텍스트 구성요소 판별 기법과 자질을 이용한 문서 요약 시스템의 개발 및 평가

(Development and Evaluation of a Document Summarization System using Features and a Text Component Identification Method)

장 동 현 † 맹 성 현 ††

(Dong Hyun Jang) (Sung Hyon Myaeng)

요 약 논 본문은 문서의 주요 내용을 나타내는 문장을 추출함으로써 요약문을 작성하는 자동 요약 기법에 대해 기술하고 있다. 개발한 시스템은 문서 집합으로부터 추출한 어휘적, 통계적 정보를 고려하여 요약 문장을 작성하는 모델이다. 시스템은 크게 두 부분, 학습과정과 요약과정으로 구성이 된다. 학습 과정은 수동으로 작성한 요약문장으로부터 다양한 통계적인 정보를 추출하는 단계이며, 요약 과정은 학습 과정에서 추출한 정보를 이용하여 각 문장이 요약문장에 포함될 가능성을 계산하는 과정이다. 본 연구는 크게 세 가지 의의를 갖는다. 첫째, 개발된 시스템은 각 문장을 텍스트 구성 요소의 하나로 분류하는 텍스트 구성 요소 판별 모델을 사용한다. 이 과정을 통해 요약 문장에 포함될 가능성이 없는 문장을 미리 제거하는 효과를 얻게 된다. 둘째, 개발한 시스템이 영어 기반의 시스템을 발전시킨 것이지만, 각각의 자질을 독립적으로 요약에 적용시켰으며, Dempster-Shafer 규칙을 사용해서 다양한 자질의 확률 값을 혼합함으로써 문장이 요약문에 포함될 최종 확률을 계산하게 된다. 셋째, 기존의 시스템에서 사용하지 않은 새로운 자질(feature)을 사용하였으며, 실험을 통하여 각각의 자질이 요약 시스템의 성능에 미치는 효과를 알아보았다.

Abstract This paper describes an automatic summarization approach that constructs a summary by extracting sentences that are likely to represent the main theme of a document. As a way of selecting summary sentences, the system uses a model that takes into account lexical and statistical information obtained from a document corpus. As such, the system consists of two parts: the training part and the summarization part. The former processes sentences that have been manually tagged for summary sentences and extracts necessary statistical information of various kinds, and the latter uses the information to calculate the likelihood that a given sentence is to be included in the summary. There are at least three unique aspects of this research. First of all, the system uses a text component identification model to categorize sentences into one of the text components. This allows us to eliminate parts of text that are not likely to contain summary sentences. Second, although our statistically-based model stems from an existing one developed for English texts, it applies the framework to individual features separately and computes the final score for each sentence by combining the pieces of evidence using the Dempster-Shafer combination rule. Third, not only were new features introduced but also all the features were tested for their effectiveness in the summarization framework.

1. 서 론

인터넷과 다양한 정보 서비스의 폭발적 증가로 인해 원하는 정보를 검색하는 작업이 점점 심각한 문제로 대두되고 있는 것이 현재 정보 시대의 현실이다. 이러한 문제를 해결하고자 다양한 정보 검색 시스템이 개발되어 사용자에게 필요한 정보를 검색하는 일을 대행하고

† 학생회원 : 충남대학교 컴퓨터과학과
dhjang@cs.cnu.ac.kr
†† 종신회원 : 충남대학교 컴퓨터과학과 교수
shmyaeng@cs.cnu.ac.kr
논문접수 : 1999년 3월 5일
심사완료 : 2000년 3월 4일

있다. 일반적으로 검색 결과는 사용자가 원하는 정보와 차이가 있게 마련인데, 제공된 정보의 유용성과는 별개로 사용자는 문서의 내용을 파악하기 위해 많은 시간을 소비하게 된다. 이러한 과다 정보의 문제로 인해서, 전문(full-text)을 읽지 않고도 문서의 내용을 쉽게 파악할 수 있는 시스템의 필요성이 대두되었는데 대표적인 것이 문서 자동 요약 시스템이다.

최근에 다양한 접근 방법의 요약 시스템이 개발되고 있는 상황에서 본 연구에서도 한국어 문서를 대상으로 하는 요약 시스템을 개발했으며, 전문에서 가장 의미 있고 대표적인 문장 추출을 주된 기능으로 하고 있다. 개발된 시스템은 통계 기반의 모델로서 학습 코퍼스로부터 추출한 어휘적, 통계적 정보를 이용하여 각 문장이 요약문에 포함될 확률을 결정하게 된다. 본 요약 모델이 갖는 핵심적인 특징은 다양한 자질과 더불어 문서의 의미적 구조(thematic structure)를 사용하는 것이다. 의미적 구조는 요약 과정과 동일한 확률 체계를 통해서 인식하게 되는데, 의미적 특성에 따라서 문서 내의 문장 또는 부분을 범주별로 분류하는데 사용한다.

본 연구에서 시도한 다른 특징으로는 이 통계기반 모델에 다양한 자질을 포함시키기 위해서 다각도의 실험을 통해 시스템을 개발하였다. 즉, 요약문 생성에 있어서 모든 자질을 사용하는 대신에 특정 자질을 전혀 갖지 않는 문장을 요약 문장 후보에서 사전에 제거하기 위해 자질을 필터(filter)로 사용하는 접근 방법을 시도하였다. 또한 필터를 적용한 기법 외에 각각의 자질이 요약문 생성에 미치는 영향을 살펴보기 위해서 실제 요약 과정에서 다양한 실험을 시도하였다.

본 연구에서 제시한 모델은 확률을 기반으로 하고 있으며, 영어에 대해서 만족할 만한 성능을 보여준 Kupiec[1]의 접근 방법과 유사하다. 그러나 모델링의 관점에서 볼 때, 모든 자질을 하나의 집합으로 간주하기보다는 각각의 자질을 통계적 분류자로 적용시킨 점과 Dempster-Shafer 규칙[2]을 사용하여 각 자질의 증거를 결합시킨 점에서 Kupiec의 방법과는 다르며, 문서의 의미적 구조를 확률 기반 모델에 적용한 점이 새로운 시도라 할 수 있다. 그리고, Kupiec의 모델이 모든 언어에 대해서 적용 가능한 것으로 보고 기본 모델로 채택하여 한국어에 적용시키고 있기 때문에, 모델의 검증 차원에서도 본 연구의 가치가 있다.

본 논문의 구성은 2장에서 관련 연구에 대해서 기술하였으며, 3장에서는 본 연구에서 개발한 시스템에 대해서 크게 학습 과정과 요약 과정으로 나누어 개괄적으로 설명하였고 4장과 5장에서는 개념적 정의를 통해서 각 과정을 단계별로 자세히 기술하였다. 자질이 요약문 생성에 미치는 영향을 알아보기 위해서 시도한 다양한 실험 결과는 6장에서 설명하고 있으며, 결론 및 향후 연구에 대한 내용은 7장에서 논의한다.

2. 관련 연구

요약 시스템은 접근 방법에 따라서 크게 두 가지 형태로 분류할 수 있다. 하나는 문장 추출(sentence extraction) 기반의 시스템이고, 다른 하나는 문장 또는 문서 이해(sentence or document understanding)를 기반으로 하는 시스템이다. 문장 추출 시스템은 일반적으로 각 문장이 요약 문장에 포함될 수 있는 가능성을 계산하여 문서의 내용을 대표하는 문장을 원시 문서로부터 추출하는 것이 핵심인데, 일반적으로 문장에 점수를 부여하여 고득점을 우선으로 하거나 순서를 재구성하기도 한다. 이러한 종류의 시스템은 구현이 용이하지만, 추출된 문장이 일관성 있는 구성을 이룬다는 보장이 없기 때문에 부자연스러운 요약문을 생성할 수 있다.

이러한 범주의 기존 연구를 살펴보면 요약 문장의 가능성을 나타내는 자질을 사용하는 경우가 대부분이다. 예를 들어, 요약 문장에 나타나는 단서 단어(cue words)와 문서 내에서 문장의 위치 등이 대표적인 자질로 사용될 수 있는데, Kupiec[1]은 이러한 자질을 사용하여 통계 기반의 학습 가능한 요약 시스템 모델을 제시하였다. 그는 학습 문서 집합으로부터 추출한 요약 문장에 대한 정보를 이용하여 임의의 문장이 요약 문장으로 추출될 확률을 예상하는 분류 함수(classification function)를 제안하였다. 요약은 확률 값에 의해 문장에 순위를 부여함으로써 생성이 되며, 사용자가 지정한 개수만큼의 상위 문장을 선택할 수도 있다. 본 연구에서도 이러한 접근 방법을 사용하였지만 모델 자체를 발전시켰고 새로운 자질을 추가하였다. 또한 텍스트 구조 개념을 포함시킨 점과 서로 다른 자질로부터 계산한 증거들을 하나의 값으로 결합시키기 위한 기법을 새롭게 적용한 점에서 기존의 연구와는 다른 특성을 갖고 있다.

문장 추출 기반 자동 요약의 또 다른 접근 방법으로 정보 검색 기법을 그대로 적용한 연구가 이에 속한다. 이 방법은 단어의 빈도수를 사용함으로써 단어가 문서를 대표하는 정도를 계산하기도 하며, 단락 단위의 검색 기법(passage retrieval)을 적용하여 단락 사이의 연관성을 판단한 후 이를 통해서 습득한 관계(relationship)를 기반으로 하여 요약 단락을 추출하는 기법도 있다[3][4].

텍스트 이해를 기반으로 하는 연구는 문장 추출 시스

템보다 자연스러운 요약 문장을 생성한다. 그러나 이러한 접근 방법은 이론적인 구현이 어려울 뿐만 아니라 현재의 자연언어 처리 기술로는 신뢰할 만한 결과를 얻을 수 없는 한계성이 있다. 텍스트 이해 기반의 요약을 하기 위해서는 의미를 표현할 수 있는 방법과 그 표현으로부터 요약문을 생성할 수 있는 기법이 있어야만 한다. 즉, 문서의 의미적 표현을 생성한다는 것은 하위 레벨의 언어 처리 기법뿐 아니라 지식 표현 기법[5][6]과 의미 분석을 요구한다. 이해 기반의 접근 기법은 나름대로의 장점에도 불구하고 특정 분야에 대한 지식을 사용해야 한다는 어려움이 있다. 이러한 문제를 해결하고자 하는 노력으로 분야에 대한 기본 지식 대신에 WordNet [7]과 같은 어휘 데이타베이스를 사용한 연구[8]도 있다.

문장 추출 기반과 텍스트 이해 기반의 요약 기법을 절충한 방법으로 요약문에 포함될 필요한 개념을 나타내기 위해서 고안된 틀(template)을 이용한 접근 기법[6][9]이 있다. 틀의 슬롯(slot)은 원시 문서를 분석함으로써 채워지게 되고 요약문을 생성하는데 사용된다. 틀은 특정 분야의 전문가나 학습 코퍼스로부터 추출된 주요 내용을 학습함으로써 직접 디자인되어질 수 있다. 슬롯을 채우는 것은 비교적 피상적인 자연언어 처리로 이루어질 수 있는 반면에 각 분야에 대해서 틀을 만들어야 하는 단점이 있다. 예를 들어, Paice[9]는 요약을 위해서 색인과 요약을 효과적으로 결합하고자 하는 연구를 시도하였는데, 고도로 정형화된 연구 논문을 대상으로 하였으며 문서의 내용은 의미 기반의 프레임(frame)을 사용함으로써 조직화하였다. 원시 문서의 내용 중에서 단서가 되는 문체나 구조는 프레임의 다양한 슬롯을 채우기 위한 후보로 선정이 되며, 슬롯에 채워질 최종적인 실제 개념은 다양한 후보와 그들의 가중치를 비교함으로써 선택이 된다. Paice의 접근 방법은 주어진 특정 분야에 대해서는 효과적일 수 있지만, 새로운 분야에 적용하기 위해서 의미적 프레임을 만드는데 많은 양의 수작업이 필요하며 일반화가 될 수 없는 단점이 있다.

McKeown[6]은 일련의 신문을 대상으로 일정 기간 동안 발생한 동일 사건의 변화된 주요 내용을 요약하는 시스템을 개발하였다. 요약을 생성하기 위해서 한편의 기사에 해당되는 각 틀의 내용을 통합하기 위한 오퍼레이터를 고안하였으며, 이미 만들어진 요약문의 분석에 의한 경험을 기반으로 오퍼레이터를 적용하게 되며 이를 통해서 간단한 요약문장을 생성하게 된다. McKeown이 제안한 구조는 다양한 길이의 요약문장을 생성할 수 있으며, 서로 독립적인 오퍼레이터를 결합하여 복잡한 요약문장을 생성하는 것이 가능하다. 이 시스템은 문서 추출 시스템(document extraction system)이 생성한 틀을 기반으로 요약을 하는데, 문서 추출 시스템은 특정 분야의 문서를 처리하고 미리 정의한 틀에 들어갈 다양한 사실을 추출하는 시스템이다.

Miike[10]는 분야에 독립적으로 적용 가능하도록 언어적 지식만을 사용하여 문서 구조를 분석하는 시스템을 개발하였다. 텍스트의 구조를 분석하는데 있어서 접속어, 지시어, 관용 표현 같은 언어적 단서를 기반으로 단락과 문장 사이의 관계를 결정한다. Miike가 제안한 시스템은 텍스트 구조를 미리 분석한 후 저장함으로써 문장 사이의 수식 관계를 나타내는 상대적 중요도에 따라서 문장을 추출함으로써 실시간으로 요약을 할 수 있다.

3. 시스템 구조

본 연구에서 개발한 시스템은 요약문에 포함될 확률이 높은 문장을 추출함으로써 문서를 요약하는 시스템이다. 각 문장에 대한 확률 값은 요약문에서 발생할 가능성이 있는 특정 자질 출현을 기반으로 추정하게 되며, 자질에 대해서 필요한 통계 정보는 학습 코퍼스로부터 추출한다. 본 연구의 접근 방법은 Kupiec이 영어 문서에 대해서 시도한 방법[1]에 기반을 두고 있지만 여러 가지 측면에서 이를 확장 발전시켰다.

확률 값은 각 자질에 대해서 독립적으로 계산이 되며, 최종 값을 계산하기 위해서 각 자질의 확률 값을 통합하게 된다. 이 과정에서 각 자질이 요약문의 특성이 될 확률과 포함되지 않을 확률이 계산되며 이 두 값은 Dempster-Shafer 규칙에 의해서 하나의 값으로 표현된다. 또한 Kupiec의 경우는 단서 단어, 색인어, 위치 정보 등을 자질로 사용하였지만 본 연구에서는 문서의 중심성(centrality), 제목에 대한 유사도, 텍스트 구성 요소(text component) 등을 추가적인 자질로 사용하고 있다. 문서의 중심성 자질은 문서에 대한 문장의 유사도 값을 의미하며, 제목에 대한 유사도 자질은 문장과 문서 제목 사이의 유사도 값을 의미한다. 텍스트 구성 요소 자질은 문장이 지니고 있는 특성에 따라 미리 정의한 구성요소 중 하나로 분류하는데, 텍스트 구성요소 중에서 중요내용에 속하지 않는 문장을 사전에 제거하는 필터(filter) 역할로 사용되었다. 또한, 확률 계산 과정에서 구성요소를 자질 중의 하나로 사용함으로써 얼마나 효과적으로 이용할 수 있는지에 대한 실험도 수행하였다.

개발한 시스템은 그림 1에서 보는 바와 같이 학습 과정(training process)과 요약 생성 과정(summarization

process)으로 구성된다. 학습 과정은 수동으로 작성한 요약 문장이 포함된 코퍼스로부터 필요한 정보를 추출하며, 요약 과정은 학습 과정에서 추출한 정보를 이용하여 원시 문서 내의 각 문장이 요약문에 포함될 확률을 계산한다.

요약 과정을 세분하면 두 부분으로 구성할 수 있다. 전반부는 텍스트 구성요소를 판별하는 부분으로 요약 과정의 후반부에 적합한 구성요소만을 요약의 대상이 되도록 함으로써, 중요하지 않은 텍스트 구성요소로 판단되는 문장을 미리 제거하여 계산량을 줄이는 효과를 얻을 수 있다. 후반부는 요약 문장으로서 가치가 있는지를 계산하는 부분으로 전반부 과정에서 중요 텍스트 구성 요소로 판별된 문장만을 대상으로 한다. 텍스트 구성요소는 일반적인 자질로서 사용될 수 있지만 본 연구에서는 필터 역할에 더 많은 비중을 두고 있다.

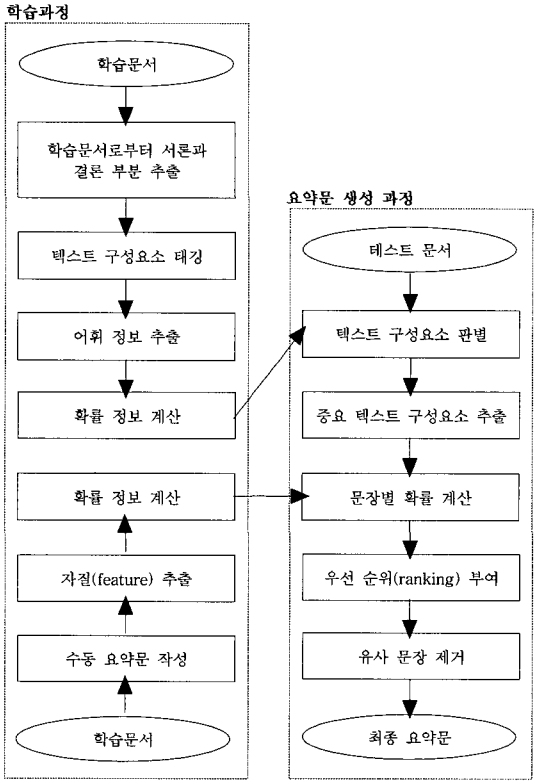

그림 1 요약 시스템 구조

요약 과정과 마찬가지로 학습 과정도 크게 두 부분으로 구성된다. 하나는 텍스트 구성 요소 판별과 관련이 있는 부분이며, 다른 하나는 요약 문장을 추출하는 과정과 관련되는 부분이다. 즉, 텍스트 구성 요소 판별 작업은 각 문장을 텍스트 구성요소 중의 하나로 분류하는 작업으로 확률을 기반으로 하고 있는데 학습 과정에서 필요한 통계적 정보를 추출하게 된다. 본 연구의 구현에 사용한 문서는 중요 요약 문장이 주로 문서의 서론과 결론 부분에 나타나는 특징이 있으며, 이 부분의 각 문장은 다음의 텍스트 구성요소 중의 하나에 포함되는 경향이 있다.

- 배경(background)
- 중요 내용(main theme)
- 문서 구조 설명(explanation of the document structure)
- 향후연구(future work)

본 연구를 통해서 실험한 결과, 가장 유용한 텍스트 구성요소는 '중요내용'이었으며 사람이 선택한 요약 문장의 96%이상이 이 부류에 속하였는데, 요약을 필요로 하는 사용자에 따라서 '중요내용'보다는 '배경'이 중요하다고 생각할 수 있으며 반대의 경우도 충분히 발생할 수 있다. 그러나, 학습문서를 대상으로 한 실험에서 평가자들이 추출한 문장의 96%가 '중요내용'에 속하였으므로 이를 근거로 실제 요약문장을 추출할 때 '중요내용'을 필터링으로 하는 실험을 시도하였다.

학습 과정에서는 학습 문서의 각 문장을 위의 텍스트 구성요소 중의 하나로 수동 태깅한 후, 각 구성요소에 해당되는 어휘 정보를 추출하게 되는데 이는 특정 구성요소를 나타내는 척도로서 사용된다. 텍스트 구성요소 판별에 필요한 정보 외에 임의의 문장이 요약문장으로 포함될 확률을 계산하기 위해서는 자질에 대한 통계적 정보를 추출하는 과정이 필요하다. 이는 학습 문서에 대해서 수동으로 요약을 작성한 후, 수동 요약문으로부터 단서 단어나 위치 정보 자질 등에 대한 통계적 정보를 추출함으로써 학습하게 된다.

요약 과정에서는 문장이 요약문장으로 포함될 확률을 계산하기 위해서 학습 과정에서 추출한 정보를 이용하게 된다. 요약 문장의 추출은 문장에 여섯 가지의 자질이 존재하는 정도에 의해서 이루어지며, 중심 자질과 제목 유사도 자질은 테스트 문서의 문장을 대상으로 하기 때문에 학습과정과는 독립적이다. 자질에 대한 여섯 가지의 확률이 계산되면 문장이 요약문장에 포함될 가능성을 나타내는 신념도를 계산하기 위해서 Dempster-Shafer 규칙을 적용하여 대표 값을 계산한다. 마지막 과정은 그림 1에서 보는 바와 같이 이상의 과정에서 추출

한 후보 문장으로부터 중복된 문장을 제거하는 단계로 서론과 결론 부분에 유사한 문장이 중복되는 경우가 종종 발생하기 때문에 필요하다. 문장 사이의 유사도는 벡터 공간 모델[11]에서 문서와 질의의 유사도를 계산하는 방법을 이용하여 계산한다.

4. 요약 문장의 추출 과정

본 장에서는 학습 과정으로부터 추출한 통계적 정보를 기반으로 하여 문장을 추출함으로써 요약문을 생성하는 과정을 자세히 기술한다. 그림 2는 요약문 생성 과정을 단계별로 보여주고 있으며, 자질에 관한 모든 필요한 통계적 정보는 5장에 기술된 내용에 의해 계산된다고 가정한다.

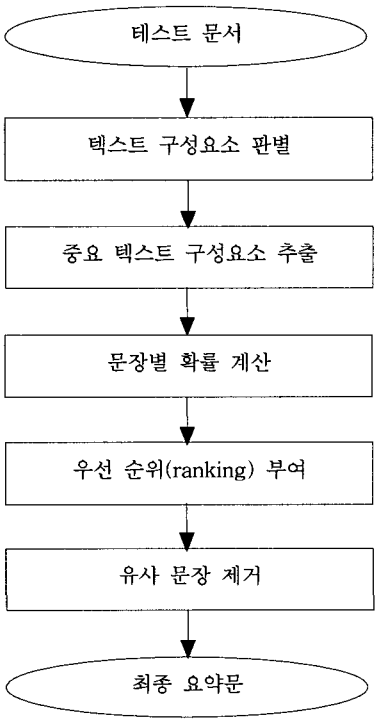

그림 2 요약 생성 과정

4.1 텍스트 구성요소의 판별

본 과정은 요약 대상이 되는 문서의 각 문장이 텍스트 구성요소 중에서 어느 부류에 속하는지를 미리 판별함으로써 요약문에 포함될 가능성이 적은 문장을 미리 제거하는 단계이다. 이러한 과정을 통해서 확률 계산 과정중 소비하는 계산 시간을 절약할 수 있을 뿐만 아니라 전문 처리 과정에서 발생하는 노이즈(noise)를 미리 제거할 수 있게 된다. 이와 같이 텍스트 구성요소 정보의 주된 역할은 필터기능이며, 앞서 언급했던 것처럼 자질로서도 사용할 수 있다.

식(1)은 베이스 규칙(Bayes' rule)으로부터 유도된 것으로서 임의의 문장 s가 4개의 텍스트 구성요소(TC_i, i=1..4)에 포함될 확률 계산식이며, TC_i는 배경, 중요내용, 문서구조설명, 향후연구를 각각 의미한다.

$$P(TC_i|s) = \frac{P(s|TC_i)P(TC_i)}{P(s)} \approx \frac{\prod_{k=1} P(t_k|TC_i)P(TC_i)}{\prod_{k=1} P(t_k)} \quad (1)$$

식(1)에서 $P(s|TC_i)$는 임의의 문장 s가 텍스트 구성요소 i에 나타날 확률이며, $P(TC_i)$는 특정 텍스트 구성요소가 나타날 확률, 그리고 $P(s)$는 요약 문장이 학습 코퍼스에서 나타난 확률이다. 문장 s가 n개의 각기 독립적인 단어로 구성이 되었다고 가정하면, 확률 값 $P(s)$는 각각의 단어 t_k에 대해서 구한 확률 값들의 곱으로 표현될 수 있다. 예를 들어, $P(t_k|TC_i)$는 단어 t_k가 텍스트 구성요소 TC_i에 속하는 모든 문장에 나타날 확률이다. 식 (1)을 사용함으로써 각 문장에 대하여 가장 큰 값을 갖는 텍스트 구성요소를 판별할 수 있게 되며, 이러한 과정을 통해서 선택된 텍스트 구성요소가 요약에 유용한 문장인지 결정할 수 있게 된다.

표 1은 각 텍스트 구성요소에 속하는 대표적인 단어들로 학습 요약문장에서의 출현 빈도수를 근거로 자동 추출한 결과이다. 일반적으로 "요약", "방법", "결과" 등과 같은 단어가 텍스트 구성요소 중 '중요내용'에 자주 출현하리라 기대되지만, 실제 학습 데이타에서는 다른 구성요소에 출현할 확률이 높았으며 본 연구의 기본적인 접근방법이 확률을 기반으로 하고 있기 때문에 학습 내용을 그대로 사용하였다.

표 1 텍스트 구성요소별로 추출된 단어

텍스트 구성요소	추출된 단어
배경	관련, 관심, 관점, 구조, 다양, 일반, 최근, 현재, ...
중요내용	논문, 방안, 문제, 모델, 목적, 제공, 조사, 처리, ...
문서구조 설명	구성, 기술, 방법, 설명, 연구, 요약, 제시, 제안, ...
향후연구	개발, 결과, 과제, 문제, 연구, 적용, 필요, 개선, ...

4.2 자질에 대한 확률 계산

본 과정은 학습 과정에서 추출한 통계적 정보와 실제 요약의 대상이 되는 문장으로부터 얻은 정보를 이용하

여 요약문에 포함될 확률을 계산하는 과정이다. 이는 문장 s에 대하여 본 연구에서 설정한 각각의 자질 특성에 대한 기여도를 고려함으로써 요약문에 포함될 신념도(belief)를 계산하게 되는데 계산과정을 자세히 살펴보면 다음과 같다.

4.2.1 단서단어 자질 (Cue Words Feature)

단서 단어란 수동으로 작성한 요약문장에서 빈도수가 많은 단어나 구를 의미한다. 한 문장에 단서 단어가 있을 경우 그 문장이 요약문에 포함될 확률이 높다고 가정하는데, 이는 단서 단어가 학습 문서로부터 수동으로 선택한 요약문장에서 빈도수가 많은 단어를 기준으로 추출하기 때문이다. "본 논문의 목적" 이나 "~을 제안한다"와 같은 구를 포함하는 문장이 이에 속하게 되는데, 이러한 구의 인식은 파싱을 하는 것이 아니라 빈도수가 많은 색인 단어만을 추출함으로써 인식하게 된다. 즉, "본 논문의 목적"과 같은 경우 "본 논문 목적"이 단서 단어 구가 된다. 임의의 문장 s에 대하여, 단서단어 CW_i가 있을 경우 요약문 S에 속할 확률은 베이스 규칙을 이용하여 식(2)와 같이 표현될 수 있다.

$$P(s\in S|\ CW_i)=\frac{P(\ CW_i|s\in S)P(s\in S)}{P(\ CW_i)} \tag{2}$$

식(2)에서 $P(CW_i|s\in S)$는 단서 단어가 요약문에 나타날 확률이며 학습문서를 대상으로 수동으로 작성한 요약문에서의 단서단어의 빈도수를 이용하여 계산할 수 있다. $P(s\in S)$와 $P(CW_i)$도 학습 문서의 데이타를 이용하여 계산하는데, 이용 방법은 5장에 기술되어 있다.

4.2.2 부정단어 자질 (Negative Words Feature)

요약문 이외의 문장에서 빈도수가 많은 단어를 부정단어(negative words)라 정의하는데, "예를 들어"와 같은 구가 이에 속한다. 단서단어 자질과 마찬가지로 임의의 문장에서 부정단어 NW_i가 요약문에 포함되지 않은 확률은 식(3)과 같이 나타낼 수 있다.

$$P(s\notin S|\ NW_i)=\frac{P(\ NW_i|s\notin S)P(s\notin S)}{P(\ NW_i)} \tag{3}$$

다른 자질이 긍정과 부정적인 증거로 사용할 수 있는데 비해서 이 자질은 부정적인 증거로만 사용되며, 확률 값의 계산은 식(2)와 같은 방법으로 계산된다.

4.2.3 위치 자질 (Position Feature)

임의의 문서에 속한 문장들은 문장이 위치한 곳에 따라서 크게 3가지 경우, 처음(처음 5개의 문장), 마지막(마지막 5개 문장), 중간(처음과 마지막 이외의 문장) 부분으로 구분할 수 있다. 문서의 내용이 짧은 경우에는 처음과 마지막 부분이 중복될 수도 있으며 중간 부분이 없을 수도 있다. 본 연구에서는 문서의 서론과 결론 부분만을 요약의 대상으로 하고 있기 때문에 각 문장에 대한 위치정보는 6개(서론 3개, 결론 3개)중에서 하나의 값을 적용할 수 있다. 수동으로 작성한 요약문을 관찰한 결과, 서론 부분의 마지막 5개 문장(0.7097)과 결론 부분의 처음 5개 문장(0.4095)이 다른 위치에 있는 문장보다 요약문에 포함될 가능성이 높은 것으로 나타났다. 문서별 문장의 수는 평균 약 20개이므로 서론, 결론 각각 10개의 문장이 평균적으로 있다고 할 수 있는데 이를 근거로 5개의 문장을 선택하였다. 6개의 위치 중 하나에 속하는 문장이 요약문에 포함될 확률은 식(4)와 같이 표현할 수 있다.

$$P(s\in S|\ P_i)=\frac{P(\ P_i|s\in S)P(s\in S)}{P(\ P_i)} \tag{4}$$

식(4)에서 $P(P_i|s\in S)$와 $P(P_i)$는 위치 i 에서 요약문장이 나타날 확률과 문장이 전체 문장 중에서 위치 i에 나타날 확률을 각각 나타낸다. 이 식에서 임의의 문장에 대한 확률 값은 문장을 구성하고 있는 단어에 의해서 영향을 전혀 받지 않고 단순히 그 문장이 속한 위치에 의해서만 결정된다. 그리고 각 문장은 하나의 위치 정보만을 갖고 있기 때문에 위치 i 에 대한 확률 값은 학습과정에서 계산되어진 6개의 값 중에서 하나를 갖게 된다.

4.2.4 키워드 자질 (Keyword Feature)

키워드나 중요 내용을 내포하고 있는 단어는 문서의 내용을 대표하기 때문에 정보검색 분야에서 중요한 역할을 해왔다. 전통적인 정보검색 기법에서는 키워드 중심의 검색 기법이 주로 이용되어 왔으며, 직관적으로도 키워드를 포함하고 있는 문장이 요약문장으로 포함될 경향이 높으므로 본 연구에서도 자질중의 하나로 사용하고 있다. 이 자질을 사용할 경우 문장에 포함된 단어가 중요하고 많을수록 그 문장이 갖게되는 확률 값도 더 커지게 된다. n개의 키워드(t)를 갖는 임의의 문장 s가 요약문 S에 포함될 확률은 식(5)와 같다.

$$P(s\in S|s)=\frac{P(s|s\in S)P(s\in S)}{P(s)}\approx\frac{\prod_{k=1}^{n}P(t_k|s\in S)P(s\in S)}{\prod_{k=1}^{n}P(\ t_k)} \tag{5}$$

4.2.5 중심성 자질 (Centrality Feature)

문장이 문서 내에서 어느 정도 구심점 역할을 하는지에 대한 지표로서 문장의 벡터(vector) 값과 문서 전체의 벡터 값을 비교함으로써 계산될 수 있다.

4.2.6 제목에 대한 유사 자질 (Resemblance to the Title)

문장이 문서의 제목과 유사한 정도를 고려한 자질로서 보통 문서 제목이 문서의 내용을 대표하기 때문에 요약에 있어서 의미가 있는 자질이다. 따라서, 본 자질은 임의의 문장과 제목에 있는 중복되는 용어의 수에 의해서 유사도가 결정된다.

4.3 증거 수집(Evidence Gathering)

요약의 길이는 사용자의 목적에 따라서 다르기 때문에 지정한 개수만큼의 상위 문장이 최종 요약문에 포함될 수 있도록 문장의 우선 순위를 부여하는 것이 하나의 방법이 될 수 있다. 우선 순위를 부여하기 위해서는 앞서 기술한 자질로부터 계산되어진 확률 값들을 통합하는 것이 필요하다. 본 연구에서는 문장이 요약문에 포함될 신념도(belief)를 계산하기 위해서 Dempster-Shafer 규칙[12]을 적용하였다. 일반적인 신념도의 개념은 명제 집합에 대한 증거의 정도를 수치로 표현한 것으로 전혀 증거가 없는 값인 0에서부터 확실한 증거를 나타내는 1 사이의 범위를 갖는다. 이러한 값은 어떤 명제에 대한 신념도를 나타낼 뿐 아니라 가지고 있는 정보의 양을 나타낸다.

Dempster-Shafer 규칙을 사용함으로써 임의의 문장이 요약문에 포함될 신념도를 얻을 수 있는데, 새로운 증거가 추가되는 경우 전체적인 신념도는 새로운 증거의 정도에 따라서 현재의 신념도와 명제의 최대 신념도 사이의 차이를 줄이는 방식으로 증가하게 된다[13]. 다수의 값을 통합하기 위해 다른 방법을 적용하는 것도 가능하지만, 규칙 자체가 직관적이고 이론적 근거가 타당성 있는 Dempster-Shafer 규칙을 적용하였다. 문장이 요약문에 포함되지 않을 신념도는 포함될 신념도와 유사한 방식으로 계산할 수 있는데, 긍정 신념도(positive belief)와 부정 신념도(negative belief) 두 값의 차이가 문장이 요약문에 포함될 최종적인 값이 된다. 본 연구에서 서로 다른 여러 개의 자질을 사용하고 있기 때문에 신뢰도나 중요도 측면에서 그 중요도가 다를 수 있지만, 모든 자질이 최종적인 신념도에 미치는 영향을 같다고 가정하고 자질로부터 얻은 값을 통합하는데 있어서 최종적인 값이 0과 1사이가 되도록 정규화하였다.

4.4 중복 문장의 제거

요약문장 추출 과정에서 마지막 단계는 후보 문장 중에서 중복되는 문장을 제거하는 단계로 문서의 서론과 결론 부분에 동일하거나 유사한 문장이 반복적으로 기술되는 경향이 있기 때문에 필요한 과정이다. 본 과정이 없다면 부자연스러운 결과가 생성될 수 있으며, 요약문의 길이가 고정된 경우 중요한 요약 문장을 제외시키는 결과를 초래할 수 있다. 이러한 문제점을 해결하기 위해서 모든 문장들 사이의 유사도를 계산하는데, 두 문장 사이에 공통된 단어가 많으면 많을수록 서로 유사한 정도는 커지게 된다. 문장 사이의 유사도가 임계값(threshold)보다 크면 두 문장 중에서 순위가 낮은 문장을 제거하는데, 학습과정에서 중복문장이라고 판단된 문장의 약 74%가 0.5 이상의 유사도를 보였으며 0.5 이하의 값을 갖는 나머지 문장들은 이음동의어를 사용한 경우로 본 연구에서는 이에 대한 처리는 하지 않았다. 따라서 중복문장 판단을 위해 사용한 임계값은 0.5를 사용하였다.

이상의 과정을 통해 최종 요약문이 추출되며 본 연구를 통해 개발한 자동 요약 시스템을 통해 추출된 요약 결과를 부록 A에 제시하였다.

5. 학습 과정

그림 3에서 보는 바와 같이 학습 과정은 두 부분으로 구성이 되어있다. 하나는 텍스트 구성요소에 존재하는 특성을 학습하는 것이고, 다른 하나는 자질이 요약문장에 미치는 영향을 학습하는 것이다.

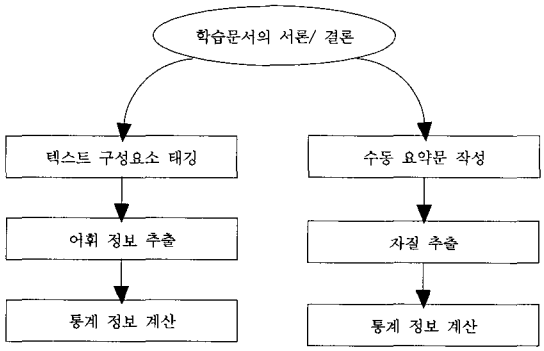

그림 3 학습 과정

본 연구에서는 한글 및 한국어 정보처리 학회에서 발표된 바 있는 논문 50건을 학습 문서로 사용하였다. 전체 학습문장의 개수는 995개이며, 텍스트 구성요소 별로는 배경 438개, 중요내용 364개, 문서구조설명 88개, 향후연구 105개의 문장이 각각 속하였다. 텍스트 구성요소가 갖는 특성을 파악하기 위해서 특정 구성요소에 속하는 문장에서 빈도수가 높은 단어나 구를 관찰함으로써 각 구성 요소의 특성을 학습하게 된다. 이는 4장에서 기

술한 식(1)에서 $P(t_k|TC_i)$에 해당한다. 단어나 구에 해당되는 t_k에 대해서 구성요소 TC_i에 속하게 될 확률은 식(6)에서처럼 베이스 규칙을 사용해서 표현할 수 있다.

$$P(TC_i|t_k)=\frac{P(t_k|TC_i)P(TC_i)}{P(t_k)} \quad (6)$$

위의 식에서 $P(t_k|TC_i)$는 용어(term) t_k 가 특정 텍스트 구성요소 i 에 나타날 확률이며, 학습 문서로부터 특정 구성요소에 속하는 모든 문장에서 용어가 발생하는 빈도수 등의 통계적 정보를 학습함으로써 계산할 수 있다. $P(TC_i)$는 학습 코퍼스 중에서 구성요소 i 에 속하는 문장 수에 대한 값을 나타내며, $P(t_k)$는 전체 학습 코퍼스에서 용어 k의 상대적 빈도수를 의미한다.

학습과정에서는 이와 같이 텍스트 구성요소에 대한 정보를 학습하는 것과 더불어 양질의 요약문을 나타내는 것으로 판단되는 문장으로부터 통계적 정보를 학습하게 된다. 이를 위해서 먼저 수동으로 학습문서로부터 중요한 요약 문장을 판별해낸 후, 그 문장들로부터 필요한 자질 정보를 추출하게 된다. 4장에서 기술한 식(2)의 확률 $P(CW_i|s\in S)$는 학습 코퍼스에서 수동으로 작성한 요약문에 나타난 단서 단어의 빈도수를 이용하여 계산할 수 있다. 식에서 $P(s\in S)$는 코퍼스의 전체 문장 수와 요약문장 수의 비율, 그리고 $P(CW_i)$는 코퍼스 전체 단어 수와 단서단어 수의 비율로서 계산할 수 있으며, 부정 단어에 대한 확률도 동일한 방법을 사용한다.

6. 평 가

본 연구에서는 개발한 시스템에 대해서 두 가지 측면으로 나누어 평가를 시도하였다. 첫째는 텍스트 구성요소 판별의 기능에 대한 정확도를 평가하였고, 다른 하나는 여러 가지 자질을 사용하여 생성한 요약문의 품질을 평가하는 것이다. 텍스트 구성요소 판별의 경우, 구성요소 각각에 대한 판별 결과를 평가하는 것이 가능하지만 판별 목적이 요약 결과에 대한 성능 향상에 있기 때문에 중요내용의 판별에 대한 평가만을 하였다.

본 연구에서 개발한 시스템을 평가하기 위해서 평가자 6명(대학원생 3명, 학부생 3명)에게 다음의 내용을 각각 수행하도록 하였다.

- 각 문장이 텍스트 구성 요소 중에서 중요내용에 속하는지 아니면 다른 구성 요소에 속하는지에 대한 분류 (모든 텍스트 구성요소의 판별이 가능하지만 본 연구에서는 중요내용에 초점을 두기로 하였다.)
- 문서에 대한 요약문 생성을 위해 위에서 중요내용 문장으로 판별된 문장들을 대상으로 요약문에 포함될 우선 순위 부여
- 요약문으로 추출된 문장들 중에서 유사한 문장 제거

실험을 위해 사용한 문서는 학습 과정에서 사용했던 문서를 제외한 30개의 문서를 사용하였으며, 동일한 문서에 대해서 두 명이 평가하였고 평가자에 의해 추출된 문장 중에서 일치하는 문장만을 최종 요약문으로 사용하였다. 이와 같이 평가자에 의해 생성된 요약문장과 시스템에 의해 자동으로 생성한 요약문을 비교함으로써 개발한 시스템의 성능을 평가하게 된다. 요약의 품질에 대해서 평가하는 것은 요약문을 읽는 사람의 요구 사항과 아주 밀접한 관련이 있어 모든 사람이 동의할 수 있는 요약문을 생성한다는 것은 어려운 일로 요약에 대한 평가 자체가 난해한 문제이면서도 중요한 부분이다. 이와 같이 개발한 시스템의 요약 기능을 평가하기 위해서 각 문서는 시스템과 평가자에 의해서 요약이 되며, 평가자로 하여금 요약문에 포함시키고 싶은 상위의 5개 문장을 추출하도록 하였지만 두 명이 공통으로 추출한 문장이 많지 않아 요약 문장의 수가 적은 경우도 있다. 시스템의 경우도 마찬가지로 우선 순위를 부여하여 5개의 문장을 추출하도록 하였다. 정확도(precision)와 재현도(recall) 계산을 위해서는 정보검색 분야에서 평가를 위해 일반적으로 사용하고 있는 방법을 적용하였다. 평가자에 의해 추출된 요약 문장을 적합 문서 집합(set of relevant document)으로 하고 시스템에 의해 추출된 문장을 검색된 문서로 취급함으로써 가능한데, 정보검색의 관점으로 볼 때 테스트 집합(test collection)으로 간주할 수 있다.

6.1 텍스트 구성요소 판별 기능에 대한 결과

텍스트 구성요소 판별에 대한 시스템의 평가를 위해서 시스템이 선택한 문장들과 평가자에 의해서 추출된 문장들과 비교하는 방법을 사용하였으며, 텍스트 구성요소 중에서 중요내용이 주로 요약문에 포함될 것이라고 가정하였기 때문에 다른 구성요소 판별에 대한 평가는 제외하였다. 시스템은 텍스트 구성요소 중에서 다른 요소보다 중요내용에 포함될 가능성이 높을 경우 그 문장을 선택하였다. 평가자가 중요내용으로 추출한 문장은 총 295개였으며 이 중에서 210개가 시스템에 의해서 추출되어 67%의 재현도와 40%의 정확도를 보였다. 텍스트 구성요소 판별을 필터의 목적으로 사용하고 있기 때문에 정확도보다는 재현도에 대한 성능을 향상시키는 것이 중요하다. 왜냐하면 텍스트 구성요소 중 중요내용으로 판별된 문장만이 요약의 후보가 되기 때문에 이 과정에서 제외된 문장은 요약문장으로 추출될 가능성이

전혀 없기 때문이다. 따라서 중요내용에 대한 확률이 다른 구성요소에 비해 작더라도 임계값 이상의 확률을 갖는 문장을 요약의 후보문장으로 포함시켜 재현도를 향상시킬 수 있다. 그리고, 구성요소의 분류 기능이 학습 문서에 의존적인 경향이 있기 때문에 요약 대상이 되는 문서를 학습 문서의 일부로 포함시켜 재학습시키는 방법이 재현도를 향상시킬 수 있는 방편이 될 수 있다.

6.2 요약 결과

3장에서 논의했던 것처럼 현재 구현한 시스템 평가의 초점은 요약문을 생성하는데 있어서 자질을 필터로 사용한 경우와 사용하지 않은 경우의 성능을 비교하는 것이다. 성능 평가치는 정보검색 평가에서 일반적으로 사용하고 있는 11-point 평균 정확도로 나타냈으며, 문서 결론 부분의 처음 5개 문장을 추출한 값인 0.3682를 비교를 위한 기본 값으로 사용했다. 기본 값은 정보가 전혀 없는 상태에서 연속적인 5개의 문장을 추출했을 때 가장 좋은 성능을 보인 값을 선택하였다.

표 2는 요약과정 이전에 서로 다른 필터 방법을 적용한 실험 결과를 보여주고 있는데, 이를 통해서 필터 방법을 사용한 후 요약한 경우와 사용하지 않은 경우를 비교할 수 있다. 괄호 안에 있는 숫자는 기본 값과 비교해서 성능 향상을 보인 결과에 대해 백분율로 나타낸 값이다.

표 2 여과 방법과 다양한 자질을 사용한 경우의 실험 결과

사용 자질	사용 필터			
	사용하지 않은 경우	텍스트 구성요소	키워드	텍스트 구성요소와 키워드
모든 자질(6개)	0.3865(+5)	0.3949(+7)	0.3913(+6)	0.3996(+9)
단서단어 자질 제외	0.3116	0.3203	0.3172	0.3259
위치 자질 제외	0.3544	0.3572	0.3594	0.3594
텍스트구성요소 자질 제외	0.4007(+9)	0.4167(+13)	0.4228(+15)	0.4228(+15)
제목유사 자질 제외	0.3439	0.3591	0.3487	0.3639
키워드 자질 제외	0.3976(+8)	0.4033(+10)	0.4022(+9)	0.4075(+11)
중심성 자질 제외	0.4318(+17)	0.4337(+18)	0.4365(+19)	0.4389(+19)
제목+단서단어+위치 자질	0.4274(+16)	0.4363(+18)	0.4420(+20)	0.4436(+20)

표 2에서 각 열은 필터로 사용한 자질을 보여주고 있는데 예를 들어, 두 번째 열은 문장이 텍스트 구성요소 중에서 주요내용으로 판별된 경우에만 요약의 대상이 되어 실험한 결과이며, 세 번째 열은 키워드를 포함하고 있는 문장만을 일차적인 요약문장의 후보로 하여 실험한 경우이다. 표에서 행은 필터를 한 후에 요약 과정에서 사용한 자질을 보여주고 있다. 예를 들어 첫 번째 행의 모든 자질은 텍스트 구성요소, 위치, 중심성, 제목 유사도, 키워드, 단서 단어, 부정 단어 자질을 사용한 경우이며, 세 번째 행은 위치 자질을 제외한 5개의 자질을 사용하여 실험한 결과다. 결과에서 보듯이 단서단어, 위치, 제목 자질을 제외했을 경우 기준 값보다 성능이 저하되는 것을 알 수 있으며, 나머지 자질을 제외했을 경우에는 성능이 향상되는 것을 알 수 있다. 자질을 제외했을 경우 성능이 향상되었다는 것은 요약의 성능에 불필요한 자질로 판단이 되며, 반대로 성능이 저하되었다는 것은 반드시 필요한 자질들이므로 이들의 조합이 가장 우수한 결과를 보일 것으로 보여 실험한 것이 마지막 행이다.

3가지 필터 방법 중에서 텍스트 구성요소와 키워드 자질을 필터로 사용한 경우가 그렇지 않은 경우보다 좋은 성능을 보여주고 있으며, 텍스트 구성요소만을 사용한 경우도 좋은 결과를 보이고 있다. 모든 자질을 사용한 실험 결과 그 차이는 작지만, 자질을 다양하게 조합해서 실험한 결과는 많은 차이를 보여주고 있다. 필터 방법을 사용하지 않았을 경우 중심성 자질을 제외했을 때 가장 좋은 성능을 보이고 있는데 이는 중심성 자질을 포함한다면 요약의 성능을 향상시킬 수 없음을 나타낸다. 또한, 중심성 자질, 텍스트 구성요소 자질, 키워드 자질을 제외했을 경우에 결과 값이 향상되기 때문에 요약의 자질로 포함시키면 악영향을 미친다는 결론을 얻을 수 있다. 그리고, 제목 유사도 자질, 단서 단어 자질, 위치 자질만을 사용하였을 경우에는 여과 방법의 종류와는 상관없이 항상 가장 좋은 결과를 보이고 있다. 그러나, 이러한 사실이 텍스트 구성요소 자질과 키워드 자질의 사용 가치가 없음을 의미하지는 않는데, 이는 필터로서 사용한 경우에 유용한 결과를 보이고 있기 때문이다. 즉, 실험 결과 중에서 가장 좋은 성능을 보인 경우는 텍스트 구성요소와 키워드 자질을 필터로 사용하고 3개의 가장 좋은 자질(제목과의 유사도, 단서 단어, 위치)을 사용한 실험이다. 이러한 사실로 미루어 볼 때 키워드 자질은 후보 문장을 선택하는데 있어서 중요한 역할을 하지만, 좋은 요약 문장을 생성하는데 있어서는 필요 조건이 아님을 알 수 있으며 이와 같은 해석은 중요 내용을 판별하는 텍스트 구성요소에 대해서도 동일하게 적용이 된다.

일반적으로 문서의 키워드는 정보검색 분야에서 중요

한 요소로 작용을 하지만 요약의 성능에는 좋지 않은 영향을 미친다고 알려져 있다[9]. 따라서 본 연구에서는 텍스트 구성요소 자질을 필터로 사용하는 것이 키워드 자질을 사용하는 것보다 우수할 것이라는 가정을 하고 각각을 필터로 사용하는 실험을 하였지만 결과에서 보듯이 유사한 결과를 보이고 있다. 결과론적인 측면에서 분석하면 필터 개수의 증가만을 야기하게 되었고 잘못된 가정을 설정한 것으로 판단될 수 있지만, 향후에는 텍스트 구성요소 판별 작업의 정확도를 현재의 결과보다 좀 더 향상시켜 각각의 실험결과를 비교할 필요가 있다.

6.3 텍스트 구성 요소의 역할

본 연구에서 시도한 다른 실험으로는 텍스트 구성요소 판별의 유용성을 알아보기 위해서 텍스트 구성요소를 필터로 사용한 경우와 사용하지 않은 경우로 나누어서 정확도와 재현도를 비교하였다. 성능 평가를 하는데 있어서 시스템에 의해서 추출된 문장 중에서 상위 5개의 문장만을 대상으로 한다는 것 외에는 6.2절에서 기술한 첫 번째 실험과 동일한 조건이다.

표 3은 시스템에 의해서 추출된 문장 수에 따른 5가지 경우에 대해서 텍스트 구성요소를 필터로 사용한 경우와 사용하지 않은 경우 각각에 대한 정확도와 재현도를 나타내고 있다. 표에서 보는 바와 같이 텍스트 구성요소의 필터 사용 여부에 따라 정확도와 재현도 두 가지 측정값에서 많은 차이가 있음을 알 수 있다. 예를 들어, 텍스트 구성요소를 필터로 사용하여 시스템이 5개의 문장을 추출한 경우, 정확도가 10%, 재현도는 48%가 각각 향상되었다. 이러한 결과로 미루어 볼 때, 텍스트 구성요소 판별이 단순한 모델이지만, 요약 과정에 있어서는 효과적으로 필터 역할을 수행하고 있음을 알 수 있다.

표 3 텍스트 구성 요소의 사용 여부에 따른 실험 결과

	텍스트 구성요소 미 사용		텍스트 구성요소 사용	
추출 문장 수	정확도	재현도	정확도 (% 변화폭)	재현도 (% 변화폭)
1	50.00	10.00	53.33 (+7)	14.55(+46)
2	36.67	14.67	43.34(+18)	21.37(+46)
3	36.67	22.00	43.33(+18)	35.46(+61)
4	35.83	28.67	41.59(+16)	45.00(+57)
5	36.00	36.00	39.53(+10)	53.19(+48)

6.4 정성적 분석(Qualitative Analysis)

본 연구에서 사용한 자질들간의 공통성을 알아보기 위해서 단서단어, 부정단어, 키워드, 제목에 속하는 단어들이 일치하는 정도를 알아본 결과는 표4에서 보는바와 같다. 예를 들어, 부정단어와 단서단어 간의 일치하는 정도는 17.75%인데 이는 $(단서단어 \cap 부정단어)/(단서단어 \cup 부정단어)$에 의해 계산된 결과이다. 표3에 나타난 4개의 자질들은 비슷한 유형의 정보를 어느 정도 공유하고 있으며 완전히 독립적인 자질은 아님을 알 수 있는데, 단서단어와 키워드 사이의 공통성이 다른 자질보다 높은 이유는 키워드를 기반으로 단서단어를 추출하기 때문이다. 그리고 표에는 나타나 있지 않지만 중심성 자질과 제목유사 자질 또한 일부 같은 정보를 포함하고 있으며, 위치 자질도 다른 자질의 특성을 어느 정도 공유하고 있는데 이는 본 연구의 접근 방법이 통계적인 확률을 기반으로 하고 있기 때문이다.

표 4 자질 단어간의 공통성

	부정단어	키워드	제목 단어
단서단어	17.75%	35.85%	20.78%
부정단어	-	17.44%	18.90%
키워드	-	-	23.13%

본 연구에서는 자질에 대한 의미적 특성을 이해하기 위해서 시스템에 의해 생성된 자질을 분석하였다. 한국어를 대상으로 하는 대부분의 자연 언어 응용에서는 문장을 이루는 단어의 순서가 자유롭기 때문에 구(phrase) 단위를 인식하는 것 자체가 문제점으로 대두되지만, 영어의 경우에는 문장이 구 단위로 표현되기 때문에 한국어에 비해 인식하기가 쉬운 편이다. 그러나 요약 과정에서 사용한 문서에는 영어에서처럼 단서 단어 구의 어순이 일정한 패턴을 유지하고 있었기 때문에 구를 인식하는데 문제가 발생하지 않았다.

그리고 본 연구를 통해서 관찰된 결과 단서 단어의 30%이상이 특정 분야에서는 중요하지 않은 개념을 나타내는 단어에 속하지만, 본 시스템을 다른 분야에 적용할 경우 해당 분야의 학습 문서를 대상으로 단서 단어를 추출하기 때문에 문제가 되지는 않는다. 예를 들어, "결과", "논문", "방법", "분석", "연구" 등과 같은 단어가 이에 속하는데 이러한 단어가 부정 단어에는 전혀 속하지 않았다.

본 연구에서 개발한 방법으로 생성하는 요약문은 사람이 일반적으로 생성하는 요약문과는 전혀 일치하지

않을 수도 있으며 자연스럽지 않을 수도 있다. 그러나, 문장을 추출하는 작업을 고려해본다면, 추출된 문장 자체는 원시 문서의 중요한 의미를 타당성 있게 대표한다고 할 수 있다. 추출한 문장의 10% 이하는 지시어를 포함하고 있는데, 대부분 이전 문장에서 언급한 "논문", "방법", "결과" 등과 같은 의미를 지시하고 있기 때문에 큰 문제는 되지 않고 있다. 이러한 사실은 선택된 문장에 나타난 지시어가 이전에 추출한 문장에서 언급한 중요한 개념이 아닌 이상 문장 추출 시 지시어의 내용을 대체할 필요가 항상 있는 것은 아님을 나타낸다.

7. 결 론

본 연구에서는 문서의 중요내용을 표현하고 있는 문장을 추출함으로서 요약문을 생성하는 방법 및 시스템을 개발하였다. 요약문장을 추출하는 방법으로써 문서 코퍼스로부터 어휘적, 통계적 정보를 고려한 확률 모델을 사용하였다. 본 연구의 의의를 살펴보면 다음과 같다. 첫째, 각 문장이 갖고 있는 특성에 따라서 문장을 분류하는 텍스트 구성 요소 모델을 사용함으로써 요약문장으로서의 특성을 갖고 있지 않는 문장을 미리 제거한다. 둘째, 본 연구에서 개발된 시스템이 영어 텍스트에 대해서 개발된 통계적 모델을 근간으로 하고 있지만, 각각의 자질을 독립적으로 적용할 수 있는 기틀을 마련하였으며 Dempster-Shafer 규칙을 이용하여 각 자질의 확률 값을 통합하여 최종 확률 값을 계산한다. 셋째, 기존의 연구에서 사용되었던 자질 외에 새로운 자질을 사용하였으며, 요약 과정에서 자질의 역할을 알아보기 위해 다양한 실험을 하였고 결론을 도출하였다.

본 연구에서 수행한 실험 결과를 통해 텍스트 구성요소를 사용함으로써 성능이 향상됨을 보였다. 텍스트 구성요소 판별을 통하여 요약 대상 데이타 양을 줄일 수 있을 뿐만 아니라 정확도와 재현도의 성능이 향상될 수 있었다. 개발된 시스템과 영어에 대해서 적용한 시스템과의 직접적인 비교를 할 수 없지만 단순한 정확도와 재현도 면에서 더 좋은 결과를 보였다.

또한 실험을 통하여 다양한 자질의 유용성에 대한 가치를 확인 할 수 있는 결과를 얻었다. 즉, 요약 문장에 대한 순위를 결정할 때 단서 단어, 위치, 제목 자질이 가장 중요한 것으로 나타났다. 또한, 텍스트 구성요소와 키워드 자질은 다른 자질과 혼합되어 사용될 경우 성능의 향상을 얻을 수 없었지만 필터로 사용될 경우에는 효과적임을 알 수 있었다.

이 분야의 연구가 아직 초기 단계이기 때문에 앞으로 연구해야 될 부분이 많이 있다. 첫째, 텍스트 구성요소 판단의 재현도가 67%로 좀 더 좋은 결과를 얻을 수 있도록 모델을 발전시켜야 한다. 둘째, 자연스러운 요약문장을 제공하기 위해서 다양한 지시어에 대한 처리를 할 수 있도록 시스템을 발전시켜야 한다. 셋째, 기본 모델이 언어에 독립적이고 모든 분야에 적용할 수 있는 일반화된 모델이지만 텍스트 구성요소 판별 부분은 언어와 분야 의존적이기 때문에 이를 보완하여 발전시킬 필요가 있다. 또한 문서의 형식이나 분야에 대한 의존성을 가급적 배제하고자 통계적인 접근방법을 사용하였지만, 본 연구의 실험에 사용하지 않는 분야의 데이타 즉, 신문기사, 일반서적, 보고서, 웹 문서 등과 같이 다양한 분야의 문서를 대상으로 실험 및 분석을 함으로써 제안하는 방법이 가질 수 있는 문서 형태에의 의존성 여부를 검증할 필요가 있다. 넷째, 좀 더 객관적이고 다양한 평가방법을 동원할 필요가 있다. 평가에 있어서 주관성이 완전히 배제될 수는 없지만 논문을 대상으로 하는 요약의 경우에는 원문 저자가 작성한 요약문과 시스템이 생성한 요약문을 비교하는 평가가 요구된다. 다섯째, 본 연구에서는 다양한 자질을 사용하고 각각의 자질에 대한 유용성을 실험을 통해 밝혔지만, 자질 인스턴스(예: 각 단서 단어)가 요약 생성에 미치는 공헌도를 분석할 필요가 있다. 여섯째, 텍스트 구성요소의 활용도를 높이기 위해서는 다양한 사용자의 관심도를 반영할 수 있는 텍스트 구성요소 단위의 요약 시스템을 개발해야 한다. 그리고, 각 요소별 가중치를 적용하여 성능을 향상시킬 수 있는 방법과 HMM, MRF 등을 이용하여 단어들의 연관관계를 활용하는 방법을 적용할 필요가 있다.

참 고 문 헌

[1] Kupiec, J., Pedersen, J., and Chen, F., "A Trainable Document Summarizer," *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68-73, 1995.

[2] Shafer, G., *A Mathematical Theory of Evidence*, p., Princeton University Press, 1976.

[3] Abracos, J., and Lopez, G. P., "Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles," *Proceedings of Workshop in Intelligent Scalable Summarization*, pp. 51-57, 1997.

[4] Mitra, M., Singhal, A., and Buckley, C., "Automatic Text Summarization by Paragraph Extraction," *Proceedings of Workshop in Intelligent Scalable Text Summarization*, pp. 39-46, 1997.

[5] Jacobs, P. S., and Rau, L. F., "Natural Language

Techniques for Intelligent Information Retrieval," *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 85-99, 1998.

[6] McKeown, K., and Radev, D. R., "Generating Summaries of Multiple News Articles," *Proceedings of Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-82, 1995.

[7] Miller, G., George, A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, J., "Introduction to WordNet: an On-line Lexical Database," *International Journal of Lexicography*, Vol.3, No.4, pp. 235-312, 1990.

[8] Hovy, E., and Lin, C. Y., "Automated Text Summarization in SUMMARIST," *Proceedings of Workshop on Intelligent Scalable Summarization*, pp. 18-24, 1997.

[9] Paice, C. D., and Jones, P. A., "The Identification of Important Concepts in Highly Structured Technical Papers," *Proceedings Of Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 69-78, 1993.

[10] Miike, S., Itoh, E., Ono, K., and Sumita, K., "A Full-Text Retrieval System with a Dynamic Abstract Generation," *Proceedings of Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 152-161, 1994.

[11] Salton, G and McGill, M. J.. *Introduction to Modern Information Retrieval*, p.123, McGraw-Hill, New York, 1983.

[12] Rich, E. and Knight, K., *Artificial Intelligence*, 2nd Ed., p.242, McGraw-Hill, New York, 1991.

[13] Jang, D. H., and Myaeng, S. H., "Development of a Document Summarization System for Effective Information Services," *Proceedings of RIAO 97 Conference*, pp. 101-111, 1997.

[14] 김철완, 장재우, "형태소 네트웍을 이용한 한글 문헌의 자동 키워드 추출", 제 6 회 한글 및 한국어 정보처리 학회, 1994.

부록 A.

테스트 문서[14]를 요약한 결과의 예: 상위 5개 문장

본 논문에서 제안하는 형태소 네트웍은 문법형태소가 어절 내에 또는 어절간에서 특정한 결합 패턴을 보이는 것에 착안하여 이를 이용하여 품사 모호성 해결이 가능함을 보이고 또한 어휘사전의 사용을 최소화하여 분석속도를 높이고 미등록어 추정이 가능함을 보인다.

본 논문은 비록 형태소 분석 방법이 구문, 의미분석 방법에 비해 정확성이 떨어지고 구단위 색인이 어렵다는 문제가 있지만, 한국어 자동 색인에 있어서 형태소 분석 방법을 가장 현실적인 방법으로 판단하였다.

본 논문에서 설정한 4가지 타입의 외부링크는 어절간의 연결을 이용한 저수준의 통사분석이라 말 할 수 있지만 조사 및 어미를 세분화하여 외부링크의 타입을 확장하는 대량의 문헌을 통해 한국어 문장에 나타나는 구문적 특성을 링크로 설정할 수 있다면 보다 정확한 분석이 가능할 것이다.

형태소 네트웍을 이용하여 조사나 어미사전이 링크로 표현될 수 있음을 보였으며 어절간에 나타나는 문법형태소의 연결패턴을 단서로 모호성을 해결을 시도하였다.

이러한 입장에서 본 논문은 형태소분석을 이용한 기존의 자동색인 시스템의 문제점을 해결하기 위해 새로운 접근 방법을 제안한다.

장 동 현

1995년 2월 충남대학교 컴퓨터과학과 학사. 1997년 2월 충남대학교 컴퓨터과학과 석사. 1997년 3월 ~ 현재 충남대학교 컴퓨터과학과 박사과정 재학중. 관심분야는 정보검색, 자연어처리, 전자도서관, 데이타 마이닝

맹 성 현

1983년 미국 캘리포니아 주립대학 학사. 1985년 미국 Southern Methodist University(SMU) 석사. 1987년 미국 Southern Methodist University(SMU) 박사. 1987년 ~ 1988년 미국 Temple University 교수. 1988년 ~ 1994년 미국 Syracuse University 교수. 1994년 ~ 현재 충남대학교 컴퓨터과학과 교수. 관심분야는 정보검색, 자연어처리, 디지털도서관, HCI.