한영 자동 번역을 위한 한국어 구문 분석 전처리

김영길[0] 양성일 서영애 김창현 홍문표 최승권
한국전자통신연구원 언어공학연구부 지식처리팀
(kimyk, siyang, yaseo, chkim, hmp63108, choisk)@etri.re.kr

Pre-Processing of Korean Syntactic Analyzer for Korean to English MT

Young-Kil Kim[0] Seong-Il Yang Young-Ae Seo Chang-Hyun Kim Mun-Pyo Hong
Sung-Kwon Choi
Knowledge Processing Team, Dept. of Linguistic Engineering, ETRI

요 약

형태소 해석 결과 생성되는 형태소 열은 구문 분석을 수행하기에는 적절하지 않은 구문 단위로 구성되어 있는 경우가 많으며 이로 인해 구문 분석기가 불필요한 연산을 수행하여 과도한 구문 트리를 생성하는 원인이 된다. 따라서 본 논문에서는 한영 자동 번역의 한국어 구문 분석기 성능 향상 및 자연스러운 대역문 생성을 위하여 시간 부사구와 명사구에 대한 구묶음을 위한 구문 분석 전처리 방법을 제안하며 이를 위한 각 구 단위의 대역 패턴을 정의한다. 방송자막 및 매뉴얼 문장을 대상으로 실험한 결과, 각 문장 구문 단위를 평균적으로 26% 정도 감소시킴으로써 불필요한 파스 트리의 생성을 배제하여 구문 분석기의 성능을 향상시킬 수 있었다.

1. 서 론

현재 한영 자동 번역 시스템은 한국어 구문 분석의 어려움으로 인하여 아직 실용화 단계에 이르지 못하고 있다. 즉 한국어 구문 분석기의 구문 분석 결과는 번역 시스템의 성능에 직접적인 영향을 미치고 있기 때문이다. 형태소 해석 결과 생성되는 형태소 열은 구문 분석을 수행하기에는 적절하지 않은 구문 단위로 구성되어 있는 경우가 많으며 이로 인해 구문 분석기가 불필요한 연산을 수행하여 과도한 구문 트리를 생성하는 원인이 된다. 그리고 이에 대한 해결을 위해 구문 단위 형태소의 구묶음에 관한 연구는 이미 활발히 진행되고 있다.

그러나 이제까지의 구문 단위에 대한 형태소 구묶음에 관한 연구는 보조 용언, 복합 조사 등을 대상으로 일부 논문에서 발표된 적이 있으며[1, 2], 전이망을 이용하여 명사구 묶음을 시도한 적이 있다[3]. 그러나 한국어 문장에서 일반 부사 어휘 이외에 특정 형태소들이 결합되어 시간 부사구 역할을 하는 경우도 빈번하며 이러한 부사구들은 어떤 사건의 시점을 나타내는 등 그 문장에서 중요한 내용을 담고 있는 경우가 많다. 따라서 시간 부사구에 대한 구묶음 처리는 한국어 구조 분석기의 성능 향상은 물론, 그 응용 시스템인 번역 시스템의 번역율 등에 영향을 끼친다.

따라서 본 논문에서는 구조 분석기의 전처리 단계로 구문 분석의 애매성이 발생하지 않는 단계까지의 명사구 묶음을 비롯하여 시간 단위 명사구 또는 시간 부사구를 인식한다. 그리고 이후의 번역을 위한 대역어 생성을 통하여 구문 분석기에서 사용할 하나의 구문 단위를 생성한다. 이를 위해 본 논문에서는 명사구 및 시간 부사구 묶음을 위한 구단위 번역 패턴을 정의한다. 인식된 부사구 및 명사구에 대해 번역 패턴의 대역부를 참조하여 대역어를 결정함으로써 이를 한국어 구조 분석기의 입력으로 사용한다. 이에 대해 방송 자막 및 매뉴얼 문장을 대상으로 실험한다.

2. 구문 분석 전처리기

그림 1은 형태소 분석기와 구문 분석기 사이에 위치하는 구문 분석 전처리기의 구성도를 나타낸다. 한국어 구조분석 전처리기에서는 형태소 정규화 처리, 시간 부사구 처리, 단위성 명사구 묶음 및 일반 명사구 묶음을 수행한다.

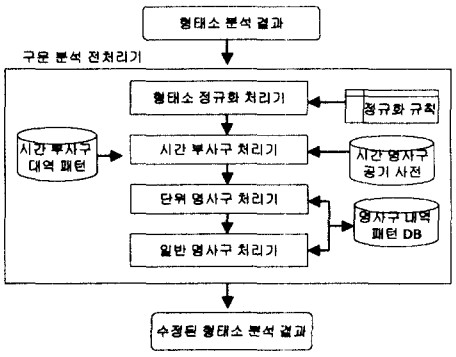

그림 1. 한국어 구문 분석 전처리기

형태소 정규화 처리는 이후 구조 분석에서 사용하는 동사구 패턴에서 사용할 동사에 대한 정규화 과정을 수행한다. 예를 들어" 공부하는 중이다" 와 같은 형태를 " 공부하고 있다" 와 같은 변환을 통하여 동사 " 공부하다" 의 동사구 패턴을 참조할 수 있게 한다. 그리고 시간 부사구 및 명사구 묶음은 자연스러운 번역 문장 생성을 위해서 뿐만 아니라 형태소 해석의 축소, 구문 해석 도중 불필요한 부분 파스 트리의 배제로 인한 속도 및 메모리 효율성 향상, 구문 모호성의 축소를 그 목적으로 한다.

3. 시간 부사구 묶음

일반 시간 부사는 시점, 시간대, 시작점, 동안, 잦기, 차례 등의 의미를 지니는[4] 단일 형태소 단위의 부사를 말하여 시간 부사구란 시간과 관련된 복수개의 형태소 열로 구성되는 부사구로 정의하며 문장이나 절 또는 동사구를 한정하는 수식어구이다.

알 왈리드 왕자가 매입하는 전환사채는 이번 달 말에 발행되고, 만기일은 2003년 3월 30일이다. (1)

문장 (1)은 방송 자막 문장에서 시간 부사구를 포함하는 일례이며 이 문장에서 " 이번 달 말에" 는 하나의 묶음으로서 시간 부사의 기능을 가지고, " 2003년 3월 30일" 은 " 이" 라고 하는 서술격 용언의 보격 성분으로써 명사구 역할을 한다.

3.1 시간 부사구 인식

시간 부사구는 먼저 시간 부사구에 해당되는 부분에 대한 구 묶음이 선행되어야 하며 그 이후 부사 또는 명사인지에 관한 품사 결정이 이루어져야 한다. 일반적으로 시간 부사구가 나타날 수 있는 형태는 상당히 다양하며 자동 번역 시스템에서는 시간 부사구가 인식되면서 이에 대한 목적언어의 대역 정보 또한 결정되어야 한다. 따라서 본 논문에서는 시간 부사구 및 이후의 명사구 인식을 위해서 각 구 단위의 대역 패턴을 정의하여 시간 부사구 및 명사구 대역 패턴 DB를 이용한다.

```
구단위 대역 패턴 ::
        한국어패턴부 > 대역패턴부 [품사정보]

        한국어패턴부 = (한국어어절)+
        한국어어절 = (변수|변수=의미코드!조사|변수!조사|한
국어어휘|한국어어휘!조사)
        변수 = A,B,C, - , N1|N2|N3 -

        대역패턴부 = 대역어절 + (' SPACE|NULL' + 대역어절)*
        어절 = 대역어휘|변수|변수:생성정보
        변수 = N1|N2|N3 - .
        대역생성정보 = th|month|10|100| -  |0.1|0.2
```

그림 2. 구묶음 패턴 형식

구묶음 패턴은 인식부인 한국어 패턴 부분과 생성부인 대역 패턴부로 나눈다. 그림 2는 한영 번역 시스템에서 사용되는 구묶음 패턴의 표현 형식이며 이는 한국어를 원시언어로 하는 모든 번역 시스템에 적용 가능한 형태이며 다음은 부사구 패턴의 일례이다.

```
N1년!부터N2년!까지 > from N1 to N2 [ADVP]
지난N월 > last N:month [ADVP|NP]
오전N시!에 > (at) N 'oclock in the morning [ADVP]
지난겨울!에 > (at) this winter [ADVP]
```

구문 분석기에서 사용될 하나의 구문 단위로의 변환을 위해서 인식된 구묶음 단위에 대한 품사 결정 및 대역어 생성이 이루어져야 한다. 품사정보에서 명사구 가능 구묶음에 대한 품사 결정은 " 오늘", " 내일" 등과 같이 시간 부사 및 일반 명사로 사용되는 어휘들에 대한 품사 결정과 동일하다. 이러한 부사 및 명사 가능 어휘들에 대한 품사 결정을 위해서는 뒤이어 나타나는 어휘들과의 공기 정보를 사용한다.

3.2 시간 부사구 묶음의 예외 사항

관계절의 수식을 받는 경우와 동사의 시간 부사격 이외의 격 성분이 되는 경우는 시간 부사구 패턴에 일치하더라도 이때는 부사구 패턴의 마지막 조사를 제외하고 명사구로 인식한다. 다음은 시간 부사구의 명사구 묶음을 위한 3가지 예외 사항이다.

[Rule 1] 관형절로 수식을 받는 경우
- 한창 더웠던 지난 8월 달에 그의 막내 아들이 태어났다.

[Rule 2] 시간을 나타내는 조사 이외의 격조사가 사용된 경우
- 지난 1950년 6월 25일을 잊지 말아야 한다.
- 올 8월이 끝나야 무더위가 한풀 꺽일텐데.
- 1988년 6월의 서울 올림픽에서 한국선수의 선전이 돋보였다.

[Rule 3] 서술격 조사 " 이" 의 보격 성분인 경우
- 만기일이 2003년 3월 30일이다

4. 명사구 묶음

격조사가 생략된 명사 뒤에 연이어 명사가 나열되는 경우 이 명사구를 구문 분석기 이전에 묶을 경우 생략된 격조사의 복원이 어렵다. 그러나, 일반 텍스트 및 매뉴얼 문장에서 실제 코퍼스 문장을 분석한 결과 이와 같이 동사와의 원거리에서 시간 부사격 등을 제외하고 격조사가 생략되는 경우는 아주 드물게 일어나는 현상이다.

그리고 구문 분석 전 단계에서 이와 같은 생략 현상으로 인하여 명사구 묶음을 하지 않고 각 명사들을 하나의 구문 단위로 취급할 경우 구문 분석기 내에서 각 명사에 대해 격 관계를 추정해야 한다. 그러나 " 카드 형상정보의 매뉴얼 동기 기능을 제공합니다." 와 같은 예문에서와 같이" 카드" 가 주격 조사 " 가" 를 격 정보로 취할 수 있기 때문에 격 복원에 있어서의 오류가 빈번히 발생할 수 있다. 즉 불필요한 격 관계의 추정으로 인해 구문 분석 처리의 효율성 저하 및 오인식 가능성이 크다.

따라서 본 논문에서는 원거리 격 조사 생략, 동사 생략 등의 현상에 의한 묶음 오류를 제외하고 구문 분석의 모호성이 발생하지 않는 범위 내에서 명사구 묶음을 시도한다.

4.1 명사구 대역 패턴

명사구 대역 패턴은 그림 2에서와 같이 구성되며 한국어 어절을 표현하는데 있어 의미코드를 사용할 수 있다. 본 논문에서는 계층 깊이가 4인 200개의 의미코드를 사용하며 이는 명사구 패턴의 다양한 대역 표현을 지원하기 위해서다. 다음은 명사구 패턴의 일례로서 각 명사들의 의미에 따라서 다른 대역어를 가진다. 의미코드로써 표현되지 않는 경우 어휘로 표현하는 것 또한 가능하다.

```
A=사람 B=지명 C=조직 > A B C
A=건축물 B=교제 C=시간 > C for B of A
A=건축물 B=상태 C=사건 > B C at A
```

4.2 명사구 묶음

먼저, 구묶음의 모호성이 발생하지 않는 " 삼만 삼천 달러" ," 12.75 파운드" 등과 같이 수사와 단위 명사가 결합된 형태의 단위 명사구를 먼저 묶는다. 그리고 일반 명사구 묶음의 대상은 원거리 격조사 생략, 동사 생략 등의 현상에 의한 묶음 오류를 제외하고 구문 분석의 모호성이 발생하지 않는 범위 내에서 명사구 묶음을 시도한다. 즉 입력 구문 단위에 대하여 관형격 조사와 공동격 조사를 포함하는 격조사 사이의 명사 열을 그 대상으로 한다.

명사구 묶음과 동시에 대역 패턴부를 참조하여 이에 해당하는 명사 대역어를 생성한다. 명사구 묶음에 의한 기본 명사구 대역 패턴 이외에 관형격 조사, 공동격 조사, 기타 의존 명사 등으로 구성되는 명사구 대역 패턴은 구문 분석에 의해 구조 모호성이 해소된 다음 대역문 생성기에서 적용된다.

4.3 명사구 묶음의 예외 사항

다음과 같은 경우에는 구문 애매성이 발생할 수 있어 처리 대상에서 제외한다.

[Rule 1] 이전 구문 단위가 관형형 어미를 포함하고 명사구 성분에 " 콤마" ," 및" , " 또는" ," 그리고" 가 포함되는 경우.
- 어머니가 즐겨보는 TV 프로그램, 아버지가 좋아하는 프로그램은 정말 차이가 많다.
- 미리 준비한 결과 그 일을 쉽게 마칠 수가 있었다.

[Rule 2] 이전 구문 단위가 관형형 조사를 포함하고 명사구 성분에 " 콤마" , " 및" , " 또는" , " 그리고" 가 포함되는 경우.
- 일본의 보수주의 경향, 미국의 군국주의 경향에 반감을 가지고 있다.

5. 실험 및 평가

실험을 위한 코퍼스를 방송 자막 문장 및 정보통신 분야의 매뉴얼 문장을 대상으로 하였다. 방송 자막 문장은 평균 어절수가 11.87, 매뉴얼 문장은 14로 매뉴얼 문장이 보다 긴 문장으로 구성되어 있음을 알 수 있다. 그리고 방송 자막 문장에는 시간 부사구가 상당수 나타나는 반면 매뉴얼 문장에서는 전혀 나타나지 않고 긴 단위의 명사구가 많이 나타나는 특징이 있다.

본 실험에 사용한 구 단위의 대역 패턴은 방송자막 문장 및 정보통신 분야의 매뉴얼 문장의 대역 말뭉치에서 추출하였으며 테스트 문장은 대역 패턴 구축에 포함하지 않은 방송 자막 250 문장과 정보통신 매뉴얼 250 문장에 대해 실험을 실시하였다. 표 1에서와 같이 명사구 또는 부사구 묶음의 정확도는 97.2%로 상당히 정확했으며 이러한 구묶음으로 인해 평균 26.4%의 구문 단위수의 감소가 이루어졌다.

표 1. 구 묶음 정확율 및 구문 단위 감소율

	방송 자막	매뉴얼	전체
문장수	250	250	500
구문단위수	2870	3500	6370
전체 후보수(명사구 + 시간부사구)	350 (285+65)	630 (630 + 0)	980
명사구묶음 성공	278	621	899
시간부사구묶음 성공	54	0	54
구묶음 정확율	332/350	621/630	953/980 (97.2%)
수정된 구문단위수	2290 (79.8 %)	2400 (68.6 %)	4690 (73.6 %)

6. 결론

본 논문에서는 한영 자동 번역의 한국어 구문 분석기 성능 향상을 위하여 명사구와 시간 부사구에 대한 구문 단위의 묶음을 제안하였다. 기존의 구묶음 방식과는 달리 본 논문에서는 이후의 대역어 생성을 고려하여 구 단위의 대역 패턴에 의한 명사구 및 시간 부사구의 구묶음 방법을 제안하였다.

이러한 구문 분석 전처리는 각 문장 구문 단위를 평균적으로 26% 정도 감소시킴으로써 불필요한 파스 트리의 생성을 배제하여 구문 분석의 효율성을 높일 수 있었다. 향후 시간 부사구 및 명사구 패턴의 지속적인 보완에 의하여 정확한 구묶음 처리 뿐만 아니라 보다 자연스러운 대역어 생성이 이루어질 수 있어야 한다.

참고 문헌

[1] 황이규,이현영,이용석, " 형태소 및 구문 모호성을 위한 구문단위 형태소의 이용" , 정보과학회논문지, 2000

[2] 강호관, 이종혁, 이근배 , " 새로운 어절 해석에 기반한 한국어 의존관계 파서" , 한글 및 한국어 정보처리 학술대회, 1997

[3] 김미영,강신재,이종혁, " 규칙과 어휘정보를 이용한 한국어 문장의 구묶음" , 한글 및 한국어 정보처리 학술대회, 2000

[4] 서정수, 현대 국어 문법론, 한양대학교출판원, 1996