

‘X하다’와 ‘X를 하다’의 상관성에 기반한 한중 자동 번역 시스템에서의 ‘하다’ 동사 번역에 관한 연구

서영애*, 황은하, 홍문표, 최승권
한국전자통신연구원 언어처리연구팀
e-mail:{yaseo, yinxia, hmp63108, choisk}@etri.re.kr

A Study on Translation of ‘Hata’ Verb Based on Relation between ‘X-hata’ and ‘X-lul hata’ in Korean-Chinese Machine Translation

Young-Ae Seo*, Yinxia Huang, MuXyo Hong,
Sung-Kwon Choi
*Natural Language Processing Team, ETRI

요 약

패턴에 기반한 자동번역 시스템에 있어서, 번역율을 결정하는 가장 중요한 요소는 패턴의 양과 질이다. 필요한 대량의 패턴을 단기에 구축하기 위해서 패턴 정보 기술상의 잉여성을 줄일 필요가 있다. ‘하다’ 동사는 한국어에서 가장 빈번하게 사용되는 동사 중의 하나로서, ‘X를 하다’와 ‘X하다’의 2가지 구문 구조를 갖는다. 본 논문에서는 동사구 패턴에 기반한 한중 자동 번역 시스템에서 ‘X를 하다’와 ‘X하다’ 구문간의 상관성을 이용하여 두 구문이 ‘X하다’ 동사에 대한 동사구 패턴을 공유하도록 함으로써 동사구 패턴 구축 시에 정보의 중복 기술을 피하고, 패턴의 적용 범위를 넓힌다. 패턴의 공유로 인해 발생하는 대역문 생성 시의 문제점은, ‘X를 하다’ 구문에 대응하는 중국어 대역 형태 별로 분류하여 해결함으로써 패턴을 공유하면서도 번역의 질은 유지시킬 수 있는 방법을 제안한다.

1. 서론

현재 한국전자 통신 연구원에서 개발 중인 한중 자동 번역 시스템 Tramie-KC는 데이터에 기반한 자동 번역 시스템으로서, 동사구 패턴을 기반으로 하여 한중 자동 번역을 수행한다.[1] 한중 동사구 패턴의 구축은 고비용 고난이도의 작업이므로, 하나의 동사구 패턴이 가능한 많은 한국어 용언의 쓰임을 처리할 수 있도록 패턴의 적용 범위(coverage)를 높일 필요가 있다.

본 논문에서는 Tramie-KC에서 한국어에서 많이 사용되는 ‘하다’ 동사에 대한 동사구 패턴의 적용 범위를 높이는 방법에 대하여 제안한다. 이를 위해 ‘X하다’와 ‘X를 하다’ 구문간의 상관성을 이용하여 두 구문이 ‘X하다’ 동사에 대한 동사구 패턴을 공유한다. 패턴의 공유로 인해 발생하는 대역문 생성시

의 문제점은 ‘X를 하다’ 구문에 대응하는 중국어 대역 형태별로 분류하여 동사구 패턴이나 형태소 사전에 필요한 정보를 기술하여 처리함으로써, 정보 기술 시의 잉여성(redundancy)을 줄이면서 번역의 질은 유지시킬 수 있는 방법을 제안한다.

2. 관련 연구

[5]에서는 언어학적 접근에 있어서 ‘X하다’와 ‘X를 하다’, ‘X를 Y를 하다’와 ‘X를 Y하다’와 같은 구성이 유지적인 상관관계가 있음을 보이고 이들이 어떻게 도출되는지에 대해 기술하였다. 이러한 연관성에 기반하여 [2]에서는 중심어 개념을 도입한 패턴 기반 번역 방식에 있어서 서술성명사를 목적으로 취하는 ‘하다’ 동사 구문에 대해 서술성명사를 중심어로 패턴 분석하는 방법을 제안하였다. 기능동사 구문은

한국어에만 국한된 현상이 아니라 범언어적인 특성으로, [3]에서는 붙어의 서술성명사와 기능동사로 구성된 문장에서도 명사가 술어적 핵요소로 작용함을 보였다. 본 논문에서는 서술성명사 뿐 아니라 비서술성명사인 경우를 포함한 '하다' 구문에서의 동사구 패턴을 공유 방법에 대해 논하고 이로 인해 발생하는 변환 및 생성시의 문제점을 해소하고자 한다.

3. 'X하다' 동사의 한중 동사구 패턴

본 장에서는 한중 자동번역 시스템, Tramie-KC에서 사용하는 동사구 패턴과 '하다' 구문의 특성에 대해 간략히 기술하고 'X하다'와 'X를 하다' 구문이 동사구 패턴을 공유할 수 있음을 보인다.

3.1 한중 동사구 패턴

Tramie-KC의 한중 동사구 패턴은 한국어 용언의 쓰임을 기술한 하위 범주화 틀의 확장된 형태로서, 한국어 구조 분석 정보와 중국어 대역 정보가 함께 기술되어 있다. [예1]은 '가다' 용언에 대한 동사구 패턴 정보이다. A, B 등은 용언의 격슬롯을 의미하며, '사람', '건축물' 등은 격슬롯에 대한 의미제약 정보이다. 한국어부의 제약조건이 모두 만족될 경우, 오른쪽의 중국어 동사구 패턴 정보를 이용하여 중국어로 변환, 생성을 수행한다.[1]

가다1 : A=사람!가 B=건축물!에 가!다
> A 去:v B [그는 법원에 갔다]

[예1] '가다'의 한중 동사구 패턴 예

이러한 한중 동사구 패턴은 매우 고비용, 고난이도의 작업이므로, 의미적으로 같은 동사구일 경우 동사구 패턴 정보를 공유함으로써 패턴 구축 비용을 줄일 필요가 있다.

3.2 'X하다'와 'X를 하다' 구문의 특징

(1-1) 그는 마침내 결정을 했다.

(1-2) 그는 마침내 결정했다.

Tramie-KC가 (1-1)과 (1-2)의 문장을 번역하기 위해서는 '하다' 동사구 패턴 [예2]와 '결정하다' 동사구 패턴 [예3]을 필요로 한다. 그러나, 이 두 문장은 (1-1)의 '결정을 하다'가 '결정하다'로 바뀌었을

뿐 형식상으로는 의미적으로 동일하다.[5] 따라서, 하나의 동사구 패턴을 공유하도록 하여 패턴의 중복 기술을 피하도록 하는 것이 바람직하다.

결정하다3 : A=사람!가 B=방법!를 결정하!다
> A 決定:v B [그가 방침을 결정하다]

[예2] '결정하다'의 한중 동사구 패턴 예

하다26 : A=사람!가 B=방법!를 결정!를 하!다
> A 決定:v B [그가 방침을 결정하다]

[예3] '하다'의 한중 동사구 패턴 예

이를 위해 본 논문에서는 'X하다'와 'X를 하다' 구문이 하나의 'X하다' 동사구 패턴을 공유하도록 하였다. 즉, 'X를 하다' 구문이 입력된 경우, 이를 'X하다' 구문으로 간주하고 'X하다' 동사구 패턴을 가져와 구문 분석을 수행한 후, 최종 분석된 결과는 입력문의 'X를 하다' 형태로 복원한다. 이러한 동사구 패턴의 공유는 '나무하다'와 같이 'X'가 서술성명사가 아니고, '하다'가 기능동사가 아닌 경우에도 동일하게 적용된다.

4. 동사구 패턴 공유 시의 문제점 및 해결 방안

동사구 패턴 공유로 인해 대역문 생성 시에 발생하는 문제점과 이에 대한 해결책에 대해 논한다.

4.1 동사구 패턴 공유 시의 문제점

(1-1)과 (1-2)가 동일한 동사구 패턴을 이용하여 번역 가능함을 보았다. 그러나, (2-1)과 같이 '하다'와 결합된 명사 'X'가 관형어나 관형절의 수식을 받는 경우 대역문의 생성에 문제가 발생한다.

(2-1) 그는 마침내 어려운 결정을 했다.

(2-2) 他終於作了不容易的決定。

(2-1)은 '결정하다'와 '어렵다' 두 개의 동사구 패턴을 이용하여 처리된다. (2-2)의 대역문으로 올바르게 번역되기 위해서는 '결정'에 해당하는 중국어 대역어, '決定'이 대역문에 표현되어 두 중국어 용언구의 결합이 가능하여야 한다. 그러나, [예2]의 동사구 패턴에는 '決定'이 번역되지 않으므로, '어려운'에 해당하는 대역어를 전체 번역문에 반영할 수 없다.

4.2 'X하다'와 'X를 하다' 대역어의 상관성에 대한 분류

관형어/절 수식 문제의 근본적 해결책은 'X'의 대역어가 나타나는 자연스런 대역문을 생성하는 것이다. 이를 위해서는 'X하다'의 중국어 대역표현이 "중국어 기능동사 + 'X'의 중국어 대역어"의 형태로 번역되거나, 중국어의 대역어가 동빈(動賓) 구조로 이루어질 경우 가능하다. 동빈 구조란 '講課'와 같은 중국어 동사의 내부 구조가 동사와 목적어의 결합에 해당하는 경우를 말한다.

본 논문의 취지에 맞추어 한국어의 'X를 하다' 용언구에 대응하는 중국어의 번역을 [표1]과 같이 분류하였다. 중국어 번역 결과에 따라, 크게 중국어 기능동사와 결합하는 유형과 결합하지 않는 유형으로 분류하고, 기능동사와 결합 불가능한 동사들은 그 내부구조에 근거하여 다시 2가지 유형으로 하위분류하여 유형화하여 총 3가지로 분류하였다.

유형	중국어 대역	한국어 'X를 하다'의 예
1	A 加以, 給以, 予以, 給予와 결합 가능	우대를 하다
	B 作, 進行와 결합 가능	연설을 하다
	C A, B 기능동사와 모두 결합 가능	비평을 하다
2	동빈구조	강의를 하다
3	기타	공부를 하다

[표1] 'X를 하다'에 대한 중국어 번역 유형

유형1의 처리를 위해서는 'X'의 중국어 대역어휘 정보에 대하여 대응하는 기능동사 정보를 별도로 형태소 사전에 기술한다. 중국어의 기능동사는 그 의미적 특징에 근거하여 A유형-'加以, 給以, 予以, 給予'과 B유형-'作, 進行'으로 나뉜다[6]. 중국어 명사와 결합 가능한 기능동사의 선택은 각 어휘의 개별적 특성에 의해 결정된다. 형태소 사전은 한국어 형태소어휘, 형태소 코드, 의미코드(SEM), 중국어 대역어(CROOT) 정보, 중국어 기능동사 정보(CSVERB) 등을 담고 있다. [예4]는 '결정'에 대한 형태소 사전 정보의 예이다.

유형2는 동사구 패턴에 중국어 동사의 내부 구조정

보를 기술하여 'X'의 중국어 대역어에 대한 링크정보를 기록한다. [예5]의 '식사하다'에 대한 동사구 패턴에서 중국어부의 ':v'는 동빈 구조에서의 중국어 동사(吃), ':n'은 목적어(飯)임을 표시한다.

120012 479
 { [(SEM 선택) (EROOT decision) (CROOT 決定) (CSVERB 作, 加以)] }

[예4] '결정'의 형태소 사전 정보의 예

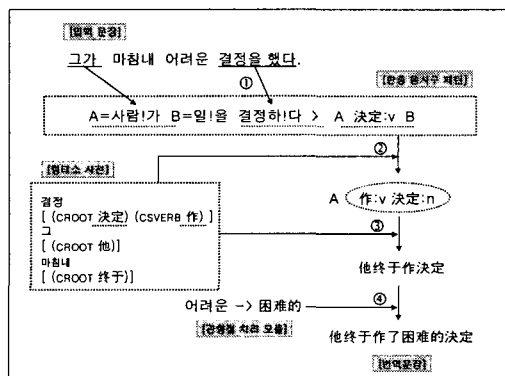
식사하다1 : A=사람!가 식사하다
 > A 吃:v飯:n [아버지께서 식사하십시오.]

[예5] '식사하다'의 한중 동사구 패턴의 예

유형3은 유형1, 2에 속하지 않는 그 외의 경우이다.

4.3 번역 유형별 관형어/절의 처리

유형1에 속하는 'X를 하다'의 'X'가 관형어/절의 수식을 받는 경우, 'X하다' 동사구 패턴을 가져온 후, 중국어 번역부의 동사 부분을 'X'의 형태소 사전에서 가져온 CSVERB와 CROOT 정보를 이용하여 중국어 '형식동사+서술성명사'의 형태로 치환한다. 그리고 나서 기존의 Tramié-KC의 관형어/절 처리 루틴을 호출하여 관형어/절이 'X'의 중국어 대역어를 수식하도록 하여 번역을 완료한다. [그림1]은 (2-1)의 문장이 [예2]의 동사구 패턴과 [예4]의 형태소 사전 정보를 이용하여 처리되는 과정을 보여준다.



[그림1] 유형1의 관형절 번역의 예

유형2의 경우는, 유형1에서의 치환과정을 거치지

않고, 'X하다' 동사구 패턴의 중국어 대역부에서 'n'의 링크정보로 기술된 'X'의 중국어 대역어를 수식하도록 관형어/절 처리 루틴을 호출한다.

유형3의 경우를 살펴보면 한국어에서의 관형어 혹은 관형절이 중국어 문장에서는 술부를 수식하는 보어로 나타나는 것이 일반적이다. (3-1)에 대한 대역문 (3-2)가 바로 그 예이며, 번역된 중국어 문장 (3-2)를 한국어로 번역하면 (3-2')와 같다.

(3-1) 그는 늦은 출발을 했지만 실망하지 않았다.

(3-2) 他出發得很晚, 但是沒有灰心。

(3-2') 그는 늦게 출발했지만, 실망하지 않았다.

4.4 실험 결과

경제 뉴스 코퍼스로부터 169개의 'X를 하다' 구문이 포함된 문장을 무작위로 추출하여 본 논문에서 제안한 방법으로 번역하였다. 실험 결과, 89.3%를 차지하는 112개의 문장은 제안된 방법으로 번역하였을 경우 자연스러운 번역 출력문을 얻을 수 있었다.

그러나, 유형 3의 경우 중 제안된 방법으로 번역할 경우 원문과는 의미적 차이가 있는 출력문을 얻음으로써 좋은 결과라 할 수 없는 경우가 18개, 즉 10.7% 발견되었다. 이 경우를 기타로 분류하여 실험 결과를 표로 나타내면 아래와 같다.

유형		출현 빈도(%)
1	A	19 (11.2%)
1	B	39 (23.1%)
1	C	52 (30.8%)
2		17 (10.1%)
3		24 (14.2%)
기타		18 (10.7%)
계		169 (100%)

[표2] 실험 결과

기타의 경우는 유형 3의 처리방법으로 번역할 의 미적인 차이를 낳았다.

(4-1) 그가 어려운 공부를 한다.

(4-2) 他學習得很難。

(4-2') 그는 어렵게 공부를 한다.

(4-3) 他學的很難。

(4-3') 그가 배우는 것은 어렵다.

번역된 중국어 (4-2)를 한국어로 그대로 직역하면 (4-2')와 같다. 언어학적인 견지에서 볼 때 (4-1)에 대한 가장 중국어다운 표현은 (4-3)이다. (4-3)을 다시 한국어로 번역하면 원문인 (4-1)과는 구조가 다른 (4-3')의 구조를 갖게 된다. 기타 경우의 해결을 위해서는 동사구 패턴과 형태소 사전 정보 이외의 보다 심화된 중국어 의미론적, 통사론적 그리고 화용론적 연구와 그에 따른 연구가 추가적으로 이루어져야 할 것이다.

5. 결론

본 논문은 동사구 패턴에 기반한 한중 자동번역 시스템 Trامية-KC에서 'X를 하다'와 'X하다' 구문의 상관성을 이용하여 'X하다' 동사구 패턴을 공유하여 처리함으로써 정보 기술의 잉여성을 줄이면서도 번역의 질은 유지할 수 있는 방안에 대하여 논하였다. 패턴의 공유 시 관형어나 관형절의 수식 문제가 발생함을 보이고, 이를 해결하기 위해 'X를 하다'에 대응하는 중국어 번역 유형을 3가지 형태로 분류하여 각 유형에 대한 번역 방법을 제안하였다. 제안된 방법은 다른 기능동사 구문 등에도 확대 적용 가능하다. 그러나, 유형 3의 경우는 보다 자연스러운 번역문의 생성을 위해서는 중국어의 표현 특성을 반영할 수 있는 추가의 연구가 이루어져야 할 것이다.

참고문헌

- [1] Kim, Hong, Huang, Kim, Yang, Seo & Choi (2002), "Korean-Chinese Machine Translation Based on Verb Patterns", AMTA.
- [2] 김혜경, 채영숙, 최기선(1999), "구 단위 패턴 기반 한영 기계 번역에서의 기능동사 구문의 중심어 선택 모델", 한글 및 한국어 정보처리 학술대회
- [3] 김혜경 (1996), 불어의 술어기능명사와 한국어 변환, 부산대학교 석사학위 논문.
- [4] 서정수 (1975), 동사 "하-"의 문법, 형설출판사.
- [5] 시정곤 (1994), "'X를 하다'와 'X하다'의 상관성", 국어학 24권.
- [6] 胡裕樹, 范曉(1995), 動詞研究, 河南大學出版社, 中國.