문서 구조 정보를 이용한 확률 모델 기반 자동요약 시스템

장 동 현, 맹 성 현
충남대학교 컴퓨터과학과

An Automatic Summarization System Based On a Probabilistic Model Using Document Structure Information

Dong-Hyun Jang, Sung Hyon Myaeng
Department of Computer Science, Chungnam National University

요 약

인터넷과 정보 서비스 기술의 발달로 일반 대중에게 제공되는 정보의 양은 기하급수적으로 증가하고 있는 추세지만 사용자가 원하는 정보를 얻기는 더욱 어려워지고 있으며, 필요한 정보를 찾은 경우에도 그 양이 많기 때문에 전체적인 내용을 파악하는 데 많은 시간을 소비하게 된다. 이러한 문제를 해결하고자 본 연구에서는 통계적 모델을 사용하여 문서로부터 문장을 추출한 후 요약문을 작성하여 사용자에게 제시하는 시스템을 개발하였다. 문서 요약 시스템의 구축을 위하여 사용된 방법은 문서 집합으로부터 중요 문장을 추출한 후 이로부터 요약문에 나타날 수 있는 특성(feature)과 중요 단어를 학습하여 학습된 내용을 이용하여 요약을 하는 방법이다. 시스템 개발 및 평가를 위해 사용된 문서는 정보 과학 분야의 논문 모음이며 이를 학습 데이터와 실험 데이터로 구분한 후 학습 데이터로부터 필요한 정보를 얻고 실험 데이터로 평가하였다.

1. 서 론

본 연구는 많은 양의 정보를 효과적으로 요약함으로써 사용자의 정보 습득 시간을 줄이고, 문서의 내용이 찾고자 하는 정보와 관련이 있는지 사용자로 하여금 신속히 판단할 수 있게 하는데 목적이 있다.

요약 시스템을 구축하기 위해서 본 연구에서는 문서 집합으로부터 추출한 어휘와 확률 정보를 사용하는 통계적 모델을 사용하였으며 실험을 통하여 성능을 평가하였다. 개발한 시스템은 수동으로 태깅(tagging)한 요약문으로부터 요약문 추출시 필요한 여러 가지 정보를 추출하는 학습과정과 이를 이용하여 각 문장이 요약문에 포함될 가능성을 계산하는 추출 과정으로 구성된다. 개발한 시스템은 텍스트의 구성 요소(component)를 판별하는 모델을 사용하여 요약문으로 포함될 가능성이 적은 문장을 제거하였으며, Dempster-Shafer 이론을 사용하여 여러 가지 특성(feature)에 대한 확률 값을 하나의 값으로 결합시키는 모델을 사용하였다.

본 논문의 2장에서는 과거에 수행된 관련연구를 살펴보고, 3장에서는 제안한 문서요약 시스템의 모델을 기술하며, 4장과 5장에서는 시스템의 학습과정과 요약문장 추출 과정 등 자세한 사항을 설명한다. 6장에서는 제안한 시스템의 성능에 대한 실험 및 평가를 하고, 이를 토대로 결론 및 향후 연구과제를 7장에서 기술한다.

2. 관련 연구

요약 시스템의 가장 일반적인 형태는 원문의 각 문장이 갖고 있는 언어적 혹은 구조적 정보를 이용하여 각 문장이 요약문에 포함될 가능성이 있는가를 판단하여 추출된 문장을 단순히 열거하거나 재정렬하는 문장 추출 기반 시스템(Passage Extraction System)이다. 이러한 종류의 시스템은 비교적 구현하기 쉽고 시스템이 간단하나 단순히 원문에 나오는

문장을 열거하기 때문에 요약이 부자연스럽고 원문에 나오는 문장에 의존하는 단점이 있다. 문장이 아닌 구(phrase)나 문단 자체를 추출하여 요약문을 생성하는 시스템도 이 부류에 속한다. 단어의 빈도수에 의거한 통계치를 사용하여 단어가 문서의 내용을 대표하는 정도를 계산하는 정보 검색 기법을 이용한 연구로 문단 단위의 검색 기법을 활용하여 문단간의 관계성을 계산한 후 관계성 패턴에 의해 추출될 문단을 결정하는 방법[1][2]이 제시되었으나 검색기법에 기반을 둔 접근 방법의 한계가 있음을 보였다. 이러한 연구를 통해서 단서 단어(clue words)나 위치 정보 등이 문서의 대표성을 지니는 문장을 추출하는데 더 중요한 역할을 할 수 있다는 점에 착안한 연구도 다수 존재 한다[3][4][5]. 또한 다량의 학습 데이터로부터 요약문에 포함되는 문장의 자질(feature)에 관한 확률 정보를 학습한 후 이를 이용하여 원문의 각 문장이 요약문에 포함될 확률을 계산하여 요약문을 추출하는 이론적인 기반과 실용성을 겸한 기법도 있다. 대표적인 연구로 미리 작성된 요약 학습 데이터로부터 특성을 추출한 후 Bayes 규칙을 이용하여 특성을 반영한 분류 함수를 통해 각 문장이 요약문에 포함될 확률을 계산한 연구[3]가 있다.

텍스트 이해 기반의 시스템(System Based on Text Understanding)은 인간이 문서를 요약하는 과정을 고도의 자연언어처리 과정을 통해 재현하려는 시도이다. 즉 요약 전문가가 문서의 내용을 파악하고 문서로부터 주제를 표현하고 있는 정보를 식별한 후 문장 생성(generation)을 통해 요약을 하는 과정[6]을 텍스트 파싱, 개념 표현, 문장 생성의 단계별로 처리하는 것이다. 이 부류의 시스템은 다양한 파싱 기술을 적용할 뿐만 아니라, 나름대로의 개념 표현 방법을 사용한다. 예를 들면 SCISORS 시스템[7]의 경우 KODIAK이라는 지식 표현 언어를 사용하여 개념 지식을 표현하고 SUMMONS[8] 시스템의 경우 정보추출 시스템에서 사용하는 틀(template)을 개념 표현 방법으로 사용한다. 텍스트 이해 기반의 시스템은 해당 분야의 지식과 문장의 문법적 구조를 기반으로 고품질의 자연스러운 요약문을 생성하나, 복잡한 자연어 처리 과정이 요구되고 적용분야마다 각각 다른 영역지식을 필요로 하기 때문에 응용분야가 한정된다는 단점도 있다. 영역 종속 지식의 사용에 따른 한계를 극복하기 위하여 단어간의 개념 관계를 나타낸 WordNet[9]과 같은 언어 지식과 자연언어처리 기법을 혼합하여 이용해서 문장이 나타내고 있는 주제를 파악하고자 하는 연구도 진행되고 있다[5].

앞서 설명한 바와 같이 문장 추출 시스템이나 이해 기반 시스템은 나름대로의 장단점을 갖고 있기 때문에, 각각의 장점은 장점대로 살리고 단점을 극복하고자 두 가지 형태의 시스템을 혼합한 형태의 시스템이 개발되고 있는 추세이다. Hovy[5]는 개념 추출을 위해서 통계적인 방법을 사용하였고 의미를 해석하기 위해서는 단어의 개념에 대한 지식을 사용하고 있는데, 이는 의미 해석을 위한 문장의 수를 줄임으로써 시스템의 성능을 향상시킬 수 있는 측면이 있으며, 일정한 수준의 요약문을 바라는 사용자의 요구와 현재 구현 가능한 기술 수준으로 볼 때 이러한 혼합된 형태의 시스템이 당분간은 주류를 이룰 것으로 보인다.

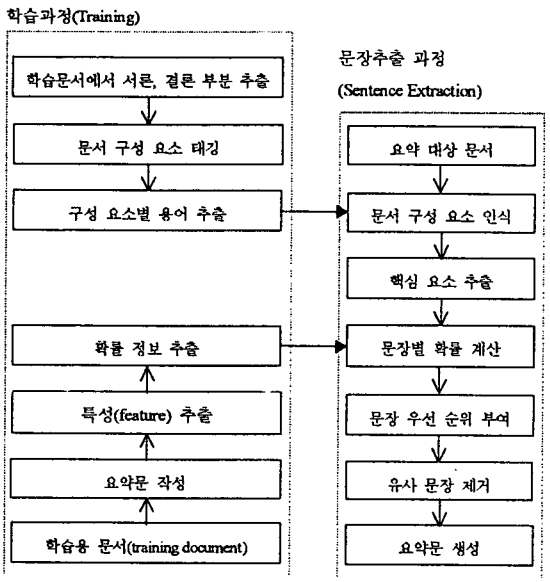
틀(template) 기반의 접근 방법[8][10]은 요약문에 포함되어야 할 개념을 수작업을 통해 틀로 정의하고 텍스트 분석을 통해 틀을 메꾼 후 요약문을 생성하는 과정을 거친다. 틀은 해당 분야의 전문가가 직접 작성하기도 하고, 학습 데이터로부터 요약문의 기반을 이루는 주제를 학습한 후 각 주제에 해당되는 틀을 만들기도 한다. 텍스트 데이터를 사용해서 틀을 채우는 방법으로는 데이터의 특성에 따라 다양한 방법을 적용할 수가 있는데, 예를 들어 각 주제별로 선택된 문장들로부터 단어의 빈도수를 계산하여 빈도수가 가장 높은 단어나 구 등을 틀의 내용으로 선택하는 방법이 적용될 수 있다. 본 접근 방법은 비교적 정확한 정보를 사용자에게 제공해 줄 수 있고 범죄 수사와 같은 특정 분야에 이용될 수 있지만, 분야마다 틀을 재정의해야 하는 문제점이 있다.

3. 요약 시스템 모델

본 논문에서 제안하는 기법은 학습 데이터로부터 3가지 특성(feature)을 추출하여 한 문장이 각 특성을 갖게 될 확률을 계산하여 순위를 부여한 후 요약 문장을 추출하게 된다. 본 시스템은 [그림 1]에서 보는 바와 같이 크게 2부분, 학습 코퍼스로부터 확률 정보를 추출하는 학습 과정과 실제 요약문을 추출하는 과정으로 나눌 수 있다. 하나의 문서로부터 요약에 사용될 문장을 추출하여 요약문을 생성하는 과정은 그림에서 보는 바와 같이 문장 구성 요소를 인식하여 핵심 요소만을 추출하는 과정과 추출된 각 문장별로 요약문에 포함될 확률을 계산한 후 우선순위를 결정하여 최종적으로 요약문을

생성하는 과정으로 대별할 수 있다. 학습 과정은 크게 두 가지로 볼 수 있다. 하나는 문장을 문서 구성요소별로 분류하는데 필요한 확률정보를 추출하는 과정이며, 다른 하나는 요약문에 포함되는 문장의 특성을 미리 학습하는 과정으로 각 특성별로 그 중요도를 확률 값으로 계산하는 과정이다.

학습 과정 중에서 구성 요소에 속한 용어를 추출하는 과정은 다음과 같다. 문장이 속하게 되는 구성 요소를 1)문서에 대한 배경 지식, 2)문서의 주요내용, 3)문서의 구조 설명, 4)향후 연구로 정의하고, 학습용 데이터의 각 문장을 구성 요소중의 하나로 수동 태깅(tagging)을 한 후, 구성 요소별로 단어를 추출하게 된다. 특성별 확률 정보 추출은 요약문장 추출과정에서 각 문장별로 요약문에 포함될 확률 계산을 위해서 필요한 학습이다. 확률 정보의 추출을 위해서 학습용 문서를 대상으로 수동으로 요약문을 작성한다. 작성된 요약문을 통해서, 1)요약문의 단서를 제공해주는 단어(cue word), 2)요약문에서 제외되는 단어(negative word), 3)문서 내에서 문장의 위치 등 3가지 특성(feature)에 대해서 확률정보를 추출한다. 추출된 각 특성에 대해서 요약문에 포함될 확률과 포함되지 않을 확률 정보를 추출하게 되면 학습 과정이 끝난다.



[그림 1] 시스템 구성도

문서의 구성 요소는 학습과정 중에서 구성 요소별로 추출한 단어를 이용해서 인식한다. 하나의 문장은 Bayes 규칙을 이용해서 각 구성 요소일 확률을 계산해서 가장 높은 확률 값을 갖는 구성 요소로 분류된다. 분류된 구성 요소 중에서 핵심 요소인 주요내용(CONT)만이 요약문 추출의 대상이 된다. 이러한 과정을 통해서 요약문에 포함될 확률 계산의 대상이 되는 문장을 줄일 수 있는 장점이 있다.

추출한 핵심 구성 요소 문장에 대해서 학습 과정 중에서 추출한 3가지 특성에 대한 확률 정보를 이용, 각 특성에 대해서 요약문에 포함될 확률을 계산하게 된다. 확률 계산 과정이 끝난 후 Dempster-Shafer 이론을 이용하여 특성별 확률 값을 결합하여 각 문장이 요약문에 포함 되어야 하는 신념(belief) 값을 계산한다.

요약문 추출의 대상이 논문이고 그 중에서도 서론과 결론이므로 중복되는 문장이 있을 수 있다. 따라서 유사한 문장이 있을 경우 문장의 우선 순위가 낮은 문장은 제외시킨다. 문장 사이의 유사도는 정보검색에서 문서와 질의간의 유사도를 측정하는 방법을 사용하여 각 문장을 구성하고 있는 단어를 비교하여 판단한다.

지금까지 제안한 요약 시스템에 대해서 개략적으로 살펴보았고 각 단계에 대한 자세한 내용은 4장과 5장에서 설명하기로 한다.

4. 학습 과정(Training Process)

제안한 모델의 학습 과정(training process)은 [그림 1]에서 보듯이 2가지를 학습하게 된다. 하나는 문서의 구성 요소를 4가지(배경지식, 문서의 주요내용, 문서의 구조, 향후과제)로 정한 후 각 구성 요소를 나타내는 단어를 학습하는 것이며, 다른 하나는 학습 데이터로부터 얻어낸 요약문의 특성별로 확률 정보를 계산하는 것이다.

4.1 구성 요소 학습 과정

4.1.1 학습 데이터 (training data)

요약 시스템의 학습용으로 사용한 데이터는 "한글 및 한국어 정보처리 학술대회"에서 발표되었던 50편의 논문을 사용하였다. 논문의 작성자가 전산학, 언어학, 심리학 등 전공분야가 다양하기 때문에 논문 양식(style) 또한 다양한 것이 학습 데이터의 특성이다.

일반적으로 논문의 경우 서론과 결론에 그 논문의 중요한 내용이 있으므로, 본 연구에서도 이를 기본 가정으로 하고 50편의 논문 중 서론과 결론 부분만을 학습 데이터로 사용하였다.

4.1.2 문서 구성 요소별 태깅(tagging)

학습 데이터와 같이 문서가 논문인 경우 글의 전개 방식이 정해져 있다. 즉, 서론 부분에서는 논문에서 다루는 내용의 배경 지식이 서술되고 이어서 연구하고자 하는 내용 그리고 논문의 구성 순으로 글이 전개되는 것이 일반적이다. 본 논문에서는 이를 문서의 구성요소(component)라고 정의하고, 문서를 요약하는 데 있어서 이러한 특성을 이용하고자 한다. 요약을 목적으로 문서의 구성 요소를 4가지(배경지식, 주요내용, 문서구조 정의, 향후연구)로 정하고 각 문장별로 속하는 구성요소를 수동 태깅을 하는데, 이는 각 구성 요소를 나타내는 단어 추출이 목적이다. 문서의 구성 요소를 결정하지 않고 문서 전체를 요약문의 대상으로 할 수도 있지만, 요약문 추출 대상을 줄임으로써 확률 계산 시간을 줄일 수 있고 요약문에서 제외 되어야 하는 문장들을 미리 제거하는 효과도 얻을 수 있다.

4.1.3 구성 요소별 용어(term) 추출

구성 요소별로 태깅된 문장으로부터 소속된 구성요소를 나타내는 단어를 추출하는 단계로, 각 단어가 주어진 구성 요소에 속하게 될 확률 정보를 사용함으로써 새로운 문장이 어떤 구성요소에 속하게 되는지 판별할 수 있게 된다. 단어 추출 방법은 형태소 분석기를 이용하는 것이 가장 일반적인 방법이나 분석기 자체가 복잡하고 불필요한 분석 결과를 제공하는 단점이 있어 정보검색의 색인 목적으로 사용될 수 있는 특수목적 프로그램을 사용하였다. 이 방법에서는 어절이 주어졌을 경우 단순명사를 반복적으로 분석하는 방법[11][12]을 사용하도록 하는데, 한국어의 경우 복합 명사 빈도수가 많고 중요도 또한 높으므로[13] 복합명사의 처리를 하는데 매우 효과적이다.

4.2 확률 정보 추출 과정

4.2.1 요약에 사용되는 문장 분리

학습 데이터로부터 각 문서별로 요약문을 수동으로 추출하게 되는데, 요약문 추출에 대한 확실한 기준이 없고 가능하면 주관적인 생각을 배제하기 위해서 논문 작성자의 요약 부분(abstract)을 기준으로 요약 문장을 추출하였다. 추출된 요약문으로부터 요약문이 갖게 되는 위치 정보, 단서를 제공하는 단어(cue word)에 대한 확률 정보를 계산하며, 요약문에서 제외된 문장을 기반으로는 제외의 요소가 된 단어(negative word)에 대한 확률정보를 추출하게 된다.

4.2.2 특성(feature) 추출

앞서 구축한 문장 집합으로부터 요약문장이 갖는 특성을 추출하는 과정으로 아래와 같이 3가지의 특성을 사용하였다.

• 단서 단어 특성 (cue word feature) : 요약문장이 갖는 단서를 제공해주는 구를 정의한다. 예를 들면, "본 논문의 목적은 ~", "~ 처리 과정을 제안~" 등과 같은 구가 이에 속한다.

• 위치 특성 (position feature) : 요약문으로 추출된 문장의 문서 내에서의 위치에 관한 특성으로 학습 코퍼스를 기준으로 볼 때 일반적으로 서론의 경우 뒷부분에, 결론의 경우 앞부분에 중요한 문장이 위치하는 것을 알 수 있는데, 이 특성을 실제 요약문 추출 시 사용한다.

• 부정어 특성 (negative word feature) : 요약문에 속할 확률을 적게 만드는 어휘적 특성으로 "예를 들면", "생각한다"와 같은 구가 이에 속한다. 이는 학습을 위해 구축한 요약문에서 제외된 문장들을 대상으로 자주 나오는 구의 패턴을 추출했다.

4.2.3 확률 정보의 추출

실제 요약문 추출 시 문장이 요약문에 속하게 될 확률을 계산하기 위해서 필요한 확률 정보를 얻어내는 단계로 Bayes 규칙을 사용하여 각 특성별로 확률정보를 계산한다.

• 단서 단어 확률 (cue word probability) : 학습 데이터(training corpus)에서 단서 단어(*CWi*)가 요약문에 속하게 될 확률은 (식 1)로부터 구할 수 있다.

$$P(s \in S \mid CW_i) = \frac{P(CW_i \mid s \in S)P(s \in S)}{P(CW_i)}$$

(식 1)

위의 식에서 $P(CWi)$는 전체 학습 데이터에서 단서 단어 i가 있을 확률, $P(CWi \mid s \in S)$는 요약문장에 단서 단어 i가 있을 확률을 나타낸다.

• 위치 확률 (position probability) : 각 문장의 위치에 따라서

요약문에 포함될 확률 값을 계산하게 되는데 본 시스템에서는 크게 6개로 위치를 나누어 각각의 확률 값을 얻어낸다. 처음 5문장이내, 끝에서 5문장 이내, 그 외 중간에 있을 경우 각각에 대해서 요약문에 포함될 확률 값을 계산하게 되는데, 서론과 결론은 내용을 전개하는 특성이 다르므로 분리하여 계산하였다. 예를 들어 서론에서 처음 5번째 이내의 문장이 요약문에 포함될 확률은 (식 2)로부터 구할 수 있다.

$$P(s \in S|P1) = \frac{P(P1|s \in S)P(s \in S)}{P(P1)}$$

(식 2)

• 부정어 확률 (negative word probability) : 부정어(*NWi*)가 요약문에 포함될 확률은 단서 단어 확률을 계산하는 방법과 동일하다.

지금까지 한 문장이 각 특성(cue-word feature, position feature, negative-word feature)을 갖고 있을 때 요약문이 될 확률 정보가 어떻게 추출되는지를 보였다. 이와 유사한 방법으로 문장이 각 특성을 갖고 있을 때 요약문에 포함되지 않을 확률 정보도 추출한다. (식 3)은 부정어(*negative word i*)가 주어졌을 경우 요약문에 포함되지 않을 확률을 계산하는 식을 보이고 있으며, 마찬가지로 다른 특성에 대해서도 이러한 식을 사용한다.

$$P(s \notin S|NW_i) = \frac{P(NW_i|s \notin S)P(s \notin S)}{P(NW_i)}$$

(식 3)

각 문장에 대해 추출된 두 가지 정보는 요약문장 추출의 최종 확률 계산 시 이용하게 되며, 이에 대한 자세한 내용은 5장에서 알아보기로 한다.

5. 요약문장 추출 과정

전처리 과정에서 학습한 정보를 통해서 요약문을 작성하는 과정으로 본 장에서는 요약문 추출의 각 단계를 설명한다.

5.1 문서 구성요소(component) 구분

본 과정은 문서 요약의 목적상 불필요한 문장을 제거하는 단계로 이후의 단계에서 처리해야 하는 문장의 수를 줄일 수 있는 장점이 있고, 구분이 되지 않았을 경우 발생할 수 있는 잡음(noise) 제거 효과도 있다. 학습과정에서 문서의 구성 요소를 4가지로 정의한 후 각 구성요소별로 추출한 용어를 이용하여 각 문장이 어느 구성 요소에 속하는가를 확률적으로 계산하는 단계이다. 예를 들어, 어느 한 문장이 문서의 구성요소 중에서 주요내용(CONT)이 될 확률은 (식 4)로부터 구할 수 있다. 여기서 문장 *s*는 *n*개의 단어(*t1, t2, ... , tn*)로 구성되며 각 단어는 서로 독립적으로 출현한다고 가정한다.

$$P(CONT|s) = \frac{P(s|CONT)P(CONT)}{P(s)}$$

$$\approx \frac{\prod_{i=1}^{n} P(t_i|CONT)P(CONT)}{\prod_{i=1}^{n} P(t_i)}$$

(식 4)

이와 동일한 방법으로 각 문장이 다른 구성 요소인 배경지식, 문서구조 정의, 향후연구에 속할 확률을 구한 후 최고의 확률을 갖는 구성요소로 각 문장은 분류가 된다. 이렇게 문장별로 각각의 문서 구성요소에 대한 확률을 구함으로써 각 문장이 문서 내에서 어느 구성요소에 속하게 되는지 자동적으로 결정한다. 일반적으로 요약문에 포함될 문장은 구성 요소 중에서 주요 내용이라고 가정하고 이에 속하는 문장만을 요약 문장의 후보로 사용한다.

5.2 특성별 확률 계산

문서의 구성 요소 중 주요 내용에 해당하는 문장을 대상으로 요약문이 될 확률을 계산하는 단계로 학습과정에서 생성한 확률 정보를 이용하여 1)단서를 제공하는 단어의 확률, 2) 위치 확률, 3)부정어에 대한 확률을 계산한다.

• 단서 단어 확률 (cue word probability) : 문장에 한 개 이상의 단서 단어(cue word)가 나올 때 그 문장이 요약문에 포함될 확률 계산은 (식 5)를 이용한다.

$$P(s \in S|cw_1,\ldots,cw_n) = \frac{P(cw_1,\ldots,cw_n|s \in S)P(s \in S)}{P(cw_1,\ldots,cw_n)}$$

(식 5)

• 위치 확률 (position probability) : (식 6)을 이용하여 문장의 위치가 *i(i=1,...,6)*에 있을 경우 요약문에 포함될 확률을 계산하게 되며, 문장은 [표 1]의 위치별 정의 중에서 하나에만

$$P(s \in S|P_i) = \frac{P(P_i|s \in S)P(s \in S)}{P(P_i)}$$

(식 6)

속하게 되므로 위치에 따라서 하나의 확률 값만 갖게 된다.

i	장 (section)	정 의
1	서론	처음 5문장 이내의 위치
2	~	처음과 끝 5문장 이외의 위치
3	~	끝 5문장 이내
4	결론	처음 5문장 이내의 위치
5	~	처음과 끝 5문장 이외의 위치
6	~	끝 5문장 이내

[표 1] 위치별 정의

- 부정어 확률 (negative word probability) : 부정어 확률은 문장에 한 개 이상의 부정어(negative word)가 나올 때 그 문장이 요약문에 포함될 확률 값을 나타내는데 단서 단어 확률을 계산하는 방식과 동일하다.

5.3 증거(Evidence) 수집 및 우선 순위 결정

5.2에서 각 특성별로 문장이 요약문에 포함될 확률, 요약문에 포함되지 않을 확률을 구하는 과정을 보였다. 문장의 우선 순위를 부여하기 위해서는 각 특성별로 구한 확률 값을 하나의 대표 값으로 나타내야만 한다. 각 특성별 확률 값은 최소 0, 최대 1이지만 특성별로 값의 차이가 많이 나므로, 각 특성 값을 요약문에 포함될 증거(evidence)로 취급함으로써 문장이 요약문에 포함될 가능성을 판가름하는 신념도(degree of belief)를 나타내는 Dempster-Shafer 이론[14]을 이용한다. 이 이론은 어떤 사실에 대한 신념도뿐 아니라 정보의 양을 나타낼 때 사용하게 되며 여러 가지 가정에 대한 증거를 하나의 값으로 축적할 경우 이용한다. 해결하고자 하는 요약문 추출의 경우도 하나의 특성에 대한 확률 값만 있을 경우보다 여러 개의 특성에 대한 확률 값이 존재하면 그만큼 문장의 요약문 포함 여부가 확실해지게 되는 것이다. 예를 들어 한 문장이 요약문에 포함될 각 특성별 확률이 다음과 같을 때,

$P(s \in S|cw_1,\ldots,cw_n)=0.4,\quad P(s \in S|p_f)=0.6,\quad P(s \in S|nw_1,\ldots,nw_n)=0.3$

Dempster-Shafer 이론을 이용하여 3개의 값을 하나로 결합시키면 최종 신념도(belief)는 0.832가 된다. 같은 방법으로 부정적인 신념도(negative belief)도 구한 후 이를 이용해서 문장이 요약문에 포함될 확률 값을 구하게 된다. 예를 들어 한 문장에 대한 긍정적인 신뢰도가 0.5이고 부정적인 값이 0.25인 경우 (0.5 - 0.25)의 결과인 0.25가 문장이 요약문이 될 최종 값이 된다. 이렇게 각 문장별로 구한 값의 크기에 따라서 요약문에 포함될 우선순위가 결정된다.

5.4 유사 문장 제거를 통한 최종 요약문 생성

논문의 경우 서론과 결론 부분에 중복되는 문장이 있는 경우가 있다. 따라서 요약문장 추출 시 이러한 문장 중에서 하나는 제거할 필요가 있다. 이를 위해서 문장 사이의 유사도를 계산하여 임계값이 0.5 이상인 경우 요약문이 될 확률이 작은 문장은 제거한다. 실험 결과 문장의 유사도가 0.5이상인 경우 거의 비슷한 문장으로 결과가 나왔다.

시스템에서 추출한 요약문장이 사용자에게 자연스럽게 제공되기 위해서는 추출한 문장의 순서를 자연스럽게 재구성할 필요가 있다. 그러나 본 연구에서는 시스템 평가의 용이성을 위해 이에 대한 연구는 차후에 미루기로 한다. 실제로 본 요약 시스템이 사용자에게 요약문을 제공하는 방법은 위에서 구한 확률 값의 크기에 따라 우선 순위를 정하여 문장을 제공한다. 이렇게 함으로써 실제로 사용자가 원하는 수만큼 요약문을 제공할 수 있을 뿐 아니라, 본 시스템을 평가하는 과정에서도 생성된 요약문의 길이를 조정하여 실제 평가자(human subject)가 선택하는 요약문의 개수와 일치시킴으로써 올바른 평가를 할 수 있다.

6. 실험 및 평가

개발한 시스템의 성능을 실험하기 위해 사용한 데이터는 학습 과정에서 사용했던 데이터와 동일 분야의 데이터로 "한글 및 한국어 정보처리학회" 논문 30건을 사용했다. 이는 학습 데이터와는 다른 데이터이며, 학습할 때와 마찬가지로 서론과 결론만을 추출하여 요약의 대상으로 이용했다.

평가를 위해서 대학원생 3명과 학부생 3명으로 하여금 실험 데이터로 사용된 문서에서 주요 내용(CONT component)에 해당하는 문장을 추출하도록 한 후, 추출한 문장에 대해서 우선 순위에 따라 요약문을 추출하도록 했다. 시스템의 평가는 두 가지 측면을 고려했다. 하나는 주요 내용 추출에 대한 성능 평가이며, 다른 하나는 요약문 추출에 대한 평가이다.

본 연구에서는 평가자가 추출한 요약문장과 시스템이 추출한 요약문장을 비교해서 정확도와 재현도로 시스템을 평가하였다. 평가자의 공정성을 기하기 위해서 하나의 문서에 대해서 두 명으로 하여금 요약을 하도록 했지만, 평가자가 추출한 요약문장에서도 다소 차이를 보이므로, 정확도와 재현도를 독

립적으로 계산한 후 평균값을 취했다.

평가 대상은 크게 2가지이다. 하나는 문장의 구성 요소(component) 인식에 대한 평가이며, 다른 하나는 요약문장 추출에 대한 평가이다.

- 구성 요소(component) 판별에 대한 평가

문서 구성 요소 인식에 대한 평가는 평가자의 구성요소 판별과 시스템이 추출한 구성 요소가 얼마만큼 재현되었는지에 대한 인식률로 한다. 이 부분에 대한 평가를 정확하게 하기 위해서는 구성 요소 각각에 대한 평가를 해야 하겠지만 본 시스템의 주목적이 구성 요소 인식에 있지 않기 때문에 구성 요소중에서 요약 시스템에 영향을 주로 미치는 "주요내용"에 대한 평가만을 하였다. 사용자가 주요내용으로 추출한 문장의 수는 295개였으며, 그 중에서 시스템이 추출한 문장과 동일한 문장의 수는 총 201개로 약 67%의 재현율을 나타냈다. 구성 요소의 오류 판별 문장은 요약문 추출 과정에서 제거될 수 있으므로 정확도에 대한 평가는 하지 않았다.

- 요약문장 추출에 대한 평가

요약문장 추출에 대한 평가 방법은 정보검색 시스템 평가 시 주로 사용하는 정확도(precision)와 재현도(recall)[15]를 이용하였으며, 본 연구에서 제안한 구성 요소 판별에 대한 효용성을 평가하기 위해서 구성 요소 판별을 한 경우와 판별하지 않은 경우 각각에 대해서 평가를 하였다.. 정확도는 요약 시스템이 추출한 문장이 요약문장으로서 적합한지 판단하게 해주는 값이며, 재현도는 전체 문서 내에서 요약문에 포함되어야 할 문장이 얼마만큼 요약 시스템이 추출하였는지에 대한 척도이다.

- 결과 분석

[표2]와 [표 3]은 각각 구성 요소를 판별을 하지 않은 경우와 판별한 경우에 대한 결과를 보여주고 있으며 결과로부터 구성 요소 판별을 한 경우의 결과가 우수함을 알 수 있다. 개발한 시스템의 평균 정확도([표 3])는 약 44%, 재현도는 약 34%의 성능을 보이고 있는데, 평가 방법이나 적용 문서의 분야뿐 아니라 언어가 다르기 때문에 다른 관련 연구와의 직접적인 비교는 큰 의미가 없으나, 본 연구와 가장 유사한 Kupiec[3]의 연구에 대한 평가 결과는 평균 43%의 정확도를 보였다.

오류의 형태를 살펴보면 두 개의 문장 중에서 한 문장이 다른 문장의 내용에 거의 완전히 포함되지만, 포함시키는 문장 내에 다른 내용의 단어가 많이 포함되어있을 경우 두 문장의 유사도가 작게 나오는 경우와 구성 요소의 오분석으로 인해 잘못된 요약문을 추출하는 경향이 있다.

# of sentence	Precision (%)	Recall (%)
1	50.00	10.00
2	36.67	14.67
3	36.67	22.00
4	35.83	28.67
5	36.00	36.00

[표 2] 구성 요소를 판별하지 않은 경우의 평가 결과

# of sentence	Precision (%)	Recall (%)
1	53.33	14.55
2	43.34	21.37
3	43.33	35.46
4	41.59	45.00
5	39.53	53.19

[표 3] 구성 요소를 판별한 경우의 평가 결과

7. 결론 및 향후과제

본 논문에서는 사용자가 많은 양의 정보를 손쉽게 파악할 수 있는 방안으로 사용될 수 있는 문서 요약 기법을 제시하고 구현한 후 그 성능을 평가하였다. 제안된 시스템에 사용된 요약 모델은 확률 이론에 근거한 것으로 학습 데이터로부터 확률 및 어휘 정보를 추출한 후 이를 이용하여 문서를 요약하는 시스템을 개발하였다. 개발된 시스템은 한국어 문서를 대상으로 문장을 구성 요소별로 분류함으로써 요약문 추출 과정에서 처리해야 하는 문장의 수를 줄일 수 있었다. 또한 정보 검색 기법을 이용하여 요약문의 단서를 제공해주는 단어(cue word)와 부정적인 단어(negative word) 추출을 자동화하였다. 실험을 통한 성능 평가를 하였으며 또한 본 연구에서 제안한 구성 요소를 판별한 후 요약을 하는 기법이 구성 요소를 판별하지 않고 요약하는 것보다 우수한 결과를 얻을 수 있음을 보였다.

본 연구에서 제안한 한국어 텍스트 요약을 위한 시스템은 고유의 모델을 수립하고 유용한 결과를 낸 가치가 있지만 앞으로 수정, 보완해야 할 점을 정리해 보면 다음과 같다. 첫째, 요약문장 추출 시 사용되는 특성(feature)을 다양화할 필요가

있다. 둘째, 문장의 구성 요소를 인식하는 부분에 대한 보다 체계적인 연구가 필요하다. 셋째, 현재는 문장이 요약문에 포함될 확률 값의 순서에 따라 제시해주고 있으나 보다 자연스러운 요약문의 생성을 위해 요약 문장의 재구성이 필요하다. 넷째, 지시어나 대명사에 대한 처리가 필요하다. 다섯째, 다양한 분야에 적용할 수 있는 방법이 연구되어야 한다.

참 고 문 헌

[1] Jose Abracos and Gabriel Pereira Lopes, "*Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles*", Proc. of a Workshop in Intelligent Scalable ex Summarization, pp.51~57, July, 1997.

[2] M. Mira, A. Singha and C. Buckley, "*Automatic Text Summarization by Paragraph Extraction*", Proc. of a Workshop in Intelligent Scalable Text Summarization, pp.39~46, July, 1997.]

[3] Julian Kupiec, Jan Pedersen, Francine Chen, *"A Trainable Document Summarizer"*, Proc. of 18th ACM-SIGIR Conference, pp.68-73, 1995.

[4] Dong-Hyun Jang and Sung Hyon Myaeng, "*Development of a Document Summarization System for Effective Information Services*", Proc. of RIAO '97 Conference, pp. 101~111, 1997.

[5] E. Hovy and C. Y. Lin, "*Automated Text Summarization in SUMMARIST*", Proc. of a Workshop on Intelligent Scalable Text Summarization, pp. 18~24, July, 1997.

[6] B. Endres-Niggemeyer, E. Maier and A. Sigel, "*How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor*", Information Processing & Management, Vol. 31, No. 5, pp. 631~674, 1995.

[7] P. S. Jacobs and L. F. Rau, "*Natural Language Techniques for Intelligent Information Retrieval*", Proc. of 11th ACM-SIGIR Conference, pp.85~99, 1988.

[8] K. McKeown and D. Radev, "*Generating Summaries of Multiple News Articles*", Proc. of 18th ACM-SIGIR Conference, pp.74~82, 1995.

[9] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "*Introduction to WordNet: An On-Line Lexical Database*", International Journal of Lexicography, Vol. 3, No. 4, pp.235~312.

[10] Chris D. Paice & Paul A. Jones, "*The Identification of Important Concepts in Highly Structured Technical Papers*", Proc. of 16th ACM-SIGIR Conference, pp.69~78, 1993.

[11] S.H.Myaeng & D.H.Jang, *"On Language Dependency in Indexing"*, Proc. of the Workshop on Information Retrieval with Oriental Languages, pp.17-23, 1996.

[12] 장동현, 맹성현, *"효율적인 색인어 추출을 위한 복합명사 분석 방법"*, 제8회 한글 및 한국어 정보처리 학술대회, pp.32-35, 1996.

[13] 윤보현, 임희석, 임해창, *"통계 정보를 이용한 한국어 복합명사의 분석 방법"*, 한국정보과학회 봄 학술발표논문집, Vol. 22, No. 1, pp.925-pp.928, 1995.

[14] Elaine Rich & Kevin Knight, Artificial Intelligence, 2nd edition, McGraw-Hill, 1991.

[15] Gerard Salton, Automatic Text Processing, Addison-Wesley Publishing Company, 1989.