적절한 동사 대역어 선택을 위한 한영 변환 사전 구성

송 정 근
서 울 대 학 교 국 문 과
jksong2@snu.ac.kr

The Composition of Korean-English Transfer Dictionary for Proper Selection of Verb Translation

Jung-Keun Song
Dept. of Korean language & Literature, Seoul National University

요 약

기계번역이 인간의 언어 능력을 기계로 구현한다는 점에서 전산학적 성격이 강하다면, 변환 사전은 인간의 어휘부(lexicon) 정보를 그대로 기계에 표상한다는 점에서 언어학적 성격이 강하다. 여기서는 다양한 어휘부 정보 중에서 한영 기계번역에서 필요한 언어학적 정보를 추출하고 이러한 정보를 바탕으로 적절한 동사 대역어 선택을 위한 변환 사전의 모형을 만들어 보고자 하였다.

한영 기계번역에서 적절한 동사 대역어 선택의 어려움은 한국어 동형어 처리 문제와 한국어에서는 포착되지 않지만 영어로 번역하는 과정에서 발생하는 영어 표현의 특수성 때문에 기인한 것으로 볼 수 있다. 이 논문에서는 이러한 문제를 논항과 문법 형태소, 선택제약, 개별 어휘 등의 기초적인 언어학적 개념을 이용한 변환사전을 통해 해결한다. 또한 동사 대역어 선택에 영향을 미치는 이러한 개별적인 요인들은 실제 변환사전의 기술에 있어서는 복합적으로 적용됨을 동사 '먹다'의 기술을 통해 확인할 수 있다.

1. 서론

일련의 자연언어 처리(natural language processing)는 인간의 '언어'를 다룬다는 측면에서 언어학과 관련이 있으며 인간의 언어를 '기계'에 표상(representation)한다는 측면에서는 전산학과 관련을 맺는 통합 학문적 성격을 갖는다. 따라서 전산학이나 언어학 어느 한쪽의 독자적인 연구만으로는 통합 학문적 성격을 갖고 있는 자연언어 처리를 성공적으로 수행할 수 없다. 단순한 규칙이나 수학적, 통계적 확률을 바탕으로 하는 기계적 접근법이나 인지적이고 추상적인 논의를 바탕으로 한 언어학적 접근법은 그 자체로서의 가치는 확보하고 있다고 할 수 있을지 모르나 자연언어 처리의 관점에서는 모두 한계성을 갖고 있는 것이다.

자연언어 처리의 대표적인 응용분야인 기계번역(machine translation)에서도 이러한 언어학과 전산학의 통합적 성격은 잘 드러난다. 일반적인 변환(transfer) 방식의 기계번역은 개별 언어를 분석(parsing), 생성(generation) 그리고 두 언어 사이의 어휘적 문법적 대응 관계를 포착해 주는 변환 단계로 크게 나누어 볼 수 있다. 기계번역을 위한 이러한 각각의 과정들은 인간의 언어 능력을 기계적으로 구현하려는 목표를 갖고 있다고 할 수 있기 때문에 전산학적 성격이 강하다고 할 수 있다. 하지만 구현을 위한 바탕은 언어학적 직관이나 이론이다. 분석을 위해 전산학자가 사용하고 있는 문법이나 두 언어 사이의 관계를 포착해주는 변환 규칙, 생성을 위한 문장 생성 정보는 모두 언어학적인 것이라고 할 수 있다.

여기서 언어학이 제공하는 정보들은 순수한 언어학적인 것일 수도 있고 기계적 응용을 위해서 기존의 이론이 수정된 것일 수도 있으며 완전히 기계를 위해 새로 고안된 이론일 수도 있다. 촘스키(chomsky)의 변형 생성 문법(transformational generative grammar)이 기계번역에서 그대로 사용될 수도 있겠지만 TAG(Tree-Adjoining Grammar), CCG(Combinatory Categorical Grammar), LIG (Linear Indexed Grammar)와 같은 새로운 문법들이 이용되기도 하며 기존의 문법을 필요에 따라 변형하여 사용할 수도 있는 것이다. 즉 기계번역에서 문법은 번역 단계의 필요에 의해 선택되거나 고안되는 것이다.

하지만 변환 과정에서 사용되는 변환 사전의 경우는 반대의 경우라고 할 수 있다. 기계번역에서 사용되는 변환 사전은

기계가 수행하는 번역을 위해 취사 선택되는 정보가 아니며 기계번역을 위한 필수적인 정보이다. 예를 들어 '나는 학생이다.'라는 문장을 번역하기 위해서는 이들이 어떤 방식으로 분석되고 변환되든 형태소 '나, 는, 학생, 이, 다'에 대한 각각의 영어 대역어와 올바른 대역어 선택을 위한 언어학적 정보가 있는 것이다. 이러한 정보는 매우 확정적이고 필수적인 것이어서 기계는 어떤 방식으로든 이 정보를 이용해야 한다. 더 나아가 변환사전에서 기술된 이러한 정보를 기계가 가장 잘 이해할 수 있는 여러 가지 방법론이 필요에 따라 선택되는 것이다. 결국 기계번역에서 변환사전은 언어학적인 측면에서의 접근이 가장 필요한 분야인 것이다.

여기서는 한영 기계번역에서 한국어 동사에 대한 적절한 영어 대역어 선택을 위하여, 변환사전에서 제공해야 하는 정보와 그 정보를 기술하는 구체적인 방식에 대해 논의하고자 한다. 기존의 연구에서는 규칙에 기반한 기계번역(rule-based machine translation, RBMT)이든 예문에 기반한 기계번역(example-based machine translation, EBMT)이든 '연어' 개념(공학적인 측면에서)을 통해 동사 중의성을 해소하고 적절한 대역어를 선택하는 방법론에 대한 연구가 많이 이루어 진 바 있다.[1,2,3,4] 그러나 언어학적으로 대역어 선택에 영향을 미치는 요인은 개별 어휘들 간의 연어 관계 외에도 다양하며, 이러한 요인들은 현재 자연언어 처리 수준에 맞추어 응용 가능한 형식으로 재정리되어야 한다. 변환 사전은 기본적으로 인간의 어휘부(lexicon)을 기계에 표상하는 작업으로 볼 수 있는데, 인간의 방대한 어휘 능력을 숙어 혹은 연어와 같은 형태로 모두 표상할 수는 없는 노릇이다.

또한 이러한 언어학적 연구를 통해 정리된 내용들을 실제 기계번역에 적용하여 그 타당성을 검증 받고자 한다. 자연언어처리는 전형적인 응용 분야로서 실제적인 실용가능성으로 그 존재 가치를 확보하는 것이기 때문이다. 여기서는 실험의 편의를 위해 구체적으로 동사 '먹다'의 변환 사전에서의 기술 방식과 이들의 번역 결과를 보이도록 하겠다[1].

2. 동사 대역어를 결정하는 요인

기계번역에서 영어 대역어 선택은 기본적으로 국어의 동형어[2] 처리 문제와 관련된다. 형태상으로는 동일하지만 의미나 용법이 서로 다른 동형어들은 각각의 경우를 별개의 단어로 취급해야 하며 별개의 영어 대역어로 번역되어야 한다. 사전마다 약간의 차이는 있지만 흔히 국어 사전에서는 이들을 어깨 번호를 써서 구별해 놓고 있다. 다음의 경우를 살펴보자.

(1) **먹다**[1] ① 음식물을 입을 통하여 넘기다. ㉠씹어서 음식을 뱃속에 들여보내다. ¶아이들이 과자를 잘 먹는다./그는 저녁밥을 먹고 텔레비전을 켰다. ㉡액체로 된 것을 마시다....
먹다[2] (귀가) 소리를 듣지 못하게 되다. ¶할머니는 귀가 먹어서 손자가 하는 말을 당최 알아들을 수 없었다....
먹다[3] 앞의 동사가 나타내는 행위를 낮추던가 얕잡는 뜻을 담아서 강조함. '실컷', 또는 '겨우 ~ 하다'의 뜻. ¶그녀의 남편은 그녀를 일생 동안 부려먹었다./ 서울서 장사해 먹기 힘듭니다...

(1)의 예에서 먹다[1]과 먹다[2], 먹다[3]은 형태상 동일하지만 서로 별개의 단어들이다. 타동사 먹다[1]과 자동사 먹다[2]는 의미 유연성(motivation)을 전혀 찾을 수 없을 뿐만 아니라, 실제 용법에 있어서도 취하는 논항의 수나 종류에서 차이를 갖는다. 보조 동사의 용법을 보이는 먹다[3]의 경우도 역시 먹다[1], 먹다[2]와는 다른 단어로 취급해야 한다. 종이 사전에서는 (1)에서와 같이, '먹다'를 통사 의미론적 성격에 따라 분리하여 기술해 놓기만 하면, 사전 사용자가 자신이 찾고자 하는 '먹다'가 위의 어느 경우인지를 확인하고 필요한 정보를 선택하면 된다.

그러나 기계번역에서는 종이사전에서와 같이 각각의 경우를 나누어 기술하는 것만으로는 부족하다. 기계는 번역하고자 하는 한국어 문장에서 사용된 '먹다'가 사전에서 기술된 먹다[1], 먹다[2], 먹다[3] 중 어떤 것인지 판단하고 선택할 수 없기 때문이다. 따라서 변환 사전은 기본적으로 한국어 동형어들의 번역에서 이들의 올바른 대역어 선택을 위한 정보를 제공해 줘야 하는 것이다.

한영 기계번역에서는 이러한 국어 동형어 문제 외에도 한국어를 영어로 번역하는 과정에서 영어 어휘가 갖는 특수성도 고려해야 한다. 예를 들어 '밥을 먹다, 사과를 먹다, 물을 먹다' 등의 예에서 '먹다'는 모두 (1)의 먹다[1]의 의미로 사용되고 있으며 동일한 영어 대역어로 번역될 것으로 기대할 수 있다. 그러나 영어에서 '물을 먹다'의 '먹다'는 다른 먹다[1]과 동일하게 'eat'로 번역되는 것이 아니라 'drink'로 번역되어야 한다. 이것은 한국어에서는 포착할 수 없는 영어가 갖는 특수성으로 볼 수 있다. 즉 먹다[1]로 사용된 경우라도 대상이 액체일 때 대역어는 'drink'로 달라지는 것이다. 이러한 언어 개

1) 실험하게 될 번역기는 (주) 언어와 컴퓨터에서 만들고 있는 한영 번역기로 분석부는 강승식 교수(국민대), 변환부는 이하규 교수(성공회대)가 설계, 제작한 것이다. 여기서 우리의 논의는 기계번역에 대한 전반적인 것이 아니라, 언어학적 관점에서 변환 사전의 구성에 관한 것으로 한정한다.

2) 국어는 한자어를 표시하지 않으면 동형이의어와 동음이의어를 구별하기 어려우며, 사전 처리에서는 동음이의어와 다의어의 구별도 쉽지 않다. 여기서는 형태를 같지만 의미가 다른 모든 단어들을 동형어라고 부르기로 한다.

별적(language specific) 정보들은 어떤 언어 쌍간의 번역이냐에 따라 추가되거나 삭제되어야 하는 것들이다. 그러나 고품질의 기계번역을 위해서는 이러한 정보 역시 변환 사전에서 충실히 제공해야 하는 정보임은 물론이다.

결국 한국어 동형어들의 올바른 대역어 선택을 위한 정보나, 영어가 갖는 특수성으로 인한 영어 대역어 선택에 관한 정보들은 변환 사전에서 제공해야 하는 기본적이며 중요한 정보들이다. 여기서는 한국어 동사의 영어 대역어를 결정하는 요인으로 동사가 취하는 논항(argument)과 논항의 문법적 기능을 표시해 주는 문법 형태소(조사, 어미), 동사의 선택제약(selectional restriction) 정보, 개별 어휘 등이 이용될 수 있음을 보이고, 변환 사전에서 이러한 정보를 어떤 방식으로 기계에 적용할 수 있는지 구체적으로 살펴보도록 하겠다[3].

2.1 논항과 문법 형태소

한 문장의 서술어(predicate)인 동사나 형용사는 자신의 의미를 완전하게 하기 위하여 자신의 어휘개념구조 속에 있는 '의미적 참여자'를 통사적으로 취하여 문장을 형성한다. 여기서 '의미적 참여자'는 서술어의 의미를 완전하게 기술해 주는 의미론적인 역할을 하는 동시에 통사론적으로도 서술어의 성분들이 된다. 일반적으로 이렇게 서술어에 요구되는 성분들은 필수적인 성격을 지닌 성분과 필수적이지 않고 부가적인 기능만을 하는 성분으로 나누어 전자를 논항(argument)[4], 후자를 부가어(Adjunct)로 명명하게 된다[5].

논항은 개별 서술어의 의미 특성에 따라 다양한 양상으로 나타날 수 있다. 하나의 서술어가 하나의 논항만을 취할 수도 있고 둘 혹은 그 이상의 논항을 취할 수도 있는 것이다. 흔히 1항(혹은 1가) 동사, 2항(혹은 2가) 동사라는 명명은 서술어가 필요로 하는 논항의 양상을 항가(valency) 개념으로 포착하여 설명한 것이다.

논항이 어떤 문장에서 갖는 의미론적 역할을 의미역(θ-role, thematic-role)이라고 하는데, 어떤 서술어가 필요로 하는 논항과 그 논항의 의미역이 표시된 틀을 의미역틀 혹은 의미역 격자(θ-grid)라고 한다. 예를 들어 (1)의 먹다[1]은 <행동주, 대상>이라는 의미역틀을 갖는 것이다. 그러나 의미역틀 정보만으로는 한국어 어휘가 갖는 통사 어휘적 특징을 모두 나타낼 수 없기 때문에 논항들의 통사적 기능이나 논항의 의미 특성을 부가적으로 표시하여 확대된 의미역틀을 만들 수도 있다.[8]

언어학적으로 동형어들은 각각 취하는 논항의 수나 논항의 의미역들이 다르기 때문에 이러한 확대된 의미역틀을 이용하게 되면 동형어들을 보다 정확하고 명시적으로 구분할 수 있다. 기계번역에서도 동형어 처리를 위해 이러한 언어학적 정보를 이용할 수 있다. 그러나 변환 사전에서 동형어 처리를 위해 동사 혹은 형용사의 의미역틀을 비롯한 다양한 언어학적 정보를 모두 이용할 수는 없다. 개별 어휘들의 언어학적 정보를 변환 사전에서 모두 기술하는 것은 현실적으로 어려울 뿐만 아니라 현재의 자연언어 처리 수준을 고려한다면 실제로 기계가 처리할 수 있는 정보는 매우 한정되어 있기 때문이다.

이러한 문제 해결을 위해서는 이용 가능한 언어학적 정보만을 선별적으로 변환 사전에 기술해야 한다. 효율적인 변환 사전의 기술을 위해 선택할 수 있는 방법은 동사의 영어 대역어 선택에 영향을 미치는 논항과 그 논항의 문법적 기능을 알 수 있는 조사나 어미 정보를 제공하고 이에 대응하는 대역어를 제시하는 방식이다[6].

---

3) 언어학적인 분류 체계에서 통사론적인 결합은 의미 결합 양상에 따라 구성 성분의 의미가 전체 의미와 투명한 자유 결합(free combination)과 불투명한 비자유 결합(non- free combination)으로 구분할 수 있다.[10] 자유 결합은 일반적인 통사론적, 의미론적 원리에 따라 설명 가능한 결합이고 비자유 결합은 그렇지 못한 결합이라고 할 수 있다. 따라서 기계번역에서 자유 결합은 기존에 소개된 언어학적인 규칙을 통해 분석, 변환, 생성이 가능한 구성인 반면, 비자유 결합은 사전을 이용한 어휘 정보를 이용해야만 하는 구성이라고 할 수 있다. 예를 들어, 자유 결합인 'X을 먹다'라는 결합은 동사 '먹-'가 영어의 'eat'에 해당하고 대상역의 논항으로 '실체성'을 갖는 명사를 갖는 다는 정보만을 알고 있으면 모든 구성을 처리할 수 있다. 하지만 구성 성분의 의미와 전체 의미가 불투명한 '미역국을 먹다, 머리를 굴리다, 판결을 하다/내리다'등은 사전에서 이들의 대역어들을 개별적으로 제시해야 한다. 변환 사전 구성에 대한 논의는 이러한 자유 결합과 비자유 결합의 분류 체계에 따라 이루어 질 수 있겠으나 여기서는 과도한 언어학적 분류 체계를 피하기 위해 이들의 구별 없이 기계 번역에 이용되는 개념을 중심으로 설명한다.

4) 논항(argument)이라는 개념이 반듯이 서술어가 필요로 하는 성분만을 말하는 것은 아니다. 일반적으로 자신이 속한 구성의 통사적 성격을 결정하는 요소인 핵(표제 혹은 머리 Head)이 필수적으로 필요로 하는 성분을 논항(argument)이라고 할 수 있다. 따라서 명사도 논항을 취할 수 있으므로 핵이 될 수 있다.

5) 언어학적으로 '필수적인' 성분과 그렇지 않은 성분이 명확하게 구분되지 않는 경우도 있을 수 있다. 최근의 논의에서는 이러한 이분법적 구분의 문제점을 지적하고 '정도성'의 문제로 설명하는 경우도 있다.

6) 여기서 논항과 그 의미역 정보를 이용하지 않고 논항과 조사 정보를 이용하여 변환 사전을 구성하는 이유는 분석(parsing) 단계를 고려했기 때문이다. 현재 분석 단계에서는 논항과 그 논항의 통사적 성분까지는 포착할 수 있으나 논항의 의미역 정보와 같은 의미 분석은 이루어지지 못하고 있다.

(1)에서 예를 들었던 '먹다'의 경우를 살펴보자. 보조 동사 먹다[3]은 일반 동사인 먹다[1], 먹다[2]와는 달리 다른 용언을 선행하는 특징을 갖는다. 체언을 논항으로 취하는 먹다[1], 먹다[2]와 용언을 논항으로 취하는 보조 동사 먹다[3]은 이들이 취하는 논항과 조사, 어미 정보만으로도 구별될 수 있는 것이다. 또한 먹다[1]과 먹다[2]는 취하는 논항이 2개로 같지만 이 논항들의 역할을 동일하지 않다. 먹다[1]이 취하는 두 개의 논항은 문장 성분 상 주어와 목적어로 설정할 수 있지만 먹다[2]의 논항은 모두 주어로 설정되어 이중 주어문을 형성하는 것으로 볼 수 있기 때문이다.[7] 결국 이러한 정보들은 다음과 같이 거칠게 정리될 수 있다.

(2) *A:NN,가 *B:NN,를 먹다 = *A eat *B
*A:NN,가 *B:NN,가 먹다 = *A become deaf
*A:VV,-어 먹다 = *A:VV

기계번역을 위한 이러한 각각의 논항과 논항의 문법적 기능을 표시하는 문법 형태소 정보는 다음과 같은 관용구의 동사 대역어 선택에서도 이용될 수 있다.

(3가) 영희가 미역국을 먹었다.
(3나) 영희가 이번 시험에서 미역국을 먹었다.

(3가)의 경우는, '미역국을 먹다'가 축자적 의미로 사용된 예이고 (3나)는 관용적 의미로 사용된 예이다. 한국어 화자가 문맥적 상황을 알지 못하는 상황에서도 (3가, 나)의 의미 차이를 구별할 수 있다. 이것은 '이번 시험에서'라는 부사구 때문이다. 즉 (3나)와 같이 '미역국을 먹다'가 관용적 의미로 사용된 경우, '이번 시험에서'는 '미역국을 먹다'의 의미 해석에 결정적인 역할을 하고 있는 것이다.

따라서 아직 의미 분석이나 화용론적인 의미 해석이 자연언어 처리에서 이용되지 못하는 상황에서 (3나)와 같은 문장은 '먹다'가 취하는 논항을 통해 대역어가 결정되어야 한다. 즉 '미역국을 먹다'가 '시험에서, 승진에서'와 같은 논항을 취할 경우, 그 의미가 'fail'로 대역되어야 하는 것이다.[8]

결국 논항과 논항의 문법적 기능을 나타내는 문법 형태소 정보는 동형어들의 영어 대역어 선택과 관용구의 의미 해석에서 중요한 역할을 할 수 있다고 할 수 있다.

2.2 선택 제약(selectional restriction)

단순히 논항의 유무나 조사, 어미 정보로 한국어 동사의 영어 대역어를 결정하지 못하는 경우도 있다. 예를 들어 논항과 문법 형태소 정보만으로는 위에서 언급한 '물을 먹다'와 같은 문장의 올바른 번역을 기대할 수 없을 뿐만 아니라, '영희네 집에서 미역국을 먹었다.'에서와 같이 단순히 논항의 유무만으로는 올바른 대역어를 선택할 수 없는 경우가 존재한다. 이러한 문제를 해결하기 위해서는 논항에 대한 보다 구체적인 정보가 필요하다고 할 수 있는데, 선택제약이 이러한 문제의 해결책이 될 수 있다.

변형문법에서 서술어는 선택제약(selectional restriction)을 통해 자신이 취하는 논항의 의미 자질(semantic feature)을 명세(specification)하게 된다. 예를 들어 '밥을 먹다'라는 구성에서 논항 '밥'은 서술어에 대해 대상에 해당하는 의미론적인 역할을 하는데, '구체성'이라는 의미 자질을 갖는 다른 어떤 명사도 그 의미론적 역할을 대신할 수 있다. 즉 동사 '먹-'은 대상에 해당하는 의미론적 역할을 하는 '구체성'이라는 의미 자질을 갖는 명사를 논항으로 취할 수 있다고 하는 선택제약을 갖고 있다고 할 수 있다.

이러한 선택제약은 변환 사전에서 올바른 영어 대역어을 선택하는 정보로 이용될 수 있다. 위에서 예를 들었던, '물을 먹다'의 경우를 살펴보자. '물을 먹다'에서 '먹다'는 다른 먹다 1의 경우와는 다르게 'drink'로 번역되야 한다. 그러나 '물' 대신에 '액체'라는 의미 자질을 갖는 다른 명사가 논항으로 올 경우에도 대역어는 동일하게 'drink'가 선택되어야 한다. 만약 변환사전에서 '먹다'가 '물'을 논항으로 취할 경우 'drink'로 번역된다는 정보만을 담고 있다면 '액체'의 의미자질을 갖는 다른 모든 명사를 사전에서 개별적으로 기술해줘야 하는 문제를 갖는다. 그러나 의미자질을 이용하게 되면 이러한 문제가 해결될 수 있다.

(4) *A:NN,가 *B:NN,를%액체 먹다 = *A *drink* *B

단순하게 논항의 유무만으로 영어 대역어를 결정할 수 있는 경우가 있는 반면, (4)에서와 같이 선택제약을 통해 논항의 의미 자질을 이용하여 올바른 영어 대역어가 결정될 수도 있는 것이다.

---

7) '영희는 귀가 먹었다.'라는 문장을 이중 주어문으로 분석하지 않고, 어휘부(lexicon)에 '귀가 먹다'가 하나의 단어처럼 등재되어 있고 '귀가 먹다' 전체가 '영희는'을 논항으로 취하는 것으로도 볼 수 있다. 그러나 '귀가 먹다'와 같은 절을 분석(parsing) 단계에서 한 단위로 분석하기 쉽지 않고 변환 사전에서 이러한 절을 표제어로 설정하여 기술해야 하는 문제가 발생한다.

8) '영희가 철수네 집에서 미역국을 먹었다.'와 '영희가 시험에서 미역국을 먹었다.'의 두 문장에서 '철수네 집에서'와 '시험에서'가 동일한 문법적 지위를 갖는 것은 아니다. 전자의 경우는 전형적인 부가어로 파악할 수 있지만 후자의 경우는 '미역국을 먹다' 관용구 전체가 하나의 논항으로 '시험에서'를 취하는 것으로 파악하는 것이 옳을 듯 싶다.

이러한 선택제약을 이용하기 위해서는 모든 명사에 의미 자질이 표시되어 있어야 한다. 즉 '물'이라는 명사에 '액체'라는 의미 자질이 표시되어 있지 않다면 (4)와 같은 선택제약 정보가 이용될 수 없기 때문이다. 따라서 선택제약을 이용한 대역어 선택을 위해서는 기본적으로 한국어에서 사용될 수 있는 의미자질들과 의미자질로 기술된 명사 사전이 필요한 것이다.

(5) 실체명사 entity
유정물 organism
신 spiritual thing
동물 animal
물고기 fish
새 bird
곤충 insect
사람 person
남성 male
여성 female
인공유정물 group

(6) 배
% 신체일부 교통수단 과일 계량
?신체일부 belly:NN
?교통수단 ship:NN
?과일 pear:NN
?계량 double:NN

(5)는 우리가 실험하게 될 한영 번역기에서 사용하고 있는 92개의 의미자질 중 일부를 보인 것이고, (6)은 그 의미자질이 부착된 명사 사전의 예이다. 변환사전에서 선택제약을 이용해 올바른 영어 대역어 선택을 하기 위해서는 기본적으로 한영 번역에 사용되는 의미 자질 집합(set)을 확정하고 각각의 명사에 이러한 의미 정보를 모두 달아 주어야 한다. 여기서 사용될 의미자질의 수는 필요에 따라 늘어나거나 줄어들 수도 있다.

명사의 의미 자질을 이용한 대역어 선택은 기본적으로 자연언어 처리의 기초적인 의미 분석과 의미 처리를 도입하는 것이다. 기계적인 의미 처리는 추상적인 인간의 의미 체계를 기계에 표상(representation)해야 하는 단계로 볼 수 있다. 따라서 이러한 작업은 전반적인 한국어의 의미 체계와 개별 단어들의 의미 특성을 일정한 체계 안에서 기술한 정보를 바탕으로 해야 하며, 앞으로 많은 연구가 필요한 분야라고 할 수 있다.

'물을 먹다'와 같이 '먹다'가 특정한 의미 자질을 갖는 명사를 논항으로 취했을 때, '먹다'의 영어 대역어가 결정되는 경우과 같은 예들이 있을 수 있다.

(7) 먹다
*A:NN,가 *B:NN,를%액체 먹다 = *A drink *B
*A:가 *B:를*%돈 먹다 = *A embezzle *B (예: 그가 뇌물을 먹었다.)
*A:가%도구 먹다 = *A bite (예: 톱이 잘 먹는다.)
*A:가 *B:를%순위 먹다 = *A become *B (예: 내가 일등을 먹었다.)

2.3 개별 어휘

한영 기계번역 과정에서 개별 어휘들이 공기(共起) 관계에 있는 다른 어휘의 대역어 선택에 영향을 미치는 경우는 매우 다양하다. 기존의 연구에서는, 공학적 측면에서의 연어(collocation)[9] 혹은 숙어(관용구, idiom)개념을 설정하고[5,6], 문장 성분 별로 주어나 목적어, 부사어 등이 서술어인 동사나 형용사의 대역어 결정에 영향을 미치는 경우에 대해 다양하게 연구된 바 있다.

(8) 먹다
*A:가 욕:를 ~ = *A be scolded
*A:가 겁:를 ~ = *A be frightened
*A:가 충격:를 ~ = *A be shocked
*A:가 마음:를 ~ = *A make up mind
*A:가 나이:를 ~ = *A grow older
*A:가 더위:를 ~ = *A be affected by the heat

(8)에서 제시된 예들은 전형적으로 '욕, 겁, 충격, 마음, 나이, 더위' 등의 어휘로 인해 동사 '먹다'의 대역어가 결정되는 경우이다. 국어 사전에서 이들은 일반적으로 (1)에서 살펴본 먹다1의 다의어로 처리한다. 그러나 변환 사전에서는 개별 어휘와 이들로 인해 대역어가 결정되는 각각의 경우를 기술해 줘야 하는 것이다.

개별 어휘를 통해 동사의 대역어가 결정되는 경우는 한국어 관용절의 처리에도 적용된다.

(9) *A:가 *B:와 한솥밥을 먹다 = *A break bread with *B

한국어 관용절에 해당하는 '한솥밥을 먹다'의 의미는 개별 어휘인 '한솥밥, 먹다'의 의미와는 전혀 관계없는 '함께 생활하며 지내다' 정도의 의미를 갖는다. 이들을 처리하기 위해서는 동사 '먹다'가 '한솥밥'을 목적어로 취한 경우, 이들 절 전체의 영어 대역어를 변환 사전에서 제공해 주면 되는 것이다.

그러나 모든 경우가 이렇게 간단히 개별 어휘로 인해 대

9) 최근 언어학에서도 '연어(collocation)'에 대한 관심이 높아지고 있다. 그러나 연어라는 용어가 공학과 언어학에서 서로 다른 개념으로 사용되고 있기 때문에 이들을 동일한 선상에서 비교할 수 없는 문제점이 있다. 게다가 언어학적 연어의 개념이나 용례들이 명확하지 않은 문제점도 지적될 수 있다. 궁극적으로 자연언어 처리가 공학과 언어학의 공동작업이라는 점을 고려한다면 동일한 용어에 서로 다른 개념을 부여하는 것은 바람직하지 못할 것으로 판단된다. 여기서는 혼란을 막기 위해 연어, 관용구 등의 용어는 사용하지 않도록 하겠다.

역어가 결정되는 것은 아니다. 위에서 살펴본 '시험에서 미역국을 먹었다'와 같은 관용절의 경우를 다시 살펴보자.

(10가) *A:가 시험에서 미역국을 먹다 = *A fail in exam
(10나) *A:가 *B:에서 미역국을 먹다 = *A fail in *B
(10다) *A:가 *B:에서%장소 미역국을 먹다 = *A eat seaweed soup in *B

'철수가 시험에서 미역국을 먹었다'라는 문장을 올바로 번역하기 위해서는 변환 사전에서 '시험'과 '미역국'이 '먹다'의 성분이었을 때 '미역국을 먹다'가 'fail'이라는 정보를 우선 주어야 한다(10가). 그러나 '미역국을 먹다'가 '실패하다'의 의미로 사용될 수 있는 경우는 시험에만 국한되는 것은 아니다. '철수는 입시에서 미역국을 먹었다, 철수는 승진에서 미역국을 먹었다.' 등과 같이 부사어 자리에 올 수 있는 어휘들은 다양할 수 있다. 따라서 (10가)는 (10나)와 같이 수정하여 기술할 수 있다. 그러나 (10나)와 같은 변환사전 기술은 '철수가 영희네 집에서 미역국을 먹었다.'와 같은 문장 역시 '미역국을 먹다'를 '실패하다'로 처리하게 된다. '미역국을 먹다'의 경우, 부사어 '*B:에서'의 '*B' 자리에 올 수 있는 어휘에 따라 '미역국을 먹다'의 의미가 달라지는 것이다. 이러한 문제는 (10다)와 같은 정보를 변환 사전에 추가하여 해결할 수 있다. 부사어로 사용되는 어휘의 의미 자질이 '장소'일 경우와 그 이외의 경우를 구분하여 처리하는 것이다.

즉 (8, 9)와 같이 개별 어휘를 통해 공기하는 단어나 절의 대역어가 결정되는 경우는 변환 사전에서 간단히 처리할 수 있다. 그러나 (10)와 같이 여러 가지 요인을 고려해야 올바른 영어 대역어를 선택할 수 있는 경우는 논항과 조사, 어미 정보, 선택제약과 같은 요인들을 적절하게 이용하여 효과적으로 변환 사전을 구축해야 하는 것이다[10].

3. 동사 '먹다'의 변환 사전 기술

위에서 살펴본 영어 대역어 선택 요인들을 이용하여 동사 '먹다'를 변환 사전에 본격적으로 기술하기 전에 고려해야 하는 사안들이 있다. 일반적으로 한국어 문장은 필수 성분이라고 하더라도 수의적으로 생략이 자유로운 특징을 갖는다. 따라서 주어가 생략된 '밥을 먹었다'라는 문장을 번역할 경우, 기계는 '밥을 먹다'와 변환 사전의 '*A가 *B를 먹다'를 동일한 형식으로 포착하지 못한다.

이러한 문제를 해결하기 위하여 동사가 취하는 최대 논항을 사전에서 제시하지 않고 번역에 영향을 미치는 논항만을 선별하여 기술하는 방식을 취할 수 있다[11]. 기계번역 과정을 고려해본다면, '*A:NN,가 *B:NN,를 먹다 = *A eat *B'라는 사전 기술에서 '*A'에 관한 정보는 잉여적인 정보로 볼 수 있다. 변환 사전에서 처리해주지 않아도, 분석부에서 '*A'는 이미 한국어 문장의 주어로 분석되고 생성부에서 영어 문장의 주어로 기계적으로 처리 가능하기 때문이다. 이러한 잉여적인 정보를 제거하게 되면 (11)과 같게 된다.

(11) 먹다
(*A:NN,를 먹다 = eat *A)
*A:NN,를%액체 먹다 = drink *A
귀:가 먹다 = become deaf
욕:를 먹다 = be scolded

변환사전에서는 또한 기정치(default)를 이용하여 대역어를 제공할 수 있다는 점을 고려해야 한다. 예를 들어, '먹다'의 기정치를 'eat'로 설정했다면 동사의 대역어가 'eat'인 경우에 대해서는 사전에서 따로 처리해 주지 않아도 된다. (11)의 경우라면, ( )로 표시된 정보는 기정치가 대신할 수 있는 것으로 변환사전에서 제공할 필요가 없는 정보라고 할 수 있다.

이러한 모든 사안을 고려하여 논의된 동사 '먹다'의 실제 기술은 다음과 같다. (지면 관계로 위에서 논의된 유형들만 보인다.)

(12) 먹
= eat
0 귀:가 먹다 = *A become deaf
0 *A:VV,-어 먹다 = *A:VV
0 *A:NN,를%액체 먹다 = drink *A
0 *A:를*%돈 먹다 = embezzle *A
0 *A:가%도구 먹다 = *A bite
0 *A:를%순위 먹다 = become *A
0 욕:를 ~ = be scolded
0 겁:를 ~ = be frightened
0 충격:를 ~ = be shocked
0 마음:를 ~ = make up mind
0 나이:를 ~ = grow older

10) 한국어 동사의 영어 대역어 선택에 있어 고려해야 하는 정보들은 매우 이질적인 것처럼 보이지만 실제 변환 사전 기술에서는 종합적으로 적용된다. 오히려 개별적인 정보만으로는 충분한 역어 선택 정보를 제공하지 못하는 경우가 발생할 수 있다. 실제로 ALT-J/E (일본어->영어 번역시스템)에서 용언사전에서 사용하는 의미자질의 수와 기술할 수 있는 격프레임의 수와의 관계를 조사한 바에 따르면, 약 500 (의미자질 체계에서는 6층위까지)개의 의미자질로는 78%의 격프레임이, 약 50(4층위의 의미자질)개의 의미자질로는 52%의 격프레임 밖에 기술할 수 없다는 것이 밝혀졌다.[9]

11) 일반적으로 정밀한 동사 사전에서는 여기서 제시한 논항과 선택제약, 어휘 정보들을 모두 갖고 있다. 그러나 기계 번역을 위한 변환 사전은 이러한 정보 중에서 번역에 필요한 핵심 정보만을 재정리하여 담고 있어야 한다.

0 더위:를 ~ = be affected by the heat
0 *A:에서 미역국을 먹다 = fail in *A
0 *A:에서%장소 미역국을 먹다 = eat seaweed soup in *B

4. 번역 결과

우리의 목적은 변환 사전에서 올바른 동사 대역어 선택을 위해 담아야 하는 정보를 정리하고 이를 바탕으로 변환 사전을 구성하여 기계번역에 적용하는 것이었다. 변환 사전의 정보가 올바르게 기계번역에 적용되었는가를 판단하기 위해서는 많은 문장들의 번역 결과를 검토해야 한다. 그러나 분석부, 변환부, 생성부에 대한 논의 없이 변환 사전 정보와 그 번역 결과만을 확인하는 것은 커다란 한계를 갖는다. 각 부분이 유기적으로 작동하는 번역 시스템에서 번역 결과는 모든 과정의 종합적 산물로 볼 수 있는데 변환 사전은 전체적인 과정의 일부분이기 때문이다.

따라서 여기서는 번역 시스템 전반을 평가하기 위한 본격적인 번역 테스트 실험을 대신하여 우리가 기술한 변환 사전 정보가 번역 결과에 적절하게 반영되고 있는지에 대한 가능성을 확인하도록 하겠다.

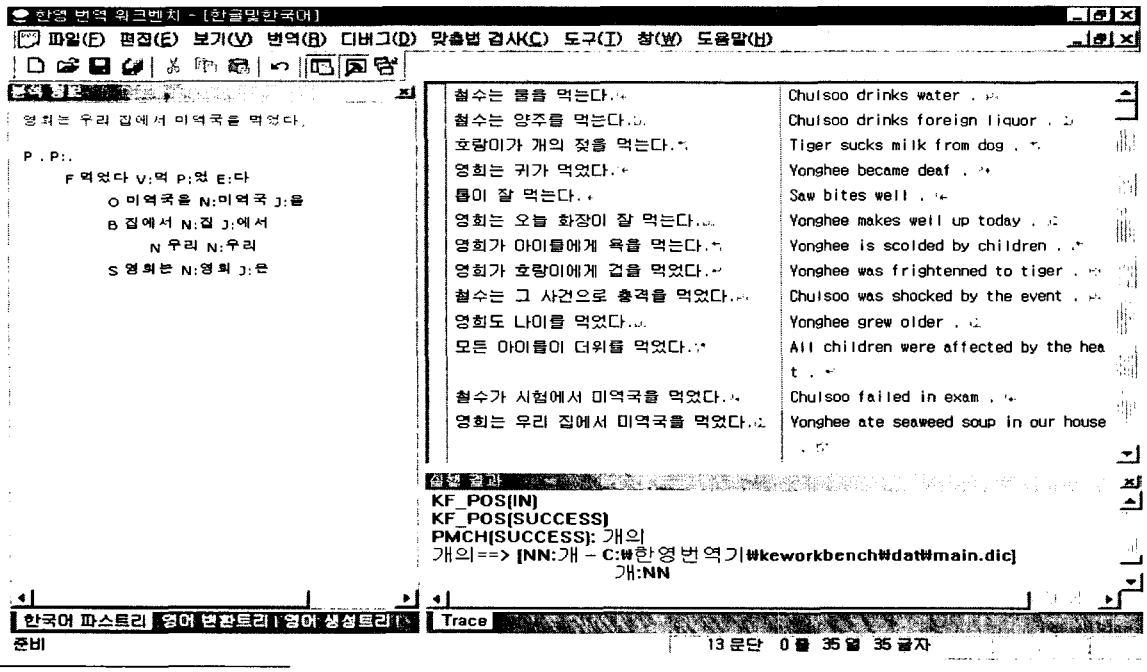
<그림1>[12]에서 확인할 수 있듯이, 기본적으로 변환 사전에서 기술된 정보들이 번역결과에 잘 반영되고 있음을 알 수 있다. 물론 의미 자질을 이용해서 동사 대역어를 선택하는 경우, 해당 명사에 적절한 의미 자질이 붙어 있지 않아 의미 자질을 추가한 예도 있다. 또한 분석부나 생성부의 오류로 인해 적절한 번역이 이루어지지 못한 경우도 있었다.

그러나 논항과 문법 형태소, 선택제약, 개별 어휘 등의 정보로 동사 '먹다'의 영어 대역어를 결정할 수 있었다는 사실에는 주목할 필요가 있다. 이들은 아주 기본적인 언어학적 개념들이지만 변환 사전에서의 용언 기술에는 필수적인 정보들이다. 기계번역은 이러한 기본적이지만 필수적인 정보들을 이용한 정밀한 변환 사전 없이 성공할 수 없다.

비교적 근래에 소개된 예문에 기반한 기계번역(example-based machine translation, EBMT)의 경우도 최종 번역문 선택에 있어서 의미 유사성(semantic similarity)을 계산하기 위해 시소러스(thesaurus)를 이용한다. 그러나 실상이 시소러스를 구축하는 작업은 선택제약을 위해 명사에 의미 자질을 표시해 주는 것과 크게 다르지 않다. 또한 EBMT에서 유사성을 계산하는 과정은 실제 문장과 변환 사전을 매칭(matching)하는 단계에 비교될 수 있다.

결국 기계번역에서는 기본적으로 인간의 어휘부(lexicon)

<그림 1 > 한영 기계번역 작업 툴



12) <그림 1>은 (주) 언어와 컴퓨터의 한영 기계 번역기 작업 툴이다.

정보를 담고 있는 사전이 필요하다. 어휘 정보 없이 번역을 할 수 있다는 것은 상상도 할 수 없는 일이다. 기계번역에서 언어학자들의 역할은 일차적으로 번역을 위해 변환 사전이 담아야만 하는 정보들을 선별하고 이러한 정보들을 이용하여 대규모의 변환 사전을 구축하는 일이 되어야 할 것이다.

5. 결론

한영 기계번역에서 적절한 동사 대역어 선택의 어려움은 한국어 동형어 처리 문제와 한국어에서는 포착되지 않지만 영어로 번역하는 과정에서 발생하는 영어 표현의 특수성 때문에 기인한 것으로 볼 수 있다. 여기서는 이러한 문제를 논항과 문법 형태소, 선택제약, 개별 어휘 등의 정보를 이용하여 변환사전에서 해결하고자 하였다.

논항과 문법 형태소 정보는 일반적인 한국어 동형어를 구별해 주고, 관용구의 의미 해석에 이용될 수 있었고 서술어가 갖는 논항에 대한 의미제약을 나타내는 선택제약을 통해서는 영어 표현이 갖는 특수성을 변환사전에서 처리할 수 있었다. 공기 관계에 있는 어휘의 대역어 선택에 영향을 미치는 개별 어휘들은 기존의 연구에서 밝혀진 바와 같이 변환 사전에서 개별적으로 처리해 줘야 하는 정보로 변환 사전에서 각각의 동사 대역어 선택에 중요한 역할을 하였다. 또한 동사 대역어 선택에 영향을 미치는 이러한 개별적인 요인들은 실제 변환 사전의 기술에 있어서는 복합적으로 적용됨을 동사 '먹다'의 기술을 통해 확인하였다.

이러한 작업은 궁극적으로 완전한 기계번역을 위해 인간의 어휘부(lexicon)를 기계에 표상하려는 극히 초보적인 시도라고 볼 수 있다. 따라서 인간 어휘부 자체에 대한 연구와 이러한 연구를 바탕으로 기계번역에 필요한 정보를 재정리하는 작업들이 계속되어야 함은 물론이다.

6. 참고 문헌

[1]. 옥철영 · 김영택, 연어에 기반한 최상의 번역 선택, 한국정보과학회 논문지, Vol. 20, No. 4, 1993.

[2]. 옥철영, 한영 기계번역을 위한 구단위 번역 사전, 서울대학교 박사학위 논문, 1993.

[3]. 김나리 · 김영택, Collocation 정보에 기반한 한 · 영 트랜스퍼 사전의 구성에 관한 연구, 서울대학교 컴퓨터 공학과 석사 학위 논문, 1992.

[4]. 이호석 · 김영택, 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전, 한국정보과학회 논문지, Vol. 20, No. 7, 1993.

[5]. 김유섭 · 김영택, 영한 기계번역에서 관용구에 기반한 의미분석, 정보과학회 논문지, 25권, 4호, 1998.

[6]. 윤성희 · 김영택, 기계번역을 위한 자연언어의 숙어적 분석, 한국정보과학회 논문지, Vol. 20, No. 8

[7]. 이상조 · 박상규 · 김영택, 한영 기계번역을 위한 중심어 기반 구 구조 변환 사전, 한글 및 한국어 처리 학술대회 논문집, 1994

[8]. 임홍빈 · 장소원, 국어 문법론1, 방송통신대학 출판부, 1994

[9]. 신효필, 전산언어학 강의 노트, 서울대 언어학과 가상 강좌, 2000.

[10]. Wanner, L. (ed.), Lexical Functions in Lexicography and Natural Language Processing, Amsterdam: John Benjamins Publishing Company, 1996.