Journal of Korea Multimedia Society Vol. 7, No. 7, July 2004(pp. 998-1008)

인접 단어들의 접속정보를 이용한 일한 기계번역 시스템

김 정 인†

요 약

일본어와 한국어는 문법적으로 많은 유사점을 가지고 있다. 이러한 유사점을 잘 이용한다면 일한 기계번역 시스템에서 구문해석이나 의미해석의 상당한 부분을 생략할 수 있다. 몇 년 전부터 우리는 유사성을 이용하여 번역율을 높이는 방법으로 번역테이블을 이용한 일한기계번역 시스템을 연구해 왔다. 그러나 이 시스템은 활용어미의 번역, 다의성 단어의 처리 등 몇 가지 문제점을 가지고 있었다. 본 논문에서는 번역테이블을 이용하는 시스템을 개선하여 이웃 하는 단어들과의 관계 정보를 이용한 일한 기계번역 시스템을 제안한다. 현재 시스템의 문제점들을 해결하기 위하여 우선 조사, 조동사의 접속 정보를 최대한 이용한다. 또한, 번역테이블을 엔트리테이블과 접속정보 테이블로 나누어 설계하여 번역의 효율을 높인다. 즉, 하나의 역어만 가지는 단어인 경우, 우리는 일한 직접 대응 방법을 이용하여 바로 번역하고 2개 이상의 역어로 번역되어야 할 경우만 접속 정보 값을 평가하여 가장 가능성이 높은 번역어를 선택하도록 한다.

Japanese-Korean Machine Translation System Using Connection Forms of Neighboring Words

Jung-In, Kim†

ABSTRACT

There are many syntactic similarities between Japanese and Korean languages. Using these similarities, we can make out the Japanese-Korean translation system without most of syntactic analysis and semantic analysis. To improve the translation rates greatly, we have been developing the Japanese-Korean translation system using these similarities from several years ago. However, the system remains some problems such as a translation of inflected words, processing of multi-translatable words and so on. In this paper, we suggest the new method of Japanese-Korean translation by using relations of two neighboring words. To solve the problems, we investigated the connection rules of auxiliary verbs priority. And we design the translation table which is consists of entry tables and connection forms tables. A case of only one translation word, we can translate a Korean to Japanese by direct matching method use of only entry table, otherwise we have to evaluate the connection value by connection forms tables and then we can select the best translation word.

Key words: Machine Translation(기계번역), Inflected Words(용언), Multi-Translatable Words(다역어), Connection Forms(접속 정보), Neighboring Words(인접 단어)

※ 교신저자(Corresponding Author) : 김정인, 주소 : 부산광역시 남구 용당동 535(608-711), 전화 : 051)610-8393, FAX : 051)610-8393, E-mail : jikim@tmic.tit.ac.kr
접수일 : 2004년 3월 24일, 완료일 : 2004년 5월 25일
† 종신회원, 동명정보대학교 공과대학 컴퓨터공학과 조교수.
※ 이 논문은 2002학년도 동명정보대학교 학술연구비 지원에 의하여 이루어진 것임.

1. 서 론

일한기계번역시스템은 두 언어의 유사성에 의존하여 형태소 해석 결과로부터 얻은 번역대상 단어로부터 직접 대응에 의해 역어를 생성시키는 가장 단순한 번역처리를 가능하게 한다. 또한 양국어는 말과

말의 접속관계를 표시하기 위하여 조사를 이용하므로 격조사에 의한 격형식 패턴매칭 값을 이용하는 일한기계번역이 많이 연구되어지고 있다[8,9,15]. 그러나, 유사성을 이용하기 때문에 구문해석과 의미해석의 상당 부분을 생략한 결과, 활용어의 번역처리와 다의성 번역처리 등에서는 번역을 위한 정보가 부족한 단점을 가지고 있다. 활용어의 처리, 다의어의 처리를 위하여 유용한 번역 정보를 추출하고 적절하게 이용하는 것이 가능하다면 단순하면서도 고품질의 번역어가 출력되는 일한 기계번역 시스템의 구축이 가능하며 거기에 대하여 몇 가지 연구결과가 발표되어 있다[6,17,18,21-24].

다의성 처리로는 말과 말의 관계를 이용한 의미해석에 의한 단문 다의성 처리[8], 체언의 의미소성을 나열한 격형식 패턴을 이용하는 동사의 다의성 처리 등이 제안되었다[5,13,15]. 그 외에도 양국어의 유사성을 이용하는 연구의 일종으로 일본어 분류 어휘표로부터 한국어 분류 어휘표를 간단히 작성하는 방법이 제안되었다[19].

활용어 처리에 대하여서는양국어 술부표현의 의미대응에 의한 활용어 처리[9], 한일 기계번역에적용한 음운 표현형식에 의한 술부의 활용처리[14] 등이 제안되었지만 한국어 용언의 불규칙적인 활용을 모두 표현할 수 없었다. 용언의 불규칙적인 활용을 분산 처리하기 위한 기법으로, 용언별로 의미접속관계에 따라 역어를 미리 테이블에 준비하는 번역테이블 방식이 제안되었다[10-14,16].

또한 번역테이블 방식의 확장성 결여, 접속관계의 표현부족, 번역시스템의 일관성 결여 등의 단점을 보완한 확장 번역테이블을 이용한 일한기계번역 방식이 제안되었다[21].

본 연구는 확장 번역테이블을 이용한 일한기계번역 방식에서 인접2단어와의 접속관계를 이용한 활용어 번역처리에 관하여 기술한 것이다. 2장에서 종래의 번역테이블방식과 확장 번역테이블 방식의 차이점을 살펴보며, 확장 번역테이블 방식에서 인접2단어를 표현하기위한 접속요소 들을 정리한다. 3장에서 인접2단어와의 접속관계를 평가하는 방법을 제안하고 4장에서 번역테이블 방식과 확장 번역테이블 방식의 번역 결과를 비교 평가하며 잘못 번역된 경우를 분석한다. 5장에서 향후 개선되어야 할 점 등을 논하며 결론을 내린다.

2. 확장 번역 테이블 방식을 이용한 일한기계번역

자연언어를 번역하는 시스템들은 인간이 사용하는 수많은 규칙을 적용해야 한다. 번역 처리시스템은 처리의 흐름, 규칙 등을 적은 번역시스템과 단어의 속성, 역어 등을 기술한 번역사전으로 나누어진다. 두 부분은 밀접한 관계를 가지고 있으며 번역시스템 속에 규칙을 상세히 설정하는 것에 의해 번역사전의 용량을 적게 만드는 것이 가능하며, 반대로 번역사전에 필요한 정보를 많이 기술하는 것에 의해 번역 시스템을 단순한 형태로 구축하는 것이 가능하다. 실제로 자연언어처리에는 여러 종류의 애매성이 존재하며 그것들을 해소하기 위하여 대량의 규칙을 사용하는데 그런 규칙들을 번역시스템에 모두 기술하는 것은 불가능에 가까우며, 번역사전의 데이터 형식으로 기술하는 방법이 여러 측면에서 보다 효과적이다[1-4].

2.1 종래의 번역테이블 방식

번역테이블을 이용한 일한 번역 방식은 품사별로 공통된 테이블 형식을 준비한다. 예를 들어 동사의 경우는 후접 단어의 의미 및 문법 정보에 따라 17가지로 나누어 각각에 맞는 대역어를 준비한다. 표1(a)는 일본어 동사 "行く"의 번역테이블이다. 일본어의 모든 동사에 아래 형식을 적용한 번역테이블을 미리 준비하고 번역 대상 문장에서 활용어 뒤에 어떤 의미가 접속되는지에 따라 대역어를 선택할 수 있도록 하였다. 형용사, 형용동사는 후접 단어와의 관계를 12가지로 각각 분류하여 테이블화 하였으며 조사, 조동사는 통일된 테이블 형식 없이 단어 하나 하나에 대하여 접속 정보를 따로 정의하고 거기에 맞는 대역어를 준비하였다. 표 1(b)는 형용사 "美しい", 표 1(c)는 형용동사 "重要だ"의 번역테이블이다.

번역테이블 방식은 품사에 의존하며 동사의 경우, 17가지 이상의 접속 관계를 표현하기 위해서는 테이블의 구조를 변경시켜야 하는데 이 작업은 모든 동사들에 영향을 미치게 되어 변경이 쉽지 않다. 또한, 일본어 동사 "行く"의 경우, 17가지 중에서 "가"로 번역되는 경우가 8가지로 대표역어의 성질을 가지지만 번역테이블 방식에서는 이들을 통합하는 작업이 불가능하다. 그림 1에 종래의 번역 테이블 방식의 개

표 1. 번역테이블의 예

(a) 동사 "行く"의 번역테이블

활용형	활용단어	접속정보에 따른 한국어 대응		
		No	후접단어의 의미 및 문법	대역어
미연형	行か	1	부정	가
	行こ	2	부정 이외	가
연용형	行き	3	정중, 희망	가
		4	정중, 과거	갔
		5	정중, 부정	가
		6	상태	갈
		7	용언	가기
		8	연용중지형	가서
		9	전성명사	행
	行っ	10	과거	갔
종지형	行く	11	추정, 전문	갈
		12	정중	가
		13	접속조사	가
		14	종료	간다
연체형	行く	15	NULL	갈
가정형	行け	16	NULL	가
명령형	行け	17	NULL	가

(b) 형용사 "美しい"의 번역테이블

활용형	활용단어	접속정보에 따른 한국어 대응		
		No	후접단어의 의미 및 문법	대역어
미연형	美しかろう	1	추량	아름답겠
연용형	美しかっ 美しく	2	부정	아름답
		3	연용중지형	아름답고
		4	과거	아름다웠
종지형	美しい	5	추정	아름다운
		6	정중	아름답
		7	접속조사	아름답
		8	종료	아름답다
		9	전문	아름답다
연체형	美しい	10	NULL	아름다운
가정형	美しけれ	11	NULL	아름다우
어간	美し	12	상태	아름다운

(c) 형용동사 "重要だ"의 번역테이블

활용형	활용단어	접속정보에 따른 한국어 대응		
		No	후접단어의 의미 및 문법	대역어
미연형	重要だろ	1	추량	중요하겠
연용형	重要で 重要に 重要だっ	2	부정	중요하
		3	연용중지형	중요하고
		4	과거	중요했
종지형	重要だ	5	접속조사	중요하
		6	종료	중요하다
		7	전문	중요하다
연체형	重要な	8	NULL	중요한
가정형	重要なら	9	NULL	중요하
어간	重要	10	추정	중요하
		11	정중	중요
		12	상태	중요한

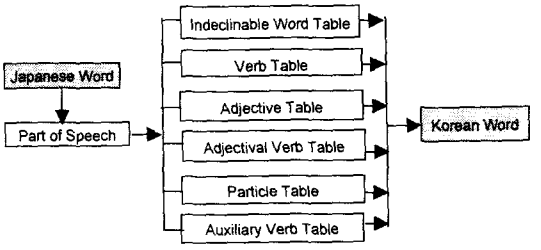

그림 1. 종래의 번역테이블 방식 개념

념을 나타낸다. 우선 형태소 분석의 결과에서 품사를 확인한 후, 품사별 테이블을 이용하여 한국어를 생성하게 된다.

2.2 확장번역테이블 방식

확장번역 테이블방식은 대표역어를 엔트리 테이블에, 후보역어와 접속규칙은 접속정보 테이블 속에 각각 기술한다. 그림 2는 확장번역 테이블 방식의 개념을 나타낸다. 확장번역테이블 방식은 우선 일본어의 번역대상 단어를 엔트리 테이블에 기술되어 있는 대표적인 역어로 번역 시킨다. 복수의 역어로 번역 가능한 경우는 접속정보 테이블에 기술되어 있는 인접단어와의 접속 규칙을 이용하여 적절한 역어를 선택한다.

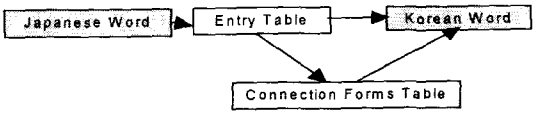

그림 2. 확장번역테이블 방식의 개념

역어 선택에 필요한 접속관계 정보는 접속정보 사전에 기술하며, 번역 대상 단어별 인접 2단어와의 접속 정보를 α′, α, β, β′로 표시한다. 예를 들면 아래 문장에서 번역할 단어가 "行っ"인 경우에 전접단어(α)는 바로 앞에 위치한 "に"가 되며 후접단어(β)는 바로 뒤에 위치하는 "て"가 된다. 또한 전접단어보다 하나 앞에 위치한 단어를 전전접단어(α′)라 부르고 아래의 문장에서는 "学校"가 되며, 반대로 후접단어보다 하나 더 뒤에 위치한 단어를 후후접단어(β′)라 하고 아래의 예에서는 "き"가 된다.

学校/ に / 行っ / て / き /まし/た/。
α′ α W β β′
전전접단어 전접단어 번역대상단어 후접단어 후후접단어

2.3 확장 번역테이블 방식에서의 접속요소

확장 번역테이블 방식에서의 접속관계를 나타내는 접속요소는 품사, 체언의 의미소성, 용언의 활용형, 조사/조동사 및 기호의 일련번호를 사용한다.

- 품사 : 명사, 준체언(の), 동사(상1, 하1, カ행변격, サ행변격, 5단)(가능, 진행, 보조), 형용사, 형용동사, 조동사, 조사(격, 접속, 부, 종), 부사, 준체언, 접속사, 감동사, 기호 등
- 체언의 의미소성 : 구체명사(동물, 인간, 조직 및 기관, 식물, 생물의 부분, 자연물, 생산물 및 도구), 현상명사, 추상명사(동작 및 작용, 정신, 언어작용, 성질, 관계, 공간 및 방향, 시간, 수량), 기타
- 용언의 활용형 : 미연1, 미연2, 미연3, 연용1, 연용2, 연용3, 종지, 연체, 가정, 명령1, 명령2, 어간
- 조사, 조동사 및 기호의 일련번호 : 조사 58개, 조동사 19개, 기호 30개

이들 접속 요소를 이용하여 번역 대상 단어의 앞뒤로 인접하는 단어들과의 접속 규칙을 이용하여 적절한 역어로 번역한다. 일본어 품사는 기본적으로 학교 문법에 나타나는 9품사 분류법을 비롯하여 언어처리에 맞게 분류된 여러 품사 분류법이 존재한다[20]. 본 논문에서는 인접 단어와의 접속 관계를 적절히 표현하기 위하여 품사의 분류와 용언의 활용형 분류를 학교 문법보다 세밀하게 분류하였다. 활용형의 분류는 품사에 따라 형태가 다르기 때문에 표 2처럼 12가지로 분류하였다. 체언의 의미소성 분류에 대해서는 IPAL동사사전의 분류방법을 인용하였다[25]. 그리고 조사, 조동사의 일련번호표를 표 3, 표 4에 나타낸다. 여기서 조사의 일련번호에는 같은 형태의 조사가 복수 등록되어 있는 것처럼 보이지만 실제는 의미가 다른 조사이다. 1번에서 10까지는 격조사, 11번에서 27번까지는 접속조사, 28번에서 45번까지가 부조사, 46번부터 나머지가 종조사이다. 예를 들어 1번의 「が」와 13번의 「が」는 각각 그 문절이 주어인 것을 나타내는 격조사(이/가)와 확정의 역접(이지만)을 나타내는 접속조사로 구별되며 7번의 「から」와 16번의 「から」는 기점/원인/이유의 격조사(로부터)와 확정의 순접(이므로)을 나타내는 접속조사로 구별된다. 조동사의 경우도 14번 「だ」와 18번 「だ」가 같은 형태로 되어 있지만, 전자는 과거(었다)를 후자는 단정(다)을 나타낸다.

표 2. 활용어의 분류

형태	코드	동사					형용사	형용동사
		상1단	하1단	カ행변격	サ행변격	5단		
미연1	1	起き	受け	こ	さ	書か	美しかろ	元気だろ
미연2	2				し	書こ		
미연3	3				せ			
연용1	4	起き	受け	き	し	書き	美しかっ	元気だっ
연용2	5					書い	美しく	元気で
연용3	6							元気に
종지	7	起きる	受ける	くる	する	書く	美しい	元気だ
연체	8	起きる	受ける	くる	する	書く	美しい	元気な
가정	9	起きれ	受けれ	くれ	すれ	書け	美しけれ	元気なら
명령1	A	起きろ	受けろ	こい	しろ	書け		
명령2	B	起きよ	受けよ		せよ			
어간	C	起	受			書	美し	元気

표 3. 조사의 일련번호

번호	조사	번호	조사	번호	조사	번호	조사
1	が	16	から	31	さえ	46	か
2	の	17	し	32	でも	47	な
3	を	18	ても	33	しか	48	な
4	に	19	でも	34	まで	49	なあ
5	へ	20	けれど	35	ばかり	50	や
6	と	21	けれども	36	だけ	51	ぞ
7	から	22	て	37	ほど	52	とも
8	より	23	で	38	くらい	53	よ
9	で	24	ながら	39	ぐらい	54	のわ
10	や	25	たり	40	など	55	ね
11	ば	26	だり	41	きり	56	ねえ
12	と	27	ものの	42	ぎり	57	さ
13	が	28	は	43	なり	58	
14	のに	29	も	44	やら		
15	ので	30	こそ	45	か		

표 4. 조동사의 일련번호

번호	조사	번호	조사	번호	조사	번호	조사
1	せ	6	ぬ	11	たい	16	ようだ
2	させ	7	ん	12	ます	17	らしい
3	れ	8	う	13	た	18	だ
4	られ	9	よう	14	だ	19	です
5	ない	10	まい	15	そうだ		

본 시스템에서는 형태소 해석 결과로부터 단어별로 접속 요소로서 이용하는 4종류, 즉, 품사, 체언의 의미소성, 활용어의 활용형, 조사/조동사/기호의 일련번호가 주어졌다고 전제한다. 4종류의 접속요소는 품사 별로 다른 요소와 결합하며 표 5에 6종류의 결합형태를 나타낸다. Type1은 체언, Type2는 용언, Type3는 조동사, Type4는 조사, Type5는 기호,

표 5. 접속관계 형태

구분	품사 (제1접속요소)	의미소성 혹은 활용형 (제2접속요소)	일련번호 (제3접속요소)
Type1	명사	체언의 의미소성	NULL
Type2	용언	용언의 활용형	NULL
Type3	조동사	조동사의 활용형	일련번호
Type4	조사	NULL	일련번호
Type5	기호	NULL	일련번호
Type6	기타	NULL	NULL

Type6는 부사, 연체사, 접속사, 감동사의 단어 정보를 나타낸다.

여기서 제1접속요소는 품사를 나타내고 제2접속요소는 제1접속요소와 연동하며, 제1접속요소가 명사의 경우는 의미소성, 용언이나 조동사의 경우는 활용형, 그 외에는 NULL을 기술한다. 제3접속요소는 제1접속요소가 조동사, 조사, 기호의 경우는 그 일련번호를, 그 외의 경우는 NULL을 표기한다.

3. 인접2단어와의 접속관계를 이용한 번역

인접2단어와의 접속관계를 이용하여 일한 번역을 행하기 위해서는 인접2단어와의 접속관계를 계량화하여 평가하는 작업이 필요하다.

3.1 인접2단어와의 접속관계 평가방법

인접2단어와의 접속정보를 이용하여 적절한 역어를 선택하기 위해서는 어느 접속 정보가 적절한지 평가하는 평가기준이 필요하다. 접속정보 평가는 비교에 따라 행하여지며, 비교연산자 Δ를 도입한다. 비

교연산자 Δ의 좌변과 우변을 비교하여 같은 경우는 1, 다른 경우는 1을 출력한다. 또한, 좌변 혹은 우변이 NULL인 경우는 비교를 생략하고 0을 돌려준다. 여기서 인접2단어와의 접속정보를 집합 C={α',α,β,β'}로 표기하고 어떤 접속정보 x를 x∈C로 한다. 그러면 비교연산자 Δ의 연산은

$$x_i \Delta x_{di} = \begin{cases} 1(x_i = x_{di}) \\ 0(x_i = \text{NULL or } x_{di} = \text{NULL}) \\ -1(x_i \neq x_{di}) \\ (\text{단, } i=1,2,3) \end{cases}$$

로 표시할 수 있다. 여기서 x_i는 분석대상 단어의 정보이고 x_{di}는 번역사전 속에 준비해준 단어의 정보이다. 또한, 접속요소별 우선순위를 고려하기 위하여 가중치를 부여한다. x_1은 품사이며 비교적 약한 속성이므로 가중치를 1로 한다. x_2는 의미소성 혹은 활용형이며 품사보다는 조금 구체적인 속성이 되므로 가중치 2를 부여한다. x_3은 단어의 일련번호, 즉 단어 그 자체를 나타내므로 가중치를 3으로 한다. 실제의 접속 관계를 평가하기 위해서

$$E_x = \begin{cases} -1(x_i \Delta x_{di} = -1, \text{ at least one I}) \\ \sum_{i=1}^{3}((x_i \Delta x_{di}) * w_i, \text{ otherwise}) \\ (\text{단, } w_1=1,\ w_2=2,\ w_3=3) \end{cases}$$

를 이용한다. 인접2단어와의 평가치는 전전접, 전접, 후접, 후후접 관계를 평가한 4개의 평가치를 합산한다.

$$E_t = \begin{cases} -1(E_x<0, \text{ at least one x}) \\ \sum_{x=C} E_x(\text{otherwise}) \\ (\text{단, } E_t : t\text{번째의 접속관계에 대한 평가치}) \end{cases}$$

평가의 결과, $E_t = -1$의 경우는 접속 불가능이며, $E_t = 0$의 경우는 접속요소가 NULL만으로 비교할 수 없는 것을 의미한다. 그리고 $E_t>0$의 경우는 접속관계의 강도를 나타낸다. 즉, 평가치 E_t가 최대치인 접속관계는

$$t^* = \arg \max_{t \in N} E_t$$

(n : 후보역어의 수, N={1,2,…,n})

이며 t^* 번째의 접속정보와 그 역어가 선택되어진다. 최대치가 같은 값으로 복수 존재할 경우는 사전에 등록된 순서에 의존하며 먼저 등록된 역어가 우선된다. 인접2단어와의 접속정보를 도입한 확장 번역테이블 번역 방식은 다음과 같은 특징을 가진다.

- 기존의 번역테이블 방식에서는 기술할 수 없었던 앞뒤 단어와의 세밀한 접속관계가 부속어(조사, 조동사)를 따로따로 처리 시킴으로써 가능하게 되었다.
- 동사, 형용사, 형용동사 등의 번역 테이블은 접속관계와 역어의 수가 고정되어 있지만 확장번역테이블에서는 접속관계를 자유롭게 기술할 수 있으며, 접속관계 수를 활용단어별로 조절할 수 있다. 따라서 활용 단어별로 독립된 접속관계를 기술할 수 있다.
- 활용 단어의 대표역어를 마련하여 같은 역어의 중복 등록을 제거하였다.
- 확장번역테이블은 엔트리 테이블과 접속정보 테이블이라는 2개의 사전에 모든 품사가 들어가도록 설계되어 사전의 통일성을 꾀하였으며, 따라서 일관성 있는 시스템 구축이 가능하다. 즉, 번역시스템의 구축 및 관리가 용이하다.

번역대상 단어를 둘러싼 인접2단어는 표 6과 같은 9가지 형태로 사용되어진다. 실제 접속관계에서 CF6, CF7, CF8 관계를 이용하는 경우는 잘 나타나지 않았다.

표 6. 인접2단어와의 접속관계 형태

구분	전전접	전접	번역대상	후접	후후접
CF0			*W*		
CF1		α	*W*		
CF2			*W*	β	
CF3		α	*W*	β	
CF4	α'	α	*W*		
CF5			*W*	β	β'
CF6		α	*W*	β	β'
CF7	α'	α	*W*	β	
CF8	α'	α	*W*	β	β'

3.2 인접2단어와의 접속관계를 이용한 번역처리의 예

일본어 문 「学校に行ってきました。」는 표 7과 같이 형태소 해석되어 그 결과로부터 차례로 적절한 역어를 선택해 간다.

우선 「学校」는 1:1 대응에 따라 "학교"로 번역된다. 그러나, 조사 「に」의 경우는 사람이나 동물의 뒤에 사용될 경우 "에게", 용언의 연용1형 뒤에 나타날 경우는 "(으)러", 연체형 뒤에 나타날 경우는 "데" 등

표 7. 「学校に行ってきました。」의 번역정보 일람

단어	구분	제1접속요소(품사)	제2접속요소(의미소성, 활용형)	제3접속요소(일련번호)
学校	Type1	명사	의미소성(장소: LOC)	NULL
に	Type4	조사	NULL	004
行っ	Type2	5단동사	활용형(연용2: 5)	NULL
て	Type4	조사	NULL	022
き	Type2	カ행변격동사	활용형(연용1: 4)	NULL
まし	Type3	조동사	활용형(연용1: 4)	012
た	Type3	조동사	활용형(종지형: 7)	017
。	Type5	기호	NULL	001

으로 번역되어야 한다. 즉, "에, 에게, (으)러, 데" 등의 복수 역어가 존재하기 때문에 전후의 접속된 단어 정보 「学校」, 「行っ」, 「て」등을 이용하여 적절한 역어를 선택한다. 「に」의 전후에 나타난 단어들을 전접, 후접, 후후접으로 구분하여 표기하면 다음과 같다. 여기서 W는 번역 대상단어이다. 「に」의 엔트리 테이블과 접속정보테이블을 표 8에 나타내었다.

学校(α) に(W) 行っ(β) て(β') きました。

「に」의 접속정보 테이블에 인접2단어와의 접속정보에 따른 복수의 역어를 등록해 두었다. 첫번째 접속정보에서는 접속형태가 CF1이어서 번역사전의 α_d와 번역대상 단어보다 앞에 접속한 단어의 정보 α를 비교한다. 예문에서는 「学校」를 전접단어 α 로 하며, 후접단어 「行っ」을 β, 「て」를 후후접 단어 β'로 하기 때문에 사전에 기술되어 있는 접속정보 α_d(α_{d1}: 명사, α_{d2}:HUM αd3:NULL)와 「学校」의 단어정보 α (α_1:명사, α_2:LOC α_3:NULL)와 비교한다. 제1접속요소는 "α_1:명사"와 "α_{d1}:명사"이며 품사가 일치하고 있으므로 평가치 $x_i\Delta x_{di}$는 1점이 된다. 제2접속요소의 비교에서 "α_2:LOC"와 "α_{d2}:HUM"이 일치하지 않으므로 이 접속은 불가능하며 평가치 $x_i\Delta x_{di}$에 1을 대입하고 더 이상의 비교를 생략한다. 「学校」를 전접단어로 하는 조건을 만족시키는 역어를 선택하기 위하여 접속정보테이블의 접속정보와 차례로 비교해 보지만 그 조건을 만족하는 접속정보는 존재하지 않기 때문에 대표적인 역어인 엔트리테이블의 "에"로 번역한다. 다음 번역단어 「行っ」은 인접2단어로

学校(α') に(α) 行っ(W) て(β) き(β') ました。

가 접속되어 있어서 「行っ」의 확장 번역테이블에 기술되어 있는 접속 형태에 맞추어 평가치를 계산한다. 「行っ」의 확장 번역테이블을 표 9에 나타낸다. 「行っ」은 8개의 접속관계에 따라 역어를 가지고 있으므로 최초의 접속정보부터 비교하여 평가치를 구한다. 우선, 최초의 접속정보는 접속형태가 CF2이므로 β와 β_d 만을 매칭시켜 평가치를 구한다. 그 결과 품사가 일치하므로 1점, NULL은 계산하지 않으며 일련번호가 일치하므로 3점을 더하여 평가치 E_1은 4점이 된다. 두 번째 접속정보는 접속형태가 CF5이므로 β와 β_d, β'와 β'_d의 매칭을 이용하여 평가치를 구하게 된다.

$\beta \cdot \beta_d$ 매칭의 평가치는 첫 번째와 같은 4점이 계산되지만 $\beta' \cdot \beta'_d$의 매칭에서 품사가 일치하지 않아서

표 8. 조사 「に」의 엔트리테이블과 접속정보테이블

원형	품사	엔트리	단어의 정보		대표역어	멀티
			활용형	번호		
に	조사	に		004	에	Yes

엔트리	품사	활용형	No	CF	전전접단어(α')			전접단어(α)			후접단어(β)			후후접단어(β')			역어
					품사	활용형	No	품사	활용형	No	품사	활용형	No	품사	활용형	No	
に	조사		004	CF1				명사	HUM								에게
				CF1				명사	ANI								에게
				CF1				용언	연용1								(으)러
				CF1				조사	연체								데

표 9. 「行っ」의 엔트리테이블과 접속정보테이블

원형	품사	엔트리	단어의 정보		대표역어	멀티
			활용형	번호		
行く	동사	行っ	연용2		가	Yes

엔트리	품사	활용형	No	CF	전전접단어(α')			전접단어(α)			후접단어(β)			후후접단어(β')			평가치	역어
					품사	활용형	No	품사	활용형	No	품사	활용형	No	품사	활용형	No		
行っ	동사	연용2		CF2							조사	NULL	022				1+0+3	가서
				CF5							조사	NULL	022	조사	NULL	007	-1	가서
				CF5							조사	NULL	022	동사	NULL	NULL	1+0+3+1	갔
				CF5							조동	NULL	013	조사	NULL	054	-1	갔
				CF5							조동	NULL	013	조사	NULL	013	-1	갔지
				CF5							조동	NULL	013	조사	NULL	015	-1	갔기
				CF5							조동	NULL	013	조사	NULL	016	-1	갔기

(카변형동사와 조사) 강제적으로 매칭의 실패를 나타내는 -1점을 평가치 E_2에 부여하고 비교 연산을 종료시킨다. 3번째 접속정보를 이용한 매칭 처리는 두 번째와 같아서 E_3 역시 -1이 된다. 4번째 접속정보는 $\beta \cdot \beta_d$ 매칭에서 4점, 거기에 $\beta' \cdot \beta'_d$의 매칭으로부터 품사가 같기 때문에 다시 1점을 더하여 평가치 E_4는 5점이 된다. 5번째 접속정보부터 $\beta \cdot \beta_d$의 최초 비교 $x_1 \Delta x_{d1}$에서 품사가 같지 않으므로 평가치 E_5에 -1을 대입하고 매칭을 중단한다. 이하 6번째부터 8번째까지는 5번째와 같은 조건이므로 E_6, E_7, E_8에 각각 -1을 대입하고 비교 처리를 마친다. 마지막으로 지금까지 매칭처리로 구한 E_i에서 최대치를 가지는 평가치를 찾아보면 E_4(5점)이므로 4번째 역어(t^*=4)가 선택되어진다.

같은 방법으로 전후 단어와의 접속관계를 이용하여 번역 대상단어의 적절한 역어를 선택해가면 "학교에 갔다 왔습니다."와 같은 자연스러운 한국어가 생성된다.

4. 번역실험의 결과 및 검토

인접2단어와의 접속관계를 이용한 일한 기계번역 방식의 유효성을 확인하기 위하여 번역시스템을 간단히 구현하였다. 번역대상 문장은 상대적인 평가를 행하기 위하여 이전에 번역테이블 방식에서 사용한 474 예문(11,695 단어)을 그대로 사용하였으며, 내용은 과학기술용의 논문, 기계의 매뉴얼, 아사히 신문 기사의 일부 등 문법적 오류가 비교적 적은 문장들로 구성되어 있다[11,12,16]. 앞으로 시스템을 더욱 확장시켜 대량의 코퍼스를 대상으로 실험을 행할 계획이며, 본 시스템의 실용화를 위해서는 각각의 단어에 대하여 접속관계를 포함한 번역정보를 사전에 기술할 필요가 있지만 대역 코퍼스로부터 자동으로 각 단어의 번역지식을 추출하는 수법 등을 이용하면 번역 지식을 늘리는 것이 가능할 것으로 생각된다.

종래의 번역테이블 방식을 평가하는 동안에 2가지 측도를 이용하였다. 하나는 생성된 한국어의 정확도이며, 다른 하나는 원문의 의미가 어느 정도 전달되었는가를 측정하는 충실도이다. 표 10은 번역결과 테이블이며 이해도를 1,2,3,4로 나누어 횡축에 기입하고 충실도를 A,B,C로 나누어 종축에 표시하였다.

이해도는

1) 한국어 문장으로 옳으며 충분히 이해가 가능하다.

2) 한국어 문장으로 조금 이상하지만 이해가 가능하다.

3) 한국어 문장으로 이상하여 상당한 추측을 해야 이해가 가능하다.

4) 한국어 문장으로 옳지 않으며 이해 불가능이다.

충실도는

A) 원문의 의미가 그대로 번역되어 있다.

B) 원문의 의미가 부족하지만 단어는 맞게 번역되어 있다.

C) 원문의 의미에 맞지 않게 번역되었다.

로 정의하여 실험의 데이터를 얻었다.

표 10(a)는 종래의 번역테이블을 이용한 번역처리의 결과를, 표 10(b)는 확장번역테이블 방식에서의 번역처리 결과를 나타낸다.

여기서 이해도 1,2와 충실도 A,B를 합격으로 하면, 양 방식의 번역률은 각각 83%(394)와 87%(415)이며, 확장 번역 테이블 방식이 약 4%정도 높은 번역률을 나타내고 있다. 그러나, 한국어 이해도의 면에서 1,2, 단계의 수치 차이에 주목해 보면 이해도 1의 경우 50%에서 66%까지올라, 역문의 이해가 되어도 상당한 후 편집이 필요했던 것이 확장번역테이블 방식을 사용함에 따라 약 16% 이상 후 편집 작업 효율이 향상된 것을 알 수 있다.

본 시스템에서 번역 실패의 원인을 살펴보면 다음과 같다.

(1) 체언의 다의성에 의한 번역실패(1,B)

その時　　그 시 (x)→그 때(0)

3時　　3 시 (0)

よんだことがない　　읽은 일이 없다(x) 읽은 적이 없다(0)

なんのことですか　　무슨 일입니까(0)

よむことができない　　읽는 일을 할 수 없다(x)→읽을 수가 없다(0)

표 10. 번역처리의 결과

(a) 번역테이블 방식에 의한 평가

	1	2	3	4	합계
A	183 (38.6)				191 (40.3)
B	53 (11.2)	150 (31.6)	52 (11.0)	4 (0.8)	259 (54.6)
C			18 (3.8)	6 (1.3)	24 (5.1)
합계	236 (49.8)	158 (33.3)	70 (14.8)	10 (2.1)	474

(b)인접 2단어를 이용한 방식에 의한 평가

	1	2	3	4	합계
A	216 (45.6)	3 (0.6)			219 (46.2)
B	97 (20.5)	98 (20.7)	35 (7.4)	3 (0.6)	233 (49.2)
C			18 (3.8)	4 (0.8)	22 (4.6)
합계	313 (66.0)	102 (21.5)	53 (11.2)	7 (1.5)	474

(2) 시제의 번역실패(2,A)

セボロの服装で行く必要がある

세비로의 복장으로 가는 필요가 있다(x)→갈 필요가 있다(0)

銀行にいくのは　은행에 가는 것은(0)

(3) 후접하는 용언의 패턴에 의한 조사의 번역 실패(2,B)

技術者と言う　기술자와 말한다(x)→기술자라고 말한다(0)

技術者と話す　기술자와 말한다(0)

(4) 어순이 맞지 않는 번역 실패(3,B)

改善されつつある　개선되어 계속있다(x)→계속 개선되고 있다(0)

死ぬまで走り續ける　죽을 때까지 달려 계속한다.(x)→계속 달린다(0)

ページを入れ替える　페이지를 넣어 바꾼다(x)→바꾸어 넣는다(0)

(5) 조동사의 다의성에 의한 번역실패(3,C)

講演が2時間も續けられる　강연이 2시간도 계속되었다(0)

彼は仕事を續けられるようになった

그는 일을 계속하여지도록 되었다(x)→계속하도록 되었다(0)

(6) 관용어의 번역실패

気が短い　기가 짧다(x)→성격이 급하다(0)

(7) 용언의 다의성에 의한 번역실패(4,C)

3億円を引いた金額　3억원을 당긴 금액(x)→3억원을 뺀 금액(0)

手の動きがうまい　손의 움직임이 맛있다.(x)→손의 움직임이 노련하다.(0)

うまいラーメン屋がある　맛있는 라면집이 있다.(0)

(8) 적절한 표현이 어려운 번역실패(4,C)

やむを得ないだろう　그만둠을 얻지 않겠지(X)→할 수 없겠지(0)

5. 결 론

본 논문에서는 인접단어와의 접속관계를 정의한 확장 번역테이블을 이용하는 일한 번역처리방식을 제안하였다. 확장번역테이블의 특징으로는 기존의 번역테이블에서 품사별로 종속되게 정의한 구조를 품사에 관계없이 엔트리 테이블과 접속정보 테이블을 이용하는 번역처리를 행할 수 있었으며, 조사, 조동사와의 접속관계를 표현함으로써 종래의 번역테이블 방식에서는 나타낼 수 없었던 조사, 조동사와의 인접관계를 보다 세밀히 표현할 수 있다. 번역 대상 단어별로 좌우 단어와의 접속관계를 자유롭게 기술할 수 있으므로 번역 시스템을 완성해 둔 후에 더욱 번역률을 높이는 번역 대상 단어별 튜닝이 쉽게 가능하다. 또한, 일한 기계번역에서 대표역어로 번역 가능한 약 70%의 경우는 엔트리 테이블에 의한 1대1

번역을 행하고, 역어가 복수인 단어만을 접속정보 테이블에서 처리하기 때문에 단어별로 고정된 양의 정보를 가지는 번역테이블 방식보다 번역사전의 용량을 어느 정도 세이브하는 효과도 있다. 마지막으로 번역 가능한 복수의 접속관계와 역어를 번역사전에 기술하고 인접2단어와의 접속관계를 이용한 역어 선택 알고리즘을 제안함으로써 인접단어의 정보만을 사용하는 1대1 대응 직접번역방식에 따른 일한기계번역 시스템을 실현하였다.

그러나, 문장의 구조나 의미해석을 생략하여 일한 번역시스템을 단순화하는 데는 성공하였지만 용언이나 조동사의 다의성에 대한 처리, 현재와 미래 시제의 적절한 표현, 양국어 어순의 차이에서 오는 역어의 부자연스러움 등은 차후에 개선되어야 할 과제이다.

참 고 문 헌

[1] 野村浩郷、言語処理と機械翻訳、講談社、1991.

[2] 田中穂積、自然言語解析の基礎、産業図書、1989.

[3] 野村浩郷、田中穂積、機械翻訳-bit別冊、公共出版、1988.

[4] 長尾真、機械翻訳サミット、オーム社、1989.

[5] 野村賢一、ほか、計算機用日本語動詞辞書 I PAL(Basic Verbs)-解説編、情報処理振興協会技術センター、1987.

[6] 田辺洵一、くわしい国文法、文英堂、1987.

[7] 日本語翻訳システム環境下での韓国語飜訳システム開発のための一考察、情処学、自然言語処理研究報、Vol.86, No 4, 1988.

[8] 李義東、中嶋正之、安居院猛、語と語の関係を用いた意味解析による日韓短文機械翻訳 システム、信学論(D-II) Vol.72, No 10, 1989.

[9] 李義東、中嶋正之、安居院猛、助詞表現の意味対応による日韓述部機械翻訳システム、情処学論 Vol.31, 801-809, 1990.

[10] T.S.Kim, Syoji.Ura, A Japanese-Korean Machine Translation based on Conjugated Words analysis, ICEIC, 199-203, 1991.

[11] 金泰錫、浦昭二、日韓両国語の類似性を利用した機械翻訳、慶応義塾大学、理工学研究科、Technical Report, No 91001, 1991.

[12] 金泰錫、浦昭二、日韓機械翻訳における意味接続関係を考慮した翻訳テーブル、慶応義塾大学、理工学研究科、Technical Report, No 92002, 1992.

[13] 金泰錫、金政仁、大駒誠一、浦昭二、意味接続関係に基づく翻訳テーブルを用いた日韓機械翻訳における日本語の形態素解析、43回情処学全大、Vol.3, 201-202, 1991.

[14] 李秀絃、小沢、韓日機械翻訳のための音韻表現形式による用言の活用処理、情処学論 Vol.33, 1-12, 1992.

[15] 金政仁、大駒誠一、日韓機械翻訳における動詞の多訳性処理、45回情処学全大、Vol.3, 97-98, 1992.

[16] 金泰錫、浦昭二、日韓機械翻訳における意味接続関係を用いた韓国語の生成方法 情処学論 Vol.33, No12, 1992.

[17] 金泰錫、浦昭二、日韓機械翻訳における否定語の処理、情処学論 Vol.34, No5, 1993.

[18] 金政仁、金泰錫、大駒誠一、日韓機械翻訳における話し言葉の翻訳処理、47回情処学全大、Vol. 3, 179-180, 1993.

[19] 黄道三、長尾真、日本語分類語彙表から韓国語分類語彙表の作成、信学自然言語処理研究報、94-11, 79-84, 1993.

[20] 宮崎政弘、白井諭、池原悟、言語過程説に基づく日本語品詞の体系化とその効果、言語処理学会論文誌、Vol.2, 2-36, 1995.

[21] J.I.Kim, T.S.Kim, S. Okoma, A Processing of Polysemy and Multi-Translatable Verbs on Japanese Korean Machine translation, Proceedings of ICCTA'94, 144-148, 1994.

[22] 金政仁、金泰錫、大駒誠一、擴張飜訳テーブルを用いた日韓機械飜訳、51回情処学全大、Vol.3, 89-90, 1995.

[23] 문경희,이종혁,김정인,양기주, 일-한 기계 번역 시스템 : 연어 패턴을 이용한 어휘 다의성 해소, 정보과학회 논문지(B)제25권 제8호, pp.1270-1280, 1998.

[24] 김정인, 문경희, 이종혁, 일한기계번역에서 진행형 "떼이루"의 번역처리, 정보처리학회 논문지 제8-B권 제6호, 685-692, 2001.

김 정 인

1996년 일본 게이오대학 계산기 과학 공학박사
1996년~1998년 포항공과대학교 정보통신연구소 연구원
1998년~현재 동명정보대학교 공과대학 컴퓨터공학과 조교수

관심분야 : 자연어처리, 정보검색, 지식공학