코퍼스로부터 구문 분석을 위한 사전 구성

정민수* · 정규철** · 박기홍***
군산대학교 컴퓨터과학과

A Dictionary Composition for Syntactic Analyzer from Corpus

Minsu Jung, Kyuchol Jung, Kihong Park
Department of Computer Science, Kunsan National University

요 약

한글은 종심어 후행성과 어순의 자유성, 격을 결정하는 조사의 생략 등으로 인해 영어권에서 연구되어진 변형 생성 문법이나 어휘 함수 문법, 구구조문법류 등이 적용되기 어려운 문제점을 가지고 있고 관형적인 표현이 많아 구문 규칙 만으론 분석하기 쉽지 않기 때문에 사전에 의존해야 하는 경우가 많으므로 이에 적합한 사전을 구성하고자 한다. 그러나 기존의 태그된 키워드만으로 구성된 사전만으로 어려운 점이 많고, 이 때문에 문법 규칙을 같이 적용하게 되는데 이 규칙을 보통 알고리즘을 이나 수작업을 통해 사전으로 구성하므로 정확성도 떨어진다. 저자는 이 과정을 코퍼스를 통해 구성하여 시간을 줄이고 결합 정보 또한 보다 신뢰하게 구성하기 위해 통계 정보-코퍼스 내에서 결합이 사용된 빈도-에 따라 순위를 결정할 수 있도록 구성하였다. 이를 보다 확장하여 구문분석 시에도 활용할 수 있도록 분석된 단어간의 결합 정보와 그 결합이 사용된 빈도를 포함하여 구문 결합 정보 사전을 구성하고자 한다. 이는 기존의 의존 문법이나 구문 관계를 이용하여 구문 분석을 할 경우 올바른 트리의 결합 관계를 검색할 때 소요될 수 있다.

1. 개 요

영어를 기반으로 이루어진 문법 체계 중에서 한글의 형태소 분석에 쓰이는 문법은 구구조문법류와 의존문법 류가 있는데, 어순이 자유로운 한글에는 의존 문법을 적용시키는 방법이 더 적합하다[1] 하지만 분석 시 의존 관계가 많이 발생하므로 이를 줄이기 위해서 문법 규칙과 사전을 이용하는데, 이 때 분석 시간의 대부분을 소비한다. 본 논문에서는 그 소요시간을 줄이고자 결합 가능한 형태소가 가질 수 있는 구문 결합 정보를 코퍼스로부터 사전으로 구축하여 해결하고자 한다. 또한 분기된 가지 중 올바른 순위를 결정하기 위해 각 결합 정보에 통계 정보-코퍼스 내에서 쓰인 빈도수-를 포함하고자 한다. 이는 형태소 분석 시 형태소간의 결합 정보를 사전으로 구성 할 때도 같이 적용된다.

사전은 형태소 분석 시 태그 사전과 형태소 결합 정보 사전을 사용하고 구문 분석 시 형태소 결합 정보 사전과 구문 결합 정보 사전을 사용한다. 태그 사전은 기존의 태그와 형태소로 이루어진 사전을 말하고, 형태소 결합 사전과 구문 결합 사전은 본 논문에서 구성한 사전으로 전자는 코퍼스로부터 얻은 형태소간의 결합으로 구성한 사전이고, 후자는 단어간의 결합 정보를 통계정보와 같이 모아 구성한 사전이다. 분석 시 사전에 의존할 때 발생하는 분기는 각 사전의 통계 정보를 통해 줄이고자 한다.

2. 사전의 구성

본 논문에서는 사용하는 사전은 태그와 그에 해당하는 키워드로 구성된 기존의 태그 사전과, 형태소 분석시 태그간의 결합 관계와 통계 정보를 포함하여 분기를 줄일 있도록 구성한 형태소 결합 정보 사전, 두 단위로 단어간의 결합을 통계적으로 구성한 구문 결합 정보 사전으로 이루어진다. 본 사전과 같이 용량이 크게 되면 탐색 시간이 성능에 중요한 영향을 미치게 되므로 탐색 시간을 O(n)이내-탐색하고자 하는 단어의 전체 수 이내-로 하기 위해 트라이로 구성을 하고[2] 이는 이중 배열을 이용해 구현한다[3]

본 사전을 구성하기 위해 사용한 코퍼스는 한국과학기술원에서 제공한 100만 어절의 품사 부착 코퍼스와 1만 문장의 구문 구조 부착 코퍼스를 사용하였다. 전자는 태그 사전과 형태소 결합 사전을, 후자는 단어 결합 정보 사전을 구성하는데 쓰였다. 태그의 분류 기준은 코퍼스에 쓰인 기준을 따르고, 사전 입력시 사용하는 코드는 Unicode 2.0으로 한다. 이는 현대 한글 글자 모두를 가나다 순으로 정렬, 배치

* 군산대학교 컴퓨터 과학과 대학원생
** 군산대학교 컴퓨터 과학과 대학원생
*** 군산대학교 컴퓨터 과학과 교수

한 것으로 현대 한글이 사용하는 자모 - 초성 19자, 중성 21자, 종성 27자 - 로 조합 가능한 글자의 수 19 x 21 x 28 (종성없음 포함) 11,172자를 모두 사용할 수가 있고, 초성, 중성, 종성으로 나누기 편하기 때문이다

2.1 태그 사전

태그 사전은 코퍼스에서 분류한 54개의 태그[4]와 그 태그에 해당하는 키워드를 트라이로 구성한다 한국어에서 대부분의 단어는 명사와 동사이고 기능어는 극히 일부이기 때문에 기능어에 대한 사전 정보를 미리 구축해 놓으면 계속적인 사전 항목의 추가는 명사와 동사에 한정된다. 기능어 사전의 문법 정보는 매우 복잡하지만 일단 기능어 사전이 구축되면 새로운 갱신이 적은 편이다 사용된 품사 부착 코퍼스는 완성형으로 되어 있기 때문에 구현 시 유니코드로 변환 후 태그별로 추출하여 구성한다. 태그와 그에 해당하는 키워드는 'space'로 구분을 한다. 탐색 시간은 최악의 경우에도 트라이의 레벨이 n일 때 O(n)이내이다

예제1. 태그 사전의 구성 예

```
...
nq 한글 ( 000001 )
태그 형태소 (linked number)
nq 세종 ( 000002 )
...

000001 - 예비
(linked number) - 예비
```

linked number는 사전 구성 시 마지막 종결문자의 Base값에 다른 정보와 연결하기 위한 Index값을 음수로 기록한다. 이 Index를 통해 배열이나 liked list형태로 형태소에 대한 정보를 추가 할 수 있다.

2.2. 형태소 결합 정보 사전

형태소 결합 정보 사전은 형태소 분석시 분기를 줄이기 위해 구성하는 것으로 기존의 경우 결합 정보를 알고리즘을 통해 구성하거나 수작업을 통해 사전으로 구성하였다. 그래서 시간도 많이 소요되고, 수작업을 거치게 되므로 오류 발생의 소지도 있다. 이를 보안하고자 코퍼스를 통해 결합 정보를 추출하여 보다 정확성을 높이고 통계 정보-코퍼스 내에서 태그 결합이 사용된 빈도수-를 추가하여 결정된 결합 정보의 순위를 정할 수 있도록 하였다 하나의 형태소가 여러 결합의 형태로 나타날 수 있으므로 우선 순위를 정하기 위해 통계 정보도 포함한다

형태소간의 결합 정보와 index, 통계 정보는 "space"로 구분을 하고, 빈도수에 대한 정보는 각 태그결합정보마다 유일하므로 배열로 구현한다 정보를 추가할 필요가 있을 경우엔 linked list로 구현한다. Index는 linked number를 통해 정하고 태그결합정보를 역으로 찾아 갈수 있도록 종결문자의 Index를 포함하여 구성한다.

예제2. 태그 결합 사전의 구성 예

```
.
nq+jca ( 000013 )
태그결합정보 (linked number)
ncps+xsm+ecx ( 000005 )
ncpa+xsv+etm ( 000008 )
..

000013 - 000089 - 000010
(linked number - 빈도수 - 종결문자의 Index)
```

2.3. 구문 결합 정보 사전

구문 분석은 크게 구구조 문법이 문장을 여러 개의 구성 성분으로 나누어서 분석하는 방법과, 각 어형간의 의존 관계에 기반을 두고 분석하는 의존 문법으로 나뉜다[1] 두 방법 모두 문장 구조가 트리를 형성하게 되고, 이때 불필요할 분기를 없애기 위해 분석된 형태소로 구성할 수 있는 결합을 찾게 되는 데, 본 논문에서는 이를 해결하고자 구문간에 결합이 이루어질 수 있는 통계 정보를 사전으로 구성 시 포함하여 해결하고자 한다.

여기에 사용된 구문 태그는 7가지로 명사 구절(NP), 동사 구절(VP), 형용사 구절(ADJP), 부사 구절(ADVP), 관형사 구절(MODP), 독립 구절(IP), 보조용언 구절(AUXP)이고[6], 입력된 문장은 구문 태그를 기준으로 다시 나누어 사전으로 구성한다 구문 태그 내의 형태소 결합 정보는 형태소 결합사전의 liked number로 구성을 하고 각각은 'space'로 구분한다

예제3. 구문 결합 정보 사전의 구성 예

```
...
ADJP 000013+000014 ( 000020 )
구문태그, 구문결합정보, (linked number)
VP NP+000014+000015 ( 000021 )
NP 000016+000017 ( 000022 )
...

000020 - 000030 - 000040
(linked number) - 빈도수 - 종결문자의 Index
```

3. 사전간의 관계

본 논문에서는 구문 분석 시 생성된 분기에서 올바른 경로를 찾을 수 있도록 구문 결합 정보 사전을 구성한다 이 구문 결합 정보 사전에서 liked number를 통해 연결된 사전이 필자가 전에 구성한 형태소 결합 정보 사전이고, 형태소 결합 사전은 다시 태그 사전과 연결되어 있다. 그래서 어떤 문장을 분석하고자 할 때 먼저 태그 사전을 통해 그 문장이 가질 수 있는 태그와 그에 해당하는 키워드를 찾게 된고 그 키워드로 조합할 수 있는 형태를 형태소 결합 정보 사전을 통해 탐색하게 된다. 이때 발생하는 분기를 줄이기 위해 하나의 단어가 가질 수 있는 결합을 통계 정보로 순위를 정한다. 이때 모든 순위가 바르게 정해지진 않는다. 예를 들어 문장 내에서 "도와" 라는 문장을

다음 두 가지로 구분할 수 있는데 이 모두 많이 사용하는 형태로 통계정보만으로 순위가 정해지지 않는다

명사"도" + 조사"와"
동사"돕" - 어미"와"

위와 같은 경우 뒤에 동사가 오는지 명사가 오는지 구문 결합 정보 사전을 통해 결합될 수 있는 다음 형태를 확인해야 바른 분석이 이루어 질 수 있다.

4. 실험 및 평가

사용한 코퍼스는 97년 7월에 한국과학기술원에서 제공한 100만어절의 품사부착 코퍼스와 1만어절의 한국어 구문구조부착 코퍼스를 사용하였다. 이는 완성형으로 작성되었기에 테스트에 앞서 유니코드로 변환하였다 아직 사전 구축 시 배열상의 문제로 100만어절을 모두 분석하진 못 하고 약 10만 어절을 분석하여 이를 기초로 프로그램상의 오류를 검사하고 있다. 구문분석기용 사전도 약 1000 문장을 사전으로 구축하여 검사하고 있다.

품사부착코퍼스는 약 10000어절 ~ 15000어절 단위로, 구문구조부착코퍼스는 100~300문장 단위로 추가하면서 구성시간과 탐색시간을 측정하였다 탐색시간은 입력한 문장을 다시 읽어들여 걸리는 시간으로 측정하였다. 시간측정을 위해 프로그램 내에 타이머를 추가하였고, 오류를 분석 위한 코드도 포함하여 측정하였으므로 실제 탐색시간은 더 빠르다.

실험 1. 결합정보사전의 구성단계

	태그 수	구성시간	탐색시간
1차	8121	485초 19	30초 54
2차	13121(+5000)	424초 25	34초 02
3차	17316(+4195)	139초 34	25초 24
4차	21607(+4291)	242초 54	32초 08
5차	26011(+4404)	251초 42	32초 11
6차	28519(+2508)	185초 24	30초 52
7차	32339(+3820)	282초 21	33초 42

실험 2. 구문정보사전의 구성단계

	구문정보 수	구성시간	탐색시간
1차	2752	382초 11	24초 54
2차	4285(+1533)	274초 54	18초 43
3차	6936(+2651)	320초 47	23초 24
4차	7691(+755)	241초 04	14초 01
5차	9378(+1687)	289초 42	19초 58

결합사전 구성 시 1, 2차에선 구성시간이 길었지만 점차 시간이 짧아진다. 추가되는 태그와 결합정보가 줄어들기 때문이다. 구문정보사전은 아직 많은 문장을 분석하진 못했지만 역시 점차 구성시간이 줄어들었다. 탐색시간은 1000어절을 기준으로 30초 전후이다

아직 C언어가 유니코드를 기반으로 만들어진 언어가 아니기 때문에 문자열 처리에 문제가 있고, 기존의 코퍼스에서 구문구조 분석시 사전에 등록되는 않는 형태소가 발생하는 문제점이 있다.

5. 결론

한글은 어순이 자유롭고, 관형적인 표현이 많이 사용되기 때문에 구문 규칙만으로는 다루기가 어렵고 그 규칙 또한 일기 힘들다 그래서 본 논문에서는 코퍼스를 통해 형태소간의 결합 규칙과 단어간의 결합 규칙을 구절 단위로 사전을 구축하여 해결한다. 그리고 사전에 의존시 발생하는 분기를 줄이기 위해 통계 정보를 형태소 결합 사전에 포함하여 구성한다. 아직 의미 정보까지는 고려하지 않았기 때문에 완전한 구문 분석은 어렵고 사전에 구성시 트라이를 구현할 때 충돌문제로 인해 시간이 많이 소요되는 문제점이 있다.

본 사전은 앞으로 맞춤법 검사기와 일한 번역기에서 사용할수 있도록 확장할 계획이다

참고 문헌

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

[2] 이승선, 송주원, 황규영, 최기선, "TRIE 구조를 이용한 한국어 전자 사전을 위한 데이타베이스 인덱스 구조," 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1, 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] 김재훈, 김덕봉 등, "통합국어정보베이스를 위한 한국어 형태, 통사 태그 설정", Computer Systems Lab. internal memo 1996

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

[6] 이공주, 김재훈, 강병규, 최기선, 김길창, "한국어 구문 트리 태깅 코퍼스 작성을 위한 한국어 구문태그" KAIST, Department of Computer Science 1996

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

[8] 양덕희, "한국어 전자사전 원형의 설계 및 구현", 연세대 대학원 석사 학위 논문, 1991.

[9] 권혁철 최준영, "단일화 기반 의존분법을 이용한 한국어 분석기", 한국정보과학회 논문지 Vol 19, 1992, 9