한국정보전자통신기술학회논문지(jkiiect)'18-10, Vol.11 No.5
논문 18-11-05-467 http://dx.doi.org/10.17661/jkiiect.2018.11.5.467

향상된 TextRank 알고리즘을 이용한 자동 회의록 생성 시스템

배영준*, 장호택*, 홍태원*, 이해연**

Automatic Meeting Summary System using Enhanced TextRank Algorithm

Young-Jun Bae*, Ho-Taek Jang*, Tae-Won Hong*, Hae-Yeoun Lee*

요 약 다양한 업무 수행에 있어서 회의나 토론 등의 내용을 정리하여 문서화하는 것의 중요성은 매우 높다. 그러나 기존에는 사람이 직접 내용에 대한 정리를 수작업으로 수행하였다. 본 논문에서는 TextRank 알고리즘을 이용하여 자동으로 회의록을 생성하는 시스템의 개발에 대하여 설명한다. 제안한 시스템은 발언자의 모든 발언 내용을 실시간으로 기록하고, 문장들을 출현 빈도수에 기초하여 유사도를 계산한 후, 문서 데이터 안에서 문장들 간의 관계를 찾아내는 비지도 학습 알고리즘을 통해 중요 단어 혹은 문장을 추출함으로서 자동으로 회의록을 생성하도록 하였다. 특히, PageRank 알고리즘을 단어와 문장에 적합하도록 재구성한 TextRank 알고리즘에 대하여 핵심어의 가중치 조정 기법을 도입함으로서 성능 향상을 모색하였다.

Abstract To organize and document the contents of meetings and discussions is very important in various tasks. However, in the past, people had to manually organize the contents themselves. In this paper, we describe the development of a system that generates the meeting minutes automatically using the TextRank algorithm. The proposed system records all the utterances of the speaker in real time and calculates the similarity based on the appearance frequency of the sentences. Then, to create the meeting minutes, it extracts important words or phrases through a non-supervised learning algorithm for finding the relation between the sentences in the document data. Especially, we improved the performance by introducing the keyword weighting technique for the TextRank algorithm which reconfigured the PageRank algorithm to fit words and sentences.

Key Words : Automatic Meeting Summary, Naver CSR, TextRank Algorithm, TF-IDF Model, Weight Adjustments

1. 서론

다양한 업무 수행에서 회의나 토론 등의 내용을 정리하여 문서화하는 것의 중요성은 매우 높다. 이와 같은 문서들은 정책 결정에 있어 책임 소재를 분명히 하고 참석자나 관련 기관 사이에 불신의 소지를 방지하며, 참석자가 책임 있는 발언을 할 수 있도록 유도한다.

이와 같이 회의록의 필요성이 대두되면서 중요한 이사회나 간담회, 주주 총회를 비롯하여 각 종 세미나, 포럼, 재개발, 재건축, 종친회 등 각종 회의에서 회의록의 작성 범위가 넓어지고 있다. 최근 회의가 끝나고 회의에서 누가 어떤 발언을 했었는지, 여러 단체들의 회의록이 전자 문서 형태로 공개되고 사회적 이슈가 되면서 회의의 내용이나 기록에 대한 일반인들의 관심

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1D1 A1B03030432).
* Department of Computer Software Engineering, Kumoh National Institute of Technology
** Corresponding Author : Department of Computer Software Engineering, Kumoh National Institute of Technology (haeyeoun.lee@kumoh.ac.kr)
Received May 31, 2018 Revised June 14, 2018 Accepted June 25, 2018

이 높아지고 있다. 이러한 회의록을 열람하는 사람들은 전체 문서를 읽기 전에 안건들에 대한 주요 의견이나 찬반 의사 등 회의의 전체 흐름을 한눈에 파악하길 원한다. 따라서 회의의 흐름을 파악하고 회의 내용의 핵심 문장을 추출하여 문서를 요약하는 연구가 필요하다.

텍스트 자동 요약은 문서를 적은 노력으로 이해할 수 있도록 입력 텍스트로부터 핵심 내용을 추출하는 연구 분야이다. 이는 크게 생성 요약과 추출 요약으로 구분되며 추출 요약으로는 학습 기반과 비학습 기반이 있다[1]. 학습 기반 추출 요약은 학습된 데이터를 기반으로 추출하는 방식으로, 매번 새로운 회의 주제를 접하고, 매번 새로운 기사가 보도되듯이 문서의 특성이 매우 광범위하기 때문에 매우 비효율적이다. 그렇기 때문에 본 연구에서는 비학습 기반의 추출 요약 방식으로 접근하였다.

본 연구에서 제안하는 회의록 생성 시스템에서는 기존의 비학습 기반 추출 요약 방식인 TextRank 알고리즘을 회의록에 초점을 맞추어서 핵심 회의 내용을 효과적으로 추출하였다. 또한 회의록은 대화의 핵심 내용은 유지하되 함축적이고 전체적인 흐름을 유지해야 함으로 전반적인 흐름과 핵심 내용을 추출해내기 위해 본 연구에서는 TextRank 알고리즘을 향상하여 회의록 특성과 양식에 맞게 접목시켜 발언자들의 중요 내용을 추출하여 요약문을 만들 수 있도록 하였다. 마지막으로 제안된 시스템은 회의록 양식에 맞게 자동으로 최종 회의록을 생성할 있도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구와 시스템에 적용된 알고리즘 배경지식에 대해 소개하고, 제안하는 시스템은 3장에서 설명하며 시스템이 적용된 어플리케이션에 관해 설명한다. 4장에서는 시스템 개발 환경 및 동작 결과를 제시한다. 5장에서는 본 연구의 결론과 개선점을 설명한다.

2. 관련연구

인터넷의 급속한 발전과 더불어 문서 요약에 대한 관심이 고조되면서 문서 요약에 대한 연구개발 투자도 꾸준히 증가하고 있으며 기존에 많은 연구들이 수행되고 있다[2][3].

텍스트 자동 요약은 크게 생성 요약과 추출 요약이 있으며, 추출 요약으로는 학습 기반 추출 요약과 비학습 기반 추출 요약이 있다.

추출 요약은 문서에 존재하는 단어나 구, 문장을 그대로 추출하는 요약 기법이다. 보다 쉬운 접근방법이나 요약문의 응집도가 다소 부족할 수 있다.

생성 요약은 문서의 내용을 압축하여 새로운 문서를 작성하는 요약 방법으로 자연어 이해 및 생성 기술이 필수적이다. 새로운 자연어를 생성해내는 요약 방법이기에 아직 자연어 처리 연구 분야에서 사람과 같은 수준의 자연어 생성이 어렵다. 그래서 기존 토픽 추출 방법을 적용하는 추출 요약이 주로 연구된다[4].

2.1. 도합 유사도(Degree Of Similarity)

김 등(Kim et. al)은 도합 유사도를 이용하여 한국어 추출 문서를 요약하는 시스템을 개발하였다. 이 시스템은 한국어 명사를 추출하여 문서를 의미적으로 관련이 있는 노드(문장 벡터)들 사이의 관계를 나타낸다. 하지만 이 요약 방식은 생성 요약방법으로 아직 사람과 같은 수준의 문장 요약이 어렵다고 판단된다[2].

2.2. 퍼지 이론을 이용한 요약 기술

추출 요약은 문장을 선택할 때 문장의 길이, 핵심 용어, 용어의 빈도수 등 여러 가지 특징들이 고려되어야 하나 최종 문장의 중요도를 반영함에 있어 애매모호한 기준이 문제가 되어왔다. 이에 이 등(Lee et. al)은 퍼지 이론을 이용하여 이러한 불확실성의 문제를 모델링하였다. 각 문장의 특징을 주제, TF-IDF, 제목, 문장 길이, 문장 위치로 나누어 각 각에 대한 중요도를 부여하여 중요도가 부여된 문장들을 퍼지 이론을 사용해서 최종 요약문을 완성한다. 하지만 원본 문서의 불필요한 용어의 처리 등의 문제점은 추후 해결방안으로 남는다[3].

2.3. TextRank 알고리즘

2.3.1. TF-IDF 모델

TextRank 알고리즘은 각 문장 간에 상대적 중요도

에 따라 가중치를 부여하여 중요 문장을 추출해내는 알고리즘이다. 가중치를 부여하는 방식으로 TF-IDF 모델을 적용하며 가중치는 다음 식과 같이 계산된다.

$$tf_{ij} = \frac{f_{ij}}{\max_{t_{kj} \in d_j} f_{ij}} \quad (1)$$

TF(Term Frequency)는 한 문서 전체에서 해당 단어의 빈도수를 나타내며 f는 해당 단어의 빈도수를 의미한다. 그림 1은 해당문서에 단어의 출현 빈도를 나타낸 것이며 추후 또 다른 문서에서도 동일 단어가 지속적으로 등장한다면 해당 단어는 문장의 중요도를 매기는 기준에서 배제된다.

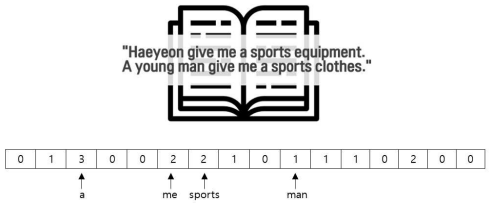


그림 1. Term frequency 예제
Fig. 1. Term frequency example

TF가 한 문서 전체에서 동일한 단어의 빈도수를 체크한다면 DF(Document Frequency)는 그림 2와 같이 한 단어가 전체 문서 집합 내에서 얼마나 공통적으로 많이 등장하는지를 나타낸 값으로서 다음의 식과 같이 계산된다.

$$df(t,d) = \frac{|\{d \in D : t \in d\}|}{D} = \frac{\text{단어 } t\text{가 포함된 문서의 수}}{\text{전체 문서의 수}} \quad (2)$$

이 값의 역수(전체 문서 수 / 해당 단어나 나타난 문서 수)로 각 문서 간 유사도를 비교하는데 이것이 IDF(Inverse Document Frequency)이다. IDF의 분모가 0일 경우를 대비하여 1을 더해주는 경우도 있으며, 값이 커질 경우를 대비해 다음 식과 같이 log로 감싸준다.

$$idf(t,d) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) = \log\left(\frac{\text{전체문서의 수}}{\text{단어 } t\text{가 포함된 문서의 수}}\right) \quad (3)$$

이와 같이 다른 문서에서도 유사도 수치를 기반으로 동일한 빈출 단어가 자주 출현한다면, 그 데이터는 TextRank 알고리즘의 Ranking 기준에서 제외된다. 이렇게 IDF를 이용하여 접속사나, 관사, 조사와 같은 단어들을 제외한 적게 나타나는 단어(명사)를 찾을 수 있게 된다. 이 단어들은 추후 해당 문서에서 중요한 문장인지 판별하는데 중요한 역할을 하게 된다[4].

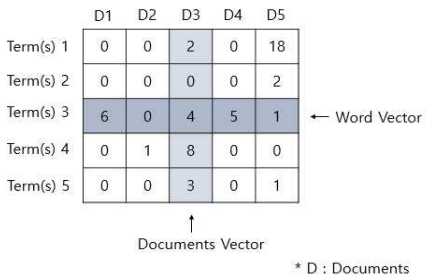


그림 2. Inverse document frequency 예제
Fig. 2. Inverse document frequency example

2.3.2 TextRank Algorithm

TextRank 알고리즘은 문장들 간의 관계를 찾아내는 비지도 학습을 통해 중요 단어 혹은 문장을 추출한다. 이 알고리즘은 Google에서 개발된 PageRank 알고리즘에서 고안해낸 문장 추출 요약 방법으로 PageRank는 하이퍼링크를 갖는 웹 문서를 정점(Vertax)로 바라보지만, TextRank는 하나의 문서에서 각 Sentence를 하나의 정점으로 인식한다.

기본적으로 문서 내의 문장 또는 단어를 이용하여 문장들 간의 관계를 찾아내는 비지도 학습이기에 어떠한 다듬어진 데이터나 어떤 특정 언어 지식 소스를 요구하지 않으며 추가적인 데이터에 대한 요구사항 없이 다른 언어로 이루어진 문서의 요약본에 효율적으로 적용되어질 수 있다[5][6][7].

그림 3은 각 정점(Sentence)들이 핵심어를 기반으로 문장 간 유사도를 참고하여 링크시킨 그래프이다. 이때 핵심어는 모든 불용어를 제거하여 굳이 의미 없는 단어는 참조해서는 안된다.

그래프는 방향성이 없는 무방향 그래프로서 각 인접한 정점은 입력 링크와 출력 링크를 하나씩 갖는다. 입력은 다른 Sentence에 현재 Sentence에 존재하는

단어가 있는지를 체크하고, 출력도 마찬가지로 현재 Sentence에서 다른 Sentence에 단어가 존재하는지 체크한다.

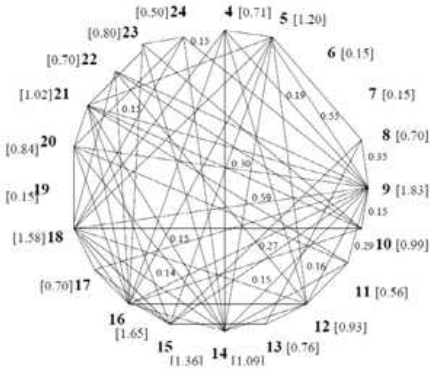


그림 3. 연결된 문장
Fig. 3. Connected sentence

그래프가 만들어지면, 문장의 중요도를 평가하기 위해 다음 식을 통하여 가중치 계산을 수행한다.

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in \ln(V_j)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (4)$$

WS(Vi)는 문장 또는 단어에 대한 TextRank값이고 Wji는 문장 또는 단어 i 와 j 사이의 가중치를 나타내며 d는 현재 문장에서 다른 문장으로 이동할 확률값으로 PageRank에서 제안한 값 0.85를 그대로 사용한다.

3. 제안하는 자동 회의록 생성 시스템

제안한 시스템은 음성인식(STT) 과 향상된 TextRank 알고리즘을 통해 전반적인 회의 내용을 추출하고 요약하여 자동으로 회의록을 생성한다. 전체 시스템의 동작은 다음과 같다. 먼저 그림 4와 같이 음성 인식을 통해 입력된 음성 텍스트가 채팅 서버로 전송되어 사용자들의 발언 내용을 모두 기록한다. 이때 발언 내용은 실시간으로 클라이언트들에게 전송되며, 그림 5와 같이 회의가 끝나면 회의 내용을 회의록 처리 서버로 전송하여 향상된 TextRank 알고리즘으로 추출한 중요 문장들을 기반으로 회의록이 작성되고 요청 클라이언트들에게 전달된다.

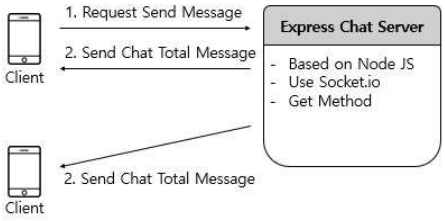


그림 4. 클라이언트 사이의 통신 (데이터 흐름)
Fig. 4. Chat between clients (data flow)

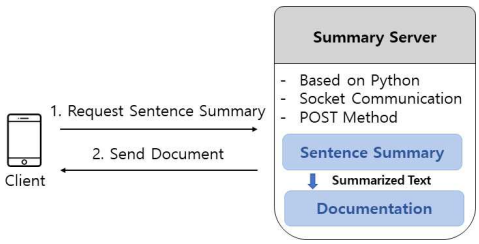


그림 5. 요약 데이터 흐름
Fig. 5. Summary data flow

3.1 자동 회의록 생성 알고리즘

향상된 TextRank 알고리즘을 이용한 자동 회의록 생성 과정은 그림 6과 같다. 먼저 서버에서 회의 내용을 전송받아 문장 단위로 분리시킨 뒤 자연어 처리(NLP) 과정을 거쳐 TF-IDF 모델을 생성한다. 그 후에 각 단어의 가중치를 계산한 후에 가중치 기반으로 그래프를 생성하게 되는데, TF-IDF에서 다 처리하지 못하는 경우를 대비하여 감탄사나 부사, 그리고 그림 7과 같은 불용어들은 다시 따로 처리하여 최종 핵심 문장으로 그래프를 구성한다.

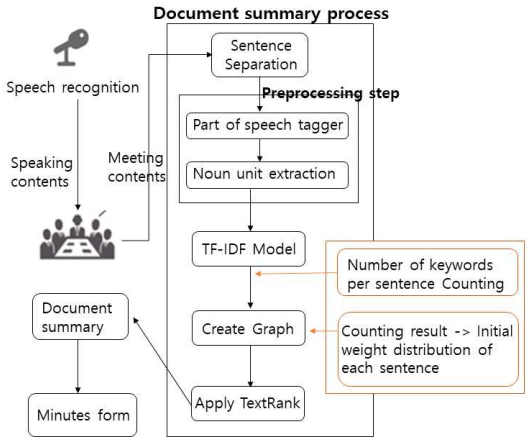


그림 6. 문서 요약 절차
Fig. 6. Document summary process

Stop word

'above', 'doing', 'too', 'can', 'd', 't', 'then', 'what', 'same', 'himself', 'but', 'with', 'on', 'when', 'so', 'isn', 'his', 'further', 'been', 'being', 'our', 'because', 'are', 'from', 'mustn', 'at', 'between', 'here', 'most', 'ours', 'again', 'shouldn', 'have', 'both', 'below', 'against', 'few', 'wasn', 'those', 'hadn', 'once', 'don', 'ain', 'for', 'under', 'o', 're', 'yourselves', 'them', 'themselves', 've', 'about', 'your', 'ourselves', 'who', 'after', 'or', 'he', 'over', 'this', 'how', 'myself', 'into', 'in', 'such', 'aren', 'hasn', 'before', 'whom', 'won', 's', 'were', 'only', 'herself', 'we', 'that', 'was', 'had', 'no', 'of', 'during', 'down', 'has', 'off', 'while', 'where', 'a', 'if', 'until', 'weren', 'be', 'having', 'theirs', 'doesn', 'will', 'to', 'just', 'her', 'ma', 'll', 'there', 'and', 'does', 'other', 'their', 'own', 'why', 'itself', 'its', 'each', 'by', 'not', 'she', 'some', 'him', 'very', 'm', 'should', 'now', 'couldn', 'yourself', 'these', 'as', 'didn', 'an', 'nor', 'is', 'yours', 'did', 'the', 'do', 'my', 'all', 'needn', 'y', 'which', 'up', 'shan', 'haven', 'through', 'me', 'out', 'mightn', 'wouldn', 'they', 'i', 'you', 'hers', 'it', 'more', 'any', 'am', 'than'

그림 7. 불용어 예제
Fig. 7. Stop word example

본 연구에서는 그림 6에 표시한 것과 같이 기존 TextRank 알고리즘의 그래프 생성 방식에서 각 노드(sentence)들의 초기 가중치 할당 방식에 변화를 주었다. 기존 TextRank 알고리즘으로 요약된 회의록은 자주 출현되는 핵심어 하나만 포함된 문장도 링크되어 있는 많은 노드로부터 일정한 가중치를 할당 받을 수 있었다. 그 결과 가중치 낮은 문장도 높은 순위가 할당되어 중요 문장으로 채택되는 경우가 생기게 되었다. 많은 핵심어를 갖는 노드들이 이러한 노드들과 링크되어 많은 노드들에게 가중치를 공유하기 때문이다. 따라서 핵심어가 다수 등장하는 노드의 초기 가중치를 핵심어 개수에 반비례하여 전체적인 각 노드 당 초기 가중치를 재조직하였다.

본 연구에서 제안하는 시스템에서는 단일 핵심어를 갖는 노드는 다수의 핵심어를 갖는 노드로부터 핵심어 개수에 반비례한 초기 가중치의 영향을 받게 된다. 즉, 다수의 핵심어를 갖는 노드는 초기 가중치 이동이 시작될 때, 입력과 관계없이 영향력이 낮은 결과를 내보내도록 설계하였다.

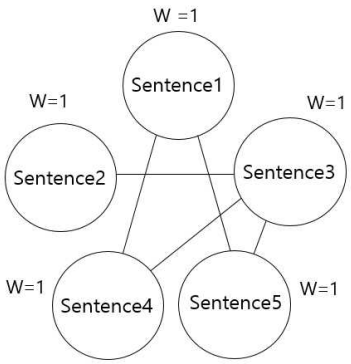
그림 8은 기존 TextRank과 제안하는 초기 가중치를 재분배한 TextRank의 그래프이다. 기존 TextRank 그래프를 보면 초기 가중치가 모두 동일하게 1이다. 여기서 Sentence1은 Sentence4와 Sentence5에게 0.5씩 가중치를 나눠준다. Sentence2는 Sentence3에게 가중치 모두를 나눠주고, Sentence3은 Sentence2에게 받은 가중치 1과 기존 가중치를 합한 가중치를 1/3(0.33)씩 나눠 Sentence2, Sentence4 및 Sentence5에게 각각 나눠준다.

본 시스템에서 제안하는 TextRank 알고리즘은 핵심 단어가 가장 많이 포함되어있는 Sentence에게 초기 가중치를 낮게 준다. 그 이유는 다른 Sentence에도 모두 포함되어 있는 핵심어나 사람의 이름을 포함한 Sentence가 핵심 문장으로 추출되어서는 안되기 때문이다. 그림 8을 보면 Sentence3은 핵심 단어를 다수 보유하고 있는 노드이다. 이 노드는 조금이라도 연관되어 있는 문장이라면 모두 링크되어 있기 때문에 Sentence3에서 나오는 출력 가중치는 각 노드들에게 핵심 문장을 판별하는데 중요한 역할을 하지 못한다. 그래서 Sentence3의 초기 가중치를 낮추어 초기 출력의 영향력을 최소화한다. 이러한 초기 가중치의 할당은 추후 학습이 끝나게 되면 회의록에 더욱 핵심 문장이 되는 문장을 추출해 낼 수 있다.

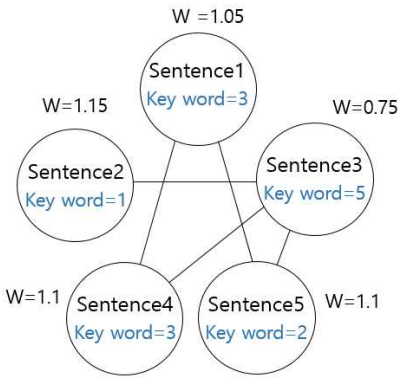
4. 실험결과

음성 인식의 결과 발언자들의 발언 내용을 약 90% 정확도로 텍스트화가 가능하였고, 제안한 시스템의 성능을 평가하기 위해 대한민국 국회 회의록 시스템의 전자 회의록 문서를 사용하였다[8]. 본 연구에서는 제20대 국회 본회의 제 343회 회의록부터 제 353회 회의록까지 총 54개의 회의록에서 개회사를 제외한 49개의 회의록에 하여 평가를 수행했다. 해당 49개의 회의록에 대해서 수작업으로 문장을 분석하여 평가 기준

데이터를 구축하였고, 요약 비율은 20%~40%로 하였다.



(a) Existing TextRank



(b) Proposed TextRank

그림 8. 기존 TextRank 및 제안 TextRank 알고리즘의 초기 가중치
Fig. 8. Initial weights of existing TextRank and proposed TextRank algorithms

4.1 시스템 개발 환경

회의 종료 후 회의록 요약을 위한 TextRank 알고리즘을 수행하기 위하여 다음과 같은 사양에서 Python 언어를 사용하여 구현하였다. 클라이언트인 Android Application은 Windows 환경에서, 회의록 요약을 위한 TextRank Server는 Ubuntu 환경에서 개발하였고 동일한 하드웨어 사양에서 진행하였다. 회의 관리를 위한 Chat Server는 Node Js를 이용하여 구현하였다. 표 1과 표 2에는 각각 하드웨어 사양, 소프트웨어 환경에 대하여 기술하였다.

표 1. 하드웨어 스펙
Table 1. Hardware specifications

Item	Spec.
CPU	Intel G4600
RAM	12GB (8GB * 1, 4GB * 1)

표 2. 소프트웨어 스펙
Table 2. Software specifications

Item	Environment
OS	Windows 10 Pro, Ubuntu 16.04
Android	Android Studio 2.3.3
Python	3.5
Chat Server	Node JS

4.2 시스템 성능 평가

본 연구에서는 입력된 음성 데이터 요약의 상용 가능성, 요약된 문장의 완성도로 성능 평가 기준을 나누어 각 과정에 맞게 따로 평가하였다. 음성 인식은 NAVER CSR(Clova Speech Recognition) Open API를 사용하였으며 상용 가능성 평가는 음성 인식된 전자 회의록 내용과 오차가 없는 완벽한 전자 회의록 내용을 각각 추출 요약하여 비교 분석하였다. 그리고 요약된 문장의 완성도 성능 평가는 수작업 된 평가 기준 데이터로 기존 TextRank와 제안하는 향상된 TextRank 두 가지 알고리즘으로 나누어 성능 평가하였다.

그림 9에는 NAVER CSR Open API를 사용하여 음성 인식의 정확도를 분석한 것으로서 평균 90.5% 정도의 인식률을 보였다. 표 3은 국회 회의록 전자문서 343~346회 회의 내용과 음성 인식을 통한 회의내용 요약본의 오차율이다. 가로축은 회의록, 세로축은 요약 비율에 따른 오차율과 오차 문장 개수(괄호 안의 수)를 나타낸다. 7개의 문서 중 높은 음성 인식률을 보인 346회 14차, 16차 회의는 추출된 문장의 오차가 거의 없었으며, 평균 오차율은 각각 10.5%, 9.8%, 12.3%로 준수한 수치가 나왔다. 음성 데이터의 요약된 문장과 오차가 없는 완벽한 문서의 요약된 문장의 오차 문장은 평균 1~2개정도로 추후 상용 가능성은

긍정적이다.

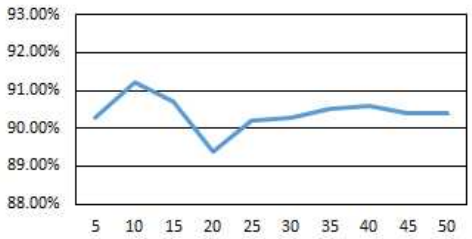


그림 9. 음성 인식 정확도
Fig. 9. Speech recognition accuracy

표 3. 음성 인식된 문서의 오류 비율
Table 3. Error rate of speech-recognized documents

Congressional meeting	10% Extraction	20% Extraction	30% Extraction
343th 4	19%(1)	22%(2)	21%(3)
346th 8	0%	11%(1)	20%(3)
346th 9	0%	0%	6%(1)
346th 10	18%(1)	27%(3)	26%(4)
346th 13	37%(2)	9%(1)	6%(1)
346th 14	0%	0%	7%(1)
346th 16	0%	0%	0%
Total	10.5%(0.5)	9.8%(1)	12.3%(1.8)

앞서 TextRank 그래프의 초기 가중치를 차등 분배하는 방식을 제안하였다. 문장 요약에 관한 성능 평가는 이와 기존 TextRank 두 가지 방법으로 결과를 측정하였다. 첫번째 기존 TextRank 요약 방법은 그래프 생성 시 모든 노드들은 초기 가중치를 일정하게 부여한다. 두 번째 앞서 제안한 요약 방법은 그래프 생성 시 각 노드의 핵심어 개수에 반비례하는 초기 가중치를 분배하여, 필요시 해당 노드의 입력과 관계없이 초기 출력의 질을 낮추면서 더욱 중요한 문장이 추출될 것이라 기대하였다. 모든 실험에서 요약 비율을 10%에서 최대 40%까지 설정하여 결과를 분석하였다.

실험 결과에 따르면 요약 정확도는 80% 이상을 웃도는 높은 정확도를 보였다. 기존 TextRank 알고리즘이 중요 문장을 잘 추출해 주었다는 것을 의미한다. 그림 10을 보면 요약 비율 20%까지 기존과 다르지 않게 높은 비율로 중요 문장이 동일하게 추출되고 있으나 추출되는 문장이 늘어날수록 초기 가중치를 변경한 TextRank가 상대적으로 더 정확한 문장을 추출하였다. 그 결과 회의 전반적인 흐름을 필요로 할 시, 높은 요약 비율로도 의미 기반 중요 문장을 추출할 수 있었다. 그러나 여전히 비중 낮은 문장이 추출되는 경우가 가끔 생겼고, 핵심어 빈도수에 의존하기 때문에 문장의 흐름을 고려하지 못하는 단점이 있었다.

그림 10. 알고리즘 정확도 비교
Fig. 10. Algorithm accuracy comparison

5. 결론

다양한 업무 수행에 있어서 회의나 토론 등의 내용을 정리하여 문서화하는 것의 중요성은 매우 높으며 본 연구에서는 회의록의 특징을 반영하여 자동으로 회의록을 요약하는 시스템을 제안하였다. 제안하는 시스템은 음성 인식으로 입력된 음성 텍스트를 향상된 TextRank 알고리즘을 이용해 중요 문장을 추출하여 자동으로 회의록을 생성한다.

TextRank 알고리즘을 회의록의 특성에 맞게 개선시켜 향상된 결과를 얻었지만TextRank 유사도 계산 방식은 문장 내 단어 간의 의미적 유사성을 완벽히 고려하지는 못하는 문제점이 존재하였다. 그 결과 여전히 비중 낮은 문장이 추출되는 경우가 생겼고, 문장의 흐름을 고려하지 못하는 단점이 있었다. 그렇기에 아직까지 이런 문제점을 해결하기 위하여 유사도 계산 방법을 그래프로 정의하는 방법 등 지속적인 연구를 필요로 한다.

REFERENCES

[1] F. Cruz, J. A. Troyano, F. Enriquez, "Supervised TextRank," Lecture Notes in Computer Science, vol. 4139, pp. 632-639, 2006.
[2] J.-H. Kim, J.-H. Kim, "Korean Indicative Sum

marization Using Aggregate Similarity," Proceedings of the Annual Conference on Human and Cognitive Language Technology, pp. 238-244, 2000.
[3] J.-P. Hong, J.-W. Cha, "Korean Important Sentence Extraction using TextRank Algorithms" Proceedings of the Korea Computer Congress, vol. 36(1C), pp. 311-314, 2009.
[4] S.-J. Moon, S. Lee, "Automatic Document Summary Technique Using Fuzzy Theory," KIPS Transactions on Software and Data Engineering, vol. 3(12), pp. 531-536, 2014.
[5] D. Hiemstra, "A probabilistic justification for using tf×idf term weighting in information retrieval," International Journal on Digital Libraries, vol. 3(2), pp.131-139, 2000.
[6] I. Mani, Automatic Summarization, John Benjamins Publishing Company, Vol. 3. 2001 (ISBN 9789027299109).
[7] F. Barrios, F. Lopez, L. Argerich, R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization," Proceedings of the Argentine Symposium on Artificial Intelligence, pp. 65-72, 2016.
[8] The National Assembly Information System, http://likms.assembly.go.kr/record/index.jsp, 2018

저자약력

배 영 준(Young-Jun Bae) **[학생회원]**

- 2012년 - 현재 금오공과대학교 컴퓨터소프트웨어공학과 (재학중)

<관심분야> 자연어 처리, 사물 인터넷

장 호 택(Ho-Taek Jang) **[학생회원]**

- 2012년 - 현재 금오공과대학교 컴퓨터소프트웨어공학과 (재학중)

<관심분야> 자연어 처리, 사물 인터넷

홍 태 원(Tae-Won Hong) **[학생회원]**

- 2012년 - 현재 금오공과대학교 컴퓨터소프트웨어공학과 (재학중)

<관심분야> 자연어 처리, 사물 인터넷

이 해 연(Hae-Yeoun Lee) **[정회원]**

- 1993년 - 1997년 성균관대학교 정보공학과 (공학사)
- 1997년 - 1999년 KAIST 전산학과 (공학석사)
- 1999년 - 2006년 KAIST 전자전산학과 (공학박사)
- 2006년 - 2007년 코넬대학교 박사후연구원
- 2008년 - 현재 금오공과대학교 교수

<관심분야> 영상처리, 멀티미디어, 디지털 포렌식