공학 연구논문집 제29권 2호 pp. 351~363, 울산대학교.
Journal of Engineering Research Vol. 29, No. 2, pp. 351~363, 1998
University of Ulsan.

언어학적인 방법론을 취하는 자동 문서 요약에 대한 연구*

배 재 학
컴퓨터 및 정보통신공학부

<요 약>

본 논문에서는, 현실적으로 사용 가능한 언어학적 도구와 지식 - 견실한 구문 분석기, 사서적 상황 지식베이스, 수사 관계, 담화 표지, 그리고 구사슬 등 - 을 적절히 활용하여, 원문의 주제 연접 관계를 밝히는 중층적 원문 이해를 시도해 보았다. 그리고 이러한 중층적 원문 이해를 바탕으로, 주어진 원문에 대한 요약문을 생성해내는 실용적인 자동 요약 방법론 개발도 모색해 보았다.

Another Investigation of Automatic Text Summarization: A Linguistic Approach

Jae-Hak Joachim Bae
School of Computer Engineering and Information Technology

<Abstract>

In this paper, the problem of mid-depth text understanding is tackled. With practical linguistic tools and knowledge - a robust parser, a lexico-situational knowledge base, rhetorical relations, discourse markers, and phrasal chains - an attempt is made to identify the topic connection of a given text. Based on the mid-depth text understanding, it is investigated to make another approach to automatic text summarization.

* 이 논문은 1998년 울산대학교의 연구비에 의하여 연구되었음.

1. 서 론

인터넷(Internet)의 상용화로 가속화된 현대의 정보화 사회에서, 개인이 수집하고 처리할 수 있는 정보량이 엄청나게 많아졌고 또한 그러한 정보는 시시각각으로 변하고 있다. 더욱이 이러한 막대한 정보량은, 개인의 정보 처리 능력을 훨씬 초과하고 있다. 스스로 원하는 주제에 대한 최신의 정보를 정확하고 빨리 그리고 요약된 형태로 얻고자 하는 바램은 예로부터 있어 왔지만 정보화 사회가 성숙되어 가고 있는 현재 더욱더 절실하다.

요약에는 두 가지 종류가 있다[36]. 통상 abstract라고 일컫는 묘사적인 요약(descriptive summary)과 summary라고 부르는 교훈적인 요약(informative summary)이 그것들이다. 묘사적인 요약이란, 원문 전체 내용을 간결히 정리하여 기술해 놓은 것을 말한다. 교훈적인 요약이란, 전체 내용에 대한 간결한 정리뿐만 아니라, 가능하다면 원문 내용에 나타난 중요 사실에 대한 작가의 판단 및 견해를 아울러 포함시킨 것을 말한다.

요약 작성의 자동화에 대한 종래의 연구는, 독자 고유의 정보 획득 욕구를 간과하고, 단지 주어진 원문의 중요 내용을, 간결하고 조리 있게 나타내는데 치중하였다. 즉, 종래의 연구는 원문지향의 묘사적인 요약을 추구했다고 말할 수 있다. 그런데 바람직한 요약이란, 원문지향의 요약이 아니라, 독자 고유의 정보 획득 욕구를 충분히 고려한, 독자지향의 교훈적인 요약[1]이어야 함은 두말할 나위가 없다. 다시 말해, 원문 요약 문제에 있어서는, 독자의 입장에서 본 요약 내용의 필요충분성에 대한 고려가 요약 내용의 압축성에 대한 고려에 못지 않게 중요하다는 것이다. 더욱이 요약 내용의 필요충분성은 요약문의 품질과 직결되어 있다는 점에서, 요약 내용의 압축성에 대한 고려에 선행해야 할 사항이다.

그러나 현재의 자연어 이해 이론과 기술 수준 그리고 가용 지식 베이스의 완전성 수준 등의 각도에서 검토해볼 때, 원문지향의 묘사적인 요약조차, 적용 분야를 한정시킨다 할지라도, 성공적이지가 않다. 이것의 근본적인 이유는 바로 자연어 이해 문제가 충분히 해결되지 않았다는데 있다. 즉, 궁극적으로 원문 요약은 원문에 대한 깊은 이해를 바탕으로 이루어져야 하는데, 이해에 필요한 이론, 기술, 그리고 지식 등에 연관된 현실적인 제약이 그것을 어렵게 만들기 때문이다. 이에 본 논문에서는, 독자 지향의 교훈적인 요약 방법론의 구현을 장기적인 최종 목표로 삼아, 현시점에서 실현 가능한 새로운 자동 문서 요약 방법론을 모색하고자 한다.

지금까지의 요약 방법론인 원문 지향의 묘사적 요약 방법론은, 크게 보아, 원문 표현(source text representation) 방식과 요약 과정(summarization step)의 각도에서 각각 분류할 수 있다[12, 13, 27]. 먼저, 요약할 원문 표현 방식에 따라 표층적 표현(shallow representation) 방법론과 심층적 표현(deep representation) 방법론으로 나눌 수 있다. 이는 원문의 내용을 표현하는데 있어 이해 과정의 수반 여부에 따라 정해진다.

표층적 표현 방법론에서는 요약할 원문 자체가 원문의 표현이다. 중요 내용 또는 핵심 문장 선별은 단어나 구에 대한 언어학적인 지식을 활용하여 수행한다. 사용하는 대표적인 언어학적인 지식과 단서를 열거하면 다음과 같다: 핵심어(key word)[13], 단서구(cue phrase)[8, 34], 어휘 사슬(lexical chain)[2, 23], 수사 관계(rhetorical relation)[18, 26, 39], 그리고 위치 표준(locational criteria)[17] 등이다.

한편, 심층적 표현 방법론에서는 요약할 원문을, 요약 시스템에 내재하는 지식을 구동시켜 이해한 후, 원문의 내용을 형식화한(formal) 지식 표현 방법으로 나타낸다. 동원하는 지식 표현

방법에 따라 각 방식을 분류해 보면 다음과 같이 된다: 각본(script) 또는 골조(frame)[7, 37, 40], 줄거리 단위(plot unit)[1, 15], 이야기 혹은 본문 문법(story/text grammar)[28, 33], 계층적 적용 영역에 기반한 구조(hierarchical domain-based structure)[29], 그리고 원문의 주제 구조와 적용 영역에 대한 지식(text thematic structure and domain knowledge)[9] 등이다.

다음은, 요약 과정으로 원문 지향의 묘사적 요약 방법론들을 분류한다. 여기에서는 요약된 내용을 추출해 내는 과정에서, 각 요약 방법론이 핵심적으로 활용하는 기법이나 기술 또는 개념에 따라 분류하였다. 따라서, 요약할 원문 표현 방식에 따라 분류한 앞서 본 결과와 몇몇은 중복하고 있다: 이야기 혹은 본문 문법(story/text grammar)[28, 33], 수사 관계(rhetorical relation)[18, 26, 39], 어휘 사슬(lexical chain)[2], 적용 분야 순응식 선별(domain-guided selection)[19], 획득할 정보에 대한 명세(specification of the information sought)[5, 7], 특질어 통계(signature word statistics)[4], 그리고 비언어적 입력 자료(non-linguistic data input)에 대한 처리[3, 20, 21] 등이다.

이상적으로 말해서, 원문 요약은 원문에 대한 깊은 이해를 바탕으로 이루어져야 한다. 그러나 현재까지 개발된 자연어 이해에 필요한 이론과 기술 그리고 축적된 지식 베이스 등이 불충분하기 때문에 그것이 용이하지가 않다. 이러한 불충분성은 단시간 내에 해결될 성질의 문제가 아니다. 따라서 현 시점에서 판단하건대, 심층적 표현 방식을 취하는 원문 지향적인 묘사적 요약 방법론은, 그것의 적용 분야를 한정시킨다 할지라도, 그다지 실용성이 없다고 하겠다.

결국, 차후의 연구에서는 표층적 원문 처리 기법(shallow text processing technique)을 동원한 요약 방법론이 보다 현실적이라 생각한다[12, 13, 27]. 즉, 원문에 나타나 있는 언어학적인 단서(cue)를 충분히 활용하는, 표층적 원문 이해(shallow text understanding)에 기반한 방법론이 그 대안이 될 것이다. 구체적으로는, 견실한 구문 분석기(robust parser) 및 유의어 사전(thesaurus)의 사용, 그리고 수사 관계(rhetorical relation)와 어휘 사슬(lexical chain)을 적절히 활용하여 주제(topic) 및 주제 연접(topic connection) 상황을 밝혀내기에 요약 방법론 개발의 초점이 맞추어져야 한다.

이에 부응한 최근 연구[2, 18]에서는 수사 관계(rhetorical relation)와 어휘 사슬(lexical chain)을 각각 활용하였다. 수사 관계에 기초한 요약 방법론은, 원문에 나타나 있는 담화 표지(discourse marker: thus, after all, that is to say, etc.)와 수사 관계의 일대일 대응을 적극적으로 활용하는 요약 방법론이다. 원문의 모든 문장에 담화 표지가 있을 경우, 이 방법론의 유용성이 극대화된다. 그러나 대부분의 원문은 그렇지가 못하고, 더욱이 담화 표지가 전혀 없는 극단적인 경우도 있을 수 있다. 따라서, 담화 표지에 전적으로 의존하는 방법론의 유용성은, 문장당 담화 표지 개수 및 담화 표지의 수사 관계와의 일대일 대응성에 따라 제한된다.

한편, 어휘 사슬에 기초한 요약 방법론은, 원문에 나타나는 명사 또는 복합 명사들의 의미적인 연관성을 나타내는 사슬을 활용하는 방법론이다. 최선의 경우, 이 사슬은 원문의 구조나 논의되는 내용을 대표할 개념들을 나타낸다. 그러나 일정한 서식을 갖춘 원문의 경우를 제외하고는, 원문의 구조로 원문 내의 각 문장의 상대적인 중요도를 판단하기는 불가능하다. 즉, 원문 안에서 한 문장의 상대적인 중요도는 다른 문장과의 관계(예: 인과 관계)로 결정되는데, 명사류로 이루어지는 사슬로는 그것을 도출하기가 불가능하다는 것이다. 따라서 명사류의 어휘 사슬에 기초한 기존 요약 방법론은, 원문 내용에 나타난 중요 사실을 추출해내는데 필요한 중요 문장 선별 기능에 그 취약점이 있다.

그럼에도 불구하고, 수사 관계와 어휘 사슬은 표층적 원문 이해(shallow text understanding)에 있어 유용한 도구임에 틀림이 없다. 따라서 수사 관계의 경우, 언어학자들의 담화 표지에 대한 최근까지의 연구 결과들을 정리하여 표층적 원문 이해로의 응용 가능성을 재검토해보아야 할 필요가 있다. 그리고 어휘 사슬의 경우는, 주제를 표현하는 명사류뿐만 아니라, 주제 전개에 개입하는 동사나 형용사류를 포섭하는 어휘 사슬이어야 할 것이다.

수사 관계 및 어휘 사슬과 더불어, 원문에 나타나 있는 언어학적인 단서를 충분히 활용하는 표층적 원문 이해에 유용한 또다른 언어학적 도구가 바로 견실한 구문 분석기 및 유의어 사전이다. 견실한 구문 분석기는 한 문장 안에서 표층적 이해에 필요한 단어나 구를 선별하는데 사용된다. 유의어 사전은 명사류 어휘 사슬을 형성하는데 사용될 뿐 아니라, 적절한 의미 처리 과정을 통해 이를 확장시켜서, 동사나 형용사류를 포섭하는 어휘 사슬을 형성하는데 이용할 수 있다.

이에 본 논문에서는, 견실한 구문 분석기, 확장된 유의어 사전, 수사 관계와 어휘 사슬 등의 언어학적 도구와 지식을 적절히 활용하여 원문의 주제 연접 관계를 밝히는 중층적 원문 이해(mid-depth text understanding)를 시도한다. 이해할 원문의 장르는 설화(narrative)이다. 설화문(narrative text)은, 그것의 이해에 대한 인지과학적 연구가 심도 있게 진행된 원문의 한 종류이다. 원문의 장르를 설화로 제한함으로써, 중층적 원문 이해 방법론의 타당성과 유용성 검증이 용이할 것이다. 그후, 이러한 중층적 원문 이해를 바탕으로, 원문에 대한 요약문을 생성해내는 실용적인 자동 요약 방법론 개발을 모색해 보고자 한다.

2. 구 사 슬

어휘 사슬(lexical chain)[23]로써 어휘 응집(lexical cohesion)[10] 상황을 표현할 수 있다. 어휘 응집은 단어 사이에 존재하는 의미적 연관 관계에 연유하는 단어간 결속이다. 어휘 응집은 인접한 두 단어 사이뿐만 아니라, 동일 주제를 다루는 문장군 안에서 광범위하게 연결되어 나타난다. 이것을 어휘 사슬이라고 한다. 이상적으로 어휘 사슬이 형성되었을 경우, 이 사슬은 원문의 구조나 논의되는 내용을 대표할 개념들을 나타낸다[23, 25].

그러나 일정한 서식을 갖춘 원문의 경우를 제외하고는, 원문의 구조로 원문 내의 각 문장의 상대적인 중요도를 판단하기는 불가능하다. 즉, 원문 안에서 한 문장의 상대적인 중요도는, 인과, 가능, 논리, 수사, 집합, 그리고 공간 등과 같은 다른 문장과의 일관성(coherence) 관계로 파악할 수 있는데, 어휘 사슬로는 그것을 도출하기가 불가능하다는 것이다.

한편, 요약 과정은 단순화 과정의 일종으로, 세부 내용을 보면, 중요 문장 선별 과정과 추상화 과정으로 이루어져 있다. 선별 과정이란, 주어진 원문에서 중요 문장을 발췌해 내는 과정을 말한다. 추상화 과정은, 발췌한 내용을 일반화시키는 과정을 말한다. 추상화란, 예를 들어, 어떤 사람이 텔레비전, 세탁기, 초음파 가습기, 전축, 그리고 전기 밥솥 등을 구입하였다'는 것을, 어떤 사람이 가전 제품들을 구입하였다로 바꾸어 표현하는 것을 말한다.

이렇듯이, 중요 문장 선별이란 요약 과정의 핵심 기능 중의 하나이다. 따라서 중요 문장 선별이 핵심 기능으로 필요한 요약 문제에 있어서, 어휘 사슬의 직접적인 기여를 기대할 수는 없다. 다만, 요약 과정에서 원문의 구조나 논의되는 내용을 대표할 개념을 추정하는

데 어휘 사슬을 적극적으로 활용할 수는 있겠다.

이와 더불어, 현재까지 알려진 바로는, 단어 이상의 수준에서 어휘 사슬을 계산하고 이를 자동 요약에 활용하는 시도는 없다. 어휘 사슬을 이용하여, 원문 안의 한 문장이 다른 문장과 어떤 일관성(coherence) 관계를 가지고 있는가를 개략적으로나마 파악할 수 있으려면, 적어도 구(phrase) 수준의 어휘 사슬을 생성해야 한다. 이러한 어휘 사슬을 구사슬(phrasal chain)이라고 부르기로 한다.

구사슬은 선행 연구[2, 23]에서 사슬 형성 후보 단어로서 자격이 없었던 고빈도어(high-frequency words)들을 포섭할 수 있다. 예를 들어 동사, 전치사, 그리고 형용사 등이 구의 구성 요소가 되고, 그 구가 문장 내외의 다른 구와 의미적 일관성 관계를 맺고 있다면, 구사슬의 요소인 구의 일부로서 그것들이 구사슬에 참가한다는 것이다. 이와 아울러, 수사 관계를 표층적으로 나타내는 담화 표지들도 구사슬의 요소가 될 수 있다. 결국 구사슬이란, 원문의 표층적 이해나 주제 전개 추적에 필요한 최소한의 계산언어학(computational linguistics)적인 도구일 것이다.

어휘 사슬을 생성하는데 동원한 언어학적 자원은 유의어 사전이다. 이것과 함께, 구사슬을 형성시키는데 추가로 필요한 도구가 견실한 구문 분석기이다. 먼저 유의어 사전은, 적절한 의미 처리 과정을 통해 확장시켜서 구사슬을 형성하는데 이용한다. 이에 대해서는 다음절에서 구체적으로 거론할 것이다. 견실한 구문 분석기로는, 한 문장 안에서 표층적 이해에 단서가 되어, 그래서 구사슬의 후보 요소가 되는 단어나 구를 선별하는데 사용한다.

3. 사서적 상황 지식베이스

사서적 상황 지식베이스(LSKB: Lexico-Situational Knowledge Base)는 사서적 지식베이스(LKB: Lexical Knowledge Base)[22, 24, 31]의 일종이라 할 수 있다. LKB란, 어휘에 대한 통사 및 의미 정보를 자연어 처리(NLP: Natural Language Processing)에 맞게 표현해 놓은 지식베이스이다. 여기에는, 자연어 처리에 유용하다고 인정되는, 어휘에 대한 주변 지식도 표현된다. 달리 말하여 LKB는, 기계 가독성 어휘 사전(MRD: Machine-Readable Dictionary)을 토대로 적절한 의미 후처리를 시행함과 병행하여 어휘에 대한 주변 지식도 이식시킨, 지식베이스 라고 하겠다. 한편, LSKB는 상태나 사건의 상호 연관성을 쉽게 확인할 수 있게 한 LKB이다. 이렇게 되기 위해서는, 적어도 사건이나 상태를 기술하는데 필수적인, 동사구(verb phrase)와 형용사구(adjective phrase)에 대한 의미 규정을 포함하여야 한다. 또한 사물의 용도에 대한 정보[16]도 수록함이 필요할 것이다.

LSKB 구축에 사용 가능한 어휘 데이터베이스(LDB: Lexical Database)로는 WordNet[22], Roget's Thesaurus[31], 그리고 MRC Psycholinguistic Database[24] 등이 있다. Roget's Thesaurus의 경우, 동사구 get away에 대한 항목은 다음과 같다:

#287. [Motion from.] Recession. -- V. *get away*

여기에서 #287은 범주(category) 번호를 나타내고, V는 get away가 동사임을 표시한다. 이에 대한 LSKB의 항목은, Prolog를 사용하여 잠정적으로 다음과 같이 표현할 수 있겠다:

```
verb_phrase(get*away,[287=[motion from]+recession%v|
                      ptrans+[agt:A,obj:O,
                              src:S:197=nearness,dst:D:196=distance]])
```

여기에서도 각 번호는 범주 번호를 나타낸다. 또한 get away에 대한 Roget's Thesaurus 항목을 그대로 수용하면서 그것에 대한 격틀(case frame)도 제시되었다. 격틀을 표현하는 데 나타난 agt, obj, src, 그리고 dst 등은 각각 행위자격(agent), 대상격(object), 원천격(source), 그리고 행선격(destination) 등을 의미한다. 이와 아울러, ptrans는, 장소의 이동을 표현하는 CD(Conceptual Dependency)[35] 이론의 근본 행동(primitive action)의 일종이다. 마지막으로, 대문자로 표시된 것들은 Prolog 변수들이다.

앞서 제시한 바와 같은 동사구 get away에 대한 LSKB의 항목은, 다음과 같은 어휘 데이터베이스와 함께,

```
lex_db(get*away,287=[motion from]+recession,v)
```

아래에 제시한, CD 이론의 근본 행동에 대한 격틀 데이터베이스에서 유도할 수 있다.

```
case_frame(ptrans,[agt:A,obj:O,src:S,dst:D])
isa(287=[motion from]+recession,ptrans)
case_role_spec(287=[motion from]+recession,[src:S:197=nearness,dst:D:196=distance])
```

Roget's Thesaurus 제 4 판(1977)[32]의 경우, 범주 개수가 1042 개이다. 격틀 데이터베이스는 낱낱의 동사구가 아닌 범주를 대상으로 구축할 것이므로, 적절한 의미처리 과정을 거친다면, 반자동적으로 앞서 보인 격틀 데이터베이스와 LSKB를 구축할 수 있을 것이다.

4. 수 사 관 계

원문의 내용 전개 구조를 나타내는 수사 관계(rhetorical relation)[6, 11, 14]는, 표층적 원문 이해(shallow text understanding)에 있어 유용한 도구이다. 수사 관계는, 원문의 최소 분석 단위들이 원문의 내용 전개에 있어서 어떠한 논리적 호응 관계를 가지는가를 나타낸다. 호응 관계의 예로서는, 정당화(justification), 증거(evidence), 양보(concession), 그리고 재언급(restatement) 등을 들 수 있다. 또한 원문의 최소 분석 단위는 통상 절(clause)이다.

수사 관계의 완전한 파악도 원문의 깊은 이해를 전제로 한다. 그러나 수사 관계의 단서를 원문의 도처에서 발견할 수 있는데, 그것들이 바로 담화 표지(discourse marker)들이다. 담화 표지의 출현 빈도는, 원문의 종류에 불문하고, 대략 두 절당 한 개꼴로 나타난다고 한다[30]. 또한 심리언어학자들의 연구 결과[6, 11, 14]에 의하면, 절보다 큰 원문 분석 단위 - 예를 들어, 문장, 단락, 그리고 원문 등 - 의 논리적 호응 관계를 담화 표지가 일관되게 나타낸다고 알려져 있다.

한편, 담화 표지는 단서구(cue phrase)이지만 모든 단서구가 담화 표지는 아니다[11]. 예

를 들어 incidently의 경우, 한 문장 안에서 다른 절을 해석하는데 영향을 주는 용법 - 문장적 용법(sentential use) - 으로는 부사로 쓰이지만, 수사적 용법(discourse use)으로는 지엽적인 논의를 이끄는 여담(digression) 관계를 나타낸다. 더욱이 단서구가 담화 표지로 확인되었다고 할지라도, 담화 표지는 일반적으로 두 개이상의 수사 관계를 표층적으로 나타낼 수 있으므로, 담화 표지의 중의성 문제를 해결하여야 한다.

이러한 점들을 고려해 볼 때, 담화 표지에 기반한 표층적 원문 분석 방법은, 담화 표지와 수사 관계가 일대일 대응을 이루고 또한 원문의 모든 문장에 담화 표지가 있을 경우, 그 방법의 유용성이 극대화된다. 그러나 대부분의 원문은 그렇지가 못하고, 더욱이 담화 표지가 전혀 없는 극단적인 경우도 있을 수 있다. 따라서, 담화 표지에 전적으로 의존하는 방법론의 유용성은, 문장당 담화 표지 개수 및 담화 표지의 수사 관계와의 일대일 대응성에 따라 제한된다.

수사 관계와 담화 표지의 이러한 대응 관계에 착안하여, 원문의 수사 구조를 표층적으로 분석하는데 그것을 이용하고, 또 그 결과를 원문 요약에 응용한 연구[18]가 최근에 있었다. 이 원문 요약기의 경우, 원문의 중요 부분에 대한 변별력이 회상(recall)도와 정확(precision)도 면에서 약 70% 정도라고 보고되었다. 이러한 변별력을 제고시키는 한 방법이 앞서 논의한 구사슬의 채용이라고 생각한다.

구사슬과 담화 표지는, 고려하는 각각의 후보 요소가 서로 중복하지 않으면서도, 원문의 주제 전개를 표층적으로 개관한다는 동일한 목표 달성에 유용한 도구이다. 따라서 원문의 주제 전개에 대한 표층적 개관이라는 목표에 상호 보완적인 이 두 도구를 적절히 활용한다면, 원문의 중요 부분에 대한 변별력 제고에 시너지 효과를 기대할 수 있다. 이에 대한 예를 다음절에서 보기로 한다.

5. 적 용 예

이 절에서는, 견실한 구문 분석기, 사서적 상황 지식베이스, 수사 관계와 구사슬 등의 언어학적 도구와 지식을 적절히 활용하여 원문의 주제 연접 관계를 밝히는 중층적 원문 이해의 예를 보기로 한다. 이해할 원문의 장르는 설화(narrative)이다. 설화문(narrative text)은, 그것의 이해에 대한 인지과학적 연구가 심도 있게 진행된 원문의 한 종류이다. 원문의 장르를 설화로 제한함으로써, 중층적 원문 이해 방법론의 타당성과 유용성 검증이 용이할 것이다. 그후, 이러한 중층적 원문 이해를 바탕으로, 원문에 대한 요약문을 생성해내는 실용적인 자동 요약 방법론 개발을 모색해 볼 수 있을 것이다.

설화문이란, 하나 혹은 그 이상의 특정 행위자의 행동으로 유발되는 일련의 사건이나 상태의 변화를 중심으로 전체 내용이 전개되는 원문의 한 종류이다[38]. 설화문의 전형적인 특징은, (1) 사건은 대부분 시간 경과순으로 기술된다. (2) 일인칭이나 삼인칭을 사용한다. 그리고 (3) 내용이 특정 행위자들을 중심으로 일련의 사건이나 상태의 변화로 전개된다.

Mike와 Paul에 관한 다음 설화문[15]을 보자. 이 원문에 대한 주제 전개 상황을 그림 1에서 확인할 수 있다:

Mike and Paul had been close friends ever since their high school days. But now

Mike wanted Paul out of town for a few days so that *he could build a patio in Paul's backyard as a surprise birthday present.* He suggested to Paul that he get away for a weekend, but *Paul said he wasn't interested.* On another occasion Mike casually spoke about the joys of fishing or camping trips. But Paul told him he enjoyed puttering around the house much more. Paul was getting very settled in his old age.

그림 1에서 알 수 있듯이, 실선으로 표시된 화살표에는 각각 담화 표지가 붙어 있고, 굵은 점선으로는 구사슬을 나타내며, 그리고 가는 점선으로는 의미적으로는 연관성이 있지만 구사슬을 계산한 방법으로는 확인할 수 없는 관련어를 연결하고 있다. 예문의 경우, 중요 부분은 (1) *he could build a patio in Paul's backyard as a surprise birthday present*와 (2) *Paul said he wasn't interested*이다. (1)의 내용이 중요하다는 것은, 그림에 나타난

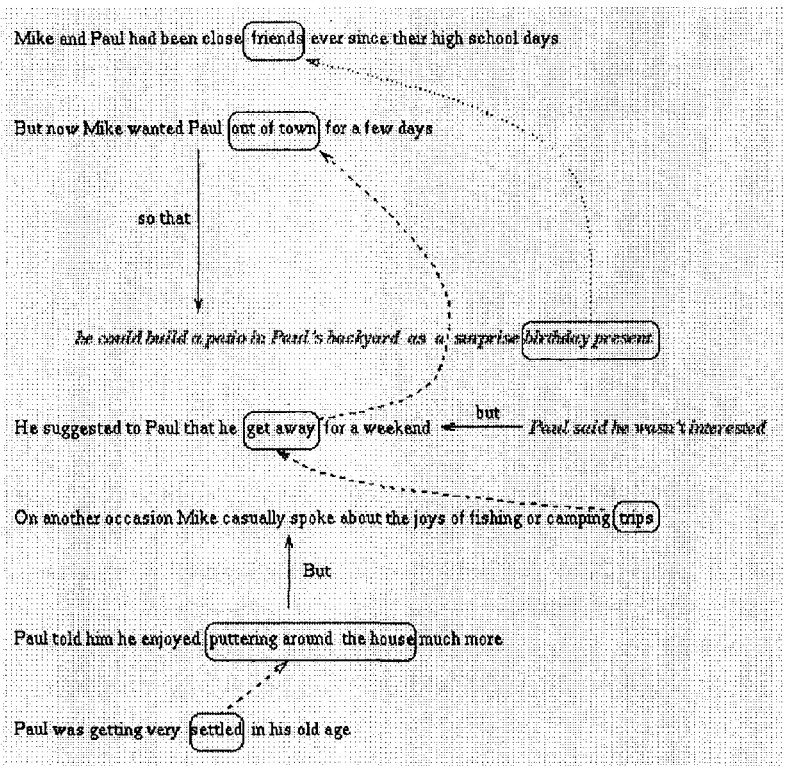

그림 1. 구사슬과 담화 표지로 표현한 원문 주제 전개 상황

구사슬과 담화 표지의 화살표가 (1)을 지향하고 있다는데서 알 수 있다. (2)의 중요도는, (1)의 중요성을 확인하는데 사용되었던 사슬과 반정립(antithesis)의 수사 관계로 연결되어 있다는 점에서, 그것을 가늠할 수 있다.

get away와 out of town을 연결하는 구사슬의 일부는, 전술한 바 있는 get away에 대한 사서적 상황 지식베이스(LSKB: Lexico-Situational Knowledge Base)의 항목과 out of town에 대한 다음 LSKB 항목을 토대로, 구사슬기(phrasal chainer)가 형성한다. 이 LSKB

항목은 out of town이라는 부사 혹은 형용사구에 대한 의미 분석 결과를 보인다:

```
prep_rel(out*of*town,
        [exteriority%adv+adj|
        [src:[town=189=[place of habitation, or resort]+abode%n]:197=nearness,
        dst:D:196=distance]])
```

전에 언급한 바 있는 Roget's Thesaurus[31]의 경우, out과 town에 대한 항목은 각각 다음과 같고,

#220. Exteriority. -- Adv. &c. adj. out
#189. [Place of habitation, or resort.] Abode.-- N. town

이들이 각각 다음에 보이는 것과 같이 lex_ind로 어휘 데이터베이스에 등록되어 있다:

```
lex_ind(out,220=exteriority,adv+adj)
lex_ind(town,189=[place of habitation, or resort]+abode,n)
```

out of town에 내재하는 전치사적 관계(prepositional relation) 분석은 out과 town의 범주 정보로 of의 의미를 유추하는 방식이다:

```
case_role_spec_for_prep(of,
        220=exteriority,189=[place of habitation, or resort]+abode,
        [src:189=[place of habitation, or resort]+abode:197=nearness,
        dst:D:196=distance])
```

out of town과 get away의 두 구가 구사슬로 연결되는 근거는, (1) 두 개의 구가 각각 동일 인물 Paul의 상태와 동작을 나타낸다. 이러한 사실은 적절한 구문 분석기를 통하여 쉽게 확인할 수 있다; (2) 인과 관계가 설정될 수 있다; 그리고 (3) 두 격역할소(case role), src와 dst에 대한 명세가 서로 모순되지 않는다:

```
prep_rel(out*of*town,
        [exteriority%adv+adj|
        [src:[town=189=[place of habitation, or resort]+abode%n]:197=nearness,
        dst:D:196=distance]])
verb_phrase(get*away,[287=[motion from]+recession%v|
                ptrans+[agt:A,obj:O,
                        src:S:197=nearness,dst:D:196=distance]])
```

6. 결 론

현재의 자연어 이해 이론과 기술 수준 그리고 가용 지식 베이스의 완전성 수준에서 볼 때, 심층적 원문 이해(deep text understanding)의 성취는 단기적으로 불가능하다. 따라서 근본적으로 원문 이해가 필요한 문서 요약 분야의 경우, 표층적 원문 이해(shallow text understanding)에 의존할 수밖에 없다. 수사 관계와 어휘 사슬은 이러한 표층적 원문 이해에 있어 유용한 도구이다. 이에 본 논문에서는 어휘 사슬을 구사슬(phrasal chain)로 확장하고, 수사 관계의 표층 단서인 담화 표지와 함께 표층적 원문 이해 - 특히, 원문의 중요 부분 선별 - 에 이들을 활용해 보았다.

구사슬과 담화 표지는, 고려하는 각각의 후보 요소가 서로 중복하지 않으면서도, 원문의 주제 전개를 표층적으로 개관한다는 동일한 목표 달성에 상호 보완적인 유용한 도구임을 본 논문에서 확인하였다. 자동 문서 요약기의 경우, 원문의 주제 전개에 대한 표층적 개관 기능은 필수 불가결한 요소이다. 따라서 실용적인 새로운 자동 문서 요약 방법론을 개발하는데 있어, 구사슬과 수사 관계를 표층적으로 나타내는 담화 표지는, 그 어느 것보다 유용한 언어학적 자원이 될 것임을 확신한다.

이러한 구사슬의 계산을 위하여, 사서적 지식베이스(LKB: Lexical Knowledge Base)의 일종인 사서적 상황 지식베이스(LSKB: Lexico-Situational Knowledge Base)를 본 논문에서 제안하였다. LKB란, 어휘에 대한 통사 및 의미 정보를 자연어 처리(NLP: Natural Language Processing)에 알맞게 표현해 놓은 지식베이스로, 어휘에 대한 주변 지식도 포함된다. 한 걸음 더 나아가 LSKB는, 동사구(verb phrases)와 형용사구(adjective phrases)에 대한 의미 규정을 포함하여, 상태나 사건의 상호 연관성을 쉽게 확인할 수 있게 한 LKB이다.

결국 본 논문에서는, 현실적으로 사용 가능한 언어학적 도구와 지식 - 견실한 구문 분석기, 사서적 상황 지식베이스, 수사 관계, 담화 표지, 그리고 구사슬 등 - 을 적절히 활용하여, 원문의 주제 연접 관계를 밝히는 중층적 원문 이해(mid-depth text understanding)를 시도했다고 할 수 있겠다. 그리고 이러한 중층적 원문 이해를 바탕으로, 주어진 원문에 대한 요약문을 생성해내는 실용적인 자동 요약 방법론 개발도 모색해 보았다.

7. 참 고 문 헌

[1] Bae, J.-H. J. and Lee, J.-H. Another Investigation of Automatic Text Summarization: A Reader-Oriented Approach. In Proceedings of ANZIIS '94 (Australian and New Zealand Conference on Intelligent Information Systems), pp. 472-476, 1994.

[2] Barzilay R. & Elhadad M. Using Lexical Chains for Text Summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, 1997.

[3] Bateman, J. and Teich, E. Selective information presentation in an integrated publication system: an application of genre-driven text generation. Information Processing & Management Volume 31, Issue 5, pp.753-767,1995

[4] Brandow, R., Mitze, K. and Rau, L.F. Automatic condensation of electronic publications

by sentence selection. Information Processing & Management Volume 31, Issue 5, pp. 675-685, 1995.

[5] Ciravegna, F. Understanding messages in a diagnostic domain. Information Processing & Management Volume 31, Issue 5, pp. 687-701, 1995.

[6] Costermans, J. and Fayol, M. Processing Interclausal Relationships. Studies in the Production and Comprehension of Text. Lawrence Erlbaum Associates, Publishers, 1997.

[7] DeJong G. F. Skimming stories in real time: an experiment in integrated understanding. Ph.D. thesis, Yale University, 1979.

[8] Edmundson, H.P. 1968. New methods in automatic extracting. Journal of the Association for Computing Machinery, Volume 16, Number 2, pp. 264--285, 1968.

[9] Hahn U. Topic parsing: accounting for text macro structures in full-text analysis. Information Processing and Management, Volume 26, Number 1, pp. 135-170, 1990.

[10] Halliday, M. A. K. and Hasan, R. Cohesion in English. Longman, London, 1976.

[11] Hirschberg, J. and Litman, D. Empirical studies on the disambiguation of cue phrases. Computational Linguistics, Volume 19, Number 3, pp. 501-530, 1993.

[12] Jones K. S., What might be in a summary? In G. Knorz, J. Krause, and C. Womser-Hacker, editors, Information retrieval '93: von der modellierung zur anwendung, pp. 9-26. Konstanz, Universitatsverlag Konstanz, 1993.

[13] KAML Research Group. Research on Text Summarization (fragments of a grant proposal). http://www.csi.uottawa.ca/~szpak/proposals/text-summ-1996_ToC.html.

[14] Knott, A. A Data Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh, 1996.

[15] Lehnert, W. G. Plot units: A narrative summarization strategy. In W. G. Lehnert and M. H. Ringle (Eds.), Strategies for natural language processing, Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.

[16] Lehnert, W. G. and Burstein M. H. The Role of Object Primitives in Natural Language Processing. In Proceedings of the Sixth International Joint Conference on Artificial Intelligence(IJCAI-79), pp. 522-524, 1979.

[17] Liddy, E. D. et al. Development, implementation and testing of a discourse model for newspaper texts. In Proceedings of the 1993 ARPA Human Language Technology Workshop, pp. 159-164, Princeton, New Jersey, March 1993. Advanced Research Projects Agency, Morgan Kaufmann.

[18] Marcu D. From Discourse Structures to Text Summaries. The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp. 82-88, Madrid, Spain, July 11, 1997.

[19] Marsh E., Hamburger H., and Grishman R. A production rule system for message summarization. In Proceedings of the National Conference on Artificial Intelligence, pp. 243-246. American Association for Artificial Intelligence, 1984.

[20] Maybury, M.T. Generating summaries from event data. Information Processing & Management Volume 31, Issue 5, pp. 735-751, 1995.

[21] McKeown, K., Robin, J. and Kukich, K. Generating concise natural language summaries. Information Processing & Management Volume 31, Issue 5, pp. 703-733, 1995.

[22] Miller, G. A. WordNet: A lexical database for English. Communications of the ACM, Volume 38, Number 11, pp. 39-41, 1995.

[23] Morris, J. and Hirst, G. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. Computational Linguistics, Volume 17, Number 1, pp. 21-48, 1991.

[24] MRC Psycholinguistic Database. http://www.dci.clrc.ac.uk/Activity.asp?Psych+280.

[25] Okumura M. & Honda T. Word Sense Disambiguation and Text Segmentation Based on lexical Cohesion. In Proceedings of The 15th International Conference on Computational Linguistics, pp. 775-761, 1994.

[26] Ono, K., Sumita K. and Miike S. Abstract Generation Based on Rhetorical Structure Extraction. pp. 1-5, COLING-94, 1994.

[27] Paice, C. D. Constructing literature abstracts by computer. Information Processing and Management, Volume 26, Issue 1, pp. 171-186, 1990.

[28] Rama, D. V. and Srinivasan, P. An investigation of content representation using text grammars. ACM Transactions on Information Systems, Volume 11, Number 1, pp. 51-75, 1993.

[29] Rau L. F. Conceptual information extraction and information retrieval from natural language input. In Proceedings of the Conference on User-Oriented, Content-Based, Text and Image Handling, pp. 424-437, Cambridge, Massachusetts, 1988.

[30] Redeker, G. Ideational and pragmatic markers of discourse structure. Journal of Pragmatics, Volume 14, pp. 367-381, 1990.

[31] Roget's Thesaurus. http://www.promo.net/pg/_authors/roget_peter_mark_. html# rogetsthesaurus.

[32] Roget, P.M. Roget's International Thesaurus, Fourth Edition. Harper and Row Publishers Inc., 1977.

[33] Rumelhart, D. E., Notes on a schema for stories. In D. G. Bobrow and A. Collins (Eds.), Representation and understanding, New York: Academic Press, 1975.

[34] Rush, J. E., Salvador, R., and Zamora, A. Automatic abstracting and indexing. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. Journal of American Society for Information Sciences, Volume 22, Number 4, pp. 260--274, 1971.

[35] Schank, R. C. Conceptual dependency: A Theory of natural language understanding. Cognitive Psychology, Volume 3, October 1972.

[36] Souther, J. W. and White, M. L., Technical report writing, second edition. Wiley-Interscience, John Wiley & Sons, New York, 1977.

[37] Tait, J. I. Automatic summarising of English texts. Thesis, University of Cambridge, 1983; Technical Report 47, Computer Laboratory, University of Cambridge, 1983.

[38] The Linguistic Glossary. http://www.sil.org/linguistics/glossary/. Summer Institute of

Linguistics, Inc., 1997.

[39] T'sou, B. K., Lin, H.-L., Ho, H.-C., Lai, T. B. Y., and Chan, T. Y. W. Automated Chinese Full-text Abstraction Based on Rhetorical Structure Analysis. In Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, pp. 259-266, 1995.

[40] Young S. R. and Hayes P. J. Automatic classification and summarization of banking telexes. In Proceedings of the Second Conference on Artificial Intelligence Applications, pp. 402-408, 1985.