문장추상화 : 개념추상화를 도입한 문장교열

김 곤[†] · 양 재 군[††] · 배 재 학[†††] · 이 종 혁[††††]

요 약

문장추상화(Sentence Abstraction)는 문장의 의사전달 기능이 보존된 단순화이다. 이는 문장교열(Sentence Revision)과 개념추상화(Concept Abstraction)를 동시에 가능하게 한다. 문장교열은 사람이 생각한 바와 문장으로 표현된 의미의 차이를 해결하는 방법이다. 개념추상화는 개념들의 공통된 요소로부터 얻은 보편적인 관념을 표현하는 것이다. 문장추상화는 문장의 주요구성성분들을 선별해 내고, 이들의 의미적인 정보를 파악하여 상위개념을 표현함으로써 문장교열과 개념추상화를 가능하게 한다. 본 논문에서는 문장추상화를 위한 구문분석기 LGPI+와, 온톨러지 OfN을 구체화하였다. 문장추상기 SABOT는 LGPI+와 OfN을 활용하며, 구문분석 결과를 처리하여 문장에서 추상화 할 후보단어를 선택한다. 문장추상화를 활용한 원문이해 시스템으로 23개 이야기의 58개 문단에 대해 중요 문장에 대한 문장재현율과 선별된 문장들의 주제관련성을 확인해 보았다. 실험결과, 문장재현율은 54~72%의 범위이었고, 주제관련성은 76~86% 정도의 비율로 나타났다. 이를 유사 시스템과 비교해 보았을 때, 약 10~20% 정도의 성능향상을 보인다. 본 논문에서는 문장추상화를 활용하여 글의 화제문을 효율적으로 선택할 수 있는 문장교열과 원문의 이해심도를 보다 더 깊게 할 수 있는 개념추상화가 가능함을 확인하였다.

Sentence Abstraction : Sentence Revision with Concept Abstraction

Gon Kim[†] · Jae-Gun Yang[††] · Jae-Hak J. Bae[†††] · Jong-Hyeok Lee[††††]

ABSTRACT

Sentence abstraction is a simplification of a sentence preserving its communicative function. It accomplishes sentence revision and concept abstraction simultaneously. Sentence revision is a method that resolves the discrepancy between human's thoughts and its expressed semantic in sentences. Concept abstraction is an expression of general ideas acquired from the common elements of concepts. Sentence abstraction selects the main constituents of given sentences and describes the upper concepts of them with detecting their semantic information. This enables sentence revision and concept abstraction simultaneously. In this paper, a syntactic parser LGPI+ and an ontology OfN are utilized for sentence abstraction. Sentence abstracter SABOT makes use of LGPI+ and OfN. SABOT processes the result of parsing and selects the candidate words for sentence abstraction. This paper computes the sentence recall of the main sentences and the topic hit ratio of the selected sentences with the text understanding system using sentence abstraction. The sources are 58 paragraphs in 23 stories. As a result of it, the sentence recall is about 54~72% and the topic hit ratio is about 76~86%. This paper verified that sentence abstraction enables sentence revision that can select the topic sentences of a given text efficiently and concept abstraction that can improve the depth of text understanding.

키워드 : 문장추상화(Sentence Abstraction), 문장교열(Sentence Revision), 개념추상화(Concept Abstraction)

1. 서 론

문장교열은 사람이 생각한 바와 문장으로 표현된 의미의 차이를 해결하는 방법이다. 문장교열과정에서, 사람 또는 컴퓨터는 철자나 문법, 문체상의 오류를 수정한다. 그리고, 부적절하거나 반복적인 용어들을 삭제한다. 또한, 의미가 명확하지 않은 경우에는 적절한 어휘를 추가하거나 대용하여 원문을 보다 더 간결하고 명확하게 만든다. 본 논문에서는 문장교열의 한 방법으로 문장추상화를 활용한다. 문장추상화는 문장의 주요구성성분들간의 문법적인 관계뿐만 아니라 의미적인 관계도 파악한다. 따라서, 문장추상화를 통하여 문장교열뿐만 아니라 문장에 내포된 상위개념을 찾아내는 개념추상화가 가능하다.

1.1 연구 동기

문장교열은 잘못 사용된 어휘를 올바른 어휘로 대치하거나, 어법에 맞지 않는 비문을 올바른 문장으로 수정하여, 원문을 보다 간결하고 명확하게 한다. 기존의 문장교열 방법론들[8, 17-20]은 문장의 통사구조를 바탕으로 얻을 수 있는 문법적인 정보를 토대로 하고 있다. 여기에는 문장간 또는 문장내 구성성분들이 가지는 의미정보가 부족하거나, 의미

※ 이 논문은 2004년 울산대학교의 연구비에 의하여 연구되었음. 본 연구는 한국과학재단 목적기초연구 R05-2004-000-12362-0 지원으로 수행되었음.
† 종신회원 : 울산대학교 대학원 컴퓨터 · 정보통신공학부
†† 준 회 원 : 울산대학교 대학원 컴퓨터 · 정보통신공학부
††† 종신회원 : 울산대학교 컴퓨터 · 정보통신공학부 교수
†††† 정 회 원 : 포항공과대학교 컴퓨터공학과 교수
논문접수 : 2004년 5월 18일, 심사완료 : 2004년 7월 3일

단계의 분석과정이 아예 포함되어 있지 않은 경우도 있다. 상대적으로 실현이 어려운 의미분석의 경우, 기존 연구에서는 말뭉치(Corpus)를 활용하거나 분야지식을 구동시켜 성취하고 있다.

문장교열과정에서, 문장내 구성성분들이 가지는 의미적인 중요도를 파악하여 활용한다면, 문장교열의 효율성을 보다 더 높일 수 있을 것이다. 본 논문에서는 이를 위하여 기존의 문장교열방법과는 달리, 문장의 구성성분들 간의 관계와 문장내 구성성분들이 가지는 온톨러지 정보를 바탕으로 한다. 논문에서 제시하는 방법은 개념추상화를 활용한 문장교열방법으로, ① 문장의 주요의미를 정형적(Formal)으로 표현하고, ② 문장구성성분들의 의미정보를 파악하여 상위개념을 표현할 수 있다.

1.2 관련 연구

자연어처리 분야에서는 가용한 언어학적 도구와 자원을 활용하여 기계에 의한 자동문서요약에 대한 연구[8, 17-20, 21]가 계속되어 왔다. 이상적인 요약이란, 전체 내용에 대한 간결한 정리뿐만 아니라, 가능하다면 원문내용에 나타난 중요사실에 대한 필자의 판단 및 견해를 아울러 포함시킨 것을 말한다. 즉, 사람이 생각한 바를 충실하게 문장으로 표현할 수 있어야 한다. 문장교열은 이러한 요약과정에서 사람이 생각한 바와 문장으로 표현된 의미의 차이를 해결하는 방법이다.

요약 전문가는 초벌 요약을 얻은 다음, 이를 개정하여 요약문의 단축성, 조응성, 유려성 등을 증진시킨다. 여기에는 다양한 종류의 첨삭(Cut-and-Paste) 작업이 수반된다[8]. ① 문장내(Intra-Sentence) 및 문장간(Inter-Sentence) 교열, ② 애매하거나 장황한 어구 제거, ③ 표현의 일반화 또는 특화, ④ 대용어 정리, 그리고 ⑤ 어휘대치 등. 이러한 첨삭작업을 주어진 글에 처음부터 바로 적용시켜 요약문을 얻을 수도 있다. 이 점에 착안하여 개발된 것이 생성형 요약(Summarization by Generation) 방법론이다. 첨삭작업을 방법론 구현의 각도에서 보면, 통사적 처리로 실현할 수 있는 것과 함께 상당한 수준의 의미분석이 뒷받침되어야 하는 것으로 나뉜다. 이 중에서 상대적으로 실현이 어려운 의미분석의 경우, 기존연구에서는 말뭉치(Corpus)를 활용하거나 분야지식을 구동시켜 성취하고 있다.

문장교열은 그 적용범위에 따라, 한 문장 내에서 이루어지는 교열(Local Revision)과 문장 간 교열(Global Revision)로 나눌 수 있다[8, 20]. 문장내 교열(Local Revision)[8]의 경우에는, 여분의 용어나, 부적절하고 애매한 용어들이 삭제된다. 어휘의 반복을 피하기 위하여, 문장 의존도[21]에 따라, 다른 용어로 대치되거나 추가될 수도 있다. 이러한 용어들의 삭제와 추가로 보다 더 간결하고 분명한 요약을 얻을 수 있다. 문장간 교열(Global Revision)[8]의 경우는, 각 문장들에 대해서 문장내 뿐만 아니라 다른 문장들로부터 얻은 정보들을 활용하고 있다.

문서요약을 위한 문장교열(Sentence Revision) 방법[8]에는 ① 원문축소(Text Compaction), ② 문장절감(Sentence Reduction), ③ 문장압축(Sentence Compression) 등이 있다. 이러한 문장교열의 방법 중에서 ① 원문축소(Text Compaction)[8, 18]는 주어진 문장에서 나타나는 용어에 대해서 개념적으로 모호한 부분을 해결하고 문장내 구성성분들의 문법적인 관계에 따라 문장을 축소하는 방법이다. 이러한 문장의 축소를 위한 기본적인 규칙에는 명사와 형용사를 다른 구성성분들보다 더 중요하다고 보고, 명확한 의미를 가진 명사를 중요시하며, 문장에서의 주절과 반의어들을 중심으로 하는 것 등이 있다. ② 문장절감(Sentence Reduction)[8, 17]은 원문을 축소하는 방법으로 문장에서 나타나는 어휘들의 연결관계를 파악하고, 그 응집도에 따라 문장에서 불필요한 부분을 제거하는 것이다. ③ 문장압축(Sentence Compression)[8, 18]은 문법적으로 문장을 분석한 자료를 토대로 문장에서의 주요어를 선별하여 요약 과정을 수행하는 방법이다.

기존의 문장교열 방법론들[8, 17-20]은 문장이나 문장내 구성성분들이 가지는 의미정보 보다는 문장의 통사구조를 바탕으로 얻을 수 있는 문법적인 정보를 토대로 문장을 표현하는 방법이다. 이와는 달리, 본 논문에서 제시하는 문장추상화(Sentence Abstraction)는 개념추상화를 도입한 문장교열(Sentence Revision)작업이다. 교열작업의 목표는 원문축소(Text Compaction)[8, 18]나 문장절감(Sentence Reduction)[17]의 그것과 유사하다. 그러나 그 차이점은 문장추상화는 처리과정에 있어서, ① 문장의 주요의미를 정형적(Formal)으로 표현함과 아울러 ② 개연사슬(Abductive Chain)[2, 3, 13]의 후보어구(Candidate Word)나 줄거리 단위(Plot Unit)[2, 13]의 구성요소가 될 수 있는 문장의 주요 구성성분들을 분리해 낼 수 있다.

1.3 연구 내용 및 기여도

요약과정은 일종의 추상화 과정[1, 2, 3, 13]으로, 원문에서 상대적으로 중요한 부분을 발췌하여 일반화시키는 과정을 말한다. 추상화에는 원문의 각 문장 구성성분들 중에서 추출할 것에 대한 선별작업이 수반된다.

문장추상화란 문장의 의사전달 기능이 보존된 단순화이다. 문장추상화를 통해서 문장의 주요의미를 정형적으로 표현하고, 주요 문장구성성분들을 선별해 낼 수 있다. 문장추상화는 구문분석기를 통해 얻은 정보와 Roget 시소러스에 기반한 온톨러지를 바탕으로 한다. 구문분석기는 주어진 문장을 구문분석하여, 구성성분에 대한 구문상 중요도를 파악한다. 활용하는 온톨러지(Ontology)는 주어진 문장에서 중요정보를 분별하고 이야기를 이해하기 위하여 고안된 것이다. 구문분석기를 통해 온톨러지의 유형으로 확인된 것만을 추

상화된 문장의 구성성분으로 채택한다. 구문분석기의 출력은 문장추상기로 처리하여 추상화할 후보단어를 선택한다. 문장추상화는 문장의 주요구성성분들을 선별해 내고, 이들의 의미적인 정보를 파악하여 상위개념을 표현함으로써 문장교열과 개념추상화를 가능하게 한다.

일반적인 원문은 그 표면적 구조로 보아 장(Chapter), 절(Section), 문단(Paragraph), 문장(Sentence), 절(Clause), 구(Phrase), 단어(Word), 글자(Letter) 등으로 세분할 수 있다. 여기에서 문단은 작문의 단위로서, 저자가 이야기하고자 하는 화제가 통상 한 개씩 들어간다[2]. 따라서, 문장추상화를 통하여 추상화된 문장들로 구성된 문단은 문단추상화라고 할 수 있다.

본 논문에서는 문장교열의 한 방법으로 문장추상화를 활용하고, 이를 원문이해 시스템에 적용하여 중요 문장에 대한 문장재현율과 주제관련성을 확인하였다. 또한, 유사 시스템과 비교하여 성능향상을 보였다. 이를 통하여, 개념추상화를 도입한 문장교열의 한 방법으로 문장추상화를 활용할 수 있음을 보이고, 원문이해에도 적용할 수 있음을 보였다.

2. 문장추상화

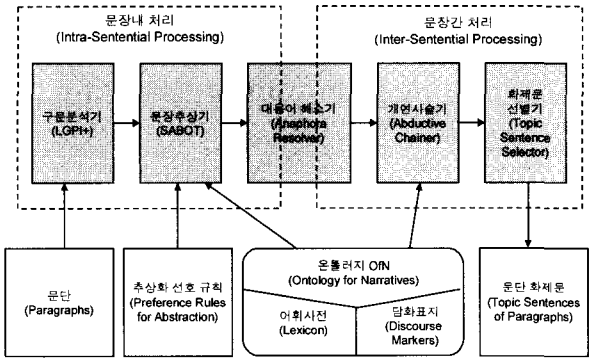
추상화(Abstraction)란 실체(Entity)의 본질적인 속성이 보존된 단순화(Simplification)이다. 추상화의 결과는 실체의 단순화된 표현이며, 여기에는 다른 실체와 구별할 수 있는 본질적인 속성이 보존되어야 한다. 한편, 문장은 생각이나 감정의 완결된 내용을 나타내는 의사전달의 최소 독립 문법적 단위이다. 이러한 문장의 본질적인 기능은 의사전달이다. 문장과 추상화의 뜻을 토대로, 문장추상화(Sentence Abstraction)는 의사전달 기능이 보존된 문장의 통사적, 의미적 단순화라고 할 수 있다.

2.1 문장추상화 원리

문장추상화는 Roget 시소러스에 기반한 온톨러지와 구문분석기를 통해 얻은 정보를 바탕으로 한다. 활용하는 온톨러지는 Roget 시소러스를 심중사전(Lexicon)으로 삼아 이를 재구성하여 얻었다. Roget 시소러스에는 체계적으로 분류된 어휘에 관한 지식이 있다. 이는 사건, 상태, 개체속성 등을 기술한 것으로 온톨러지가 갖추어야 기본지식을 내장하고 있다. 이와 더불어 다양한 숙어와 복식어휘(Multiword Unit)도 포함되어 있다.

한편, 구문분석기는 원문의 각 문장을 구문분석하여, 구성성분에 대한 구문상 중요도를 파악한다. 구문분석기의 출력은 문장추상기로 처리하여 문장의 요점어(Pivot Word)의 위치를 확인한다. 요점어는 문장추상화 후보단어로서, 문장을 추상화시킨 후에도 계속 잔류할 문장의 주요 구성성분이다. 이것의 의미범주는 Roget 시소러스를 재구성한 온톨러지에 등록되어 있는 것이어야 한다.

문장추상기는 구문분석기의 출력내용에서 주어, 동사, 목적어, 그리고 동사 수식어구에 주목한다. 대체로 최상위 어구(Top-Level Phrase)의 주요어(Head Word)가 우리의 주된 관심대상이 된다. 그러나 동사가 의미적으로 변화(Change)를 내포하고 있거나 검토중인 단어가 심상(Affect State)에 연관되어 있을 때에는, 문장 구성성분에 대한 추상화 검토심도를 한 단계 깊게 한다. 이 경우, 현재 검토중인 단어의 목적어구나 수식어구의 주요어를 검토대상에 포함시킨다. 이러한 과정에서 복식어휘는 한 단어로 취급한다. 검토대상이 되는 어구는 온톨러지 범주가 결정된 후, 문장추상화에 참가할 요점어(Pivot Word)가 된다. (그림 1)의 원문이해 시스템에서 문장추상화는 주어진 원문의 문장내 처리를 담당한다.



(그림 1) 원문이해 시스템 구성도

(그림 1)의 원문이해 시스템은 문장내 처리와 문장간 처리로 나누어져 있다. 문장추상화는 문장내 처리에서 활용된다. 이러한 문장추상화를 통해서, ① 문장의 주요의미를 정형적(Formal)으로 표현함과 아울러 ② 개연사슬(Abductive Chain)[2, 3, 13]의 고리(Link)가 될 수 있는 후보어(Candidate Word)나 줄거리 단위(Plot Unit)[2, 3, 13]의 심상(Affect State)이 될 수 있는 문장 구성성분을 분리해낼 수 있다.

2.2 문장추상화 알고리즘

문장추상기 SABOT
입력 : 구문분석기 LGPI+의 출력
출력 : 문추상식
[1] 추상화 검토대상 선정 : 문장의 주어, 동사, 목적어, 그리고 동사 수식어구에 대해서 다음 작업을 실시한다.
  (1) 최상위 어구(Top-Level Phrase)의 주요어(Head Word)를 파악한다. 이 때 복식어휘(Multiword Unit)는 한 단어로 취급한다.
  (2) 전치사구의 경우, 전치사와 그 목적어만 고려한다.
  (3) 동사가 의미적으로 변화(Change)를 내포하고 있거나 검토중인 단어가 심상(Affect State)에 연관되어 있을 때에는, 문장 구성성분에 대한 추상화 검토심도를 한 단계 깊게한다. 이 경우, 현재 검토중인 단어의 목적어구나 수식어구의 주요어를 검토대상에 포함시킨다.
[2] OfN 범주 결정 : 검토대상으로 선정된 어휘들에 대해서 다음 작업을 실시한다.
  (1) 단어의 OfN 범주를 결정한다.

(2) 동사가 변화의 의미를 포함하고 있고 이 동사의 목적어나 수식어구의 OfN 범주가 시간(Time) 또는 공간(Space)이라면, 이들의 OfN 범주는 시공의 변화(delta-time 또는 delta-space)가 된다.

[3] 문장의 주체와 객체를 파악한다.

[4] 문추상식으로 표현한다.

2.3 문추상식

추상화된 문장은 다음과 같은 문추상식(Sentence Abstraction Formula)으로 표현한다.

$$SID : [Pred1 : Words1 / (Chars1), \cdots, Predn : Wordsn / (Charsn)]$$

여기에서 ① SID는 절(Clause) 번호를 포함한 문장 식별자(Sentence Identifier)이다. ② Predi는 원자식(Atomic Formula)으로서 pred(args)의 Prolog 술어형태를 취한다. 이 식은 OfN에서 분류된 개념체(Conceptualization)의 표현이다. ③ Wordsi는 문장의 요점어를 나타낸다. 그리고 ④ Charsi에는 문장의 주체와 객체를 기록한다.

문추상식이 시사하고 있지만, 문장추상화 과정에서 등장인물의 교차관계(Cross-Character Relationship)를 파악한다. 일반적으로 글에는 여러 주인공이 등장한다. 등장인물이 여럿인 글을 이해하기 위해서는 그들의 역학관계를 파악하여 추적할 수 있어야 한다. 이러한 역학관계에는 등장인물간 역할관계 뿐만 아니라, 등장인물 사이의 상호작용을 포함하고 있다. 즉, 문추상식에는 등장인물간의 이러한 상호영향을 표현할 장치가 마련되어 있는 것이다.

2.4 활용한 도구와 자원

본 논문에서 이용한 문장추상화 도구들은 ① 주어진 문장에서 중요정보를 분별하기 위한 온톨러지 OfN(Ontology for Narrative), ② 문장의 구성성분들 사이의 관계를 파악하는 구문분석기 LGPI+, ③ 문장의 구성성분들이 가지는 온톨로지 정보와 추상화를 위한 선호규칙(Preference Rules)을 적용하여 문장추상화 작업을 수행하는 문장추상기 SABOT이다.

2.4.1 설화용 존재론 : OfN

본 논문에서는 주어진 문장에서 중요정보를 분별하고 이야기를 이해하기 위한 존재론으로 OfN을 사용한다. OfN 구현은 Roget 시소러스[4]를 심중사전(Lexicon)으로 삼아 이를 재구성하여 얻었다[2, 5]. Roget 시소러스에는 체계적으로 분류된 어휘에 관한 지식이 있다. 이는 사건, 상태, 개체속성 등을 기술한 것으로 온톨러지가 갖추어야 할 기본지식을 내장하고 있다. 이와 더불어 다양한 숙어와 복식어휘(Multi-word Unit)도 포함되어 있다.

OfN은 문장추상화 과정에서 추출할 문장구성성분에 대한 선택기준을 제공한다. 온톨러지 OfN은 Roget 시소러스와 대응하면서 다음의 7가지 범주로 구성되어 있다. 등장인물(Character), 심상(Affect State), 사건(Event), 상태(State), 시간(Time)과 공간(Space), 담화표지(Discourse Marker). 이렇게 설정한 OfN을 구축하기 위해서 먼저 Roget 시소러스[4]의 범주를 심상, 시간과 공간, 사건, 그리고 상태 등으로 재편성하였다. 등장인물 유형에 속하는 어휘들은 고유명사 자원[11]을 이용하여 선정하였다. 담화표지의 경우는 수사구조의 연구결과[7]를 활용하였다. 이와는 달리 시공의 변화는, 구문분석 후 문장의 구성성분간의 상호작용에 의하여 확인되는 유형인 바, 그 기본유형은 시간과 공간이다. OfN은 Roget 시소러스를 재구성하여 새로운 분류체계를 부여하고 관련 정보를 추가한 것이다. 이는 실체의 속성과 상관관계를 표현하는 이야기 이해용의 온톨러지를 향하여 Roget 시소러스를 한 단계 향상시킨 것이라고 할 수 있다.

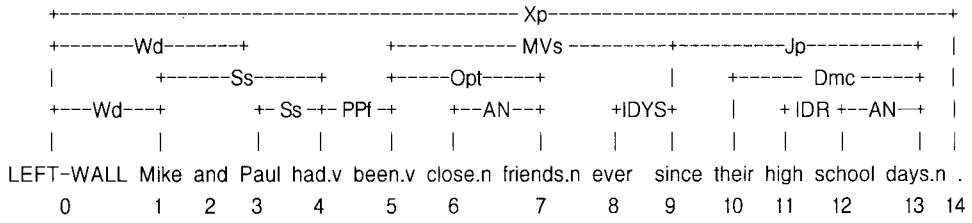
OfN과 함께 견실한 구문분석기도 활용하는데 그 과정은 다음과 같다. ① 주어진 문장을 구문분석하여, 구성성분에 대한 구문상 중요도를 파악한다. ② 중요 구성성분에 대한 OfN 유형을 확인한다. ③ 확인된 OfN 유형을 토대로, 구문상 중요도를 평가한다. ④ OfN 유형으로 확인된 것만을 추상화된 문장의 구성성분으로 채택한다. 이 과정에서 문장구성성분이 OfN의 복수 범주에 해당될 경우에는 다음과 같은 범주 우선순위를 따라 해당범주를 지정한다. Character > Affect State > Cue Phrase > Event > State > Space > Time.

2.4.2 구문분석기 : LGPI+

원문의 각 문장을 구문분석하는 데는 LGPI+(Link Grammar Parser Interface+)[1]를 사용하였다. LGPI+는 Link Grammar Parser[23]에 대한 SWI-Prolog API[6]를 제공한다. LGP는 6만 어형을 수록한 사전을 내장하고, 다양한 구문구조를 처리할 수 있다. 이 사전은 필요에 따라 확장이 가능하다. 입력문장에 대한 LGP 구문분석 결과는, 표식고리의 집합으로 문장의 통사구조가 표현된다. 표식고리는 한 쌍의 단어를 연결하며 그것들의 문법적인 기능을 표시한다. 이러한 LGP 구문분석 정보는 (그림 2)와 같이 나타난다. 이를 문장추상기에 적용하기 위해 LGPI+를 이용하여 (그림 3)과 같은 출력을 얻을 수 있다.

LGPI+의 출력은 문장추상기 SABOT(Sentence Abstracter Based on Ontology)[1, 2]으로 처리하여 문장의 요점어(Pivot Word)의 위치를 확인한다. 요점어는 문장추상화 후보단어로서, 문장을 추상화시킨 후에도 계속 잔류할 문장의 주요 구성성분이다. 이것의 의미범주는 OfN에 등록되어 있는 것이어야 한다. LGPI+는 입력문장에서 나타나는 표식고리들을 문장의 단어에 대한 색인번호로 대상을 지정하여 나타낸다.

다음 문장을 생각해 보자. *Mike and Paul had been close friends ever since their high school days.* 이에 대한 구문분석 결과는 (그림 2)와 같다.



(그림 2) 예문에 대한 구문분석 결과 : LGP

(그림 2)의 구문분석 결과는 문장구성성분들간의 관계를 표식고리로써 나타내고 있다. 각 표식고리의 역할은 다음과 같다.

① *Xp*는 문장의 마침표와 좌벽을 연결한다.
② *MVs*는 동사를 접속사와 연결시킨다.
③ *Jp*는 전치사와 그것의 복수형 목적어를 연결한다.
④ *Opt*는 *be* 동사를 복수명사에 연결하고 *there* 구문 후처리에 참가한다.
⑤ *Dmc*는 정관사와 복수명사를 연결한다.
⑥ *Wd*는 평서문에서 주부를 좌벽에 연결한다.
⑦ *Ss*는 단수명사를 단수동사형과 연결한다.
⑧ *PPf*는 *have* 동사를 과거분사형과 연결하고 *it* 나 *there* 구문 후처리에 참가한다.
⑨ *A*는 한정형용사를 명사와 연결한다.
⑩ *ID*[*X*][*Y*]는 숙어단어를 일렬로 연결한다. 여기에서 *X*, *Y*는 임의의 영문자이다.
⑪ *AN*은 수식명사를 명사와 연결한다.

(그림 2)의 결과에서 표식고리 ID 유형은 구문분석기 LGPI+가 숙어 문자열을 만났을 때 발생시키는 것이다. 물론 숙어는 구문분석기의 사전에 등록된 것에 한정된다. 이 사전은 용도에 맞게 확장이 용이하게 되어있다. 예문의 경우, IDYS가 ever와 since를 연결하고 있고 high와 school도 IDR로 연결되어 있다. 일반적으로 부사인 ever와 전치사인 since는 각각 그 자체로는 문장추상화 단계의 후보단어로서 자격이 없다. 물론 어휘사슬 형성시에도 마찬가지이다. 그 이유는 이들이 빈출어일 뿐만 아니라 중의적이기 때문이다. 그러나 복식어휘(Multiword Unit)로서의 ever since는 단서구(Cue Phrase)의 역할을 할 수 있어서, 필자의 어떤 의도를 함축하고 있는 것이다. 즉 담화정보를 담고 있다. 이와 함께 형용사 high도 명사 school과 함께라면 후보단어구성원의 자격을 부여받을 수 있는 것이다.

이 예를 통해 전치사, 부사, 형용사 등과 같은 것들이 문장추상화 과정에서 후보단어의 자격을 취득할 수 있음을 알았다. 이것들과 함께 구동사(Phrasal Verb) 또한 유망한 후보단어이다. 구동사는 일반적으로 복식어휘이고 이 안에는 동사나 불변화사(Particle)가 빈출어로 포함이 되어 있다. 이러한 품사의 단어들은 종래의 어휘사슬을 형성하는데 단지 빈출어 또는 중의성이 농후하다는 이유로 간과되었던 것이다. 그러나 개별적인 단어로 보면 출현빈도가 높고 중의적이나, 복식어휘의 구성요소로 나타날 때에는 그 중의성이 대체로 해소된다.

(그림 3)은 위 예문에 대한 LGPI+의 출력결과이다. LGPI+는 LGP의 구문분석 결과를 기계가독형으로 바꾸어 주며, 문장구성성분들 간의 연결유형은 문장의 각 단어에 대한 색인번호로 그 대상을 지정한다.

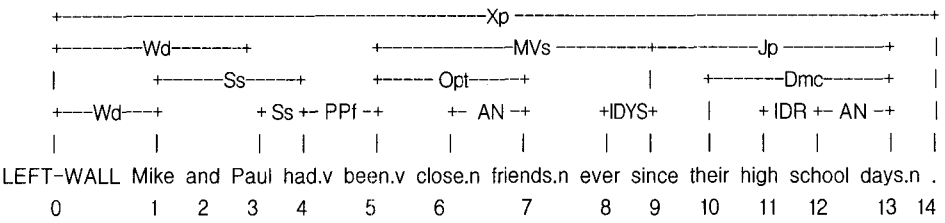
```
[link([m], connection(3-4, s-[s], paul(_G1443), had(v))),
 link([m], connection(0-3, w-[d], 'left-wall'(_G1413), paul(_G1415))),
 link([m], connection(11-12, idr-[], high(_G1383), school(_G1385))),
 link([m], connection(12-13, an-[], school(_G1356), days(n))),
 link([m], connection(10-13, d-[m, c], their(_G1329), days(n))),
 link([m], connection(9-13, j-[p], since(_G1296), days(n))),
 link([m], connection(8-9, idys-[], ever(_G1266), since(_G1268))),
 link([m], connection(6-7, an-[], close(n), friends(n))),
 link([m], connection(5-7, o-[p, t], been(v), friends(n))),
 link([m], connection(5-9, mv-[s], been(v), since(_G1181))),
 link([m], connection(4-5, pp-[f], had(v), been(v))),
 link([m], connection(1-4, s-[s], mike(_G1119), had(v))),
 link([m], connection(0-1, w-[d], 'left-wall'(_G1089), mike(_G1091))),
 link([], connection(0-14, rw-[], 'left-wall'(_G1059), 'right-wall'(_G1061)))]
```

(그림 3) 예문에 대한 구문분석 결과: LGPI+

2.5 문장추상화의 예

(그림 4)는 예문 "Mike and Paul had been close friends ever since their high school days"에 대한 문장추상기 SABOT의 출력이다. Prolog로 구현한 문장추상기 ***SABOT***가 요점어 *mike*, *paul*, *been*, *friends*, 그리고 *ever since* 등을 선별해내었다. 문장표식 *sent(X,Y)*에서 *X*는 문단 내에서 위치를 그리고 *Y*는 절의 위치를 각각 나타낸다. 술어 *affect_state*와 *cue_phrase*는 각각 ***OfN***의 심상과 담화표지에 대응한다.

(그림 4)에서 SABOT는 주어진 예문에 대한 문장추상화 결과로 Mike, Paul, friends, been, ever since를 선별해 내었다. 이는 문장의 통사구조 분석을 통해 이루어진다. 또한, 각 어휘들이 가지는 의미정보를 OfN 범주로써 정하게 된다. (그림 4)의 friends는 심상(Affect State)에 해당하며, 그 대상은 등장인물 Mike와 Paul이다. friends의 의미적 정보는 사회적으로 마음이 통하는 관계임을 나타낸다. 주어진 문장에 대한 추상화 결과를 요약해 보면 "친구관계가 지속되



[sent(1, 1/2) : [affect_state ([friend, social, sympathetic]) : friends / (mike ↔ paul),
state ([identity, absolute, relation]) : been / (mike ↔ paul)],
sent(1, 2/2) : [cue_phrase ([temporal, durative]) : [ever, since] / (mike ↔ paul)]]

(그림 4) 추상화된 예문 : 문장추상기 SABOT의 출력

었다."임을 알 수 있다. 이러한 추상화 수준은 구문분석기의 단계(Level)를 조정함으로써 가능하다. 구문분석의 최상위(Root)는 좌벽(Left-Wall)에서부터 시작한다. 즉, 최상위와 가까울수록 추상화 수준이 높게 나타난다.

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 5) 추상화시킬 문단

[sent(1, 1/2) : [affect_state([friend, social, sympathetic]) :
friends / (mike ↔ paul),
state([identity, absolute, relation]):
been / (mike ↔ paul)],
sent(1, 2/2) : [cue_phrase([temporal, durative]) :
[ever, since] / (mike ↔ paul)]]

[sent(2, 1/2) : [affect_state([requirement, conceptional,
prospective]) : wanted / (paul ← mike),
delta(space) : [out, of, town] / (paul ← mike),
delta(time) : days / (paul ← mike),
cue_phrase([temporal, repetitive]) :
[but, now] / (paul ← mike)],
sent(2, 2/2) : [affect_state([wonder, contemplative]) :
surprise / (paul ← mike),
affect_state([giving, intersocial]) :
present / (paul ← mike),
delta(space) : patio / (paul ← mike),
delta(space) : backyard / (paul ← mike),
event([production, power, causation]) :
build / (paul ← mike),
cue_phrase([causal, specific, purpose]) :
[so, that] / (paul ← mike)]]

(그림 6) 추상화된 문단

일반적으로, 문단 안에서 각 문장은 다른 문장과의 상관관계에 따라 문장간 연결도(Connectivity Degree)를 가진다. 따라서 문장이나 절의 연결도는 문단 내에서 그것의 중요도를 반영한다.

문장추상기 ***SABOT***을 문단을 구성하고 있는 각 문장에 적용하여 추상화된 문단을 얻는다. Mike와 Paul에 대한 이야기[2]의 경우를 보자. (그림 5)에 추상화시킬 한 문단이 있다. ***SABOT***이 문단을 처리한 결과가 (그림 6)에 보인다.

3. 문장추상화 활용

문장추상화(Sentence Abstraction)는 문단기반의 중층적 원문이해 방법론 MIDTERM(Mid-Depth Text Reasoning Methodology)[2, 3]을 설계하는 중에 구체화되었다. 이 방법론은 다음과 같은 절차를 가지고 있다. ① 문단에 문장추상화를 실시하고 ② 추상화된 문장의 구성성분 중에서 문장간에 개연적(Abductive) 연결성을 가지는 구성성분 쌍을 확인한 뒤에 ③ 문장 중에서 개연적 연결정도가 상대적으로 높은 것을 문단의 화제문(Topic Sentence)으로 선정한다.

3.1 중층적 원문이해

문장추상화를 활용한 원문이해 방법은 계산 가능한 심층적 원문이해라고 생각할 수 있다. 그런 의미에서 중층적 원문이해(Mid-Depth Text Understanding)라고 할 수 있다. 심층적 원문이해는 사람이 글을 읽을 때 진행되는 것으로서 최소한 다음의 두 가지 사실을 내포한다. ① 문장이 기술하는 상황(Situation)을 파악하여 사건과 상태의 인과성(Causality)을 인식한다. ② 이러한 상황적 제약을 토대로 문장간 인과연쇄를 면밀하게 추적한다. 이에 본 논문에서는 원문(Text, 텍스트) 이해심도를 표층·중층·심층 등의 3단계로 구분한다. 표층적 원문이해는, 단어간의 직·간접적인 의미결속을 토대로 성취하는 피상적인 원문이해라고 간주한다.

한편, 문장추상화를 활용한 중층적 원문이해는 문장간 의미

결속을 설명하는 데 있어서 심층적 원문이해의 경우와 같이 철저하지는 않다. 언어학적 가용자원과 전산학적 도구의 제약하에서 심층적 원문이해를 지향한다는 것이 그 특징이다. 원문에 대한 중층적 이해는 문장추상화(Sentence Abstraction)와 개연규칙(Abductive Rules)을 통하여 성취한다. 그 후, 발췌요약을 위하여 원문의 화제문(Topic Sentence)을 선별한다. 여기에서 '중층적(Mid-Depth)'이라는 말은 요약대상이 되는 원문을 심층적(In-Depth)으로 이해할 수 있는 것은 아니나, 가용한 자연언어 처리도구나 자원을 적극적으로 활용하여, 가능한 깊이 원문을 이해할 수 있게 설계된 것이라는 뜻이다. 원문이해용 계산에 사용하는 자원과 도구는 다음과 같다. 온톨러지 OfN, 견실한 구문분석기(Robust Parser) LGPI+, 문장추상기(Sentence Abstracter) SABOT, 그리고 개연사슬기(Abductive Chainer) SICHA 등.

(그림 1)은 중층적 원문이해 방법론 MIDTERM(Mid-Depth Text Reasoning Methodology)을 구현한 시스템의 기능적 구조이다. 시스템은 대용어 해소기(Anaphora Resolver)를 중심으로 문장내(Intrasentential) 처리와 문장간(Intersentential) 처리로 나뉜다. 그 구성요소는 다음과 같다. 온톨러지(Ontology) 및 심중사전(Lexicon), 구문분석기(Syntactic Parser), 문장추상기(Sentence Abstractor), 대용어 해소기(Anaphora Resolver), 화제문 선정기(Topic Sentence Selector). 이 중에서 심중사전은 Prolog 데이터베이스(Database)로 구현한 것으로, 공개된 Roget 시소러스[4]를 변환하여 얻었다. 구문분석기는 Link Grammar Parser [23]의 원시 프로그램(Source Program)을 개작한 것이다. 마지막으로, 대용어 해소기는 선행사 구조[16]에 관련된 알고리즘의 구현이다.

(그림 1)의 문장내 처리에서는 문단을 구성하는 문장의 구문분석 정보와 온톨러지 OfN 정보를 활용하여 문장추상화를 수행한다. 그 후, 문장간 처리에서는 문단의 주제문장을 선정한다. 이를 위하여, 문장추상기 SABOT와 개연규칙(Abductive Rules)을 적용하여 문단을 구성하는 문장간 개연적 연결상황을 파악한 후, 개연사슬기 SICHA를 통하여 문장간 연결도가 높은 문단의 화제문들을 선택한다.

3.2 개연사슬 형성

주어진 글의 화제와 관련이 깊은 문장을 선택하는 과정에서 어휘사슬(Lexical Chain)[10]을 활용할 수 있다. 그러나 명사류 유의어의 연쇄인 어휘사슬로써는 글에서 반복되는 개념이나 내용전개 구조를 파악할 수 있을 뿐, 문장간의 의미적 연관성을 포착할 수는 없다. 이에 종래의 어휘사슬의 기능을 포함하는 한편, 줄거리 단위, 단서구 용법, 문장 사이의 개연성 등을 감지할 수 있는 새로운 어휘사슬을 정의하였다. 이것을 본 논문에서는 개연사슬(Abductive Chain)이라고 부른다. 개연사슬은 개연고리(Abductive Link)로 이루어져 있다. 개연고리는 개연규칙(Abductive Rule)의 용례(Instance)이다. 글에서 개연사슬을 형성하는 것은 글의 내용을 이해하는 과정에 대응한다. 개연규칙은 문장간 구성성분들의 개연적인 의미 결속성을 나타내며 문장 구성성분들이 가지는 OfN 정보로써 표현된다. 2항 또는 3항인 개연규칙의 일반적인 모습은 다음과 같다.

Ante <= Post [{=+>, =->, =*>} Cons]

여기에서 ① Ante, Post, Cons는 pred(args)의 형태를 가진다, ② pred(args)는 OfN에 명시된 개념으로 7가지의 범주로 표현된다, ③ =+> and =->는 Post와 Ante에 제한사항이 있음을 나타내고, ④ =*>는 담화표지가 있음을 나타낸다. (그림 7)은 Mike와 Paul의 이야기[2]를 처리할 때 사용한 개연규칙의 예이다.

```
① % 마음이 통하면 주고 싶어진다
   affection(sympathetic) <= affection(offer)
② % 풀이 죽으면 소극적이 된다
   event(inactivity) <= event(descent).
③ % 장소를 바꾸고 싶을 때 여행을 한다
   delta(space) <= event(journey) =-> affection(prospective).
④ % 싫은 것을 권하면 흥미를 끌지 못한다
   affection(advice) <= affection(cause(pleasure)) =*>
   cue_ phrase(adversative)
```

(그림 7) 개연규칙

개연규칙은 크게 2항 규칙과 3항 규칙으로 나눌 수 있다. (그림 7)에서 규칙 ①과 규칙 ②는 2항 규칙, 규칙 ③과 규칙 ④는 3항 규칙이다. 이러한 개연규칙들은 문장구성성분들간 인과관계를 나타낸다. 개연규칙은 문장추상화 과정에서 문단의 주제가 될 수 있는 화제문장들을 선정하기 위해 적용하는 유용한 언어학적 도구이다.

문장추상화와 개연규칙을 적용하여 문단의 주제와 가장 연접한 화제문을 선정하기 위하여 Prolog로 구현한 개연사슬기 SICHA(A Situation Chainer)를 활용하였다. SICHA는 문단을 구성하는 문장간 연결집중도를 보여준다. 문장들의 연결집중도는 개연규칙에 부합하는 연결유형의 급수(Degree)로써 나타낼 수 있다. 연결집중도의 일반적인 모습은 다음과 같다.

$$SDs = \{D_1\text{-sent}(N_1), \cdots D_n\text{-sent}(N_n)\}$$

여기에서 SDs는 문장의 연결집중도(Sentence Degrees)를 나타내며 D는 그 문장의 급수(Degree)를, sent(Nn)는 문단내에서 각 문장의 순서(Sentence Number)를 나타낸다. 개연규칙에 의해 한 문장이 다른 문장과 연결되는 경우에 Degree를 1로 한다. (그림 8)은 SICHA를 통해 얻은 (그림 5)에서 보인 예문에 대한 문장간 연결 집중도를 나타낸다.

```
SDs = [1-sent(8), 3-sent(10), 3-sent(12), 4-sent(4),
       4-sent(6), 4-sent(15), 5-sent(1), 5-sent(3),
       5-sent(11), 6-sent(14)] ;
```

(그림 8) 문단내 문장들의 연결 집중도

SICHA는 (그림 8)과 같이 문단내 문장들의 연결집중도를 오름차순으로 정렬하여 보여준다. (그림 8)의 경우에는, 문단의 14번째 문장의 연결집중도가 6으로 가장 높게 나타났다. SICHA의 결과에서는 문단의 주제를 내포하고 있는 문장들의 연결집중도가 높게 나타난다. 본 실험에서는 각 문단별로 미리 그 문단의 주제와 연접하는 화제문들을 준거요약문으로 선정한다. 그 후 SICHA를 통해 얻은 화제문들을 준거요약문과 비교하여 문단에 대한 문장재현율(Sentence Recall)과 주제관련성(Topic Hit Ratio)을 구하였다.

문단에 대한 문장재현율은 사람이 선택한 중요 문장을 문장추상화와 개연규칙을 적용했을 때 화제문 선별기가 재현해 내는 비율이다. 문단에 대한 주제관련성은 재현해 낸 문장이 의미적으로 문단의 주제와 직접 관련이 있는지를 나타내는 비율이다. 여기에서 미리 정한 준거요약문의 개수와 상위의 연결집중도를 가지는 문장의 개수를 일치시킴으로써 비교 대상의 범위를 정하였다. 본 논문에서는 중요 문장 선정의 객관성 확보를 위하여 ① 준거요약문의 설정을 Dear Abby 상담 이야기[22]의 제목에 근거하였고, ② 참여한 실험자들(박사과정생 3명, 석사과정생 3명, 학부생 1명)이 개별적으로 선택한 관건 문장에 대해 각자 재평가를 한 후, ③ 다수결로 준거요약문을 선정하였다.

(그림 8)의 연결집중도에서, 미리 설정한 준거요약문들이 1, 4, 11, 14번째 문장이라고 했을 때 비교 범위는 연결집중도가 높은 상위 4개의 문장이 되며, 문장재현율은 다음과 같다.

$$\textit{Sentence Recall} = \{ 1, 3, 11, 14 \} / \{ 1, 4, 11, 14 \} = 0.75$$

여기에서 3번 문장이 준거요약문에는 해당되지 않지만 문단의 주제와 의미적으로 직접 관계가 있을 경우, 주제관련성에 포함시키게 된다.

4. 실험결과 및 평가

실험을 통해 선정된 화제문을 사람이 선택한 관건문장과 비교하고 그 결과를 <표 2>에 보였다. <표 2>에서 '자신규칙'은 실험대상의 상담문에서 유도해낸 개연규칙들이다. '전체규칙'은 각 상담문으로부터 얻은 모든 개연규칙들을 말한다. 23개 이야기에 있는 58개 문단의 경우 문장재현율(Sentence Recall)은 최소 54%, 최대 72% 정도이었고, 주제관련성(Topic Hit Ratio)은 최소 76%, 최대 85%로 나타났다. 최소값은 실험대상의 이야기에서 수집한 개연규칙을 제외하고 개연사슬을 계산한 경우이다. 한편, 최대값은 수집된 전체 개연규칙을 적용하였을 때 나타난 결과이다.

<표 1> 실험 대상 이야기의 문단별 문장 구성표

이야기	문단 개수	문단의 문장개수	문단의 단어개수
23개	58개	568개 (문단별 평균 : 10개)	5010개 (문단별 평균 : 86개)

<표 2> MIDTERM 성능평가

구 분	문장재현율 (Sentence Recall)	주제관련성 (Topic Hit Ratio)
1. 자신규칙만 적용	66%	86%
2. 자신규칙 제외	54%	76%
3. 전체규칙 적용	72%	85%

실험 결과, 실험대상의 이야기로부터 수집한 개연규칙(자신규칙)만을 적용하여 얻은 문장재현율은 66%, 주제관련성(Topic Hit Ratio)은 86%로 나타났다. 자신규칙만을 적용하였으므로, 문장재현율을 100% 기대하였으나, 실험 결과는 그 기대치에 부족하다. 이러한 결과의 원인을 분석해 보면 다음과 같다. ① 각 문장에 대한 구문분석과 문장추상화 과정에서 개연성 규칙에 적용된 단어들이 문장내 주요어로 선택되지 않을 수 있다. 문장의 주요구성성분은 주어, 목적어, 술어에 해당되는 단어나 구이며, 전치사구나 부사구에 해당되는 단어들은 추상화과정에서 생략될 수 있기 때문이다. ② 개연규칙들이 추상화된 문장들과 대등하게 연결되지 않는 경우가 발생하는데, 이는 개연규칙을 생성하는 과정에서 범할 수 있는 오류로 추정할 수 있다. 이러한 문제점들을 최소화하기 위해서는 개연규칙의 생성과정에서 문장의 주요 구성성분들을 그 대상으로 선정해야 할 필요가 있다.

문장추상화를 활용한 원문이해 방법 MIDTERM은 어휘사슬 및 수사구조를 활용한 방법론의 장점을 취하고 개량과 접목을 시도한 것이다. 따라서 이 시스템의 성능을 다른 것보다도 이들 두 방법론을 구현한 시스템의 결과와 비교하는 것이 의미가 있을 것이다. 그러나 세 가지 방법론의 성능평가 결과를 해석하는데 있어서 다음과 같은 점을 감안해야 한다. ① 실험에 사용한 말뭉치(Corpus)의 유형이 다르다. ② MIDTERM의 경우와는 달리 다른 두 방법론의 결과에는 재현문장 등급의 표시가 없다. ③ 준거요약문을 설정할 때 사람의 주관성을 완전히 제거하지는 못한다.

<표 3>에 성능비교 내용을 보였다. 이야기의 표제와 의미적으로 일치하는 화제문 재현율은 다른 두 시스템의 결과와 비교하여 최소 9%, 최대 19% 정도의 성능향상이 있었다.

MIDTERM 성능평가에는 다른 시스템과는 달리 말뭉치로 사용한 이야기에 표제가 있어서 시스템이 선정한 화제문과 자구적인 일치도 평가해 볼 수 있었다. 그 결과는 시스템이 선정한 화제문의 최소 55%, 최대 65% 정도가 자구적으로 일치한다는 결과를 보였다.

〈표 3〉 MIDTERM 성능평가 비교

방법론 / 평가척도	어휘사슬[10] (Lexical Chain)	담화분석[9] (Rhetorical Analysis)	중층적 원문이해 (MIDTERM)
화제문 재현 (Topic Sentence Recall)	67%	66%	55%~65% (Sentence Recall) 76%~85% (Topic Hit Ratio)

MIDTERM 시스템의 성능평가 결과는 방법론과 시스템의 장래를 낙관적으로 보이게 한다. 그 이유는 ① OfN은 개방형이어서 새로운 개연성 규칙을 쉽게 추가할 수 있다. ② 개연규칙에는 담화기능을 추가하여 체계적으로 정교하게 만들 수 있다. 이렇게 하면 추상화된 문장 사이에서 조응하는 문장 구성성분들을 보다 세밀하게 감지할 수 있고, 실험 결과의 정확도와 신뢰도를 더욱 높일 수 있다.

5. 결 론

문서요약 자동화는, 원문에 대한 심층적인 이해(In-Depth Understanding)과정의 실현을 통해서 이상적인 결과를 얻을 수 있다. 그러나 현실적으로는 심층이해에 대한 차선책으로 원문의 글 유형이나 독서관련 경험적인 지식을 활용하여 통계·언어학(Statistical Linguistics)적인 방법으로 자동요약 문제에 접근하기도 한다. 이렇게 하는 이유는 현재의 자연어 처리 이론과 기술수준 그리고 축적된 기계가용 지식자원의 완전성 수준에서 볼 때, 원문에 대한 심층적 이해의 성취는 응용영역을 제한하지 않는 한, 단기적으로 불가능하기 때문이다. 그럼에도 불구하고 문제의 본질상 깊은 원문이해가 필요한 문서요약 자동화 분야의 경우, 원문에 대한 표층적 이해(Shallow Understanding) 수준에 머무는 방법론에만 안주할 수 없다.

본 논문에서는 표층적 원문이해의 한계를 극복하여, 심층적 원문이해를 도모하고자 문장교열방법으로 문장추상화를 활용하였다. 문장추상화는 개념추상화를 도입한 문장교열 작업이다. 이를 통하여 요약과정에서 형태 뿐만 아니라 의미적인 접근을 시도할 수 있다. 추상화된 문장들의 주요 구성성분들을 대상으로 문장간 연결상황을 확인하고, 그 연결도가 높은 문장을 화제와 밀접한 관계가 있는 문장으로 선택함으로써 문단추상화도 가능함을 알 수 있었다. 이러한 과정에서 존재론 *OfN*과 실용적인 구문분석기 *LGPI+*, 그리고 문장추상기 *SABOT*, 문단 화제문을 선정하는 개연사슬기 *SICHA* 등을 활용하였다.

문장추상화를 활용한 원문이해 실험에서, 선별된 문단의 주제문을 사람이 선택한 관건문장(Key Sentence)과 비교하였다. 적용한 23개 이야기의 58개 문단의 경우, 주요문장의 재현율이 약 72%, 그리고 선별된 문장의 주제관련성이 약 85% 정도임을 확인할 수 있었다. 이는 유사 시스템의 성능보다 약 10%에서 20% 정도의 성능향상을 보이는 것이다. 본 논문을 통하여 개념추상화를 도입한 문장교열방법으로 문장추상화를 보였다. 또한, 문장추상화를 활용한 원문이해 방법은 계산가능한 심층적 원문이해를 지향할 수 있으며, 원문에 대한 이해심도를 높일 수 있는 실용적인 문서 요약의 도구로 이용할 수 있음을 알았다.

참 고 문 헌

[1] 김곤, 김민찬, 배재학, 이종혁, "ISAAC : 문장분석용 통합시스템 및 사용자 인터페이스", 정보처리학회논문지B, 제11-B권 제1호, pp.107-116, 2004.

[2] Bae, J. -H. J. and Lee, J.-H., "Topic Sentence Selection with Mid-Depth Understanding," Proc. of ICCPOL, pp. 199-204, 2001.

[3] Bae, J. -H. J. and Lee, J. -H., "Mid-Depth Text Understanding by Abductive Chains for Topic Sentence Selection," IJCPOL, Vol.15, No.3, pp.341-357, 2002.

[4] Roget's Thesaurus, http://promo.net/cgi-promo/pg/t9.cgi?ftp://ibiblio.org/pub/docsftp://ibiblio.org/pub/docs/ books/gutenberg/.

[5] 양재군, 배재학, "온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우", 정보처리학회 춘계학술발표대회 논문집, 제9권 제1호, pp.515-518, 2002.

[6] SWI-Prolog, http://www.swi-prolog.org/.

[7] 배재학, "언어학적인 방법론을 취하는 자동 문서요약에 대한 연구", 울산대학교, 공학연구논문집, 제29권 제2호, pp. 351-363, 1998.

[8] Mani, I., 'Automatic Summarization', John Benjamins Publishing Company, 2001.

[9] Marcu, D., "From Discourse Structures to Text Summaries," in *Proc. ACL'97 and EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, pp. 82-88, Jul., 1997.

[10] Barzilay, R. and Elhadad, M., "Using Lexical Chains for Text Summarization," in *Proc. ISTS'97 (The Intelligent Scalable Text Summarization Workshop, ACL)*, Madrid, Spain, pp.10-17, Jul., 1997.

[11] Proper Names Wordlist. http://clr.nmsu.edu/cgi-bin/Tools/

CLR/clrcat#I4.

[12] C. Fellbaum (ed.), WordNet : An Electronic Lexical Database, MIT Press, 1998.

[13] 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.

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

[15] Luhn, H. P., "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, Vol.2, No.2, pp.159-165, 1993.

[16] Huls, C., Bos, E. and Claassen, W., "Automatic Referent Resolution of Deictic and Anaphoric Expressions," *Computational Linguistics*, Vol.21, No.1, pp.59-79, 1995.

[17] Jing, H., "Sentence Reduction for Automatic Text Summarization," In *Proceedings of The 6th Applied Natural Language Processing Conference and the First Meetion of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL'2000)*, pp. 310-315, 1999.

[18] Grefenstette, G., "Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the blind," In *Working Notes of the Workshop on Intelligent Text Summarization*, pp.111-117, 1998.

[19] Knight, K. and Marcu, D., "Statistics-Based Summarization - Step One : Sentence Compression," *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, Outstanding Paper Award, Austin, Texas, July-August, 2000.

[20] Cremmins, E. T., '*The Art of Abstracting*,' Information Resources Press, 1996.

[21] Robin, J., 'Revision-based generation of natural language summaries providing historical background : corpus-based analysis, design and implementation,' Ph.D. Thesis, Columbia University, 1994.

[22] DearAbby, http://www.dearabby.com/.

[23] Link Grammar, http://www.link.cs.cmu.edu/link/.

김 곤

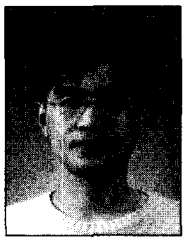
e-mail : gonkim@mail.ulsan.ac.kr

1997년 울산대학교 전자계산학과(공학사)

2000년 울산대학교 대학원 컴퓨터·정보통신공학부(공학석사)

2002년~현재 울산대학교 대학원 컴퓨터·정보통신공학부 박사과정

관심분야 : 자동프로그래밍, 인공지, 문서요약, 전자상거래



양 재 군

e-mail : jgyang@mail.ulsan.ac.kr

1997년 울산대학교 공과대학 건축학과(공학사)

2001년 울산대학교 정보통신대학원 정보통신공학전공(석사)

2001년~현재 울산대학교 대학원 컴퓨터·정보통신공학부 박사과정

관심분야 : 자동프로그래밍, 인공지능, 온톨러지, 문서요약

배 재 학

e-mail : jhjbae@ulsan.ac.kr

1981년 중앙대학교 전자계산학과(이학사)

1983년 한국과학기술원 전산학과(공학석사)

2003년 포항공과대학교 컴퓨터공학과(공학박사)

1985년~현재 울산대학교 컴퓨터·정보통신공학부 교수

관심분야 : 자동문서요약, (자동, 논리)프로그래밍, 지식경영 및 기술, 전략경영정보시스템, 교육인적자원정보시스템

이 종 혁

e-mail : jhlee@postech.ac.kr

1980년 서울대학교 수학교육과(이학사)

1982년 한국과학기술원 전산학과(공학석사)

1988년 한국과학기술원 컴퓨터공학과(공학박사)

1991년~현재 포항공과대학교 컴퓨터공학과 교수

관심분야 : 자연어처리, 한국어정보처리, 한중일영 기계번역, 교차언어 정보검색, 문서요약, 문서분류