구문해석을 이용한 색인어 자동 추출 시스템

한성현, 박혁로, 최기선, 김길창
한국과학기술원 전산학과(KAIST)
인공지능 연구센터(CAIR)

요약

본 논문에서는 자동 색인 시스템 구현에 있어서 형태소 해석뿐만 아니라 구문해석을 응용하면 통계적 방법이나, 간단한 단서에 의한 색인어 추출보다 훨씬 나은 색인어 추출이 가능하다는 것을 보이고 한국어 필수격이 색인어로써 충분한 자질이 있다는 제안을 한다. 또 시스템의 전체적인 흐름과 필수격 처리 과정, 예외적인 자유격의 처리 등에 대한 부분을 설명하고, 결론에서는 사람이 추출한 색인어와 본 시스템의 결과를 비교, 분석 한다.

I. 서론

현대 정보화 시대에서, 정보가 사람이 관리할 수 없을 정도로 쏟아져 나오게 됨에 따라 많은 정보들이 컴퓨터내에 저장 관리되고 필요에 의해 사용자에게 서비스되는 정보관리 시스템이 널리 이용되고 있다. 이러한 정보 서비스 시스템의 성공여부는 정보 서비스의 정확성과 정보추출의 속도에 의해서 결정된다. 정보 서비스 시스템의 정확성과 속도는 그 정보 서비스 시스템이 유지, 관리 하는 색인어의 정확성에 의존한다.

정보 서비스 시스템에서 정보의 내용을 표현하는 색인을 추출하는 방법으로 이전에는 사람에 의한 색인이 사용되었으나, 정보량의 방대함, 색인어의 불일치성등의 문제에 따라 컴퓨터에 의한 자동색인(Automatic Indexing)에 대한 많은 연구가 수행되었다. 자동색인이란 데이타 베이스에 저장된 많은 정보 중에서 탐색어에 따라 필요한 정보를 즉시 찾을 수 있도록 각각의 정보에다 그 정보를 대표할 수 있는 특정한 색인어들을 자동으로 컴퓨터가 부여하는 시스템을 말한다.

본 논문에서는 한국어 문서에 대해 형태소분석과 구문분석을 이용한 색인어 자동 추출 시스템의 설계와 구현을 보인다.

II. 관련 연구

1. 자동 색인 시스템(Automatic Indexing System)

자동 색인 시스템은 컴퓨터에 입력된 문헌의 텍스트를 분석한 후 이 문헌의 주제 내용을 대표할 수 있는 단어나 단어구를 추출해 내는 것[9]이다. 자동 색인기법은 크게 통계적 방법, 언어학적 방법과 문헌 구조적인 방법으로 나눌 수 있다[9]. 통계적 방법은 문헌에서 단어의 출현 빈도를 색인어의 선택 기준으로 삼는다. 문헌 구조적인 방법은 문헌의 단서를 이용 즉, 서론 부분이나, 첫 문장등 특정한 부분의 주제어를 색인어로 하는 방법으로 문헌전체를 대상으로 하는 자동색인 방법으로는 유용하지 않다.

언어학적 방법은 문장 분석 단계에 따라 형태소분석, 구문분석, 어미분석 단계로 나눌 수 있다. 형태소분석을 이용한 자동추출은 문헌에 형태소분석을 행하여 불용어(stop word)를 제거한 후 통계적 방법을 행하는 방법을 사용하고 있다[8]. 형태소분석 자동색인은 색인어 선택이 단순한 빈도수나, 가중치에 의해 정해지므로 정확한 색인어 추출이 어렵고 단일어에 대한 처리만 가능하다[12]. 따라서 문장의 분석을 통한 정확한 색인어 추출이 자동색인에서는 필요하다.

자동 색인에서의 구문분석이란 특정한 구문적 기능을 수행하는 단어나 단어구가 문헌의 내용을 나타낸다는 가정 아래 이러한 구문단위를 식별하는 작업을 의미한다. 외국의 실용 자동 색인어 추출 시스템으로는 구문분석을 이용한 SMART[10][11][12], PHRASE[13]가 있는데 이들은 완전한 문장분석을 통하여 색인어를 추출하는 시스템이다. 그리고 일본에는 한정된 특허문에서 색인어를 추출하는 시스템[14]이 있다. 한국어에 대한 색인어 시스템으로는 아직까지 구문분석을 이용한 실용적인 자동색인 시스템은 없다.

의미 분석을 이용한 방법은 지식 표현이나 어미 사전을 이용해 완전히 문장의 이해를 하는 시스템으로써 지식표현 데이타 베이스(knowledge representation data base)와 방대한 시소러스(thesaurus)를 필요로 하기 때문에 현재의 자동색인 시스템에서는 구현하기는 무리가 있다.

그러므로 본 논문은 형태소해석을 이용한 방법의 단점을 최대로 없애고 간단한 구문분석 기법을 이용함으로써 형태소분석을 이용한 방법의 색인어 추출보다 더 정확한 색인어를 추출하는 시스템의 설계와 구현을 하였다.

2. 구문분석 기법

구문분석은 문장의 구조를 체계적으로 그 문장의 구성성분을 분석하는 작업이다. 일반적인 기계번역(machine translation)이나 자연어 이해 시스템에서는 구구조규칙(Phrase structure rule)이나, 격해석법을 이용하여 구문분석하고 있다. 그러나 한국어는 부분자유어순 형태를 띠는 언어이기 때문에 구구조규칙을 사용하여 구문분석을 하기 위해서는 복잡한 구구조규칙이 필요하고 문장해석 결과에 많은 애매성이 생기므로 이러한 방법으로는 해석이 어렵다[7]. 그리고 구구조규칙을 이용한 구문해석 결과의 구문구조는 본 시스템에서 원하는 중요한 단어의 선택을 더욱 복잡하게 한다. 따라서 본 시스템에서는 격해석을 이용하여 간단히 구문분석을 행하여 색인어 추출 시스템에 이용했다.

III. 한국어의 격과 색인어

1. 격과 색인어

색인어의 중요한 역할 중 하나는 어떤 문헌에 대해 그 문헌이 나타내는 의미와 내용을 표시한다[9]는 것이다. 그런데 그 문헌이 나타내는 의미와 내용이라는 것은 그 문헌을 구성하는 각각의 문장에서 중요한 의미를 지닌 말들 또는 문장에서 중요한 역할을 하는 말들을 모아서 만들어 질 수 있다. 즉 색인어는 문헌 내 각 문장에서 중요한 의미를 지닌 단어, 또는 단어구들 것이다. 한편 Fillmore가 주장한 격의 원리를 이용한 격문법은 한 문장에서 의미표시를 서술어(한국어에서는 용언)와 문장 내의 그와 연관된 명사구들의 의미역할(semantic role)을 설정하는 문법을 말하는데[3] 이러한 격해석에서 정의된 격의 기본적인 개념이 앞에서 말한 색인어의 추출 원리와 맞아 떨어진다고 볼 수 있다. 다음에 제시되는 Fillmore의 격에 대한 정의를 보면 더욱 위의 사실이 확실해 진다.

격문법에 대한 나의 제안에서, 나는 술어의 의미구조를 기술하는 데 있어 우리가 말할 수 있는 역할 형태는, 보편적으로 타당하고 이치에 맞게 잘 명시된 조의 개념을 형성한다고 했다. 역할형태 자체는 분석할 수 없는 것이며, 누가 그 일을 했는지, 누가 그것을 경험했는지, 어디서 그것이 일어났는지, 결과는 어떻게 되었는지, 무엇이 움직였는지, 그 밖의 몇가지에 관한 인간사 측면의 기본적인 인지에 해당하는 것이라고 했다.

2. 색인어와 필수격에 대한 실험

이 부분에서는 앞에서 내린 가정, 즉 술어에 대한 필수격은 색인어 추출개념과 같은 의미를 지닌다는 것에 대한 실험적 결과를 제시해 보겠다. 조사 대상문은 현재 실제적으로 '천리안 서비스'에 직접 사용되고 있는 각종 신문기사를 선택했다. 이 기사들은 크기가 약 200에서 400단어로 된 각종 사실들을 보도한 것으로 약 300개의 기사를 대상으로 했다. 이들 대상문의 색인어는 사람이 직접 뽑은 것으로 보통 하나의 기사에 대해 40개에서 60개 정도의 색인어를 가진다. 이런 대상문에서 필수격이 아닌 것이 색인어로 추출되는 비율을 조사 해 보았다. 첫번째 대상문 150개 중 자유격은 474개가 그 중 41개가 색인어로 채택되었고, 두번째 대상문 150개 중 자유격으로 400개가 검색되었고, 그 중 42개가 색인어로 선택되었다. 약 10% 가량이 자유격이면서 색인어로 선택되었다. 자유격으로 색인어가 된 경우는 주로 고유명사인 경우로써 색인어로 선택되거나(60%), 문장 내에서 격의 사용이 잘못된 경우가 대부분이었다.

VI 시스템 구현

1. 전체 시스템의 개요

시스템 구조를 간단히 살펴 보면 전체 시스템은 입력문을 받아들이는 형태소 해석기, 형태소 해석기 결과를 가지고 구문해석을 하는 구문해석기, 그리고 불용어를 처리하는 부분으로 나누어져 있다. (그림 4-1).

2. 형태소 해석기

본 시스템에서 이용하는 형태소 해석기는 접속정보를 이용한 한국어 형태소 해석기[1]이다. 이것은 한국어 철자검색 시스템의 한 부분으로 자동 색인 시스템의 구문해석에 필요한 부분만 이용했다. 해당 한국어 형태소 해석기는 좌우접속정보표[1]의 작성과 사전의 구성을 위하여 25가지로 품사를 세분류 하고 있다.

형태소 해석기는 각 형태소의 품사분류가 세분류되어 구문분석에서 문장성분의 판별을 더욱 쉽게 한다. 형태소 해석에서 발생되는 단어의 애매성은 형태소 해석 다음 단계에서 가능한 애매성을 해소하도록 하였다. 형태소 해석은 두가지 사전, 주 사전과 보조 사전을 사용하고 있다. 주 사전의 형태는 ('형태소' ('좌접속 번호' '우접속 번호'))의 형태를 가지고 실질적 형태소 해석에 사용된다. 보조 사전은 주사전에서 발견되지 않은 사전 미등록어를 처리하기 위해 단순한 형태소 만을 수록한 명사 사전이다. 형태소 해석기는 한 어절 내에서의 형태소를 한국어 어절 구조[5]에 따라 파악하는 방식을 취했다. 좌접속 정보를 단어의 품사로 간주하여 구문해석을 했다.

3. 구문해석기

구문해석기는 완벽한 한국어에 대한 구문해석을 하기 보다는 본 시스템이 목표로 하는 한국어 색인어 추출에 목표를 둔 구문해석기를 만들고자 했다. 구문해석기에 대한 전체적인 흐름도(그림 4-2)와 자세한 설명은 아래와 같다.

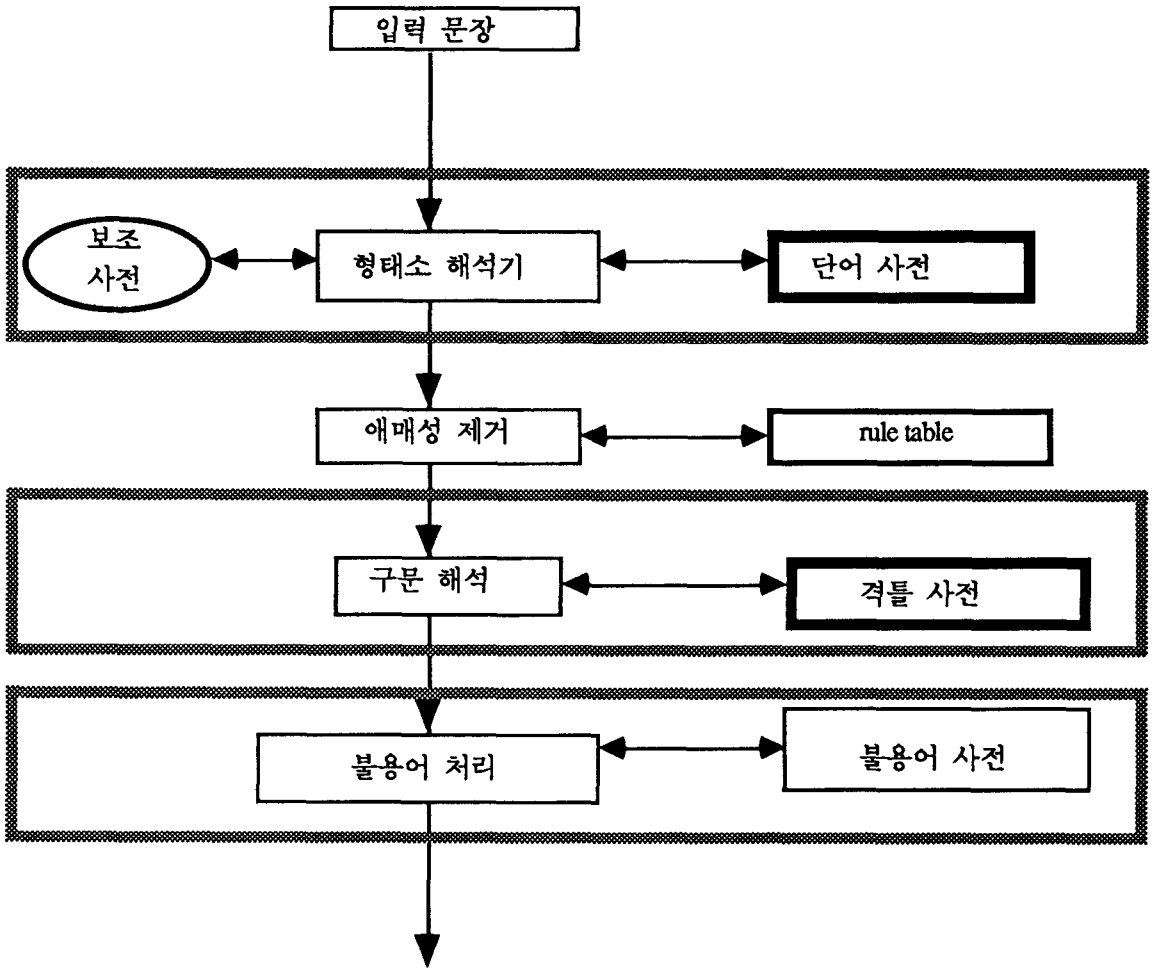

그림 4-1 <전체 시스템 흐름도>

3.1 판형사, 판형격 조사 처리

구문해석기에서는 맨 처음으로 판형사에 대한 처리를 하고 동시에 판형격 조사 '의'를 처리 했다. 그리고, 명사가 명사를 수식 하는 경우도 판형사 수식과 똑같은 방법으로 처리했다.

3.2 용언, 조용사, 형용사 판정

형태소 해석 결과로 만들어진 각각의 형태소를 차례로 읽어 용언 어간, 형용사 어간, 조용사(하다, 되다, 이다, 있다, 없다..) 어간을 찾는다. 찾아진 어간을 가지고 격사전에서 해당 용언, 조용사, 형용사에 대한 격틀을 읽어 온다.

위에서 읽혀진 격구조를 가지고 해당 용언이나 형용사로 부터 문장 뒤에서 읽어 오면서(back search) 각각의 격틀에 할당된 대표조사와 비교하여 형태소들을 격틀에 넣는다. 여기서 문장을 거꾸로 다시 비교하는 이유는 한국어에서의 내포문 처리에 훨씬 유리 하기 때문에 이와 같은 방법을 채택하여 사용했다.

3.3 격틀 사전(case frame dictionary)

격틀 사전의 데이타 저장은 '(날다 (1 ((2 1))))'와 같은 형태로 되어 있다. 맨 처음의 '1'은 해당 용언의 격틀이 한개가 존재한다는 것을 표시하고, 나머지 숫자는 격틀에서 분류한 대표조사를 의미한다. '날다'의 격틀을 다시 풀어 쓰면 '(날다 (1 ((AGT '은, 는, 이, 가, 에서, 께서,..') (OBJ '을, 를'))))'과 같이 된다.

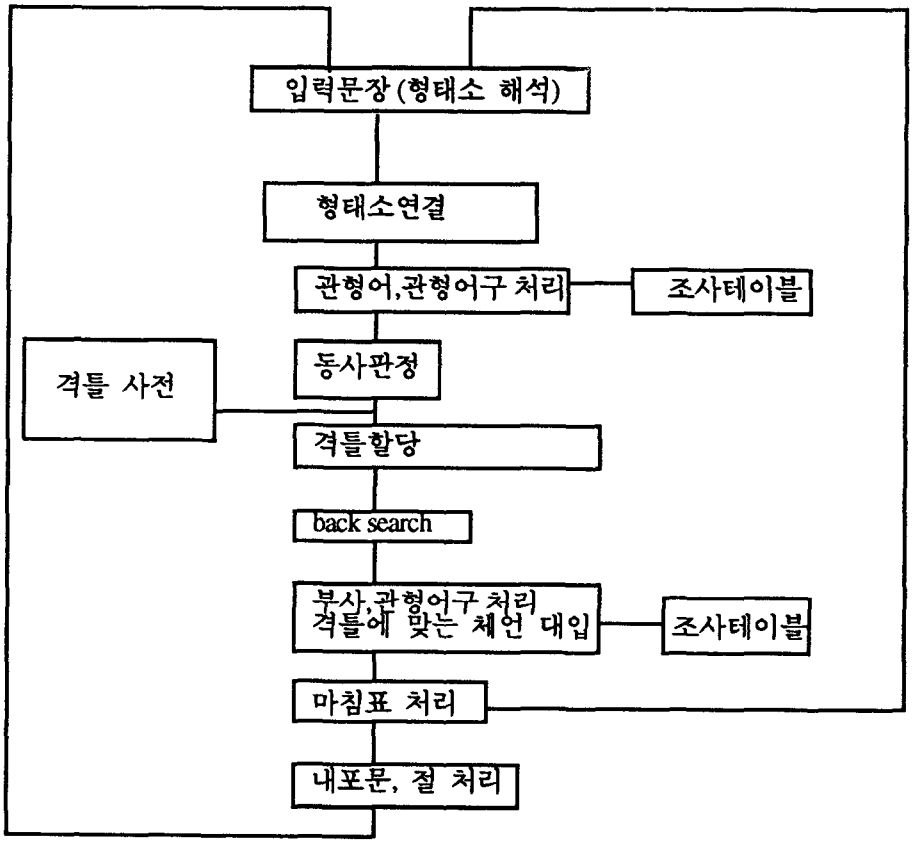

그림 4-2 <구문 분석 흐름도>

3.4. 한국어 통사 규칙 처리

한국어에서 주로 일어나는 통사규칙은 탈락 규칙, 첨가 규칙, 대치 규칙, 이동 규칙이 있다.[4] 첫번째, 첨가규칙에서 부정법은 단순히 조용사로 처리 가능하다. 피동규칙과 사동규칙의 경우는 피동, 사동문이 용언에서 확인되면 새로운 격틀을 만든다. 관계절화, 보문화는 내포문처리와 같은 방법으로 행한다. 두번째, 탈락 규칙은 의미 파악이 없이는

처리하기 불가능하므로 문장중에 나온 필수격만 처리하도록 했다. 세번째, 대치규칙도 마찬가지로 의미 파악이 없이 불가능하다. 각각의 대명사는 불용어로 제거된다. 마지막의 이동 규칙은 문장내에서 어떤 곳으로 이동하더라도 조사에 따라 격의 판별이 이루어 질 수있다. 부사의 이동은 문장에서 자유격으로 처리된다. 상승규칙의 경우는 격조사의 변화가 발생하므로 내포문처리와 같은 방법으로 처리한다. 이동규칙 중에서 다의성을 주는 주제화규칙의 경우는 본 시스템에서는 여러개의 주어가 올 수 있도록 하여 모든 가능한 격을 다 찾도록 하였다.

4. 불용어 처리

본 시스템에서의 불용어는 일반적인 의미의 불용어, 즉 문장중의 의미 없는 단어(조사, 어미, 접두어, 접미어, 보조어간...)와는 달리 자주 출현하는 단어로써 색인어가 될 수 없는 명사(common noun)로 정의 했다. 현재 본 시스템에서 이용하는 불용어 사전은 DACOM에서 발췌한 불용어 사전중에서 명사부분만 그리고 한국 산업 연구원에서 조사한 한국어 불용어[6]중에서 명사 부분을 이용하고 있다.

V. 결론 및 성능 평가

본 시스템의 평가는 직접 사람이 추출해 낸 결과를 가지고 자동색인 시스템의 결과와 다음과 같은 두가지 방법으로 비교 하였다.

REL = (HI - AIT) / HI
UNREL = (AI - HIT) / AI

REL = 자동색인 방법으로 뽑은 색인어 중 사람이 뽑은
색인어와 비교하여 올바르게 뽑힌 색인어의 비율
UNREL = 자동색인 방법으로 뽑은 색인어 중 사람이 뽑은
색인어와 비교하여 쓸데 없는 색인어가 뽑힌 비율

AI = 자동 색인으로 추출된 색인어
AIT = 자동 색인으로 추출된 색인어 중 사람이 뽑은 색인어와 같은 색인어
HI = 사람이 직접 추출한 색인어
HIT = 사람이 직접 추출한 색인어 중 자동색인으로 뽑은 색인어와 같은 부분

본 시스템에서 비교 대상으로 하는 색인어는 1990년 8월에서 10월 사이의 각종 신문기사 약 300개를 가지고 현재 DACOM '천리안 서비스'에 이용되도록 사람이 직접 색인어를 추출한 것이다. REL은 자동색인으로 추출된 색인어 중에서 사람이 색인어로 선택한 것과 비교하여 색인어 추출이 성공한 것을 말하고, UNREL은 자동색인어가 색인어로 잘못 선택한 것을 말한다.

앞에서 '색인어와 필수격에 대한 실험'과 마찬가지의 대상기사를 가지고 본 시스템의 최종결과와 비교하였다. REL의 경우는 약 15%의 결과가 나왔다. 즉 사람이 추출한 색인어의 85%정도가 자동색인으로 추출 가능하였다. UNREL의 경우 130%정도의 사람이 뽑은 색인어 보다 많은 색인어가 나왔다. 이러한 결과는 본 시스템에서 가정한 필수격의 색인어 자질이 크게 어긋나지 않고 잘 맞아 떨어 진다는 것을 알 수 있다.

완전한 문장의 의미 파악과 시소러스, 지식 베이스 기반을 사용한다면 더 좋은 색인어를 추출할수 있지만 현재의 여건으로 볼 때는 사용이 한정되어 있다. 따라서 본 시스템의 방법과 같은 것을 이용하되 더 좋은 제한조건과 해석방법을 이용한다면 훌륭한 자동색인어 시스템을 만들 수 있을 것이다.

VI. 참고문헌

[1] 강재우 "접속정보를 이용한 한글 철자 및 띄어쓰기 검사기의 설계 및 구현" : 한국과학기술원 석사논문 : 1990

[2] 남기심 "언어학 개론" : 탑출판사 : 1988

[3] 남용우 "격문법이란 무엇인가" : 을유문화사 : 1978

[4] 박영순 "한국어 통사론" : 집문당 : 1985

[5] 성균관 대학교 대동 문화 연구원 : "고등학교 문법" : 대한 교과서 주식회사 : 1988

[6] 안현수 "한글문헌의 자동색인에 대한 실험적 연구" : 한국산업연구원 : 1987

[7] 야촌조향 "자연언어처리의 기초기술" : 일본 전자정보통신학회 : 1988

[8] 이영주 "자동색인을 위한 한국어 형태소분석 알고리즘" : 한글날 기념 학술대회 발표논문집 : 240~246 : 1989

[9] 정영미 "정보 검색론": 정음사 : 1986

[10] Gerard Salton "Automatic Information Oragainzation and Retrieval" : McGraw-Hill Book Company : 1965

[11] Gerard Salton "Automatic Text Indexing Using Complex Identifiers" : ACM : 1988 : vo. 12

[12] Gerard Salton "On the Application of Syntactic Methodologies in Automatic Text Analysis" : Cornell University TR : 1988

[13] L.L. Earl, "Use of Word Government in Resolving Syntactic and Semantic Ambiguities" : ISR 9 : 639~664 : 1973

[14] NIKKEI ELECTRONINCS BOOK "구문해석 이용으로 특허문에서 키워드 96% 자동추출" : 1986

APPENDIX

1. 입력 문장

내년 중반부터 서울지역을 중심으로 전자전화번호부 서비스(EDS)가 실시되어 95년 전국으로 확대되며 이를 위하여 정보검색단말기 3백만대가 공급된다.

2. 형태소 해석 결과

(((내년 2))((중반 4)(부터 106))((서울 1)(지역 2)(을 103))((중심 4)(으로 101)) ((전자 4)(전화 2)(번호 4)(부 4)(서비스 8)(EDS 8)(가 100))((실시 2)(되다 210) (어 307))((95 370)(년 365))((전국 2)(으로 101))((확대 2)(되다 210)(며 284)) ((이 50)(를 100))((위하다 140)(어 307))((정보 4)(검색 2))((단말기 4)) ((3백만 370)(대 6)(가 100))((공급 2)(되다 210)(ㄴ 다 251)))

3. 구문해석 결과

```
==========================================
verb = 210 되다
   left = 실시
   right = 어
adverb ---------------------------------
   으로
   중심
------------------------------------------
   을
   서울/지역
------------------------------------------
   부터
   내년/중반
------------------------------------------
(2 - 1) case
   가
   전자/전화/번호/부/서비스/EDS
========= next connect ====================
verb = 210 되다
   left = 확대
   right = 며
adverb ---------------------------------
   으로
   95년/전국
------------------------------------------
(2 - 1) case not exit
========= next connect ====================
verb = 210 위하다
   right = 어
(1 - 1) case not exit
(2 - 2) case
   를
   이
========= next connect ====================
verb = 210 되다
   left = 공급
   right = ㄴ 다
(2 - 1) case
   가
   정보/검색/단말기/3백만/대
==========================================
```

4. 추출된 색이어

서울/지역
전자/전화/번호/부/서비스/EDS
정보/검색/단말기