한글 문헌 자동축약 시스템에 관한 연구

김 세 중 / 조 성 호
(한 국 경 제 신 문)

The Study of Automatic Extracting Systim on Korean Fulltext

Kim Se Jung / Cho Sung Ho
(The Korea Economic Daily)

요 약

본 연구는 한글 문헌을 컴퓨터를 이용하여 축약하는 시스템 구축에 관한 연구로서, 기존의 '완전 자동축약'에 따른 축약문 생성의 편협성을 해결하기 위하여 '자동축약 + 후통제 처리'라는 절충형 시스템 관리 형태로 실제 실현 가능한 시스템을 설계한다는데 그 큰 목적이 있다. 대상 문헌에 대한 구체적 적용 문법은 언어학적 문법 이론인 '격문법 이론'과 '성분 이론'을 그 핵심으로 이용하여 문장을 '의미 있는 어절' 단위로 추출, 해당 문헌을 축약하는 방법을 택하였다.

Ⅰ. 들어가는 말 (연구 목적 및 방법)

본 연구는 언어학자인 C.J.Filmore의 'A Case for Case'라는 논문을 통해 발전된 언어학적 문법 이론인 '격문법 이론' 및 이와 깊은 관련이 있는 '성분 이론'을 그 핵으로 이용하여, 문장의 구조적 짜임새를 분석, 원문에서 '가장 의미 있는 어절'만을 자동으로 추출(=발췌)하여 원문을 축약하는데 그 목적이 있다.

격문법 이론은 문장에서 가장 중심이 되는 성분을 '서술어'로 보고 명사들이 어떤 '일정한 격'에 의해 서술어를 중심으로 기본적인 하나의 문장을 이룬다는 이론이다.

본 연구는 격문법의 이러한 특성에 착안을 하여 우리말의 개별 서술어에 기본적(=필수적)으로 따라붙는 기본격(또는, 기본성분, =필수격)을 미리 설정('기본문형 사전' 설정)해 놓은 후에 , 해당 서술어가 본문에 나타나면 이를 중심으로 문장을 분석, 해당 문장을 자동으로 축약하는 방법을 사용한다.

그후 적절하지 못한 축약문은 원문과 대조하며 제거하거나 첨가하는 후통제 방식을 적용, 축약문을 완성한다.

본 자동 축약 시스템은 이와 같이 ① 자동축약 시스템과 ② 자동축약 후통제 처리 시스템으로 만들어진다.

자동색인이 그러하듯 완전자동에 따른 비정확성을 줄이기 위해, 인력과 시간 투입이라는 제한적인 요소가 있음에도 불구하고, '자동축약 + 후통제 처리'의 방식을 채택하였다. 그러나 후통제 처리의 부담을 최소화하는데 역점을 두었다.

Ⅱ. 개요

하나의 문장은 문장 내에서 각기 다른 역할을 하는 구성요소인 성분들로 구성된다.

문장의 성분에는 주성분과 그에 딸린 부속성분이 있다. 주성분은 문장성립에 필수적인 것으로 그것이 빠지면 불완전한 문장이 된다.

1. 바람이 분다.
2. 아이들이 공을 던진다.
3. 물이 어름이 된다.
4. 철수가 영희에게 빵을 준다.

위에서 밑줄 그은 부분(어절)들은 모두 해당 서술어들에 없어서는 안 될 말들이다. 이들 중 어느 것이 하나라도 빠진다면 불완전한 문장이 된다. 이처럼 서술어를 중심으로 해당 서술어에 반드시 따라와야만 문장이 성립되는 구성성분들을 문장의 필수성분 또는 기본성분이라 한다.

5. 그가 새 옷을 몽땅 집어갔다.

밑줄 그은 말, '새, 몽땅'은 각각 목적어 '옷'과 서술어 '집어갔다'를 꾸며서 뜻을 더해 주는 말인데 이들이 없어도 위의 5는 '그가 옷을 집어갔다'와 같이 온전한 문장이 된다. 이들은 이렇게 문장의 뼈대를 이루는데 아무 기여를 하지 못하고 다른 성분에 딸려 있는 까닭에 부속성분이라 한다.

우리말 문장에서 부속성분은 보통 관형어와 부사어를 말하기는 하지만 이들은 해당 서술어에 따라 문장에서 반드시 필요한 필수성분으로 나타날 때도 있다.

본 축약 시스템은 이상에서 살펴본 우리말 문장의 구성요소인 성분 분류를 중심으로 이루어진다. 서술어의 종류에 따라 필수적으로 따라붙는 성분들은 추출하고, 없어도 되는 부속성분들은 제외하면서 원문을 자동으로 축약하려고 한다.

해당 서술어의 필수성분을 추출하기 위해 우리는 우리말의 문장을 다음과 같이 9개의 문형으로 분류하여 시스템에 적용한다.

Ⅲ. 자동축약을 위한 우리말의 기본문형 설정

1문형

주어	+	서술어
이/가/은/는/에서/께서		일어나다, 터지다, 발생하다, 있다, 우수하다 네모지다, 둥글다, 내리다, 희다...

2문형

주어 +	(직접)목적어 +	서술어
이/가/은/는/에서/께서	을/를	생각하다, 부수다, 찌르다, 지나다, 통과하다, 두려워하다, 먹다...

3문형

주어 +	(간접)목적어 +	서술어
이/가/은/는/에서/께서	와/과	만나다, 싸우다, 결혼하다, 충돌하다...

4문형

(간접)목적어 +	주어 +	서술어
에게(는)	이/가/은/는/에서/께서	무섭다, 즐겁다, 두렵다, 잡히다, 들리다, 먹히다, 좋다...

5문형

주어 +	보어 +	서술어
이/가/은/는/에서/께서	이/가	아니다, 되다, 내려오다, 튀어나오다...

6문형

주어 +	위치어 +	서술어
이/가/은/는/에서/께서	에	있다, 걸리다, 당다, 면하다, 접촉하다. 부딪히다, 충돌하다,...

7문형

주어 +	위치어 +	부사어 +	서술어
이/가/은/는/에서/께서	에서	(으)로	바뀌다, 변하다, 떠나다, 출발하다, 가다, 도착하다...

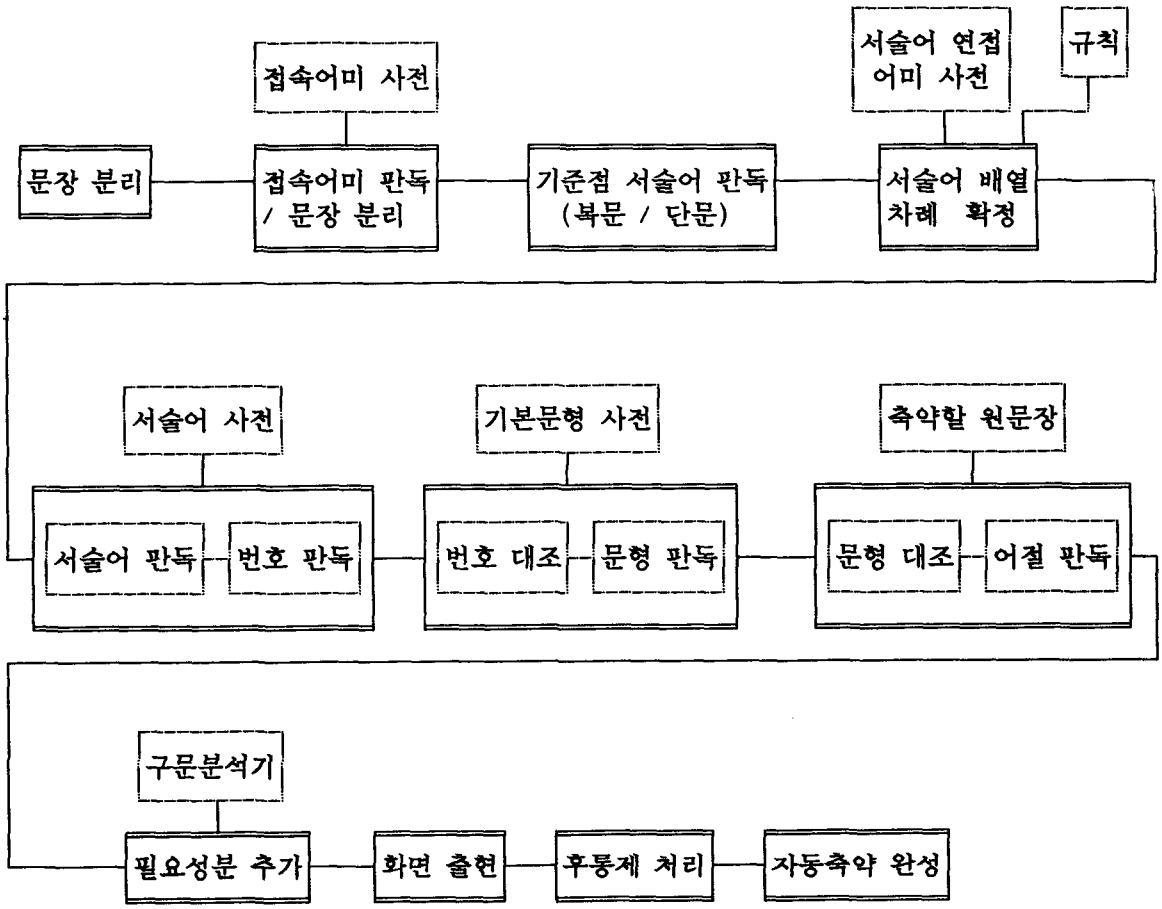
8문형

주어 +	(직접)목적어 +	부사어 +	서술어
이/가/은/는/에서/께서	을/를	(으)로	고치다, 수리하다, 취급하다, 간주하다...

9문형

주어 +	위치어 +	(직접)목적어 +	서술어
이/가/은/는/에서/께서	에(게)	을/라고	주다, 받다, 팔다, 말하다, 명령하다, 설명하다...

Ⅳ. 시스템 구성 및 흐름도



Ⅴ. 시스템 구성요소들의 각 기능과 특성

(과정1) 문장 분리 : 하나의 기사는 각각의 문장들로 이루어져 있다. 기사 본문을 문장 단위로 축약하기 위해서 마침표(.)를 인식, 기사를 문장 단위로 분리한다.

(과정2) 접속어미 사전 통과 / 어미 판독 : 두 개 이상의 주어, 서술어가 접속어미로 연결되어 있는 문장의 경우, 이들의 호응 관계를 알기 위하여 접속어미를 중심으로 문장을 분리한다. 분리 위치는 <접속어미 + 비어미어절>이다.

(보기) 새가 하늘을 날고 물고기가 물 위에서 뛴다.

<—접속어미 사전과 대조

<————————접속어미 '-고' 발견

<————비어미어절 '물고기가' 판독

(과정3) 기준점 서술어 판독 : 두 개 이상의 문장이 연결되어 하나의 문장을 이룬 복합문의 경우, (과정2)에서 접속어미를 중심으로 문장이 분리된다. 어미가 있는 서술어를 중심으로 이렇게 문장을 분리하는 것은 각각의 서술어에 따르는 해당 성분 요소들이 서로 다르기 때문이다. 문장들 중에는 이러한 서술어들이 여러 개 겹쳐져 있는 경우가 있다. 그런데 이들은, 문장과 문장이 대등하게 이어져 있는 복합문이든, 하나의 문장이 다른 여러 문장들을 안고 있는 문장이든 일정한 순서에 의해 연결돼 있다. 이들을 순서대로 분리, 해석해야만 문장의 본뜻을 이해할 수 있게 된다. 그러므로 컴퓨터가 이들의 순서를 파악하는 일은 매우 중요하다. 본 (과정3)은 (과정4)에서 하게 될 서술어의 배열 차례를 확정하기 위하여 규칙 적용을 시작하는 '기준점'을 잡기 위한 것이다.

<기준점 서술어>는 단문의 경우, 문장 맨끝의 마침점 바로 앞에 있는 서술어이고, 복합문의 경우는 이미 (과정3)에서 접속어미에 의해 분리된 개개 문장들 중의 맨끝에 있는 서술어가 된다.

<복합문의 경우>

(보기) 다람쥐가 달아나면 / 새들도 덩달아 하늘 높이 날아간다.

기준점 서술어 ↑ 기준점 서술어 ↑

이 (보기)에서 '기준점 서술어'는 '달아나다'와 '날아가다' 둘로 판독된다. 그 까닭은, 이 문장이 (과정2)에서 <접속어미 어절 + 비어미어절> 부분인 <달아나면 + 새들도>를 판독하고 접속어미 '-면'이 있는 어절 '달아나면'을 중심으로 둘로 나누어졌기 때문이다.

<기준점 서술어> 판독은 나누어진 문장들에서 맨끝의 어절이 되기 때문이다.

<단문의 경우>

(보기) 마침 계곡 저쪽에서 다람쥐 두 마리가 뛰어왔다.

기준점 서술어 ↑

이 (보기)에서 서술어는 '뛰어오다' 하나로 판독된다.

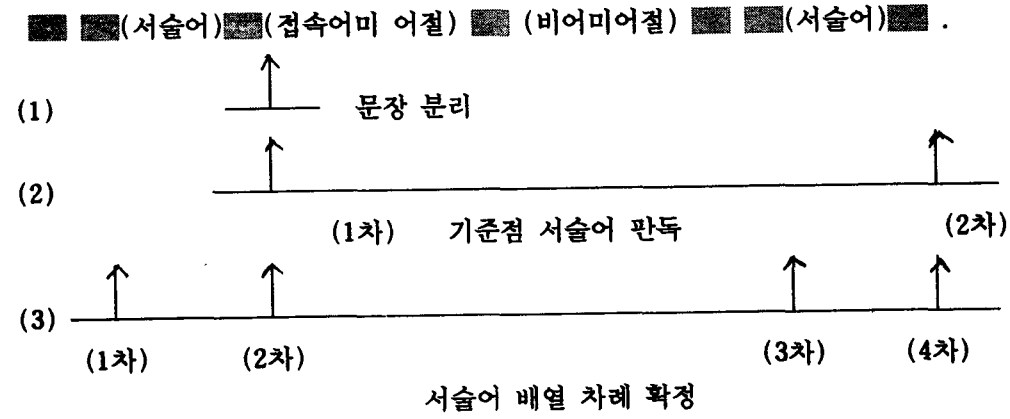
(과정4) 서술어 배열 차례 확정 : 본 처리 과정은 잇따라 쓰여진 서술어의 배열 차례를 확정해 내므로서 각 서술어가 반드시 필요로 하는 '필수격 어절', 즉 〈기본문형〉을 찾아내도록 하기 위한 것이다.

두 개 이상의 서술어들이 잇따라 이어져 있는 경우 이들 중 마침표에서 가장 먼 곳에 있는 서술어를 1차 서술어로, 1차 서술어 다음에 마침표쪽으로 바로 이어오는 서술어를 2차 서술어로, 또 3차 서술어로 한다. 이 과정은 〈어미사전〉을 문장의 각 어절 뒷부분과 대조하여 결정한다. 여기서의 서술어는 '관형형 어미'가 붙은 것은 제외된다.

〈원칙 1〉 복합문의 경우

< '접속어미 + 비어미어절'에 의해 문장을 분리, 서술어 확정 > : 복문

'접속어미 + 비어미어절'에 의해 문장을 분리한다. 그 다음 분리된 문장들 중마침표에서 먼 쪽에 있는 문장부터 어절 뒷부분을 비교, '서술어 연접 어미'가 있는 어절을 모두 판독, '기준점 서술어'에서 먼 쪽의 '연접 어미 어절'부터 배열 차례를 정한다. 기준점 서술어는 절단된 문장의 맨끝 어절이다.

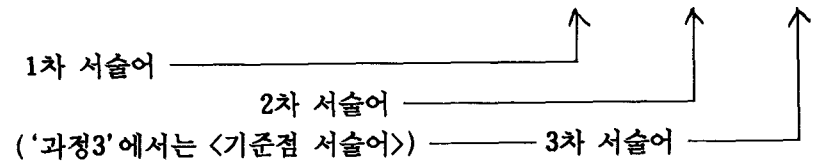


〈그림〉 '과정1'부터 '과정4'까지의 종합 설명도

(원문 보기) 수친다 총리는 국왕 알현 후 잠롱과 함께 관영 TV에 나와 잠롱을 포함한 나흘간의 시위 도중 체포된 모든 사람들을 즉각 석방할 것이라고 밝혔고 / 잠롱도 국민들에게 시위를 자제해달라고 촉구했다.

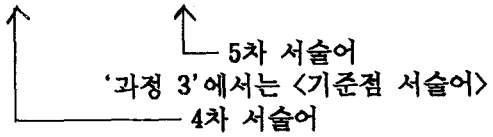
[적용 1] (접속어미 문장)

수친다 총리는 국왕 알현후 잠롱과 함께 관영 TV에 나와 잠롱을 포함한 나흘간의 시위 도중 체포된 모든 사람들을 즉각 석방할 것이라고 밝혔고 /



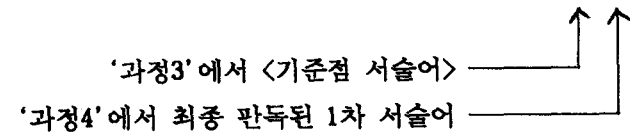
[적용 2] (마침표 문장)

잠롱도 국민들에게 시위를 자제해달라고 촉구했다.



만약 다음 문장과 같이 <기준점 서술어> 바로 앞에 있는 어절을 <서술어 연접어미 사전>과 비교한 결과 다른 어미가 발견되지 않으면 처음의 <기준점 서술어> 하나만이 이 문장의 서술어로 판독된다.

(보기) 시위를 벌였던 군중들도 자진 해산, 방콕시는 평온을 되찾았으며 /



〈원칙2〉 단문의 경우

< 문장의 맨끝에 마침표를 중심으로 서술어 확정 > : 단문

마침표 앞에 있는 <기준점 서술어>를 중심으로 그 앞쪽으로 <서술어 연접어미 사전>과 대조, 배열 차례를 확정한다.

(보기) 태국의 치안당국은 방화와 기물파괴 등 폭동을 일으킨 불량배들을 제외한 일반 시민들은 이날 모두 귀가시켰다고 밝혔다.

1차 서술어 ——↑
'과정 2'에서는 〈기준점 서술어〉 —— 2차 서술어 ——↑

이상과 같이 서술어를 분리, 찾아내면 문장을 자동으로 축약할 수 있는 가장 기본적인 작업이 완료된 것이다. 왜냐하면 서술어에 따른 문장의 필수적인 기본구조를 알아낼 수 있어 격조사가 붙은 어절을 판독, 이들만을 찾아 문장의 기본구조를 완성(제 7 과정)시킬 수 있기 때문이다. 이후 문장의 필요성분을 추가하면(제 8 과정) 원 문장을 자동축약하는 시스템 모델이 완성된다.

(과정5) 서술어 판독 / 해당번호 판독 : 앞 과정에서 확정된 순서 중 1차 서술어부터 서술어 사전을 판독, 같은 형태의 서술어를 읽은 뒤 해당 번호를 판독한다.

(과정6) 기본문형 사전 번호 대조 / 문형 판독 : 판독된 서술어의 번호를 〈기본문형 사전〉에 있는 번호와 대조하여 같은 번호를 판독한 뒤, 해당 문형을 읽어들인다.

(표) 서술어 사전과 기본문형 사전의 DB 구성도 및 도치화일로의 연결 관계

서 술 어 사 전

서 술 어	품사	번호
도착하다	자	7
들리다	자	4
말하다	타	9
바뀌다	자	7
발생하다	자	1
생각하다	타	2
설명하다	타	9
수리하다	타	5
아니다	자	8
있다	자	6
충돌하다	자	3
통과하다	타	2
.		
.		

기 본 문 형 사 전

번호	기 본 문 형
1	--이/가/은/는 ____
2	--이/가/은/는 --을/를 ____
3	--이/가/은/는 --과/와 ____
4	--에게 --이/가 ____
5	--이/가/는/은 --을/를 --(으)로 ____
6	--이/가/는/은 --에/(으)로 ____
7	--이/가/는/은 (--에서) --이/(으)로 ____
8	--이/가/는/은 --이/가 ____
9	--이/가/는/은 --에(게) --을/라고 ____

(과정7) 축약할 원문장과 문형 대조 / 어절 판독 : (과정6)에서 판독된 기본문형의 각 조사들을, 축약할 원문장과 대조하여 원문장에서 추출할 기본문형의 어절을 판독한다.

(과정8) 필요성분 추가 : '기본문형'만으로는 의미 해석에 문제가 있는 경우가 있다. 그러므로 <구문분석기>에 있는 규칙을 이용하여 만족할만한 축약문장이 되도록 필요성분을 추가한다.

진행되는 과정이 본 (과정8)에서 적용받는 규칙
규칙1) 서술어 배열 차례에서 정해진 서술어의 수가 두 개 이상이면 제1차 서술어가 적용된 뒤 다시 제2차 서술어가 서술어 판독을 거쳐 필요성분 추가까지 적용된 뒤 다음 과정으로 이동한다. 규칙2) 하나의 문장이 (과정2)인 '접속어미 사전 통과 / 어미 판독'에서부터 (과정8)인 '필요성분 추가'까지 거쳤으면 곧이어 다음 문장을 (과정2)부터 (과정8)까지 문장 단위로 <되풀이 적용>을 한 뒤, 이들을 원문장 순서대로 화면에 출현시킨다.

(과정9) 화면 출현 : 화면 상단에 <원문>, 화면 하단에 <축약문> 출현

(과정10) 후통제 처리 : 화면의 상단과 하단에 출현한 <원문>과 <축약문>을 보고 <축약문>을 정리, 확정한다.

(과정4)의 <원칙2>에 제시된 보깃글의 경우는 반드시 후통제가 필요하다. 그 까닭은 다음과 같다.

해당 보깃글에 대한 <서술어 판독> 결과, 1차 서술어는 '귀가시키다'가 되고, 2차 서술어는 '밝히다'가 된다. 이들 서술어들을 <기본 문형 사전>에 대입하면 '귀가시키다'의 경우 2번 문형인 < --가 --를 ____ >으로 판독된다. 이를 원문에 적용, 해당 어절을

추출하고 필요 성분을 자동으로 부여하면 '태국의 치안당국은 폭동을 일으킨 불량ㅅ배들을 귀가시켰다고 밝혔다'와 같은 정반대의 문장이 생성된다. 그러므로 이와 같은 문제를 해결하기 위하여 자동 추출 시스템에 덧붙여 '후통제 처리'라는 제어장치가 필요하게 되는 것이다.

Ⅵ. 검 증

(본문)한편 주태 한국대사관과 교민회는 이번 방콕의 소요사태에서 피해를 입은 교민은 한명도 없는 것으로 이날 오전 확인됐다고 밝혔다.

-----> 1. 교민은 것으로 확인됐다.
2. (추가) 피해를 입은 교민은 없는 것으로 확인됐다.

(풀이1) '1'의 어절이 추출된 것은 <1차 서술어>로 확정된 '확인되다'가 <기본 문형 사전>을 통과하여 해당 문형인 < --가/은/는/이 --(으)로 ____ >를 읽어온 뒤 이를 원문에 적용하여 해당 어절을 추출하였기 때문이다. '은/는'의 형태가 붙은 곳이 몇 군데 더 잇는데 그 가운데 '교민은'이란 어절을 읽어올 수 있었던 것은 해당 서술어를 중심으로 그 앞의 어절 중 가장 가까운 곳에 있는 어절을 추출하라는 규칙이 구문분석기에 또한 있기 때문이다. '가까운 곳'이라는 전제로 문장을 분석한 이유는 우리말 복합문장의 특수성을 고려한 것이다.

(풀이2) '1' 다음에 '2'에 추가된 규칙은 '명사 앞의 관형어는 추출하라'와 '그 관형어의 품사가 타동사이면 바로 앞의 목적어도 추출하라'는 규칙 때문이다.

-----> 3. 교민회는 밝혔다.

(풀이3) (본문)에서 판독된 서술어는 '확인되다'와 '밝히다'의 두 개이다. 이 가운데 <1차 서술어>인 '확인되다'는 이미 앞에서 적용이 됐으므로 이번에는 남은 <2차 서술어>가 <기본 문형 사전>과 대조, 문형을 읽어온 뒤 원문에 적용하여 해당 어절을 읽어온다. 이때 '밝히다'의 기본 문형은 < --가/이/는/은 --를/을 ____ >이다.
이때 의심의 여지가 생기는 부분이 있는데 '은, 는' 등이 들어간 곳이 위에서처럼 몇 군데 있는데 어째서 '교민회는'만 추출되었는가?에 대한 물음일 것이다.
이에 대한 규칙은 한 번 적용된 문장은 다른 규칙에 더 이상 적용되지 않는다는 규칙이 또한 있기 때문이다.

-----> 4.(추가) 한편, 고민희는 피해를 입은 고민은 없는 것으로 확인됐다고 밝혔다.

'한편'이 삽입된 것은 문장과 문장을 접속하는 접속사는 문 연결의 부드러움을 위하여 필요할 것으로 판단되어 이를 구문분석기의 규칙으로 삼았기 때문이다.

Ⅶ. 맺음말

이상으로 컴퓨터를 이용한 자동축약 시스템 구축에 '격문법 및 성분 이론'을 적용하는 방법을 간략하게 살펴보았다. 살펴본 결과, '자동축약 시스템'에 이들을 적용하는 것이 상당히 효과적임을 알 수 있었다.

그러나 다음과 같은 문장에서 보이는 일부 서술어는 위에 설정한 기본문형을 벗어나는 것 같다.

(보기1) 이 실험장치는 크게 기계적인 부분과 전자적인 부분으로 분리된다(나뉜다).
(보기2) 그 결과를 기계번역 시스템에 충분히 활용할 수 있다.

위의 문장들에서 핵심이 되는 부분(성분)은 서술어 '분리된다(나뉜다)'와 '활용할 수 있다'이다.

이들을 중심으로 한 필수(기본)성분은 밑줄을 그은 부분으로, (보기1)은 '실험장치는 기계적인 부분과 전자적인 부분으로'이고, (보기2)는 '결과를 기계번역 시스템에'이다.

결국 (보기1)의 서술어 '분리되다(나뉘다)'의 기본문형은 < --은/는 --과/와 --으로 ___ >이 되고, (보기2)의 서술어 '활용하다'의 기본문형은 < --를 --에 ___ >가 될 것이다. 그러나 이 두 문형은 위의 9개 문형에 나타나지 않고 있다. 그러므로 위에 설정한 기본문형 9개는 모든 부분에 일제히 적용된다고 볼 수 없으므로 추가 조사 후 이의 재조정이 반드시 필요한 것으로 생각된다.

그러나 기본문형의 틀 안에 넣기에는 그 규칙성 정도가 혼란스럽다고 판단되면 이들은 구문분석기를 보완하거나 후통제 처리를 통하여 해결될 것이다.

♠ 참 고 문 헌 ♠

[1] 고영근/남기심, <표준 국어문법론>, 탑출판사, 1989

[2] 김승곤, <우리말 토씨 연구>, 건국대 출판부, 1989

[3] 김영희, '국어의 격문법 연구', 연세대 국어국문학과 석사논문, 1973

[4] 남용우/임선호/이통진/황봉주 역, <격문법이란 무엇인가>, 을유문화사. 1987

[5] 문동섭, '구문정보 및 의미정보를 이용한 한국어 격문법 PARSER의 설계 및 구현에 관한 연구', 한양대 전자공학과 석사논문, 1986

[6] 백혜승/이승미/최기선, '한국어 문서 축약 시스템의 설계', <인간과 기계와 언어>, 한국정보과학회 제3회 학술발표회, 1991

[7] 서경주, '언어학적 분석기법에 의한 신문기사 자동색인 시스템 설계에 관한 연구', 숙명여대 도서관학과 석사논문, 1990

[8] 서정수, '변형 생성 문법의 이론과 국어 V(동사)-류어의 하위분류 연구', 연세대 국어국문학과 석사논문, 1968

[9] 이익섭/임홍빈, <국어문법론>, 학연사, 1988

[10] 정영미, <정보검색론>, 정음사, 1988

[11] 조성호, '컴퓨터를 이용한 한글 자동색인 시스템 구축에 관한 연구', 연세대 산업대학원 산업정보학 석사논문, 1988

[12] 최원태, '격문법을 이용한 자동색인 및 탐색 확장에 관한 연구', 연세대 도서관학과 석사논문, 1986

[13] 최현배, <우리말본>, 정음문화사, 1983

[14] 한영목, '한국어 구문 도해 연구', 충남대 국어국문학과 박사논문, 1988