한영 자동 번역을 위한 동사구 번역패턴의 활용

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

Design of Verb-Phrase Patterns for Korean-to-English MT

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

요 약

원시언어 문장의 구조 분석을 기반으로 하는 기계번역 시스템에서 원시언어의 최소 의미 단위는 동사를 중심으로 한 단문으로 생각할 수 있다. 단문 단위 대역어를 지정하기 위해서는 동사구 번역패턴의 사용이 요구된다. 본 논문에서는 한국어 단문 내 격 정보와 번역을 위한 의미 제약조건을 기술하여 한영 기계번역 시스템에서 사용하는 동사구 번역패턴을 정의하고, 문장 정규화를 통한 동사구 번역패턴의 활용 방법을 제안한다. 동사구 번역패턴은 단문 구조 파악을 위한 제약 조건부와 대역어 선정부로 나뉜다. 제약 조건부는 단문 구조 번역을 위한 최소한의 의미 제약만으로 기술되며, 격조사로 구분되는 격 정보를 갖는다. 이러한 격 정보는 원시언어인 한국어의 단문 분석을 위해 사용되며 분석결과에 대해 단문 단위 대역어를 지정한다. 동사구 번역 패턴은 실제 말뭉치에서의 사용을 반영하기 위해 병렬 말뭉치로부터 구축되며 실험을 통해 예측되는 패턴의 규모를 알아볼 수 있다.

1. 서 론

한영 기계번역 시스템에서 갖는 어려움은 원시언어인 한국어와 대상언어인 영어의 구조적 차이에서 기인한다. 구조분석을 기반으로 하는 번역 시스템에서 이러한 문제점을 해결하기 위해 원시언어의 최소 의미 단위를 동사를 중심으로 한 단문으로 생각하고, 구조분석 이후 단문 단위의 대역어를 붙이는 방법이 있다. 그러나 원시언어인 한국어가 자유로운 어순과 조사생략, 보조사의 쓰임으로 격 정보 결정에 어려움이 있고, 이로 인한 단문 구조 분석 애매성은 부자연스러운 대역어로 연결된다. 이러한 단문 구조 애매성을 최소화하기 위해서 부가적인 격 결정 정보의 구축이 필요해 지고, 실용적인 시스템 구현을 위한 지식 구축의 어려움이 발생한다.

문장의 구문관계를 결정하는 방법은 의미적 제약과 통계적인 추출 방법으로 크게 나누어 볼 수 있다. 의미적 제약은 하위범주화 정보와 함께 의미제약을 사용한 방법[1]으로 지식 구축의 어려움이 대두된다. 통계적인 추출 방법은 대량의 말뭉치로부터 명사와 동사간의 공기 정보와 하위범주화 정보를 사용할수 있도록 제안[3]되었다. [4]에서는 이러한 의미적 제약과 통계적 방법을 병행하여 2단계 결정 방법을 취한다.

본 논문에서는 한영 번역을 위해 단문 구조를 파악하여 단문 단위 대역어를 선정할 수 있는 의미 기술 동사구 번역 패턴을 정의하고, 패턴 구축을 위한 작업을 최소화하기 위해 문장 정규화를 통한 패턴의 재 사용 방법을 제안한다.

동사구 번역 패턴은 실제 말뭉치의 사용을 고려하기 위해 병렬 말뭉치를 사용한 실례를 바탕으로 구축되며, (동사, 조사, 명사)의 의미쌍을 이용한 개념패턴[4]과 달리 단문을 구성하는 모든 격 정보를 함께 기술한다. 명사 의미제약은 지식구축의 어려움을 최소화하기 위해 계층구조를 배제한 200개 미만의 의미코드를 사용하며 실험을 통해 지식구축의 규모를 예상할 수 있다.

2. 단문 내 제약 조건

단문이 갖는 의미는 중심어인 동사와 동사에 연결되어 격 정보를 채우는 명사가 갖는 의미로 파악할 수 있다. 한국어는 격조사와 함께 쓰이는 명사의 구문적 역할에 의해 격 정보를 결정할 수 있으며, 명사가 갖는 의미는 중심어인 동사에 해당 격 정보를 채우며 연결이 가능한지 여부를 결정하는 제약조건이 된다. 이러한 조건을 기술한 번역패턴을 사용할 때 원시언어인 한국어의 단문구조를 파악하고 이에 따르는 단문단위의 대역어 지정을 할 수 있다.

단문 내 의미구조를 표현하기 위한 제약 조건은, 동사가 취할 수 있는 격조사, 격 정보를 채우는 명사의 의미와 동사가 취하는 형태에 따른 의미구분으로 크게 구분하여 기술한다.

첫번째, 격조사는 격 정보를 나타낼 수 있는 대표조사를 선정하여 사용하고, 주격조사의 대표형은 '가'를 사용하며 목적격 조사는 '를'을 사용한다. 그외 조사는

부사격 조사로 기술하도록 한다.

두번째, 명사의 의미는 세밀한 의미정보보다 한영 번역을 위해 단문의미를 전달할 수 있는 최소한의 정보만 나타낼 수 있도록 계층화를 배제한 200개 미만의 의미코드를 사용한다. 명사의 의미는 형태소 사전에 어휘와 함께 기술되어 형태소 분석 단계에서 얻어지며, 중의적 의미는 입력문에서 같이 쓰인 동사에 의해 구분되어 결정된다.

마지막 제약조건으로 동사가 취할 수 있는 의미구분과 함께 사동/피동형의 변화형과 보조용언에 의한 의미변화를 표현할 수 있어야 한다.

```
타1-1!다 :C
```

그림 1. 동사 제약 조건의 예

그림 1은 "타다" 동사의 제약조건을 표기한 것이다. 먼저 "타다" 동사의 의미 중의성을 구분하기 위해 어간 "타"에 의미코드 "1-1"이 붙었으며 기호 "!"에 의해 동사를 나타내는 종결어미 "다"와 구분된다. 마지막으로 보이는 ":C" 표기는 이 동사의 제약조건이 사동형임을 표기한 것이다.

3. 동사구 번역패턴

단문의 의미적 제약조건들은 한영 번역에 적용하여 동사구 번역 패턴을 정의할 수 있다. 동사구 번역 패턴은 제약조건을 명사-동사 쌍에 국한하지 않고 단문 단위의 제약조건을 함께 기술하도록 한다. 본 논문에서는 단문 단위의 모든 제약 조건 집합을 단문 제약 조건으로 간주하고, 단문 제약 조건과 대역어를 함께 기술한 형식을 동사구 번역 패턴으로 정의한다. 한영 번역을 위해 단문 제약 조건은 원시언어인 한국어를 대상으로 지정되며, 대역어는 영어를 기술한다.

A=Sn_1!P_1 B=Sn_2!P_2 ... Sv_1!다:{C|P|V}
-> TL(A, B,...)

A, B: 의미코드 지정 변수
Sn_i : 명사의 의미코드
P_j : 대표 격조사
Sv_k : 동사의 의미코드
C : 사동형 제약조건
P : 피동형 제약조건
V : 형용사의 동사형 제약조건
TL(A, B, ...) : 의미변수 A, B를 취하여 선정되는 대역어

그림 2. 한영 동사구 번역 패턴의 형식

한영 동사구 번역 패턴은 그림 2와 같은 형식으로 지정된다. 동사구 번역 패턴은 크게 '->' 기호를 기준으로 앞쪽은 단문 제약 조건부를 나타내고, 뒤쪽은 대역어 정보를 갖는다.

명사 의미코드, 격조사, 동사 제약 조건으로 이루어진 단문 제약 조건부에서 명사 의미코드 부분은 격 정보를 채우는 명사의 의미와 대역어 선정부에서 사용하기 위해 의미코드를 지정하는 변수로 기술되며, 격을 나타내는 대표조사를 사용하여 주격/목적격/부사격을 표기한다.

동사 제약 조건 부분은 사동/피동형 제약과 함께 "어_하"와 같이 형용사를 동사형으로 변환하여 용언이 취하는 격 정보가 바뀌는 동사화 보조용언을 위해 제약조건 "V"를 표기한다.

```
A=사람!가 B=사람!를 부럽1-1!다:V -> A envy:v B
```

그림 3. 한영 동사구 번역 패턴의 예

그림 3에서 지정되는 동사제약 "V"는 "부러워_하다"와 같이 동사화 된 형용사에 쓰일 수 있는 패턴임을 지정하며, 대역어 선정부의 ":v" 표기에 의해 대상 동사 envy로 번역된다.

4. 동사구 패턴 적용

입력문의 분석은 동사구 패턴을 적용하여 격 정보 분석과 동사 제약 조건의 적용을 함께 한다. 입력문의 구조를 파악하기 위하여 의존문법 파서를 사용하며, 동사구 패턴에 의해 단문 구조를 파악한다.

동사구 패턴의 적용은 동사구 패턴에서 기술하는 제약조건을 입력문과 비교하여 전체의 조건이 입력문과 일치하는 완전매칭과 일부 조건이 일치하는 부분매칭에 따라 다른 가중치를 할당하여 단문 분석을 시도한다.

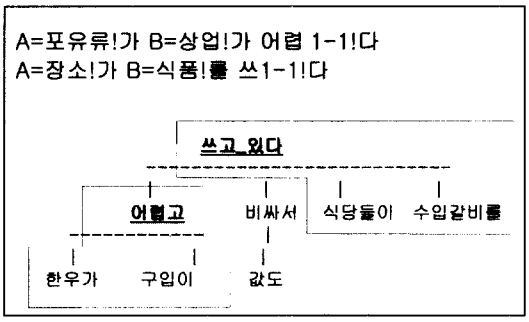

그림 4. 단문 제약 조건부의 적용

그림 4는 "한우가 구입이 어렵고 값도 비싸서 식당들이 수입갈비를 쓰고 있다"의 문장을 입력으로 받아, "어렵다"와 "쓰다" 용언에 각각 제약조건을 적용하는 모습을 구문 구조에서 보여준다.

관형절의 경우 피수식어가 수식절의 격 정보를 채우게 되므로 피수식어를 포함한 수식절을 별도의 단문으로 취급하여 동사구 패턴을 적용해 보아야 한다.

5. 문장 정규화

단문의 의미 구조를 나타내기 위한 동사구 번역 패턴의 구축은 수동으로 기술되어야 하는 어려움이 있다. 한국어의 경우 동사는 선어말 어미와 보조용언, 접미사 등을 통해 다양한 의미변화가 생길 수 있다. 이러한 동사의 변화형에 따라 각각을 모두 독립된 패턴으로 구축하는 것을 방지하기 위하여 문장 정규화 과정을 거쳐 동사 원형만을 사용한 동사구 패턴을 구축할 수 있도록 한다.

격 정보의 이동이 일정한 사동/피동형 변환은 정규화 규칙을 사용하여 변환될 수 있다. 동사의 변화형은 "먹이다"와 같이 접미사 "이,히,리,기,우,구,추"가 붙어 사동/피동형을 나타내는 경우, "먹게 하다"와 같이 보조용언 "게_하", "게_되"가 붙는 경우, 보조동사 "시키다", "되다", "당하다" 등을 사용하여 변환되는 경우의 세가지로 나누어 정규화 규칙을 적용한다.

표 1. 사동/피동형 정규화 규칙

피동문 → 주동문	사동문 → 주동문
① 타동사 대상	① 자/타/형용사 대상
② 주어 → 목적어	② 주어 → 부사어
③ 부사어 → 주어	③ 목적어,부사어(에게)
④ 주어, 부사어	→ 목적어,주어
(에서) → 부사어	④ 부사어(에게) → 주어
(로), 목적어(를)	⑤ 목적어 → 주어

표 1은 정규화 규칙을 표기한 것이다. 피동문의 변환 규칙 중 부사어의 조사는 "에 의해", "에게", "한테", "로", "에"를 대상으로 하고, 사동문의 변환 규칙 중 주동문에서 만들어지는 부사어의 조사는 "에 의해"를 붙여 생성하거나 생략되도록 한다.

6. 보조사 및 조사 생략 처리

격 조사에 의해 구분되는 격 정보는 입력문에서 보조사의 쓰임이나 빈번한 조사의 생략으로 명확하지 못한 경우가 많다. 본 논문에서는 미지격 처리의 대상을 주격과 목적격만으로 제한하고 보조사와 조사 생략을 동일하게 분석한다. 관형절의 격 정보 역시 피수식어의 수식절 내 격 관계를 조사생략의 형태로 처리하여 분석한다.

조사가 생략된 형태의 입력문에는 임의적으로 주격조사 '가'와 목적격 조사 '를'을 복원하여 동사구 패턴과 비교하여 본다. 동사구 패턴에서 취하는 명사 의미를 만족하고 해당 격조사가 취하는 격이 이미 채워져 있지 않다면 복원된 격조사에 의해 격 정보를 결정한다.

7. 실험 및 평가

실험에는 52000 개의 해당 분야 전공자들에 의해 구축된 한영 동사구 번역 패턴을 사용하여 패턴의 매칭 성공률과 번역률을 계산하였다. 입력문장은 KBS 뉴스 문장을 사용하였으며, 12어절 이상 21어절 이하의 문장을 대상으로 200 문장을 임의 선택하여 실험하였다. 번역문의 평가는 표 2와 같은 기준으로 이루어 졌다.

표 2. 번역문 평가 기준표

점수	기준
4 (Perfect)	- 문장의 의미가 명확 - 개별 단어의 번역도 정확
3 (Good)	- 문장의 의미는 대체로 명확 - 개별 단어의 번역 오류가 일부 존재 (문장의 단어수 20% 이내)
2 (OK)	- 문장의 의미는 몇 번 읽어야 파악됨 - 개별 단어의 번역 오류가 일부 존재 (문장의 단어수 30% 이내)
1 (Poor)	문장의 의미는 추측을 통해 이해됨
0 (Fail)	여러 번 읽어도 텍스트의 의미를 알수 없거나 번역 실패

동사구 번역 패턴의 완전 매칭 횟수가 높을수록 정확한 대역어의 선정이 가능하나 구축된 52000 패턴에 대해 임의 선택된 입력문장의 매칭 비율은 12%로 낮았다.

표 3. 실험 결과

분류	횟수	비율
입력문장수	200	-
전체 동사구 패턴 시도	631	100%
완전 매칭 성공	76	12%
부분 매칭 성공	410	65%
평균 번역 평가 점수	3.04	

8. 결 론

한영 번역을 위한 원시언어인 한국어의 단문은 최소 의미단위로 해석되어 대역어를 선정할 수 있다. 이러한 한영 번역을 위한 단문 분석은 한영 동사구 번역 패턴을 사용한다. 단문 제약 조건을 갖는 동사구 번역 패턴은 병렬 말뭉치를 이용하여 수동으로 구축되며, 지식 구축의 효율성을 위해 번역을 위한 최소한의 의미 코드를 지정하여 패턴을 구축하고, 문장 정규화를 통해 번역 패턴을 재사용한다.

참고 문헌

[1] 나동렬,"한국어 파싱에 대한 고찰",정보과학회지,Vol.12,No.8,pp.33-46,1994.
[2] 서영애 외3명,"용언구에 기반한 한영 기계번역 시스템: CaptionEye/KE",한국정보처리학회 추계학술대회 논문집,Vol.7,No.2,pp.269-272,2000
[3] 양재형,김영택,"통계 정보를 활용한 한국어 미지격 명사구의 구문관계 결정",정보과학회 논문지,Vol.21,No.5,pp.808-815,1994.
[4] 이휘봉, 강인수, 이종혁, "개념 패턴과 통계 정보를 이용한 한국어 미지격의 구문관계 결정 방법", 한글 및 한국어 정보처리 학술대회,pp.261-266,1998.