번역학연구
2010년 가을 제11권3호

번역 말뭉치로부터 추출한 어휘 번역 패턴의 의미 분류와 자동번역시스템에의 활용

최승권 · 김영길
(한국전자통신연구원)

1. 서론

자동 번역기는 대용량의 문서를 고속으로 일관성 있게 번역하는 장점이 있다. 하지만 인간 번역가보다는 자연스럽지[1] 못한 번역 결과를 제공한다는 단점을 가지고 있다. 자동 번역기가 자연스럽지 못한 번역 결과를 제공하는 원인은 자동 번역기의 번역 방법에 따라 차이가 있겠지만 출발 언어(Source language)의 구조를 그대로 목표 언어(Target language)의 구조로 가져와서 목표 언어의 표현을 만들기 때문이다. 이러한 자동 번역기의 번역 결과를 더욱 자연스럽게 만들기 위해서는 번역가의 번역 능력을 명시화하여 자동 번역 시스템에 활용하면 가능

1) 번역의 품질을 평가하는 방법으로 정확성(accuracy), 명료성(clarity), 스타일(style)이 있다(Hutchins, 1992: 163). 본 논문의 "자연스러운" 번역이란 정확성과 명료성을 합쳐놓은 개념으로 번역된 결과가 원문의 의미와 일치하면서 독자가 쉽게 이해할 수 있음을 의미한다.

www.kci.go.kr

할 것이다. 이런 번역가의 번역 능력 중 하나가 어휘 번역 패턴이라 할 수 있다.

본 논문에서는 어휘 번역 패턴을 번역 말뭉치로부터 반자동으로 추출하고 자동 번역 시스템에 활용함으로써 부자연스럽던 자동 번역 결과를 자연스럽게 하는 방법을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 자동 번역의 부자연스런 번역 결과를 소개하고자 한다. 부자연스런 번역 결과는 본 논문에서 제시하고자 하는 어휘 번역 패턴의 필요성을 설명하는 토대가 될 것이다. 3장에서는 어휘 번역 패턴의 기존 연구에 대해 살펴보고자 한다. 4장에서는 어휘 번역 패턴을 반자동으로 구축하는 방법에 대해 기술하고자 한다. 5장에서는 어휘 번역 패턴을 의미적으로 분류한 결과를 기술하고자 하며, 6장에서는 어휘 번역 패턴을 적용한 실험 결과를 제시하고자 한다.

2. 부자연스런 자동 번역 결과

부자연스런 자동 번역 결과는 출발 언어 문장을 목표 언어 문장으로 구조적으로 의미적으로 알맞게 번역하지 못하기 때문에 발생하는 경우가 많다. 자동 번역이 부자연스러운 이유는 크게 두 가지로 나누어 볼 수 있다. 하나는 원문과 관련된 것이며, 다른 하나는 번역문과 관련된 것이다. 원문과 관련된 것은 관용구와 같은 비조합적(non-compositional) 문자열을 인식하지 못하여 발생하며 번역문과 관련된 것은 원문과 번역문의 언어적 차이를 인식하지 못하여 발생한다. 예 (1a-b)가 원문과 관련되며, (1c-d)가 번역문과 관련된 것이다.

(1) a. Pilot program goes into effect.
[자동번역] 시범 프로그램이 효과로 들어간다.
b. The heavy snowfall prevented me from arriving on time.
[자동번역] 폭설이 내가 정각에 도착하는 것을 방해했다.
c. A few of my girlfriends may come.
[자동번역] 내 여자 친구의 몇몇이 올 것이다.
d. Learning a foreign language helps you find a solution.
[자동번역] 외국어를 배우는 것은 당신이 해법을 발견하게 돕는다.

www.kci.go.kr

(1a)는 ‘go into effect’라는 영어 관용구를 인식하지 못한 예이고 (1b)는 ‘prevent A from B’의 관용구에서 영어 주어 ‘the heavy snowfall’를 한국어 주어로 번역함으로써 부자연스럽게 번역된 경우이다. (1c)는 ‘a few of’가 한국어 번역에서 명사구 밖으로 나와 번역되는 것이 자연스러우나 그렇지 못하여 부자연스럽게 된 경우이다. (1d)는 ‘help’ 동사를 한국어에서 ‘돕다’라는 어휘로 번역함과 동시에 ‘learning a foreign language’라는 동명사구를 한국어의 주어로 번역함으로써 한국어 문장이 어색해진 경우이다.

3. 관련 연구

규칙에 기반하여 자동 번역이 이루어지는 것이 규칙 기반 자동 번역 방법이다. 이 방법의 문제점은 자연스런 번역을 기술하기 위한 규칙을 작성하는데 어려움이 있으며 작성된 규칙일지라도 서로 충돌 현상이 발생해 규칙을 관리하는데 문제를 가지고 있다. 이러한 규칙 기반 자동 번역 시스템의 문제점을 해결하기 위해 개발된 방법이 어휘 번역 패턴에 기반한 패턴 기반 자동 번역 방법이다 (Takeda, 1996: 1155). 어휘 번역 패턴이란 반복적으로 나타나는 원문의 표현 구조와 그것에 대응되는 번역문의 표현 구조를 어휘와 구문노드를 이용하여 쌍으로 기술한 형태를 말한다. 예를 들어 “take a look at”이라는 영어 숙어는 “take:VERB:1 a look at NP:2 => VP:1 VP:1 <= NP:2 를 보:V:1.”과 같은 번역 패턴에 의해 한국어로 “~를 보다”로 번역할 수 있다는 것이다.

패턴 기반 자동 번역 시스템에 의해 자연스런 자동 번역 결과를 만들려는 노력은 언어간 번역 차이를 극복하려는 노력으로 이어졌으며, 이러한 노력은 Dorr(1994)와 양승현(1997)에서 잘 정립되었다.

Dorr(1994)에서는 번역할 언어 쌍에 따른 번역 차이를 7가지 유형으로 나누고 이들 번역 차이를 통사 구조(syntactic structure)와 중간 언어 표현(interlingual representation) 간의 매핑 규칙에 의해 설명하려고 하였다. 그녀에 따르면 ‘I like eating <=> Ich esse gern ‘I eat likingly’’이라는 영어<->독일어 번역은 핵심어와 논항의 위치가 바뀌는 Demotional divergence라는 매핑 규칙에 의해 번역될 수 있다는 것이다.

www.kci.go.kr

(2) Demotinal divergence의 번역 예

$$[_{C-MAX}\ [_{I-MAX}\ [_{N-MAX}\ I]_i\ [_{V-MAX}\ [_V\ \text{like}]\ [_{C-MAX}\ PRO_i\ \text{to eat}]]]]$$

$$\Updownarrow$$

$$[_{State}\ BE_{Circ}\ ([_{Thing}\ I],\ [_{Position}\ AT_{Circ}\ ([_{Thing}\ I],\ [_{Event}\ EAT\ ([_{Thing}\ I],\ [_{Thing}\ FOOD])])]\ [_{Manner}\ LIKINGLY])]$$

$$\Updownarrow$$

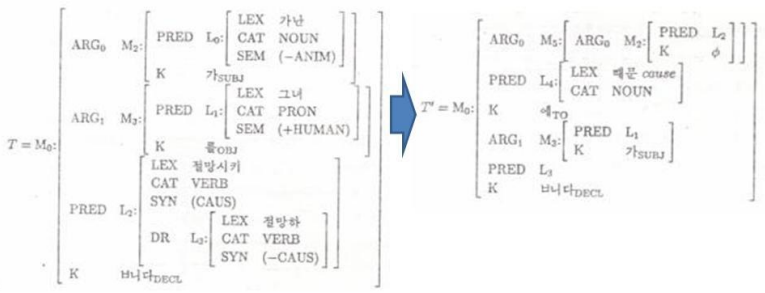
$$[_{C-MAX}\ [_{I-MAX}\ [_{N-MAX}\ \text{Ich}]\ [_{V-MAX}\ [_V\ [_{ADV}\ \text{gern}]\ [_V\ \text{esse}]]]]]$$

그러나 이 방법은 모든 단어를 중간 언어 방식으로 기술하는 어려움과 매핑 규칙이 적용되었을 때 적절한 어휘를 선택하기가 어렵다는 문제점 때문에 더 이상 개발이 되지 않고 있다.

양승현(1997)에서는 번역할 언어 쌍에 따른 번역 차이를 7가지의 스타일 변환(style transfer) 규칙으로 설명을 하였다. 스타일 변환 규칙은 출발 언어 구조를 그대로 목표 언어로 가져온 상태에서 이 구조를 목표 언어의 스타일에 맞게 바꾸는 것을 말한다. 예를 들어 “Poverty drove her to despair.”라는 영어 문장이 한국어로 그대로 오면 “가난은 그녀를 절망시켰다”인데, 이 문장을 한국어 스타일로 바꾸면 “가난 때문에 그녀는 절망했다.”로 되어야 한다는 것이다. 이 예문에 적용되는 스타일 변환 규칙을 소개하면 다음과 같다.

(3) Intransitivization의 스타일 변환 규칙

가난은 그녀를 절망시켰습니다 → 가난 때문에 그녀는 절망했습니다



스타일 변환 규칙은 번역 품질을 개선하는 좋은 아이디어를 제공하긴 하였으나 두 가지 점에서 문제가 있었다. 첫 번째는 스타일 변환 규칙의 위치이다. 스타일 변환 규칙이 자동 번역 결과가 나온 뒤에 적용되기 때문에 자동 번역

www.kci.go.kr

결과가 개선되면 잘 적용되던 스타일 변환 규칙도 잉여적인 규칙이 될 수 있다. 두 번째 문제는 스타일 변환 규칙의 구축과 관련된 것으로 스타일 변환 규칙의 표현이 예 (3)에서와 같이 확장된 자질 구조(extended feature structure)로 되어 있기 때문에 너무 복잡해 대량으로 구축하는 데 어려움이 있다.

기존의 어휘 번역 패턴들과 본 논문에서 제시하려는 어휘 번역 패턴의 차이점은 다음과 같다. Takeda(1996)는 통사 위주의 어휘 번역 패턴으로 기술하는 편리성은 존재하지만 의미가 배제되어 번역을 표현하는 데 한계성이 있는 반면, 본 논문의 어휘 번역 패턴은 통사 및 의미를 부여할 수 있어 기술의 편리성 및 번역의 한계성이 더욱 개선되었다. Dorr(1994)와 양승현(1997)은 의미가 부여되어 기술의 한계성은 해결하였지만 앞의 예에서 본 바와 같이 기술하는 표현이 복잡하여 가독성이 떨어지는 반면, 본 논문의 어휘 번역 패턴은 의미뿐만 아니라 전산 처리가 가능한 통사 구조를 가지기 때문에 기술의 편리성, 번역 한계성의 극복 및 전산 처리의 편리성을 제공하는 장점을 가진다.

4. 어휘 번역 패턴의 반자동 구축 방법

자동 번역 시스템에서 어휘 번역 패턴은 높은 재사용성, 편리한 유지 및 관리, 어휘 단계에서의 기술 가능, 자동 번역률의 점증적 증가(최승권 2002: 1), 원문과 번역문의 동시성(Shieber et. al. 1990: 253), 자동 번역 속도의 향상(Watanabe et. al. 1998: 1370)과 같은 장점을 가진다. 하지만 이러한 장점들은 어휘 번역 패턴을 대량으로 구축할 수 있다면 더욱 효과가 날 수 있다.

어휘 번역 패턴을 대량으로 구축하기 위한 방법으로 번역가의 번역 말뭉치로부터 번역 패턴을 통계적 정렬(statistical alignment) 방법에 의해 자동으로 추출하여 구축하는 방법이 있다(Hearne et. al. 2007: 85)(Ohara et. al. 2003: 150). 그러나 이 방법은 영어와 한국어 같이 구조 차이가 많이 나는 어족 간에는 그 정확도가 높지 않기 때문에 현재도 세계적으로 연구 중에 있다.

본 논문에서 소개하는 어휘 번역 패턴 구축 방법은 통계적 방법과 수작업을 혼합한 반자동 구축 방법이다. 원문 말뭉치로부터 원문 어휘 패턴 후보를 통계적으로 추출하여 수작업에 의해 어휘 번역 패턴을 구축하는 방법은 ETRI에

www.kci.go.kr

서 현재 개발 중에 있는 방법이다. 이에 반해 대역 말뭉치와 자동 번역 결과를 비교하여 어휘 번역 패턴을 수작업으로 구축하는 방법은 기존의 방법(Yamada et. al. 2002)(Choi et. al. 2008: 161)을 응용한 방법이다.

4.1. 어휘 번역 패턴의 포맷

어휘 번역 패턴이 되기 위해서는 다음과 같은 조건을 만족해야 한다.

(4) 어휘 번역 패턴이 되기 위한 조건
a. 어휘 번역 패턴은 원문부와 대역부로 이루어진다.
b. 원문부에는 원문 어휘가 최소한 1개 존재하여야 한다.
c. 원문부와 대역부는 의미적으로 완성된 형태이다.
d. 원문부의 변수들은 대역부에 동일한 형태로 나타난다.
e. 원문부와 대역부가 일대다가 되면, 대역부의 패턴들은 서로 변별력을 가져야 한다.

이상의 어휘 번역 패턴이 되기 위한 조건을 만족하는 어휘 번역 패턴의 포맷을 개략적으로 기술하면 다음과 같은 형태가 된다.

(5) 어휘 번역 패턴의 포맷
{원문부_패턴} <-> {대역부_패턴1} ({대역부_패턴n})*

위의 포맷에서 패턴은 '{..}' 안에 기술이 되며, '<->'는 번역 방향을 의미하며 '*'는 0개 이상을 의미한다. 따라서 패턴은 화살표의 방향에 따라 양방향으로도 단방향으로도 적용될 수 있으며 대역부 패턴은 1개 이상이어야 한다.

4.2. 어휘 번역 패턴의 반자동 구축 방법

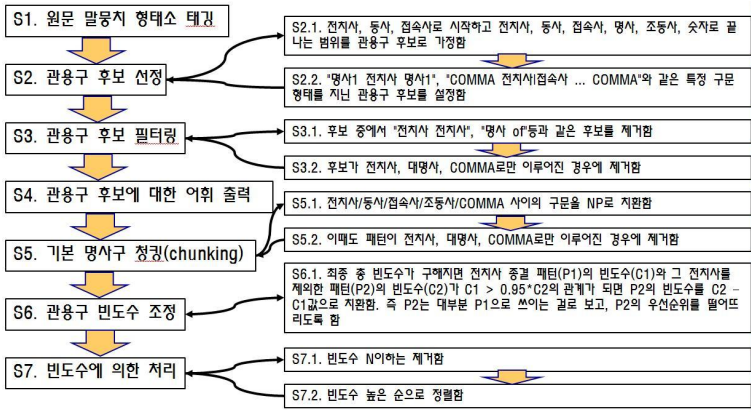
4.2.1. 원문 관용구 인식과 어휘 번역 패턴의 구축

원문의 관용구 인식과 그것의 어휘 번역 패턴을 구축하는 방법은 크게 두 단계로 나눌 수 있다. 첫 번째 단계는 번역 말뭉치로부터 원문의 관용구 후보를

www.kci.go.kr

자동으로 추출하는 것이며, 두 번째 단계는 추출된 관용구 후보에 번역 말뭉치를 보면서 어휘 번역 패턴을 구축하는 것이다.

〈그림 1〉 원문 관용구 후보의 자동 추출 방법



원문 관용구 후보를 추출하기 위해서 할 첫 번째 작업은 원문에 대해 해당 언어 형태소 분석기로 품사를 부착하는 형태소 태깅 작업이다(S1). 형태소 태깅된 결과를 기반으로 관용구 후보를 선정한다(S2). 관용구 후보는 전치사로 시작하면 전치사구, 동사로 시작하면 동사구, 접속사로 시작하면 접속사구이며(S2.1), 명사로 시작하면 명사구, 콤마로 시작하면 삽입구로 가정한다(S2.2). 예를 들어 이에 해당하는 것이 “step by step” 또는 “, as you know”와 같은 것이다. 이와 같이 만들어진 관용구 후보에 대해 필터링을 하는데(S3), 전치사구 중에 “전치사+전치사”로 시작하는 것이나 명사구 중에 “명사+of”로 되어 있는 구문은 제거하고(S3.1), 패턴이 전치사, 대명사, 콤마로만 되어 있는 경우, 예를 들어 “for it” 같은 것은 제거한다(S3.2). (S3.2)까지 만들어진 품사로 되어 있는 관용구 후보에 대해 어휘를 출력한다(S4). 출력된 어휘와 품사의 조합을 대상으로 기본 명사구 청킹을 시도해 명사구를 변수로 치환한다(S5). 전치사, 동사, 접속사, 조동사, 콤마들 사이의 구문을 NP로 치환한다(S5.1). (S5.1)의 단계에 의해 ‘accuse a person of theft’라는 관용구 후보는 ‘accuse NP of NP‘로 바뀔 것이다. 또한 이때도 후보가 전치사, 대명사, 콤마로만 이루어진 경우는 삭제한다

www.kci.go.kr

(S5.2). 이러한 예가 ’in NP of’와 같은 경우에 해당한다. 명사구인 NP까지 치환된 형태의 관용구 후보를 대상으로 빈도수 조정을 한다(S6.) 빈도수를 조정하는 이유는 포함 관계를 조정하기 위해서다. 예를 들어 ‘in spite of’의 빈도수가 1,000이고, ‘in spite’의 빈도수가 990이면 ‘in spite’의 빈도수에서 ‘in spite of’의 빈도수를 빼서 ‘in spite’의 빈도수를 10으로 만들어서(S6.1) ‘in spite’는 대부분 ‘in spite of’로 쓰이는 걸로 보고 나중에 빈도수에 의한 필터링에 의해 제거하려는 것이다.(S7). 추출된 관용구 후보 중 저빈도로 사용되는 후보는 제거하고(S7.1) 빈도수 높은 순으로 관용구 후보를 정렬하여 향후 번역가에 의해 검증하는 단계를 밟도록 한다.(S7.2).

위에서 설명한 관용구 후보의 자동 추출 방법에 의해 영어 기업 문서를 대상으로 관용구 후보를 자동으로 추출한 예를 보이면 (6)과 같다.

(6) 영어 기업 문서에서 자동으로 추출된 관용구 후보들

빈도수	원문관용구후보
17,090	for more information
4,091	in term of
3,277	see NP on page NUM
2,556	for information about
1,664	use the following command
1,125	for more detail
207	make it possible for NP to
37	in the early day of

자동으로 추출된 원문 관용구 후보에 대해 대역 말뭉치를 토대로 어휘 번역 패턴을 수동으로 구축한 예가 (7)이다.

(7) 기업 문서에서 구축한 어휘 번역 패턴

빈도수	원문관용구후보 <=> 대역어휘패턴
17,090	for more information <=> 더 많은 정보를 원하면
1,125	for more detail <=> 더 세부적인 사항을 원하면
207	make it possible for NP to VP <=> NP:가 VP:는 것을 가능하게 하다
37	in the early day of NP <=> NP:의 초창기에

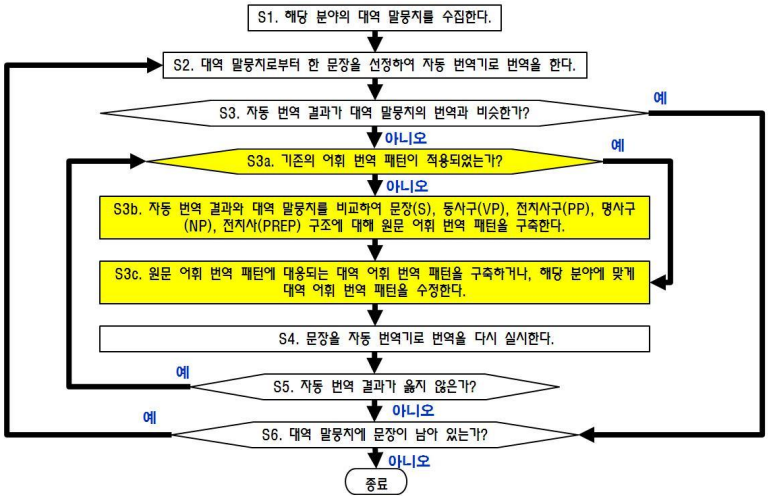
www.kci.go.kr

(6)의 예 중에 (7)에서 없어진 예들은 올바른 어휘 번역 패턴이라 여겨지지 않기 때문에 삭제한 것이다. (7)에서 동사구를 나타내는 VP가 나타난 이유는 해당 원문 어휘 번역 패턴의 구조를 의미적으로 완성시키기 위한 것이다.

4.2.2. 원문-번역문의 언어 차이 파악과 어휘 번역 패턴의 구축

본 절에서는 어휘 번역 패턴을 실제적으로 구축할 수 있는 방법으로서 번역 말뭉치와 자동 번역 결과를 비교하면서 어휘 번역 패턴을 구축하는 방법인 Yamada(2002)와 Choi(2008: 161)를 응용한 방법을 기술하고자 한다.

〈그림 2〉 원문-번역문의 언어 차이 파악과 어휘 번역 패턴 구축 흐름도



예를 들어 (8)과 같은 기업 문서의 한 문장이 입력되었다고 가정해 보자.

(8) Setting this DWORD registry key to 1 prevents the Transport Scan Job from scanning IPM replication messages. <=> 이 DWORD 레지스트리 키를 1로 설정하면 Transport Scan Job은 IPM 복제 메시지를 스캔하지 않아도 된다.

www.kci.go.kr

<그림 2>의 절차에 의해, (8)의 영어 문장은 (9a)와 같이 초기 자동 번역에서는 어색하게 번역되지만 구축한 (9c)와 같은 어휘 번역 패턴에 의해서 최종 자동 번역에서는 (9d)와 같이 자연스런 번역 결과로 만들어진다는 것이다.

(9) a. S2: 1에 이 DWORD 레지스트리 키에서 설정하는 것은 수송 스캔 업무를 주사 IPM 복제 메시지에서 막는다.
b. S3b: {set NP1 to NP2}
{VP1 prevent NP from VP2}
c. S3c: {set! NP1 to NP2} <-> {NP1:을 NP2:로 설정하!다}
{VP1 prevent! NP from VP2} <-> {VP1:면 NP:은 VP2:지 않아도 되!다}
d. S4: Setting this DWORD registry key to 1 prevents the Transport Scan Job from scanning IPM replication messages. 이 DWORD 레지스트리 키를 1로 설정하면 Transport Scan Job 은 IPM 복제 메시지를 스캔하지 않아도 된다.

5. 어휘 번역 패턴의 의미 분류

어휘 번역 패턴은 패턴이 담당하는 표현 범위에 따라 통사적으로 크게 세 가지로 분류할 수 있다(최승권 2007: 301). 첫 번째는 구문 어휘 번역 패턴으로서 명사구, 전치사구, 부사구 등 동사를 포함하지 않는 구 단위의 어휘 번역 패턴이다. 두 번째는 동사구 어휘 번역 패턴으로서 하나의 동사를 포함하는 동사 관용구와 같은 어휘 번역 패턴이다. 세 번째는 문장 전체를 대상으로 하는 문장 어휘 번역 패턴이다. 어휘 번역 패턴의 통사적 분류와 포맷은 최승권(2007)에서 자세히 기술한 바 있어 본 논문에서는 어휘 번역 패턴을 의미적으로 분류하는 데 초점을 두어 기술하고자 한다.

5.1. 관용구

대역 말뭉치의 원문으로부터 구축할 수 있는 전형적인 어휘 번역 패턴이 관용구다. 관용구는 어떤 구의 전체적 의미가 그것을 구성하고 있는 각 형태소

www.kci.go.kr

의 의미의 조합으로 설명할 수 없는 형태를 말한다. 본 논문에서는 관용구에 문법적 가변 요소의 필요 여부에 따라 세분화된 2가지 유형을 설명하고자 한다.

5.1.1. 가변 요소가 불필요한 관용구

가변 요소가 불필요한 관용구는 어휘로만 이루어진 관용구를 말한다.

(10) [원문] I think of you every now and then.
[자동번역] 저는 모든 지금 그리고 그러면 당신을 생각합니다.
[수동번역] 가끔 네 생각을 한다.
[어휘번역패턴] {every now and then} <-> {가끔}

5.1.2. 가변 요소가 필요한 관용구

가변 요소가 필요한 관용구는 문법적으로 가변 요소가 필수적으로 들어가야 하는 관용구를 말하며 가변 요소에 의해 관용구가 분리되는 지의 여부에 따라 다시 두 가지 형태로 세분할 수 있다.

(11) a. 가변 요소에 의해 분리되지 않는 관용구
[원문] Don't make light of this computer.
[자동번역] 이 컴퓨터의 불을 만들지 마시오.
[수동번역] 이 컴퓨터를 얕보지 마라.
[어휘번역패턴] { make! light of NP } <-> {NP:을 얕보!다 }
b. 가변 요소에 의해 분리되는 관용구
[원문] I put his bad manners down to his ignorance.
[자동번역] 나는 그의 무지 아래로 그의 나쁜 태도를 둡니다.
[수동번역] 나는 그의 나쁜 태도를 그의 무지 탓으로 여긴다.
[어휘번역패턴] {put! NP1 down to NP2} <-> {NP1:을 NP2:의 탓으로 여기!다}

예문 (11a)의 어휘 번역 패턴은 'make light of'가 명사구 NP라는 가변 요소 1개를 필요로 하는 관용구라는 것을 보인다. 예문 (11b)에서는 'put ... down

www.kci.go.kr

to ...'라는 관용구가 두 개의 명사구 가변 요소인 NP1과 NP2를 필요로 한다는 것을 의미한다.

5.2. 무생물 주어 구문

5.2.1. 무생물 주어와 타동사 구문

영어를 한국어로 번역할 때 가장 흔하게 접할 수 있는 자연스럽지 못한 번역 예는 영어의 무생물 주어와 타동사 구문의 번역일 것이다. 문서의 종류에 따라 영어의 무생물 주어와 타동사 구문을 영어 구조 그대로 한국어로 번역하는 것이 자연스러운 경우도 있겠으나(장영준 2009: 128) 대부분은 영어의 무생물 주어를 한국어의 주어로 그대로 직역하면 부자연스런 경우가 더 많다. 영어에서는 무생물이 주어가 되는 경우가 흔히 있지만 한국어에서는 무생물 주어가 행위자를 나타내는 표현은 거의 없기 때문이다. 그렇다고 영어의 무생물 주어가 한국어에서 모두 주어가 아닌 다른 문장 성분으로 옮겨지는 것은 아니다. 영어의 무생물 주어가 한국어의 다른 문장 성분으로 바뀌는 경우는 타동사의 특별한 어휘와 결합하였을 때만 가능한 것이다. 어휘 번역 패턴에서 무생물 주어와 타동사 구문의 예를 보이면 다음과 같다.

(12) [원문] That book contains many pictures.
[자동번역] 그 책은 많은 그림을 포함한다.
[수동번역] 그 책에는 많은 그림이 있다.
[어휘번역패턴] {NP1:무생물 contain! NP2} <-> {NP1:에는 NP2:이 있!다}

(13) [원문] The report says that more than 180,000 women are incarcerated.
[자동번역] 보고서는 18만 명 이상의 여성이 수감되어 있다고 말한다.
[수동번역] 보고서에 따르면 18만 명 이상의 여성이 수감되어 있다고 한다.
[어휘번역패턴] {NP1:무생물 say! S} <-> {NP1:에 따르면 S:다고 하!다}

www.kci.go.kr

(14) [원문] Learning a foreign language will help you find a solution.
[자동번역] 외국어를 배우는 것은 당신이 해법을 발견하도록 도울 것이다.
[수동번역] 외국어를 배움으로써 당신은 해법을 발견할 수 있을 것이다
[어휘번역패턴] {VP help! S} <-> {VP:ㅁ으로써 S:ㄹ 수 있!다}

예 (12)는 무생물 주어와 타동사 'contain'의 조합에 대한 자연스런 어휘 번역 패턴에 관한 것이다. (12)에 따르면 영어 무생물 주어 'NP1'이 타동사 'contain'과 함께 쓰이면 무생물 주어는 한국어 번역에서 장소나 공간을 나타내는 'NP1:에는'으로 번역되고 목적어 'NP2'가 주어가 되며 contain은 '있!다'로 번역되는 것을 알 수 있다. 예 (13)는 무생물 주어와 타동사 'say'에 대한 것으로 예 (12)에서는 목적어가 명사구인 반면, 예 (13)에서는 목적어가 문장인 경우이다. 따라서 '무생물 주어 + say + 문장'의 조합은 한국어로는 '무생물 주어:에 따르면 문장:다고 하!다'와 같이 번역되는 것이 자연스러운 한국어 번역이라는 것을 의미한다. 예 (14)은 동사구인 VP가 주어이고 동사는 'help'이고 목적어가 문장일 때 주어인 VP는 한국어에서는 'VP:ㅁ으로써'와 같이 서술적으로 기술되고 help는 'ㄹ 수 있!다'와 같이 문장의 뒤에서 양상을 나타내도록 하는 것이 자연스럽다는 것이다.

5.2.2. 무생물 주어와 사동 구문

영어에서는 'make, have, get, let' 등과 같은 사역 동사가 발달되어 있다. 영어에서 사역 동사의 주어가 무생물 주어일 때 영어의 무생물 주어는 한국어에서 이유의 부사격 조사나 동사 어미를 사용하여 표현하는 것이 훨씬 자연스럽다(이영옥, 2001: 67).

(15) [원문] The humiliation made me shudder.
[자동번역] 수치심이 나를 부르르 떨게 하였다.
[수동번역] 나는 수치심으로 부르르 떨었다.
[어휘번역패턴] {NP1:무생물 make NP2:사람 shudder} <-> {NP2:는 NP1:으로 부르르 떨!다}

www.kci.go.kr

예 (15)에서 무생물 주어 ‘NP1:무생물’와 목적어 ‘NP2:사람’이 사역 동사 ‘make ... shudder’와 함께 사용되면 한국어 번역에서는 목적어인 ‘NP2:사람’이 주어가 되고 주어인 ‘NP1:무생물’이 부사격 조사가 붙은 ‘NP1:으로’로 되며 ‘make ... shudder’는 ‘부르르 떨!다’로 기술되어야 자연스런 번역이 된다.

5.2.3. 무생물 주어와 감정 동사 구문

이영옥(2001: 57)에 따르면 “영어에서는 감정을 나타내는 동사들이 대부분 목적어를 동반하는 타동사를 기본형으로 가지는 데 비하여 한국어에서는 목적어가 표면에 나타나지 않는 자동사 또는 형용사 구문의 형태로 나타난다.”는 것이다. 따라서 이영옥(2001: 60)은 “감정을 표현하는 타동사 구문에서 감정을 유발하는 주체를 주어 자리에 확연하게 표현하는 영어 구문들을 자연스러운 한국어로 표현하기 위해서는 무생물을 주어로 하지 않는 대신 주요 논항으로 등장한 그 감정의 원인을 어떤 식으로든 나타내 주어야 하...”와 같은 결론에 도달하고 있다. 이러한 결론은 본 논문에서 제시하는 어휘 번역 패턴의 조건 및 실행 형태를 모두 충족하는 것이기도 하다. 다음은 무생물 주어와 감정 동사 구문의 자연스런 번역의 예이다.

(16) [원문] Did the noise frighten you?
[자동번역] 그 소음이 너를 놀라게 했니?
[수동번역] 그 소리 때문에 놀랐니?
[어휘번역패턴] {NP1:무생물 frighten! NP2:사람} <-> {NP1:때문에 NP2:가 놀라!다}

예 (16)의 어휘 번역 패턴에 따르면 무생물 주어 ‘NP1:무생물’이 감정 동사 frighten과 같이 나타나면 한국어 번역에서는 무생물 주어는 원인을 나타내는 부사구 ‘NP1:때문에’로 되며 목적어 ‘NP2:사람’가 주어가 된다는 것이다.

5.3. 양화사(Quantifier) 구문

양화사란 수량을 나타내는 한정사를 말한다. 영어 양화사의 자연스런 한국

www.kci.go.kr

어 번역은 양화사가 부사나 서술어로 번역되는 것이다.

5.3.1. 양화사 서술 구문

영어 양화사 중에는 한국어로 번역되었을 때 서술어로 번역하는 것이 더욱 자연스러운 양화사들이 있다(양승현 1997). (17)가 양화사의 서술 구문 예이다.

(17) [원문] Many people cling to false beliefs.
[자동번역] 많은 사람들이 잘못된 믿음을 고수한다.
[수동번역] 잘못된 믿음을 고수하는 사람들이 많다.
[어휘번역패턴] {many NP VP} <-> {VP:는 NP:가 많!다}

예 (17)의 수동 번역에 따르면 'many'가 영어에서는 'people'을 수식하는 한정사이지만 한국어 번역에서는 '많다'라는 서술어로 번역되는 것이 자연스럽다는 것을 알 수 있다. 이런 양화사의 서술 구문을 패턴으로 기술하는 방법은 양화사는 어휘로 기술하되 명사구(NP)와 동사구(VP)는 변수로 기술하고 한국어에서는 동사구가 명사구를 수식하고 양화사는 서술어로 만드는 것이다.

5.3.2. 양화사 부사 구문

영어 양화사 중에는 부사구로 번역하는 것이 더욱 자연스러운 양화사들이 있다(양승현 1997). (18)이 양화사의 부사 구문의 번역 예를 보여준다.

(18) [원문] Each of us will start for London soon.
[자동번역] 각각의 우리는 London으로 곧 출발할 것이다.
[수동번역] 우리는 각각 London으로 곧 출발할 것이다.
[어휘번역패턴] {each of NP} <-> {NP 각각}

(18)의 패턴은 'each'라는 양화사가 번역에서는 명사구의 영역 밖으로 나와 '각각'이라는 부사로 번역됨으로써 더욱 자연스런 번역이 된다는 것을 말한다.

www.kci.go.kr

5.4. 수동태 구문

한국어에는 수동을 표현하는 문장이 잘 발달되어 있지 않으며, 수동태의 모양도 통일되어 있지 않은 반면에 영어에는 수동태가 잘 발달되어 있다. 영어에서 수동태 구문을 사용하는 목적은 주장하는 바를 일반화하기 위해서인데, 즉 누가 행위를 수행했는가를 보이기보다는 행위나 행위의 결과를 일반화하려 하는 것이다. 일반적으로 수동태 구문의 사용 빈도는 대화, 문학 작품, 신문, 학술 논문의 순으로 높아지며, 특히 신문과 학술 논문은 대화와 문학 작품에 비해서 수동태가 훨씬 많이 사용된다(조인정 2005: 121).

영한 번역에서 영어 수동태 구문을 한국어 수동태 구문으로 그대로 번역하면 부자연스런 경우가 종종 있다. 따라서 영어의 수동태 구문을 가능한 한 한국어에서는 능동태로 바꿔주는 것이 필요하며 영어 수동태의 뉘앙스를 굳이 전달하고자 할 때는 행위의 결과가 주어로 나오도록 자동사를 활용하여 번역하는 것이 바람직하다.

(19) a. 수동태 구문의 타동사화
[원문] The solution is heated gently until the sugar dissolves.
[자동번역] 설탕이 녹을 때까지 용액은 약하게 가열된다.
[수동번역] 설탕이 녹을 때까지 용액을 약하게 가열하다.
[어휘번역패턴] {NP1 be! heated} <-> {NP1:을 가열하!다}
b. 수동태 구문의 자동사화
[원문] He was killed in the war.
[자동번역] 그는 전쟁에서 죽여졌다.
[수동번역] 그는 전쟁에서 죽었다.
[어휘번역패턴] {be! killed} <-> {죽!다}

5.5. 동사파생 명사 구문

동사파생 명사란 동사에서 파생된 명사를 말한다. 동사파생 명사구문의 특징은 그 명사가 파생되기 전의 동사가 가지는 하위범주화의 논항 구조를 명사구문에 포함하고 있다는 것이다. 동사파생 명사 구문은 영한 번역과 관련해서는 절로 풀어서 번역하는 것이 자연스럽다.

www.kci.go.kr

(20) a. [원문] They have an intuitive grasp of the basic elements of strategy.
[자동번역] 그들은 전략의 기본 요소의 직관적인 이해를 가지고 있다.
[수동번역] 그들은 전략의 기본 요소를 직관적으로 이해한다.
[어휘번역패턴] { have! DET ADJ grasp of NP } <-> {NP:을 ADJ:으로 이해하!다 }
b. [원문] Since their repositioning as independent computers, sales have begun to tick upward.
[자동번역] 독립적인 컴퓨터로서의 그들의 재배치 이후로, 판매가 증가했다.
[수동번역] 독립적인 컴퓨터로 다시 자리매김을 한 후 판매가 증가했다.
[어휘번역패턴] { DET repositionining! as NP } <-> {NP:으로 다시 자리매김을 하!다 }

6. 어휘 번역 패턴의 효과 실험

어휘 번역 패턴은 웹신문으로부터 특허, 과학기술 논문, 기업문서, 대화체 문장 등 다양한 영어 문서들로부터 구축되었다. 각 문서로부터 약 10만 문장 정도를 각각 임의로 뽑아 4장에서 소개한 방법에 따라 어휘 번역 패턴을 구축하였다. 현재 구축된 어휘 번역 패턴을 통사적 분류 방법에 따라 구축된 현황을 기술하면 다음과 같다.

〈표 1〉 구축된 어휘 번역 패턴의 통사적 현황

통사적 분류	패턴수
구문 어휘 번역 패턴	16,863
동사구 어휘 번역 패턴	60,366
문장 어휘 번역 패턴	18,077
계	95,306

표 1에 따르면 동사구와 관련된 어휘 번역 패턴이 가장 많으며 그 뒤를 문

www.kci.go.kr

장 패턴이 잇고 있다. 동사구 패턴이 가장 많은 이유는 영어 phrasal verb 숙어가 대량으로 구축되었기 때문이다.

전체 95,306개의 어휘 번역 패턴의 의미적 영향을 확인하기에는 인적, 시간적 비용 문제가 있어 임의의 100문장을 선정하여 어휘 번역 패턴이 어떻게 의미적으로 적용되었는 지 분석하였다. 어휘 번역 패턴의 의미 분류를 알아보기 위한 평가문 추출 방법과 평가 방법은 다음과 같았다.

- 평가문 추출 방법
  영어 대화체 문장, IT 웹신문 문장, CNN 뉴스 문장, 일반 문장으로 각각 25문장씩을 추출하여 100문장을 만든다.
- 평가 방법
  1) 번역가가 100문장의 영어 원문을 한국어로 번역한다.
  2) 패턴 개발자가 번역가 번역문과 어휘 번역 패턴이 적용되기 전의 자동 번역 결과를 비교하면서 적용 가능하다고 여겨지는 어휘 번역 패턴을 수작업으로 설정한다.
  3) 패턴 개발자가 어휘 번역 패턴이 적용된 후의 자동 번역 결과에 적용된 어휘 번역 패턴을 확인한다.
  4) 2)번과 3)번의 내용을 비교 평가한다.

<표 2>가 100문장에 대해 어휘 번역 패턴이 적용된 결과를 보여준다. 이 표에서 괄호 안의 숫자는 자동으로 적용된 어휘 번역 패턴의 개수를 말하며 괄호 밖의 숫자는 어휘 번역 패턴 개발자가 적용 가능하다고 판단한 어휘 번역 패턴의 개수를 말한다.

〈표 2〉 어휘 번역 패턴의 적용 통계

어휘번역패턴	문장수	관용구 패턴	무생물주어 패턴	양화사 패턴	수동태 패턴	동사파생 명사패턴	계
대화체문장	25	13(7)	4(3)	4(4)	0(0)	0(0)	21(14)
IT웹신문	25	14(12)	2(2)	0(0)	1(1)	1(1)	18(16)
CNN뉴스	25	20(20)	1(1)	3(3)	0(0)	0(0)	24(24)
일반문장	25	15(15)	2(2)	2(2)	3(3)	0(0)	22(22)
계	100	62(54)	9(8)	9(9)	4(4)	1(1)	86(76)

www.kci.go.kr

<표 2>에 따르면 패턴 개발자가 설정한 어휘 번역 패턴의 총수는 86개였으며, 자동으로 적용된 어휘 번역 패턴의 수는 총 76개였다. 따라서 어휘 번역 패턴의 커버리지는 88.37%에 달하지만 11.63%에 달하는 10개의 어휘 번역 패턴이 아직도 부족하며 계속 구축해야 한다는 것을 알 수 있다. 적용된 어휘 번역 패턴 중 관용구가 54개로 가장 많이 적용되었지만 또한 앞으로 관용구 패턴이 지속적으로 구축되어야 함을 알 수 있다. 관용구 패턴에 비하면 다른 어휘 번역 패턴은 적게 적용되었으나 커버리지가 높은 것으로 나왔다. (21)은 <표 2>에서 아직 구축되지 않은 패턴의 예를 보여준다.

(21) a. [원문] My parents used to take us to different cities in Korea when we were younger.
[자동번역] 제 부모님은 우리가 더 젊었을 때 한국에서 다른 도시에 우리를 잡곤 했다.
[수동번역] 어렸을 때 부모님께서 저희들을 데리고 국내 여러 도시를 여행하셨습니다.
[미구축 관용구 패턴] { take! NP1:[인간] to NP2:[장소] } <-> {NP1:을 NP2:으로 데리고 가!다 }
b. [원문] A vase will do much for my dry room.
[자동번역] 꽃병이 제 마른 방을 위해 많은 것을 할 것이다.
[수동번역] 꽃병이 건조한 방에 크게 도움이 될 것 같다.
[미구축 무생물 주어 패턴] { NP1:[무생물] do! much for NP2 } <-> {NP1:때문에 NP2:가 크게 도움이 되!다 }

<표 3>은 어휘 번역 패턴이 적용은 되었으나 자동 번역 결과가 나빠진 경우를 보여준다. 괄호 안의 숫자가 자동으로 적용된 어휘 번역 패턴의 개수를 말하며 '-' 뒤의 숫자가 어휘 번역 패턴이 과적용되어 개악된 개수를 나타낸다.

〈표 3〉 어휘 번역 패턴이 과적용되어 개악된 수

어휘번역패턴	문장수	관용구 패턴	무생물주어 패턴	양화사 패턴	수동태 패턴	동사파생명사 패턴	계
대화체문장	25	(7)-1	(3)-0	(4)-1	(0)0	(0)0	(14)-2
IT웹신문	25	(12)-0	(2)-0	(0)-0	(1)0	(1)0	(16)-0

www.kci.go.kr

CNN뉴스	25	(20)-1	(1)-1	(3)-3	(0)0	(0)0	(24)-5
일반문장	25	(15)-0	(2)-0	(2)-1	(3)0	(0)0	(22)-1
계	100	(54)-2	(8)-1	(9)-5	(4)0	(1)0	(76)-8

<표 3>에 따르면 적용된 76개의 어휘 번역 패턴 중에 8개가 과적용되어 개악된 경우임을 보여준다. 특히 양화사 어휘 번역 패턴은 개악되는 경우가 많았으므로 향후 개선할 필요가 있음을 보여준다. 어휘 번역 패턴이 적용되었을 때 개선된 경우와 개악된 경우는 각각 아래의 (22)와 (23)과 같았다.

(22) a. 무생물 주어와 양화사 어휘번역패턴이 적용되어 개선된 예
[원문] Currently the Cite Multimedia counts about one hundred businesses that employ more than 5,500 employees.
[적용전] 현재 Cite Multimedia는 5,500명 이상의 직원을 고용하는 약 100 사업을 계산한다.
[적용후] 현재 Cite Multimedia에는 5,500명 이상의 직원을 고용하는 사업이 약 100개 있다.
[어휘번역패턴] {NP1:무생물 count! NP2:무생물} <-> {NP1:에는 NP2:이 있!다},{about NUM NOUN! } <-> {NOUN! 약 NUM 개}

b. 수동태 어휘번역패턴이 적용되어 개선된 예
[원문] A consortium was formed by the SDM.
[적용전] 컨소시엄은 SDM에 의해 형성되었다.
[적용후] SDM이 컨소시엄을 형성했다.
[어휘번역패턴] {NP1 be! formed by NP2 } <-> {NP2:이 NP1:을 형성하!다}

(23) a. 양화사 어휘번역패턴이 적용되어 개악된 예
[원문] I've visited many Asian countries and Europe too.
[적용전] 나는 또한 많은 아시아 국가와 유럽을 방문했다.
[적용후] 나는 또한 아시아 국가를 많이 와 유럽을 방문했다.
[어휘번역패턴] {many NOUN! } <-> {NOUN! 많이}

b. 무생물 주어 어휘번역패턴이 적용되어 개악된 예
[원문] The need for action will create the opportunity for major accomplishment.

www.kci.go.kr

[적용전] 활동에 대한 필요성이 주요 성과를 위한 기회를 창출할 것이다.
[적용후] 활동에 대한 필요성으로 주요 성과를 위한 기회가 창출될 것이다.
[어휘번역패턴] {NP1:[무생물] create! NP2 } <-> {NP1:으로 NP2:이 창출되!다}

7. 결론

본 논문에서는 부자연스런 자동 번역 결과를 자연스런 자동 번역 결과로 바꾸는 어휘 번역 패턴을 반자동으로 구축하고 의미적으로 분류하는 방법을 기술하는 것을 목표로 삼았었다. 어휘 번역 패턴은 궁극적으로는 자연스런 번역을 위한 것이지만 또한 높은 재사용성, 편리한 유지 및 관리, 어휘 단계에서의 기술 가능, 원문과 번역문의 동시성, 자동 번역 속도의 향상이라는 장점을 가지고 있다.

어휘 번역 패턴을 반자동으로 구축하는 것은 두 가지 방법에 의해 구축될 수 있음을 보였다. 첫 번째 방법은 대역 말뭉치로부터 원문의 관용구 후보를 자동으로 추출한 후 어휘 번역 패턴을 수작업으로 구축하는 것이며, 두 번째 방법은 번역 말뭉치와 자동 번역 결과를 비교하면서 어휘 번역 패턴을 수작업으로 구축하는 것이다.

어휘 번역 패턴은 의미적으로 크게 두 부류로 분류할 수 있었다. 한 부류는 첫 번째 반자동 구축 방법에 의해 얻어진 관용구 어휘 번역 패턴이었으며 다른 부류는 두 번째 반자동 구축 방법에 의해 얻어진 무생물 주어 구문을 위한 어휘 번역 패턴, 양화사 구문을 위한 어휘 번역 패턴, 수동태 구문을 위한 어휘 번역 패턴, 동사파생 명사 구문을 위한 어휘 번역 패턴이었다.

향후에 본 논문은 다음과 같은 계획을 가지고 있다: 1) 번역가에 의해 번역된 대량의 번역 말뭉치를 대상으로 통계적 정렬 기법을 사용하여 더욱 다양한 어휘 번역 패턴을 자동으로 추출하는 것과 2) 어휘 번역 패턴에 사용된 동사들과 유사한 동사를 가지고 어휘 번역 패턴을 반자동으로 확장하는 것이다.

www.kci.go.kr

참고문헌

양승현 (1997)「영한 기계번역을 위한 언어 스타일의 변환」.『박사학위논문』. 서울대학교 컴퓨터공학과.

이영옥 (2001)「무생물 주어 타동사구문의 영한번역」.『번역학연구』제2권 1호. 53-76.

장영준 (2009)「영.한 번역에서의 주어선택과 행위자성」.『번역학연구』제10권 2호. 105-132.

조인정 (2005)「영한 번역의 문제점: 수동태를 중심으로」.『번역학연구』제6권 1호. 121-142.

최승권 (2002)「번역패턴 기반 한독 자동번역」.『독어학』제6집. 1-18.

최승권 (2007)「영어 특허문서 자동번역을 위한 특허번역패턴 연구」.『번역학연구』Vol.8 No.1. 301-322.

Choi, Sung-Kwon, Lee, Ki-Young, Roh, Yoon-Hyung, Kwon, Oh-Woog and Kim, Young-Gil (2008) "How to Overcome the Domain Barriers in Pattern-Based Machine Translation System". *Proceedings of the 22nd Pacific Asia Conference on Language, Information, and Computation.* 161-168.

Dorr, Bonnie (1994) "Machine Translation Divergences: A Formal Description and Proposed Solution". *Computational Linguistics, 20:4.* 597-633.

Hearne, Mary, Tinsley, John, Zhechev, Ventsislav and Way, Andy (2007) "Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner". *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*. 85-94.

Hutchins W. John (1992) *An Introduction to Machine Translation*. Academic Press.

Ohara, Makoto, Matsubara, Shigeki, and Inagaki, Yasuyoshi (2003) "Automatic extraction of translation patterns from bilingual legal corpus". *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*. 150-157.

www.kci.go.kr

Shieber, Stuart.M. and Schabes, Yves (1990) “Synchronous Tree-Adjoining Grammars”. *Proc. of the 13th International Conference on Computational Linugistics*. 253-258.

Takeda, K(1996), Pattern-Based Machine Translation. *16th International Conference on Computational Linguistics*, 1155-1158.

Watanabe, Hideo and Takeda, Koichi (1998) “A Pattern-Based Machine Translation System Extended by Example-Based Processing”, *COLING 1998*. 1369-1373.

Yamada, Setsuo, Imamura, Kenji, and Yamamoto, Kazuhide (2002) “Corpus-Assisted Expansion of Manual MT Knowledge. *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation*.

www.kci.go.kr

[Abstract]

Semantic Classification of Lexical Translation Patterns Extracted from Bilingual Corpus and Application of Lexical Translation Patterns to MT System

Choi, Sung-Kwon and Kim, Young-Kil
(Electronics and Telecommunications Research Institute)

We could make the results translated by a machine translation system natural if we make an implicit translation competence of professional translator explicit and apply it to the machine translation system. The lexical translation pattern can be taken as the explicit translation competence of professional translator.

The purpose of this paper is to describe semi-automatic construction and semantic classification of the lexical translation patterns that make the machine translation results natural.

The lexical translation patterns can be built semi-automatically by two ways. One is to extract automatically the idiomatic expressions of source language from bilingual corpus and construct the lexical translation pattern manually. The other way is to construct manually the lexical translation pattern by comparing the translation result of professional translator with the result translated by machine translation system.

We are able to classify the lexical translation patterns into two semantic groups. One group is related to the idiomatic lexical translation patterns. The other group contains the lexical translation patterns with non-animate subject phrase, quantifier phrase, passive phrase, and verb-derived noun phrase.

www.kci.go.kr

▸ Key Words: Translation Pattern, Lexical Translation Pattern, Machine Translation, English-Korean Machine Translation, Automatic Translation

최승권
한국전자통신연구원 책임연구원
choisk@etri.re.kr
관심분야: 자동번역, 언어처리

김영길
한국전자통신연구원 책임연구원
kimyk@etri.re.kr
관심분야: 자동번역, 언어처리

논문투고일: 2010년 7월 19일
심사완료일: 2010년 8월 27일
게재확정일: 2010년 9월 14일

www.kci.go.kr