구조 변환을 겸한 영어 구문 분석기

*여상화[O] **서정연

*경인여자대학, **서강대학교

*shyuh@kic.ac.kr, **seojy@ccs.sogang.ac.kr

Structural Transfer through English Parsing

*Sanghwa Yuh[0] **Jeongyuh Seo

*Kyungin Women's College

**Sogang University

요 약

본 논문에서 제안하는 영어 Parser는 Bottom-Up Best-First Chart Parser를 기반으로 영어 구문 분석과 동시에 영->한 구조변환을 수행한다. 따라서, 영어 입력문에 대한 파서의 결과는 한국어 구문 Tree가 된다. 구문 분석과 변환을 동시에 수행하므로, 번역 과정을 단순화 하고, 번역 지식 관리가 용이하여 번역기의 튜닝이 용이하다. 구현된 파서는 대규모 문법 규칙에 대응하기가 용이하고, 새로운 규칙의 추가가 용이하여 번역기의 점진적인 성능 향상이 가능하다.

1. 서 론

영어(SVO)와 한국어(SOV)와 같이 구조가 상이한 언어간의 자동번역에 있어서 원시언어의 구조를 밝히는 구문분석기(Parser)는 필수적이다. 대부분의 영한 기계번역시스템은 번역 모델로서, 원시 언어의 구문 구조로부터 상대 언어(Target Language)로의 구조 변환 과정을 수반하는 변환 방식(Transfer Approach)을 채택한다[8].

영한 기계번역을 위한 영어 구문 분석기에 대한 연구는 규칙 기반, 통계적 구문 분석, 숙어 기반 구문 분석, 문틀기반 분석 등 다양한 연구가 진행되어 왔다. 이들 시스템들은 변환 방식의 번역 모델에 기초하여, 입력 문장인 영어로부터 정확한 구문 구조(Parse Tree)를 얻는 것을 목표로 한다. 그러나, PP-Attachment 등 다양한 구문적 중의성(Syntactic Ambiguity)으로 인해 원시 문장으로부터 정확한 구문구조를 밝히는 것은 대단히 어려운 문제이다. 이로 인해, 원시언어에 대한 분석의 정확도(Accuracy)에 따라 번역의 질(Quality)이 좌우되는 기계번역의 특성상 현재까지도 상품성 있는 제품이 개발되지 못하는 실정이다.

이러한 문제점의 돌파구로서, 다양한 통계 정보를 이용하여 전역적/국부적인 중의성을 해결하여 분석의 정확률을 높이는 시도[1,6]와 함께 원시 문장의 분석의 깊이를 낮추기 위한 시도로 숙어 기반 구문 분석([2,3]과 PBMT(Pattern-Based MT), TDMT(Transfer-Driven MT)[9], EBMT(Example-Base MT) 그리고 문틀 기반 분석방법[4,5] 등의 다양한 번역 방법론들이 제안되었다.

본 논문에서는 이들 두 가지 접근 방법들의 장점을 취하여 Bottom-Up Best-First Chart Parser를 기반으로 영어 구문 분석과 동시에 영->한 구조변환을 수행하는 방법을 제안한다. 즉, 분석의 중의성 해소를 위해 Parsed Corpus로부터 추출한 통계정보를 이용하고, 분석의 결과는 상대언어의 구문 구조를 곧바로 생성한다.

구문 분석과 변환 과정을 분리하는 것은 각각의 모듈의 독립성을 확보할 수 있으나, 영어와 한국어와 같이 구조 변환 과정이 필수적인 시스템에서는 다음과 같은 비경제적인 요인이 발생한다.

- 원시언어(예: 영어)와 상대 언어(예:한국어)에 대해 각각 별개의 문법 모형이 필요하며, 서로 다른 문법 모형은 서로 다른 번역 지식을 요구한다.
- 구문 분석 문법(Grammar)과 구조 변환 규칙을 각각 개발하고 유지하여야 하며 이들 규칙을 유지/관리 하는 데는 상당히 고비용이 필요하고, 이로 인해 튜닝이 매우 어렵다.

2. 기존의 연구

원문의 분석과 함께 상대언어의 생성을 수행하는 방법론으로는 PBMT(Pattern-Based MT), TDMT(Transfer-Driven MT)[9], EBMT(Example-Base MT), IBMT(Idiom-Based MT)[10] 그리고 문틀(Sentence Frame)기반 방법[4,5] 등이 있다. PBMT는 CFG Parser에 기반하여 어휘화된(Lexicaliized) CFG문법을 사용한다. TDMT는 변환기에 의해 번역이 이루어지며, 변환 지식과 입력문장의 유사도 계산에 의해 최적의 변환 규칙들이 선택되는 예제기반 번역의 일종이다. 변환정보는 원시언어표현(SLE; Source Language Expression)과 상대언어 표현(TLE)를 포함한다.

EBMT는 병렬 코퍼스로부터 번역 틀(Translation Template)를 자동으로 습득하고, 입력 문장과의 유사도 계산에 의해 최적의 번역 예를 선택하고, 이로부터 상대언어의 번역 문장을 생성한다. EBMT의 번역 지식은 어휘화된 번역틀에 기반하므로, Coverage가 매우 낮아 제한된 영역에서만 활용된다. 숙어기반 방법은 기존의 변환방식(Transfer)의 진보된 형태로, 구문 분석 전단계에서 숙어들을 인식하고, 인식된 숙어 정보[1]를 이용하여 구문분석기의 분석의 부담을 줄이고, 숙어 정보에 포함된 상대 언어 표현을 이용하여 변환기가 변환을 수행한다. 숙어기반방법은 전역적인 분석에 앞서 지역적으로 숙어를 인식하고, 인식된 숙어는 평면적인 구조를 가져 구문 분석의 효율을 높인다. 문틀기반의 방법은, 구문 분석 전 단계에서 동사, 접속사, 기호 등을 Protector로 인식하고, Protector 와 Protector 사이의 성분들에 대해서 부분적인 구문분석 PBP(Partial Parsing between Protectors)을 수행하여 입력 문장에 대한 문틀을 생성하고, 미리 구축된 문틀정보를 이용하여 상대언어의 대역 구조를 생성하는 방법이다.

기존의 방법론과 본 논문에서 제안된 방법을 정리하면 표 1과 같다.

표 1. 변환 방식 기계번역 방법론

EBMT	NO	변환	TE	유사도	매우낮음
TDMT	NO	변환	TE	유사도	낮음
PBMT	YES	파서	TE	가중치	높음
IBMT	YES	파서	ST	가중치	높음
SFBMT	NO	변환	TE	속성매칭	낮음
제안된 방법	**YES**	**파서**	**TE**	**통계적 CFG**	**높음**

(주): SE: 원시언어 표현, TE: 상대언어표현, ST: 원시언어 트리

3. 구문분석/변환기

3.1 통계적 구문분석/변환기

본 논문에서 제안하는 영어 구문분석/변환기는 Bottom-Up Best-First Chart Parser를 기반으로 한다[6]. 문법 규칙은 Penn Treebank(이하 PTB)로부터 추출한 통계적인 CFG에 기반하며 Table Look-up 방식으로 동작한다. 사용하는 Non-Terminal은 [6]에서와 같이 5개로 제한하여, 각각의 문법 규칙이 충분한 Context를 갖도록 하여, 분석의 정확률을 높인다.

표 2. Non-Terminal

NP	명사구	S	문장(Root)
NPL	Base NP	SS	절
TOINF	부정사구		

[1] 숙어 표현에는 상대언어 표현을 포함하고 있다

NPL(NP lowest)은 Base NP라고도 하며, NP의 성분 내에 다른 Non-Terminal을 포함하지 않는 것이다.

동일한 성분 열에 대해 상대언어 표현이 여럿이 있는 경우, 일반적인 통계적 구문분석과 동일하게 문법 규칙이 가진 확률값과 각 단어에 대한 품사 태깅 확률을 이용하여 가장 값이 큰 것을 선택한다.

3.2 통계적 문법 규칙의 추출

PTB의 Parsed Corpus의 각각의 구문 트리에 대해 TXL(Tree Transformation Language)[7]로 작성한 Tree 변환 규칙을 적용하여, 영어 문장에 해당하는 한국어 구구조 Tree Corpus를 작성하고, 이로부터 영->한 구문 변환을 포함하는 통계적인 CFG(이하 PCFG)를 추출한다.

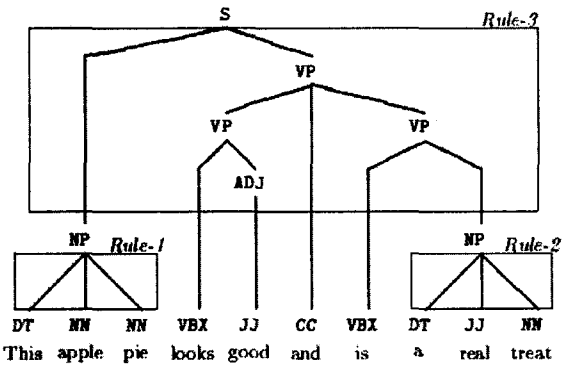

그림1. PTB 형식의 Parsed Corpus의 예

```
NP: DT NN NN                    //Rule-1
:score 1
:struct KR "(NP <1> <2> <3>)";[2]
NP: DT JJ NN                    //Rule-2
:score 1
:struct KR "((NP <1> <2> <3>)";
S: NP VBX JJ CC VBX NP    //Rule-3
:score 1
:struct KR "(S (<1> (KCASE subj))
                (VP (<3> <2>))
                <4>
                (VP <6> <5>))) ;
```

그림 2. 추출된 PCFG의 예

Rule3에서 첫번째 NP는 문장의 주어이므로, 주어 성분임을 표시하기 위해 "(KCASE subj)" 속성 정보가 추가되었으며, 두개의 VP는 어순이 변경되므로 각각 (VP (<3> <2>)) 와 "(VP

[2] 영어 파서로서의 기능을 유지하고, 상대어로의 구문 구조 작성시 원문의 구조를 참조하기 위해 영어 구조도 ":struct EN"의 형태로 보존한다.

<6> <5>)"의 형태가 된다. 구축된 PCFG에 의한 구문 분석 결과는 영어와 한국어 간의 구조 변환에 한정하는 것으로 어휘 자체에 따라 서로 다른 구조를 갖는 경우는 현재 반영하지 않는다. 그림 1의 영어 원문에 대한 본 구문분석/변환기의 분석 결과는 다음과 같다.

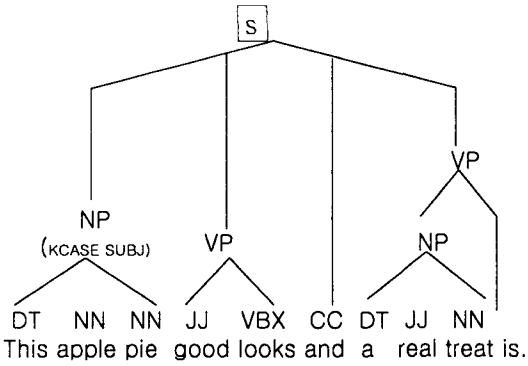

그림 3 분석 예

4. 구현 및 실험

본 시스템은 VisualC++ 6.0을 이용하여 WindowsXP에서 NYU의 Apple Pie Parser[6]를 기반으로 구현되었다. 문법 규칙은 Penn Treebank(PTB) Ver. 2.0에서 추출하였다. PTB 00-23 section에서 추출한 PCFG 규칙의 수는 다음과 같다.

표3. 문법 규칙

규칙의 종류	갯수	설명
NP	7,549	명사구
NPL	24,908	Base NP
S	16,624	문장(Root)
SS	16,941	절
TOINF	2541	부정사구
합계	68,533	

문제점으로는 5개의 Non-Terminal을 채택하는 경우, VP, ADJP,ADVP, PP 등을 Non-Terminal로 추가하는 경우보다 Context가 충분하여 Recall과 Precision은 높지만, 추출된 문법이 68,533개로 매우 많아 분석 시간이 오래 걸리고, 문장이 길어지면 Memory Limitation으로 인한 분석실패의 가능성이 높아진다. 또한, 학습 Corpus의 증가에 따른 문법 규칙의 수가 PTB 전체에 대해서도 수렴하지 못하는 단점이 있다.

5. 결론

본 논문에서는, 자동번역을 위한 구문분석기로서 구문분석과 구조 변환을 동시에 수행하는 영어 구문분석기를 제안하였다. 본 논문에서 제안한 방법은, [5]와 달리 통계적인 CFG 문법을 사용하는 Parser에 기반하며, PCFG에 의해 영어 원문의 분석을 위해 복잡한 해석 정보(예: 동사 타입 정보) 없이도 PTB 스타일의 완벽한 구문 트리를 생성하는 [6]과 달리 분석의 깊이는 상대언어의 구문 생성에 초점이 맞춰져 있으며, 규칙의 추가 삭제가 용이하도록 추출된 PCFG의 가독성(Readability)을 높였다

본 논문에서 제안한 방법은 다음과 같은 장점을 가진다.

- 구문 분석의 범위를 자동번역의 관점으로 제한하여 불필요한 심층 분석을 방지한다.
- 분석과 변환을 동시에 수행하므로 번역 절차를 간소화하여 번역시스템을 단순화, 고속화가 가능하다
- 정보의 중복 기술에도 불구하고, 분석 규칙을 직관화 하여 새로운 규칙의 추가가 용이하다. 이로 인해, 번역시스템의 튜닝이 쉽고 점진적인 성능향상을 이룰 수 있다.

향후 연구과제는 다음과 같다.

- 구문 생성을 포함하는 PCFG의 자동 생성
- 최적의 Non-Terminal Set 결정
- 구문분석기의 Coverage를 향상
- 동일한 품사라 하더라도 어휘에 따른 생성문의 변화를 반영

감사의 글

본 논문을 위해 APP의 사용을 허락하고 도움주신 NYU의 Satoshi SEKINE 교수께 감사드립니다.

참고문헌

[1]강원석, 영한 기계번역에서의 전치사구의 의미해석, 한국과학기술원 전산학과 박사학위논문, 1995

[2]운성희, 영어-한국어 기계번역을 위한 숙어 기반의 효율적 문장 분석, 서울대 컴퓨터공학과 박사학위논문, 1993.

[3]H.Jung, Sanghwa Yuh, Taewan Kim, and Sangkyu Park, A Pattern-based Approach using Compound U&nit Recognition and Its Hybridization with Rule-based Translation, in Journal of Computational Intelligencrce, Vol. 15, No.2, 1998.

[4]Y.Seo, Et al., "CaptionEye/EK:English-to-Korean Cation Translation System Using the Sentence Pattern," in Proce. Of MT Summit VIII, 2001.

[5]최승권 외 5인, "문틀기반 영한 자동번역 시스템," 제 12회 한글 및 한국어 정보처리 학술대회 발표논문집

[6]Satoshi SEKINE, Ralph GRISHMAN, "A Corpus-based Probabilistic Grammar with Only Two Non-Terminal, " IWOPT95.

[7] James R. Cordy, *The TXL Programming Language (Ver 10.2)*, http://www.txl.ca

[8]Chul-Min Sim, Hanmin Jung, Sanghwa Yuh, Taewan Kim, and Dong-In Park, "An Implementation of English-to-Korean Machine Translation System for HTML Documents," IASTED1998

[9] Osamu FURUSE and Hitoshi IIDA, "Tranfer-Driven Machine Translation," FGNLP92

[10]Hagyu Lee and Young Taek Kim, "An Idiom-based Approach to Machine Translation,"