영한 기계 번역에서 한국어 부사의 어순 결정에 관한 연구

*이신원(李信源), 안동언, 정성종
*정인대학 컴퓨터정보계열, 전북대학교 컴퓨터공학과
전화 : (063) 530-9237 / 팩스 : (063) 532-3768
H.P 번호 : 016-658-3781

A Study of Korean Adverb Ordering in English-Korean Machine Translation

*Shin Won Lee, Dong Un An, Seong Jong Chung
*Computer Information of Chongin College,
School of Computer Engineering Chonbuk National University
*E-mail : swlee@mail.chongin.ac.kr

Abstract

In the EKMT system, the part of Korea generation makes Korea sentence by using information obtained in the part of transfer. In the case of Korea generation, the conventional EKMT system don't arrange hierarchical word order and performs word order in the only modifier word.

This paper proposes Korean adverb odering rule in English-Korean Machine Translation system which generates Korean sentence.

I. 서론

컴퓨터를 이용하여 정보를 공유하는 인터넷이 많이 보급되면서 인간의 언어를 처리하는 기계번역에 관심을 많이 갖게 되었다. 그러한 분야로서 영어와 한국어같이 언어 형태가 굴절어와 교착어로 서로 크게 다른 두 언어를 번역하는 기계번역 시스템에서는 두 언어간의 차이점을 보완하고 이어주는 것이 매우 중요하고 어렵다. 이러한 과정을 위해 변환과 생성과정이 있다. 변환부문 역시 매우 중요하지만, 변환만으로는 완전하고 올바른 목적 문장인 한국어 문장의 생성이 어려우므로 목적언어 고유의 특성들은 생성부분에서 처리해야 한다. 즉, 변환의 결과에 한국어 고유문법에 맞는 의존구조로 변환하거나 자연스러운 본연의 한국어 자질을 수정, 추가, 생성하여야 한다.

영한기계번역을 위한 한국어 생성기는 하나의 번역단위마다 크게 한국어구문 생성단계와 한국어형태소 생성단계로 나누어 처리한다. 생성기에서 전적으로 한국어의 문법과 특성에 근거를 두고 한국어 처리를 요구하게 된다. 한국어구문 생성단계에서는 한국어 특성에 맞게끔 문체에 따른 의존구조를 변환하고 변환에서 넘어오는 의존구조를 탐색하여 수식어들의 어순을 결정하고 의존구조의 노드 속성으로 표현한다.

이 중에서 수식어인 부사가 연속해서 나올 경우 어떤 자질의 부사를 먼저 나타내고 나중에 나오는 부사의 자질은 어떠한 것인지에 따라 보다 더 자연스러운 문장을 생성해 낼 수 있다. 이러한 부사의 자질에 대한 분석이 국문학자들에 의해서 분류되고 있어서 기계번역을 하기 위한 일정한 규칙을 찾아낼 필요가 있다.

이러한 과정을 위해 변환과 생성과정이 있다. 변환부문 역시 매우 중요하지만, 변환만으로는 완전하고 올바른 목적 문장인 한국어 문장의 생성이 어려우므로 목적언어 고유의 특성들은 생성부분에서 처리해야 한다.

본 논문은 영한기계번역을 위한 한국어 생성단계에서 한국어 특성에 맞게끔 문체에 따른 의존구조를 변환하고 변환에서 넘어오는 의존구조를 탐색하여 수식어들

의 어순을 결정하고자 한다.

Ⅱ. 본론

영어에는 한국어에 없는 전치사가 있어서 이를 번역시 부사로 바뀌는 경우가 있다. 그리고 같은 전치사이지만 pentree_bank corpus에 태깅되어 있는 부사를 보면 전치사도 부사로 태깅되어 있다.

이처럼 영어의 부사가 한국어의 부사로 그대로 번역이 되지 않기 때문에 한국어에서 부사로 번역이 되는 경우 수식어인 부사가 연속해서 나올 경우를 살펴보고자 한다. 부사가 연속해서 나오는 경우 어떤 자질의 부사를 먼저 나타내고 나중에 나오는 부사의 자질은 어떠한 것인지에 따라 보다 더 자연스러운 문장을 생성해 낼 수 있다는 것을 알았다. 이러한 부사의 자질에 대한 분석이 국문학자들에 의해서 분류되고 있으나 기계번역을 하기 위해서 일정한 규칙을 찾아낼 필요가 있다.

Similarly, highway engineers agreed to keep the old railings on the Key Bridge in Washington, D.C., as long as they could install a crash barrier between the sidewalk and the road.

Similarly/RB ,/, highway/NN engineers/NNS agreed/VBD to/TO keep/VB the/DT old/JJ railings/NNS on/IN the/DT Key/NNP Bridge/NNP in/IN Washington/NNP ,/, D.C./NNP ,/, as/RB long/RB as/IN they/PRP could/MD install/VB a/DT crash/JJ barrier/NN between/IN the/DT sidewalk/NN and/CC the/DT road/NN ./.

그림 1 pentree_bank corpus 예

다음의 예를 보자.

I ofen go to the movies alone.
나는 자주 혼자서 영화보러 간다.

위의 예를 영어 의존구조로 그리면 다음과 같다.

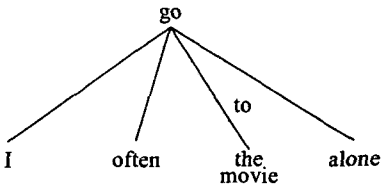

그림 2 영어 의존구조

위의 의존구조를 생성시 한국어 의존구조로 그리면 다음과 같다.

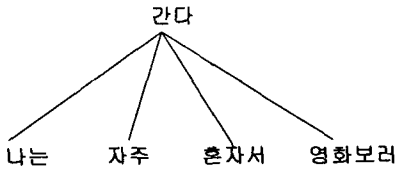

그림 3 생성된 한국어 의존구조

다른 예를 보면 다음과 같다.

He was working hard there then.
그는 그때 거기서 열심히 일하고 있었다.

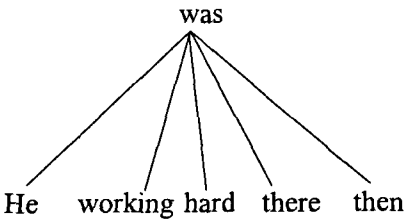

그림 4 영어 의존구조

그림 5 생성된 한국어 의존구조

위의 예를 보면 영어의 부사의 위치 정보가 번역 생성 과정을 거치면서 연속된 부사로 생성이 되고 생성된 순서도 바뀌어서 생성되었다. 첫 번째 문장을 보면

'ofen', 'alone' 부사가 떨어져 있는 위치에 있음에도 불구하고 한국어 생성시 '자주 혼자서'로 연속 생성됨을 알 수 있다. 두 번째 문장을 보면 'hard there then' 부사가 한국어 생성시 '그때 거기서 열심히'로 연속 생성되고 순서도 바뀜을 알 수 있다.

이와 같이, 영어 부사가 해석될 때 위치정보가 바뀌어 한국어 부사로 해석되는 경우가 있다. 한국어 생성시 연속 부사의 위치가 바뀌면 비문이 될 수 있다. 이처럼 한국어 연속 부사에도 위치 정보가 있음을 알 수 있다. 국문학자들이 분류해 놓은 한국어 연속 부사에 대해서 알아보고 국어정보베이스 시스템의 코퍼스를 이용하여 부사 자료를 추출하여 부사의 위치 정보를 알아보고자 한다.

국문학자 중에서 김민수[1]와 김창호[3]가 부사의 어순에 대해서 분류해 놓았다. 둘 다 부사의 종류를 6가지로 분류하였다. 부사의 종류를 부르는 명칭은 다르나 그 성질이 비슷하여 여기서는 하나의 이름으로만 사용을 하도록 한다. 김민수는 부사에 대한 어순을 예로 부사의 종류를 분류하였고 김창호는 보다 구체적인 예를 나열하면서 어순을 정하고 부사의 종류를 분류하였다. 그 종류를 보면 다음과 같다.

김민수분류 양태, 시간, 처소, 수량, 정도, 상태
↓ ↓ ↓ ↓ ↓ ↓
김창호분류 서법, 시간, 장소, 수량, 정도, 양태

그림 6 부사의 분류

양태(서법)→화자의 태도를 나타내는 부사
시간→서술의 시점이나 시간단위를 나타내는 부사
처소(장소)→사건이 일어나는 장소나 방향을 나타내는 부사
수량→행동이나 개체의 수량을 나타내는 부사
정도→상태나 동작의 정도를 나타내는 부사
상태(양태)→행동이나 모양의 양태를 나타내는 부사

그림 7 부사의 자질

김창호는 예를 들어서 설명하고 있다. 부사의 예를 보면 다음과 같다.

(1-1) 지금 막 도착했다.
(1-2) 막 지금 도착했다.
(1-3) 현재 막 도착했다.

(2-1) 밥을 빨리 못 먹는다.
(2-2)* 밥을 못 빨리 먹는다.
(2-3) 밥을 잘 못 먹는다.

(1-1), (2-1)은 사용 가능한 문장이고 자연스럽다. (1-2)는 사용은 가능하나 부자연스러운 문장이다. (2-2)는 비문법적인 문장이다. (1-3)은 비문법적인 문장이다. 위의 예를 보더라도 연속해서 부사가 나올 경우 부사의 어순이 존재한다는 것을 알 수 있다.

김창호는 부사의 어순을 예를 들어 설명하고 다음과 같이 정하였다.

서법→시간↔장소↔수량→정도→양태

Ⅲ. 실험 및 평가

본 논문에서는 한국어 말뭉치를 통하여 부사의 어순을 추출하고자 1997년에 제작된 '통합 국어정보베이스'에 수록된 '한국어 구문구조 부착 말뭉치' 1만여 문장을 가지고 분석하였다. 코퍼스에 나타난 부사의 자질은 2530개 정도이고 연속 부사는 10018개 정도이다. 세 개 연속된 부사는 140개 정도이고 네 개 연속된 부사는 13개 정도이다. 이 중에서 빈도수가 5번 이상 연속해서 나온 부사에 대해서만 추출하여 분석해 보았다.

코퍼스 분석을 통하여 빈도수와 앞 뒤의 연속 부사를 고려하여 180개의 부사만을 추출하였다. 이 연속 부사에서 주로 앞에 나오는 부사는 다음과 같다.

거의, 결국, 그러다가, 그럼, 꽤, 너무나, 늘, 달리, 도대체, 도저히, 물론, 벌써, 보다, 서로, 아니, 아마, 아무리, 아주, 어째서, 어쩌면, 어찌, 언제, 언제나, 역시, 오히려, 이리, 이미, 이제, 이젠, 전혀, 점점, 정말, 제대로, 제발, 제일, 좀더, 참, 채, 특히, 한꺼번에, 훨씬

주로 뒤에 나오는 부사는 다음과 같다.

가까이, 가만히, 계속, 깊이, 널리, 높이, 덜, 많이, 말하자면, 먼저, 멀리, 못, 바싹, 분명히, 스스로, 안, 열심히, 오래, 완전히, 일찍, 자세히, 자주, 정확히, 천천히

이 부사들에 대해서 김창호가 분류해 놓은 기준으로 빈도수를 추출하고 그 비율을 나타내었다.

빈도수	서법	시간	장소	수량	정도	양태
서법	277	114	71	212	279	1203
시간	34	276	24	437	434	355
장소			26	43		76
수량	13	64	10	510	227	547
정도	10	68	201	1272	1385	1850
양태	116	36	44	336	619	1230

표 1 김창호 어순에 따른 빈도수 추출

빈도수	서법	시간	장소	수량	정도	양태
서법	12.8	5.3	3.3	9.8	12.9	55.8
시간	2.2	17.7	1.5	28	27.8	22.8
장소	0	0	17.9	29.7	0	52.4
수량	0.9	4.7	0.7	37.2	16.6	39.9
정도	0.2	1.4	4.2	26.6	28.9	38.7
양태	4.9	1.5	1.8	14.1	26	51.7

표 2 김창호 어순에 따른 비율

위의 비율을 살펴보면 빈도수와 비율을 고려해 볼 때 아래와 같은 순서가 더 고른 비율을 가지고 있음을 알 수 있다.

빈도수	장소	시간	서법	수량	양태	정도
장소	26	0	0	43	76	0
시간	24	276	34	437	355	434
서법	71	114	277	212	1203	279
수량	10	64	13	510	547	227
양태	44	36	116	336	1230	619
정도	201	68	10	1272	1850	1385

표 3 빈도수에 따른 어순 재정렬

빈도수	장소	시간	서법	수량	양태	정도
장소	17.9	0	0	29.7	52.4	0
시간	1.5	17.7	2.2	28	22.8	27.8
서법	3.3	5.3	12.8	9.8	55.8	12.9
수량	0.7	4.7	0.9	37.2	39.9	16.6
양태	1.8	1.5	4.9	14.1	51.7	26
정도	4.2	1.4	0.2	26.6	38.7	28.9

표 4 비율에 따른 어순 재정렬

다음과 같은 순서로 부사의 어순을 결정할 때 보다 더 자연스러운 문장을 생성해 낼 것이다.

장소→시간→서법→수량→양태→정도

Ⅳ. 결론

본 논문은 영한 기계 번역 시스템에서 보다 더 자연스러운 번역을 위해서 생성할 때 국문학자가 분류해 놓은 분류기준을 토대로 '통합 국어정보 베이스'에 등록되어 있는 1만여 문장의 corpus를 분석하여 부사의 자질 정보와 연속 부사를 추출하여 부사의 어순 결정을 시도하였다.

위의 실험 결과를 보면 알 수 있듯이 수량부사, 정도부사, 양태부사는 좀 더 세분화되어 분류를 하면 더 나은 부사의 어순을 생성해 낼 것으로 생각된다.

예를 들면, 양태부사 중에서 '안', '못' 부사는 부정의 뜻을 가지고 있으므로 특수한 경우이다. 의성어와 의태어도 양태부사로 분류되어 있다. '아니' 부사는 서법 부사 등을 가리키면서 앞에 나온다.

앞으로 생성 단계에서 부사의 어순을 사용하면 영한 기계번역 시스템에서 수식어의 생성이 보다 더 자연스럽게 될 것이다.

참고문헌(또는 Reference)

[1] 김민수, 국어문법론, 일조각, 1984

[2] 남기심, 표준국어문법론, 탑출판사, 1985

[3] 김창호, 국어 부사어의 어순에 관한 연구, 계명대학교

[4] 손남익, 국어 부사 연구, 박이정, 1995

[5] 박성재, 영한 번역기에서의 부사구 처리에 관한 연구, 서울대학교 컴퓨터공학과 석사학위논문, 1992

[6] 조준모외 1인, 한·영 기계 번역을 위한 부사의 위치 및 순서제약 해결의 방안 및 구현, 제 6 회 한글 및 한국어 정보처리학회, pp. 163-167, 1994

[7] 서정수, 국어문법, 한양대학교 출판원, 1996.

[8] 통합 국어정보베이스, 과학기술처, 1997.

[9] 영한 기계 번역 시스템에서 계층적 한국어 어순 생성, 전북대학교 컴퓨터공학과 석사학위논문, 2001.

[10] http://my.netian.com/~beedman/adverd.htm

[11] J.S. Chang and K.Y. Su, "A Corpus-Based Statistics-Oriented Transfer and Generation Model for Machine Translation," TMI-93: The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, pp.3-14, 1993.