한국어 문서 축약 시스템의 설계

백혜승 이승미, 최기선
한국원자력연구소 한국과학기술원 전산학과

A Design of Korean text CONdensing System(KCONS)

Haeseung Paik Seungmi Lee, Keysun Choi
Korean Atomic Energy Research Institute Department of Computer Science, KAIST

요 약

본 논문에서는 한국어 문서를 대상으로 한국어에 관한 형태소 및 구문정보를 이용하고 또한 문장구조상에 나타난 특징들을 고려한 휴리스틱(Heuristic)을 이용하여 각 문장 단위로 축약하는 시스템을 설계한다. 그리고 이 축약 시스템을 평가하기 위한 방법들을 제안한다.

I. 서론

오늘날 우리는 '정보화 시대'에 살고 있다는 말을 흔히 듣고 있다. 하루가 다르게 쏟아지는 엄청난 양의 데이타를 접하면서, 그 모든 데이타로부터 유용한 정보를 얻는 일은 쉽지 않다. 이와 같은 데이타의 과부하 상태를 효과적으로 다루지 못 한다면 우리는 산더미 같은 데이타의 늪에 빠져 혼란이 가중될 것이다.

컴퓨터에 의해 많은 양의 데이타를 관리하고 유용한 정보를 추출하기 위한 방법들이 오래 전부터 연구, 개발되어 왔으나 이는 대부분 그 대상이 구조화된 데이타 - 가장 일반적인 예로 개별 항목이 2차원 관계형 테이블로 할당된 DBMS - 에 관한 것들이었다. 그러나 우리가 친숙하게 대하는 대개의 정보는 신문, 잡지, 전문 기술서적, 전자 우편, 전자 신문 등에 의해 전달되며 이들은 우리가 사용하는 언어에 의해 다양한 길이의 문장으로 이뤄진 문서 형태이다. 많은 내용의 문서 전체를 모두 읽기 전에 그 문서가 어떤 정보를 담고 있는 지를 미리 알 수 있다면 불필요한 문서를 읽는 데 소모되는 시간을 줄일 수 있고 문서 전체를 참조할 것인지에 대한 단서를 찾을 수도 있을 것이다.

이와 같이 문서 상의 중요정보를 추출하기 위한 노력이 자동색인 시스템 (Automatic

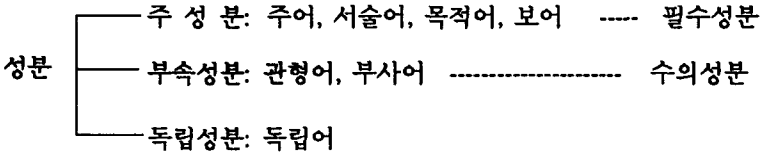
Indexing System)에서 자동요약 시스템(Automatic Summarization System)으로 이어지고 있다. 그러나 자동색인을 통해 얻을 수 있는 정보인 주요어휘들(Key words)은 단지 그 단어 자체가 갖는 의미만을 포함할 뿐 그 외 문서 상의 내용에 관한 정보는 전달할 수 없다. 문서가 전하고자 하는 중요내용은 요약(summary)이나 개괄(abstract)을 통해 알 수 있는데, 이를 자동으로 추출하기 위해서는 해당 언어에 대한 언어학 지식(Linguistic knowledge)뿐 아니라 문서가 다루는 대상영역(Field domain)에 관한 지식도 필요하다. 다시 말하면, 문장 자체에 대한 구문분석(Syntactic analysis)외에 대상영역의 지식에 관한 의미분석(Semantic analysis)까지도 이뤄져야 한다는 것이다.

본 논문에서는 한국어에 관한 언어학 지식인 형태소 및 구문 정보들을 이용하여 특정 대상영역의 제한이 없는 일반적인 문서를 축약하는 시스템을 설계하고자 한다.

II장에서는 본 시스템을 정의하고 설계하는 목적을 밝혔으며 III장에서는 시스템 설계에 관해 각 단계별로 설명하였고 IV장에서는 이 시스템의 평가기준을 제안하였다.

II. 정의 및 목적

한 문장은 문장 내에서 각기 다른 역할을 하는 구성요소인 성분들로 구성되며 그 분류는 다음과 같다[1].



문장의 주요 내용들은 주로 필수성분으로 표현되며 수의적 성질을 갖는 부속성분들은 주성분이 나타내고자 하는 것을 더 구체적으로 가리키거나 묘사한다고 볼 수 있다. 그러므로 예 1) - 3) 에서와 같이 큰 정보의 손실 없이 문장 내의 필수성분만을 추출하여 보다 간단한 문장을 만들 수 있다.

예 1) 개나리가 담장 아래에서 노랗게 피었다. (주어 + 서술어)
예 2) 코끼리가 과자를 맛있게 먹는다. (주어 + 목적어 + 서술어)
예 3) 액상의 화합물이 노란 기체가 되었다. (주어 + 보어 + 서술어)

그러나 다음의 또 다른 예 4) - 7) 에서 보듯이 문장의 필수성분만을 추출했을 경우 부속성분에 의해 표현된 다른 의미있는 정보들이 크게 손실되는 문제가 발생한다.

예 4) 그는 어제 책을 자동차로 서울에서 대전으로 옮겼다.

예 5) 이 실험장치는 크게 기계적인 부분과 전자적인 부분으로 분리된다.
예 6) 그 결과를 기계번역 시스템에 충분히 활용할 수 있다.
예 7) 이 기후는 농사에 적합하다.

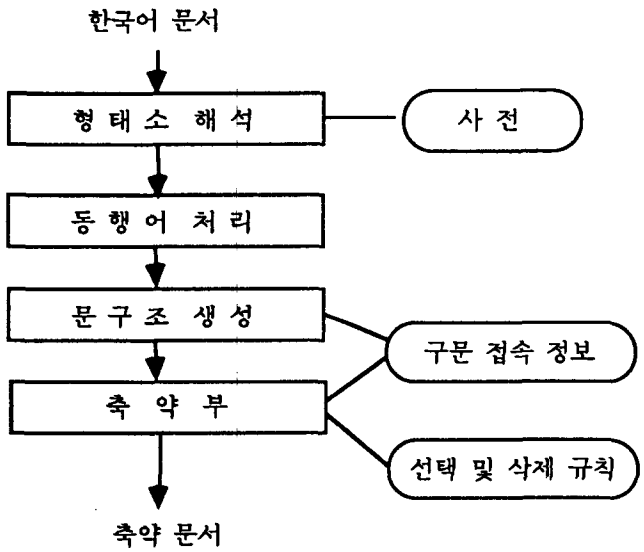
이런 현상들을 볼 때 문장의 주성분 뿐 아니라 용언의 필수격에 해당하는 부사어 및 여러 종류의 관형어들도 고려할 필요가 있음을 알 수 있다.

본 논문에서 설계하는 한글 문서 축약 시스템이란 위와 같은 한국어에 관한 언어학 정보를 이용하여 문서 내의 모든 문장에 대해 각 문장 단위로 주요성분들을 선택하여 축약된 문장을 만드는 시스템으로 정의할 수 있다. 축약된 문서가 축약되기 전의 원래 문서가 갖는 모든 정보를 전달한다고는 볼 수 없지만 거의 모든 문장이 전달되어 전체 문맥을 통해 원래 문서가 나타내고자 하는 바를 알 수 있도록 하는 것이 본 시스템의 목적이다.

III. 한글 문서 축약 시스템의 설계

1. 전체 구성도

본 시스템의 전체 구성은 다음과 같이 크게 4단계로 나눌 수 있다.



축약 시스템의 각 단계를 아래의 예문을 들어 설명해 보겠다.

예) 엘렉스 컴퓨터에서는 그 동안 시스템과 함께 무료로 제공하였던 워드프로세서 엘렉스북과 엘렉스워드에 이어 기능이 훨씬 보강된 Nisus라는 워드프로세서를 발표했다.

2. 형태소 해석

형태소 해석과정은 어절 단위의 분석을 통해 다음과 같은 정보들을 얻을 수 있다.

1) 격 조 사: 컴퓨터 + 에서는 (주격), 무료 + 로 (부사격)
2) 접속조사: 시스템 + 과, 엘렉스북 + 과
3) 관형사형: 보강 + 되 + ㄴ, 제공 + 하 + 였 + 던
4) 연결어미: 이 + 어
5) 의존명사: 동안
6) 부 사 어: 훨씬 (정도 표현)

3. 동행어 처리

동행어란 둘 이상의 어절이 모여 하나의 문장 성분을 구성할 때, 이들은 항상 함께 처리되므로 이를 동행어라고 하였다. 형태소 해석을 통해 분석된 각 어절은 아래와 같은 경우 한 성분으로 다룬다.

1) 복합명사구

엘렉스 + 컴퓨터에서는 -> 엘렉스 컴퓨터에서는
대치 + 기능 -> 대치 기능

2) 의존명사구 (서술격 조사 '이다'가 이어질 때는 제외)

그 + 동안 -> 그 동안
여러 + 개의 -> 여러 개의
다른 + 곳에 -> 다른 곳에
첨가할 + 때까지 -> 첨가할 때까지

3) 본용언 + 보조용언

믿어지지 + 않을 -> 믿어지지 않을

받지 + 않아도 -> 받지 않아도

올라와 + 있는 -> 올라와 있는

4) 용언의 이어짐

작성할 + 수 + 있다 -> 작성할 수 있다.

저장해야 + 한다 -> 저장해야 한다.

편집하고 + 저장할 수 있도록 한다.

5) 접속조사 '와', '과'

검색과 + 대치기능 -> 검색과 대치 기능

시스템과 + 함께 -> 시스템과 함께

6) 명사 + , + 명사

문자 + , + 단어를 -> 문자, 단어를

4. 문구조 생성

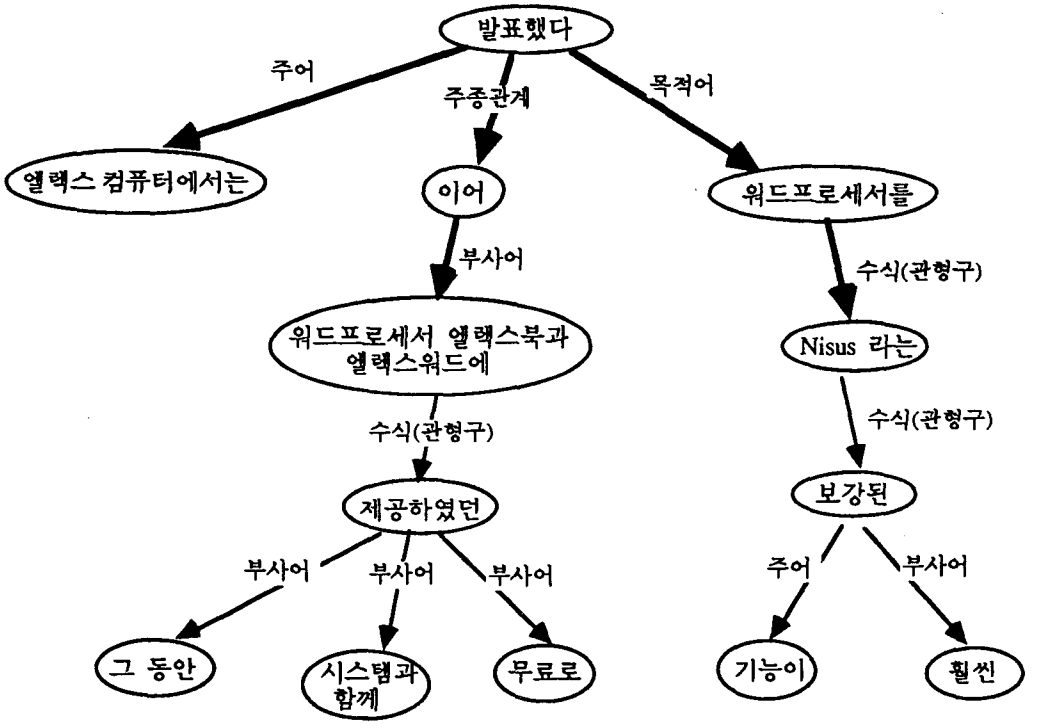
잡지나 논문 등에 실린 일반적인 한글 문서를 조사해 본 결과 아래와 같은 특징이 두드러지게 나타남을 발견할 수 있었다. 이들을 본 문서 축약시스템에 적용하였다.

1) 주어가 자주 생략된다.
2) 대부분 조사에 의해 격을 표현한다.
3) 일반적으로 수식어는 피수식어 앞에 온다.
4) 단문을 구성하는 핵심요소는 서술어이다.
5) 어순이 비교적 자유롭지만 일반적으로 서술어는 문장 끝에 위치한다.
6) 한 서술어는 그 앞에 다른 서술어가 있기 전까지의 모든 문장요소와 관계된다고 볼 수 있다.
7) 연결어미에 의해 대등관계 및 주종관계를 알 수 있다.
8) 주절의 서술어는 종속절의 서술어와 연결관계에 있다.
9) 명사절, 인용절, 관형절들이 많이 나타난다.

문장을 단문 단위로 나누면 서술어를 중심으로 모든 다른 성분들이 Head-dependent관계를 형성하고, 또 수식어가 피수식어의 dependent가 된다고 가정하자.
그러면, 형태소 해석과 동행어 처리 단계까지의 결과를 통해 구문구조트리(Sentence Structure Tree: SST)를 생성할 수 있다.

Head	Dependent
서술어	주어, 목적어, 보어, 부사어, 독립어
주어, 목적어, 보어	관형어
주 서술어	종속 서술어
체언	관형사

문구조트리(SST)에서 각 노드(node)는 문장 성분을 이루는 하나 혹은 둘 이상의 어절이거나 수식, 피수식어로 구성된다. 따라서 head와 dependent간의 관계는 문장성분의 구성요소이거나 수식, 피수식 관계 또는 주종관계를 갖는다. 이 트리는 Left-to-Right- Root 탐색에 의해 모든 문장 구성 요소를 찾을 수 있다. 앞에서 택한 예문의 문구조트리는 아래와 같다.



5. 축약부

앞에서 얻어진 구구조트리는 형태소 해석에서 찾은 사전 정보와 문장성분 정보를 유지하고 또 수식, 내포, 주종관계를 갖는다. 축약부에서는 이들 정보를 이용하여 실제로 문장을 축약하는 일을 한다. 여기서는 축약될 가능성이 있는 경우를 문장 성분에 의한 것과 그 외의 휴리스틱을 사용한 방법을 알아본다.

5.1. 문장 성분에 의한 축약

1) 체언을 수식하는 관형구나 절
- 이것은 일본으로 수출할 통신부품이다.

2) 체언의 성질이나 상태를 나타내는 성상 관형사
- 어머니는 설날이 되면 오례 새 옷을 만드셨다.

3) 성상부사 중에서 정도를 나타내는 부사 및 의성어, 의태어
- 그것은 매우 크게 확장되고 있었다.

4) 말하는 이의 마음먹기나 태도를 표시하는 양태부사
- 과연 그는 우리에게 꼭 필요한 기술자였다.

5) 문장접속부사
- 개나리가 노랗게 피었다. 그래서 우리는 사진을 찍었다.

6) 서술기능을 갖는 파생부사가 이룬 구나 절
- 그녀는 펄듯이 기뻐했다.
- 그는 돈도 없이 극장에 갔다.

7) '체언+부사격 조사'형태가 용언의 필수격이 아닐 때
- 우리는 버스로 학교에 갔다.

8) 부사성 의존 명사구
- 그는 옷을 입은 채 물 속에 뛰어 들었다.

9) 형용사 어간 +'게' 가 문두에 올 경우
- 불행하게도 그의 상태는 더욱 악화되었다.

10) 감탄사
- 아, 이렇게 아름다운 풍경을 다시 볼 수 있다니 가슴이 벅차다.

그러나 위와 같은 경우에도 예외적인 사항들이 있으므로 이들은 선택규칙으로 정의해 주고 그럼으로써 필수적인 정보가 손실되지 않도록 하고 또 축약문의 형태가 구조적으로 바르게 되도

록 한다. 한 예로 1)의 경우, 체언을 수식하는 관형구가 의존명사 '것'을 수식할 경우는 삭제하지 않고 축약문에 포함시켜야 한다.

5.2. 휴리스틱을 이용한 방법

1) 주어를 수식하는 관형절은 남겨둔다.
- DPMI라고 불리우는 이 표준은 11개의 PC업계 리더들로 구성된 위원회의 산물이다.

2) 부정문 뒤에 긍정문이 연결될 때 앞의 부정문은 생략한다.
- 이러한 정보화는 사회적 합리화가 아니라 비합리화의 역효과를 초래한다.

3) 우열을 비교하는 문의 표현에서는 나타내고자 하는 구만을 남겨둔다.
- 대부분의 기관들은 동격형의 랜보다는 서버형 랜을 주로 택한다.

4) 예의 나열에서는 그 예들을 대표하는 단어만 남긴다.
- 화면기기, 광 디스크 등 컴퓨터 하드웨어의 가격이 하락되고 있다.

5) 첨가, 부연 설명문임을 명시하는 부사구가 문두에 오면 문장 전체를 생략하기도 한다.
- 예를 들면, 즉, 다시 말하면

6) 둘 이상의 문장이 연결어미에 의해 결합된 경우 아래와 같은 종속절은 생략할 수 있다.
- 결과가 예상과 반대됨을 나타내는 '-아/라도', '-지마는', '-(으)나', '-지만'
- 어떤 일의 배경을 나타내는 '-는데', '-(으)ㄴ데'
- 양보의 의미를 갖는 '-ㄹ지라도'

7) 중요한 내용을 전달함을 명시적으로 나타내는 단어인 '목적', '장점', '특징'들이 사용된 문장은 다른 문장보다 선택규칙을 강화한다.

앞에서 택한 예문은 축약부를 거치면 다음과 같은 문장으로 된다.

"엘랙스 컴퓨터에서는 워드프로세서 엘랙스북과 엘랙스워드에 이어 Nisus라는 워드프로세서를 발표했다."

IV. 평가 방법

본 시스템을 평가하기 위한 방법으로는 크게 두 가지를 들 수 있다. 첫째는 원래 문

서와 축약된 문서 사이의 양적 평가를 하는 방법이 있고 두 번째 방법은 축약된 문서가 원래 문서가 갖는 중요의미를 얼마나 충분히 잘 전달했는가를 측정하는 것이다. 이들의 평가 기준을 다음 표와 같이 제안해 볼 수 있다.

양 적 평 가	질 적 평 가
1. 어절 수	1. 중요어휘(Key words) 의 갯수
2. 문장 수	2. 원문 중심의 질문에 답하는 정도
3. 문장 내의 평균 어절 수	3. 의미없는 문장들의 갯수

V. 결론

본 문서 축약 시스템이 좋은 결과를 얻기 위해서는 형태소 해석 및 동행어 구성도 잘 되어야 하겠지만 보다 중요한 것은 문구조트리를 생성하는 부분이다. 문구조트리가 수식-피수식관계, 내포관계, 주종관계를 명확히 나타내 주면 트리 내의 노드들에 관한 성분정보들을 이용하고 또 휴리스틱 방법을 적용하여 좋은 축약문을 생성할 수 있다.

그러나 이 시스템을 설계하면서 더 해결해야 할 문제점들을 발견할 수 있었는데, 이는 대부분 지시 대명사나 지시 관형사 및 생략된 문장 성분의 처리이고 또한 선택과 삭제 규칙 사이의 충돌이 생길 경우 이 규칙들 사이의 우선순위를 적절히 부여함으로써 이를 해결해야 한다는 것이다. 이 시스템에서 대상으로 하는 문서의 문장들은 한국어 문법에 맞는 것이어야 하며 비문(非文)의 경우는 축약할 수 없는 것으로 보았다. 설계된 본 시스템이 구현되면 많은 데이타로부터 유용한 정보를 얻는데 적지 않은 도움이 될 것으로 본다.

[참고문헌]

1. 고영근, 남기심, "표준 국어문법론", 탑출판사, 1985.
2. 김효준, " 의미구조로부터 내포문의 생성에 관한 연구", 한국과학기술원 석사 논문, 1989.
3. 박용운, 조혁규, 권혁철, "의존문법을 이용한 한국어 분석기의 구현", 한국어 정보과학회 봄 학술발표 논문집, Vol. 17 No.1, pp. 191~194, 1990.
4. 최운천, "변환방식의 기계번역을 위한 한국어 생성기의 설계 및 구현", 한국과학기술원 석사논문, 1991.
5. Irving W. Miller, Jun Ibuki, Fumihito Nishino, "The Construction and Evaluation of ECON- An English Text Condensing System", PRICAI90, pp. 328~333, 1990.
6. Eiji Komatsu, Yasuhiko Kato, Hiroshi Yasuhara, "Summarization Support System COGITO - Structure analysis of text" , 自然言語 處理 64-11, pp. 85~91, 1987.
7. Soutarou KITA, "A system for summarization of an explanatory text" , 自然言語 處理 63-6 , pp. 1987.