한중 기계번역 시스템을 위한 동사구 패턴 반자동 확장 방안 연구

언어처리연구팀, 음성/언어정보연구센터, 한국전자통신연구원
홍문표 · 류 철 · 김영길 · 박상규

A Study on Semi-Automatic Construction of Verb Patterns for a Korean-Chinese MT System

Mun-Pyo Hong, Cheol Ryu, Young-Kil Kim, Sang-Kyu Park
NLP Team, Speech/Language Technology Research Center, Daejeon, Korea

요 약

본 논문에서는 한-중 기계번역 시스템에서 사용되는 한중 동사구 패턴의 반자동 생성을 위한 방법론을 제안한다. 한중 동사구 패턴은 한국어와 중국어간의 변환을 위한 정보를 제공할 뿐만 아니라, 한국어의 구문분석과 중국어의 생성을 위해 중요한 정보를 제공하는 고급 언어자원이다. 본 논문에서 제시하는 새로운 패턴 반자동 확장 방안은 기존의 한중 동사구 패턴으로부터 대역어 정보를 이용하여 새로운 동사구 패턴을 생성해내는 방법이다. 본 방법론은 시스템 개발 초기에 일반적으로 이루어지는 사전기반 패턴 구축이 끝난 후, 패턴의 커버리지 문제를 해결하기 위해 실용적으로 적용할 수 있는 방법론으로서, 한국어와 중국어 같이 활용 가능한 대역 코퍼스가 아직 많지 않은 경우에 효과적이다. 본 논문에서 제시한 방법론은 실험 결과 67.15%의 정확률과 4.58배의 패턴 확장률을 나타냈다.

서 론

패턴 기반 기계번역 방법론(Pattern-based MT)은 번역 성능의 점진적인 향상(Scalability), 다양한 도메인에 대한 쉬운 적용성(Customization), 번역 알고리즘의 효과적인 디자인(Efficiency) 등의 이유로 인해 많은 언어쌍에 대해 적용되고 있다.[1,6,7] 그러나 패턴 기반 기계번역 방법론은 여타 데이터 기반 기계번역 방법론과 마찬가지로 대규모 번역지식의 획득을 위해 많은 시간과 비용이 든다는 단점을 가지고 있다. 이러한 지식 획득상의 문제점들을 해결하기 위해 여러 가지 지식 자동/반자동 구축 및 관리 방안이 제안되고 있다.[1,7,8] 또한 순수 데이터 기반 방식의 문제점들을 극복하기 위해 규칙기반 방식과 데이터 기반 방식의 장점들을 혼합하는 하이브리드 번역방식도 제기되고 있다.[3,4]

E-mail : hmp63108@etri.re.kr
E-mail : ryuch@etri.re.kr
E-mail : kimyk@etri.re.kr
E-mail : parksk@etri.re.kr

본 논문에서는 현재 한국전자통신연구원(ETRI)에서 개발중인 방송자막대상 한-중 기계번역 시스템 Tellus-KC에서 사용되는 한중 동사구 패턴의 반자동 생성을 위한 방법론을 제안한다. 한중 동사구 패턴은 한국어와 중국어간의 변환에 뿐만 아니라 한국어 구문 분석에도 유용하게 사용되는 중요한 언어자원이다. 이러한 동사구 패턴을 구축하기 위해 기계 가독형 사전(machine readable dictionary), 대역 코퍼스(bilingual corpus) 등이 사용된다. 기계 가독형 사전은 리소스 구축 초기에 엔트리 구성과 기본 패턴 구축을 위해 유용하게 사용될 수 있으나, 그 용례 등의 불충분함으로 인해 높은 커버리지의 패턴을 구축하기에 적합하지는 못하다. 대역 코퍼스를 사용하는 경우 번역 단위 간의 정렬(alignment) 등과 같은 기술적인 어려움은 차치하고서라도, 한국어와 중국어의 경우 대용량 코퍼스를 구하기가 쉽지 않다는 문제점이 있다. 또한 대용량이 아닌 대역 코퍼스를 가지고 작업을 하게 되면 패턴 구축이 도메인에 지나치게 치우치게 되는 문제점이 발생한다. 이러한 문제점들을 해결할 수 있고 사전 등을 토대로 하여 기구축된 동사구 패턴을 이용하는 방법이 대역어 정보를 이용한 패

턴 반자동 생성 방법론이다.

본 논문의 2장에서는 패턴의 반자동 생성을 위한 관련 연구들이 소개된다. 3장에서는 동사구 패턴의 개념과 동사구 패턴이 Tellus-KC 한중기계번역 시스템에서 차지하는 부분에 대해 언급한다. 4장에서는 대역어 정보를 이용하여 한중 동사구 패턴을 반자동으로 생성하는 방법론을 예제를 통해 설명한다. 5장에서는 앞 장에서 소개된 방법에 대한 실험 결과가 보고된다. 이 실험에서 드러난 본 방법론의 문제점도 언급된다. 끝으로 6장에서는 향후 연구개발 방향에 대해 논의한다.

관련연구

한중 동사구 패턴을 구축할 때 사전 편집자(lexicographer)의 언어지식에만 기반할 경우, 패턴의 완전성(completeness)이 결여되는 문제가 생긴다. 또한 일반 종이사전에 기반하여 작업을 하는 경우에도 종이사전에 등록되어 있는 예제의 불충분함으로 인해 유사한 문제가 발생한다. 이러한 문제점을 체계적으로 해결하기 위해 사전의 용언 엔트리에 대해 사/피동 링크를 부착하여 사전을 재구성하는 방법론이 [1]에서 제기되었다. 동사 '먹다'는 그 사/피동 활용형인 '먹이다', '먹히다'와 연결이 되어있고, 이러한 문법관계는 사전 작업자가, 예를 들어 'A=사람!가 B=육류!를 먹!다'라고 하는 패턴을 구축할 때, 워크벤치로 하여금 사동/피동 패턴 생성 메커니즘에 의해 'A=사람!가 B=사람!에게 C=육류!를 먹이!다', 'A=육류!가 먹히!다'와 같은 패턴을 생성하여 사전 작업자에게 제시할 수 있게 한다. 이와 같은 반자동 생성 메커니즘이 없을 경우, 패턴 구축 시 사/피동 형태 동사들간의 정보 공유가 일어나지 않기 때문에 동사구 패턴 구축의 효율성이 떨어지고 작업 결과의 일관성이 떨어지게 된다.

[1]에서는 또한 한국어 용언에 대한 중국어 용언의 대역 정보를 이용하여 패턴을 반자동 생성하는 방법론에 대해서도 일부 언급을 하였으나, 그 적용방법에 있어서 단순히 태(voice)정보만을 사용하여 패턴 자동생성의 정확률이 떨어진다.

[2]에서는 기존의 일본어 패턴에 대해 조사를 변이형으로 대치해 새로운 패턴을 생성해내는 방법론을 제안하고 있으나, 현재 한중 기계번역 시스템에서는 이미 모든 조사에 대해 각각의 대표형을 정의하고 그 대표값으로 패턴을 구축하고 있으므로 이 방법론을 바로 적용하기에는 어려운 점이 있다.

한중 동사구 패턴과 Tellus-KC

한중 번역시스템 Tellus-KC는 데이터 기반 방식을 따른다고 말할 수 있을 정도로, 입력문의 정확한 번역을 위하여 여러 가지 형태의 대용량 번역지식이 사용된다. 입력문의 분석 및 목표 언어로의 변환을 위하여 동사구 패턴이라고 하는 지식이 사용된다. 동사구 패턴은 일종의 하위 범주화 사전으로 볼 수 있으나, 차이점은 첫째, 용언의 하위범주 패턴에 대해 항상 대역 패턴이 존재한다는 점이고 둘째, 언어학적 관점의 하위범주 패턴과는 달리 동사구 패턴의 원문부에서는 용언의 논항 뿐만 아니라 부사와 같은 첨가어까지도 대역어 선택에 영향을 미칠 경우에는 기술된다는 점이다.

예를 들어 '자다' 동사에 해당되는 패턴의 일부는 다음과 같다,

자다1 : A=기상!가 자!다>A 停 : v[바람(A)이 자다]

자다2 : A=사람!가 자!다>A 睡觉 : v[아이(A)가 자다]

자다3 : A=계측도구!가 자!다>A 停 : v[괘종시계(A)가 자다]

자다4 : A=자연현상!가 자!다>A 平静 : v [파도(A)가 자다]

자다5 : A=재화!가 자!다>A 冻结 : v [자금(A)이 자다]

위의 예에서 보면 한중 동사구 패턴은 "원문부>대역부"와 같은 포맷으로 구성되어 있고, 원문부에는 한국어 용언이 갖는 논항 위치에 변수(A)와 이 변수자리를 차지하는 명사 논항의 의미정보가 기록되어 있다. ! 부호는 명사 논항과 조사를 구별해주거나 동사의 어간과 어미를 구별해주는 일종의 표지자(marker)이다. ">" 부호 오른편에 위치하는 대역부에는 중국어 대역어와 명사 논항의 위치정보를 담은 변수가 나온다. 중국어 용언에는 " : v" 표지자가 붙는다. 각 한국어 패턴은 해당되는 예문과 링크가 되어 있는데, 특히 명사 논항 변수와 예문에서 그에 해당하는 단어들도 정렬이 되어 있다. 이러한 링크관계는 향후 세밀한 대역어 선택 등을 위하여 명사 의미분류가 더욱 세밀해질 필요가 있을 때, 기존의 패턴을 새로운 의미체계에 따라 자동으로 수정하는데 기여한다.

명사의 논항 위치에 변수와 함께 사용되는 의미표지자들(semantic marker, 본고에서는 편의상 '의미코드'로 부르기로 함)은 Tellus-KC에서 사용되는 명사의 의미 분류체계에 따른 의미코드들이다. 현재 사용되고 있는 명사 의미코드는 총 8레벨, 414개의 노드를 포함한다. 최상위 노드는 '구체명사', '추상명사', '활동명사'로 나뉘어 있으

Table 1.

도 구	작업용도구	작업공구
		작업도구
	계측 도구	
	미용 도구	세면도구
		화장도구
	주방 도구	그 릇
		취사도구
		수 저 류
	수납 도구	상 자
		통
		병
		가 방
		봉 지
	가 구	
	조 명 등	전 등
		양 초

며 8레벨까지 자세히 분류되는 것은 '구체명사'이다. 현 의미체계는 기계번역이라는 특수한 관점에 맞춰 디자인 되었으나, 특정 목표언어에 특별히 튜닝된 점은 거의 없다고 할 수 있다.

다음은 '구체명사-물체-무생물-인공물-도구'에 속하는 일부 의미코드의 예이다(Table 1).

동사구 패턴은 이상에서 설명한 바와 같이 대역어로의 변환 및 대역문의 어순 생성과 같은 역할을 담당하지만, 또한 패턴의 원문부는 한국어 구문 분석에서 매우 중요한 정보로 사용된다. 한국어 동사구 패턴에 따라 문장의 의존구조(dependency structure)가 생성된다. 따라서 한중 동사구 패턴은 시스템의 번역률 및 구조 분석 정확률에 결정적인 영향을 미치게 된다.[1]

현재 한중 동사구 패턴은 약 1만 7천개의 한국어 용언에 대해 약 11만 5천개의 패턴이 구축되어 있다. 이 중 약 8만개의 패턴은 사전과 방송뉴스 코퍼스에 기반하여 수동/반자동의 방법으로 구축되었으며, 나머지 3만 5천개의 패턴은[1, 8] 등에서 제안한 방법과 용례로부터의 자동 추출 방법 등을 통해 반자동으로 구축되었다.

동사구 패턴의 완전 매칭 커버리지 측정을 위해 50문장의 방송 뉴스 자막과 50문장의 고등학교 교과서, 총 100문장으로 실험을 실시하였다. 100문장 내에 등장하는 총 259개의 용언에 대해 패턴 매칭 등을 위한 엔진 상의 오류가 없다는 가정 하에 42.8%의 완전 매칭률이 조사되었다. 현재 구축된 동사구 패턴의 학습 데이터가 방송 뉴스 자막이었다는 점을 감안해 볼 때 현 패턴의 커버리지는 다른 도메인 상에서는 좀 더 떨어질 것으로 예상된다.

따라서 현재 약 60% 정도인 한중 번역률과 구조분석 정확률을 획기적으로 높이기 위해서는 패턴의 커버리지를 단기간에 대폭적으로 높일 수 있는 또 다른 패턴 자동/반자동 구축방안이 마련되어야 한다.

패턴 반자동 확장 방안

본 논문에서 제시하는 새로운 패턴 반자동 확장 방안은 기존의 한중 동사구 패턴으로부터 대역어 정보를 이용하여 새로운 동사구 패턴을 생성해내는 방법이다.

우리는 일반적으로 다음의 예에서 볼 수 있는 바와 같이 어떤 동사들이 서로 의미가 유사한 경우 논항으로 취하는 명사들에 대한 의미적인 제약도 거의 유사함을 알 수 있다.

1) 사람!가 장소!로 진격하다
2) 사람!가 장소!로 돌진하다
3) 사람!가 멋있다
4) 사람!가 멋지다

따라서 만약 이러한 용언 관련 유의어 리소스가 충분히 존재한다면 이것으로부터 많은 패턴을 반자동으로 생성해 낼 수 있을 것이다. 그러나 현재 이러한 용도로 사용할 수 있는 유의어 리소스가 충분하지 않은 데 문제가 있다.

용언간의 의미적 유사도에 관해 단서를 제공하는 다른 것은 대역 정보이다. 즉, 어떠한 용언들이 대역어를 공유한다면, 이것은 두 용언의 의미가 동일하거나 유사하다는 단서가 될 수 있다.

한국어-영어의 예를 들면 영어 동사 "to give"를 대역어로 취할 수 있는 한국어 동사들은 대략 "주다/드리다/수여하다/기부하다" 등으로 볼 수 있다. 이 동사들은 모두 유사한 의미를 공유하며, 취하는 논항 명사들에 대한 의미제약도 거의 같다고 볼 수 있다. 마찬가지 예로 "to watch"를 대역어로 갖을 수 있는 한국어 동사들은 "보다/관찰하다/시청하다/쳐다보다" 등이며, 이들도 유사한 의미와 논항 명사들에 대한 의미제약을 공유한다.

이와 관련된 연구로서 [5]는 영어 동사의 의미를 동사가 취하는 격틀 및 논항에 대한 의미제약 등으로 분류하였다.

이러한 경험적 사실을 바탕으로 Tellus-KC 시스템을 위한 한중 동사구 패턴을 기존의 한중 동사구 패턴으로부터 대역 정보를 사용하여 반자동으로 생성한다. 대역 정보를 사용한 패턴 반자동 생성은 대역어를 공유하는 동사구

1 입력문장에 대해 정확히 매칭하는 동사구 패턴이 존재하지 않을 경우 가장 유사한 패턴이 적용된다. 유사한 패턴도 존재하지 않는 경우에는, 대역 사전에 기본값으로 존재하고 있는 동사의 대역어가 생성되게 된다.

패턴들로부터 상호 참조 과정을 통해 새로운 패턴을 만들어 내는 방법이다 : 예) "捐赠 동사의 경우

기증하다 : A=사람!가 B=사람!에게 C=내부기관!를 기증하!다[그 남자(A)가 그 환자(B)에게 신장(C)을 기증했다]

드리다 : A=사람!가 B=통신기기!를 C=사람!를위해 드리!다[나(A)는 핸드폰(B)을 어머니(C)를 위해 드렸다]

자동생성 결과 :

기증하다 : A=사람!가 B=사람!에게 C=내부기관!를 기증하!다⇒A=사람!가 B=통신기기!를 C=사람!를위해 기증하!다 [나(A)는 핸드폰(B)을 그 아이들(C)을 위해 기증했다]

드리다 : A=사람!가 B=통신기기!를 C=사람!를위해 드리!다⇒A=사람!가 B=사람!에게 C=내부기관!를 드리!다 [그 남자(A)는 어머니(B)에게 신장(C)을 드렸다]

패턴 반자동 생성 프로세스는 다음과 같다 :

1단계 : 한중 동사구 패턴을 중국어 동사에 따라 재소팅한다.

예) 중국어 동사 1(给)

A=사람!가 B=자동차!를 드리!다

A=사람!가 B=사람!에게 C=채소!를 주!다

A=사람!가 B=재화!를 수여하!다

중국어 동사 2(停止)

A=사람!가 B=건축토목활동!를 그만두!다

A=조직!가 B=위반행위!를 관두!다

2단계 : 동일한 중국어 동사(대역어)를 가지며 표제어가 다른 한국어 패턴끼리 1 : 1 비교한다.

예) 중국어 동사 1(给)

A=사람!가 B=자동차!를 드리!다

A=사람!가 B=사람!에게 C=채소!를 주!다

3단계 : 다음의 세가지 조건을 모두 만족하면, 다른 (한국어) 표제어 밑에 있는 (한국어) 동사구 패턴을 자신의 표제어를 사용하여 치환한다.

세가지 조건

* 비교 대상인 두 한국어 동사 표제어의 태(Voice)가 같다

* 비교 대상인 두 한국어 동사 중 어느것도 어휘패턴이 아니다

* 중국어 동사가 加以, 进行, 做, 作 가 아니다

예) A=사람!가 B=사람!에게 C=채소!를 드리!다

4단계 : 생성된 한국어 패턴을 기존의 동사구 패턴에서 찾아본다. 만약 기존에 동일한 패턴이 존재하면 새로 생성된 패턴은 버린다

이제 3단계에서 왜 상호참조를 통하여 패턴을 복사하기 전에 세가지 조건을 만족해야 하는지에 대해 살펴보자.

첫번째 조건인 두 한국어 동사 표제어의 태(Voice)가 같아야 한다는 조건은 다음과 같은 예에 기인한다.

5) 漂 : v

뜨다50|A=식물기관!가 B=위치!에 뜨!다>A 漂 : v 在 B 上[나뭇잎(A)이 물위(B)에 뜨다]

띄우다16|A=사람!가 B=장소!에 C=식물기관!를 띄우!다>A 使 C 漂 : v 在 B 上[아이(A)가 물위(B)에 나뭇잎(C)을 띄우다]

6) 滥用 : v

사용되다38|A=사람!에의해 B=약품!가 사용되!다 > B 被 A 滥用 : v[한국 사람들(A)에의해 약(B)이 함부로 사용되다]

사용하다17|A=사람!가 B=약품!를 사용하!다 > A 滥用 : v B [한국 사람들(A)은 약(B)을 함부로 사용한다]

현재 한중 동사구 패턴에서 중국어 본동사에 대해 ' : v' 와 같은 표지자가 사용되고, 패턴 소팅 시 이 표지자를 기준으로 한국어 패턴들이 정렬된다. 그러나 "띄우다16" 패턴과 같은 경우 실제로 본 동사 앞에 있는 조동사 "使"에 의해 논항구조가 동사 기본형의 경우와 달라지게 된다. 따라서 비교 대상이 되는 두 동사의 태(Voice) 정보에 대한 비교 없이 단순히 대역 정보만으로 패턴을 생성해내는 방법은 매우 위험하다고 할 수 있다. 어떠한 동사가 기본형(base form)인지, 아니면 사/피동 활용형인지를 구별하는 정보는 [1]에서 제시된 재구성된 사전과 동사구 패턴 간의 링크 관계를 통해 얻어진다.

두번째 조건인 어휘패턴 여부에 대해 살펴보기로 하자. 어휘패턴이라 함은 한국어에 나타나는 일종의 연어(Collocation)에 대한 패턴으로서, 특정 동사와 특정 명사가 일반적인 동사/명사간의 결합 정도보다 긴밀히 결합하여 쓰이는 경우를 말한다. 특정 중국어를 대역어로 취하는 패턴이 여러 개 있을 때, 그 중 어휘패턴이 존재한다면, 대역어를 공유하는 패턴과 이 어휘패턴간에는 상호참조를 통하여 패턴을 복사/생성해낼 수 없다. 다음의 예를 보자.

7) 下决心 : v⇔결심하다/결의하다/다짐하다/먹다/서다

A=사람!가 B=작업!를 결심하!다

A=조직!가 B=싸움!를 결의하!다

A=조직!가 B=약속!를 다짐하!다

A=사람!가 마음!를 먹!다

A=사람!가 결정!가 서!다

"下决心"을 대역어로 취하는 한국어 패턴 중 마지막 두 개의 패턴은 어휘패턴으로 분류된다.[2] 어휘패턴은 앞서 설명한 바와 같이 연어관계를 갖는 패턴이므로 동사가 다른 아무 명사와 결합할 수 없음은 자명하다. 거꾸로 다른 동사들도 연어관계에 등장하는 특정명사와 결합할 수 없음도 마찬가지이다.

8) A=사람!가 마음!를 결심하다 (X)
A=사람!가 마음!를 결의하다 (X)
A=사람!가 마음!를 다짐하다 (X)
A=사람!가 결정!가 결심하다 (X)
A=사람!가 결정!가 결의하다 (X)
A=사람!가 결정!가 다짐하다 (X)

마지막으로 중국어 동사가 加以, 进行, 做, 作가 아니어야 한다는 조건은 위의 동사들이 중국어에서 대표적인 기능동사(Support Verb)이기 때문이다. 중국어에도 다른 언어들과 마찬가지로 "기능 동사+술어성 명사" 의 구조가 존재하는데, 이러한 구조에서 문장의 논항구조를 결정하는 것은 기능 동사라기 보다는 술어성 명사(predicative noun)라고 볼 수 있다. 따라서 이러한 동사들이 본동사로 표지되어 있는 경우 패턴을 상호 참조하여 생성하게 되면 잘못된 패턴이 생성될 수 있다.

9) 作 : v
딸랑거리다2|A=방울!가 딸랑거리!다>A 作:v 厥当声[방울(A)이 딸랑거리다]
싸우다13|A=사람!가 B=무생물속성!와 싸우!다>A 为作 : v 斗争[주부(A)가 물가(B)와 싸운다]
운동하다7|A=사람!가 B=장소!에서 운동하!다 > A 在 B 作:v 运动[두명(A)이 땅(B)에서 운동을 하다]

위의 경우 "딸랑거리다", "싸우다", "운동하다"는 모두 대역어로서 "作" 동사를 취하지만 실제 각각의 논항구조를 결정하는 것은 술어성 명사인 "厥当声", "斗争", "运动"이다. 따라서 이와 같은 특징을 고려하지 않고 생성하게 되면 다량의 잘못된 패턴이 만들어지게 된다.

실험 및 분석

4장에서 제안된 방법론의 성능을 분석하기 위해 간단한 실험을 실시하였다. 실험을 위해서 우선 20개의 중국어 동사를 무작위로 추출하고, 이 동사들을 대역어로 취하는 총 323개의 한국어 패턴들을 추출하였다. 이 패턴들에 대해 4장에서 제안한 방법론으로 패턴을 생성한 후 그 정확률과 패턴 확장률을 조사하였다. 실험 결과는 다음의 Teble 2에 나타나있다.

Table 2. 대역어 기반 동사구 패턴 확장 실험 결과

중국어 동사		20
자동생성 패턴		323
자동생성 패턴	전 체	2,399
	올바른 패턴	1,611
	새롭게 얻어진 패턴(기존 패턴 제외)	1,478
	잘못된 패턴	788
정 확 률		67.15%
확 장 률		4.58배

실험결과 본 논문에서 제시한 방법론은 67.15%라는 비교적 높은 정확률과 기존의 패턴과 겹쳐져서 삭제된 패턴을 제외하고서도 4.58이라는 매우 높은 패턴 확장률을 보였다.

이 실험에서 드러난 오류 패턴의 유형들을 살펴보면 다음과 같다.

- 한국어 동사의 의미도 유사하고 취하는 논항 명사의 의미제약도 유사하나 취하는 조사가 다른 경우(가장 많은 오류), 예) ~에게 맞서다/~와 대결하다
- 한국어 동사의 의미는 유사하나 논항으로 취하는 명사의 의미제약이 다른 경우, 예) [서류]를 교부하다/ [돈]을 납부하다.[3]
- 한국어 동사가 의미적으로 유사성이 거의 없는 경우(중국어 대역어만 우연히 동일한 경우), 예) 스무살 남짓하다/언덕을 오르내리다.
- 어휘패턴은 아니지만 은유적이거나 숙어적인 한국어 패턴, 예) 표결에 부치다/학생에게 서류를 교부하다.

결 론

한중 동사구 패턴은 한국어와 중국어간의 변환을 위한 정보를 제공할 뿐만 아니라, 한국어의 구문분석과 중국어의 생성을 위해 중요한 정보를 제공하는 고급 언어자원이다. 이러한 동사구 패턴을 구축할 때 일반적으로 기계 가독형 사전 등을 이용하여 초기 패턴 엔트리의 구성 및 표제어에 대한 대표적인 패턴을 구축하게 된다. 시스템 개발의 초기단계에서 이와 같이 기계 가독형 사전과 이에 부착된 대표적 용례를 사용하여 패턴을 구축하게 되면, 대표적인 표제어를 빠짐없이 구축할 수 있는 장점이 있으나, 그

2 어휘패턴은 논항 자리에 변수(A, B, C 등)가 없이 바로 어휘가 사용된다는 점에서 일반 패턴과 구별이 된다
3 "교부하다", "납부하다"의 의미는 유사하나 "서류를 납부하다", "돈을 교부하다" 라는 어색한 표현이 생성됨.

용례의 부족함으로 인해 패턴의 커버리지가 떨어지게 된다. 이러한 문제점을 해결할 수 있는 방법은 대용량 대역코퍼스를 이용하거나 본 논문에서 소개한 대역정보를 이용한 패턴 반자동 생성 방법이다. 현재 확보 가능한 한-중 대역코퍼스가 충분하지 않음을 고려할 때, 본 논문에서 제시한 대역어를 이용한 패턴 반자동 생성 방법론은 언어쌍에 상관없이 시스템 개발 중반기에 충분히 고려할 수 있는 방법론이라 생각된다.

현재 ETRI에서는 기구축된 11만 5천 한중 동사구 패턴에 대해 본 논문에서 소개한 방법론을 적용하여 패턴을 자동으로 생성하고, 한중 사전편집자(lexicographer)에 의해 패턴을 정제하는 작업을 진행 중이다. 이러한 방법으로 대용량의 한중 동사구 패턴이 구축되면 한중 기계번역 시스템의 번역률 향상은 물론 한국어 구조분석기의 일반적인 성능 향상에 크게 기여할 수 있으리라 예상된다.

REFERENCES

1) 홍문표, 김영길, 류 철, 최승권, 박상규(2002) : "사전 재구성과 대역어 정보를 통한 동사구 패턴의 확장 및 관리", 제 14회 한글 및 한국어 정보처리 학술대회
2) T Baldwin, F Bond(2002) : *Alternation-based Lexicon Reconstruction, in TMI*
3) M Hong, K Lee, Y Roh, S Choi, S Park(2003) : *Sentence-Pattern based MT revisited, in ICCPOL*
4) D Kim, J Lee(2003) : *Full Interaction between Example-based and Rule-based Engines in a Hybrid Chinese-to-Korean MT, in ICCPOL*
5) B Levin(1993) : *English verb classes and alternation, The University of Chicago Press*
6) K Takeda(1996) : *Pattern-based Machine Translation, in COLING*
7) S Yamada, K Imamura, K Yamamoto(2002) : *Corpus-Assisted Expansion of Manual MT Knowledge, in TMI*
8) S Yang, M Hong, Y Kim, C Kim, Y Seo, S Choi(2002) : *An Application of Verb-Phrase Patterns to Causative/Passive Clause, in IASTED*