MATES/CK 중한기계번역시스템의 구문분석규칙[1]

송영미 강원석 김지현 송희정 황금하 최기선
{ymsong,korterm}@world.kaist.ac.kr
한국과학기술원, 전산학과, 전문용어언어공학연구센터, 첨단정보연구센터

Parsing Rules for MATES/CK

Young-Mi Song, Won-Seok Kang, Ji-Hyoun Kim, Hee-Jeong Song, Jin-Xia Huang
Key-Sun Choi
AITRC, KORTERM, Dept. of Computer Science, KAIST

요 약

중한기계번역시스템(MATES/CK)의 구문분석은 1120개의 구문분석규칙과 통계적 정보에 의한 확률기반에 따라 그 문장에 가장 적합한 구문트리를 찾아져 적용되는 방식으로 이루어지고 있다. 기존 구문분석규칙은 자체에 오류가 많고, 새로운 규칙의 생성도 필요하다. 규칙에 대한 제약조건에도 좀 더 구체적이고 정확성을 높일 수 있는 상태로의 전환이 필요하다. 본 논문에서는 중한기계번역시스템(MATES/CK)의 구문분석의 정확도를 높이기 위하여 구문분석규칙을 수정하는 방법에 관하여 알아보고 그 연구과정을 살펴본다.

1. 서 론

중한기계번역시스템 MATES/CK (Machine Translation Environment System/Chinese-Korean) 의 구문분석 과정은 문장에 적합한 각 단어의 정확한 품사 태깅을 전제로 한다. 품사가 정확하게 태깅되고나면, 1120개의 구문분석규칙 및 확률기반에 따라 그 문장에 가장 적합한 구문트리가 찾아져 적용되고 적합한 번역 변환규칙에 따라 번역이 이루어진다. [1]

우선 가장 기본적이면서, 최소단위의 규칙을 구성하는 구문분석규칙의 적용된 상태를 살펴보기 위해서, 우리가 가지고 있는 중국어 6만 코퍼스 중 길이 4음절이하의 완전한 문장 100개를 테스트셋으로 하여 기존의 중한기계번역시스템(MATES/CK)의 정확도를 테스트했다. 테스트셋의 문장은 주로 두, 세 단어로 구성되며, 각각 하나의 문장으로 인식되어지는데, 테스트 결과는 57%에 불과했다. 만약 이 길이 이상의 문장이나, 복문 혹은 특수구문인 경우라면 정확도는 더 낮아질 것이다.

본 논문에서는 중한기계번역시스템(MATES/CK) 구문분석의 오류를 파악하고, 각 오류에 따른 해결 방안과 그 방법 그리고 실험 후의 결과에 대해서 구문분석규칙을 중심으로 분석해 볼 것이다.

[1] 본 연구는 첨단정보기술연구센터 과제 "다국어 정보검색 연구"와 전문용어언어공학연구센터 과제 "중국어-한국어 정렬을 통한번역지식의 획득'을 통하여 과학재단의 지원을 받았음.

구문분석규칙은 단어의 품사 및 문법정보속성을 이용하여 그 제약조건이 만들어지는데, 각 단어마다 품사 및 문법정보속성이 기록된 **现代汉语语法信息词典**(이하 PK-dictionary라고 지칭)을 기준으로 한다. [2]

본 작업에서는 위에서 언급한 테스트 셋 100문장을 통해 기존의 구문분석규칙이 놓치고 있는 문법정보속성을 활용하여 좀 더 완벽한 구문분석규칙을 형성시키고자 했다.

2. 중한기계번역시스템의 구문분석규칙의 기준과 구성

구문분석의 오류를 파악하기에 앞서 중국어의 각 품사와 구문분석규칙의 기준, 구문분석규칙이 어떻게 구성되어 있는지 알아보겠다.

2.1. 구문분석규칙의 기준

현재 품사표기와 구문분석표기는 북경대학의 **<汉语短语标注标记集>**을 따른 것이다. [3] (**<汉语短语标注标记集>** 송희정 역 : 부록첨가)

시스템에서 사용되는 중국어의 기본 품사는 26개이다. 구문분석규칙의 ap, dp, mp, np, pp, sp, tp, vp, 등은 각각 형용사구, 부사구 , 수사구, 명사구, 전치사구, 처소구, 시간사구, 동사구를 나타낸다. 중국어에는 abar, mbar, nbar, vbar라는 구문이 있는데, 각각 준형용사구, 준수사구, 준명사구, 준동사구라고 번역할 수 있다. 이 구문은 품사간의 매우 밀접한 관계로써 형성된다. 또한, dj, fj, yj, zj 등은 각각 단문, 복문, 인용문, 완전문장을 나타낸다. 또 하나 특이한 사항은 독립성분표기인 dlc로 문장 내의 독립성분이 특수 표점의 부가로 구성된 것이다.

2.2. 구문분석규칙의 구성

MATES/CK의 구문분석규칙은 모두 1120개이다. 규칙은 알파벳 순으로 정리되어 있으며, 다음과 같은 구성을 가지고 있다.

```
#VP-->v+v[964]      규칙[규칙번호]
CenterNode=0        중심단어
@NextCate=!v        제약조건①
@0:体谓准:谓        제약조건②
<<<<<               번역변환규칙
要+1:v-->1:v+아야 한다|기를 원한다|고 싶다
会+1:v-->1:v+을 수 있다
.....
>>>>>
```

「그림1」

「그림1」의 [규칙번호]는 알파벳 순으로 정리된 규칙의 위치이다. CenterNode=0은 중심어가 화살표의 바로 다음인 0번째 오는 단어란 의미인데, 구문을 형성하는 품사의 순서는 화살표의 바로 다음이 0순위이며, + 기호로 이어져 1,2,3.....순으로 표시한다. 제약조건은 그림의 제약조건①과 같이 약속된 기호로 만들어지기도 하고, 제약조건②와 같이 단어의 품사별 속성을 이용하기도 한다. <<<<<와 >>>>>의 사이에는 예외적으로 번역되어야 할 번역변환규칙과 기본적인 변환규칙이 있다.

PK-dictionary에는 단어마다 각 품사의 속성이 제시되어 있는데, 이를 이용한 것이 제약조건②이며 현재 구문분석규칙에서 쓰이는 품사별 속성은 형용사 11, 구별사 3, 접사 2, 부사 1, 방향사 4, 수사 2, 명사 3, 양사 2, 대명사 3, 처소사 2, 시간사 1, 동사 16가지등 총 50개가 쓰이고 있다.

3. 중한기계번역시스템의 구문분석오류의 원인

중한기계번역시스템(MATES/CK)의 구문분석 정확도가 57%로 낮은 것에는 여러가지 원인이 있다. 구문분석 오류의 원인을 살펴보자.

3.1. 구문분석이 실패한 경우

그 원인은 문장을 구문분석할 수 있는 규칙이 없기 때문이다. 또한, 단어정보가 PK-dictionary에 없을 때,

품사태깅에 오류가 생겨 적합한 규칙이 찾아지지 않는 것도 그 이유이다. 단어정보가 PK-dictionary에 없는 경우, 그 단어는 1음절씩 끊어져서 제 뜻을 잃은 불완전품사인 어소자(g)로 태깅되는데, 이런 경우 100%가 구문분석에 실패한다. 어소자(g)에 관한 구문분석규칙이 제대로 적용되지 못하기 때문인데, 현재의 구문분석기는 어소자(g)를 인식하지 못하고 있는 것으로 사료된다.

예문1) 是的/l!/w
예문2) 波/n特/d兰/g./

예문1)은 zj->l+w의 규칙이 구문분석규칙 속에 없기 때문에 구문분석에 실패했다. 예문2)는 波特兰이 고유명사로 波特兰/n./w으로 태깅된 후 zj->n+w의 규칙이 적용되어 구문분석이 되어야 한다. 그러나 PK-dictionary에 없는 단어이기 때문에 1음절의 글자를 각각 하나의 형태소로 인식했는데, 兰이 어소자(g)이므로 구문분석에 실패했다.

3.2. 구문분석이 오류인 경우

구문분석이 되지 않는 경우와 마찬가지로 사전에서 비롯된 원인들이 있다. 단어정보가 PK-dictionary에 없음으로 인해서 정확한 품사태깅에 실패했을 때 구문분석에 오류가 발생할 수 있다. 또한, 단어가 다품사인 경우, 문장에 적합하지 못한 품사로 태깅되었을 때도 구문분석은 틀려진다. 구문분석오류를 발생시키는 또 다른 이유는 구문분석규칙의 오류 때문이다. 현재의 구문분석규칙은 구체적인 제약조건이 부족한 상태로, 구문을 형성하는 품사가 같을 때 문장에 적합한 구문트리를 찾지 못한다.

예문3) [SENT[ZJ[NP真/a的/u]?/w]
예문4) [SENT[ZJ[VP包/v号/n]./w]
예문5) [SENT[ZJ[DJ爱好/v集邮/v]./w]

예문3)에서 真的는 하나의 단어인데 PK-dictionary에 하나의 단어로 등록되어 있지 않기 때문에 품사태깅과 구문분석 모두 오류가 발생하였다. 예문4)는 包의 품사가 잘못 태깅되어 생긴 구문분석 오류이다. 包는 g-n-v의 품사를 가지는데, 이 문장에 적합한 품사는 n이어야 한다. 예문5)은 dj->v+v의 규칙과 vp->v+v의 규칙이 문장에 적합하게 적용되지 못하고 있다는 것을 보여준다. 예문5)의 문장은 동사구로 구문분석되어야 하는 것이 맞다.

구문분석오류의 유형과 그 원인을 예문을 들어 분석해 보았다. 품사태깅오류에서 비롯된 구문분석오류는 사전을 수정하거나 품사태깅 제약조건을 다는 등으로 해결되는 문제이므로, 테스트셋 100문장에서 이 원인으로 구문분석이 실패하거나 틀린 10문장을 제외하면, 태스트셋의 33%가 구문분석규칙을 수정함으로써 정확한 구문분석을 유도해내어야 할 부분으로 나타난다.

4. 구문분석규칙의 수정

품사태깅오류의 문제를 제외한 구문분석규칙에서 비롯된 오류를 해결하기 위해서 구문분석규칙을 세가지 형태로 수정했다. 먼저, 문장에 적합하지 못한 구문트리를 추출한다면, 그 문장의 구문트리를 분석하여 적용되어야 할 규칙과 잘못 적용된 규칙, 또는 잘못 적용될 가능성이 있는 규칙 모두에 제약조건을 주거나 제약조건에 수정을 가했다. 필요없다고 여겨지거나 정확한 구문분석에 방해가 된다면, 그 규칙은 삭제했다. 마지막으로, 적합한 구문분석규칙이 없다면 새로운 규칙을 생성시켰다.

4.1. 강하고 구체적인 제약조건

문장 안에서 적합하지 못한 규칙으로 구문분석이 되었다면, 적용되어야 할 규칙과 잘못 적용된 규칙, 또는 잘못 적용될 가능성이 있는 규칙 모두에 제약조건을 주거나 제약조건에 수정을 가해야 한다. 아래의 구문분석된 문장을 보자.

예문5) [SENT[ZJ[DJ爱好/v集邮/v]./w]

예문5)은 구문분석이 잘못되었다. **爱好**는 '좋아하다'라는 의미의 동사이고, **集邮**도 '우표를 수집하다'는 뜻을 가진 동사이다. '우표를 수집하는 것을 좋아하다'라는 의미의 주어가 없는 동사구로 이루어진 하나의 완전한 문장으로 동사구 구문분석이 먼저 이루어져야 하는데, 동사와 동사로 이루어진 단문으로 구문분석되었다. VP->v+v의 규칙이 적용[2]되도록 하기 위해서 DJ->v+v의 규칙과 VP->v+v의 규칙, 그리고 잘못 구문분석될 가능성이 있는 동사구로 이루어진 규칙인 VBAR->v+v도 함께 수정했다.

```
VP—>v+v[964]
CenterNode=0
@NextCate=!v
@0:体谓准:谓
```

「규칙1」

```
DJ—>v+v[172]
CenterNode=1
@Begin, End
```

「규칙2」

먼저 「규칙1」을 보면 간단하게 제약조건이 걸려 있는 것을 볼 수 있다. **爱好**는 동사(**谓语**)를 목적어로 취할 수 있는 동사[3]인데 원래의 규칙에도 이 속성으로 제약조건이 있지만, 「규칙2」 때문에 적용되지 않았다. 그래서 「수정1」과 같이, VP—>v+v를 이룰 수 있는 해당 제약조건을 좀 더 자세하고 정확하게 주었다. 또한 「규칙2」를 「수정1」의 규칙과 겹치지 않고 각 규칙의 제약조건 적용을 방해하지 않도록 「수정2」와 같이 수정했다.

```
VP—>v+v[964]
CenterNode=0
@0:体谓准:谓
@0:助动词:助
@1:有宾:有
@0:体谓准:准, 1:准谓宾:准
@0:单作状语:可
@0:动宾:动
```

「수정1」

```
DJ—>v+v[172]
CenterNode=1
!0:体谓准:谓
@0:单作主语:可, 1:单作谓语:可
@Nodeword[0]=|请|
```

「수정2」

또 다른 동사와 동사로 이루어진 규칙을 보자. VBAR는 VP보다 밀접한 동사끼리의 구성이다. 1자리의 동사는 결과를 나타내는 보어이거나 0자리의 동사와 같은 단어로 두 단어가 결합되어 완벽한 의미를 가진다. 비록 예문5)에는 방해가 되지 않는 규칙이나, VP와 혼동을 일으켜 구문분석 오류를 발생시키는 경우가 많았다.

```
VBAR—>v+v[848]
CenterNode=0
@NodeWord[1]=|来|, NodeWord[0]==NextWord ———①
@0:动趋:趋, 1:趋向动词:趋
@0:动结:结, 1:补语:粘 ———②
@0:动结:结, 1:补语:兼 ———③
```

「규칙3」

「규칙3」의 경우, 제약조건 자체에 이미 오류가 있

---

[2] 적합한 구문분석은 [SENT[ZJ[VP爱好/v集邮/v]./w]의 형태이다.

[3] 이런 동사의 품사적 속성은 体谓准:谓라고 표기한다. (俞士汶등,现代汉语语法信息词典详解, 清华大学出社, 1998)

는 상태로 이 조건이 제대로 적용될 리 없다. ①의 조건은 두가지 조건이 하나의 범위에서 쓰여져 제 구실을 못하고 있다. 또한 ②와 ③의 조건에서 1자리에 주어진 품사속성은 동사의 품사속성이 아니라, 형용사의 품사속성이다. 「수정3」과 같이 다시 정리했다.

```
VBAR—>v+v[848]
CenterNode=0
@0:动趋:趋
@NodeWord[0]==NodeWord[1]
@NodeWord[1]=|掉|
@0:动结:结
```

「수정3」

위와 같이 수정하여 시스템에 적용시키고 다시 테스트해 보았더니, 규칙 1,2,3의 혼선으로 생겼던 구문분석 오류가 테스트셋 100문장 내에서는 모두 없어졌다.

4.2. 규칙의 삭제 및 새로운 규칙 생성

문장에 적합한 규칙을 찾고 수정해가는 과정에서 쓰임이 없으면서 다른 규칙의 적용에 방해가 되는 규칙을 몇 개 발견할 수 있었다. 이런 경우 그 규칙이 적용되지 않도록 하기 위해서 그 규칙이 Error임을 표시했다.

```
DJ—>d+VP[100]
CenterNode=1
@BEGIN,END
```

「규칙4」

```
DJ—>d+VP[100]
CenterNode=1
ERROR
```

「삭제1」

규칙이 없으므로 구문분석이 되지 못한 경우, 새로운 규칙을 만들었다. 예문1)을 보자.

예문1) 是的/l!/w

앞에서 잠깐 언급한 대로 zj->l+w라는 규칙이 없으므로 구문분석에 실패했다. 해당하는 제약조건을 생각해서 「생성1」과 같은 규칙을 만들었다.

```
ZJ—>l+w
CenterNode=0
@BEGIN,END
```

「생성1」

5. 결과 분석

하나의 규칙제약조건을 수정하고 정리할 때마다 시스템에 적용시켜 결과를 확인했다. 먼저 강하고 구체적인 제약조건을 해당하는 규칙에 적용했을 때 원하는 구문분석결과를 확인 할 수 있었다. ERROR로 처리한 규칙의 적용상태도 확인하여 구문분석의 정확도를 높였다. 이런 방법으로 33개 구문분석오류문장 중 규칙이 없어서 구문분석되지 않는 4문장을 제외한 29개 문장의 구문분석을 정확하게 유도했다.

그러나, 새롭게 생성된 5개의 규칙을 첨가시키고 시스템에 적용시키기 위해 규칙을 재정렬하고 테스트해 보았을 때 구문분석의 정확도가 떨어졌다. 새로운 규칙이 적용되도록 하기 위해서 규칙을 재정렬시키면서, 각 규칙과 구문트리에 대한 확률값에 변화가 생겼기 때문이라고 한다. 아래 「표1」은 결과를 정리한 것이다

	기존규칙	규칙수정후	규칙생성후
구문분석정확	57	86	82
품사태깅오류	10	10	10
무규칙오류	4	4	0
구문분석오류	29	0	8
합 계	100	100	100

「표1」

규칙의 상태에 따른 구문분석오류의 형태별 변화이다. 기존의 규칙에서는 구문분석의 정확도가 57%에 불과했지만, 수정된 규칙을 적용시킨 후에는 86%의 정확도를 보였다. 그러나, 수정된 규칙이 적용된 상태에서 생성된 규칙을 적용시켜보았을 때, 정확도는 82%로 떨어졌다. 규칙이 없는 이유로, 구문분석에 실패했던 4문장은 정확하게 구문분석이 되나, 수정된 규칙이 적용되었던 7문장과 기존의 정확하게 구문분석되었던 1문장에서 오류가 발생하였다.

6. 결과에 대한 해결방안과 앞으로의 연구방향

생성된 규칙에 대한 결과에서 오류를 보였던 8문장의 구문분석결과를 정확하게 하기 위해서는 구문분석을 이루는 규칙들 중에서 오류의 원인이 무엇인가를 파악하고 다시 적합한 제약조건으로 수정하고 정리해야 한다. 더 구체적인 제약조건이 필요할 것이다. 규칙의 제약방법에서 오류가 해결되지 않으면 통계적 정보에 의한 확률기반에서 해결방안을 찾아야 할 것이다.

테스트셋 100문장으로는 기존 구문분석규칙의 정확도를 명확히 따질 수 없으며, 4음절 이상의 문장이나, 복문, 그리고 특수구문을 테스트셋으로 한다면, 시스템에서 정확한 구문트리를 찾는 것은 더욱 어려울 것이다. 앞으로 다양한 테스트셋에 대하여 정확한 구문분석을 할 수 있도록 규칙을 수정해 나가는 연구가 필요하다.

기준사전이 정립되지 못함에 따라 구문분석규칙 수정 연구도 진전되지 못하고 있다. 만약 PK-dictionary에 품사나 단어가 추가된다면, 그에 따른 품사속성의 삽입도 뒤따라야 한다. 이는 구문분석규칙이 각 단어의 품사속성을 기반으로 하고 있기 때문이다.

기존 규칙에 수반된 번역변환규칙도 재생성되어야 한다. 현재의 번역변환규칙은 번역의 다양한 형태를 뒷받침해주지 못하고 있기 때문이다. 번역변환규칙 마련은 구문분석규칙수정과 동반되어야 할 것이다.

7. 결 론

중한기계번역시스템(MATES/CK)의 구문분석의 정확도를 높이기 위하여 구문분석규칙을 수정하고 정리하는 방법에 관하여 알아보고 그 연구과정을 살펴보았다. 본 연구는 기존 규칙에 보다 구체적인 제약규칙을 주어 구문분석기가 규칙을 기반으로 하여 정확한 구문분석을 할 수 있도록 유도하는 일이다. 이러한 과정 속에서 중한기계번역시스템(MATES/CK)의 구문분석기가 정확도가 높은 그것으로 계속해서 발전하기를 기대한다.

참 고 문 헌

[1] 장민,황금하,서충원,최기선, 중한 기계번역기 MATES/CK:파이프라인 번역. 제11회 한글 및 한국어 정보처리 학술대회, 1999.10

[2] **俞士汶등, 现代汉语语法信息词典详解, 清华大学出版社, 1998**

[3] 북경대학, **汉语短语标注标记集**, 송희정 역, 1999

부 록

汉语短语标注标记集(중국어구문분석표기집)

1. 현재 사용하고 있는 품사 표기집

(1) 기본 품사 표기집 : 26개

名词	명사	n	**叹词**	감탄사	e
时间词	시간사	t	**助词**	조사	u
处所词	처소사	s	**简称略语**	약어	j
方位词	방위사	f	**习用语**	습관용어	l
量词	양사	q	**前接成分**	접두사	h
区别词	구별사	b	**标点符号**	표점부호	w
形容词	형용사	a	**后接成分**	접미사	k
状态词	의태어	z	**语素词**	어소자	g
动词	동사	v	**非语素词**	비어소자	x
数词	수사	m	**介词**	개사	p
代词	대명사	r	**副词**	부사	d
语气词	어기사	y	**连词**	접속사	c
象声词	의성어	o	**成语**	성어	i

2. 각 구문의 표기

(1) 명사구 np

(2) 준명사구 nbar

(3) 준동사구 vbar

(4) 동사구 vp

(5) 준형용사구 abar

(6) 형용사구 ap

(7) 시간구 tp

(8) 처소사구 sp

(9) 개사구 pp

(10) 수량구 mp

(11) 부사구 dp

(12) 구별사구 bp

(13) 단문 dj

(14) 복문 fj

(15) 정문 zj

(16) 준수사구 mbar

(17) 독립성분표기 dlc

(18) 직접인용표기 yj