KIPS Tr. Software and Data Eng.
Vol.3, No.12 pp.531~536 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.12.531

Automatic Document Summary Technique Using Fuzzy Theory

Sanghoon Lee† · Seung-Jin Moon††

ABSTRACT

With the very large quantity of information available on the Internet, techniques for dealing with the abundance of documents have become increasingly necessary but the problem of processing information in the documents is still technically challenging and remains under study. Automatic document summary techniques have been considered as one of critical solutions for processing documents to retain the important points and to remove duplicated contents of the original documents. In this paper, we propose a document summarization technique that uses a fuzzy theory. Proposed summary technique solves the ambiguous problem of various features determining the importance of the sentence and the experiment result shows that the technique generates better results than other previous techniques.

Keywords : Document Summary Technique, Fuzzy Theory, Topic Model, Sentence Importance, Information Processing, Information Retrieval

퍼지이론을 이용한 자동문서 요약 기술

이 상 훈† · 문 승 진††

요 약

인터넷에서 사용 가능한 수많은 정보로 인해서 대용량의 문서를 다루는 기술은 점차 그 필요성이 증가되어 왔지만, 효과적으로 문서 내 정보를 처리하기 위한 기술의 문제는 여전히 풀어야 할 과제로 남아 있다. 자동문서 요약 기술은 문서 내 중요한 부분을 유지하고, 중복된 내용을 제거함으로써 이러한 대용량의 문서를 처리하는 데 중요한 방법으로 인식되어 왔다. 본 논문에서는 이러한 요약문을 만들 때 중요도를 결정하는 문제를 해결하기 위해서 퍼지 이론을 이용한 문서 요약 기술을 제안한다. 제안된 요약 기술은 중요도를 결정하는 여러 특징들의 애매모호한 문제를 해결하고, 그 실험결과는 기존의 다른 방법과 비교해서 전반적으로 높은 결과를 보인다.

키워드 : 자동문서 요약 기술, 퍼지이론, 토픽 모델, 문장의 중요도, 정보처리, 정보검색

1. 서 론

인터넷을 통한 정보의 급속한 증가로 인해서 개개인이 수많은 정보를 처리하기에는 한계가 있기 때문에, 컴퓨터로 이러한 대용량의 문서를 자동적으로 처리하고 요약하는 기술은 점차 그 효율성이 증가 되어왔다. 흔히 문서를 요약한다는 것은 그 문서의 일관성을 유지하면서 중복을 제거하고 응축된 정보를 생산하는 것을 말하는데, 자동문서 요약 기술은 컴퓨터를 사용해서 문서 내 중요한 부분을 유지하고, 중복된 내용을 제거함으로써 처리하고자 하는 대용량의 문서를 자동적이고 효율적으로 처리하는 방법을 말한다.

일반적으로 문서를 요약하는 기술은 요약하고자 하는 용어나 문장을 선택할 때 문장의 길이, 용어 빈도수, 문장의 위치 등 여러 가지 특징들을 고려해야 하는데, 이러한 특징들을 함께 사용할 때 최종 문장의 중요도를 어떻게 반영해야 하는지가 애매모호한 문제로 제기되어 왔다. 퍼지 이론은 이러한 불확실성의 문제를 모델링하는 데 장점을 가지고 있다. Rene Witte[1]는 자연어 처리에서 어떻게 불확실성의 문제가 퍼지 이론을 사용해서 모델링 될 수 있는지 제안했었고, Ladda[2]는 제목(Title), 문장 길이(Sentence Length), 용어 가중치(Term Weight), 문장의 위치(Sentence Position), 문장들 간의 유사도(Sentence to Sentence Similarity), 적절한 명사/개체명(Proper Noun/Named Entity), 주제어(Thematic Word), 그리고 숫자로 나타낸 데이터(Numerical Data) 등의 8개의 특징을 사용해서 요약 문장을 추출하는 데 퍼지이론

† 준 회 원 : 조지아주립대학교 대학원 컴퓨터학과 박사과정
†† 종신회원 : 수원대학교 컴퓨터학과 교수
Manuscript Received : June 13, 2014
First Revision : August 22, 2014
Accepted : October 18, 2014
* Corresponding Author : Seung-Jin Moon(sjmoon103@hotmail.com)

을 이용했다. 또한 Ravindra[3]는 퍼지이론을 사용해서 자동 평가 방법을 제안했다.

본 논문에서 제안된 자동문서 요약 기술은 퍼지 이론을 사용한 방법으로 두 가지 측면에서 의의를 지니고 있다. 첫 째, 본 논문에서는 Latent Dirichlet Allocation (LDA) 모델[4] 사용해서 문서상에서 주제어를 추출하였다. Ladda의 경우 주제어를 선택할 때 가장 높은 빈도수를 가지는 10개의 용어를 주제어로 분류하였지만, 빈도수만으로 주제어를 결정하는 것은 너무 직관적인 방법으로 주제와 관련 없는 용어의 빈도수가 증가할 때, 요약된 문서의 질은 낮아질 수 있는 단점이 있다. 또한 적절한 용어/개체명의 비율이 높을수록 문장의 중요도가 높다는 방법 역시 직관적인 방법으로 관련이 없는 개체명의 비율이 높아질 때 요약된 문서의 질 역시 낮아지게 된다. LDA 모델에 기반한 주제어의 검출은 이러한 문제들을 해결하고 최종적으로 요약된 문서에서 특정 주제에 대한 검색의 향상을 기대할 수 있게 한다. 또한 숫자로 나타낸 데이터의 중요도 역시 문서의 종류나 주제에 따라 달라질 수 있는 부분이기 때문에, 본 논문에서는 고려하지 않기로 한다. 둘째, 제안된 문서 요약 시스템은 컴퓨터 평가방법을 사용한다. 현재 가장 많이 알려진 자동문서 요약의 평가 방법은 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[5] 방법으로 특정한 사람들이 요약한 문서(Gold Model)와 컴퓨터가 요약한 문서(System Model)를 비교해서 Precision과 Recall을 0과 1 사이의 값으로 정한 후 1에 가까울수록 좋은 요약문이라고 결정하는 방법이다. 하지만 기존의 방법은 개인의 평가방법에 의존했기 때문에 요약에 참여하는 개인의 관심 분야나 전공분야가 다를 때 결과 역시 변하는 문제를 가지고 있다. 본 논문에서 사용한 평가 방법은 이러한 개인별 성향에 따라 다를 수 있는 치명적인 평가 오류를 제거하기 위한 대안으로 원본 문서와 요약문서의 분배(distribution) 사이의 발산(divergence)를 계산함으로써 요약된 문서에 대한 평가를 시도하였다.

2장에서는 제안된 요약 기술에 대한 단계적 정의와 설명을 하도록 하고, 3장에서는 제안된 요약 기술을 평가하고 그 결과에 대해서 기술하기로 한다. 마지막으로 4장에서는 결론과 향후 연구과제에 대해서 설명하도록 한다.

2. 본 론

2.1 퍼지이론을 이용한 문서 요약 기술

일반적으로, 자동문서 요약은 두 가지 방법으로 이루어져 있다. 첫 번째는 문서의 원본을 그대로 유지한 채 발췌하는 추출식 요약(extractive summary) 방법이 있고, 두 번째는 문서의 원본과 다른 새로운 문장을 만들어 내는 추상적 요약(abstractive summary) 방법이 있다. 본 논문에서의 방법은 추출식 요약에 기반하고 있다. 추출식 요약은 문서의 원본을 바꾸지 않기 때문에 오류가 적고 새로운 문장을 만들어 낼 비용을 줄이기 때문에 대부분의 문서 분석 연구에서 사용되고 있다. 이번 장에서는 이러한 추출식 요약 방법을 근간으로 한 자동화 문서 요약 기술에 대해서 설명하도록 한다.

1) 제안된 기술 흐름도

그림 1은 제안된 기술의 단계별 흐름을 보여준다. 흐름은 크게 세 부분으로 나누어진다. 첫 번째는 전처리(Preprocessing) 단계로 수집된 문서를 문장으로 나누고(Sentence Boundary Detection), 불필요한 용어를 제거해서(Stop-word Removing), 추출된 용어의 어근을 분리한다(Stemming). 두 번째 단계는 각 문장의 특징을 주제(Topic), 용어 빈도수와 역 문장 빈도수(TF・ISF-Term Frequency and Inverse Sentence Frequency), 제목(Title), 문장 길이(Length), 문장 위치(Position)로 나누어 각각에 대한 중요도를 부여한다. 세 번째 단계는 중요도가 부여된 문장들을 퍼지 이론을 사용해서 최종 요약문을 완성한다.

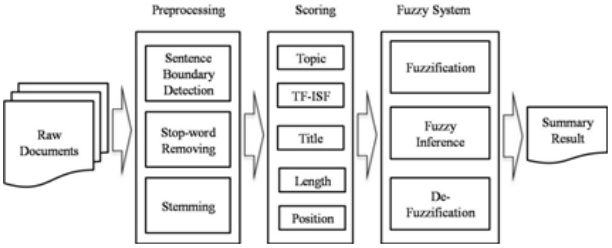

Fig. 1. Data flow of proposed technique

2) 전처리 단계

전처리 단계는 문장 영역 구분(Sentence Boundary Detection), 불필요한 단어 제거(Stop-word Removing), 제거된 단어의 어근 추출(Stemming)의 세 부분으로 구성된다. 본 논문에서는 문장 영역을 구분하기 위해서 발견법(Heuristic)을 사용하지 않고 문장 영역을 구분하는 지도 학습 시스템(Supervised Learning System)을 이용하였다. 학습 데이터는 Brown Corpus와 함께 Wall Street Journal news를 사용하였고, 사용된 news의 경우 0.25%의 error 비율을 나타내었다[6]. 이 모델은 mxTerminator[7]와 동일한 훈련 데이터를 사용하였다. 이후 생성된 요약문은 3장에서 다양한 news의 문서로 평가될 것이다. 일단 문장이 구분이 되면 문장에서 불필요한 용어를 제거할 필요가 있다. 이 작업은 이후 문서 사이의 유사도를 측정하는 데 중요한 부분을 담당하기 때문에 대부분의 문서 분석 영역에서 사용하고 있다. 하지만 전통적으로 불필요한 단어의 정확한 의미를 부여하기 어렵기 때문에, 이 부분은 여전히 문제점으로 남아 있다. 본 논문에서는 현재 가장 많이 쓰이고 있는 불필요한 단어 리스트를 사용해서 제거했다. 문장이 구분되고 불필요한 단어가 제거되면 전처리 단계의 마지막 단계인 어근 추출의 단계로 들어가게 된다. 어근은 단어의 가장 중심이 되는 형태소로 뜻이 같은 말의 최소단위이다. 본 논문에서는 영문을 평가 기준으로 삼았기 때문에 현재 가장 많이 쓰이고 있는 알고리즘인 Porter Stemmer[8]를 사용해서 어근을 추출하였다.

3) 중요도 부여 단계

이 단계에서는 문서내의 각 문장에 대해 주제, 용어 빈도수와 역 문장 빈도수, 제목, 문장 길이, 그리고 문장 위치 등으로 어떻게 중요도를 정의했는지 자세히 설명하도록 한다.

・주제: LDA 모델링을 위한 툴 박스는 David Newman[9]에 의해서 개발되었다. 본 논문은 원본 문서에서 주제를 추출하기 위해서 개발된 툴 박스를 사용하였다. 훈련 단계를 위한 입력으로 Newsblaster 시스템 저장소[10]를 사용하였고, 각 주제별로 최대 20개의 주제어를 정하였다. 원본 문서로 부터 추출된 각 주제어들은 각각의 문장에서 중요도를 부여하는데 사용된다. 주제에 대한 문장 중요도는 아래와 같이 계산된다.

$$Topic_s = \frac{N_t}{Max(N_t)} \tag{1}$$

N_t는 한 문장에 있는 주제어 수를 나타내고, $Max(N_t)$는 한 문장에서 나타날 수 있는 주제어 수의 최대치를 나타낸다. 주제에 대한 문장의 중요도는 주제어의 최대치에 대한 주제어의 수가 많아질수록 높은 값을 가진다. 본 논문에서는 주제어의 최대치가 너무 커질 경우 실제 주제에 대한 중요도가 의미가 없어질 수 있기 때문에, 한 문장에서 최대로 나타낼 수 있는 주제어 수를 threshold 값으로 정하고 실험을 할 때는 이 threshold 값을 경험적으로 판단해서 20으로 정하고 실험을 하였다.

・용어 빈도수와 역 문장 빈도수: 일반적으로 사용되는 가중치 부여 방법으로 용어 빈도수와 역 문서 빈도수(TF・IDF)가 있다. 용어 빈도수(TF)는 문서에서 발생하는 용어의 빈도수를 의미하고, 역 문서 빈도수(IDF)는 용어가 몇 개의 문서에서 나왔는지를 의미한다. 본 논문에서는 TF・IDF[11]의 변형인 TF・ISF[12]를 이용하도록 한다. 여기서 ISF는 역 문장 빈도수로 용어가 몇 개의 문장에서 나왔는지 계산하는 척도로 사용된다. 아래는 본 논문에서 사용될 TF・ISF에 대한 중요도를 정의한다.

$$TF \cdot ISF = TF(t,s) \times ISF(t) \tag{2}$$

$$ISF = 1 + \log\frac{T_s}{N_s + 1} \tag{3}$$

$$TFISF_s = \frac{\sum_{i=1}^{n} TF_i \cdot ISF_i}{Max(\sum_{i=1}^{n} TF_i^k \cdot ISF_i^k)} \quad (k = 1,2,3,\cdots,m) \tag{4}$$

TF(t, s)는 한 문장 s에서 용어 t의 개수이고, ISF(t) 는 용어 t의 역 문장 빈도를 말한다. k는 한 문서 내의 문장의 수이고, T_s는 한 문서에서 총 문장의 수를 말하며, N_s는 용어 t가 발생하는 문장의 총 개수를 말한다.

・제목: 두 집합 사이에 유사점을 측정하는 대표적인 방법으로 Jaccard[13] 유사도 계수가 있다. 본 논문에서는 각 문서별 제목의 중요도를 측정하기 위해서 Jaccard 유사도 계수를 사용하였다. 즉 한 문장에 포함된 용어가 제목에 더 포함될 때, 그 문장의 중요도는 높아진다. 아래는 본 논문에서 정의한 제목에 대한 중요도를 측정하는 방법이다.

$$Title_s = S_t \cap T_t / S_t \cup T_t \tag{5}$$

S는 문장에서의 용어를 나타내고, T는 제목에서의 용어를 나타낸다.

・문장의 길이: 문장의 길이 역시 문서 요약에 많은 영향을 줄 수 있다. 본 논문에서는 문서의 길이에 따른 문장의 중요도를 측정하기 위해서 문장의 길이에 따른 정규화를 정의하였다. 즉 문장의 길이가 길어질수록 그 문장의 중요도는 낮아진다고 가정한다. 일반적으로 용어의 수가 클 때 중요도는 적어지고, 용어의 수가 적을 때 중요도는 높아진다. 아래는 본 논문에서 정의한 문장의 길이에 따른 중요도 측정방법이다.

$$Length_s = 1 - (L/Max(L)) \tag{6}$$

L은 문장에서 용어의 길이를 나타내고, Max(L)은 한 문장에서 나타낼 수 있는 용어의 최대 길이를 말한다. Max(L)은 각 문서마다 다르게 나타나지만 해당 문서의 Max(L)은 일정하다.

・문장의 위치: 문장의 위치별 중요도는 문서에서 나타난 첫 헤드라인이 문장의 중요도에 긍정적인 영향을 끼치고, 주제어가 한 문서에서 가장 일찍 혹은 가장 늦게 나타난다는 가정을 한다. 따라서 문장의 위치에 대한 중요도는 아래와 같이 계산된다.

$$Position_s = F_p + L_p \tag{7}$$

F_p와 L_p는 각각 처음과 마지막 문장까지의 중요도를 말한다. 다시 말해서 F_5는 시작부터 다섯 문장까지의 중요도를 말하고 L_5는 끝에서 다섯 문장까지의 중요도를 말한다. 각 중요도는 차등적으로 0과 1 사이의 값을 가지는데, 여기서 첫째 문장의 중요도는 둘째 문장의 중요도 보다 높고, 둘째 문장의 중요도는 셋째 문장의 중요도보다 높다. 이런 방법으로 문장의 위치별 중요도에 차등을 두게 된다면 가장 먼저 나오는 첫째 문장과 가장 나중에 나오는 마지막 문장이 가장 높은 중요도를 가지게 된다. 결국 주제어가 문장의 위치에 따라 가장 일찍 혹은 가장 늦게 나타난다는 가정에 따라 계산된 각각의 중요도의 값들은 최종적으로 더해져서 Score에 반영하게 된다.

4) 퍼지 이론 적용 단계

퍼지 이론은 애매모호한 상태를 이진논리가 아닌 다치성으로 표현하는 논리로 L.A. Zadeh[14]에 의해서 발표되었다. 퍼지 이론에서 각 집합은 소속 함수(membership function)로 나타낼 수 있는데 각각 그 집합에 해당하는 정도를 수학적으로 나타낼 수 있다. 본 논문에서는 이러한 퍼지 이론의 특징을 이용해서 자동문서 요약 기술을 구현하였다. 일반적으로 퍼지 시스템의 입력 값은 퍼지집합이라고 불리는 소속함수로 변환될 수 있고, 확실한(crisp) 입력 값은 퍼지화(fuzzification) 작업 동안 퍼지 값으로 변환될 수 있다. 또한 이러한 값들은 규칙에 근거해서 (rule base) 퍼지 추론(fuzzy inference) 을 거쳐 비퍼지화(defuzzifier) 작업을 거친 후 출력 값이 결정된다. 본 논문에서는 이러한 퍼지 기술을 이용해서 원본 문서에서 중요한 문장을 추출한다. 퍼지 입력을 위한 소속 함수로 3)의 중요도 부여 단계에서 정의한 다섯 가지 특징을 이용하였다. 각 특징별 퍼지 집합은 아래와 같이 정의된다.

$$S_{fuzzyset} = \{(w, \mu(w)) \mid w \in W\} \tag{8}$$

S는 퍼지 집합을 나타내고, w는 중요도의 정도를 나타낸다. μ(w)는 소속 함수를 말한다.

그림 2는 주제의 소속 함수를 그래프로 표현한 것이다. X축은 주제의 중요도를 나타내고 Y축은 소속 함수의 값을 나타낸다. 퍼지 추론을 위해서 우리는 최소 t-norm을 아래와 같이 정의한다.

$$\mu_{s1} AND \mu_{s2} = \min(\mu_{s1}, \mu_{s2}) \tag{9}$$

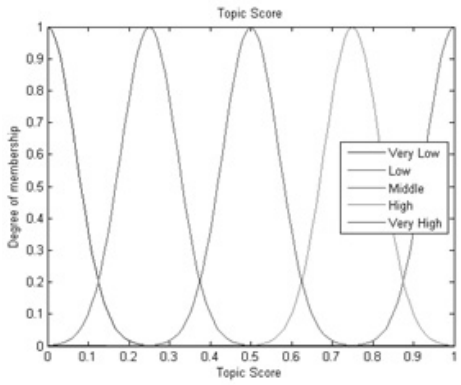

Fig. 2. Topic score membership function

각 소속 함수에 대한 규칙은 교집합으로 표현되고 이것은 출력 퍼지 집합을 생성한다. 문장의 중요도를 결정하는 다섯 가지의 요소 각각은 {Very Low, Middle, High, Very High}의 같은 스케일로 구성하였고, 이를 결합하기 위한 규칙은 주제의 중요도가 High AND TF·ISF의 중요도 값이 High AND 제목의 중요도가 High AND 문장 길이의 중요도 값이 High AND 문장의 위치 값이 High, Then 중요도의 결과는 Very High라고 정하였다. 마지막으로 비퍼지화 작업을 위해서 본 논문은 최대 최소 추론을 사용하는 무게중심 비퍼지화기(centroid defuzzifier)를 아래와 같이 정의한다.

$$Center\ of\ Gravity = \frac{\sum_{w=1}^{n} (\mu_s(w) \times L_w)}{\sum_{w=1}^{n} \mu_s(w)} \tag{10}$$

μ(w)는 각 가중치의 질량(mass)를 의미하고, L은 그 질량의 위치를 나타낸다.

3. 실 험

일반적으로, 많은 문서 요약 시스템을 평가하기 위한 방법으로 개인이 요약한 문서를 기반으로 하는 방법이 사용되고 있다. 즉 연구자들이 개인이 만든 요약문을 그들이 만든 시스템의 요약문과 비교하여 평가를 내리는 방법을 말한다. 하지만 이러한 방법은 각 개인마다의 차이로 인해 결과가 너무 유동적이라는 단점이 있다. Annie Louis[15]는 이러한 단점을 극복하기 위해서 개인의 평가방법에 의존하지 않고 요약문을 평가하는 방법을 제시하였다. 그녀는 입력 문서와 특정한 요약문을 비교해서 평가된 연관성에 기반해서 인간을 이용한 방법과 비교하였고, 그 결과 개인과 거의 비슷한 결과를 보여주었다. 본 논문에서는 이 평가방법을 사용해서 원본 문서와 요약문을 비교하도록 한다.

3.1 실험 방법

SIMetricx[15]은 위에서 언급한 평가 프로세스를 위해 사용된 개발 툴이다. 이 툴을 사용함으로써 원본 문서와 요약문서의 용어 분배(distribution) 사이에 발산(divergence)을 계산할 수 있다. 실험 평가를 위해 사용된 데이터는 Annie Louis[15]가 사용했던 Newsblaster 시스템 저장소[10]에서의 데이터를 그대로 사용하였다. 먼저 모든 문서에서 HTML 태그 등을 제거한 순수한 텍스트로만 사용하였고, 요약된 문서를 비교하기 위해서 두 개의 baseline 요약문과 두 개의 다른 시스템을 사용한 요약문을 비교 대상으로 정하였다. Baseline1은 각 문서에서 첫 번째 문장 순으로 추출한 요약문으로 문장의 길이가 총 5문장이 되면 문장의 길이에 따라 다른 문서의 첫 번째 문장이 기존 요약된 문장과 교체되는 요약문이다. 이것은 문장의 순서가 요약문에 영향을 끼치는 정도를 알아보기 위한 실험요소이다. Baseline2는 각 문서에서 중요도가 가장 높은 문장만으로 이루어진 요약문이다. 이 방법은 문장의 중요도가 요약문에 끼치는 영향을 알아보기 위한 실험요소이다. 또한 News Blaster 시스템에서 사용된 요약문과 현재 가장 많이 쓰이는 비교방법인 MEAD[16]를 사용한 요약문을 본 논문에서 구현된 요약문과 비교하도록 한다. MEAD는 요약문에 가장 많이 쓰이는 플랫폼으로써 여러 요약문을 비교하는 데 사용되어왔다.

본 논문에서는 Annie Louis가 사용한 측정 방법들을 사용하였다. Kullback Leibler (KL) 발산[17]은 두 개의 확률 분포 사이의 차이점을 측정하기 위해 사용되는 비대칭 방법으로 이를 위해서 본 논문에서는 두 확률 분포를 가지는 원본 문서와 요약문서 사이에 낭비되는 용어의 수의 평균으로 발산을 측정하였다. 또한 KL 발산은 대칭이 아니기 때문에 본 논문에서는 반대의 경우도 고려하였다. 즉 원본문서와 요약문서 그리고 요약문서와 원본문서의 순으로 KL 발산을 측정하였다. Jensen Shannon (JS) 발산[18]은 두 확률 분포 사이에 유사도를 측정하기 위한 가장 잘 알려진 방법이다. JS 발산은 분배 집합을 고려하기 위해서 제안된 KL 발산의 확장형 개념으로 본 논문에서는 JS 발산 역시 smoothed와 unsmoothed의 두 가지로 나누어 평가하도록 한다. JS 발산을 위한 smoothed 측정 방법은 요약문과 원문 문서의 확률이 0이 될 때의 문제를 피하기 위해서 고려된 방법이다. 모든 값은 0에 가까울수록 요약된 문서의 성능이 높다는 것을 의미한다. Cosine-S는 코사인 유사도를 측정하기 위한 방법으로 모든 원본 문서와 요약된 문서 사이의 코사인 유사도를 계산해서 두 문서 사이를 평가하는 데 사용하였다. 유사도는 0에서 1 사이의 값을 갖고 1에 가까울수록 유사도가 크다는 것을 의미한다. P-Topic 는 요약문에서 주제어의 비율을 계산하는 방법으로 입력 문서의 주제어들 중에 얼마나 많은 주제어가 요약문에 포함되는지의 정도를 나타낸다. T-Overlap은 코사인 유사도와 개념은 같지만 주제어만 포함해서 유사도를 측정하였다. P-Topic과 T-Overlap 둘 다 0과 1 사이의 값을 가지며, 1에 가까울수록 좋은 결과를 나타낸다. 아래는 앞서 언급한 평가를 위한 각 측정 방법을 정리한 내용이다.

- KL: 원본 문서와 요약문 사이의 Kullback Leibler divergence
- KL (R): 요약문과 원본 문서의 Kullback Leibler divergence
- JS: Jensen Shannon divergence (unsmoothed)
- JSS: Jensen Shannon divergence (smoothed)
- Cosine-S(Cosine Similarity): 모든 원본 문서와 요약문 사이의 Cosine 유사도
- P-Topic(Percent of Topic): 요약문에서 주제어의 비율
- T-Overlap(Topic word Overlap): Cosine 유사도 단, 원본으로부터의 주제어만 포함

3.2 실험 결과

표 1은 요약문들에 대한 실험 결과를 보여준다. KL, KL (R), JS, 그리고 JSS 등의 높은 발산 숫자는 요약문의 퀄리티가 낮다는 것을 의미하고, 반대로 낮은 발산 숫자는 요약문의 성능이 높다는 것을 의미한다. Cosine-S, P-Topic, 그리고 T-Overlap 등은 0과 1 사이의 값을 갖고 1에 가까울수록 요약문의 성능이 높다는 것을 의미하고, 반대로 0에 가까울수록 요약문의 성능이 낮다는 것을 의미한다. 본 논문에서 제안된 기술로 생성된 요약문은 KL에서 가장 낮은 발산이 측정되었기 때문에 KL 측정에서 가장 높은 요약문의 성능을 나타내고 있다. 그리고 KL(R)과 JSS에서는 각각 두 번째로 낮은 발산이 측정되었기 때문에, MEAD 다음으로 높은 성능을 나타냈지만, JS에서는 상당히 낮은 성능을 나타내었다. 이것은 KL과 JSS 발산 둘 다에서 낭비되는 평균 용어의 숫자가 가장 낮다는 것으로 원본 문서와 요약문서 사이의 유사도가 크다는 것을 의미한다. 반면에 JS에서는 비교적 낮은 성능의 결과를 보이는데, 이것은 낭비되는 용어의 수가 JS에서 상당히 많다는 것을 의미한다. 즉 MEAD나 NewBlaster 등 다른 요약기술과 비교해서 성능이 비교적 떨어진다고 볼 수 있다. 원본 문서와 요약된 문서의 유사도를 평가하기 위해서 본 논문에서는 코사인 유사도를 측정하였는데, 제안된 기술로 요약된 문서의 유사도가 가장 높은 결과를 보여주었다. 하지만 코사인 유사도는 원본 문서와 요약된 문서의 모든 용어와 비교하기 때문에, 이 방법을 이용한 결과가 반드시 중요하다고는 판단을 할 수 없다. 따라서 우리는 요약문에서 원본으로부터의 주제어만 포함하는 코사인 유사도를 추가적으로 측정하였다. 그 결과 역시 제시된 요약문이 다른 요약문과 비교할 때 가장 높은 분포를 가지고 있었다.

Table 1. Experiment result

	Base line1	Base line2	New Blaster	MEAD	**FuzzyS**
KL	2.22195	1.85853(2)	2.01218	1.87027	**1.81214(1)**
KL(R)	1.8105	1.51969	1.59936	**1.38714(1)**	1.43262(2)
JS	0.47793	0.43(2)	0.43693	**0.41262(1)**	0.46325
JSS	0.36072	0.31211	0.32786	**0.30364(1)**	0.30576(2)
Cosine-S	0.65752	0.631	0.75471(2)	0.7351	**0.77755(1)**
P-Topic	0.36441	0.45073	0.43845	0.54043(2)	**0.61111(1)**
T-Overlap	0.64263	0.62159	0.75159(2)	0.73456	**0.7796(1)**

4. 결 론

본 논문에서는 퍼지이론을 사용한 자동문서 요약 기술을 제안했다. 제안된 기술은 원본 문서로부터 주제어를 추출하기 위해서 LDA 모델을 사용해서 원본 문서의 각 문장에 대한 중요도를 계산해서 퍼지 시스템에서 규칙 기반 모델을 통해서 최종 문장의 중요도를 결정했다. 결정된 문장들은 요약문에 추가되었고 KL, KL(R), JS, JSS, Cosine-S, P-Topic, 그리고 T-Overlap 등 여러 측정방법을 사용해서 원본 문서와 함께 평가되었다. 실험 평가를 통해서 제안된 요약방법은 비교적 좋은 결과를 나타내었고, 특히 유사도 측정에서 가장 좋은 결과를 보여주었다. 하지만 원본 문서의 불필요한 용어의 처리 등의 문제를 해결하기 위해서는 향후 Type 2 퍼지 이론[19] 이 원본 문서와 요약문 사이의 유사도 사이의 간격을 줄이는 데 효과적일 것이라고 기대된다. 최근 들어 문서를 요약할 때 의사연관피드백 등을 이용한 문서요약 방법[20] 등이 제시되면서 의미론적 분야로도 확대되고, 또한 많은 문서 요약 방법들이 의학적 분야로도 확대되고 있음에 따라서[21][22] 의학 용어 사이의 애매모호한 용어들을 처리하는 데도 기여를 할 수 있을 것이라고 기대한다.

References

[1] R. Witte and S. Bergler, "Fuzzy coreference resolution for Summarization," In *Proceedings of International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS).* Venice, Italy: Universit Ca Foscari, pp.43-50, 2003.

[2] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization," *International Journal of Computer Science and Information Security (IJCSIS),* Vol.2, No.1, pp.65-70, 2009.

[3] G. Ravindra, N. Balakrishnan, and K.R. Ramakrishnan, "Automatic Evaluation of Extract Summaries Using Fuzzy F-score Measure," In *Proceedings of 5th International Conference on Knowledge Based Computer Systems,* pp. 487-497, 2004.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[5] C.Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In *Proceedings of Workshop on Text Summarization of ACL*, Spain, 2004.

[6] D. Gillick, "Sentence Boundary Detection and the Problem with the U.S," *The Association for Computational Linguistics*, pp.241-244, 2009.

[7] J. C. Reynar and A. Ratnaparkhi, "A Maximum Entropy Approach to Identifying Sentence Boundaries," In *Proceedings of 5th Conference on Applied Natural Language Processing*, pp.16-19, 1997.

[8] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, Vol.14, No.3, pp.130-137, 1980.

[9] D. Newman, Topic modeling tool, Available in: <http://code.google. com/p/topic-modeling-tool>.

[10] K. McKeown, R. Barzilay, J. Chen, D. K. Elson, D. K. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman, "Columbia's Newsblaster: New Features and Future Directions," *HLT-NAACL*, pp.15-16, 2003.

[11] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, Vol.24, pp.513-523, 1988. Reprinted in: Sparck Jones K. and Willet P. (eds.), *Readings in Information Retrieval, Morgan Kaufmann*, pp.323-328, 1997.

[12] I. Dhillon, S. Mallela, and R. Kumar, "Enhanced word clustering for hierarchical classification," In *Proceedings of 8th ACM Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.

[13] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Socit Vaudoise des Sciences Naturelles*, Vol.37, pp.547-579, 1901.

[14] L. A. Zadeh, "Fuzzy Sets," *Information and Control* 8, Vol. 8, No.3, pp.338-353, 1965.

[15] A. Louis and A. Nenkova, "Summary Evaluation without Human Models," *Text Analysis Conference*, 2008.

[16] D. R. Timothy, T. Allison, S. Blair-goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, A. Winkel, and Z. Zhang, "MEAD a platform for multidocument multilingual text summarization," In *Proceedings of International Conference on Language Resources and Evaluation*, pp.1-4, 2004.

[17] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol.22, No.1, pp.79-86, 1951.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, NY, 1991.

[19] L. A. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning," *Information Sciences*, Vol.8, pp.199-249, 1975.

[20] C. W. Kim and S. Park, "Document Summarization using Pseudo Relevance Feedback and Term Weighting," *Journal of Korea Institute of Information and Communication Engineering(JKIICE)*, Vol.16, No.3, pp.533-540, 2012.

[21] R. L. Summerscales, S. Argamon, S. Bai, J. Huperff, and A. Schwartzff, "Automatic Summarization of Results from Clinical Trials," *BIBM*, pp.372-377, 2011.

[22] S. Kiritchenko, B. Bruijn, S. Carini, J. Martin, and I. Sim, "Exact: automatic extraction of clinical trial characteristics from journal publications," *BMC Med Inform Decis Mak*, Vol.10, No.1, pp.56-17, 2010.

이 상 훈

e-mail : shlee8020@gmail.com
2004년 수원대학교 컴퓨터학과(학사)
2006년 수원대학교 컴퓨터학과(석사)
2013년 조지아주립대학교 대학원 컴퓨터학과(석사)
2013년~현 재 조지아주립대학교 대학원 컴퓨터학과 박사과정
관심분야 : 데이터 마이닝, 시맨틱 웹, 빅 데이터기반 텍스트 마이닝

문 승 진

e-mail : sjmoon103@hotmail.com
1986년 텍사스대학교 컴퓨터학과(학사)
1991년 플로리다주립대학교 대학원 컴퓨터학과(석사)
1997년 플로리다주립대학교 대학원 컴퓨터학과(박사)
1997년~현 재 수원대학교 컴퓨터학과 교수
관심분야 : 데이터베이스, 실시간 데이터베이스, 모바일 데이터베이스, 실시간 시스템, 임베디드 시스템