코퍼스로부터 형태소 분석을 위한 사전 구성

정민수 · 정규철 · 조원홍

군산대학교 컴퓨터과학과
전북 군산시 미룡동 산 68번지
우:573-701
kihong, kcjung, wonhong@cs.kunsan.ac.kr

A Dictionay Composition for Morphological Analyzer from Corpus

Minsu Jung, Kyuchol Jung, Wonhong Cho
Department of Computer Science, Kunsan National University

요약

한국어나 일본어처럼 문법형태소의 기능에 의해 단어의 통사적, 의미적 역할이 결정되는 교착어에서는 형태소 분석이 통사 분석과 의미 분석에 미치는 영향이 크기 때문에 한국어의 분석에 있어서 형태소 분석은 아주 중요하다. 관형적 표현이 많은 한글은 문법 규칙만으론 분석하기가 쉽지 않고, 분기가 많이 생성되므로 오류가 발생할 확률도 높다. 이러한 문제점을 해결하기 위해 본 논문에선 사전을 중심으로 해결하고자 한다. 그러기 위해선 방대한 용량의 사전이 필요로 하게 되고 이를 구축하기 위한 시간과 노력이 요구되므로 이미 구성된 코퍼스를 이용해 사전을 구성하여 많은 시간과 노력을 줄일 수 있도록 한다. 그리고 생성되는 많은 분기 가운데 올바른 경로를 찾아가기 위해 코퍼스내의 각 태그 결합정보를 추출하고 추출한 결합정보의 통계정보-코퍼스내에서 사용된 빈도수-포함하여 우선순위를 정하도록 한다.

1. 서론

영어를 기반으로 이루어진 문법체계-변형생성문법, 지배속박이론, 어휘함수문법, 구구조문법류-를 한글에 적용시키기엔 무리가 따른다. 한글 자체가 어순이 자유롭고, 격을 결정하는 조사나 주체가 되는 단어등이 생략되는 현상이 빈번하기 때문에 형태소 분석을 하기 매우 어렵운게 현실이다. 그래서 형태소 분석은 간단한 구문규칙에 따라 사전을 중심으로 이루어지게 되므로, 사전의 역할은 매우 크다.[1] 예를 들어, 용언의 경우 규칙용언은 간단한 구문 규칙만으로도 분리할수 있지만, 규칙을 적용하기 어려운 불규칙 용언은 모든 불규칙 용언을 분석하거나 아니면 모두 사전에 등록시키는 방법 밖에 없다. 복합명사의 경우도 구분이 모호하기 때문에 구문분석이나 의미분석이 되지않는 한, 사전에 등록시키는 방법이 가장 현명하다. 이런 이유로 형태소 분석시 사전의 비중은 커지게 되고, 사전의 용량과, 탐색시 걸리는 시간이 형태소 분석기의 성능을 좌우하게 된다.

그래서 본 논문에서는 형태소 분석시 필요한 사전을 코퍼스를 통해 구성하고 분기를 줄이기 위해 통계정보를 사용하고자 한다. 이미 한국과학기술원에서 100만어절의 품사가 부착된 코퍼스가 구축되어 있으므로 이를 사용하면 적은 시간과 노력으로 대용량의 사전과 결합정보를 얻을수 있다. 2장에서는 사전의 구성과 각 사전의 역할에 대해, 3장에서는 결합정보와 각 사전사이의 관계, 사전의 추가에 대해, 마지막으로 구문분석에 적용하기 위한 계획으로 결론을 맺었다.

2. 사전의 구성

기존의 형태소 분석시 사용하는 주로 사용되는 사전은 태그사전이다. 태그 사전은 태그와 그에 해당하는 키워드로 구성된 일종의 품사사전으로 보통 수작업을 통해 구현한다[2]. 본 논문에서는 수작업을 통한 문제점을 보완하기 위해 이미 구축되어 있는 코퍼스에서 태그 사전을 구성하여 시간과 노력을 줄이고자 한다. 덧붙여서 태그간의 결합정보와 그 결합이 코퍼스내에서 사용된 빈도수를 나타낸 통계정보를 포함하여 결합정보사전를 구성한다. 분석시간을 줄이기 위해 한번 분석된 문형은 형태소 사전에 등록하여 재분석시 바로 사전을 통해 분석이 이루어지도록 한다.

이 세 사전에 사용한 코퍼스는 한국과학기술원에서 제공한 100만어절의 한국어 품사 부착 코퍼스이고, 여기에 사용된 54개 태그를 태그 사전의

분류기준으로 한다. 사전의 용량이 크므로 탐색시간을 최소화 하기위해 이중배열로 구현을 한다[3]. 이중배열은 트라이 형태이므로 최악의 경우에도 탐색시간은 O(n)을 넘지 않는다[4]. 입력시 사용하는 코드는 Unicode 2.0을 사용한다. 이는 현대 한글 글자 모두를 가나다 순으로 정렬, 배치한 것으로 현대 한글이 사용하는 자모 - 초성 19자, 중성 21자, 종성 27자 - 로, 조합 가능한 글자의 수 19 x 21 x 28 (종성없음 포함) = 11172자를 모두 사용할수가 있고, 초성, 중성, 종성으로 나누기 편하기 때문이다. 뿐만 아니라 일본이나 중국쪽의 코드도 포함하고 있기 때문에 후에 사전을 확장시 많은 잇점을 내포하고 있다. 하지만 사용할 코퍼스는 완성형으로 되어 있고, 현재 사용되고 있는 윈도우 역시 아직은 유니코드를 지원하지 않기 때문에 입력된 문장이나 찾고자 하는 단어를 입력 받을때 이를 유니코드로 변환시키는 알고리즘이 추가해야 한다.

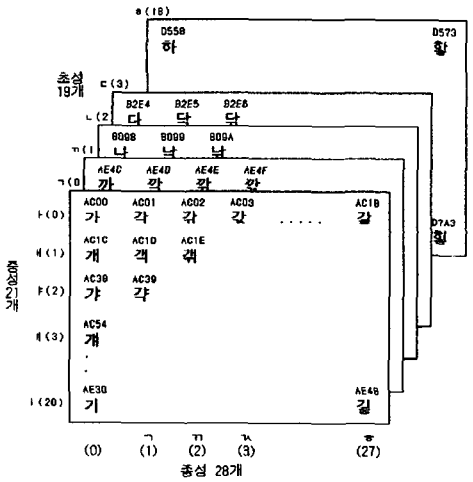


그림1. 유니코드 테이블

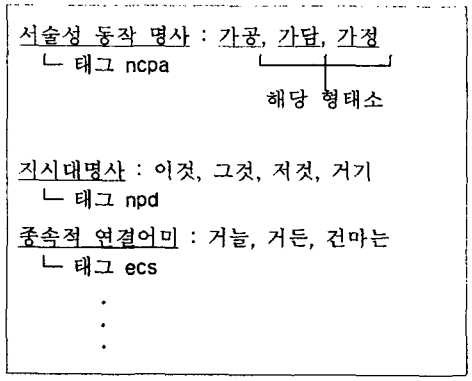
코드값 = OxAC00 + 초성 * 21 * 28
+ 중성 * 28
+ 종성

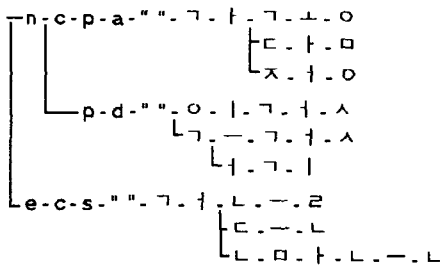
2.1 태그 사전

형태소 분석을 구성하는 데 있어서 가장 기본이 되는 사전으로, 본 논문에서 사용한 태그는 자연어 정보 처리 연구부에서 분류한 54개의 태그를 따르기로 한다[5].

한글은 대부분의 단어가 명사와 동사이고 기능어는 극히 일부이기 때문에 기능어에 대한 사전 정보를 미리 구축해 놓으면 계속적인 사전 항목의 추가는 명사와 동사에 한정 된다. 기능어 사전의 문법 정보는 매우 복잡하지만 일단 기능어 사전이 구축되면 새로운 갱신은 적다. 문제점은 하나의 형태소가 여러개의 태그로 사용되기때문에 형태소 분석시 많은 분기를 발생하게 되는데, 이를 해결하기 위해 새로이 결합 정보 사전을 구축하기로 하고 이는 뒤에 다시 설명한다. 사용한 품사부착코퍼스는 완성형으로 구성되어 있으므로 이를 유니코드로 바꾼후 2byte - 유니코드는 2byte 단위로 글을 나타내기때문 - 단위로 각 태그의 사전을 구성한다.

표1. 태그 사전 구성의 예





태그와 해당 형태소는 'space'로 구분을 하고 일반 트라이로 구현한다. 더블 트라이로 구현을 하면 사전의 용량을 줄일수 있지만 다른 사전과 link시키기가 어려우므로 일반 트라이를 택했다.

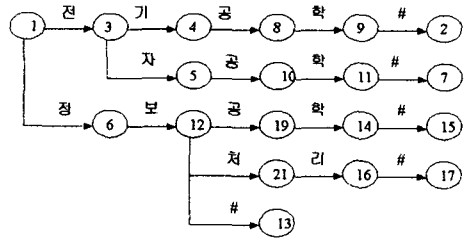


그림2. 트라이 예제

<표2. 각 문자에 해당하는 한글 표현치>

문자	#	전	기	자	정	보	공	학	처	리	사
내부표현치	1	2	3	4	5	6	7	8	9	10	11

<표3-1. 트라이에 대한 더블 배열>

Index	1	2	3	4	5	6	7	8	9	10
BASE	1	0	1	1	3	6	0	1	1	3
CHECK	21	9	1	3	3	1	11	4	8	5

<표3-2. 트라이에 대한 더블 배열>

Index	11	12	13	14	15	16	17	18	19	20	21
BASE	6	12	0	14	0	16	0	0	6	0	6
CHECK	10	6	12	19	14	21	16	0	12	0	12

다른 사전과 연결시 종결기호 "#"의 베이스값인 "NULL"대신에 해당하는 Index값을 음의 값으로 넣어 구현한다.

2.2. 결합정보사전

결합 정보 사전은 형태소 분석시 분기를 줄이기 위해 구성하는 것으로 기존에는 결합정보를 알고리즘을 통해 구성하거나 수작업으로 구성하였다. 이때에도 수작업을 거치므로, 시간도 많이 소요되고 오류 발생의 소지도 있다. 그래서 이를 보안하기 위해, 구축된 코퍼스에서 태그 사전을 구성시 결합정보를 추출하고, 보다 정확성을 높이기 위해 통계정보-코퍼스내에서 태그결합이 사용된 빈도수-를 추가하여 결정된 결합정보에 순위도 결정지을 수 있도록 한다. 태그간의 결합정보는 앞의 태그 사전과 구문규칙만을 사용할때 발생하는 분기를 줄이기 위해 코퍼스로 부터 태그를 추출할 때 같이 구성하여 불필요한 시간을 줄인다. 결합정보는 뒤에서 형태소 사전과 연결되므로 index를 포함하여 구성을 한다. 하나의 단어를 여러개의 결합정보로 분석되는 경우도 있으므로 통계정보에 따른 형태소 분석 순위는 반드시 필요하다.

결합정보사전은 형태소사전과 태그사전과는 달리 그리 크지 않고 탐색 기준이 Index와 결합정보, 양쪽에서 행해질수 있도록 다루기편한 더블링크로 구현을 한다. 이는 이중 배열로 구현된 사전의 종결 기호의 베이스값에 음의 값으로 연결하고, 태그의 결합정보와 index, 통계정보는 각각 'space'로 구분한다.

표4. 결합정보사전의 예

```
nq 1 1
   │ └─ 코퍼스에서 쓰여진 횟수
   └─── index
└──── 태그
nq+ncpa+jcm 2 1
      └─ 결합정보
nq+ncn 3 1
nq+jcm 4 2
nq+ncn+jxc 5 3
ncn 6 2
```

```
n-q-" "-1-" "-1
  └+-n-c-p-a-+-j-c-m-" "-2-" "-1
       │ └n-" "-3-" "-1
       │   └+-j-x-c-" "-5-" "-3
       └j-c-m-" "-4-" "-2
└c-n-" "-6-" "-2
```

2.3. 형태소 사전

기존의 경우 많은 문장을 분석할 때에는 같은 문형도 매번 다시 분석이 되고, 형태소 분석을 할 경우엔 한 문장내에서도 같은 형태가 매번 같은 규칙에 따라 분석되므로 시간이 낭비된다. 이러한 점을 해결하기 위해 한번 분석되어진 형태를 사전으로 구축하여 같은 형태의 경우 이미 분석되어진 결합형태를 사전에서 찾아 시간을 줄이기 위해 구성한다. 또 관형적, 관습적인 글이 자주 쓰이는 한글의 경우 이 형태소 사전은 많은 분석 오류와 시간을 줄일 수 있다.

분석된 형태소와 결합정보 사전의 index를 연결하여 구성하고 각각은 'space'로 구분한다. 한글을 포함하므로 코드 변환후 트라이로 구성한다.

표5. 형태소 사전의 예

```
남극 1
한국+의 4
남극+진출+의 2
세종+기지+는 5
서남극 6
```

```
ㄴ-ㅏ-ㅁ-ㄱ-ㅡ-ㄱ-" "-1
                └+-ㅈ-ㅣ-ㄴ-ㅊ
                      -ㅜ-ㄹ-ㅇ-ㅓ-" "-2
ㅎ-ㅏ-ㄴ-ㄱ-ㅜ-ㄱ-+-ㅇ-ㅓ-" "-4
ㅅ-ㅐ-ㅈ-ㅗ-ㅇ-ㄱ-ㅣ-ㅈ-ㅣ-ㄴ-ㅡ-ㄴ-" "-5
ㅅ-ㅓ-ㄴ-ㅏ-ㅁ-ㄱ-ㅡ-ㄱ-" "-6
```

3. 사전간의 관계

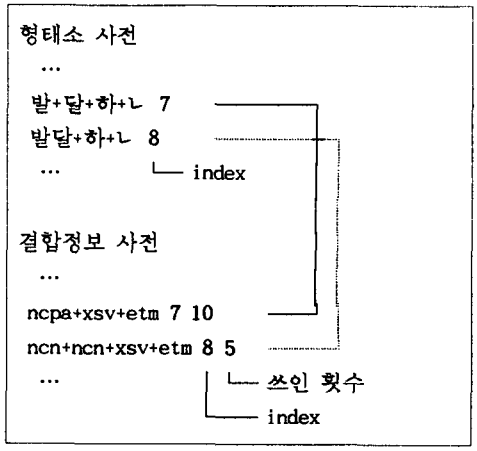
3.1 결합정보사전과 형태소 사전

결합 정보 사전과 형태소 사전은 index를 통해 연결한다. 이것은 어떤 형태소를 분석시 먼저 형태소 사전을 검색하여 분석되어진 형태를 모두 찾아내고, 그 결과를 결합 정보 사전의 index를 통해 얻을 수 있기 때문이다. 같은 형태소가 서로 다른 결합정보를 가질 경우 결합정보 사전의 통계 정보-코퍼스내에서 쓰여진 빈도수-에 따라 순위를 결정한다.

예를 들어 "발달한"을 형태소 사전에서 '발달+하

+ ㄴ' 과 '발+달+하+ㄴ"으로 분석했을 경우 index를 통해 결합 정보 사전에서 검색하면 "ncpa + xsv + etm" 과 "ncn + ncn + xsv + etm"을 찾을 수 있고 각각의 통계정보를 비교하여 선택한다. 예제에서는 전자의 경우가 많이 쓰이므로 선택될 확률이 높게 된다.

표6. 형태소사전과 결합정보사전의 관계



3.2 태그 사전과 결합정보 사전

기존의 형태소 분석시 주로 사용하는 것은 태그 사전이고 부수적으로 분기를 줄이기 위해 결합정보가 사용된다. 이 결합정보는 의존문법류나 구구조문법류에서 사용되지만 보통 수작업이나 간단한 구문규칙의 적용에 그치고 있다. 그래서 결합정보를 보다 정확히 하기 위해 코퍼스를 통하여 많은 양의 통계정보를 포함하는 결합사전을 구성하였다.

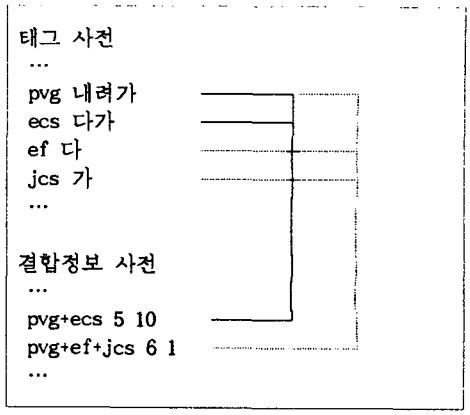
이는 코퍼스에 등록되어 있지 않은 단어를 분석하고자 할때 사용된다. 형태소 사전에서 발견되지 않는 단어는 먼저 문법규칙에 따라 성립할 수 있는 태그를 태그사전을 통해 검색하고, 검색된 태그 중 문법규칙에 맞는 결합정보를 결합정보 사전에서 찾게 된다. 이 과정에서 발생하는 여러 분기를 결합정보의 통계정보에 따라 순위을 설정하여 선택하게 된다.

예를 들어 "내려가다가" 라는 단어가 등록이 안되어 있을 경우 이를 분석하면, 먼저 태그 사전을 통해 다음 태그들을 검색된다.

내려가/pvg, 다가/ecs, 다/ef, 가/jcs

이 태그들로 구성할 수 있는 결합을 결합정보사전을 통해 검색하고 그결과를 통계정보에 따라 순위을 정하게 된다. 위의 태그는 보통 내려가/pvg+다가/ecs, 또는 내려가/pvg+다/ef+가/jcs 등으로 검색이 되는데, 보통 쓰이는 빈도가 전자가 많으므로 선택될 순위가 높게 되는 것이다.

표7. 태그 사전과 결합정보 사전의 관계



3.3 사전의 추가

코퍼스를 통해 사전에 추가할 경우엔 이미 분석되어 있으므로 문제가 없지만 일반 문장을 분석하여 추가 할 때에는 다른 사전과 마찬가지로 사용자의 동의를 얻어야 한다. 사전의 등록은 품사부착 코퍼스에서 태그가 먼저 등록이 되고, 그 과정에서 태그와 태그와의 결합정보를 등록하고, 분석된 실제 단어를 형태소 사전에 등록하는 순으로 이어진다. 코퍼스가 아닌 실제 문장을 등록할 때는 형태소 사전에서 먼저 형태를 검색하고 발견되지 않으면 새로운 유형이거나 코퍼스에서 분석되지 않은 유형이므로 태그를 검색하고 그결과로 구성할수 있는 결합을 사전 결합정보를 통해 분석하고, 마지막 통계정보를 사용하여 각분기의 우선순위에 의해 사용자가 선택하여 사전에 등록하게 된다.

4. 결론

한글을 기존의 사전으로 분석하기엔 관형적인 표현을 다루기 어려운 문제점이 있고, 또 이를 구축할 때에 수작업을 거치므로 많은 시간과 노력이 필요로 한다. 이러한 문제점을 해결하기 위해 이미 구축된 코퍼스를 이용하여 관형적인 표현과 결합정보를 구성하여 시간과 노력을 줄이고 한다. 한글은 영어나 기타 다른 언어에 비해 매우 자유로운 언어이다. 그래서 형태소 분석하기 어려운게 현실이고 사전에 의존하는 방법이 현제로써는 가장 올바르다고 본다. 아직 구문간의 정보까지는 고려하지 않았기 때문에 완전한 형태소분석은 어렵고 형태소 사전에 등록할 때에도 코퍼스로 부터 구성하지 않는 경우엔, 사용자의 확인을 거쳐야 하는 문제점이 있다.

후에 본 사전은 구문 분석에서 활용할 수 있도록 구문간의 관계를 추가하여 형태소 분석과 구문분석에서 활용할 수 있도록 할 계획이다.

감사의 글

본 논문을 쓰기까지 도와주신 박기홍 교수님과 조원홍 교수님에게 감사 드립니다.

참고 문헌

[1]김덕봉, 최기선, "효율적인 한국어 형태소 해석 방법", 한국과학기술원 전산학과, 1994

[2] 김병희, 임권묵, 송만석, "형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기", 연세대학교 전산과학과, 1994

[3] Katsushi Morimoto, Hirokazu Iriguchi And Jun-Ichi Aoe, "A Method Of Compressing Trie Structures", University Of Tokushima, Pp.265-278, Mar 1994.

[4] 이승선, 송주원, 황규영, 최기선, "TRIE구조를 이용한 한국어 전자 사전을 위한 데이타베이스 인덱스 구조," 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1, 1994

[5] 김재훈, 김덕봉 등, "통합국어정보베이스를 위한 한국어 형태, 통사 태그 설정", Computer Systems Lab. internal memo. 1996

[6] 강 승식, 김 영택, "사전 정보에 기반한 효율적인 한국어 형태소 분석기의 설계 및 구현", 한국정보과학회 봄 학술발표논문집 18권 1호, pp 529-532, 1991.

[7] 최재혁, 이상조, "양방향 최장일치법을 이용한 한국어 형태소 분석기," 한국정보과학회 학술발표논문집, 제 20권 1호, pp. 769-772, 1993.

[8] 나동렬, "한국어 파싱에 대한 고찰", 연세대학교, 구문분석관련 연구논문집 p33 ~p46