텍스트 이해 모델에 기반한 정보 검색 시스템

송인석, 박혁로
연구개발정보센터

Text Undestanding System for Summarization

In Seok Song, Hyuk Ro Park
Korea Research and Development Information Center
{issong,hrpark}@kordic.re.kr

본 논문에서는 인지적 텍스트 이해 모형을 제시하고 이에 기반한 자동 요약 시스템을 구현하였다. 문서는 정보의 단순한 집합체가 아닌 정형화된 언어 표현 양식으로서 단어의 의미적 정보와 함께 표현 양식, 문장의 구조와 문서의 구성을 통해 정보를 전달한다. 요약 목적의 텍스트 이해 및 분석 과정을 위해 경제 분야 기사 1000건에 대한 수동 요약문을 분석, 이해 모델을 정립하였고, 경제 분야 기사 1000건에 대한 테스트 결과를 토대로 문장간의 관계, 문서의 구조에서 요약 정보 추출에 사용되는 정보를 분석하였다. 본 텍스트 이해 모형은 단어 빈도수에 의존하는 통계적 모델과 비교해 볼 때, 단어 간의 관련성을 찾아내고, 문서구조정보에 기반한 주제문 추출 및 문장간의 관계를 효과적으로 사용함으로서 요약 정보를 생성한다. 그리고 텍스트 이해 과정에서 사용되는 요약 지식과 구조 분석정보의 상관관계를 체계적으로 연결함으로서 자동정보 추출에서 야기되는 내용적 만족도 문제를 보완한다.

1. 서론

전자 출판 및 인터넷을 통한 온라인 문서정보 제공은 기하급수적으로 증가하고 있다. 체계적으로 분류되어 있지 않은 정보소스로부터 원하는 정보를 찾는 것은 매우 힘들며, 검색된 정보의 적합성을 분석을 위해 많은 시간이 요구된다. 이러한 어려움을 해결할 수 있는 한 방편으로서 문서들을 분석 요약한 정보가 필요한데, 텍스트 요약은 문서정보의 주요 내용을 분석하고 이를 사용자가 요구에 적합한 정보 검색과 필터링을 위한 도구로서 필요성이 요구되는 분야이다.

본 논문에서는 이론적 기반으로서 인지적 텍스트 이해 모델에 대해 기술하고, 한국어 문서로부터 요약정보를 생성하는 실험적 자동요약 시스템 일반에 대하여 기술하고자 한다. 인지적 텍스트 모델은 문서를 생성하고 이해하는데 행하여지는 인지적 행위를 모듈화하여 시스템 전체를 구성하고, 각 모듈을 실행하기 위한 지식의 습득과정, 그리고 핵심 정보의 추출 체계를 기반으로 하는 문서 요약 이론을 바탕으로 한국어 문서의 특징을 고려하여 설계 되었다. 따라서 이 모델은 자동 요약 모델 일반적인 관점에서 문서가 이해 요약되는 과정을 일반 언어 지식과 문서 구조정보에 기반하여 기술한다.

본 논문의 구성은, 2 장에서 문제 정의 및 관련 연구에 대하여 간략하게 기술하고 3 장에서는 모델의 구성과 구현에 필요한 지식의 습득과정 및 문서 정보의 유형에 대하여 기술한다. 4 장에서는 실험 데이터의 분석을 통한 요약 결과 평가와 평가기준에 대하여 논의하고, 5 장에서는 결론 및 문제에 대하여 논의한다.

2. 문제 정의 및 관련연구

텍스트는 지식정보를 언어를 도구로 일정한 형태의 구조로 표

현된 양식이다. 텍스트를 생성하는 사람은 주어진 언어의 표현 양식에 따라 문장을 기술하고, 문장을 일정한 패턴에 의해 문서 안에 배열하며, 이는 표현된 문장의 구조, 어휘, 그리고 이를 구성 단위로 하는 문서의 구성에서, 분석 될 수 있다. 텍스트 요약은 역시 위의 정보를 이용하는 텍스트 이해를 병행함으로서 문서정보를 재구성, 요약정보를 추출하는 과정이다. 이 과정은 문서의 축약, 색인, 분류를 골자로 하는 인지적 행위로 각각 이해될 수 있으며 인지적 모델의 핵심은 이 과정과 각 단계를 기술하는데 있다. 문서 요약 생성에 관한 연구들을 살펴 보면 목적과 접근방식에 따라 통계적 확률적 정보에 기반한 방법과 문서내용 관련지식에 기반한 방법으로 나눌 수가 있다.

♦통계적 방법 – 정보 검색분야에서 색인어를 추출하는데 사용되는 기법으로 문서에서 나타난 단어의 빈도수를 측정, 문서를 대표하는 단어의 집합을 설정, 이 집합에 속하는 단어들이 나타나는 문장들 중 상위순위의 문장을 추출하는 하는 방법이다. 문서의 표현양식이나 주제 분야에 구애되지 않으며 항상 일정 수준의 만족도를 유지할 수 있는 방식이다. 문서내에서 단어의 중요도와 출현 빈도수가 직접적인 상관관계가 있는 것은 분명하지만, 출현빈도수와 무관하게 결정되는 문서 내의 단어간의 의미적 관련성 및 문장간의 관계를 처리하지 못함으로 인해 의미적으로 연결되지 못하는 문장들을 요약 결과로 추출하기 쉽다.

♦ 지식기반 방법 – 정해진 주제분야에서 추출 요약정보의 생성 패턴을 설정, 문서로부터 필요한 정보를 추출하는 방식으로 주어진 분야에 종속되어 있고 일반적으로 문장의 의미구조를 생성한 후 이를 기반하여 정보를 처리함으로 방대한 양의 전문지식 기반의 구축을 전제로하고 있어 제한성을 가지고 있다.

♦ 수사구조에 기반한 방법 – 통계적 문서내의 문장들을 수사어구 및 수사어구가 설정하는 수사관계에 기반하여 요약문을 생성하는 방법이다. 문서의 표현통계적 방법의 단점을 보완하여 주지만 문서의 표현 양식이 단순한 신문기사등의 문서 형태에서는 요약정보를 추출이 제한된다.

이러한 접근 방법들이 제한성을 가지는 것은 응용중심의 문제 접근과 텍스트 요약을 복합적인 인지적 행위로 정의하지 않은데 기인하고 있다. 따라서 인지적 텍스트 이해 모델에 기반한 요약 시스템은 처리 기법상의 새로운 제안이 아닌 보다 근본적인 문제 접근 방식으로서 2 장 서두에 정의한 바 대로 각 인지적 행위의 수행과정에서 요구되는 지식의 유형을 정의하여 모델을 정립하는데 그 목적을 두고 있다.

문서가 가지고 있는 정보 중에서 요약정보생성을 위한 일반적으로 고려되어야 할 정보의 요소는 다음과 같다(Paice & Johnes[5])

♦ 단어의 출현빈도수와 문서 내의 분포도(위치정보)

♦ 문서 및 문단 내에서의 문장의 위치

♦ 문장 내의 주제 색인어 및 포함여부

♦ 문서내의 특수 수식어를 포함한 문장(영어 예: it is important, Urgently,..)

3 장에서는 이해 모델의 구성과 인지적 처리과정에서 필요한 위의 요소들이외에 요약문 추출에 필요한 정보의 형태를 기술한다.

3.인지적 텍스트 이해모델 한국어 문서 요약 시스템

문서정보 처리를 위한 텍스트 이해 모델 구성은 Endres-Niggemeyer[3]의 개념적 요약 모델에 기반하였다.

인지적 개념 모델에서는 문서 정보를 표면과 의미정보, 문서의 구조와 주제로 구분하고 인식과 적합성 검사와 문장 생성 그리고 요약문서 생성부분으로 나누었고 각각 정보를 교환하는 인지적 행위의 타입으로 연결하고 있다. 각 행위의 단위를 과제를 수행하는 기능들을 하나의 독립된 모듈로서 정의하고 기능을 수행하는데 필요한 정보를 서로 주고 받을 수 있도록 연결되어 있다.

각 인지적 행위는 독립된 모듈로서 그 성격에 따라 각각 다음과 같이 하나의 에이전트로 분류 해석 될수 있다..

♦정보습득 에이전트들(explore, top-level, by-form, unit, browse and read-form)
♦적합성 평가 에이전트들(hold, relevant-hit, relevant-call, relevant-

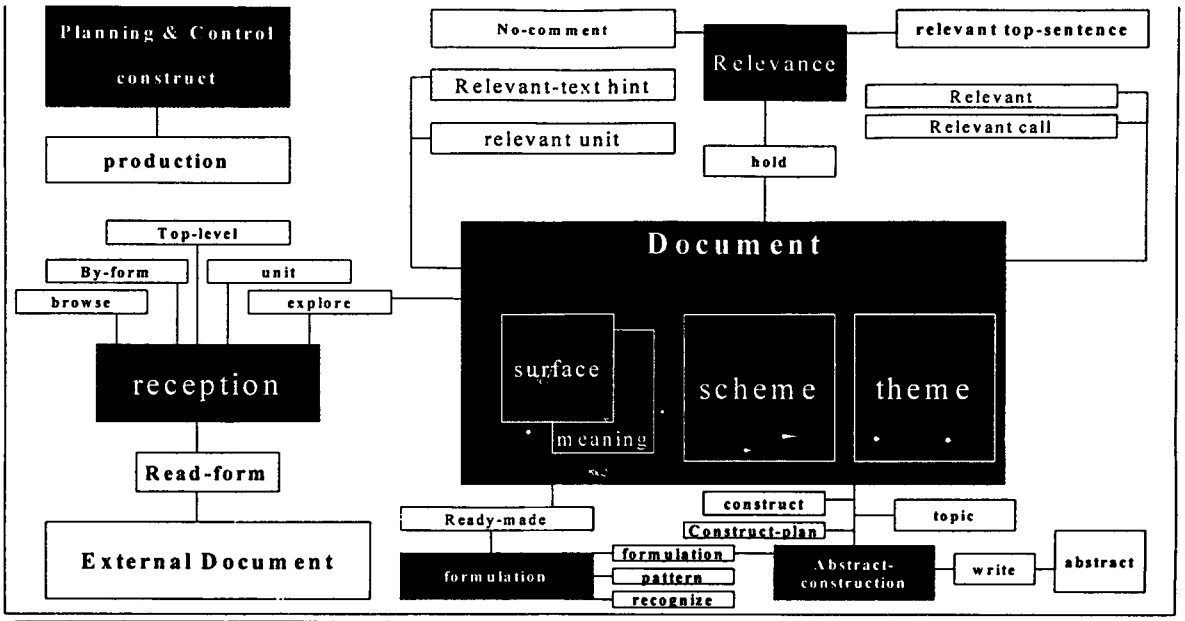

그림 1. 인지적 처리과정과 필요정보요소

topic-sentence, no-comment and relevant)

◆ 생성 에이전트들(construct-plan, topic,formulation, reday-made, reorganise, pattern and write)

각 모듈들이 정보를 교환하고 임무를 수행하는 과정을 위의 그림에 보여진 바에 기초하여 서술하자면,

- **Explore** 는 지식의 습득처리를 담당하는 모듈로 **Reception** 과 관련된 모듈들인 **By-form**(위치정보를 분석하는 모듈)또는 **Unit** (문단관련 정보 수집 모듈)등을 각각 활성화 시켜서 넘어 오는 결과를 **document** 의 **surface** 에 저장한다.라고 기술할 수 있다.

본 논문에서 제안하는 인지적 텍스트 이해 모델 기반 요약 과정은 [그림 2]의 시스템 구성도에 제시된 바 같이 크게 세 단계로 구분되며, 각 단계별로 위의 개념 모델이 제시하는 지식습득, 적합성 평가, 그리고 생성과정을 실행한다. 첫번째 단계에서 자동 요약 시스템은 문서로부터 품사정보와 문장 구조, 그리고 문서구조정보를 습득한다. 그리고 두번째 단계에서는 첫번째 단계에서 수집된 정보를 분석하여 요약정보 추출에 필요한 문장 및 핵심어를 분리하여 이들의 중요도를 계산하여 문서의 주제문 및 핵심문장의 순위를 설정, 요약문을 생성하기 위한 문장을 결정한다. 세번째 단계에서는 이들 주요 문장으로부터 요약문을 생성한다

첫단계에서 이루어지는 처리 모듈은 두번째 인지적 처리과정에서 지식 습득행위의 하부 실행단계이다. 이는 대상문서의 성격 및 문서 구조 형태의 분류를 위한 기초적인 정보를 가공하는 전처리 단계로서, 각 어절단위의 형태소 분석과 각 형태소의 태깅과정을 통하여 단어의 정보를 얻고, 문장 및 문서의 표면 구조 정보를 각각 처리한다. 이 과정에서 가장 핵심이 되는 것은 각 처리 단계에서 문장의 중요도와 적합성 판단에 적용되는 문서정보의 유형 분석이다. 특정 문서 정보 유형이 단어 및 문장의 문서 주제와 적합성의 관련성 및 중요도를 결정하는 판단 기준으로 선택 되려면 이 정보 유형을 이용한 요약문 추출이 내용적 만족도를 보일 수 있어야 한다. 이런 문서 정보 유형을 추출과 일반화를 위해서는 문서에 자주 또는 공통적으로 나타나는 현상을 분석하는 것이 필요한데, 이를 위해 우선적으로 제목 검색으로 추출된 경제 주제 분야의 기사 데이터 50건의 수동 요약문서를 작성 기사작성 전문가에게 작성하도록 하였다. 실험대상으로 선정된 문서는 인터넷 상에서 다양하게 제공되고 있는 경제 분야의 다양한 주제를 다룬 신문 기사 데이터를 선정하였다. 실험대상 데이터를 선정하는데 있어서 기준으로는 요약의 결과분석에 적절한 문서의 크기와 내용적 난이도, 그리고 데이터 수집의 용이성을 삼았다. 또한 신문 기사 데이터는 제목을 가지고 있으므로 문서의 내용 파악이 쉽고 정확한 결과분석이 가능하다는 장점을 가지고 있다. 방법은 실제 수동 요약과정에서 이루어지는 보편적 인지적 처리과정을 분석하기 위해서 제목을 제거한 기사 데이터를 제공하여 수동 요약문 작성토록 하였다, 아울러 문서 주제어 및 주제 문장들도 함께 추출토록 하였고, 요약문서의 크기도 원 문서 크기 비례 약 25%의 크기

로 일관되게 작성토록 하였다. 또한 통계적 방법 기반 모델과의 상대 비교를 위하여 해당 문서 50건과 주제별 검색된 원문데이터 300건에 대하여 실험한 결과를 비교 평가하였다. 후자의 평가방법은 검색된 기사의 제목 없이 수동 작성된 실험 데이터를 분석한 결과 2장에서 언급한 요약 정보 요소을 고려하여 다음과 같은 결론 및 유추 해석을 할 수 있었다.

어휘정보:

◆ 어휘간 관련성: 주제어 추출의 분석 결과, 동일 문장에서 같이 출현한 빈도수가 많은 단어가 문서의 주제를 나타내는 색인어로 함께 추출되었으며 의미적으로 관련이 있는 것으로 분석되었다.

◆ 어휘 기능: 추출된 주제어의 문장 내의 기능을 조사한 결과 술어부를 제외한, 주어, 목적어, 관형절이나 보어에서 다양하게 분리 추출되었으나 술어부에서 추출된 경우는 상대적으로 현저히 적었다.

◆ 복합 명사의 의미구조: 주제어의 경우 대부분 단순명사 보다는 복합 명사로서 전체 50건 중에서 그 의미가 구조가 문장 성격을 띤 형태(예; 금융시장개방, 금융실명제실시여파)를 포함한 경우 주제문 추출에서 항상 일정한 출현율을 보였다.

문장관계

◆ 주제어를 포함한 문장이 주제 적합성 평가 상위 순위로서 주제문으로 추출되었을 경우 이와 대립이나 병렬관계를 나타내는 수사로 연결된 문장의 경우 주제문으로 추출 되었다.

문서 비종속 키워드:

◆문서 내에서 중요도가 명확하게 나타나는 표현을 포함한 문장(예; 결론적으로..., 첫째, 둘째 ..., 확실한 것은 등)이 주제어 포함 여부에 종속되지 않고 추출되었다. 무서의 유형이나 성격에 관계 없이 항상 높은 중요도를 예상할 수 있는 키워드로 문서 비종속 주제어로 분류될 수 있다.

문서구조:

◆ 추출된 주제문의 분포의 분석결과, 문서의 앞부분이나 뒷부분에 위치 하였으며, 대부분 앞부분에 위치 하였다.

이와 같은 분석결과를 토대로 적합성과 중요도 판단의 기준을 갖춘 텍스트 이해 모델이 요약정보 추출과정에서 실행하는 행위와 각 지식 기반의 연결하여 요약정보를 추출하는 방법을 제안한다.

STEP1. 관찰 분석 - 문서의 구조적 지식과 통계적 정보 습득을 위하여. 문단의 수, 문장의 위치, 어절의 길이와 위치, 총 단어의 갯수 등 을 형태소 분석과 품사 태깅을 위해 입력된 문서를 읽어 들일 때 동시에 수행한다..

STEP2 분류 – 형태소 분석과 품사태깅 등 전처리의 결과를 토대로 문서의 단순 구문정보에 따라, 단순명사 및 고유명사 분류와 문장내의 기능에 따라 주어절, 목적어절, 관형어절, 목적어절, 보어절, 술어절로 분류한다. 술어부에서 추출된 명사와 그렇지 않은 경우의 명사 가중치를 차별화 하고 기능의 빈도수를 계산한다.

STEP3 평가 – 관찰분석과 분류 결과를 토대로 문서의 주제를 대표하는 단어들을 추출하고 단어와 문장의 중요도를 결정, 그 순위를 결정한다. 모든 명사의 문서 내 출현 빈도수의 합의 값에 기초한 문장 중요도의 기초 값을 설정하고 문장의 위치정보를 계산하여 기초 값을 보정한다. 중요성를 나타내는 단어를 포함한 문자의 값을 보정한다.

STEP4 추출 – 평가의 결과를 토대로 요약문 생성을 위해 적합한 문장을 선별 수집하고 결합 가능 여부를 분석한다. 주제어가 포함된 술어부를 제외한 어구의 길이를 평가에 반영하여 계산한다.(예:복합명사의 경우)

. **STEP5** 생성 – 주제문과 대립, 병렬관계로 연결된 문장을 하나의 주제문으로 결합한다. 결합된 문장들로부터 가중치에 따라 요약문을 생성한다.

관찰분석은 개념모델에서 인식(reception)부분을, 분류 및 평가는 적합성 분석(relevance)부분을, 수집 삭제 결합부분은 문장과 요약문서 생성부분을 구현하였다. 위의 모델을 토대로 각각의 단계 별로 요구되는 정보의 타입과 습득 방법을 실제 데이터에 실험을 통하여서 분석한다.

4. 실험

실험의 결과 평가는 수동 요약문을 기준으로 생성 요약문의 주제문과의 문장번호 일치수로 정하고, 자동 요약문의 경우 요약문에 제목이 주제문으로 포함 여부를 참고하였다. 문장의 가중치의 보정 옵션 중에 수사구조정보와 문서중립적 주제어 포함 문장을 제외한 다른 정보의 가중치 값, 술어부 추출 명사와 그 외의 어구에서 추출된 명사의 차별화 값, 문서 위치정보에 따른 가중치 계산을 위한 보정 포물선값등은 실험결과에 맞도록 변수로 지정, 실행 옵션에서 변경할 수 있도록 하였다.

♦어휘단위 가중치
주어 가중치 = 1.0,
목적어 가중치 = 1.0
관형절에 포함된 명사의 가중치 = 1.0
보어절에 포함된 명사의 가중치 = 1.0
단어길이 보정계수 = 1.0

♦문장위치 보정계수

$Y = C(x-0.6)^{*}-1$

포물선 변이값 = 0.6
포물선의 2 차상수값 = 4
문장길이 보정계수 = 1.0
X = 문장의 위치

일반 단순 명사 의 경우
단어의 가중치(W)=(문서내 빈도수*기능어 가중치*단어 길이값)

복합 명사의 경우,
단위 어절로 분리된 경우;
가중치(W) = (각 단위명사의 가중치의 합*단위명사중 2 개 이상이 가 한문장에 동시에 나온 빈도수)

문장의 가중치(W)=(문장내 단어의 빈도수의 합*)문장위치 보정계수

이 결과, 수동 요약문서 비교 25%로 자동요약한 문서의 주제문 포함율이 평균 40%였고, 제목을 포함한 경우가 전체문서 50 건 중 45 건이었다.
반면에 단순이 단어의 출현 빈도수만 고려하여 요약문을 추출하는 하위 버전에서는 주제문 포함률 25%에 제목을 포함한 경우가 전체 50 건중 38 건이었다.
제목의 재현률에 비해 상대적으로 낮은 주제문 재현률은 단어의 출현 빈도수와 문서의 구조정보에 주로 의존한 적합성 평가 기준에 기인한다. 이는 전체적으로 요약이 문서의 크기의 압축에 있어 균형적으로 이루어지지 못하고 주제문과 관련 문장들의 의미적 연결을 보존하지 못하기 때문으로 판단된다.

5. 결론

본 논문에서는 인지적 텍스트 이해 모델에 기반한 정보 추출 방법을 제안하고 실험하였다. 기존의 응용 중심의 시스템들과 비교하여 볼 때 모델의 개념을 정립하고 통계적 적합성 판단 기준이 되는 언어적 정보 및 문서 구조 정보를 문서로부터 습득함으로서 요약문의 내용적 만족도와 요약문의 구조적 완성도를 향상 시켰다.
앞으로 연구되어져야 할 과제로는 생성된 요약문으로부터 관련 색인어를 추출하여 자동적으로 하나의 클러스터 형태의 관련어 휘군을 생성하고 이를 다시 학습하여 요약 정보 추출에 적용, 단어간의 관련성과 문장간의 관련성을 고려할 수 있는 연구를 들 수 있겠다.

참고문헌

[1] Baker, K and J. -F. Delannoy, S. Matiwin and S. Szapakowisz, "Preliminary Validationof text summarization algorithm," University of Ottawa, Department of Computer Science, TR-96-04,1996

[2] Copeck,T.S. Delisle and S. Spakowiccz, Parsing and Case Analysis in Tanka", 15th Intl Conf on Computational Linguistics COLING 92, Nantes, 1008-1012.

[3] Endres-Niggemeyer,B, Hobbs, J., & Sparck Johnes, K(Eds) Summarizing Text for Intelligent Communication, Dagstuhl Seminar Report79, 13.12.-17.12.93(9350)

[4] Norris, J. Extracting the Essence from Text: a Computational View, Proc. of Intl.on lexically Driven Information Extraction, Frascati,

Italy, 63-80

[5] Paice, C.D and P.A. Johnes, "A Select and Generate' Approach to Automatic Abstractng," in Proc 14th Information Retrieval Colloquium, Lancaster 1992, Series 'Workshop in Computing' T. McEnery, C. Paice (ed). Springer Verlag, 1993, 114-154

[6] Salton, G., J.A.,C.Buckley and A.Shinghai, "Automatic Analysis, Theme Generation, and Summarization of Machine Readable Text," Science, Vol.264,1994,1412-1426