영-한 기계번역에서 문형에 의한 조사 및 대역어 선택

박 영진*, 김 남수**, 이 지선**, 이용석**
정인대학 컴퓨터과*, 전북대학교 컴퓨터과학과 언어정보공학실**

Selection of Postpositions and Translated Words by Sentence Pattern in the English-Korean Machine Translation

Y. J. Park, N. S. Kim, J. S. Lee, Y. S. Lee
Dept. of Computer, ChongIn College,
Dept. of Computer Science, Chonbuk National University

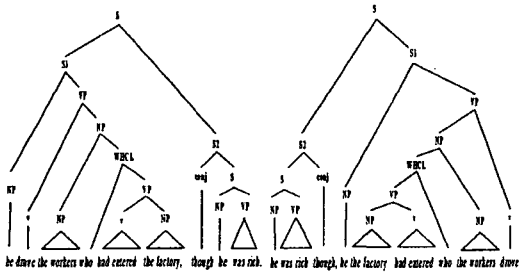
Abstract

영-한 기계번역 중 변환 단계에서 한국어 문장을 생성하기 위해서는 구구조 변환 후 조사 및 대역어 선택으로 이루어진다. 그러나 하나의 영어 단어는 여러 개의 한국어 의미들을 가지고 있기 때문에 문장에서 사용된 영어의 정확한 의미에 해당하는 한국어 대역어를 선택하는 것은 번역의 질을 높이고 시스템의 성능에 매우 중요한 역할을 한다. 특히 용언 및 체언의 대역어 선택은 문장에서 서로 간의 의미적인 관계를 고려하여야 올바른 대역어를 선택할 수 있다.

기존에는 전자 사전에 용언과 체언간의 연어 정보(collocation information)를 구축하여 대역어 선택의 문제를 해결하려고 하였으나 연어 정보가 사전에 존재하지 않을 때 올바른 대역어를 선택할 수 없었다. 또한 용언과 체언의 관계를 나타내는 조사를 선택하기 위하여 격(case)을 세분화하여 사전을 구축하였으나 격의 분류 및 사전을 구축할 경우 격을 선택하는 어려움이 있었다. 이에 따라 본 논문에서는 문형(sentence pattern)에 의한 방법으로 용언의 대역어 및 용언이 갖는 필수격 체언의 조사와 대역어 선택방법을 제안한다. 문형의 구조적인 정보에는 용언과 체언의 의미적 역할(thematic role)을 하는 조사 및 용언이 갖는 필수격 체언의 의미 자질(semantic feature)을 갖고 있다. 이러한 의미 자질을 wordnet과 한/영 및 영/한 사전을 이용하여 의미 지표(semantic marker)를 갖는 문형 사전을 구축한다. 또한 의미 지표를 갖는 문형 사전을 기반으로 조사 및 대역어 선택 알고리즘을 개발한다.

1.서론

변환 방식의 영-한 기계번역 시스템은 형태소 분석, 구문 분석, 변환, 그리고 생성 과정으로 처리되어 진다. 이러한 단계 중 변환 단계의 역할은 크게 구구조 변환과 영어 단어를 올바른 한국어로 선정하는 대역어 및 조사 선택으로 이루어진다. 구구조 변환은 영어와 한국어의 언어적 구조가 상이하기 때문에 영어를 한국어로 번역하는 기계번역에서는 필수적이다. 예를 들어, 예문 "He drove the workers who had entered the factory, though he was rich(그는 부자임에도 불구하고 공장에 들어간 노동자들을 혹사했다)."의 구구조 변환은 분석된 영어 구구조 [그림 1]의 a)가 한국어 구구조 [그림 1]의 b)와 같이 이루어져야 한다. 이러한 구구조 변환은 변환 규칙을 적용한 변환 프로그램이나 변환 문법(grammar)[김남수99]을 이용하여 영어와 한국어의 상이한 구구조를 자연스럽게 변환할 수 있다.



a) 분석된 영어 구구조 b) 변환된 한국어 구구조

[그림 1] 구구조 변환의 예

위와 같이 구구조 변환이 이루어지면 생성된 한국어 구구조 트리를 순회(traverse)하면서 용언 및 체언 등의 영어 단어를 올바른 한국어 단어로 변환하는 대역어 및 조사 선택으로 한국어 문장을 생성한다. 그러나 하나의 영어 단어는 여러 개의 한국어 의미들을 가지고 있기 때문에 문장에서 사용된 영어의 정확한 의미에 해당하는 한국어 대역어를 선택하는 것은 번역의 질을 높이고 시스템의 성능에 매우 중요한 역할을 한다. 예를 들어 변환된 구구조인 S1(setence1) 문장 'he the factory had entered (who) the workers drove'에서 동사 'drove'의 원형 동사 'drive'는 일반적으로 '(차를) 운전하다' 뿐만 아니라 '(동물을) 몰다'와 '(사람을) 혹사하다'라는 여러 가지의 의미를 가지고 있어 동사 'drive'만을 고려하여 문장 전체에서 'drive'가 쓰인 의미를 파악하여 대역어를 선택하기는 힘들다. 따라서 용언의 대역어 선택은 문장에서 용언이 갖는 목적어를 고려하여야 올바른 선택을 할 수 있다. 이에 따라 [그림 1]의 S1 문장에서 'drive' 대역어는 목적어 'worker'가 '사람 명사'의 의미 자질(semantic feature)을 가지고 있기 때문에 '혹사하다'로 선택되어져야 한다. 또한 관계사 절인 'the factory had entered'일 경우 동사 'enter'의 대역어는 용언이 필수격으로 갖는 목적어 'factory'의 의미 자질이 '장소 명사'이기 때문에 '(공장에) 들어가다', '(자료를) 입력하다' 및 '(못을) 박다'등 여러 가지 중에서 '들어가다'로 선택한다. 또한 용언과 체언의 의미적 역할(thematic role) 관계 [NLU]를 나타내는 조사의 선택이 적절하게 이루어져야 한다. 따라서 용언 'enter'의 대역어 '들어가다'와 'factory'의 대역어 '공장' 체언과의 의미적 역할을 표현하는 조사는 장소격 '-에'가 선택되어야 '공장**에 들어간다**'로 올바른 한국어 문장을 생성할 수 있다. 그러나 분석된 영어의 구구조에서 'factory'가 목적격으로 사용되고 있어 목적격 조사 '을/를' 대신에 장소격 조사 '에'를 선택하는 것은 쉽지 않은 문제이다.

이와 같은 대역어 선택의 문제를 해결하기 위하여 기존에는 연어 정보(collocation information)를 이용한 방법[김나리91, 이호석93, 이호석94, 이현아98]이 많이 연구되었다. 그러나 연어 정보를 이용한 대역어 선택은 용언이 갖는 체언들을 사전에 구축하여 처리하는 방법으로 만약 영어 문장에서 용언이 갖는 체언이 사전에 존재하지 않으면 올바른 대역어를 선택할 수 없는 문제점이 있다. 예를 들어 'drive'의 대역어를 '운전하다'로 선택하기 위하여 연어 정보를 'car, vehicle, jeep'로 사전에 구축되어 있을 때 문장'He drives a plane'에서 'drive'의 대역어는 '운전하다'로 선택되지 않는다. 따라서 이러한 방법은 사전을 구축하는 어려움 및 사전의 크기가 방대해지는 단점을 가지고 있다. 또한 용언과 체언의 관계를 나타내는 조사를 선택하기 위하여 격(case)을 세분화하여 사전에 구축하였으나, 격의 분류와 사전을 구축할 경우 격을 선택하는 어려움이 있었다.

따라서 본 논문에서는 구조적 형식에 따라 용언과 필수격 체언의 의미적 역할(thematic role)에 대한 조사를 가지고 있고 필수격 체언의 의미 자질을 포함하는 문형을 이용하여 조사 및 대역어 선택에 대하여 논하고자 한다. 예를 들어, 용언 '운전하다' 문형은 'N1이 N2를 운전하다'로 N1의 의미 자질은 '사람(person)'이며 N2의 의미 자질은 '교통수단(vehicle)'으로 한정하고 있다. 따라서 'He drives a plane'에서 'plane'이 가지고 있는 의미 자질은 'vehicle'이기 때문에 'drive'의 대역어를 '운전하다', '몰다' 및 '혹사하다' 중에서 '운전하다'로 선택한다. 이를 위하여 문형에서 용언이 한정하고 있는 체언들과 영어 단어들의 의미 자질들을 대표할 수 있는 의미 지표(semantic marker)로 구축하여야 한다. 따라서 본 논문에서는 한국어 체언의 의미 자질을 구축하는 방법으로 wordnet과 한/

영 및 영/한 사전을 사용하고 이러한 의미 지표를 갖는 문형 사전을 기반으로 조사 및 대역어 선택 알고리즘을 제안하여 올바른 한국어 문장을 생성할 수 있도록 한다.

본 논문의 구성은 2장에서 의미 지표를 갖는 문형 사전에 알아보고 의미 지표를 추출하는 방법에 대하여 논한다. 3장에서는 의미 지표를 갖는 문형 사전을 기반으로 조사 및 대역어를 선택하는 알고리즘을 제시하고 마지막으로 제 4장에서 결론 및 향후 연구 과제에 대해 알아본다.

2. 의미 지표를 갖는 문형 사전

문형이 갖는 특징은 각 용언의 고유한 성질에 따라 구문적 구성요소인 명사 항에 올 수 있는 명사들이 한정되어 있다. 즉, 용언에 따라 주격 명사로 가질 수 있는 명사들과 목적격 명사로 가지는 명사들이 제한되어 있다. 예를 들어, 용언 '들어가다'의 문형은 'N1이 N2에 V(들어가다)'로 용언이 갖는 체언의 필수격 N1 및 N2에 나타나는 단어들은 아래와 같다.

N1 = {말, 개, 고양이, 철수, 영희, ...}

N2 = {집, 회사, 공장, 동굴, 우리, ...}

이 때 N1 및 N2에 나타나는 단어들을 대표하는 것으로 의미적(semantic)인 지표(marker)로 표현하여 문형을 나타내면 아래와 같다.

N1(animal)이 N2(location)에 들어가다

이에 따라 용언이 갖는 필수격 체언의 의미 지표를 추출하여 문형 사전을 [표 1]과 같은 구조로 구축한다. 예를 들어 'enter'의 의미 지표를 갖는 문형 사전은 [표 2]와 같다.

[표 1] 의미 지표를 갖는 문형 사전

```
E_verb
    (Translate_word_1     K_verb)
    (Sentence_pattern     SP_type)
    (Semantic_marker      (N1 sm1)
                          (N2 sm2)
                          (N3 sm3))
    (Translate_word_2      K_verb)
    (Sentence_pattern     SP_type)
    (Semantic_marker      (N1 sm1)
                          (N2 sm2)
                          (N3 sm3))
                  :
```

[표 2] enter의 문형 사전

```
enter
    (Translate_word_1     들어가다)
    (Sentence_pattern     N1이 N2에)
    (Semantic_marker      (N1 animal)
                          (N2 location))
    (Translate_word_2     입력하다)
    (Sentence_pattern     N1이 N2을)
    (Semantic_marker      (N1 person)
                          (N2 information))
    (Translate_word_3     박다)
    (Sentence_pattern     N1이 N2을)
    (Semantic_marker      (N1 animal)
                          (N2 artifact))
                  :
```

위와 같이 용언이 갖는 필수격 체언들의 의미 지표를 갖는 문형 사전을 구축하기 위하여 본 논문에서는 wordnet과 한/영 및 영/한 사전을 이용하여 구축한다. WordNet은 영어 단어의 동의어 집합(synonym set)으로 표현되고 있으며, 이 동의어 집합들 간의 상하위 개념 관계는 계층구조를 표현하고, wordnet은 이 동의어 집합을 사용하여 시소러스의 역할도 수행할 수 있도록 하였다. 본 논문에서 체언의 의미 지표 추출 단계는 아래와 같이 이루어지고 있다.

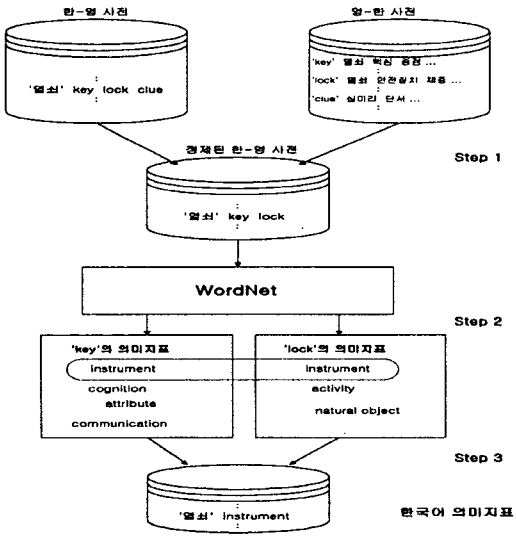
Step1] 한/영 사전의 엔트리와 영/한 사전의 엔트리를 비교하여 공통의 엔트리를 추출하여 정제된 한/영 사전을 구축한다.

Step2] Step1에서 구축된 한/영 사전에서 엔트리의 표제어인 한국어를 설명하는 영어 단어들

을 wordnet에 적용시켜 각각의 의미를 추출한다.

Step3] Step2에서 추출된 각각의 의미들 중 공통된 의미만을 추출하여 해당 한국어 표제의 의미지표로 확정한다.

위의 단계를 '열쇠'라는 단어로서 예를 들면 [그림 2]와 같다.



[그림 2] 체언의 의미 지표 추출

위와 같이 [연세사전]에 문형 정보로 기입되어 있는 단어들을 한/영 사전과 영/한 사전을 이용하여 영어로 변환한 뒤, 변환된 영어단어를 wordnet에 적용시켜 산출된 의미들의 공통적인 속성들을 추출하여 문형의 의미 지표로 확정하였다.

3. 대역어 및 조사 선택

영어의 구구조를 한국어 문형 구조로 변환할 수 있는 여러 가지 문형 형태 중에서 용언이 갖는 필수격의 개수에 따라 문형 구조를 추출하였다. 이때 각 문형 구조에서 필수격들의 일정한 순서를 부여하여 문형 구조가 갖는 의미 지표 및 조사 정보에 따라 대역어 및 조사를 선택하는 알고리즘은 다음과 같다.

[STEP1] 영어 구구조를 한국어 문형 구구조로 변환
[STEP2] 용언의 대역어 및 조사 선택
[STEP2-1] 입력된 영어문장의 체언의 의미지표 (*입력_*)를 얻어옴
[STEP2-2] 문형 사전으로부터 의미지표(*문형_*)와 [STEP2-1]의 의미지표와 비교하여 대역어 선택
주어_후보개수 := 0 *목적어_후보개수* := 0
if (*사전_엔트리개수* < 2) 이면
  현재 사전 엔트리 ID의 대역어로 결정
else 사전_엔트리개수가 0일 때 까지 각각의
  사전엔트리에 대하여
  입력된 문장의 문형 유형(TYPE)에 따라
  1. TYPE1 :
  *입력_주어*의 각 요소에 대하여 *문형_주어*를 비교
  주어_일치개수 := 비교하여 일치된 개수
  if 모두 일치하면
    현재 사전 엔트리 ID의 대역어로 결정
  else if (*주어_후보개수* < *주어_일치개수*)
    주어_후보개수 := *주어_일치개수*
    후보 := 현재 사전 엔트리 ID의 대역어
  사전_엔트리개수 := 사전_엔트리개수 - 1
  2. TYPE2 :
  *입력_목적어*의 각 요소에 대하여 문형_목적어를 비교
  *입력_주어*의 각 요소에 대하여 *문형_주어*를 비교
  if 주어, 목적어 모두 일치하면
    현재 사전 엔트리 ID의 대역어로 결정
  else if (*목적어_후보개수*<*목적어_일치개수*)
    목적어_후보개수 := *목적어_일치개수*
    후보 := 현재 사전 엔트리 ID의 대역어
  else (*목적어_후보개수* = 0) and
    (*주어_후보개수* < *주어_일치개수*)
    주어_후보개수 := *주어_일치개수*
    후보:=현재 사전 엔트리 ID의 대역어
  사전_엔트리개수 := 사전_엔트리개수 - 1
  3. TYPE3 : . . .
if (사전_엔트리개수 = 0) and (후보 <> ' ')
  return *후보*
else if (사전_엔트리개수 = 0) and (후보 = ' ')
  return 디폴트_대역어

[STEP2-3] 결정된 용언의 대역어 문형으로부터 조사 결정

[STEP3] 체언의 대역어 선택

결정된 용언의 대역어 문형이 갖는 필수격 체언의 의미 지표와 명사 사전의 의미지표를 비교하여 체언의 대역어 선택

예를 들어 'He entered the nail.'이라는 입력 문장이 들어 왔을 때, 'enter'의 문형사전(사전_엔트리 개수)은 [표 2]와 같이 '들어가다', '박다', '입력하다'로 있다고 가정하자. 'enter'의 대역어를 선택하기 위해 먼저 [STEP2-2]를 통해 입력된 문장의 필수격 체언인 주어의 의미지표(입력_주어)는 'person'으로, 목적어의 의미지표(입력_목적어)는 'artifact'와 'part_of_body'를 가지게 된다. 'enter'의 문형사전은 위의 [표2]에서 언급한 바와 같이 각각 목적어의 의미 지표(문형_목적어)로서 'location', 'artifact', 'information'를 가지고 있어 [STEP2-2]에서 'enter'의 대역어를 선택한다. 위의 알고리즘에서 유형(TYPE)은 목적어의 개수에 따라 유형을 나누었으며, 위의 입력 문장은 TYPE2에서 주어와 목적어의 의미지표가 비교된다. 처음 알고리즘의 적용된 대역어 사전은 첫 번째 사전인 '들어가다'의 주어와 목적어가 비교되는데, 사전의 슬롯(slot) 'semantic_marker'의 N1인 필드(field)가 '문형_주어'로 '입력_주어'와 비교가 되고, N2인 필드가 '문형_목적어'로 '입력_목적어'와 비교된다. 이와 같이 3번의 비교를 끝내면, '후보'에는 'translate_word2'가 되어 'enter'의 대역어는 '박다'로 선택이 되고, 문형은 'N1이 N2을 V'의 형태로 필수격 체언의 조사는 각각 '이'와 '을'이 부착된다. 또한 영어 단어를 대표할 수 있는 디폴트 대역어(가장 많이 쓰이는 대역어)를 선정하여 만약 일치되는 의미 지표가 없다면 선정된 디폴트 대역어를 결과로 내줄 수 있도록 하였다. 또한 체언의 대역어는 '박다'가 가지는 체언들의 의미 지표와 명사 사전의 'nail'이 가지고 있는 의미 지표를 비교하여 처리한다. 즉, 문형 '박다'가 갖는 목적어의 의미 지표는 'artifact'이므로 'nail'의 대역어 및 의미 지표가 각각 '못-artifact'과 '손톱-part_of_body'이므로 같은 의미를 갖는 '못'을 선택하여 올바른 한국어 문장 "그는 못을 박다.'를 생성한다.

4. 결론 및 향후 연구방향

본 논문에서는 용언이 갖는 필수격들의 조사 및 대역어를 선택하는 방법으로 문형을 이용하였다. 이를 위하여 wordnet과 한/영 및 영/한 사전을 이용하여 체언의 의미 지표를 갖는 문형 사전을 구축하였고 조사 및 대역어 선택 알고리즘을 구현하였다. 이와 같이 문형에 의한 조사 및 대역어 선택 방법은 구축이 어려운 것으로 알려진 연어 정보 사전 없이도 문장의 의미에 맞는 한국어 문장을 생성할 수 있는 장점이 있다. 또한 실험을 통하여 문형에 의한 조사 및 대역어 선택이 올바르게 처리되는 것을 알 수 있었다. 앞으로의 연구 방향은 보다 정확한 의미 지표를 구축할 수 있는 방법에 대하여 많은 연구를 하여야 할 것이며 문형의 구조가 능동태를 기반으로 되어 있어 영어의 수동태 문장을 처리할 수 방법에 관한 연구이다.

참고 문헌

[김나리91] 김나리, 김영택, "Collocation 정보에 기반한 한-영 기계번역 사전의 구성", 서울 대학교, , 한국정보과학회 학술발표논문집 91년 가을

[이호석93] 이호석, "영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전", 서울 대학교, , 정보과학회 논문지 93년 7월

[이호석94] 이호석, 김영택, "영한 변환사전 생성을 위한 말뭉치에 기반한 연어와 관용어의 자동 추출",서울대학교, 한국정보과학회 논문지 94년 11월

[이현아98] 이현아, 이재원, 장병규, 김길창, "한국어 구문 공기 정보와 사전 규칙을 이용한 영-한 기계번역에서의 역어 선택", 한국과학기술원, 정보과학회 학술발표논문집 98년 봄

[연세사전] 두산동아, 연세한국어 사전, 연세대학교 언어정보개발연구원 편

[김남수99] 김남수, "영한 기계번역에서 조건 단일화 기반 변환 문법 해석기", 전북대학교 석사학위논문, 1999

[NLU] James Allen, Natural Language Understanding, The Benjamin/Cummings Publishing Company, Inc