

# 문장 수반 관계를 고려한 문서 요약

## (Document Summarization Considering Entailment Relation between Sentences)

권 영 대 <sup>†</sup>  
(Youngdae Kwon)

김 누 리 <sup>\*\*</sup>  
(Noo-ri Kim)

이 지 형 <sup>\*\*\*</sup>  
(Jee-Hyong Lee)

**요 약** 문서의 요약은 요약문 내의 문장들끼리 서로 연관성 있게 이어져야 하고 하나의 짜임새 있는 글이 되어야 한다. 본 논문에서는 위의 목적을 달성하기 위해 문장 간의 유사도와 수반 관계(Entailment)를 고려하여 문서 내에서 연관성이 크고 의미, 개념적인 연결성이 높은 문장들을 추출할 수 있도록 하였다. 본 논문에서는 Recurrent Neural Network 기반의 문장 관계 추론 모델과 그래프 기반의 랭킹(Graph-based ranking) 알고리즘을 혼합하여 단일 문서 추출요약 작업에 적용한 새로운 알고리즘인 TextRank-NLI를 제안한다. 새로운 알고리즘의 성능을 평가하기 위해 기존의 문서요약 알고리즘인 TextRank와 동일한 데이터 셋을 사용하여 성능을 비교 분석하였으며 기존의 알고리즘보다 약 2.3% 더 나은 성능을 보이는 것을 확인하였다.

**키워드:** 문서 요약, 수반 관계 추론, 자연어 처리, 텍스트 랭크, 순환 신경망

**Abstract** Document summarization aims to generate a summary that is consistent and contains the highly related sentences in a document. In this study, we implemented for document summarization that extracts highly related sentences from a whole document by considering both similarities and entailment relations between sentences. Accordingly, we proposed a new algorithm, TextRank-NLI, which combines a Recurrent Neural Network based Natural Language Inference model and a Graph-based ranking algorithm used in single document extraction-based summarization task. In order to evaluate the performance of the new algorithm, we conducted experiments using the same datasets as used in TextRank algorithm. The results indicated that TextRank-NLI showed 2.3% improvement in performance, as compared to TextRank.

**Keywords:** document summarization, entailment relation inference, natural language processing, TextRank, recurrent neural network

- 이 논문은 2014년도 미래창조과학부의 재원으로 한국연구재단-차세대정보 컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2014M3C4A7030503)
- 이 논문은 2016 한국컴퓨터종합학술대회에서 [언어공학] 문장 수반 관계 추론을 통한 문서 요약의 제재목으로 발표된 논문을 확장한 것임

<sup>†</sup> 학생회원 : 성균관대학교 반도체시스템공학과  
kydchonje@gmail.com

<sup>\*\*</sup> 학생회원 : 성균관대학교 컴퓨터공학과  
pd99j@skku.edu

<sup>\*\*\*</sup> 종신회원 : 성균관대학교 컴퓨터공학과 교수  
(Sungkyunkwan Univ.)  
john@skku.edu  
(Corresponding author임)

논문접수 : 2016년 9월 20일  
(Received 20 September 2016)  
논문수정 : 2016년 11월 14일  
(Revised 14 November 2016)  
심사완료 : 2016년 11월 14일  
(Accepted 14 November 2016)

Copyright©2017 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.  
정보과학회논문지 제44권 제2호(2017. 2)

## 1. 서론

인터넷과 스마트폰의 대중화로 온라인 상에 존재하는 데이터가 빠르게 증가하고 있다. 2010년 이후 전 세계 데이터 생성량은 매년 2배 가량 성장하고 있으며, 2012년 기준 데이터량은 2.8제타바이트(ZB, 1ZB = 1조GB)에서 2020년까지 40제타바이트까지 늘어날 것으로 전망되고 있다[1].

이렇듯 현재 폭발적으로 늘어나고 있는 데이터의 양에 비해서 대부분의 데이터들은 잘 정제되어 있지 않기 때문에 간결하게 요약된 정보는 찾기 어려워졌다. 결국, 이러한 데이터의 증가로 인해서 사용자들은 원하는 정보를 찾는데 더욱 많은 시간을 투자해야 한다. 정보란 기본적으로 사람이 쉽게 이해할 수 있는 의미 있는 데이터 이어야 한다. 이에 문서요약 기술의 필요성이 대두되고 있으며 그 중요성 또한 점점 커지고 있다.

대표적인 문서요약 방식인 TextRank[2] 방식은 하나의 문서를 그래프로 변환한 다음, 단어의 빈도수를 계산하여 각각의 정점의 중요도를 산출해낸다. TextRank에서 문장 간의 유사도를 계산할 때 가정하고 있는 부분은 겹치는 단어의 수가 많을수록 두 문장이 유사하고 겹치는 단어의 수가 적을수록 두 문장은 유사하지 않는 것이다. 하지만 만약 예를 들어서 표 1의 Example 1과 2 같은 두 개의 문장이 있다고 한다면 두 문장 사이의 겹치는 단어의 수를 기준으로 문장 간의 유사도를 판단하기 어려울 것이다. 이 두 문장은 겹치는 단어의 수는 아예 존재하지 않지만 서로 연관성이 높은 문장이라고 할 수 있다. 그러므로 TextRank 방식은 정점 중요도를 계산할 때, 정점 즉 문장 내의 단어들의 빈도수만을 기준으로 해당 정점의 중요도를 계산하기 때문에 두 문장의 의미, 개념적 연결성을 제대로 고려해주지 못한다.

또한 단어의 빈도수와 함께 문서의 빈도수를 고려한 가중치들을 이용하여 문서 내의 문장들의 중요도를 결정하는 TF-IDF(Term Frequency - Inverse Document Frequency)[10] 방법이 존재한다. TF-IDF방법은 TextRank모델과 마찬가지로 두 문장 사이에 겹치는 단어가 없다면 유사성을 판단하기 어렵다는 단점이 존재한다. 결국 TF-IDF 방법도 두 문장 간의 의미, 개념적 연결성을 제대로 고려해주지 못한다.

그 밖에도 동의어, 유의어, 상위어, 하위어 관계 등의 단어 개념을 정리해 둔 시소러스(Thesaurus)를 활용하는 방법과 을 사용한 방법이 있다. 시소러스를 활용한 방법의 경우, 해당 단어가 시소러스에 등록되어 있지 않거나 연관 관계가 설정되어 있지 않다면 단어 간의 연관성을 파악하기 어렵다. 마찬가지로 표 1에서 Example

1과 2의 경우, 두 문장 사이에 연관성 있는 단어가 나타나지 않기 때문에 시소러스를 사용해서는 두 문장 간의 연관성을 파악하기 힘들다.

본 논문에서는 단어의 의미와 맥락을 고려하여 단어를 벡터로 표현한 Word Embedding을 기반으로 한 문장 관계 추론 모델을 사용한다. 문장 관계 추론 모델을 이용해 구한 수반 관계를 문장 추출 방식의 단일 문서 요약 작업에 적용하였다. 문서의 요약은 요약문 내의 문장들끼리 서로 연관성이 높아야 하고 의미, 개념적 연결성이 높은 글이 되어야 한다. 본 논문에서는 이 문제를 해결하기 위해 문장 간의 유사도를 고려하여 문서 내에서 연관성이 크고 중요한 문장들을 추출할 수 있도록 하였다. 또한, 문서 내의 문장들 간의 수반 관계를 사용하여 요약문의 의미, 개념적 연결성을 고려하였다.

본 논문에서는 단일 문서 내에서 문장 간의 수반 관계를 고려하며 TextRank를 기반으로 한 새로운 알고리즘인 TextRank-NLI를 제안한다. Document Understanding Conferences(DUC, 2002)[4]을 사용하여 제안한 알고리즘의 성능을 비교, 분석하였다.

2절에서는 문장 관계 추론에 대해서 소개한 뒤, 제안하고자 하는 알고리즘의 기반이 되는 알고리즘을 소개한다. 3절에서는 본 논문에서 제안하는 새로운 단일 문서 요약 방법을 설명하고 4절에서는 제안한 방법을 실험을 통해 평가한다.

## 2. 관련 연구

### 2.1 문장 관계 추론

소위 Natural Language Inference(NLI)라고 불리는 문장 관계 추론 작업은 P와 H라는 문장이 주어졌을 때, 수반 관계(Entailment), 모순 관계(Contradiction), 그리고 관계 없음(Neutral) 이렇게 세 가지 관계 중 하나로 추론하는 것이다[3]. 문장 간의 관계는 다음과 같이 정의된다. 두 문장 간의 의미적 연결성이 높고 논리적인 모순이 존재하지 않으면 "수반 관계", 반대로 모순이 존재하면 "모순 관계", 그리고 모순이 존재하는지 알 수 없다면 "관계 없음"이라고 정의된다. 세 가지 문장 간의 관계에 대한 예시는 표 1에서 확인할 수 있다.

본 논문에서는 문장 관계 추론 작업을 위해 Stanford Natural Language Inference(SNLI)[6] 데이터 셋을 사용하였다. SNLI 데이터 셋은 한 쌍의 문장들과 그에 대응하는 세 가지 문장 간의 관계로 구성된다. 본 논문에서는 Cheng et al.[5]의 Long Shot-Term Memory-Networks(LSTMN) 모델을 사용하였다. 위 모델은 두 문장이 주어졌을 때, 입력 값으로 두 문장의 단어 벡터를 받아들이고 출력으로 세 가지 문장 간의 관계를 클

표 1 수반, 모순, 관계 없음이란 문장 간의 관계를 지니는 문장 쌍들에 대한 예시들

Table 1 Examples of sentence pairs with Entailment, Contradiction and Neutral relations

	ID	Sentence	Relation
Premise	Example 1	The pitcher wearing black heaves the ball toward the catcher.	
Hypothesis	Example 2	People are playing a game.	Entailment
	Example 3	The Seahawks win the Super Bowl.	Contradiction
	Example 4	The Yankees are at practice.	Neutral

래스로 하여 각 클래스에 대한 확률 값을 내보낸다. Cheng et al.은 입력 단어들의 길이에 따라 문장 벡터의 크기를 변화시킴으로써 Recurrent Neural Networks (RNN)모델을 개선시켰다. Cheng et al.은 SNLI데이터셋을 사용하여 모델을 학습시켰고 해당 모델은 문장 관계 추론 작업에서 매우 우수한 성능을 보이고 있다.

## 2.2 TextRank

본 논문에서 제안하는 모델이 TextRank모델을 기반으로 하기 때문에 TextRank모델에 대해서 우선적으로 설명한다. TextRank모델은 "투표" 개념을 활용하여 각 연결선에 부분 가중치( $\omega_{ij}$ )를 부여한 모델로써 식 (1)과 (2)로 표현된다. TextRank모델은 PageRank[9]모델과 동일하게 정점( $S_i$ )의 점수( $WS(S_i)$ )를 계산하기 위해 수식 (1)을 사용한다. (감쇠율( $d$ ) = 0.85) 여기서 TextRank 모델은 두 개의 문장  $S_i$ 와  $S_j$ 가 주어졌을 때, 연결선의 가중치 계산을 위해 식 (2)를 사용한다. ( $S_i = \omega_1^i, \omega_2^i, \dots, \omega_{N_i}^i$ ,  $N_i = S_i$ 의 단어의 총 개수)

$$WS(S_i) = (1 - d) + d * \sum_{S_j \in IN(S_i)} \frac{\omega_{ji}}{\sum_{S_k \in OUT(S_i)} \omega_{jk}} WS(S_j) \quad (1)$$

$$\text{Similarity}(S_i, S_j) = \frac{[|\omega_k| \omega_k \in S_i \& \omega_k \in S_j]}{\log(|S_i|) + \log(|S_j|)} \quad (2)$$

## 2.3 TF-IDF

TF-IDF 방식은 정보 검색 분야에서 이용되는 가치로 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 즉, TF-IDF는 단어 빈도(Term Frequency)와 역문서 빈도(Inverse Document Frequency)의 곱이다.

본 논문에서는 TextRank의 간단한 변형으로 Similarity 계산시 TF-IDF방식을 적용하여 단일 문서 요약 작업에 적용한 모델을 비교 분석한다.

TF-IDF방식을 사용한 모델처럼 단어의 빈도만이 아닌 문서의 빈도까지 고려하여 문장 간의 유사도를 측정하는 모델도 문장 간의 의미, 개념적 연결성을 고려해 주지 못한다. 본 논문에서는 그래프 기반 랭킹 방식의 모델에 문장 관계 추론 모델을 적용하여 문장 간의 연결성을 고려할 수 있도록 개선하였다.

## 3. TextRank-NLI

문서요약으로 생성한 요약문을 문서 전체의 중요한 부분을 담고 있으며 연결성 있는 글로 정의한다면 단어의 빈도수만을 고려한 TextRank모델로는 연결성을 고려해 주기 어렵다. 그러므로 본 논문에서는 요약문의 연결성을 보장해주기 위해 세 가지 문장 간의 관계 중에서 수반 관계를 사용하여 다른 문장들과 더 긴밀히 연결되어 있는 문장일수록 더 높은 점수를 가지도록 하였다.

문장 관계 추론 모델은 두 문장 사이의 관계를 확률 값으로 나타낸다. 세 가지 관계 중 수반 관계 확률 값이 크다면 두 문장은 의미, 개념적으로 연결되는 문장이고 확률 값이 작다면 연결되지 않는 문장이라고 볼 수 있다. 이러한 관점에서 요약문을 구성하는 문장들이 유사도 뿐만 아니라 높은 수반 관계 확률도 가질 수 있도록 제안 모델을 구현하였다. 본 논문에서 제안하는 TextRank-NLI 모델은 다음과 같다.

$$PR(S_i) = \sum_{j=i+1}^N \sum_{j=i+1}^N P(S_i, S_j) \quad (3)$$

$$\text{where } P(S_i, S_j) = 0 \text{ if } |S_i| < k \text{ or } |S_j| < k$$

식 (3)에서는 단일 문서 내의 전체 문장들에 대해서 각 문장( $S_i$ )이 그 뒤에 이어지는 문장( $S_j$ )에 대해 가지는 수반 관계 확률( $P(S_i, S_j)$ )의 합( $PR(S_i)$ )을 구한다. 수반 관계 확률의 합( $PR(S_j)$ )의 값은 각 문장이 뒤에 이어지는 문장들에 대해 얼마만큼 의미적, 개념적 연결성을 지니고 있는지를 나타낸다. 이 때 만약 문장의 길이가 특정 문턱값( $k$ ) 이하라면 고려하지 않는다.<sup>1)</sup> ( $N$  = 문장의 총 개수)

$$NPR(S_i) = \frac{PR(S_i)}{\text{count}(\sum_{j=i+1}^N P(S_i, S_j)) + 1} \quad (4)$$

$$\text{score}(S_i) = WS(S_i) + NPR_n(S_i) \quad (5)$$

식 (3)의 조건에 의해 수반 관계 확률을 합하는 개수가 각 문장마다 다르므로 식 (4)를 이용해 수반 관계 확률의 합( $PR(S_i)$ )을  $NPR(S_i)$ (Normalized  $PR(S_i)$ )로 정규화시킨다. 두괄식 구성이 많은 영어 문서의 특성을 고려하여 문서의 시작 문장을 선호하도록  $\text{count}+1$ 으로 나

1) 특정 문턱값( $k$ )은 60 글자 내외이다.

누어주고 추가적으로 가중치를 부여하였다. 식 (5)에서 TextRank모델의 점수( $WS(S_i)$ )에 정규화된 수반 확률의 합( $NPR(S_i)$ ) 중에 상위  $n$ 개에 해당하는 값( $NPR_n(S_i)$ )을 더하여 각 문장에 대한 최종 점수( $score(S_i)$ )를 구한다. 결국 전체 문장들에 대해서 높은 유사도를 가지는 문장들 중에 의미적, 개념적 연결성 또한 높은 문장들이 최종적으로 높은 점수를 가지게 된다. 마지막으로 전체 문장들을 최종 점수 순으로 정렬하여 상위에 위치하는 문장들을 추출하여 요약문을 생성한다.

## 4. 실험

### 4.1 문장 관계 추론

SNLI 데이터 셋에는 총 570,152개의 문장 쌍이 존재한다. Bowman et al.은 550,152개의 문장 쌍을 training set으로, 10,000개의 문장 쌍을 각각 development set과 test set으로 사용하였다. SNLI 데이터 셋의 각각의 문장 쌍들은 위에서 언급된 세 가지 문장 간의 관계들 중 하나를 가지거나 "-"기호를 가진다. "-"기호는 사람들에게 의해 문장 쌍에 부여된 관계들이 과반수 일치율 보이지 않을 때 부여된다. 본 실험에서는 정확한 수반 관계 확률을 구하기 위해 "-"기호가 부여된 문장 쌍들을 제외한 나머지를 가지고 실험을 진행하였다. "-"기호가 부여된 문장 쌍들을 제외하면 training set으로 549,367개의 문장 쌍, development set으로 9,842개의 문장 쌍, test set으로 9,824개의 문장 쌍이 남는다(표 2).

우선 LSTMN모델을 학습시키기 위해서 사용한 매개 변수에 대해서 설명한다. 먼저 Mikolov et al.[11]의 word2vec과 Pennington et al.[7]의 GloVe의 미리 학습된 단어 벡터를 각각 사용하여 LSTMN모델을 학습시킨 후 해당 모델들의 성능을 비교하였다. 본 논문에서는 GloVe 단어 벡터를 사용한 모델의 정확도가 더 우수하였기 때문에 GloVe 단어 벡터를 실험에 활용하였다. 또한, 임의로 정의된 단어 벡터에 의한 성능 저하를 막기 위해 SNLI 데이터 셋과 DUC 2002 문서 데이터에는 나타나지만 GloVe에서 미리 학습되어 있지 않은 단어 벡터는 해당 문장 내의 다른 단어 벡터들의 평균으로 정의하였다. LSTMN모델은 단어 벡터를 제외한 채 학습시켰다(표 3).

표 2 학습, 검증, 시험 집합에서 사용된 문장 쌍의 개수  
Table 2 Numbers of sentence pairs in training, development and test sets

SNLI	Sentence Pairs
Training set	549,367
Development set	9,842
Test set	9,824

표 3 LSTMN 모델의 정확도(%)

Table 3 Accuracies(%) of LSTMN model

LSTMN (our implementation)	Accuracy
Development set	82.6
Test set	79.5

### 4.2 문장 추출 요약

본 논문에서는 DUC 2002에서 사용된 문서요약 데이터 셋을 사용하였다. DUC 2002은 567개의 뉴스 기사를 이루어져 있다. 각각의 뉴스 기사에 대해서 사람이 직접 뉴스 기사를 읽고 작성된 약 100단어 분량의 요약문이 2개씩 있다.

표 4는 요약문 평가시 각 모델이 만든 요약문과 사람이 만든 요약문의 (a)all words, (b)stemmed words, (c)stemmed words, and no stop-words에 대해서 각각 비교한 것이다. TOP( $n, m$ )은 수반 관계 확률 TOP  $n$ 개의 문장에 가중치로  $m$ 을 곱해준 점수를 더해서 얻은 결과이다. 표 4의 실험에서 사용한 문턱값은 60 글자이다.

실험에서 사용한 모델들은 각 문서에 대해서 100단어 분량의 요약문을 생성한다. 요약된 문서를 검증하기 위한 평가 척도로 사람의 평가와 높은 상관 관계가 있다고 알려져 있는 ROUGE[8]를 이용하였다. ROUGE 점수는 모델의 요약문과 사람이 만든 요약문 사이의 겹치는 단어수( $n$ -gram)에 따라 95%의 신뢰구간으로 계산된다. 표 4는 unigram을 사용한 ROUGE-1의 Precision, Recall, F1-score를 각 모델 별로 보여준다. 제안 모델인 TextRank-NLI의 성능 평가를 위하여 세 가지의 모델을 비교군으로 선정하였다. Only Entailment는 TextRank-NLI모델에서 문장 간의 유사도를 고려하지 않고 수반 관계 확률만을 이용한 결과이다. TextRank는 본 논문에서 구현한 원래의 TextRank 모델의 성능 결과이고 TextRank(TF-IDF weight)는 원래의 TextRank모델에 TF-IDF weight를 각 점점의 연결선의 가중치로 사용하여 얻은 결과이다.

TextRank의 점수에서 수반 관계 확률을 더해줄 때 수반 관계 확률에 어느 정도의 가중치를 주어야 하는지를 알아보기 위해 여러 가지 경우의 수에 대해서 실험을 진행하였다. 표 4를 보면 알 수 있듯이 TOP( $n, 1$ )의 경우처럼 작은 가중치를 부여했을 때는 성능의 변화가 크지 않다. TOP( $n, 20$ )같이 큰 가중치를 부여했을 때보다 TOP( $n, 5$ ), TOP( $n, 10$ )의 경우처럼 적절한 가중치를 주었을 때 대체로 TextRank-NLI의 성능이 가장 높은 것을 확인할 수 있다. TOP(1,  $m$ ), TOP(3,  $m$ ) 모델과 TOP(5,  $m$ ) 모델을 비교했을 때, 수반 관계 확률을 더 많은 문장에 대해서 적용할수록 성능을 더 높일 수 있다는 걸 알 수 있다. TextRank와 수반 관계 확률 둘

표 4 Precision, Recall, F1-score에 대한 ROUGE-1 Score 비교 분석. 95% 신뢰구간으로 t-test를 진행하여 TextRank모델의 성능에 비해 유의미한 성능 향상인 경우 \*표시

Table 4 ROUGE-1 Score in terms of Precision, Recall, F1-score. \* represents a significant improvement over the TextRank model according to t-test at 95% confidence interval

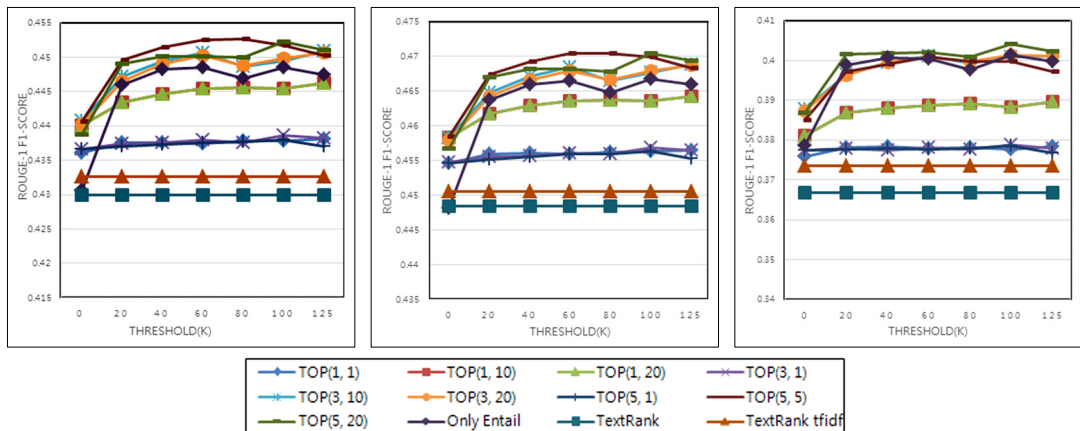
Method	ROUGE-1 Score								
	Basic (a)			Stemmed (b)			Stemmed no-stopwords (c)		
	P	R	F	P	R	F	P	R	F
TOP(1, 1)	0.4400	0.4352	0.4374	0.4586	0.4536	0.4559	0.3915	0.3675	0.3779
TOP(1, 10)	0.4479*	0.4433*	0.4447*	0.4662*	0.4614*	0.4629*	0.4001	0.3803*	0.3879*
TOP(1, 20)	0.4479*	0.4433*	0.4447*	0.4662*	0.4614*	0.4629*	0.4001	0.3803*	0.3879*
TOP(3, 1)	0.4405	0.4358	0.4379	0.4588	0.4539	0.4561	0.3914	0.3677	0.3780
TOP(3, 10)	0.4526*	0.4491*	0.4507*	0.4705*	0.4669*	0.4685*	0.4092*	0.3955*	0.4011*
TOP(3, 20)	0.4522*	0.4488*	0.4503*	0.4699*	0.4665*	0.4680*	0.4086*	0.3957*	0.4009*
TOP(5, 1)	0.4402	0.4354	0.4376	0.4586	0.4536	0.4559	0.3917	0.3669	0.3776
TOP(5, 5)	<b>0.4543*</b>	<b>0.4510*</b>	<b>0.4525*</b>	<b>0.4724*</b>	<b>0.4690*</b>	<b>0.4705*</b>	<b>0.4095*</b>	0.3947*	0.4009*
TOP(5, 20)	0.4517*	0.4489*	0.4501*	0.4698*	0.4669*	0.4682*	0.4088*	0.3974*	0.4020*
Only Entailment	0.4498*	0.4476*	0.4485*	0.4678*	0.4656*	0.4665*	0.4060*	<b>0.3969*</b>	<b>0.4003*</b>
TextRank (our implementation)	0.4325	0.4276	0.4299	0.4512	0.4460	0.4484	0.3817	0.3556	0.3669
TextRank (TF-IDF weight)	0.4346	0.4310	0.4326	0.4527	0.4489	0.4506	0.3900	0.3609	0.3734

중에 하나만을 적용한 모델보다 TextRank와 수반 관계 확률을 동시에 적용한 모델의 성능이 더 높은 것을 고려했을 때, TextRank와 수반 관계는 상호 보완적이라고 할 수 있다. 하지만 수반 관계 확률만을 사용한 Only Entailment 모델이 대부분의 경우 TextRank-NLI에 비해 성능이 떨어지는 것은 문장 관계 추출 모델로부터 구해내는 수반 관계 확률의 정확도가 높지 않아 수반 관계 확률만으로는 연관성 있는 문장들을 구성하기 힘들다고 볼 수 있다.

문턱값의 크기에 따른 TextRank-NLI의 성능을 비교

분석하기 위해 실험을 수행하였다. 그림 1은 요약문 평가시 각 모델이 만든 요약문과 사람이 만든 요약문의 (a)all words, (b)stemmed words, (c)stemmed words, and no stop-words에 대해서 각각 비교한 것이다. 또한, 표 4에서와 동일한 실험 경우들에 대하여 문턱값의 크기를 매우 작은 값(0글자)에서부터 매우 큰 값(125글자)까지 변화시켜 가면서 각 경우의 성능을 알아내었다.

먼저 그림 1을 보면 알 수 있듯이, 문턱값의 크기가 20글자 이하로 매우 작을 때, 전체적으로 TextRank-NLI모델의 성능이 좋지 않다. 각 테스트 케이스 별로 약



(a) Basic

(b) Stemmed

(c) Stemmed no-stopwords

그림 1 특정 문턱값(k)에 따라서 ROUGE-1 F1-score를 표시한 그래프

Fig. 1 ROUGE-1 F1-score graph corresponding to threshold(k)

표 5 실험용 데이터 셋에 대한 통계  
Table 5 Statistics on the experimental dataset

	DUC 2002
Type	Articles
Language	English
Documents	567
Tokens/document	3738.6
Sentences/document	37.0
Tokens/sentence	101.0

#### Document Example

1. u a AM-SuperBowl-Newsage 01-28 0265
2. AM-Super Bowl-Newsage,0277
3. San Francisco Routs Broncos in Super Bowl
4. NEW ORLEANS (AP)
5. The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever.
6. The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade and tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries.
7. San Francisco won the National Football League championship game in 1989, 1985 and 1982.
8. The Broncos, on the other hand, lost the last four Super Bowl games they have played.
9. San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions.
10. He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever.
11. Montana left the game with nearly 11 minutes to play.
12. In four Super Bowls, he has thrown 11 touchdowns and no interceptions.
13. For Denver quarterback John Elway, it was a day of futility, ending with his third Super Bowl defeat.
14. He missed eight of his first 10 passes and was intercepted twice and fumbled once.
15. By halftime the score was 27-3.
16. With their third loss in four years, the Broncos have now been outscored 136-40.
17. San Francisco was boringly perfect, doing more than even the experts who made them favorites by nearly two touchdowns after a 14-2 season and a waltz through the playoffs.

#### TextRank-NLI Summary

The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade and tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries. He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever. The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. The Broncos, on the other hand, lost the last four Super Bowl games they have played.

#### Manual Summary 1

The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. This was the 49ers fourth win in four tries, tying them with the Pittsburgh Steelers for the NFL championship. San Francisco quarterback Joe Montana set several Super Bowl records: five touchdown passes; 13 straight pass completions; a third Super Bowl MVP award. San Francisco's point total was the most ever. In four Super Bowls, Montana has thrown 11 touchdowns and no interceptions. It was a day of futility for Broncos' quarterback John Elway as his team suffered its third Super Bowl defeat.

#### Manual Summary 2

The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. The 24th Super Bowl also generated other statistics. The 49ers, having also won in 1989, 1985 and 1982, tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries. San Francisco quarterback Joe Montana, broke the record for touchdown passes, 5. He also set a record for straight pass completions, 13. Montana set five Super Bowl career records, including his third Super Bowl MVP award, and San Francisco's point total was the most ever.

그림 2 요약에 사용된 문서와 요약문 예시들  
Fig. 2 A document and its example summaries

간의 변동성이 존재하지만 문턱값의 크기를 증가시키에 따라서 TextRank-NLI모델의 성능이 상승하는 것을 확인하였다. 하지만 문턱값의 크기가 100글자를 초과하여 매우 커지는 경우엔 오히려 TextRank-NLI모델의 성능이 하락한다. 작은 문턱값을 적용하였을 때, 성능 향상이 이뤄지는 것은 요약문 생성에 큰 도움이 되지 않지만 비교적 높은 수반 관계 확률을 가지는 매우 짧은 문장들이 걸러지기 때문이다. 매우 큰 문턱값을 적용하였을 경우에는 필요한 정보를 지니는 문장들까지 제외되는 경우가 발생하기 때문에 모델의 성능 저하를 유발시킨다.

그림 1의 그래프를 확인하였을 때, 문턱값을 적절히 조절하여 수치를 약 60내외로 두었을 때 대부분의 경우가 가장 높은 성능을 보인다. 하지만, 모든 경우에서 위와 같은 경향성을 보이는 것은 아니다. 예를 들어, TOP(5, 20)처럼 몇몇의 경우 문턱값이 100일 때, 가장 높은 성능을 보인다. 또한, 표 4에서 분석하였듯이 그래프를 통해서도 TOP(n, m)에서 n값이 동일할 때 적절한 가중치 m값(5 or 10)을 가지는 것이 대체적으로 성능이 더 좋은 것을 확인하였다. 마지막으로 TextRank-NLI가 비교군으로 선정한 세 개의 모델들보다 (a), (b), 그리고 (c)의 경우에서 문턱값이 20 글자 이상인 경우 더 나은 성능을 보이는 것을 확인하였다.

위의 그림 2를 살펴보면 TextRank-NLI에 의해 만들어진 요약문은 예시 문서의 핵심적인 내용을 두루 포함하고 있다. 또한, TextRank-NLI Summary는 사람에 의해 작성된 요약문들과 비교했을 때, 매우 비슷하다는 것을 확인할 수 있다.

## 5. 결론

본 논문에서는 기존의 TextRank에 수반 관계라는 요소를 추가한 새로운 단일 문서 요약 방법을 제안하였다. 기존의 TextRank와 제안한 모델인 TextRank-NLI를 DUC 2002 문서 데이터 셋에 대하여 실험을 진행한 결과 TextRank-NLI이 더 나은 성능을 보이는 것을 확인할 수 있었다. 이는 문장의 의미, 개념적 연결성을 고려하는 것이 단일 문서 추출 요약에 유용함을 의미한다.

향후 문장 간의 관계 중에 수반 관계뿐만 아니라 모순 관계 등 다양한 문장 간의 관계들을 함께 고려할 수 있는 연구가 필요할 것이다.

## References

- [1] John Gantz, David Reinsel, The Digital Universe in 2020, [Online]. Available: <https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>, 2012.
- [2] Mihalcea, Rada, and Paul Tarau, "TextRank: Bringing order into texts," Association for Computational

- Linguistics, 2004.
- [3] MacCartney, Bill, "Natural language inference," PhD diss., Stanford University, 2009.
  - [4] DUC. Document understanding conference 2002, [Online]. Available: <http://www-nlpir.nist.gov/projects/duc/>, 2002.
  - [5] Cheng, Jianpeng, Li Dong, and Mirella Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.
  - [6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning, "A large annotated corpus for learning natural language inference," *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
  - [7] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, "Glove: Global Vectors for Word Representation," *EMNLP*, Vol. 14, pp. 1532-1543, 2014.
  - [8] Lin, Chin-Yew, and Eduard Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Vol. 1, pp. 71-78, 2003.
  - [9] S.Brin and L.Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107-117, 1998.
  - [10] Ramos, Juan, "Using tf-idf to determine word relevance in document queries," *Proc. of the first instructional conference on machine learning*, 2003.
  - [11] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, pp. 3111-3119, 2013.



이 지 형

1993년 한국과학기술원 전산학과 학사  
 1995년 한국과학기술원 전산학과 석사  
 1999년 한국과학기술원 전산학과 박사  
 2000년~2002년 SRI International, International Fellow.  
 2002년~현재 성균관대학교 컴퓨터공학과 교수. 관심분야는 기계학습, 추천시스템, 텍스트마이닝, 딥러닝



권 영 대

2016년 성균관대학교 반도체시스템공학과 학사. 2016년~현재 삼성전자 주임연구원. 관심분야는 데이터 마이닝, 기계학습, 인공지능망



김 누 리

2013년 성균관대학교 전자전기컴퓨터공학과 학사. 2013년~현재 성균관대학교 전자전기컴퓨터공학과 석박사 통합과정 관심분야는 추천시스템, 텍스트마이닝, 기계학습