서는 이 말뭉치에서 외부의 다른 지식 없이 문서와 사람이 작성한 요약만으로 레이블 할당 방안에 대해 제안하였다. 또한 문장 선택 확률을 해석할 수 있도록 설계하였다. 본 연구에서는 이 연구에서 제안한 SummaRuNNer모델을 기반으로 언어 분석 정보의 효과를 실험하였다.

언어 분석 정보를 단어와 함께 자질로 사용했던 [1]의 연구가 있었다. 이 연구에서는 sequence-to-sequence를 이용한 생성 문서요약에 이를 활용하였다. 이 연구에서 문서의 주요 키워드를 모델이 인식할 수 있도록 POS, TF, IDF, Named Entity와 같은 언어 분석 정보를 단어를 표현하기 위해 함께 사용하였다. 본 연구에서도 이와 유사한 언어 분석 정보 목록과 개념을 사용하였다. 본 연구에서는 이러한 정보들이 전통적으로 추출 문서요약에 적용되어온 점에 기반하여 추출 문서요약에서 이러한 정보들의 효과가 더 극대화 될 것으로 보고 이를 추출 문서요약에 적용해 보고자 한다.

3. 단일 문서 추출 요약

3.1 SummaRuNNer

본 연구에서 제안하는 언어 분석 자질을 기존의 추출 요약 모델에 적용하여 그 유용성을 살펴보고자 한다. 기존 모델로는 SummaRuNNer[2]을 사용하였다. 이 모델은 RNN 2계층으로 구성된 신경망 모델로 기존의 여러 연구들에서 성능 비교 대상으로 사용되어 온 추출 요약 모델이다. 모델을 도식화하면 Fig. 1과 같다.

이 모델은 n개의 문장으로 구성된 문서 $X=[S_1, S_2, ..., S_n]$에 대한 레이블 $Y=[Y_1, Y_2, ..., Y_n]$을 순차적으로 부여하여 추출 요약을 수행한다. 이때 $y_k \in [0,1]$는 k번째 문장이 요약에 포함되어야 할지(1) 아닌지(0)를 나타내는 것으로 Equation (1)을 통해 결정된다.

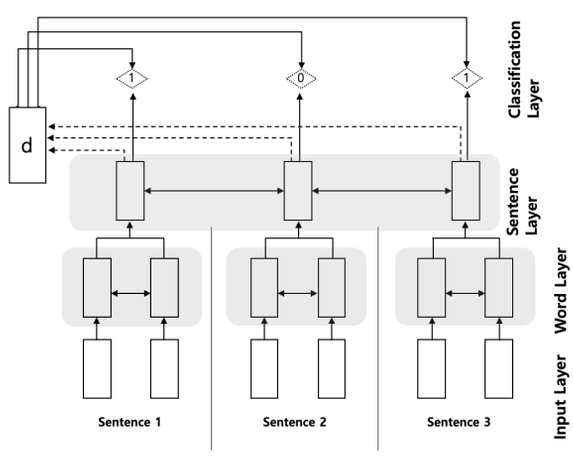


Fig. 1. Architecture of SumaRuNNer

$$
\begin{aligned}
p(y_k=1|h_k,g_k,d) = \sigma(\; & W_c h_k && \#\ content \\
& + h_k^T W_k d && \#\ salience \\
& - h_k^T W_r \tanh(g_k) && \#\ novelty \\
& + W_{ap} P_k^a && \#\ absolute\ position \\
& + W_{rp} P_k^r && \#\ relative\ position \\
& + b) && \#\ bias\ term
\end{aligned} \tag{1}
$$

Equation (1)에서 h_k는 k번째 입력 문장에 대한 벡터 표현이며, d는 입력 문서 전체를 표현하는 벡터이다. g_k는 k번째 이전까지 만들어진 부분 요약을 표현하는 벡터이다. 이들 정보

와 각 항목에 대한 가중치를 이용해 문장 자체의 중요성(content), 전체 문서에 대한 해당 문장의 대표성(salience), 이전까지 생성된 요약과 해당 문장과의 중복도(novelty), 문장의 절대적, 상대적 위치(P^{a_k} P^{r_k})에 따른 중요성을 수치화하고 이를 합하여 해당 문장이 요약으로 선택될 확률을 결정한다.

입력 문장은 2계층의 RNN을 통해 벡터로 표현된다. 첫번째 층은 개별 문장을 구성하는 단어들을 이용하여 문장을 표현한다(Fig. 1의 Word Layer). 문서가 입력되면, 문서를 구성하는 각 문장들은 개별적으로 양방향 GRU(Bi-directional Gated Recurrent Unit)[13] 층에 입력된다. 이 GRU 층의 입력은 문장을 구성하는 단어들로, p번째 단어의 입력은 단어의 임베딩 벡터($E_p = E^{surface_p}$)로 표현된다. m개의 단어로 구성된 k번째 문장의 벡터 s_k는 각 단어의 양방향 GRU 출력에 차원 별 평균(element-wise average)을 취해 생성한다.[1)]

$$s_k = \frac{1}{m}\sum_{p=1}^{m}[\overrightarrow{GRU}(E_p \mid E_1,...,E_m), \overleftarrow{GRU}(E_p \mid E_1,...,E_m)] \quad (2)$$

문장을 구성하는 단어로부터 만들어진 벡터 s_k는 또다른 GRU 층에 입력되어 주변 문장들과의 문맥정보가 반영된 문장 벡터 h_k를 생성한다.

$$h_k = [\overrightarrow{GRU}(s_k \mid s_1,...,s_n), \overleftarrow{GRU}(s_k \mid s_1,...,s_n)] \quad (3)$$

전체 문서를 표현하는 벡터 d가 Equation (4)를 통해 계산된다.

$$d = \tanh(W_d \frac{1}{n}\sum_{j=1}^{n} h_j + b) \quad (4)$$

요약은 문서에 대한 짧은 글을 생성해야 하기 때문에 그 길이에 제한을 받는다. 이때 중복된 내용이 요약에 입력되면 공간을 낭비하게 되므로 문장 선택에서 중복을 방지할 수 있는 장치가 필요하다. Equation (1)의 novelty항목은 선택 후보 문장과 지금까지 만들어진 요약의 유사도를 수치화하고 이것이 높은 경우 후보 문장이 요약으로 선택되지 않도록 전체적인 점수를 낮추는 역할을 한다. 현재까지 만들어진 요약의 표현 g_j는 Equation (5)와 같이 생성된다. 이는 이전까지 검증한 각 문장들이 요약으로 선택되었을 확률과 문장을 표현하는 벡터의 곱으로 계산된다.

$$g_j = \sum_{i=1}^{j-1} h_i P(y_i = 1 \mid h_i,\ s_i,\ d) \quad (5)$$

3.2 언어 분석 자질

본 연구에서 활용한 언어 분석 자질은 Table 1과 같다.

Surface 자질은 단어 그 자체를 의미한다. 딥러닝 이전의 연구에서 단어를 이용하여 문장이나 문서를 표현하기 위해서는 문제에 따라 수만 차원의 one-hot 벡터 또는 그 조합을 필요로 하였다. 하지만 인공신경망 기반 단어 표현 연구를 통해 다른 단어들과의 관계정보를 내포한 고정된 차원의 벡터로 표현 할 수 있게 되었다[14, 15]. 인공신경망 기반 문서요약은 이러한 단어 임베딩을 기본자질로 사용하였다. 본 연구에서도 단어의 임베딩 벡터를 기본 자질로 사용하였다. 이와 함께, 단어의 중요도나 문장 내부의 패턴 등을 모델에 좀더 직접적으로 반영할 수 있도록 추가적인 언어 분석 자질을 Surface 자질과 함께 사용하였다.

Table 1. Linguistic Analysis Features

Feature Name	Description	Category
Surface	word	number of vocab
TF	Word frequency in a document	0-50
POS	POS tag of a word	number of POS tag
NE	Whether a word is a named entity	binary
STOP	Whether a word is a stop word	binary

[5]의 연구를 통해 문서에서 자주 나타난 단어들이 요약에서도 자주 쓰인다는 것을 알 수 있다. 이러한 사실에 기반하여 단어 빈도수(Term Frequency, **TF**)는 추출 문서요약과 키워드 추출관련 연구에서 중요한 자질로 사용되어 왔다. 이를 토대로, 단어가 문서에서 가지는 중요도를 모델에 반영할 수 있도록 단어 빈도수를 단어를 표현하는 자질 중 하나로 사용하였다.

문장을 구성하는 모든 단어가 문장 안에서 중요한 역할을 하는 것은 아니다. 문법 구조를 위해 필요한 기능어의 경우, 문장의 의미 표현에 중요한 역할을 하지 않을 수 있기 때문에 많은 연구에서 전처리 단계로 제거 되었다[16]. **STOP** 자질은 어떤 단어가 의미가 있는 단어인지 아닌지 여부를 모델에 반영할 수 있게 해준다.

단어의 품사(part-of-speech, **POS**)도 이와 유사한 정보를 제공한다. 주요 의미를 담고 있는 동사, 명사와 같은 품사의 단어와 심볼이나 전치사, 관사 등 문법적인 역할을 하는 단어를 구별할 수 있게 해준다. 또한 문장 품사열의 패턴은 문장 안에서 주요 키워드를 식별하는데 도움을 줄 수 있다[17]. 문장의 품사열 패턴을 자질로 사용함으로써 모델이 이러한 정보를 학습에 직접 반영할 수 있도록 하였다.

문서요약을 이벤트(event) 중심으로 수행하는 시도들이 있었다[18, 19]. 이들 연구는 '누가', '무엇을', '누구에게', '언제' 등의 개체명(named entity, **NE**) 정보를 중심으로 요약을 생성하였다. 본 연구에서는 어떤 단어가 개체명인지를 알 수 있도록 단어의 개체명 여부를 자질로 사용하였다. 이를 통해 문서 전체에 얼마나 많은 개체명이 있는지, 어떤 문장이 중요한 개체명을 가지고 있는지 등을 모델에 직접 반영할 수 있다.

3.3 언어 분석 자질 결합

본 연구에서는 3.2에서 설명한 언어 분석 자질을 벡터화하여 기존의 단어 자질과 함께 결합하고 이를 인공신경망 문서요약 모델의 입력으로 사용하였다. 이러한 언어 분석 자질을 이용한 단어 표현은 Equation (6)과 같다.

1) [·]는 벡터의 연결(concatenate)를 의미한다.

$$E_p = \tanh(W_l[E_p^{surface}, E_p^{TF}, E_p^{POS}, E_p^{NE}, E_p^{STOP}] + b) \quad (6)$$

Equation (6)에서 E_p^{TF}, E_p^{POS}, E_p^{NE}, E_p^{STOP} 는 각각 p번째 단어의 단어 빈도수, 품사 종류, 개체명 여부, stop word 여부를 나타내는 벡터이다.

4. 실험 및 결과

4.1 학습 및 평가 데이터

언어 분석 자질의 효과를 검증하기 위하여 기존의 추출 문서요약 모델인 [2]의 SummaRuNNer 모델을 기반으로 비교 실험을 수행하였다. 이 실험에서 사용한 데이터는CNN/DailyMail 말뭉치이다. 이 말뭉치는 [20]의 연구에서 passage-based 질의응답을 위해 개발한 데이터로, [1, 12]의 연구에서 문서요약에 사용한 이래로 문서요약 연구에 널리 사용되고 있는 말뭉치이다. 이는 신문기사와 그 기사 내용에 대해 사람이 작성한 3-4줄의 요약문으로 구성된다. 이 데이터는 익명화(Anonymized)된 버전과 비익명화된(Non-Anonymized)버전이 제공된다. 익명화 데이터는 문서에서 나타난 개체명을 "@entity#"으로 대체한 데이터이다. 본 연구에서는 [2]의 연구와 동일하게 익명화된 버전의 데이터를 사용하였다. 본 연구에서는 [20]에서 제공한 스크립트를 이용하여 신문기사를 수집하고 함께 제공된 데이터를 이용해 토큰화(tokenization)와 개체명 인식(named entity recognition)을 수행하였다. 추가적으로, Stanford CoreNLP[21]와 원문을 이용하여 추가적인 문장 분리와 형태소 분석을 수행하였다. 이 데이터에 대한 수치 정보는 Table 2와 같다. 학습 데이터(Training)의 문서는 평균 약 40문장, 767단어로 구성되어 있고 약 4개의 요약문을 가지고 있다.

Table 2. Number Documents IN CNN/Dailymail

	Training	Validation	Test
CNN	90,151	1,220	1093
DailyMail	196,962	12,148	10397
Total	287,113	13,368	11,490

4.2 학습 설정

본 연구에서 사용한 단어 정보는 word2vec 알고리즘[15]을 이용하여 128차원의 벡터로 표현하였고 학습은 CNN/DailyMail 말뭉치의 학습데이터를 사용하였다. 언어 분석 자질은 Table 1의 category열과 같이 숫자 범주로 표현된다. POS 자질은 각 태그에 대응하는 64차원의 임베딩 벡터로 자질을 표현한다. NE와 STOP 자질은 개체명 여부와 stop word여부를 0과 1로 보고 이에 해당하는 64차원의 벡터로 표현하였다. TF의 경우 문서에서 단어가 등장한 빈도수의 절대값에 64차원의 벡터를 대응시켜 표현하였다.

학습의 효율을 위해 입력 문서 길이를 100문장으로 제한하고 각 문장의 길이는 50단어로 제한하였다. GRU 레이어들의 hidden 크기는 128차원으로 설정하였다. 4개의 GPU에서 각각 32개의 batch 크기를 이용하여 학습하였다. 파라미터 업데이트 알고리즘으로 Adam[22]알고리즘을 사용하고 Learning rate 0.001로 학습을 시작하였다. 파라미터 정규화를 위한 L2 penalty는 1e-6 로 설정하였다.

추출 문서요약을 위해서는 각 문장을 요약에 포함시켜야 할지 아닌지를 나타내는 정답 레이블이 필요하다. [2]에서 제안한 신문기사와 요약문을 이용한 레이블링 방법을 이용하여 각 문장에 대한 레이블을 할당하였다. 이러한 레이블과 모델 출력의 Binary Cross Entropy를 손실함수로 사용하여 모델의 파라미터를 학습하였다.

모델의 예측 결과로, 각 문장에 대한 선택 확률이 출력된다. 각 문장을 선택 확률에 따라 정렬하고 높은 순서대로 실험 조건에 맞춰 선택하여 요약을 생성한다. 학습 중 일정 주기로 validation 데이터를 이용하여 모델을 평가하고 그 중 가장 좋은 성능을 나타낸 모델을 비교실험에 사용하였다.

4.3 언어 분석 자질 성능 실험

Table 3은 기존 SummaRuNNer결과와 언어 분석 자질을 추가한 모델의 실험 결과를 비교한 표이다. 이 실험은 시스템에 의해 예측된 문장들 중 예측 점수 상위 3문장을 요약으로 결정하고 이 요약과 사람이 작성한 요약문과의 Rouge-n[23] F1점수를 계산한 결과이다. Table 3에서 LEAD-3는 기사의 첫 3문장을 요약으로 간주하여 만든 요약으로 [2]에서 제시한 결과이다. SummaRuNNer는 [2]의 논문에서 제시한 모델의 결과이고 SummaRuNNer-re는 본 연구에서 비교를 위해 재구현한 모델의 실험 결과이다. 여기에 언어 분석 자질을 추가하여 실험한 결과가 SummaRuNNer-Linguistic이다.

Table 3. Results of Full-Length f1

Models	Rouge-1	Rouge-2	Rouge-L
LEAD-3	39.2	15.7	35.5
SummaRuNNer	39.6	16.2	35.3
SummaRuNNer-re	40.2	16.8	36.7
SummaRuNNer -Linguistic	40.7	17.3	37.2
SWAP-NET[4]	41.6	18.3	37.7

Table 3의 결과에서 SummaRuNNer-re가 SummaRuNNer보다 더 높은 점수를 나타내는 것을 볼 수 있다. 이는 기존 SummaRuNNer 실험에 사용한 데이터의 경우 문서가 평균 28개의 문장으로 구성되어 있는 반면, 본 연구에서는 추가적인 문장 분리를 수행하여 문서의 문장 수가 평균 40문장으로 구성된 것과 관련 있는 것으로 보인다. 본 연구 데이터의 문장 길이가 평균적으로 더 짧기 때문에, 비슷한 Recall이라면 Precision 점수에서 좀더 이득을 볼 수 있다. 이를 통해 추출 요약의 단위를 좀더 의미 있게 나눌 수 있다면, 더 간결하고 의미를 충분히 담은 요약을 생성할 수 있다는 것을 알 수 있다. Table 3의 마지막은 본 연구에서 사용한 데이터와 같은 종류의 데이터를 사용하고 Pointer Network를 활용한 [4]의 실험 결과이다. 본 연구의 결과와 Rouge-2점수에서 약 1.0의 성능차이를 보이고 있다.

Table 3의 SummaRuNNer-Linguistic와 SummaRuNNer-re의 Rouge-1, Rogue-2, Rouge-L의 F1점수 차이는 약 0.5점씩

Table 4. Example of Summary

Gold Reference
- There are no English clubs left in the Champions League or Europa League - Diego Simeone admits he is surprised at the plight of Premier League sides - He believes Barcelona and Real Madrid are the top two clubs in Europe - Atletico Madrid face Real Madrid in their Champions League quarter-final
Summary using All features
- Chelsea, Arsenal and Manchester City were eliminated at the Champions League last 16 stage, and Everton were dumped out of the Europa League in the same round. - Atletico Madrid boss Diego Simeone believes English football needs to 'wake up' after this season's poor showing in Europe. - [Atletico Madrid]@entity1 boss [Diego Simeone]@entity0 has admitted his surprise at the plight of [English clubs]@entity3 in [Europe]@entity6
Summary using Surface
- Chelsea, Arsenal and Manchester City were eliminated at the Champions League last 16 stage, and Everton were dumped out of the Europa League in the same round. - Atletico Madrid boss Diego Simeone believes English football needs to 'wake up' after this season's poor showing in Europe. - [Simeone]@entity0 believes his side will face one of [Europe]@entity6's top two teams when they take on [Real Madrid]@entity15

차이가 나는 것을 볼 수 있다. 이를 통해 언어 분석 자질을 추가한 경우 더 높은 성능을 나타내는 것을 볼 수 있다. 이를 토대로 어떤 문장이 요약으로서 더 중요한 문장인지 결정하는 것에 언어 분석 자질이 중요 역할을 한다는 것을 알 수 있다.

4.4 언어 분석 자질 조합 실험

Fig. 2는 validation데이터에 대한 언어 분석 자질 조합별 실험 결과이다. 각 자질 조합의 모델 학습 중 일정 주기마다 validation데이터를 이용해 모델을 평가하였다. 그래프를 통해 모든 언어 분석 자질을 조합한 경우가 가장 좋은 결과를 나타냄을 알 수 있다. 또한 단어와 TF를 조합한 것이 다른 조합보다 더 높은 성능을 보이고 있다. 이를 통해 TF정보가 추출 요약에 중요하게 작용한다는 것을 알 수 있다. 모든 경우, 단어만 사용한 것보다는 언어 분석 자질을 함께 사용한 것이 대체로 더 높은 결과를 보이고 있다. 이를 종합하면 본 연구에서 제안한 언어 분석 자질이 추출 문서요약에서 효과가 있음을 알 수 있다.

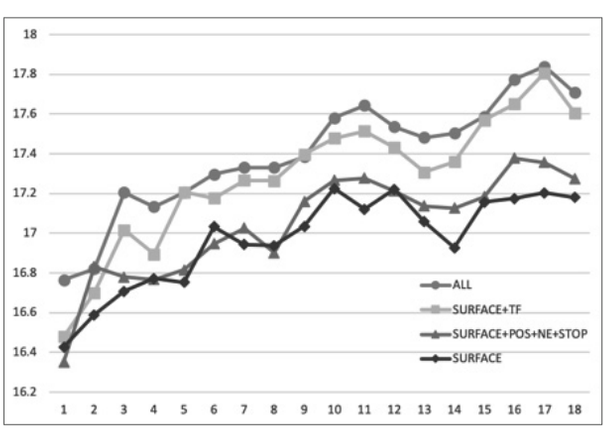


Fig. 2. Test of Validation

4.5 요약 예제 분석

Table 4에 추출 요약의 결과 중 하나를 나타내었다. 이 표의 'Summary using All features'은 모든 언어 분석 자질을 이용한 모델이 생성한 요약이고 'Summary using Surface'는 단어만 사용한 모델이 생성한 요약이다. Gold Reference는 해당 문서에 대해 사람이 쓴 요약문이다. 가독성을 위해 익명화된 개체명을 복원하여 나타내었다. 이 예에서 모든 언어 분석 자질을 통해 만들어진 요약(All features)과 단어 자질만을 이용해 만들어진 요약(Surface)은 처음 두 문장은 같은 문장이 선택되었고 세 번째 문장에서 다른 선택을 하였다. 서로 다른 3번째 문장을 살펴보면, All features의 경우 Surface보다 개체명 수가 1개 더 많다.

또한 문장에 나타난 개체명들의 문서에서 빈도수 합은 33회, 20회로 All features에 빈도수가 높은 개체명이 더 많이 사용되었다. 또한 문장에 등장한 단어들 중 stop word가 아닌 단어들의 TF평균은 2.5와 1.8로 All features이 더 높다.

5. 결 론

본 연구에서는 인공신경망 기반 추출 문서요약 모델과 언어 분석 자질을 결합하고 그 효과를 검증해보았다. 이를 위해 추출 문서요약에서 효과가 있을 언어 분석 자질 목록을 정하고 이를 기존의 인공신경망 기반 추출 문서요약 모델인 SummaRuNNer에 적용해 보았다. 그 결과, 단어만 사용한 경우보다 언어 분석 자질을 함께 결합하여 사용하는 것이 더 높은 성능을 보였다. 이를 통해 본 연구에서 제안한 언어 분석 자질이 추출 문서요약에 유용한 것을 보였다.

하지만 본 연구에서 제시한 결과는 현재 문서요약 시스템의 최고 성능과 비교하였을 때 Rouge2 Full-length F1에서 1~1.5정도 낮은 성능을 보이고 있다. 이는 기본 모델로 사용한 SummaRuNNer가 가지는 한계에서 비롯한 것으로 보인다. 또한 언어 분석 자질도 가장 기본적인 전방향 인공신경망을 이용하여 결합하였다. 이는 본 연구의 목적이 언어 분석 자질의 효과를 검증하는 것이므로, 기초적인 모델을 이용하여 그 효과를 검증하였다. 향후 더 나은 모델과의 결합이나 새로운 추출 요약 모델을 연구해 나갈 계획이다.

References

[1] R. Nallapati, et al., Abstractive textsummarization using sequence-to-sequence rnns and beyond. arXiv preprintarXiv: 1602.06023, 2016.

[2] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization ofdocuments," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, 2014.

[4] A. Jadhav and V. Rajan, "Extractive Summarizationwith SWAP-NET: Sentences and Words from Alternating Pointer Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). 2018.

[5] A. Nenkova, and L. Vanderwende, "The impact offrequency on summarization," Microsoft Research, Redmond, Washington, Tech. Rep.MSR-TR-2005, 2005. 101.

[6] E. Filatova and V. Hatzivassiloglou, "Event-based extractive summarization," 2004.

[7] G. Erkan, and D. R. Radev, "Lexrank: Graph-basedlexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, Vol.22, pp.457-479, 2004.

[8] D. R. Radev, et al., "Centroid-based summarizationof multiple documents," *Information Processing & Management*, Vol.40, No.6, pp.919-938, 2004.

[9] R. McDonald, "A study of global inferencealgorithms in multi-document summarization," in *European Conference onInformation Retrieval*, 2007. Springer.

[10] D. Shen, et al., "Document summarization usingconditional random fields," *IJCAI*, Vol.7, pp.2862-2867, 2007.

[11] H. P. Edmundson, "New methods in automaticextracting," *Journal of the ACM (JACM)*, Vol.16, No.2, pp.264-285, 1969.

[12] J. Cheng and M. Lapata, "Neural summarization byextracting sentences and words," arXiv preprint arXiv:1603.07252, 2016.

[13] K. Cho, et al., "Learning phrase representationsusing RNN encoder-decoder for statistical machine translation," arXiv preprintarXiv:1406.1078, 2014.

[14] Y. Bengio, et al., "A neural probabilisticlanguage model," *Journal of Machine Learning Research*, Vol.3(Feb), pp.1137-1155, 2003.

[15] T. Mikolov, et al., "Efficient estimation of wordrepresentations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[16] S. Menaka and N. Radha, "Text classificationusing keyword extraction technique," *International Journal of Advanced Researchin Computer Science and Software Engineering*, Vol.3, No.12, 2013.

[17] A. Hulth, "Improved automatic keyword extractiongiven more linguistic knowledge," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003. Association for Computational Linguistics.

[18] M. Wu, et al., "Event-based summarization usingtime features," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2007. Springer.

[19] W. Li, et al., "Extractive summarization usinginter-and intra-event relevance," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of theAssociation for Computational Linguistics*, 2006. Association for ComputationalLinguistics.

[20] K. M. Hermann, et al., "Teaching machines to readand comprehend," in *Advances in Neural Information Processing Systems*, 2015.

[21] C. Manning, et al., "The Stanford CoreNLP naturallanguage processing toolkit," in *Proceedings of 52nd annual meeting of theassociation for computational linguistics: system demonstrations*, 2014.

[22] *D. P. Kingma* and J. Ba, "Adam: A method forstochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[23] C.-Y. Lin, "Rouge: A package for automaticevaluation of summaries," *Text Summarization Branches Out*, pp.74-81, 2004.

이 경 호

https://orcid.org/0000-0002-3639-3155

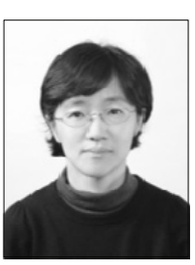
e-mail : gyholee@gmail.com

2011년 충남대학교 정보통신공학과(학사)

2013년 충남대학교 정보통신공학과(석사)

2013년~현 재 충남대학교 전자전파정보통신공학과 박사과정

관심분야 : 자연언어처리, 기계학습, 인공지



이 공 주

https://orcid.org/0000-0003-0025-4230

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~2003년 한국마이크로소프트(유) 연구원

2003년 이화여자대학교 컴퓨터학과 대우전임강사

2004년 경인여자대학 전산정보과 전임강사

2005년~현 재 충남대학교 전파정보통신공학과 교수

관심분야 : 자연언어처리, 기계학습, 인공지능, 정보검색