한.영 기계번역을 위한 중심어 기반 구 구조 변환 사전

이 상조* 박 상규* 김영택**
*경북대학교 컴퓨터공학과 **서울대학교 컴퓨터공학과

Head-based Pharse Structure Transfer Dictionary for Korean_English Machine Translation

San-Jo Lee* Sang-Kyu Park* and Yung-Taek Kim**
*Dept. of Computer Engineering Kyung-Pook National University
**Dept. of Computer Engineering Seoul Natiuonal University

한국어로부터 자연스러운 영어 역어문장을 생성하기 위한 정보를 사전에 일관성있게 수록하는 방법을 제시하였다. 기계번역의 각 과정에서 필요한 정보는 가장 적당한 형태로 사전으로부터 제공되어야 하는 것이 일반적인 방법이다. 그러나 한국어는 어순의 부분적 자유성, 어미의 복잡한 활용규칙, 조사의 다양한 쓰임새로 인해 이러한 규칙들의 정보를 일관되게 사전에 수록하기가 어려운 실정이다. 본 논문에서는 한국어 문장과 역어 문장을 단어나 구 혹은 절등의 구성요소들의 다대다 매핑규칙을 찾고 이들 규칙을 적당한 형태로 사전에 수록하여야하는 어려움에서 벗어나 문장대 문장구조를 직접대응시켜 구구조단위로 분석된 형태의 부분 파서트리 형태의 트리구조를 역어와 함께 사전에 수록하므로서 사전정보를 손쉽게 구축, 유지하고자 하였다. 또 이들 정보를 추출해내는 알고리즘을 사용함으로써 주어진 한국어 문장에 대해 사전에 수록된 가장 자연스러운 형태의 역어문장을 생성할 수 있도록 하였다.

1. 서 론

기계번역은 하나의 언어로 표현된 입력 문장으로 부터 동일한 의미를 갖는 다른 언어로 된 문장을 출력하는 작업을 컴퓨터에 의해 일어나게 하는 것을 말한다.

본 연구의 동기는 질적으로 우수한 기계번역을 위하여 사전에서 제공할 수 있는 정보를 어떻게 용이하게 줄 수 있으며 또한 쉽게 이용하는 방법론을 발견함에 있다. 기계번역에서 쓰이는 사전에서 제공되는 정보는 기계번역의 각 단계의 쉽고 어려움 뿐만 아니라, 기계번역된 결과의 질을 좌우하므로 특히 많이 연구되고 있다[2,3,7,8,9,12,13,16]. 기계번역을 위해 고안된 지식과 언어학에서 설명하는 언어현상에 관해 연구된 지식들이 사전에 적당한 형태로 구축되어 있어야 하므로 이에 관한 연구가 중요한 관심사가 되어 왔다. 본 연구에서는 영어 문장과 그에 상응하는 한국어 문장을 두고 이들을 분석하여 가급적 문법적 지식이 없이 표층구조를 그대로 사전에 두어, 단일화 연산 및 제약조건에 의한 역어선택 등 기계번역에 필요한 절차를 간단히 하면서도 자연스러운 영어 문장이 생성되도록 하였다.

지금까지 한·영 기계번역을 위해 고안된 번역 사전을 살펴보면 다음과 같은 점에서 개선되어야 한다.

1. Collocation등의 정보를 주는 방법에 일반적인 규칙을 부여함이 바람직하다.
2. 목적언어 생성 면에서 일관성을 유지할 수 있는 방안이 마련되어야 한다.
3. 새로운 정보가 사전에 수록하는 방법의 일관성을 유지해야 한다.

이상의 사실들로 부터 중심어에 기반한 구 구조 사전은 이러한 점에 대해 중심어 주도 구 구조 문법(HPSG)의 자질틀(자질값의 집합)이라는 개념[14,15]의 의미를 하나의 원자적 자질 요구(하위범주나 수식, 피수식)로부터 분자적 자질 요구 이상으로 확대시키므로써 모든 언어적 현상을 HPSG의 확대된 자질틀 내에 융합시켜 표현하도록 하였다. 기계번역에서의 역어 문장의 생성을 숙어 및 관용어에 대한 역어, 정형적인 방법에 의한 역어를 별도로 처리하는 다원적인 처리가 아닌 하나의 융합된 방법으로 일반화 하였다. 그래서 하나의 문장에 대한 역어가 완전히 숙어적 표현으로 된 그러한 것에서부터 부분적인 구나 절이 숙어적 표현인 경우와 완전히 문법적 처리에 의한 정형적 역어 생성이 가능한 경우까지 하나의 일관된 처리 방식을 통해 역어를 만들어 내자는 것이다. 단일한 하나의 문법기제를 이용하여 형태소 분석, 구문 분석 그리고 역어 선택까지 처리하므로써 이에 필요한 모든 사전이 단일한 구조를 가지게 되어 자연어 처리시의 많은 예외적 상황등을 포함한 모든 처리를 하나의 일관된 처리 속에서 해결할 수 있다. 끝으로 이 연구에서는 (1)의 일반성과 (2)의 일관성의 획득으로 인해 누구나가 필요에 따라 사전 정보를 쉽게 첨가할 수 있는 융통성 있는 사전의 신축적 운용을 위한 방법론을 제안하는 것이다.

2 변환 사전의 구축

자연스러운 영어 문장의 생성을 위해서 한국어 문장을 영어 표현에 대응시켜 보면, 구분되는 번역의 단위로는 단어, 관용적 표현을 포함하는 숙어, 구 구조, 또는 문장 전체가 될 수 있다. 한국어의 문장을 구문 분석한 후 번역의 대응 단위 별로 부분 파스 트리를 얻고 중심어에 해당하는 단어 또는 한 단어처럼 쓰이는 숙어를 표제어로 하여 부분 파스 트리를 구조화한 결과와 영어 표현을 사전에 수록하도록 한다. 이 논문에서 제시하고자 하는 구 구조 변환 사전의 구축에 대해서 사전 구축을 위한 구 구조 설정과 사전 정보의 구성에 관하여 구체적인 내용을 기술한다.

2.1 구 구조 설정시 고려해야 할 사항

한국어 문장의 구 구조를 영어 문장으로 번역하는 관점에서 목적 언어로의 변환시에 변환 단위는 각각의 단어가 아니라 분류된 구 구조를 그 단위로 한다. 이러한 구는 구 그 자체만으로 인식되어야 하며, 그들이 하나의 단위를 이루어 번역되어야 자연스러운 번역을 얻을 수 있다.

따라서, 보다 자연스러운 영어 문장의 생성을 위하여 구 구조의 설정은, 실제 널리 사용되는 영어 문장을 기준으로 하여 그에 대응되는 한국어 문장을 비교하면서 역어 선택에 필요한 구성 성분들을 가급적 길게 하여 의미의 모호성까지 해결할 수 있도록 하였다. 이러한 구의 처리를 위하여 구 구조의 범위와 그들의 유형을 다음과 같이 정의한다.

■ 한·영 기계 번역에서의 구 구조

구 단위로 번역되어야만 적절한 역어를 얻을 수 있는 연속된 어절들로 된 구.

구 구조의 유형은 다음과 같이 분류하였다.

(1) 숙어

한국어 문장에서 나타나는 표현들 중 일부는 쓰임이나 형태가 고정되어서 하나의 품사처럼 쓰이는 표현이 있다.

이러한 표현들에 대해서 굳이 더 작은 단위로 분석하여 그에 대한 파스 트리 형태의 구 구조를 변환 사전에 수록해 둔다고 해도 번역에 많은 도움이 되지 못하므로, 이러한 형태의 구는 하나의 단위로 취급하는 것이 바람직하다[18].

(2) 역어 선택 성분을 가진 구 구조

한국어의 표현에 대한 영어의 단어를 선택할 때 한국어의 몇 단어들이 함께 사용되어 영어의 한 표현에 대응되는 경우가 있다. 즉, 번역에 대응되는 단어들의 수적인 관계가 다대일 혹은 다대다의 번역이 되는 구를 역어 선택 성분을 가진 구 구조의 범주에 넣었다. 이러한 구의 경우 중심어인 술어가 그를 보충해 주는 주어, 목적어 혹은 그를 수식하거나 행위의 방법을 설명하는 수식어구와 함께 하나의 역어로 번역된다. 이것은 단어가 그들의 보충어 성분 혹은 수식 성분과 함께 여러 의미로 다르게 사용되고 있음을 보이고 있다. 하지만 이러한 유형의 구 구조를 고정된 형태로 인식하여 처리할 경우에 구성 성분들 사이에 수식어 혹은 다른 보충어 성분들이 자유롭게 끼어 들 수 없게 된다.

(3) 문 형식 결정 성분을 가진 구 구조

한국어의 일반적인 특성들 중의 하나로 용언의 다양한 어미 변화와 조사의 쓰임의 다양함이다. 한·영 기계번역의 관점에서 이들을 살펴보면 이 중 몇 가지는 영어로 번역될 때 특별한 의미의 역어를 선택하는 데 필요한 정보를 주기 보다는 문장의 형식을 결정하는 데 필요한 정보를 제공한다. 부사형 어미와 몇몇 부사들이 특수한 형태의 서술어와 호응하여 사용되는데 부사의 경우는 부정이나 비교, 추측의 의미를 지닌 다른 부사들이 있고 부사형 어미도 마찬가지이다. 이러한 활용 어미나 부사는 중심어가 아니므로 중심어의 자질에 그의 자질들을 첨가할 뿐이지만, 실제로 번역하는 경우 영어 문장 생성에서는 큰 의미를 가지는 것들이다.

(4) Collocation

Collocation에 대한 정의는 사전적 의미로부터 기존의 기계번역 시스템의 적절한 처리를 위해 여러 가지로 언급되고 있지만 쓰임에는 큰 차이가 없다. 하나의 단어가 다의어이거나 동형이의어일 때 그 의미는 문장을 구성하는 다른 단어와의 의미 관계에서 결정된다고 본다. 이러한 정보들을 이용하여 한국어와 영어가 일대일 대응 관계를 이룰때, 다의어나 동형이의어의 의미 모호성 뿐만 아니라 어휘 모호성을 해결할 수 있다.

이러한 구 구조의 유형 분류는 올바른 번역문을 생성하기 위하여 사전에 있어야 할 입력 문장에서의 구 구조 설정과 그에 따른 표제어 설정등을 결정할 뿐이다.

2.2 변환 사전의 구성

입력된 한국어 문장에 대해서 문장내에서의 문법적 관계를 규명하고, 각 어휘의 의미를 파악하기 위해서는 문장의 분석이 필요하다. 이 분석은 그 자체로 의미를 부여하기 보다는 분석된 결과의 쓰임에 의존하므로, 여기서는 영어 문장 생성을 위한 문장의 분석이 궁극적 목표가 된다. 이 논문에서는 구문 분석을 의미론에 근접한 정보 기반의 문법 이론으로서 HPSG(중심어 주도의 구 구조 문법)를 채택하고 일반적인 방법에 의한 구문 분석등[5,6]의 기본 골격에 역어 생성에 필요한 정보를 얻을 수 있도록 일부 변형을 가하였다. 한국어의 조사, 어미의 처리가 역어의 생성에 꼭 필요하므로 이의 처리를 위해 중심어-조사 구조를 도입하여 조사를 문법적 처리의 범주에서 다룰 수 있는 통합 규칙을 추가하였다[15]. 아울러 어미에 관해서도 독립된 범주 자질을 부여하기에는 무리가 있지만 역어 생성에 최대한 많은 정보를 제공하도록 형태소의 분석을 시도하였다. 그 외의 구절은 중심어-보어 구조를 중심어-보충어 구조, 수식어-중심어 구조를 중심어-수식어 구조, 그 외의 접속 구조, 체언-조사의 형태를 중심어-조사 구조로서 처리하였다.

2.3 사전 구성 형식 및 내용 기술

중심어 중심의 사전에는 구 구조, collocation등의 정

보들이 구문 분석의 결과인 파스 트리의 형태 그대로 수록되며, HPSG의 특성상 자질들의 구조가 같은 형식으로 얼마든지 중첩될 수 있는 순환적 구조이다. 그러므로 아무리 복잡한 구 구조에 대한 역어 변환 정보라 할지라도 일부 메타 기호를 첨가한 순환적 구조를 써서 표시할 수 있다. 여기서 메타 기호는 주어나 목적어등이 아직 결정되지 않은 구 구조 형태를 표시하게 될 때 쓰인다. 따라서 이러한 단일한 기제로서 순환적 검색및 비교(recursive retrieval and comparison)연산을 통해 가장 적절한 역어를 선택할 수 있다. 변환 사전은 크게 한국어 분석 결과부와 역어부로 나누어 진다.

[사전 형식]

[$표제어 ^ 표제어의 자질 값] #(역어);

한국어 분석 결과부는 역어 선택을 위한 표제어와 역어 선택에 이용될 정보를 나타내는 표제어의 자질 값으로 다시 나눌 수 있다. 사전은 표제어를 기점으로 순환적 구조를 가진 사전의 한 엔트리를 역시 순환적 검색 및 비교 연산에 의해 구 구조부의 최말단에 이르기까지의 총괄된 구조로 포착하게 되는 것이다.

표제어의 자질 값은 구문 분석에서 이용하는 자질 값의 내용과 동일한 것으로 번역될 구의 부분 구문 분석 결과 중 그의 역어를 선택하는 데 필요한 자질들을 모두 나열할 수 있다. 구 구조의 역어를 결정하기 위하여 필요에 따라서는 표제어의 자질 값이 계층적인 구조를 가질 수 있다. 사전 형식에서 보듯이 역어 선택 정보를 위한 형태는 모두 같다. 역어부에는 구의 단순한 역어 외에도 그의 전후에 올 수 있는 문장 성분들을 표시하여 둠으로써 역어 생성부에서 구 단위 변환 사전에서 선택된 역어들의 순서 관계를 따로 설정하지 않더라도 완전한 영어 문장을 생성할 수 있다. 이제, 표제어의 자질 값으로 올 수 있는 한국어 문장 분석의 결과는 다음과 같은 기본 구조들을 가질 수 있다.

◈ 중심어 - 조사 구조의 사전 구성
[중심어^중심어자질값^조사:[조사^조사자질값]]

◈ 중심어 - 수식어 구조의 사전 구성
[중심어^중심어자질값
^수식어:[수식어^수식어자질]] #(역어);

◆ 중심어 - 보충어 구조의 사전 구성
[중심어^중심어자질값
^보충어^[보충어^보충어자질값]] #(역어);

◆ 접속 구조의 사전 구성
[중심어(연결어)^접속구1의자질값
^접속구2의자질값] #(역어);

◈ 숙어 구조의 사전 구성
[숙어^숙어자질값] #(역어);

◈ 단일단어 사전 구성
[표제어^표제어 자질 값] #(역어);

2.4 사전 구성의 실제

지금까지 설명한 내용들을 바탕으로 실제 사전을 구성하는 예를 보이고자 한다.

(2-1) "I will take a cold shower."

예문 (2-1)에 대해서 한국어 문장 (2-2)을 대응시키려고 한다.

(2-2) "나는 찬물로 샤워를 하겠다."

분석 결과에 의하면 '나는', '찬물로', '샤워를', '하겠다'등으로 어절들이 구분되어 그 각각에 대해서 'I', 'with cold water', 'shower', 'will do'등으로 역어를 대응시킬 수 있겠지만 'take a shower'에 대해서 '샤워를 하다'가 대응되므로 '하(다)'의 표제어 밑에 '샤워를'을 보충어 성분(문장성분:목적어)으로 갖게 하여 (2-3)처럼 사전에 수록한다.

(2-3) [$하^문장성분:서술어^보충어:[*^문장성분:주어]
^보충어:[샤워^문장성분:목적어]] #(*take a shower)

다음 '찬물로'를 고려하면 이 어휘는 '하다'의 수식어 성분(문장성분:부사어)이어서 'with cold water'로 역어를 처리할 수 있으나, 이렇게 하면

(2-4) "I will take a shower with cold water."

로 문장이 되므로 원 문장과 거리가 있다. 따라서 번역의 대응 단위를 '나는', '찬물로 샤워를 하다'로 정하면

(2-5) [$하^문장성분:서술어^보충어:[*^문장성분:주어]
^보충어:[샤워^문장성분:목적어]^수식어:[찬물^품사:명사^문장성분:부사어^조사:[로^품사:조사^조사기능:방법]]] #(*take a ^A.A cold shower);

이 된다.

이와 같이 영어에 보다 가깝고 자연스러운 표현의 번역문을 생성하기 위한 중심어 기반 구 구조 변환 사전을 구축하기 위해서는 수집된 영어 문장에 대한 한국어 번역문을 살펴 보면서 어떤 정보를 어디까지 사전에 수록할 것인가를 검토하여야 한다. 이 사전은 한국어 분석 결과를 그대로 사전에 수용하였으므로 구 구조를 이루는 구의 성분들이 고정되지 않은 부분 자유 어순을 허용하는 한국어의 처리에 적합하며, 또한 구 구조의 구성 성분들 사이에 부사어나 관형어의 삽입된 형태의 구를 인식하는 데에도 큰 무리가 없으며 적절한 의미의 역어를 선택할 수 있다.

3 한·영 기계번역의 실제

이 절에서는 한국어 분석기에 의한 입력 문장의 분석 결과를 입력으로 하여 영어 문장이 생성되는 과정을 설명하므로써, 이 논문에서 제시한 사전의 유용성을 보이고자 한다. 앞에서 언급하였듯이, 변환 사전의 내용은 구 구조에 대한 부분 구문 분석 결과이며 또 사전에 영어 문장 생성에 관한 정보가 아울러 수록되어 있으므로 영어 문장의 생성이 이루어 질 수 있다.

3.1 역어 선택의 단계

이 단계에서는 입력된 한국어 문장의 분석 결과를 보고 영어 문장 생성을 위한 역어를 찾아 올바른 위치에 두는 일을 한다. 그 과정은 다음과 같다.

【역어 선택 알고리즘】
입력 : 구문 분석된 한국어 분석 결과의 트리구조
출력 : 선택되어진 중간 형태의 영어문장
단계 1: tree가 비어 있으면 null 스트링을 반환;
단계 2: 현 tree에 해당하는 역어를 찾아서 저장
(Match_Phrase_Structure);
단계 3: 생성해야 할 하위구가 있는 동안
단계 3.1: 하위구에 대한 sub tree을 가져옴
(Traverse_Input_Parsertree);
단계 3.2: 하위구에 대한 tree로 Pass1을 순환적으로 실행하여 하위구에 대한 역어를 가져옴;
단계 3.3: 하위구의 역어를 이미 찾은 영어 단어들 내의 올바른 위치에 넣음;
단계 4: 역어를 반환;

이 논문에서 구축한 중심어 기반 변환 사전 형식은 하나의 일관된 형태, 즉 구 구조의 부분 파스 트리 혹은 전 파스 트리를 가지게 된다. 이러한 형태의 사전이 가지는 장점은 한꺼번에 처리해야 할 번역 단위에 대해서 구 구조 별로 일관성 있는 표현이 가능하고 그 처리 방법 또한 일관성이 있다.

알고리즘 Match_Phrase_Structure의 기능은 문장 분석에서 얻어진 결과에 대응되는 적절한 영어 표현을 찾아내는 것이다. 입력으로 받은 tree 형태의 부분 문장 분석 결과와 사전에 있는 역어 후보들을 보고 제약조건과 하위 어구의 최장 일치를 통하여 역어 후보들 중 가장 적절한 역어를 선택한다.

단계 2에서 역어를 선택하는 기준은 다음과 같다.

◆ **자질구조일치.**

하나의 표제어에 대해 문장성분, 품사, 서법등이 표현되어 나타난 부분 또는 전파서트리의 모든 자질구조와 구조적으로 최장일치하는 것을 선택하여야 한다. 여기서 구조적이란 뜻은 어순에 관계없이 분석된 구 구조가 일치한다는의미이다.

즉 사전에 1. S[a,b,c,d,e] #A;
2. S[a,b,c,d] #B;
3. S[a,b,c] #C; 와 같이 표제어 S에 대한 항목들이 있을 때(이 때 a,b,.등은 구의 자질구조) 입력으로 S[a,b,c,d,e]가 들어오면 1번과 2번이 입력에 subsume 되지만 1번이 최다일치를 만족시키므로 1번을 선택한다.

위의 Traverse_Input_Parsertree 알고리즘은, Match_Phrase_Structure에서 각 구 구조의 중심어의 역어가 먼저 찾아진 후 아직 역어가 선택되지 않은 성분들이 있을 때, 이들 하위구의 역어 선택을 위해 입력 프레임을 가져오는 역활을 맡은 루틴이다.

역어 선택 알고리즘의 단계3.2 는 이전에 찾은 하위구의 sub tree로 다시 Pass1을 recursive하게 수행시킴으로써 입력 tree를 깊이 우선 탐색한다. 이러한 처리 방법으로 접속 구조와 여러 내포구들을 찾아 내어 특별한 문법 규칙없이 번역문을 생성해 낼 수 있다.

3.2 영어 문장 생성의 단계

첫 단계에서 얻어진 정보들로부터 영어 문장을 생성한다. 이 단계에서는 선택된 역어들이 문법적인 문장을 이룰 수 있도록 시제, 인칭, 수의 일치 문제 뿐만 아니라 재귀 대명사의 처리 등의 여러 가지 사항을 고려한 변형을 한다.

【영문 생성 알고리즘】
입력 : 역어 선택 단계에서 구한 중간 형태의 영어 문장.
출력 : 일치에 의한 변환을 거친 영어 문장.
단계 1: 역어 선택 단계에서 받은 문장에 변형이 필요한 단어가 있는 동안
단계 1.1: 변형이 필요한 단어의 기본형을 보고 사전에서 찾는다.
단계 1.2: 변형이 필요한 단어가 명사이면
수의 일치에 관한 변환
단계 1.3: 변형이 필요한 단어가 동사이면
시제의 일치에 관한 변환
단계 1.4: 위에서 찾은 변형된 영어단어를 문장내 변형되기 전 단어와 바꾼다. 이때 필요한 영어 변형 사전에는 각각의 영어 단어에 대한 인칭, 수, 시제등의 정보가 있어야 한다.

4 실험결과

이 사전의 구축을 위해서 중심어 주도 구 구조 문법에 바탕을 둔 한국어 문장 분석기를 구현하였으며 또 구축한 사전의 유용함을 보이기 위하여 입력된 한국어 문장에 대한 구문 분석의 결과를 입력으로 하여 영어문장을 생성해 내는 부분을 구현하였다. 여기서 구축한 한국어 문장 분석기는 한국어 문장에 대해서 역어 선택에 필요한 정보를 얻는데 중점을 두었고, 영어 문장 생성 프로그램도 역어의 선택 및 생성이 이루어 질 수 있다는 사실을 입증하는데 중점을 두었다.

4. 1 사전의 구성

앞서의 한·영 기계번역을 다루는 논문에서 언급한 예문들과 중학교 영어 교과서에 나오는 문장들, 그리고 생활 영어 표현들로부터 대상 문장 및 어휘들을 취했다. 단어 '먹다', '목소리', 및 '마구'에 대해서 사전 정보의 예를 '시사영어사' 출판의 "한영 사전"에서 해당 용례를 취하여 제안된 사전 형식으로 표현한 예를 다음에 보였다.

(먹다):
[$먹^문장성분:서술어^보충어:[귀^품사:명사^문장성분:주어2]] #(^S *become deaf ^S2.A);
[$먹^문장성분:서술어^보충어:[귀^품사:명사^문장성분:주

어2]^수식어:[갑자기^품사:부사^문장성분:부사어]]#(^S *be struck deaf);
[$먹^문장성분:서술어^보충어:[귀^품사:명사^문장성분:주어2]^수식어:[일시적으로^품사:부사^문장성분:부사어]]# ^S *be deafened);
[$먹^문장성분:서술어^보충어:[{조반,아침}^품사:명사^문장성분:목적어]] #(^S *eat breakfast);
...
(목소리):
[$목소리^품사:명사^수식어:[크^문장성분:관형어]] #(a loud voice);
[$목소리^품사:명사^수식어:[굵^문장성분:관형어]] #(a full voice);
[$목소리^품사:명사^수식어:[{곱,아름답,달콤하}^문장성분:관형어]] #(a sweet voice);
...
(마구):
[$마구^품사:부사] #(recklessly)
...

이 사전에 들 정보는 다음과 같은 과정을 거쳐서 얻어진다. 먼저 한국어 문장을 분석하여 문장 성분 간의 관계를 규명한다. 그 결과를 대응되는 영어 문장과 비교하여 번역의 단위 즉 구 구조가 정해진다. 이 구 구조는 한국어 문장이 그 문장에 대응되는 영어 문장으로 번역이 될 수 있게끔 정해지는데, 단어, 숙어, 관용적 표현, collocation, 구 구조, 그리고 문장이 될 수 있다. 사전에는 구 구조 별로 1)여러 단어가 함께 한 단어처럼 쓰여서 영어의 한 표현으로 번역이 가능한 숙어, 2)한국어의 용언에 대하여 보충어에 따라서 역어가 결정되는 중심어-보충어 관계, 3)한 단어를 수식하는 단어에 따라 역어가 결정되는 중심어-수식어 관계, 4)체언이 조사와의 결합에 의해 역어가 결정되는 중심어-조사 관계, 그리고 5)번역된 후의 영어 문장의 문장 형식을 결정하는 연결어를 중심어로 한 관계등의 정보를 구조화된 부분 파스 트리에 표시하여 역어와 함께 중심어를 표제어로 하여 사전에 두게 하였다.

4. 2 제안된 사전의 제약점 및 앞으로의 연구 방향

이 절에서는 이 논문에서 제안한 사전으로 영어 문장을 생성할 때 문제가 되는 예들을 통해서 앞으로의 연구 방향을 제시하고자 한다.

1) 분석단계에서의 파서트리의 과다 생성

표층 구조만 사전에 수록하는 이 논문의 사전에서는 의미자질이나 의미표지등의 자질을 사용하지 않았기 때문에 한국어 분석단계에서의 생성가능한 부분파서트리가 아주 많아 진다. 따라서 의미자질까지를 포함하는 자질들을 사전의 정보로 주는 방법과 파싱의 전(앞)단계에서 사전의 정보를 이용 선택적 파싱을 하는 등의 연구를 하고 있다.

2) 형태소 분석시의 비효율성

근본적으로 Tabular방식의 형태소 분석과 같은 처리를 하고 있으며 이는 사전의 검색횟수를 줄이려는 [1,17]등의 연구결과를 충분히 이용하지 못했다. 그 이유는 근본적인 본 논문의 처리흐름이 그 방법들을 따를 수 없었기 때문인데 1)의 문제점을 해결하면서 자연스럽게 이용을 하게 될 것같다.

3) 관용적 표현의 처리 면에서

접니다.(누구냐에 대한 답으로) : It's me.
접니다.(전화 통화에서 본인을 바꾸어 달라고 했을때) : speaking

위의 예는 상황에 따라 역어가 다른 경우이다. 이 경우는 상황의미론 등에서 고려할 문제지만 역시 이 논문에서 제시한 정보로는 적절한 역어를 선택하기가 어렵다.

4) 사전 작성의 자동화 문제

비록 구 구조의 자질구조가 그대로 수록되어 사전의 항목에 대한 추가나 수정이 일관되게 수행이 되지만 그 작업자체는 번거로운 점이 많았다. 따라서 이러한 부분에서 어느정도 자동화시킬 수 있는 방법을 모색하고 있으며 앞으로의 사전은 인간의 개입이 보다 적은 쪽으로 가야만 실용적인 수준에까지 사전을 만들 수 있다.

지금까지 본 논문에서 미흡하거나 불완전한 부분과 앞으로의 대략적인 연구방향을 제시했거니와 사전 정보의 확충과 아울러 생략된 성분의 처리, 문단 단위의 의미해석과 같은 연구가 계속되어야 하고 또 이 정보들이 사전에 적합한 형태로 들 수 있어야 보다 적절한 영어 문장을 얻을 수 있다.

5 결 론

두 언어 간의 대응되는 번역의 단위가 단어, 숙어, 관용적 표현, 그리고 문장이 되기 때문에, 성격이 전혀 다른 영어 문장으로 변환을 위한 일관성 있는 정보를 사전에 두기가 어려웠다. 본 논문에서는 자연스러운 영어 문장을 얻기 위하여 중심어에 바탕을 둔 구 구조 변환 사전이 제안되었다.

이 사전을 다음과 같은 면에서 유용하다.

(1) 사전의 용도 면에서

기계번역의 단계 별로 필요로 하는 정보가 다르기 때문에 각 단계마다 다른 사전이 쓰였지만, 이 논문에서 제안한 변환 사전은 단순히 역어 선택을 위한 용도 뿐만 아니라 기계번역의 형태소 분석, 구문 분석, 그리고 역어 생성 단계에 까지 쓰인다. 따라서 단일 사전의 사용으로 인한 사전 관리와 유지가 보다 용이하다.

(2) 사전의 구성 형식 면에서

번역의 단위 별로 구문 분석의 결과를 부분 파스 트리의 형태로 구조화시켜 역어와 함께 저장하면서 그 형식을 같게 하였기 때문에 번역의 단위에 무관하게 정보 표현 형식이 같으므로 사전의 작성 및 정보의 추가 작업이 용이하다

(3) 부분 자유 어순에 대한 처리 면에서

구 단위 변환 사전이 표제어가 포함된 숙어를 평면 구조로 나타내지 않고 이를 파스 트리의 형태로 구조화하여 수록하기 때문에, 숙어를 구별해 내기 위한 별도의 처리 과정을 거치지 않고도 한국어 문장의 구 구조에 대한 역어 선택이 가능하다.

(4) 처리의 일관성 면에서

기존의 변환 사전에서 제공되는 정보로 역어를 선택하고자 하면 번역의 단위 별로 사전에 수록된 정보의 형태가 달랐으므로 처리 방법 또한 달라야 했다. 그러나 이 논문에서 제시한 사전의 정보로서 한 문장을 구성하는 구 구조가 어떤 유형이든지 혹은 구 구조를 수식하는 또 다른 구의 개수가 몇 개이든지 영향을 받지 않고 처리되므로 무제한 수식 관계에 따른 처리를 가능하게 한다. 이는 구구조의 순환적 관계구조와 사전의 어떠한 엔트리와도 구조적 일치를 보장함으로써 가능하다.

(5) 표제어 설정 기준의 명확성 면에서

기존의 사전에서는 어떤 구 단위에 대한 역어가 있을 때, 이를 어느 표제어 밑에 둘 것인가에 대해서 일관된 규칙을 부여하지 않았다. 표제어 설정 기준의 명확성은 사전 정보의 중복성을 피할 수 있게 하며, 이로 인해 사전 검색및 사전의 정보의 일치성과 유지, 보수에 보다 효율적이다.

(6) 구 구조를 구성하는 단어들의 융통성 면에서

영어의 한 표현에 대해서 한국어의 여러 표현들이 대응될 수 있다. 이러한 한국어 표현들이 구 구조의 구성성분일 경우, 이들에 대한 역어 정보를 모두 개별로 수록한다는 것은 비효율적이다. 따라서 이 논문에서 제시한 사전은 이러한 단어들을 같은 역어를 생성하도록 집합으로 묶어 처리하였다. 이러한 정보의 표현은 기존의 의미 분류의 고정된 역어 선택에 따른 문제점을 해결하는 동시에 같은 의미의 여러 단어에 대한 정보 수록의 중복성을 피할 수 있는 장점을 아울러 가진다.

(7) 자연스러운 영어 문장 생성 면에서

한국어로 된 문장을 입력으로 받아 번역된 영어 문장을 얻으려 할 때, 한국어의 문장에 쓰인 단어의 의미, 단어들의 나열 순서, 어미 변화 및 조사의 쓰임에 대한 분석도 물론 중요하지만, 그 문장이 나타내고자 하는 의미만 정확히 파악된다면 그에 대응되는 영어 문장을 얻을 수 있어야 한다. 그러나 모든 문장을 번역의 단위로 고려할 수는 없기 때문에, 그에 근접하는 효과를 얻기 위해서는 영어 표현에 대응되는 한국어 표현을 가급적 길게 잡는 것이 한 방법이 될 수 있다.

이 논문에서 제안된 사전에는 의미 표지와 같은 어휘들의 분류에 대한 정보등을 감안하지 않고 오직 표층 구조에 나타난 정보만을 취급하였으므로, 영어 문장 생성시 부사구와 같은 필수적이 아닌 수식어 성분들의 위치 정보를 두지 않았다. 그래서 이 사전이 실용화 되게 하기 위해서는 더 많은 용례에 대한 사전 정보의 확충과 아울러 영어 문장 생성에 대한 계속적인 연구가 필요하다. 아울러 영어 문장과 한국어 문장을 이용하여 자동적으로 변환 사전이 구성될 수 있도록 하는 연구가 필요하며 이를 이용하여 실질적으로 사용가능한 사전의 구축이 가능해 지리라 본다.

참 고 문 헌

[1] 강승식,"음절정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 박사학위 청구논문, 1993

[2] 한국과학재단,"자연언어처리의기초연구",KOSEF860207,1987

[3] 김나리,김영택,"Collocation정보에 기반한 한-영 트랜스퍼 사전의 구성에 관한 연구," 서울대학교 컴퓨터공학과 석사 학위 청구논문,1992

[4] 박원철,"형태론적 연산으로서의 통합", 서울대학교 석사학위 청구논문,1989

[5] 서 영훈,"의미정보를 이용하는 중심어 주도의 한국어 파싱",서울대학교 박사학위 청구 논문,1991

[6] 양재형, "HPSG에 기반한 한국어 분석기의 연구", 서울대학교 석사학위 청구논문, 1990

[7] 옥철영, "한영 기계 번역을 위한 구단위 번역 사전", 서울대학교 박사학위 청구논문, 1993

[8] 이상조, "기계번역 시스템을 위한 사전구성," 정보과학회지,Vol.7,No.6,pp.25-30,1989

[9] 이상조,"한영 번역을 위한 전자 사전 구성에 관한 연구," ETRI 위탁과제 연구보고서,1989.7

[10] 이상조,"한국어 자연어 인터페이스를 위한 사전 구성에 관한 연구,"ETRI, 위탁과제 연구보고서,1991. 6

[11] 이상조,"한국어 자연어 인터페이스를 위한 사전구성에 관한 연구(II)," ETRI 위탁과제 연구보고서,1992.6

[12] 이호석, 김영택, " 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전" 한국정보과학회지논문지, vol 20, No.7, 93.7월 pp 976-986

[13] 임수연,김권양,이상조,"한국어 내포문 해석을 위한 사전의 구성", 한국 정보과학회 인공지능 연구회 춘계학술 발표 논문집,1990,pp:151-162

[14] 장석진,"한영 동사의 하위범주화와대응에 관한 연구", 제1회 기계번역 workshop발표 논문집,1989,pp:169-173

[15] 장석진, "정보기반 한국어 문법", 언어와 문법, 1993

[16] 정희성,"한영 기계 번역 시스템의 한국어 파서 및 사전 시스템의 개발", 국책연구개발 사업과제 보고서, 8SC15004805F, 과학기술처,1989

[17] 최재혁, 이상조,"양방향 최장 일치법을 이용한 한국어 형태소 분석기", 한국 정보과학회 춘계학술발표논문집,1993,pp:769-772

[18] Yoon,S.H.,"Idiomatical and Collocational Approach to Machine Translation," *Proceedings of the 2nd Pacific Rim International Conference on Artificial Intelligence*, pp. 49-63, Seoul,Korea,1992 Vol.11, No.1,pp. 1-17,1985