□ 기술해설 □

자동 요약 시스템

충남대학교 장동현*·맹성현**

1. 서 론

자동 요약이란 한 마디로 문서의 내용을 압축하는 것이라고 볼 수 있다. 즉, 본래 문서가 가지고 있는 기본적인 의미를 유지하면서 문서의 길이나 정보의 복잡도 등을 줄이는 작업이다[1~4]. 정보가 많지 않았던 시대에는 도서관 등에서 전문적으로 요약을 하는 사람이 초록이나 요약문을 생성했었으나, 인터넷(internet)과 정보 서비스 기술의 발달로 유통되는 정보의 양이 기하급수적으로 늘어남에 따라 자동 요약의 필요성이 급증하고 있다.

자동 요약은 정보검색 환경에서 그 필요성이 가장 절실하게 나타난다. 예를 들어 인터넷 상에서 검색엔진을 사용하여 질의를 하는 경우 검색된 문서의 수가 10,000건을 넘는 경우가 보통인데, 사용자가 이들을 모두 읽어보면서 적절성을 판단하는 것은 거의 불가능하다고 하겠다. 이때, 여러개의 관련된 문서를 종합하여 요약 정보를 제시하는 기능(문서간 요약)이라든지 수십 쪽의 문서를 몇 줄로 요약하는 기능(문서내 요약)은 정보의 과적재(information overload)를 피하기 위해 필수적이다.

한편 DIALOG와 같은 온라인 정보 서비스 시스템에서는 검색의 결과로 대개 전문(full text)보다는 초록이나 제목 정보를 제시하게 되는데, 이러한 정보검색 시스템 환경에서는 검색 효율(efficiency)을 위해 초록이나 제목 등의 요약 정보를 사용하여 색인을 하게 된다. 이제 까지는 검색 대상이 되는 초록 정보의 생성이 수작업으로 이루어져 왔으나, 향후 생성되는 정보의 양이 증가함에 따라 자동 요약 시스템의 필요성도 증대될 것이라는 것은 쉽게 생각할 수 있다. 즉, 자동요약은 정보검색의 결과 제시 단계에서 뿐만 아니라 검색을 위한 DB의 구축에도 필수적인 기능이다.

자동 요약에 대한 연구는 자동 초록생성(automatic abstracting)으로부터 시작하여 그 역사가 비교적 길지만, 근래에 일반 사용자가 접할 수 있는 정보의 양이 급증하면서 그 중요성이 새롭게 인정되기 시작하였고 새로운 요구사항과 접근 방법으로 연구도 활기를 띠기 시작하였다. 요구사항면에서 보면 하나 이상의 문서로부터 중복된 부분을 제거하고 자연스러운 요약을 생성하는 고난이도 요약으로부터, 하나의 문서에서 중요한 문장 구성 성분을 추출하여 나열함으로써 사용자가 문서의 전반적인 내용을 손쉽게 파악할 수 있게 해주는 비교적 단순한 요약까지 매우 다양하다.

본 논문에서는 이렇게 다양한 자동 요약 기술을 접근 방법에 따라 분류하여 이 연구 분야를 체계적으로 정리하고, 자동요약 시스템을 구축하는데 있어서의 전반적인 어려움을 기술하여 그 특성을 알아본다. 그리고 대표적인 시스템의 사례를 소개하고 앞으로의 개발 방향을 전망해 본다.

2. 요약 시스템의 분류

문서 자동 요약은 정보검색 및 자연어처리 응용 분야의 하나로서 문서의 요약 기법에 대한 접근 방법은 크게 네 가지 형태로 나눌 수

* 학생회원
** 정 회 원

있다. 본 장에서는 이미 개발되었거나 현재 연구되고 있는 요약 시스템을 문장 추출 기반, 문장 이해 기반, 혼합된 형태, 틀(template) 기반의 시스템으로 분류하여 각각의 특성 및 장단점을 기술한다. 이 분류는 사용되는 기술과 접근 방법에 근거를 둔 것으로 하나의 시스템이 두 개 이상의 부류에 속할 수 있다.

2.1 문장 구성 요소 추출 시스템(Passage Extraction System)

문장 구성요소 추출 기반 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 방법이다. 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.

이러한 시스템의 구축에 있어 중요한 이슈는 크게 두 가지로 볼 수 있는데, 어떤 함수를 이용하여 각 문장 구성요소가 문서의 내용을 대변하는 정도를 계산하는가 하는 문제와 어떤 휴리스틱을 사용하여 추출된 문장 구성요소를 정렬하여 요약문의 가독성 및 응집도(coherence)를 높이는가 하는 문제이다. 대부분 연구의 초점이 되어 온 첫번째 문제에 대한 해결 방법으로는 정보검색 기법을 사용하여 문서의 주요 내용을 담고 있는 문장 구성요소를 식별하는 기법과, 학습 데이타와 확률 이론을 사용하여 일종의 분류기(classifier)를 구축함으로써 요약문에 포함될 문장 구성요소와 포함되지 않을 문장 구성요소를 분리하는 기법이 있다.

2.1.1 검색 기반

정보 검색 시스템의 경우 단어의 빈도수에 의거한 통계치를 사용하여 단어가 문서의 내용을 대표하는 정도를 계산한다. 자동 요약에 대한 초창기 연구에서는 주제어의 빈도수가 많은 문장을 위주로 문서의 자동 요약을 시도하였으나[4], 요약문으로서의 가치가 없는 문장에 주제어가 포함되는 경우가 많기 때문에 좋은 결과를 얻지 못했다. 근래에는 문단단위의 검색 기법을 활용하여 문단간의 관계성을 계산한 후 관계성 패턴에 의해 추출될 문단을 결정하는 방법[3, 5]이 제시되었으나 검색기법에 기반을 둔 접근 방법의 한계가 있음을 보였다. 이러한 연구를 통해서 얻은 결과 중의 하나는 주제어보다는 "이 글의 목적은", "요약하면"과 같은 실마리 단어(clue words)나 위치 정보 등이 문서의 대표성을 지니는 문장을 추출하는데 더 중요한 역할을 할 수 있다는 것인데, 이점에 착안한 연구도 다수 존재 한다[1, 2, 6].

2.1.2 분류 기반

분류 기반의 접근 방법은 계산 언어학 분야에서 대량의 코퍼스(corpus)를 사용하는 자연언어처리 문제를 해결하려는 시도와 그 맥락을 같이 하는 도메인(domain) 독립적인 접근 방법이다. 이 방법은 다량의 학습 데이타로부터 요약문에 포함되는 문장의 자질(feature)에 관한 확률 정보를 학습한 후 이를 이용하여 원문의 각 문장이 요약문에 포함될 확률을 계산한다. 문장 구성요소 추출에 의존한 요약 기법으로서의 한계는 있으나 이론적인 기반과 실용성을 겸한 기법이다.

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[1]를 들 수 있고, 이를 확장한 방법으로 문서를 구조화하여 문서의 특정 부위에 속한 문장에 대해서만 확률 값을 계산하는 방법[2]이 있다. 후자의 경우 이론적으로 견고한 방법을 한국어 문서에 적용하여 그 실효성을 보인 연구 결과로서 국어공학적인 관점에서 볼 때 가장 직접적으로 관련있는 연구라 하겠다.

2.2 텍스트 이해 기반 시스템(Systems Based on Text Understanding)

텍스트 이해 기반의 시스템은 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[7]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[8]의 경우 KODIAK이라는 지식 표현 언어를 사용하여 개념 지식을 표현하고 SUMMONS[9] 시스템의 경우 정보추출 시스템에서 사용하는 틀(template)을 개념 표현 방법으로 사용한다. 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[10]과 같은 언어 지식과 자연언어처리 기법을 혼합하여 이용해서 문장이 나타내고 있는 주제를 파악하고자 하는 연구도 진행되고 있다[6].

2.3 혼합 형태

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[6]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.

2.4 틀(template) 기반 시스템

틀 기반의 접근 방법[9, 11]은 요약문에 포함되어야 할 개념을 수작업을 통해 틀로 정의하고 텍스트 분석을 통해 틀을 메꾼 후 요약문을 생성하는 과정을 거친다. 틀은 분야의 전문가가 직접 작성하기고 하고, 학습 데이타로부터 요약문의 기반을 이루는 주제를 학습한 후 각 주제에 해당되는 틀을 만들기도 한다. 텍스트 데이타를 사용해서 틀을 채우는 방법으로는 데이타의 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. 자동 요약의 난이성

본 장에서는 지금까지의 자동 요약에 대한 연구 결과를 통해 나타난 대표적인 문제점들을 살펴본다. 이를 통해 요약 시스템의 구축의 난이성을 보이고 연구 개발에 있어서의 주요 이슈를 부각시키는데 그 목적을 둔다.

첫째, 모든 사람의 요구에 맞는 요약문을 작성할 수 없다. 요약문의 질(quality)은 요약문을 읽는 사람의 요구에 따라 좌우된다. 문서의 대표적인 주제어의 제시만으로도 만족하는 사람, 주제어가 포함된 문장을 요구하는 사람, 그리고 자연스러운 문장으로 이루어진 완전한 요약문을 요구하는 사람 등 다양한 형태의 요약문을 요구한다. 이러한 점을 극복하고자 하는 노력의 일환으로 사용자에게 주제어, 구 그리고 문장, 세 가지 형태의 요약문을 제시하여 평가한 연구가 있다[6].

둘째, 문서에서 다양하게 쓰이고 있는 지시어를 처리하기가 어렵다. 문장 추출 기반의 자동 요약 시스템에서 지시어가 포함된 문장을 요약문장으로 추출하였을 경우, 다른 문장들을 참조하여 지시어가 의미하는 직접적인 단어로 해당 지시어를 대체해야 매끄러운 요약문이 된다. 그러나, 지시어가 문장 또는 단어일 경우, 사람 또는 사물 등을 의미하는 경우 등 다양하게 사용될 수 있어 처리하기가 쉽지 않다. 더

욱이 은유법으로 사용되는 단어의 경우 이를 자동적으로 처리하기란 현재 자연어 처리 기술로는 거의 불가능하다. 또한, 문서 내에서 반복적인 단어의 사용을 피하기 위해서 사용되는 대용어(anaphora)의 처리도 고려해야 한다. 각기 다른 문장에서 같은 의미로 쓰인 이형 동의어를 다르게 처리하면 주제의 분산을 초래함으로써 같은 내용의 문장이 요약문에 중복되어 나타날 수 있는데, 이러한 문제점을 해결하기 위해서 단어 사이의 개념 관계를 나타낸 WordNet이나 Signature를 이용한 연구가 있다[6, 12].

셋째, 요약된 내용을 평가하는 작업은 평가자의 주관성을 배제할 수 없기 때문에 객관적인 평가 결과를 도출하기가 쉽지 않다. 요약 시스템을 평가하기 위해서 두 가지 태스크, 사용자가 요약문을 보고 정확하게 문서 분류(categorization)를 할 수 있는가와 사용자가 제시한 주제를 효과적으로 요약문에 표현하고 있는가로 나누어 요약 시스템을 평가한 연구에서는 문서의 분류를 위해 필요한 주제 제시, 요약문 생성에 걸리는 시간, 요약문의 길이, 사용자가 제시한 주제를 어느 정도 잘 나타내고 있는지를 평가 기준으로 사용하고 있다[13]. Paice[11]는 요약 전문가를 이용한 평가 방법을 시도했으나 공정한 평가를 내릴 수 없어서 요약될 개념이 요약문에 포함되었는지에 대한 확률로 평가를 시도했다. 즉, 시스템의 요약 결과에 대해서 평가자로 하여금 요약해야 될 내용이 어느 정도 포함되었는지를 확률값으로 나타내도록 하여 평가를 하였지만 평가자가 내리는 확률도 불확실한 것이므로 논란의 여지가 있다. Kupiec[1]은 원시 문서의 문장과 사람이 수동으로 요약문을 추출한 문장 사이의 매칭 유형 중에서 원래의 문장과 같거나 약간의 변화만 있는 유형, 그리고 두 개 이상의 문장의 조합으로 매칭되는 유형을 시스템이 추출한 요약문과 비교하는 평가 방법을 시도하였다.

넷째, 요약문을 생성하는데 있어서 최적 길이에 대한 문제는 아직 구체적으로 연구된 바가 없다. 요약문의 길이가 길수록 사용자에게는 많은 정보를 줄 것은 자명한 사실이지만 그만큼 요약문의 가치는 저하된다. 사용자로 하여금 요약문의 길이를 선택하도록 하는 시스템이 있지만 자동 요약 시스템의 원래 목적과는 부합되지 않으므로 근본적인 해법은 아니라고 할 수 있다.

4. 개발 사례

본 장에서는 위에서 언급한 접근 방법을 구체화하기 위해 지금까지 개발된 시스템 중에서 대표적인 네 가지 시스템의 기능 및 사용된 기법을 기술한다.

4.1 통계적 접근 방법

Kupiec[1]은 요약문이 갖는 특성(feature)을 학습 코퍼스로부터 추출한 후, 문서 내의 각 문장에 대하여 요약문의 특성을 갖는 확률을 계산하여 일정한 값 이상이면 요약문에 포함시키는 접근을 시도하였다. 학습 코퍼스로부터 추출한 특성은 다음의 5가지다.

- 문장 길이에 따른 특성(Sentence Length Cut-Off Feature) : 문서 내에서 길이가 짧은 문장(5단어 이하로 구성된 문장)은 요약문에 포함되지 않는 경향을 갖는다는 특성으로 예를 들면 section의 heading이 이에 속한다. 문장의 길이가 정해진 임계값(threshold)보다 크면 특성 값은 "true"이고 그렇지 않으면 "false"가 된다.
- 상용구 특성(Fixed-Phrase Feature) : "this letter", "in conclusion"과 같은 상용구를 갖는 문장이나 section heading의 바로 다음 문장에 "conclusions", "results", "summary", "discussion"과 같은 단어가 포함된 경우 요약문이 될 가능성이 많다는 특성이다.
- 단락 특성(Paragraph Feature) : 문서에서 처음 10개와 마지막 5개의 단락을 대상으로 각 단락 내에서의 위치를 세 가지(앞부분, 세개 이상의 문장으로 이루어진 단락인 경우 중간 부분, 두 개 이상의 문장으로 이루어진 경우 끝부분)로 분류하여 특성 값을 부여한다.
- 주제어 특성(Thematic Word Feature) : 문서의 내용(content)을 대표하는 단어

중에서 빈도수가 많은 단어를 주제어라고 정의하고 각 문장에 주제어가 포함되어 있는지에 대한 특성이다.

- 대문자 특성(Uppercase Word Feature) : 대문자로 이루어진 단어가 요약문에 포함될 확률이 높다는 특성으로 이러한 단어는 문서 내에서의 빈도수가 일정 횟수 이상이어야 하며 약어(예 : F.C, Kg, etc)이어서는 안된다.

이 연구에서는 이상의 5가지 특성을 Bayes 규칙을 이용하여 각 문장이 요약문에 포함될 확률을 계산하였다. k개의 특성 Fj(j=1, …, k)가 정의된 경우 임의의 문장 s가 요약문 S에 포함될 확률은 다음과 같이 계산한다.

본 접근 방법은 영역(domain) 독립적으로

$$P(s\in S \mid F_1, F_2, \cdots F_k)=\frac{P(F_1, F_2, \cdots F_k \mid s\in S)P(s\in S)}{P(F_1, F_2, \cdots F_k)}$$

적용할 수 있는 장점이 있지만 각 특성에 대해서 이진(binary) 값을 사용하기 때문에 정확한 의미의 확률 정보를 이용하지 못한다는 단점이 있다.

4.2 주제 기반 접근 방법

주제 기반 접근 방법은 일반 자연언어처리에서 사용되는 의미분석 과정을 거치지 않고 어휘 분석을 통해 원문의 주제 전개를 파악하여 요약을 생성하는 기법이다. Brazilay와 Elhadad에 의한 연구에서는 WordNet, 품사 태거(tagger), 간단한 파서 등을 이용하여 텍스트 상의 어휘 사슬(lexical chain)을 식별한 후 강력한 사슬이 있는 문장을 선정하여 요약문 생성을 한다[12].

어휘 사슬을 구성하기 위해서는 먼저 문서 내에서 후보 단어(명사, 복합명사)를 선택한 후, WordNet을 이용하여 각 후보 단어와 같은 개념을 갖는 단어가 포함된 사슬을 찾는다. 사슬을 찾으면 단어를 연결하고 실패하면 새로운 사슬을 만든다.

예를 들어, 그림 1은 문장 "Mr. Kenny is the person"을 사슬로 구성한 것을 보여주고 있다. 그림에서 "Person"은 두 가지의 의미를 갖고 있음을 보여 주고 있으며 "individual", "someone"의 의미를 갖고 있는 경우에 "Mr."와 연결이 되고 그렇지 않은 경우는 연결이 되지 않음을 알 수 있다. 이와 마찬가지로 문서 내의 모든 후보 단어에 대해서 이와 같은 과정이 끝나면 개념이 같은 단어가 집합되어 있는 사슬들이 만들어지게 된다. Length가 어휘 사슬에 나타나는 단어의 빈도수이고 Homogeneity는 유일한 단어의 빈도수를 Length로 나눈 값인 경우 각 사슬은 식, Score(Chain) = Length * Homogeneity를 이용해서 구한 점수를 갖게 되며, 다음의 식을 만족하는 사슬에 대해서 이를 대표하는 단어를 포함하고 있는 첫번째 문장이 요약문장으로 선택된다.

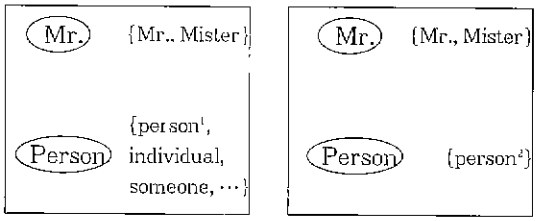

그림 1 어휘 사슬(Lexical Chain)의 예

Score(Chain) > Average(Scores) + 2 * Standared Deviation(Scores)

이 방법은 같은 주제어를 갖는 문장이 요약문에 중복적으로 나타나지 않지만, 문장의 길이가 긴 것일수록 요약문에 포함될 가능성이 많으며, 지시어나 요약문의 길이 조절에 대한 처리를 하지 않았다.

4.3 혼합 기법

Hovy[6]는 문서의 주제를 파악하기 위해서 통계적인 방법을 사용하였고, 내용의 이해 차원에서 문장의 개념을 파악하기 위해 WordNet, Concept Wavefront[14], Concept Signature [15] 등도 사용하였다. 이 연구에서 개발한 요약 시스템은 컴퓨터와 관련된 신문 기사를 대상으로 하고 있으며 3단계 즉, 문서의 주제 파악, 개념 통합, 요약문 생성의 과정으로 구성된다.

- 1단계(Topic Identification) : 학습 데이타로부터 요약문에 포함될 최적의 위치에 대한 확률 값을 학습하는 단계로 최적위치

에 관한 휴리스틱을 정의하여 사용하였다.

- 2단계(Concept Fusion) : 1단계의 확률 값을 이용해서 추출된 문장에 대해서 같은 개념인 경우 하나의 개념이나 상위 레벨의 개념으로 통합하는 단계이다. 예를 들어 문장 "John bought some vegetables, fruit, bread, and milk."을 "John bought some groceries."으로 이해하기 위해서 WordNet과 Signature를 사용한다.
- 3단계(Summary Generation) : 필요에 따라서 세 가지 형태 즉, 1) 주제를 나타내는 단어만으로 요약문을 대신하거나, 2) 명사구나 절로 이루어진 간단한 문장, 3)1, 2단계를 통해 생성되는 완전한 문장으로 구성된 요약문을 제시한다.

이와 같이 본 연구에서 개발된 시스템은 통계적 기법과 문장의 의미를 해석하고자 했으나, 다양한 확률과 지식을 이용하는 방안이 계속적으로 연구되어야 할 부분이다.

4.4 틀(template) 기반 접근 방법

Paice는 고도로 정형화된 문서로부터 의미상의 역할(semantic role)과 패턴(pattern)을 추출한 후, 정보검색 시스템에서 주로 사용하는 빈도수(frequency)와 가중치(weight)를 이용하여 의미상의 역할에 해당하는 적합한 단어를 선택함으로써 요약문을 작성하는 접근 방법을 시도하였다[9]. 이 연구에서 사용한 학습 및 실험용 데이타는 농작물에 관련된 논문으로 그림 2는 학습 데이타로부터 추출한 의미상 역할과 텍스트의 패턴을 보여주고 있다. 요약문 추출은 학습 과정중에서 추출한 텍스트 패턴과 같은 문장을 추출해서 패턴내의 의미상 역할에 해당하는 단어를 선택한다. 틀에 들어갈 단어는 추출한 패턴들의 빈도수와 중요도를 이용하여 선택한다.

의미상 역할 (semantic roles)	역할에 대한 설명
SPECIES	the crop species concerned
CULTIVAR	the cultivars use
HIGH-LEVEL PROPERTY	the property being investigated
PEST	any pest which infests the crop
AGENT	chemical or biological agent applied
INFLUENCE	e.g. drought, cold, grazing
LOCALITY	where the study was performed
TIME	years when the study was conducted
SOIL	description of soil

그림 2 농작물 논문에 대한 의미상 역할

이 방법은 요약과 색인 방법을 잘 결합시켰지만, 적용 분야마다 의미상 역할(semantic role)과 텍스트 패턴을 추출해야 하는 단점이 있다. 그리고 생성할 요약문의 패턴이 정해져 있으므로 요약문의 형식도 획일적이다.

같은 종류의 연구로는 신문기사로부터 같은 사건의 기사를 수집한 후, 이를 대상으로 요약문을 작성하는 연구[9]가 있다. 이 연구에서는 미리 정의된 틀(template)에 신문 기사의 내용을 채운 후, 각각의 틀에 있는 내용이 같은 사건에 대한 내용이면 사건이 발생된 시간과 사건의 내용을 비교 종합하여 하나의 완성된 결과를 제시한다.

5. 결 론

지금까지 자동 요약 시스템의 종류, 자동 요약의 어려움, 개발 사례 등을 중점적으로 알아보았다. 사용자 입장에서는 좀 더 사람이 작성한 요약문과 비슷한 수준의 요약문을 요구하고 있고, 시스템 개발자의 입장에서는 복잡하지 않고 어느 정도의 수준을 유지하는 시스템을 개발하고자 한다. 최근에 개발되고 있는 시스템은 이러한 경향을 반영하여 사용자의 요구를 수용하기 위해 문장 추출 기반의 시스템과 이해 기반 시스템을 혼합하는 형태를 지니고 있다. 특히 상용화 단계의 시스템들은 계산량이 비교적 적고 실용적인 문장 구성요소를 추출하는 기법을 채택하고 있다.

외국의 경우 자동 요약 시스템의 필요성을 인식하고 축적된 자연언어 처리 기술을 활용하여 자동요약에 대한 연구가 꾸준히 진행되어 오고 있다. 본 논문에서 중점적으로 기술한 영어권에서의 연구이외에도 일본[16], 중국[17], 독일[18] 등도 각기 나름대로의 연구를 진행해 오고 있지만, 국내의 경우는 최근에서야 요약의 전단계로 볼 수 있는 정보 검색에 관심을 기울이고 있는 상황이므로 이에 대한 연구는 이제 시작 단계에 있다고 볼 수 있다. 한글과

영어의 경우 언어의 특성이 많이 다르기 때문에 정보 검색 분야에서 영어에 적용되고 있는 여러 가지 기술들이 한글에 그대로 적용될 수 없는 것과 마찬가지로, 자동 요약의 경우도 영어를 비롯한 외국어에 적용된 기술이 한국어의 경우에 성공적으로 적용될지는 미지수이다. 따라서 한편에서는 영어권에서 연구된 기법들을 조심스럽게 적용하면서 동시에 한국어 텍스트가 가지는 구조 및 특성을 파악하여 이에 적합한 모델 및 알고리즘을 개발하는 것이 중요하다.

참고문헌

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

[2] 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.

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

[4] H. P. Edmundson, "New Methods in Automatic Extracting", *Journal of the Association for Computing Machinery*, Vol. 16, No. 2, pp. 264-285, 1969.

[5] M. Mitra, A. Singha and C. Buckley, "Automatic Text Summarization by Paragraph Extraction", *Proc. of a Workshop on Intelligent Scalable Text Summarization*, pp. 39-46, July, 1997.

[6] 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.

[7] 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.

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

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

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

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

[12] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization", *Proc. of a Workshop on Intelligent Scalable Text Summarization*, pp. 10-17, July, 1997.

[13] T. F. Hand, "A Proposal for Task-Based Evaluation of Text Summarization Systems", *Proc. of a Workshop on Intelligent Scalable Text Summarization*, pp. 31-38, July, 1997.

[14] C. Y. Lin, "Topic Identification by Concept Generalization", *In Proc. of the 33rd ACL Conference*, Boston, MA.

[15] C. Y. Lin and E. H. Hovy, "Identifying Topics by Position", *In Proc. of the Applied Natural Language Processing Conference*, Washington, DC, 1997.

[16] Seiji Miike, Etsuo Itoh, Kenji Ono, Kazuo Sumita, "A Full-Text Retrieval System with a Dynamic Abstract Generation Function", *Proc. of 17th ACM-SIGIR Conference*, pp. 152-161, 1994.

[17] Benjamin K. T'sou, et al., "Automated Chinese Full-Text Abstration Based on Rhetorical Structure Analysis", *Proc. of the 1995 International Conference on Computer Processing of Oriental Languages*, pp. 259-266, 1995.

[18] Ines-a. Busch-Lauer, "Abstracts in German Medical Journals : A Linguistic Analysis", *Information Processing & Management*, Vol. 31, No. 5, pp. 769-776, 1995.

장 동 현

1995 충남대학교 전산학과(이학사)
1995~1997 충남대학교 대학원 전산학과(이학석사)
1997~현재 충남대학교 대학원 컴퓨터과학과 박사과정 재학중
관심분야 : 정보검색, 자연어처리, 디지털도서관

맹 성 현

1983 California State University(B.S.)
1985 Southern Methodist University(M.S.)
1987 Southern Methodist University(Ph. D.)
1987~1988 Temple University 교수
1988~1994 Syracuse University 교수
1994~현재 충남대학교 컴퓨터과학과 교수로 재직중
관심분야 : 정보검색, 자연어처리, 인간과 컴퓨터 상호작용, 디지탈도서관