자동 구축된 구문패턴사전과 규칙을 이용한 구묶음

임지희 최호섭 이정철 옥철영
울산대학교 컴퓨터정보통신공학과 한국어처리연구실
jhlim@mail.ulsan.ac.kr, {hoseop,jclee,okcy}@ulsan.ac.kr

Chunking Using Automatic Constructed Syntactic Pattern Dictionary and Rule

Jihui Im Hoseop Choe Jungchul Lee Cheulyoung Ock
Korean Language Processing Laboratory,
Dept. of Computer Engineering and Information Technology,
University of Ulsan

요 약

본 논문은 실용적인 구문분석기의 전단계로서, 자동 구축된 구문패턴사전과 규칙을 이용하여 구묶음하는 방법을 제안한다. 우선 규칙은 구문분석 말뭉치(30,875어절)를 대상으로 자동 추출된 고빈도의 규칙(Rewriting Rule)을 본 논문에 맞게 수동으로 구축하였다. 규칙은 조건부, 행위부로 이루어진 이진 규칙(binary rule)의 형태를 이루며, 명사구(NP), 수식어구(AP, DP), 인용구(X), 용언구(VP, VC)을 대상으로 15개를 구축하였다. 그리고 구문패턴은 중심어와 중심어 선행 요소의 특성뿐만 아니라 중심어 후행 요소도 고려하여 형식화시킨 것으로, 중심어의 복합용언 여부에 따라 일반용언패턴과 본+보조용언패턴으로 구분한다. 부분적인 언어 현상의 처리보다는 실세계에서 사용되는 수많은 문장들에 내재되어 있는 매우 광범위한 언어 현상의 처리를 하기 위해, 구문패턴은 형태소주석 말뭉치(460만 어절)을 대상으로 자동 구축하였다. 구축된 구문패턴사전과 규칙을 이용하여 구묶음을 수행한 결과 정확율 83.09%가 나타났다.

1. 서 론

자연언어처리 시스템은 최근까지 형태소분석, 구문분석, 의미분석 등 효과적인 자연언어의 분석과 활용을 위해 많은 연구가 진행되어 왔다. 이 중 구문분석 기술은 기계번역을 비롯한 정보검색, 언어이해기술 등에서 많은 연구가 진행되었으나, 분석 기술의 문제와 정확성 문제 등의 이유로 좋은 연구 성과를 얻지 못하고 있는 실정이다. 특히 한국어의 경우, 문장 표현의 자유로움, 각 어휘에 따른 개별적인 특징, 일반적인 문법 구조를 벗어난 문장 등의 문법적, 의미적 특성으로 인해, 다른 언어의 구문분석기 개념 모델을 차용할 수 있지만 그대로 활용할 수는 없다. 그러므로 다양한 한국어처리 시스템의 개발을 위한 기초 작업으로 실용적인 구문분석기의 개발은 필수적이다.

실용적인 구문분석기는 부분적인 언어 현상의 처리보다는 실세계에서 사용되는 수많은 문장들에 내재되어 있는 매우 광범위한 언어 현상의 처리를 목적으로 한다. 이를 위해서 대용량의 말뭉치에서 개별 어휘에 따른 다양한 정보들을 수집하여 규칙화하고, 규칙화된 정보들을 자동으로 구축하는 방법을 모색함으로써 한국어 문장의 구문 분석에 활용할 필요가 있다.

구문분석에 이용될 정보를 자동으로 구축하고자 하는 시도가 있었는데[1-3], 문장의 중심요소인 용언의 특정 후행요소에 따라 형성되는 패턴이 다른 점을 간과하였으며, 작은 말뭉치를 대상으로 구축하여 그 결과가 특정 용언에 한정되는 문제가 있다. 또한 구문분석 말뭉치에서 자동으로 추출된 Rewriting Rule이 문맥을 고려하여 적용되기보다 빈도만을 이용하여 적용되는 문제점이 있다.

따라서 본 논문에서는 규칙화된 정보를 이용하기 위해 460만 어절을 대상으로 구문패턴을 자동으로 추출하여 구문패턴사전을 구축하는 방법과 자동 구축된 구문패턴사전과 규칙을 이용하여 구묶음(Chunking)하는 방법을 제안한다. 구문패턴은 중심어와 중심어 선행 요소의 특성뿐만 아니라 중심어 후행 요소도 고려하여 규칙에 따라 형식화시킨 것으로, 중심어의 복합용언 여부에 따라 일반용언패턴과 본+보조용언패턴으로 구분하며, 각 패턴은 패턴을 이루는 요소와 쓰임새에 따라 실질어

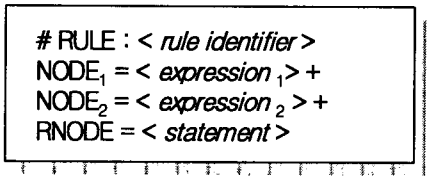
를 포함한 어휘패턴, 조사의 나열로만 이루어진 품사패턴, 표준패턴으로 구분한다. 표준패턴은 구문패턴의 부족문제를 해결하기 위해 동사/형용사에 따른 고빈도 품사패턴으로 구성된다. 그리고 규칙은 실용적인 구문분석의 전단계를 수행하기 위한 본 논문의 목적에 맞게 구문분석 말뭉치를 기반으로 하여 수동으로 구축하였다. 이렇게 구축된 구문패턴사전과 15개의 구묶음 규칙을 구문 분석 말뭉치(30,875어절)를 대상으로 구묶음(chunking)을 수행하였다.

3. 규칙(basic Rule)

본 논문은 구문분석 말뭉치(30,875어절)를 대상으로 자동 추출된 고빈도의 규칙(Rewriting Rule)을 수동으로 구축하였다. 본 논문의 구묶음은 구문분석의 전단계로 진행되었으며 수동구축된 규칙도 이를 목적으로 15개가 구축되었다. 각 규칙은 분석할 노드의 종류에 따라 다른 규칙을 먼저 호출할수 있는 형태로 구축되어 적용범위가 넓다.

각 규칙은 명사구(NP), 수식어구(AP, DP), 인용구(X), 용언구(VP, VC)를 대상으로 하며, 여기서 용언구는 동사, 형용사, 지정사, 및 보조용언이 중심어가 되는 구(phrase)를 가리킨다. 특히, 지정사와 보조용언은 동사, 형용사와 달리 결정적으로 구묶음을 수행할 수 있으므로, 그에 대한 규칙을 따로 설정·적용한다.

[그림 1]는 이진 규칙(binary rule)으로 표현된 규칙의 형식이다. 규칙의 종류를 나타내는 규칙 식별자(rule identifier)는 구묶음을 수행할 때 분석 어절에 따라 적용할 규칙을 선별하게 한다. 그리고 expression은 각각 노드(NODE1, NODE2)에 대한 조건을, statement는 노드(NODE1, NODE2)에 대한 조건을 만족시켰을 때 생성될 결과 노드(RNODE)에 대한 행위를 기술한다. statement는 단순하게 NODE1, NODE2의 정보를 조합하여 결과 노드에 데이터를 입력할 뿐만 아니라, 단방향으로 구묶음을 수행하였을 때 생기는 문제점을 해결하기 위해 다른 규칙을 호출하는 함수를 포함하기도 한다.



[그림 1] 규칙의 형식

4. 구문패턴

구문패턴은 말뭉치에서 나타나는 특정 개별 어휘(중심어)와 이와 공기하는 일련의 성분들을 규칙에 따라 형식화시킨 것으로, 중심어의 복합용언 여부에 따라 일반용언패턴과 본+보조용언패턴으로 구분한다. 중심어가 단일용언인 일반용언패턴과 달리, 본+보조용언패턴은 본용언에 보조용언이 부착되어 논항 정보가 달라지거나 추가적인 성분을 요구하게 되는 특징을 고려하였다. 조정미(1998)에서는 이런 현상을 '격이동 현상'이라 하여 선택 제한 지식 추출의 방해 현상으로 분류하였다. 즉, '본용언+보조용언'를 하나의 구문 단위로 묶어준 뒤, 구문패턴을 생성하여야 정확한 구문패턴을 생성할 수 있다.

패턴을 이루는 요소, 쓰임새에 따라서 실질어를 포함한 어휘패턴(LP : Lexical Pattern), 조사의 나열로만 구성된 품사패턴(PP : POS Pattern), 표준패턴(SP : Standard Pattern)으로 구분한다. 품사패턴과 표준패턴은 패턴을 이루는 요소는 동일하지만, 생성 방법과 쓰임새가 다르다. 즉, 품사패턴은 개별 용언에 따라 패턴을 그룹핑하여 고빈도 패턴만을 선별하고, 표준패턴은 품사(동사, 형용사)에 따라 패턴을 그룹핑하여 패턴을 선별한다. 표준패턴은 구묶음(Chunking) 수행시 어휘패턴과 품사패턴을 이용한 패턴 매칭에 실패하거나 분석하고자 하는 용언에 대한 어휘패턴, 품사패턴이 없을 때 이용된다.

5. 구문패턴사전 구축 방법

구문패턴사전은 대량의 말뭉치에서 주변 문맥 정보를 이용하여 형태소주석 말뭉치(350만 어절)과 금성사전(110만 어절)을 대상으로 구축되었다.

5.1 구문패턴의 구성요소 추출

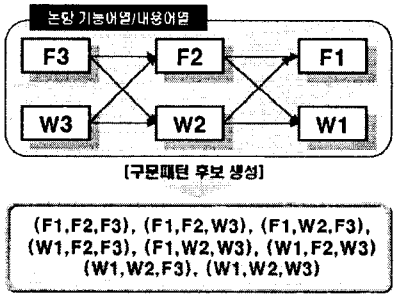
구문패턴의 구성요소는 중심어 후행의 원리에 의해, 본용언과 본용언 사이로 추출 범위를 지정하여 용언의 인접어절 정보를 중심으로 추출한다. 용언의 선행 요소 중에서 주격조사(JKS), 부사격조사(JKB), 목적격조사(JKO)를 취하는 어절만을 추출하고, 용언의 후행 요소 중에서 어미와 다음어절을 추출하였다. 선행요소가 보조사(JX)를 취하는 어절은 그 격의 모호성 때문에 추출대상에서 제외하였다. 즉 구문패턴의 구성요소는 논항 기능어열/내용어열, 어간 품사/어휘, 어미 품사열/어휘열, 다음 어절의 기능어열/내용어열이며, [그림 2]는 위와 같은 방법으로 구문패턴 구성요소를 추출한 모습이다.

[그림 2] 구문패턴 추출기(USPE)

5.2 구문패턴 후보 생성

추출된 구문패턴의 구성요소 중 '논항 기능어열/내용어열'을 조합하여 구문패턴 후보를 생성한다. 다음 [그림 3]은 구문패턴 후보를 생성하는 방법이다. 여기서 F는 조사정보를 ,W는 어휘+조사정보를 나타내며, 선행요소가 K개이면 확장·추가된 구문패턴 후보는 2K-1개이다. 생성된 구문패턴 후보는 어휘화된(lexicalized) 구문패턴의 자료 희소성의 문제를 해결할 수 있다.

본 연구에서는 용언 후행요소가 관형형 전성어미(ETM)인 경우에는, 용언과의 연관도가 선행 요소보다 피수식어가 더 높다고 판단하여 위 구문패턴 후보 확장 방법을 적용하지 않고 구문패턴 생성에서 제외한다.



[그림 3] 구문패턴 후보 생성 방법

5.3 구문패턴사전 구축

구문패턴 후보를 어미의 품사열/어휘열, 다음 어절의 기능어열/내용어열의 용언 후행요소에 의해 일반용언패턴 후보, 본+보조용언패턴 후보로 분류한 다음, 이것을 다시 어휘패턴 후보, 품사패턴 후보, 표준패턴 후보로 분류한다. 그리고 패턴확률(P)이 임계치 θ이상이고 빈도(F)가 ρ이상인 구문패턴 후보를 추출하여 구문패턴사전에 구축한다.

-어휘패턴 & 품사패턴

$$P_{kj} = \frac{F_{kj}}{\sum_{i=1}^{N} F_{ki}} > \theta$$

P_{kj} : 용언 k의 j번째 구문패턴 후보의 패턴확률

F_{ki} : 용언 k의 I번째 구문패턴 후보의 빈도

-표준패턴

$$P_j = \frac{F_j}{\sum_{i=1}^{N} F_i} > \theta$$

패턴을 이루는 요소에 따라 품사패턴과 어휘패턴으로 구분하여 구문패턴을 생성하는 이유는, 품사패턴에 비해 어휘패턴의 빈도가 상대적으로 너무 낮기 때문에 다수의 의미있는 어휘패턴이 구문패턴사전에 추가되지 않고 버려지는 경우가 발생하기 때문이다. 또한 표준패턴은 품사패턴/어휘패턴의 부족 현상, 패턴 매칭 실패에 대한 해결책으로 제시한다. 다음 [그림 4]는 '가하다'_V의 구문패턴을 나타내면, [표 1]은 위와 같은 방법으로 구축된 구문패턴사전의 구문패턴 갯수이다.

[가하다_V]
<어휘패턴>
- N+에/JKB 신중/NNG+을/JKO
- N+에/JKB 만전/NNG+을/JKO

<품사패턴>
- N+에/JKB N+을/JKO
- N+을/JKO

[그림 4] '가하다' V의 구문패턴 예

[표 1] 구문패턴사전의 구문패턴 개수

	일반용언패턴	본+보조용언패턴
어휘패턴(LP)	1052개	93개
품사패턴(PP)	11999개	5997개
표준패턴(SP)	74개	69개

6. 구묶음 알고리즘(Chunking Algorithm)

본 논문에서는 품사 부착된 문장을 입력으로 하며, [그림 5]는 구묶음 알고리즘의 개요이다.

우선 입력 문장 내의 각 어절에 대해 CNode형 인스턴스를 생성함으로써, 각 어절에 대한 초기 phrase tag(NP,VP,...), function tag(SBJ, OBJ,...) 등의 정보를 설정하여 U_chart에 기록한다(2-3행). U_chart[x][y]에는 문장의 x위치에서 전방향으로 y개의 범위에 이르

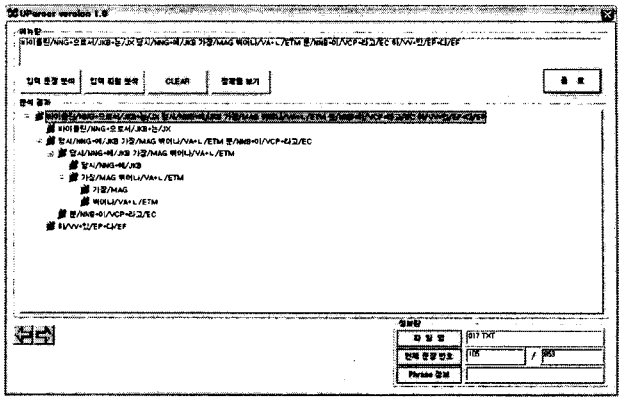
는 Node가 기록되어 있다. 이와 같은 초기 할당은 학습 말뭉치에서 각 형태소 태그가 어떤 phrase tag, function tag를 가장 자주 할당받았는지를 근거하여 설정한 규칙에 의해 이루어졌으며, 보조사(JX)와 같이 function tag를 할당하기 애매한 경우, 별도의 function tag를 설정하였다.

초기 할당이 이루어진 다음, left-to-right 방향으로 반복하여 구묶음을 한다(4-27행). 각 노드에 대해 우선 인용구문에 대한 검사를 한다. 만약 Node가 인용부호를 포함한다면, 인용구를 먼저 구묶음한 후, 인용구 내의 Node들에 대해 구묶음한다(7-8행). 그리고 각 분석 노드의 phrase tag에 따라 수동으로 구축된 규칙을 적용한다. 만약 분석 노드의 phrase tag가 'VP'이면, 지정사 혹은 보조용언 구문인지를 검사하여 별도의 규칙을 적용하고, 그렇지 않은 경우, 구축된 구문패턴사전에서 어휘패턴, 품사패턴, 표준패턴을 차례대로 검색 · 매칭시킨다(13-22행). 문장의 각 노드에 대해 분석을 마쳤을 때, U_chart에 Node수가 증가하지 않으면 구묶음 알고리즘은 종료된다. [그림 4]은 위 알고리즘을 적용한 프로그램의 실행 모습이다.

본 논문의 구묶음 알고리즘은 비재귀적인 구를 인식하고자 하는 구묶음과 달리, 구문분석기의 전처리 단계로서 구현되었다. 그래서 분석이 애매한 부분에서는 범위가 작은 구(phrase)를 생성하기보다 큰 구(phrase)를 생성함으로써, 분석을 보류하여 구문분석기에서 담당하도록 한다.

```
1)  void chunking() {
2)  for all words in the sentence
3)    make a node and put into U_chart[node_position][1];
4)  loop {
5)    for node_position-1 to sentence_size {
6)      current_node = U_chart[node_position][1]->act_node;
7)      if qutation is included current_node
8)        apply quotation rule to current_node;
9)      switch current_node's phrase tag {
10)       case NP : apply NP_rule to current_node; break;
11)       case AP :
12)       case DP : apply modifier_rule to current_node; break;
13)       case VP :
14)         if current_node's content_word include 'VCP' or 'VCN'
15)           apply VC_rule to current_node;
16)         else if current_node's next node's phrase tag == VX
17)           apply VX_rule to current_node;
18)         else {
19)           search syntactic patterns from the pattern dictionary;
20)           try to match syntactic patterns;
21)           if current_node's function tag == ETM
22)             apply modifier_rule to current_node;
23)         }
24)     }
25)   if count of U_chart's cell is unchanged
26)     break;
27) }
```

[그림 5] 구묶음 알고리즘



[그림 6] 구묶음 프로그램 실행 모습

7. 실험 및 분석

본 논문은 구문분석 말뭉치(30,875어절)를 이용하여 학습 말뭉치는 28,740어절, 검증 말뭉치는 2,135어절으로 나누어 사용하였다.

구축된 구문패턴사전과 규칙을 이용하여 학습 말뭉치와 검증 말뭉치에 실험한 정확률은 [표 2]와 같다.

[표 2] 실험결과

	학습 말뭉치	검증 말뭉치
규칙	93.67%	93.24%
규칙+구문패턴	84.26%	83.09%

규칙만을 이용하여 구묶음한 결과가 93.24%로 높게 나타났다. 그러나 이것은 구문분석의 전단계로서 구묶음을 수행하므로, 정확하게 구묶음할 수 있는 경우만을 규칙으로 설정하여 새로운 구를 생성하고자 하는 원칙에 의해, 명사 연결 형태와 같이 분석이 모호한 경우에는 보다 큰 구를 생성하거나 분석을 다음 단계로 보류하였기 때문이다. 즉, 아래와 같은 명사 연결구의 경우에는 보다 큰 구를 생성하며, 생성된 구 내부의 상세한 분석은 구문분석 단계에서 처리하고자 하였다.

예) 산업/NNG 공동/NNG+화/XSN 우려/NNG
→[산업/NNG 공동/NNG+화/XSN 우려/NNG]_NP

그러므로, 구묶음 결과 생성된 구의 개수가 규칙+구문패턴을 이용한 경우보다 규칙만을 사용한 경우가 더 적게 나타났기 때문에, 정확률은 더 높게 나타났다.

8. 결론 및 향후 연구 방향

구묶음의 처리 대상과 그 범위는 실제 사용될 시스템의 목적에 따라 달라질 수 있다. 본 논문에서는 실용적인 구문분석기의 전단계로서, 구문패턴사전과 수동으로 구축된 재귀적인 규칙을 이용하여 구묶음하는 방법을

제안하고자 하였다. 수동으로 구축된 규칙의 수는 적지만 재귀적인 형태를 가짐으로써, 분석 노드의 특징에 따라 분석 방향을 달리하여 융통성을 가진다. 그리고 대량의 말뭉치에서 자동으로 구축된 구문패턴사전은 하위범주화 사전, 선택제약 사전 등을 구축하는 기본 자료로서 활용할 수 있다.

본 논문에서 수동 구축된 규칙에 대한 세밀한 분석을 통한 규칙 추가와 알고리즘의 성능 향상, 구문분석기와의 연동 방법 등 계속적인 노력과 연구가 필요하다.

9. 참고문헌

[1] 김나리, 패턴 정보를 이용한 한국어 구문분석, 서울대 박사학위논문, 1997.

[2] 박준식, 품사 패턴을 이용한 한국어 병렬 구문의 해석, 한국과학기술원 석사학위논문, 1998.

[3] 송만석 외, 한국어 처리를 위한 격틀의 자동 구축, 정보통신부 연구개발 결과 보고서, 1998.

[4] 조정미, 코퍼스와 사전을 이용한 동사 의미 분별, 한국과학기술원 박사학위논문, 1998.