인지과학, 제 14 권 제 2 호, 2003년

Glossary에 기초한 시스템에서의 적형태 영어문장 생성을 위한 한영 대역어 전자사전구축

Constructing A Korean-English Bilingual Dictionary For Well-formed English Sentence Generations In A Glossary-based System

신 효 필*
(Hyopil Shin)

요 약 본 논문은 자연언어처리 (Natural Language Processing), 특히 한영 기계번역에서 필수적인 한영 대역어 사전을 구축함에 있어 영어 생성시 정확한 문장형태를 도출하기 위한 방법에 대해 논의한다. 기간의 연구는 주로 한국어와 영어의 의미적 모호성이 해결된 정확한 번역을 위한 대역어 내지 변환사전 구조에 초점이 맞추어져 왔고 상대적으로 형태적 또는 구문적으로 정확한 영어문장을 생성하는 것은 간과되어져 왔다. 기존 자원의 활용이라는 측면에서는 텍스트화된 한영사전을 그대로 이용한다고 하면 그 기술방식과 영어표현은 다양한 형태로 나타나기 때문에 정확한 의미의 대역어 뿐만 아니라 적격한 영어문장의 생성을 위해서는 어떠한 정보들이 대역어 사전에 기술되어야 하는지 고려해 볼 필요가 있다. 따라서 본 논의에서는 기존의 인쇄된 한영사전을 구조분석하여 자동으로 변환하여 최소한의 인간의 간섭으로 정확한 영어생성에 필요한 형태적 정보를 자질로 부여하는 방법을 기술한다. 기본적으로 이 방법은 단어 대 단어 번역시스템 등 glossary에 기초한 얕은 층위의 번역이 필요한 시스템을 위한 사전을 구축에서 시작하며 더 나아가 대규모의 전자사전 구축작업에서 어떻게 응용될 수 있는지 논의한다.

주제어 자연언어처리, 한영대역어사전, 자동사전구문분석, 형태적자질구조, 적격영어문장생성

Abstract We introduce a way to generate morphologically and syntactically well-formed English sentences when building Korean to English bilingual dictionary for Machine Translation Systems. It has been proved that basic inflectional or structural descriptions for English sentences are by no means enough to generate proper English sentences because of traditional dictionary structures. Furthermore, much research has been focused only on how to disambiguate semantic ambiguities of words in a bilingual dictionary. To take advantage of existing paperback Korean to English bilingual dictionary, its automatic conversion to an electronic version and methodologies to assign proper features to the descriptions for well-formed English sentences with minimum human effort have been proposed on the basis of the dictionary-specific structures. This approach was originally motivated for a glossary-based machine translation system, but it can be also applied to large scale dictionary work.

* 서울대학교 인문대학 언어학과 조교수
Tel: 02-880-6170 Fax: 02-882-2451

1. 서론

자연언어처리(Natural Language Processing) 또는 어휘지식 데이터베이스 (lexical knowledge database) 연구의 일환으로 전자사전에 대한 관심과 실제 연구가 많이 행해져 왔고 현재도 진행 중이다. 그 중에서 기계번역 또는 다국어 (multilingual) 처리 관점에서 대역사전에 관한 작업이 많이 행해져 왔다.[2] 기계번역의 경우에 번역사전은 출발이 되는 원시언어(source language)에서 목표언어(target language)의 자연스러운 생성을 위해 목표언어의 특징적인 활용이나 형태, 의미적 정보가 미리 번역어 사전에 기술되어야 한다.

기간의 한국에서의 기계번역과 관련된 연구는 외국어-한국어, 특히 영어/일어/중국어에서 한국어로 주로 행해졌다. 따라서 한국어는 분석보다는 생성관점에서 형태소적 정보, 구문적 정보가 우선적으로 요구되었다. 상대적으로 그동안 한국어에서 외국어로의 연구는 많이 행해지지 않았는데 그 이유는 생성과 달리 한국어 분석에 많은 어려움이 있기 때문이다. 한국어-외국어(특히 영어, 중국어, 일본어)로의 연구에서 외국어 생성을 위한 대역사전은 필수적이나, 많은 경우에 대역사전은 단순히 번역어를 나열하는 정도에 그치고 있고 실제로 어떠한 형태적 정보도 목표어 생성적 관점에서 체계적으로 기술되고 있지 않다. 또한 대규모의 번역사전 구축에 있어서도 다의어적 쓰임이나 용법에 초점이 맞추어져 그 형태, 구문적 정확한 문장의 생성은 간과되었다.

한편 한국어-영어 대역사전을 구축함에 있어서 기초자료로 텍스트화된 한영사전이 이용된다. 그러나 이런 종류의 인쇄사전은 그대로 자연언어처리에 응용되기에는 문제점을 지닌다. 사전기술과 항목분류 등 사전학적인 기본적인 문제는 차지하더라도 대역어, 표현의 질적인 문제 - 번역어의 영어답지 못한 표현, 잘못된 표현 - 와 여러 기술적인 문제 - 동형어와 다의어 구분의 자의성, 파생어 기술의 부적절 등-를 포함하고 있다.

이 논의에서는 요즘 그 요구가 증대되고 있는 정보검색, 요약, 그리고 기계번역의 통합시스템에서 그리고 단어 대 단어 정도의 번역에서 필요한 사전을 구축할 때 그리고 기존의 텍스트화된 사전을 사용하여 대역사전을 구축할 때 어떻게 영어의 생성적 관점에서 필요한 형태통사적 자질(feature)들을 기술하고 최소한의 사전기술자들의 간섭으로 텍스트 사전에서 일차적으로 구문분석 또는 패턴매칭된 사전항목들이 정리되어 유용한 사전이 될 수 있는지 고찰해 본다. Grossary에 기초한 단순 시스템 및 통합시스템에서의 각 모듈은 고도의 질적인 측면보다는 속도면에서 그 중요성이 강조되기 때문에 이런 관점에서 대역사전을 구축하고 이렇게 구축된 사전이 다시 어떻게 최소한의 사전기술자들의 노력으로 의미적, 형태적, 통사적으로 풍부한 정보가 기술되는 다른 전자사전과 결합될 수 있는지 살펴본다. 여기서 근간으로 삼고 있는 방법은 미국 뉴멕시코 주립대학(New Mexico State University)의 CRL(Computing Research Lab)에서 행해진 Glossary-에 기초한 한영기계번역을 위한 사전구축 작업이며 이를 바탕으로 더 필요한 수정과 첨가가 이루졌다. [2], [3], [4]. 이렇게 구축된 대역사전이 기존의 출판사전의 정보를 그대로 쓰는 사전에 비해 단어 대 단어 번역시스템에서 번역의 질을 어떻게 향상시키는지 살펴보고 또 대규모 전자사전 구축 작업인 21세기 세종전자사전 구축에서 어떻게 통합될 수 있는지 살펴본다.

2. 기존의 관련 연구 및 한영사전의 구조

한영기계번역 관점에서 변환사전과 관련된 초기 대표적인 연구로 옥철영을 [5] 들 수 있다. 우선 이론적인 토대로 이 연구를 살펴보고 이 논의의 특징을 바탕으로 대역어 사전이 자연언어처리 관점에서 어떻게 구축되는지 살펴본다. 또한 본 논의에서 기본 자료로 사용되는 텍스트화된 한영사전은 [6] 인간의 편의를 위해 기술된 것으로 그대로 자연언어처리에 사용되기 어렵기 때문에, 기존 대역어 사전구조 및 기술을 살피고 자동적으로 구문분석 또는 패턴매칭(pattern matching)하여 필요한 정보를 도출하는 방법을 고찰한다.

2.1 옥철영(1993)의 연구

옥철영[5]은 한영 변환사전의 초기 연구로 한국어와 영어와의 통사적 모호성, 의미 및 모호성 그리고 구조변환의 문제로 인해 발생하는 한영기계번역에서의 모호성을 해결하기 위해 구 단위 변환사전을 제안한다. 여러 의미를 가지는 용언의 의미 및 어휘 모호성을 해결하기 위해서, 용언을 특정역어로 번역되게 하는 단어를 용언의 collocation으로 정의하고 그 통사적 역할에 따라 SBJ, OBJ, ADV 세 유형으로 분류한다. 그리고 여러 다른 형태소로 분석되는 단어의 통사적 모호성을 해결할 수 있는 구와 한영 번역에서 구성성분이 1:1의 구조 대응관계를 갖지 않는 구를 숙어로 정의하여 변환사전에

1) 자연언어처리와 같은 특수목적의 전자사전에 관한 한국의 대표적인 연구는 21세기 세종계획의 일환으로 문화관광부와 국립국어연구원 주도로 진행되는 세종전자사전구축 작업이다. 자세한 것은 21세기 세종계획 전자사전 개발분과 (2002) 참조 [1].

이 숙어의 기술형식을 정의하고 있다. 이런 기본적인 방법으로 한 어휘가 가지는 다양한 용법을 변환사전에 기술하고 있는데 다음은 '먹다'의 기술이다.

(1)

```
(먹다
  ((OBJ
    (((밥 rice)(고기 meat)(생선 fish)(빵 bread)(조반 breakfast)(eat))
    (((약 medicine)(좀 pawn))          (take)
    (((술 wine)(물 water))             (drink)
    (((담배 cigarette))                (smoke)
    (((젖 milk))                       (suck)
    (((겁 scared)(욕 scolding)(나이 older))   (get))
    (((이문 profit))                   (make))
    (((상 prize)(판돈 wager)           (win))
    (((녹 stipend))
  ((SBJ
    (((톱 saw)(칼 knife)(대패 plane))  (cut)))
  (IDM
    ((귀!가 =!다)                      (become deaf)
    ((A!을 한입! =!다)                 (take a bite of A!OBJ))
    ((A!를 한입!에 =!다)               (eat A! OBJ at a mouthful))
    ((A!을 게걸스럽!게 =!다)           (devour A! OBJ)
    ((A!을 맛있!게 =!다)               (eat A!OBJ with relish))
    ((A!를 =!어 보!다)                 (try A!OBJ))
    ((A!를 =!어 치우!다)               (eat up A!OBJ))
    ((A!를 =!고 남기!다)               (leave A!OBJ half~eaten))
    ((배부!리 =!다)                    (eat *POS fill))
    ((=!고 살아가!다)                  (manage to live)
    ((놀!고 =!다)                      (live an idle life))
    ((=!을 수 ! 있!는)                 (eatable)
    ((=!을 수! 없!는)                  (not good to eat))
    ((이문!을 =!지 않!고)              (without profit))
  (DFT (eat)))
```

'먹다'의 기술에는 우선 그 해당 목적어의 종류에 따라 다양한 역어가 선택되고 있다. 또 일대일의 대응관계를 갖지 않는 표현을 숙어로 하여 해당되는 영어의 적절한 표현으로 기술하고 있다.

옥철영(1993)은 한영기계번역 초기의 연구로 의의가 있으나 몇 가지 한계점을 지적할 수 있다. 우선 동형어 및 다의어의 구분에 기초하지 않은 어휘 그 자체의 나열적인 방법이라 이 어휘가 조금만 확장되어도 위의 번역사전의 의도대로 변역되기 어렵다. 가령 '밥을 먹다'라는 문장에 대해서는 위에 사전에 따라 'eat rice'라는 표현으로 번역되나[3] 조금만 확장된 표현들 '점심을 먹다, 오곡밥을 먹다, 김치를 먹다' 등으로 된 번역은 위의 사전에서 그 항목이 다 나열되지 않는 한 제대로 번역되기 어렵다. 이런 세밀한 용법은 실제 영어문장생성에서는 의미적으로 아주 중요한 역할을 한다. 그러나 본 논의에서는 이런 의미적 모호성 해결은 논의로 하고 여기서 형성된 정보가 세종전사사전의 기술에 통합되어 다의어, 동형어의 구분된 기술로 어떻게 결합되는지 논의한다.

형태 구문적인 측면에서도 이 사전기술은 충분하지 못하다. 가령 (1)에서 'A!를 =!어 치우!다'의 대역어로 'eat up A!OBJ'로 기술된다. 즉 'A!'에 해당하는 영어 단어가 'eat up' 다음에 목적어로 나타난다는 것이다. 구조적인 관점에서 필요한 형태정보 (A!OBJ)를 적절한 위치에 부여한다는 점에서 의의가 있으나 실제로 영어생성에 필요한 다른 정보들은 기술되고 있지 못하다. 즉 이 표현에서 머리어(head)가 되는 것은 'eat'이며 이것이 인칭, 수, 시제 등에 의해 실제 문장에서는 활용되어야 하는데 적절한 정보가 주어져 있지 않고 목적어가 대명사인 경우에는 'eat' 과 'up' 사이에 위치해야 하지만 그런 정보는 명시되어 있지 않다. 또한 위의 사전기술에서 'eatable', 'not good to eat', 'without profit'과 같은 대역어들은 위와 같은 형태로만으로는 정확한 영어문장을 생성할 수 없다. 즉 이 대역어 앞에 오는 서술어가 어떤 것인지 알 수 없다.

본 연구에서 일차적으로 초점을 맞추는 부분은 이런 구조적인 측면이지 한국어의 표현을 영어로 옮길 때 나타나는 모호성 해결은 논외로 한다. 구조적으로 제대로 된 영어문장이 생성될 수 있도록 그 대역어들에 적절한 정보를 명시하는 방법에 초점을 맞춘다. 따라서 옥철영(1993)과는 다른 관점의 논의이다.

2.1 기존 한영사전의 기술의 특징

2.1.1. 구조적 관점에서의 특징

기존 한영 사전은 주요품사 - 명사, 동사, 부사 등- 을 표제어로 하고 여기서 다시 파생될 수 있는 성분들- 동사, 부사 -을 같이 기술하는 특징이 있다. 다음은 '고전'이라는 항목에서 기술되어 있는 번역어이다.[6]

(2)

고전(古典) an old book; a classic; the classics; [고전 문학] classical literature. ¶ ~적인 classic; classical / ~적인 용모다 have clear-cut

3) 사실 이 번역도 좋은 영어번역이라 할 수 없으나 여기서는 논외로 한다.

features; look classic

⊙ ~극 classical drama. ~문학 classical literature; (the) classics. ~미(美) classical beauty. ~음악 classical music; long hair (美俗). ~주의 classicism; ~주의자 a classicist. ~파 경제학 classical economics. ~학 (study of) classics; ~학자 a classical scholar; a classicist / ~학파 the classical school; the classicists.

고전(古錢) an ancient coin; an old coin; an old coin.

⊙ ~수집가 a collector of old coins

고전(苦戰) hard fighting; a hard (severe) fight; a desperate [hard-fought) battle; [경기, 경쟁의] a close game; a tight play (game match); hard (tough) going. ~하다 fight hard; struggle desperately. fight against heavy odds; fight hopeless odds; have a tough (close) game. ¶ ~끝에 이기다 win a bitter fight(hard-fought game) ((from)) / 선거에서 상당한 ~을 하다 have a close contest in the election / 경기는 상당한 ~이었다 The game was close and tough.

이 사전의 기술에 따르면 '고전'은 세 동형어 (古典, 古錢, 苦戰)로 되어 있으며 여러 파생어 '고전극, 고전문학, 고전수집가, 고전파 경제학' 등이 각 항목에서 ~에 의해 도출된다. 이런 인쇄사전이 전자화되고 여기서 필요한 정보를 추출하여 역어사전을 구성할 때, 우선 가능한 번역어들을 자동으로 획득하고 그 다음 인간의 간섭으로 교정하고 재분류하는 방법이 필수적이다. 왜냐하면 이 기술에서 보듯이 다의어의 구분이 이루어지지 않고 이를 자동화하기는 어렵기 때문이다. 또 같은 의미라도 여러 가능한 표현들을 문체적 차이에 따라 나열된다. 가령 'hard fighting'이나 'hard fight' 와 같이 품사적 차이에 따른 기술뿐만 아니라, 'solid state physics', 'physics of the solid state' (고체물리학)과 같이 연결구조에 따른 변이형들 더 나아가 'tenacity', 'pertinacity' (집착력) 과 같이 영어자체의 동의어까지 나열된 기술이 상당히 많기 때문이다. 이 경우 이 모든 것을 다 나열하는 것은 사전의 불필요한 항목을 증대시키고 비효율적으로 만들 수 있기 때문에 문체적인 차이가 생기는 것까지 목록화 할지는 더 고려해 보아야 한다.

이 글에서 논의하는 대상은 일차적으로 전자화된 인쇄사전에서 패턴매칭에 의해 자동적으로 획득된 대역어들을 어떻게 최소한의 사전기술자들의 교정으로 빠르게 번역사전을 구축하는지 그리고 형태적으로 정확한 대역어가 도출되도록 어떤 정보를 어떻게 기술해야 하는지에 대한 것이다. 따라서 분석대상이 되는 기존의 한영 인쇄사전의 구조에 의존적이며 그 대역어 기술 방식에 따른다. 위와 같이 문체적으로 나열되는 표현의 경우 가급적 하나의 대역어를 취하고 그 선택은 먼저 나온 것 그리고 짧은 것을 우선 순위로 한다. 왜냐하면 다음 절에서 논의하겠지만 구로 된 표현보다는 한 단어로 된 대역어들이 영어생성에 있어 훨씬 더 유용하기 때문이다. 한편 자동 추출의 경우 표제어에서 파생되는 단어들이 한 항목에서 같이 기술되면 - '사랑'이라는 표제어 항목에 '~하다', '~받다' 의 경우나 위의 예의 '고전'에서 파생되는 '고전극' 등- 이런 구조에 품사를 달리하거나 (사랑하다) 또는 그렇지 않은 경우(고전극) 등이 적절히 구별되어야 한다. 그러나 이 보다 더 복잡한 구조를 보이는 경우도 있다.

(3) 가. 감흥: fun; interest; inspiration
~을 자아내다: stimulate(excite) interest
~이 일어나다: be deeply stirred; be inspired

이 경우는 '사랑하다' 등과 달리 '감흥을 자아내다'가 전체적으로 표제어가 될 수 없기 때문에 '감흥'에서 기술되기 보다는 '자아내다'의 한 의미로 기술되어야 한다.[4]

더 나아가 기술적 복잡성을 보이는 것으로 다양한 종류의 괄호에 의해 표시되는 정보들이다. 이 표식에 의해 동일한 표현들이 나열되기도 하고, 생략될 수 있는 성분들이 명시되기도 하며 또 해당분야를 명기하는 등 설명을 위한 구조로 사용되기도 한다. 이는 사전을 구문분석(parsing)할 때 문제를 야기한다. 즉 단순한 나열, 생략될 수 있는 성분들 그리고 설명들을 서로 쉽게 구분할 수 없는 문제가 생긴다. 또한 그 괄호의 범위 (scope)가 애매한 경우가 많다. 바로 직전의 단어에 영향을 끼치는 경우와 몇 단어에 걸쳐 영향을 끼치는 경우가 있다.

(4) 곗돈 : money for (from/owned by) the mutual assistance society (the credit union)

이 예에서 괄호 안에 있는 성분은 전치사 for를 대치할 수 있는 경우와 'the mutual assistance society' 전체를 대치할 수 있는 경우를 보이고 있다. 실제작업에서는 이 괄호안의 성분들은 분석하지 않고 생략하였다.

4) 실제로 다음 절에서 살펴 볼 세종전자사전에서는 서술어가 취하는 명사와 관련하여 어휘의 의미(sense)가 별도로 구분되기 때문에 이렇게 구별되는 역어를 기술할 수 있는 기재가 마련되어 있다.

2.1.2. 내용적 관점의 특징

논의의 자료가 되는 민중서림의 한영사전의 내용적 특징을 간략히 살펴보자. 첫째, 번역어의 두 언어의 문화적 차이에 기인한 개념구조의 상이와 부재에서 오는 기술의 비형식성 그리고 정확하지 못한 역어를 선택하는 질적인 문제를 지적할 수 있다. 두 문화차이에서 오는 해당 개념의 부재는 아주 다양한 형태로 나타난다. 다음은 그 한 예이다.

(5) 가. 선문답: question-and-answer exchange between Zen priests and their followers
나. 단오: the *Tano* Festival (on the fifth day of the fifth lunar month)

위의 두 어휘는 한국어에서는 존재하나 해당 영어에는 이와 직접 관련된 개념을 찾기가 어렵다. 따라서 대부분의 역어사전에서 위의 어휘를 설명하는 표현을 취하는 방법으로 기술한다. 실제로 이와 같은 기술은 기존 한영사전에서 아주 광범위하게 나타난다.

그러나 이와 같은 기술은 인간을 위한 사전에서는 가능할지 몰라도 자연언어처리라는 관점에서는 부적절한 기술이다. 가령 기계번역시에 위와 같이 구(phrase)또는 더 확대된 문장으로 된 표현은 제대로 된 의미를 전달하지 못할 뿐 아니라 영어문장 생성에도 어려움을 야기한다. 이런 경우에는 가급적 가장 적절한 어휘를 찾아야 한다. 그러나 기존 사전을 그대로 이용하는 본 논의에서는 이와 같은 질적인 문제는 자세히 논의하지 않지만 번역사전의 질을 향상시키기 위해서는 계속해서 이런 표현들을 더 정제할 필요가 있다. 이 경우 직접적이고 가장 간단한 현재 상태의 해결방법은 위와 같은 설명적인 기술보다는 차라리 로마자로 표기(romanize /transliterate)하는 것이 자연언어처리 관점에서는 더 유용하다. 영어 생성시 의미는 전달되지 못해도 형태와 구문상에서 나타나는 부적절성은 피할 수 있기 때문이다.

둘째, 앞에서 지적한 대로 하나는 대역어가 한 단어로 대응되는 경우보다는 한 단어 이상의 구(phrase)나 문장으로 구성되는 경우가 많다. 복합어인 경우는 제외하더라도 두 언어의 차이에 의한 이런 구 대역어는 필연적일 수 밖에 없다. 문제는 이런 구 표현에서 영어 생성을 고려해 각 성분에 필요한 정보를 표시해야 한다는 점이다. 즉 복합명사구성인지 아니면 단순한 구구성인지 등이 명시되어야 한다. 왜냐하면 두 구성에 따라 머리어가 위치하는 곳이 다르며, 영어의 문법활용과 복수표지 등은 이 머리어에 부가되는 속성이기 때문이다.

셋째, 사람을 지칭하는 'one, person, oneself, one's'와 사물을 지칭하는 'thing, something', 어떤 일을 가리키는 'matter, affair', 그리고 동사의 일반형을 지칭하는 'do'를 사용하는 기술이 상당히 많다는 점이다.[6]

(6) 가. 언제까지 : ~고 as long as *one* likes
나. 상기: ~시키다 recall (*something*) to (*a person's*) mind
다. 모시다: 부모를 ~ have *one's* parents with *one*
라. 교묘: ~하게 처리하다 manage (a *matter*) cleverly
마. 교분: 옛 ~을 새로이 하다 renew *one's* old acquaintance with (*a person*)

실제로 이 논의의 기초가 되는 미국 뉴멕시코 주립대학에서 개발한 한영 대역어 사전에는 총 83,249 어휘항목이 있으며 그 중 동형어와 다의어를 포함하여 총 284,909개의 번역어가 기술되어 있다. 산술적으로는 한 어휘에 평균 3.4개의 대역어가 나타난다. 이 중에서 "one's"형이 역어에 나타나는 것은 7,815개이며 "oneself"형이 역어에 나타나는 경우는 모두 3436 가지이다.

이 표현이 문제가 되는 것은 역어의 정확성 측면에서 뿐만 아니라 영어의 생성시 문제를 야기하기 때문이다. 즉 위의 표현을 바탕으로 한국어가 영어로 변환될 때 의미를 기술하기에 사용되는 위와 같은 표현들은 실제로 부적절하거나 아니면 정확해도 그 주어 등의 일치자질에 따라 바른 형태로 활용되어야 하기 때문이다. 실제로 한영사전에서 사용되는 'one's'의 경우는 상당부분 실제 사용에서는 필요치 않다. 이는 개념을 기술할 때 사용되는 표현이지 실제 발화에 사용되는 표현은 아니다. 이를 처리하기 위해서는 일차적으로 정리된 자연언어처리용 대역어 사전에서 계속 정리해 나갈 때 삭제하거나 아니면 제한된 경우로 사용을 유지할 필요가 있다.

네 번째 문제는 관사의 사용이다. 인쇄된 한영사전에는 영어의 정관사와 부정관사가 같이 기술되어 있는 경우가 많다. 기본원칙은 가산명사에는 관사를 붙이고 불가산 명사에는 관사를 붙이지 않는다는 원칙이나, 정관사의 경우는 그렇다고 치더라도 부정관사의 사용은 일관적이지 못하고 영어문장의 생성시 불필요한 문제를 야기할 수 있다.[5] 따라서 정관사가 필요한 경우를 제외하고는 부정관사의 사용은 전자사전에서 제외할 필요가 있다. 다만 부정관사가 꼭 필요한 경우나 구 표현인 경우 중간에 나타나는 것은 예외로 한다.

---

5) 실제로 복수형태로 생성하는 경우에 부정관사가 붙어 있는 경우에 부정관사 + 복수형의 형태가 도출될 수 있으며 정관사가 필요한 경우에 그 형태를 제대로 도출할 수 없다.

(7) 결론내다 : draw a conclusion

다섯째로 지적할 수 있는 문제점은 한국어와 그 대역어인 영어의 품사가 일치하지 않는 경우이다. 즉 한국어 어휘로는 명사부류에 속하나 영어의 경우에는 전치사구나 부사구로 되는 경우 등이다. 이는 한국어 명사가 (조사를 수반하거나 아니면 그렇지 않거나 해서) 영어의 전치사구나 부사구에 해당하는 표현이 되기 때문이다. 또 품사차원을 벗어나 많은 영어 대역어가 일대일로 대응되지 못해 구로 대응되는 경우가 많다.

(8) 가. 비상용: for emergency

나. 방과후: after school

다. 실행상: practically; in practice

한국어 어휘에서도 기존의 한국어 사전과의 기술상의 차이가 생기기도 한다. 가령 기존의 한국어 사전에 등재되어 있는 모든 어휘들이 다 기술되지 못하고 또 많은 접사들이 따로 기술되어 있기 때문에 실제로 이 접사와 어휘들이 결합할 때 어떤 의미로 쓰이는지 결정하기가 어렵다. 왜냐하면 대부분이 한자어에서 오는 이런 접사들은 다양한 의미를 갖기 때문이다. 다음은 접두사로서의 '구-'와 접미사로서의 '-구'가 한영사전에 기술되어 있는 예이다.

(9) 가. -구: a tool, an implement, an opening, a mouth, a window, a home, a wicket, a globe, a sphere, a ball, a bulb, a tube

나. 구-: former, one-time, ex-, old, outgoing

2.2. 전자사전 구축에서의 대역어

그동안 자연언어처리 관점에서 전자사전은 여러 방면에서 추구되어 왔다.[6] 현재 진행되고 있는 대표적인 전자사전 작업은 21세기 세종계획 연구 중 하나인 세종전자사전 개발이다. 이 연구는 한국어 전산처리에 필수적인 대규모 한국어 전산어휘부의 구축을 본격적으로 연구 개발하여 기계 가독형(machien readable)사전개발을 추구하고 있다. 이 사전은 효율적인 자연어 자동처리를 위해서 특정 분야에만 응용될 수 있는 자동처리방식에서 탈피하여 언어전반에 걸쳐 종합적으로 적용할 수 있는 방대한 규모의 정보자료를 구축하려 한다.[1] 여기서는 이 세종사전에서 대역어가 어떻게 기술되고 있는지 간단히 살펴보고 그 문제점을 살펴보도록 한다.

2.2.1. 세종전자사전의 정보기술

세종전자사전은 각 품사별로 기술되고 있고 각 품사별로 풍부한 통사, 의미적 정보를 XML을 바탕으로 한 기술로 표시하고 있다. 본 논의와 관계되는 대역어에 초점을 맞추면, 이 품사의 경우는 의미정보의 일부로 @trans=[]라는 항목에 기술된다. 이 사전은 동형어와 다의어의 엄격한 구분에 기준하고 있기 때문에 각 항목(entry)은 하나의 대역어를 가져 자연언어처리시 유용하게 사용될 수 있다. 그러나 한국어 통사의미적 정보의 철저한 기술과 달리 실제 대역어 기술은 기존의 인쇄사전에서처럼 그냥 대역어를 나열하는 정도에 불과하다. 다음은 그 한 예이다. 복잡한 구조를 단순화하기 위해 대역어가 있는 부분만을 표시하였다.

(10) 가. 살통주머니

<sense n=1> @eg=[화살을 살통주머니에 넣다] @trans=[a kind of bag in which one put bow etc] @domain=[] @reg=[] @con=[] @curs=[C] @sem=[용기]]

나. 보이다.

<eg>바지 아래로 복사뼈가 그대로 내보인다</eg>
<eg>속옷이 내보이는 빛 바랜 옷을 입었다</eg>
<sem idval="01">
<trans>be seen</trans>
<eg>신령이 은도끼를 내보이며 물었다</eg>
<eg>동생은 흰 이를 내보이며 밝게 웃었다</eg>
<eg>그는 우리들에게 초조한 기색을 내보였다</eg>
<eg>그렇게 말하면서 그는 은근히 자신의 속내를 내보였다</eg>
<trans>show something</trans>
<trans>show (one's thought)</trans>

위의 기술에서 <trans>에 기술되어 있는 번역어는 기존의 사전과 크게 다르지 않다. 즉 앞에서 지적된 여러 문제들이 그대로 다 나타난다. 몇 가지를 더 살펴보면 다음과 같다.

(11) 가. 첫국밥: the first seaweed soup and rice taken after childbirth

나. 출세주의자: a person who has an ambition to make his mark in the world

다. 하이틴: one's late teens

라. 화성인: a Martian; an inhabitant of Mars

마. 휑하다: extremely empty

바. 별수없다: there's no help

어떤 영어의 생성을 위한 활용자질도 명시되지 않은 그리고 일관적이지 못한 기술이 이루어 진다. 실제로 자

6) 시스템공학센터, 한국전자통신, 그리고 KAIST 및 자연언어처리 전문 회사에서 이와 관련된 연구 및 실제 사전들이 구축되어 왔다.

연언어처리를 위한 전자사전에서 대역어 사전은 아주 중요하나 한국어 어휘의 통사/의미 기술에 초점이 맞추어져 있어 이에 대한 논의와 기술이 제대로 되어 있지 않다. 다만 이 전자사전은 한국어 항목이 동형어와 다의어의 구분에 의해 명확히 구분되어 있어 번역어도 이 관점에서 잘 구분되어 있는 점에서는 의의가 있다.

3. 생성적 관점에서의 대역 자질

이상과 같이 살펴 본 바와 같이 기존의 인쇄사전을 바탕으로 한영대역어 사전을 구축하기 위해서는 기계가 독형 입력외에 내용과 구조를 정리하는 작업이 필요하고 또 더 나아가 영어 생성에 필요한 정보들을 적절히 명시하여야 한다. 여기서는 이런 전자사전을 구축할 때 최소한의 노력으로 최대의 효과를 얻기 위해 필요한 정보가 무엇인지 그리고 이를 어떻게 효과적으로 기술할지 논의한다. 또 어느 정도까지 인쇄사전에서 정보가 부가되어 있는 전자사전으로 자동화할 수 있는지에 대해서도 논의한다. 영어생성에 필요한 정보는 자질(feature)이라는 관점에서 기술되며, 한영 사전의 기술에서 나타나는 특성과 영어 자체의 특성을 고려한 제한된 수의 자질체계가 품사에 따라 설정될 수 있다. 각 품사에서 다양한 특성이 나타날 수 있으나 가장 문제가 되는 명사와 동사 (형용사 포함)를 중심으로 설명한다.

3.1. 명사

명사는 보통 한 단어로 대응되어 기술될 것 같지만 실제로는 그렇지 못하다. 대역어가 단일어라면 별다른 자질이 기술될 필요가 없다. 따라서 하나 이상의 단어로 되어 있는 구(phrase)를 바탕으로 영어생성에 필요한 자질을 명시하는 방법을 살펴보도록 한다. 우선 영어는 단수 복수에 따른 활용이 명사구성에 나타난다. 따라서 이런 활용을 자질로 표시하기 위해서는 명사구(phrase)로 되어 있는 구성에 머리어(head)를 자질로 표시하는 것이 필요하다. 왜냐하면 이 머리어가 실제로 복수형에 따라 활용하기 때문이다. 본 논의에서 머리어는 <#>로 표시한다. 따라서 대역어에 이 표시가 붙어있는 것은 이 부분이 머리어라는 것을 나타내며, 한 단어로 되어 있는 대역어는 물론 이 표시가 필요없다. 구 명사란 번역어가 한 단어로 되어 있지 않고 구로 되어 있는 경우로 다음과 같은 예들이다.

(12) 가. 실행력: power<#> of execution<phn>
나. 제도법: method<#> of making pottery<phn>
다. 가구배치법: plan<#> of furniture arrangement <phn>
라. 가열분해: decomposition<#> by heating <phn>

이 경우 'power, method, plan, decomposition'이 각각 이 명사구의 머리어라는 것이 표시된다. 이런 구조가 구 명사라는 것을 표시하기 위해 (phn) 자질이 필요하다. 왜냐하면 이 구 명사에서 머리어는 다시 수식성분에 의해 수식될 수 있기 때문에 그 위치로 인해 추정할 수 없기 때문이다. 따라서 다음의 명사복합구성과 대조된다.

한편 이 구명사와 대조되는 것으로 명사복합구성(noun compound)이 있다. 명사복합구성은 구명사와 달리 '명사와 명사'의 복합구성이나 '동명사(또는 현재/과거분사) + 명사' 구성, 또는 '형용사 + 명사' 구성 등으로 되어 있는 것으로 그 자질 <nc>만 명시하고 특별히 머리어를 표시하지 않는다. 왜냐하면 영어의 특성상 마지막 명사가 머리어가 되기 때문이다. 다음은 그 예들이다.

(13) 가. 가열면적: heating area<nc>
나. 가장순양함: converted cruiser<nc>
다. 범죄자형: criminal type<nc>
라. 법과대학생: law student<nc>

대역어로 영어표현이 복수인 것이 자연스러운 것이 많다. 이런 어휘들을 위해 <plur> 자질을 사용하고 그 단수형을 대역어로 기술한다.

(14) 가. 성서: Bible; Scripture<plur>
나. 소득: earning<plur>

또한 이 복수표현은 명사복합구성과 구 명사 구성에도 같이 나타날 수 있다. 이를 위해 <phn plur>, <nc, plur> 자질이 필요하다.[7] 에는 영어의 특성상 언제나 복수형으로 쓰여야 하는 구성이 있다. 이 구성을 위해 다음과 같은 <nc plur>이라는 자질이 필요하다.

(15) 가. 10종 경기: ten event<nc plur>
나. 가랑잎: fallen leaf<nc plur>
다. 소득공제: deduction<#> from income<plur phn>;
라. 소량거래: transaction<#> in small lots<plur phn>

실제로 구명사인 경우에 복수는 머리어 명사에 복합명사구인 경우는 마지막 명사에 부여된다.
복합명사구 관점에서 구분되어야 할 것이 소위 고유명사항목들이다. 고유명사 -인명, 지명 등-는 그 속성상 한 단어 이상으로 되어 있으나 복합명사구나 구 명사구와는 다르다. 따라서 이 고유명사가 별도의 사전으로 구축되어 있는 경우는 별 문제가 없으나 통합된 사

7) 실제로 자질은 집합의 개념으로 기술된다. 따라서 어떤 순서도 가정되지 않는다.

전을 고안하는 경우는 이 경우에 <proper> 등의 자질을 붙여 복합명사구에 적용되는 생성규칙이 적용되지 않도록 할 필요가 있다.

마지막으로 명사와 관련한 자질을 살펴 볼 때 지적할 수 있는 것으로 한국어에서는 명사구이나 실제 번역어가 전치사구가 되는 영어표현들로 다음과 같다.

(16) 가. 관할내: within the jurisdiction<#><pp>
나. 교섭중: in negotiation<#><pp>
다. 근무중: on duty<#><pp>
라. 담화중: in conversation<#><pp>
마. 대인용: for adult<#><pp>

많은 경우 명사+한글자 한자 접사 구성에서 가능하며 그 번역상 전치사구로 대응되기 때문에 위아 같이 전치사구의 머리어 부분에 <#>과 전체구성에 <pp> 자질을 명시한다. 한편 위와 구성이 비슷하나 전치사구성이 아닌 다음과 같은 경우는 특별한 주의를 요한다. 기계적으로 '전치사 + 명사' 구성을 위와 같이 처리한다면, 부사 또는 형용사형으로 쓰이는 다음과 같은 표현들도 전치사구로 오인될 소지가 있지만 명사구 복합구성으로 표시되어야 한다.

(17) 가. 내벽: inside wall<nc>
나. 과대평가: over estimation<nc>
다. 대변: opposite side<nc>
라. 뒷맛: after taste<nc>

3.2. 서술어

서술어(여기서는 한국어 동사, 형용사 포함)는 명사의 경우와 달리 더 복잡한 자질이 필요하다. 명사의 활용이 제약되어 있고 그 구성도 몇 되지 않지만 용언의 경우는 여러 활용이 가능하고 그 구조도 복잡하기 때문이다. 우선 대역어가 한 단어로 되어 있는 경우는 명사의 경우와 같이 별 문제가 되지 않는다. 그러나 대부분의 역어는 한 단어 이상의 구로 되어 있어 명사에서처럼 구동사(phrasal verb) 자질 <phv>을 설정한다.

(18) 가. 뒷정리하다: put in order<phv>
나. 드러눕다: lie down<phv>
다. 살랑거리다: blow gently<phv>
라. 살찌다: gain weight<phv>
마. 위임하다: entrust with<phv>

이 구동사는 다양한 구조로 나타나나, 동사가 맨 앞에 위치하고 뒤에 다시 부사, 전치사, 또는 명사 등의 성분이 나타나는 구조로 되어 있다. 이런 구조적인 특성으로 인해 명사와 달리 머리어를 명시할 필요가 없다. 명사의 경우는 수식성분이 앞에 위치하거나 다른 명사가 머리어 앞에 위치하기 때문에 구명사의 경우에도 머리어를 표시할 필요가 있었다.

한편 서술어 구성에서 특징적인 것은 소위 'be+형용사구/분사구/전치사구'로 표시되는 구성이다. 이런 구성은 아주 생산적이며 사전기술에서 그 주의를 요한다. 머리어가 되는 be 동사는 주어의 인칭과 수에 따라 변화하기 때문에 사전기술에서는 제외할 필요가 있다. 이런 구성을 위해 <pred>라는 자질을 부여한다. 이 <pred>는 서술적 의미로 쓰이는 성분을 나타내며 이 자질에 의해 실제 영어 문장생성시에 적절한 be 동사가 복원된다. 다음은 그 단편적인 예들이다.

(19) 가. 유념하다: mindful of<pred>
나. 값나가다: valuable<pred>
다. 붐비다: crowded<pred>
라. 걱정되다: worried about<pred>
마. 상사하다: in love with each other<pred>
바. 오탁하다: polluting<pred>

그러나 실제 자질부여시 이 <pred> 구성은 다음의 수동표현 구성을 위한 자질 <passive>와 그 구분이 모호한 경우가 많다. <passive> 자질은 수동표현으로 대역어가 이루어지는 경우이며 역시 왕성한 출현을 보인다.

(20) 가. 옮다: infect<pass>
나. 완료되다: complete<pass>
다. 즉사하다: kill instantly<pass>
라. 가득해지다: fill up<pass>
마. 복원되다: restore to the original state<pass>

이 예들은 영어의 'be+과거분사'의 형으로 수동을 이루는 구성이다. 이 구성은 <pred> 자질의 과거분사구성과 혼동되기도 하나, 주로 동사에서 동작의 피동을 이루는 구성에는 <pass>라는 자질을 형용사구성에서 서술적으로 쓰이는 (과거)분사형에는<pred>를 부여하도록 한다. <pred>와 <passive> 자질은 위의 특성으로 자동추출할 때 자질을 부여하지만 구분이 애매해서 잘못될 가능성이 많기 때문에 반드시 사전기술자들이 점검해야 할 항목이다.

다음으로 한영대역어 사전에서 많이 나타나는 표현으로 사역적인 의미를 지닌 것들이 있다.

(21) 가. 견고히하다: make firm
나. 꿇리다: make kneel down
다. 누이다: make urinate
라. 늙히다: make old
마. 미화하다: make beautiful
바. 약화시키다: make weak

주로 'make'가 사용되어 사역적인 의미들을 갖는 경

우이다. 이런 표현을 위해서 <cause>라는 자질을 설정하고 make를 생략한 자질구조로 표시한다.

(22) 가. 견고히하다: firm<cause>
나. 꿇리다: kneel down<cause>
다. 누이다: urinate<cause>
라. 늙히다: old<cause>
마. 미화하다: beautiful<cause>
바. 약화시키다: weak<cause>

실제로 영어 생성시는 이 <cause>라는 자질에 의해 make를 복원하고 더 나아가 목적어가 나타나는 경우 적절한 곳에 위치하게 한다.

(23) 가. make him weak
나. make it beautiful

그러나 경우에 따라 make 다음에 목적어가 명시되어야 하는 경우가 있다. '진천하다'의 번역어인 'make the whole world wonder'의 경우 어떻게 대역어 사전에 기술해야하는가가 문제가 된다. 이 경우에 'the whole world'는 이 '진천하다'라는 의미에 기여하고 다른 목적어가 이 위치에 오는 경우는 그 의미를 상실하기 때문에 'the whole world<#> wonder<cause>' 등으로 자질을 표시하여 정확한 영어로의 생성이 이루어지게 한다.

3.3. 인칭/사물을 위한 자질

앞 2.1.2에서 한영사전에서 많이 나타나는 표현으로 인칭, 사물을 지칭하는 'one, oneself, one's, thing' 등의 사용을 언급했다.[8] 이런 구성은 실제 생성에서 문제를 야기함을 이미 언급했다. 이에 대한 아주 완벽한 해결은 없지만 자질구조 관점에서는 이 해당어휘 대신에 자질을 명기하고 실제 영어 생성시 적절한 형태가 도출되도록 한다. 'oneself'를 위해 <self>라는 자질이 one's를 위해 <poss>가 *thing/something*을 위해 <thing>이라는 자질을 설정한다.

(24) 가. 상기: ~시키다 recall (something) to (aperson's) mind
가'. 상기시키다: recall <thing> to <poss> mind

이 경우 다른 자질과 달리 실제 어휘를 생략한다는 점에서 차이가 난다.[9]

4. 사전의 자동 구문분석

일반 텍스트 형태의 사전을 기계가독형태의 사전으로 만들려는 시도는 1960년대 후반의 Systems Development Corporation에서 시작된 이 후 전산학, 언어학, 전산사전편찬학 등에서 단어의 의미를 정형화된 형태로 나타내려는 방법을 찾으려 하고 있다. 이런 작업은 기존의 자원을 이용한다는 점에서 유용한 것으로 간주되나 그 구조의 복잡성에 의해 효율성이 문제되기도 한다. 한편 어휘풀이, 정의, 예문에서 의미관계를 발견하기 위해서는 어휘를 인식하고 풀이 텍스트의 구조를 파싱할 수 있는 프로그램이 개발되어야 하는데 Bograev[7][8], Copestake(1990)[9] 등이 그 예이다. [10]

한편 최근에는 사전의 뜻풀이말에서 의미정보를 사용하여 동형이의어의 중의성 해결을 노력하려는 여러 시도들이 행해지고 있다[11]. 허정, 옥철영 [11]에 의하면 사전의 뜻풀이말을 두 유형으로 구분하여 상-하위어 관계의 유형과 동형이의어가 상-하의어의 중간에 나타나는 구조로 구분하여 명사와 용언을 동시에 고려하고 있다. 본 논의에서의 주된 작업은 이런 사전의 정보를 이용하여 자동으로 의미적 중의성을 해결하는 작업이 아니라, 기존의 텍스트 형태의 사전을 분석하여 필요한 정보를 도출하고 최소한의 인간의 간섭으로 정확한 대역어를 구축하고 영어 생성적 관점의 형태적 정보가 명시된 사전의 구축이다. 이런 작업에서 의미적 중의성 해결은 많은 부분 사전을 정리하는 사람의 손에 달려있다. 따라서 자동적으로 이를 구축하려는 기존의 여러 연구와 방향을 달리한다. 또한 이 작업은 대규모 전자사전 구축작업인 세종사전에서 항목을 기술할 때 그 해당정보로 기술되기 때문에 사전기술자에 의해 중의성이 해결되게 된다. 이에 대해서는 다음 절에서 논하도록 한다.

여기서 사전의 자동정보 추출은 텍스트 형태 사전이 그대로 입력되어 있는 기계가독형 원시자료에서 출발한다. 즉 인쇄사전을 입력하는 것에서부터 시작하는 것이 아니라 이미 기본자료로 입력된 사전에서 사전의존적으로 필요한 자료를 도출하는 방법에 대해 논의한다. 여기서 시도하는 사전의 구문분석은 이 사전정보를 모두 도출하는 것이 아니라 대역어로 필요한 정보만이다. 본 논의에서 사용되는 민중서림의 영한사전[6]을 기계가독형 형태의 어휘데이터로 바꾸는 논의는 최병진[10]에 자세히 기술되어 있다.

4.1. 기본 항목 도출

인쇄사전이 그대로 전자화되어 있는 경우 이를 바탕으로 필요한 형태로 변환하는 것은 각 사전구조에 전적으로 의존한다. 그러나 사전기술이 앞에서 지적한 대로 복잡하기 때문에 이를 필요한 형태로 직접 변환하기는

8) 실제 예는 2.1.2의 (6) 예들을 참조
9) 이런 기술방법 외에 다른 가능한 방법들이 있을 수 있다. 어떤 방법을 취하든지 이런 특징들이 포착될 수 있어야 한다.

어렵다. 또한 품사정보가 명시되어 있지 않는 사전이 대부분이기 때문에 각 항목을 자동으로 품사별로 분류할 수 있는 것도 중요하다.

첫 단계로 우선 표제어와 그 번역어들을 그대로 획득하는 것이 필요하다. 앞에서 살펴 본대로 '~' 로 확장되는 표제어는 제한된 동사 -경동사(light verb) 또는 기능동사[10]-와 결합하는 것은 자동으로 품사를 설정하고 새로운 표제어로 구성하는 것이 가능하지만 그렇지 않은 경우는 직접 확인하여 적합한 표제어인지 점검할 필요가 있다. 표제어와 그 번역어들로 구성된 기본구조를 형성할 때 품사정보에 대한 고려가 필요하다. 대부분의 인쇄사전에서는 품사가 명시되지 않는다. 따라서 이 품사를 어떻게 자동으로 부여할 수 있는지 살펴 보는 것이 중요하다. 여기서 채택하는 방법은 기존의 품사별로 분류된 어휘항목 목록을 사용하여 인쇄사전을 변환할 때 이 목록과 대조하여 품사를 부여한다. 여기에도 물론 중의성이 있는 경우와 항목이 발견되지 않아 품사부여를 할 수 없는 경우가 있다. 이런 경우는 따로 특별한 표시를 하여 사전기술자가 직접 사전을 정제할 때 다시 점검한다. 다만 위에서 언급한 서술명사+기능동사의 구조는 동사나 형용사로 변환될 수있다.

또한 앞에서 기술한 대로 괄호나 여러 사전고유 표시들은 구문분석 과정에서 다 생략되어 가장 단순한 형태들만을 도출한다. 이렇게 해서 일차적으로 구문분석에 의해 도출된 구조들은 다음과 같은 구조로 표시된다.

(25) 가. $H 가늘다
$C Adjective
$M be thin;be narrow;be feeble;be small;be slight;be fine;
$$
나. $H 대세
$C Noun
$M general trend;general tendency;great power;serious condition;critical state;
$$
다. $H 독립하다
$C Verb
$M become independent;stand on one's own legs;separated from;separate from;rely on oneself;help oneself;
support oneself;

이 일차적인 구문분석의 결과 품사정보가 첨가되고 여러 불필요한 표식들을 제거한다. 이 구조는 뉴멕시코 주립대학의 CRL에서 기술되는 구조이다. $H 는 표제어 정보를 $C는 품사정보를 $M 은 첫 번째 구문분석에서 나온 일차구조들을 나타내기 위한 식별자들이다. $$는 이 항목의 끝을 나타내어 다른 항목과 구분한다. 이러한 구조는 동형어 및 다의어가 여전히 구분되지 않고 한 항목 내에서 같이 기술되는 구조라 실제 사용에 있어 정확한 번역을 얻기 어렵다. 이 다음 작업은 이런 기초분석된 자료를 바탕으로 사전기술자들이 직접 정확한 역어를 선택하고 형태적 관점의 자질을 부여한다.[11]

4.2. 자질부여 및 정리

일차적으로 구문분석되어 나온 결과들은 그대로 사용하기에는 아직 거칠고 여러 문제들이 있다. 이제 실제로 각 번역어들에 앞 절에서 논한 자질들을 부여하고 구문분석에서 나타난 문제들 -잘못된 품사부여, 번역어 구문에서의 오류 등-을 교정하는 작업이 필요하다. 이 부분은 완전히 자동화되기 어려우며 인간의 간섭이 절대적으로 요구된다. 따라서 구문분석이 체계적으로 이루어 졌다면 이 단계에 많은 시간이 소비된다. 또한 이 단계에서 중요한 작업은 다의어와 동형어에 따른 역어구분이다. 앞에서 살펴본 대로 완전한 동형어의 경우는 텍스트형태의 사전에서도 그 항목이 달리 기술되어 있기 때문에 여기서도 다른 항목으로 분리되어 있다. 문제는 다의어에 따른 대역어의 구분인데 사전의 자동구문분석에 의한 다의어의 구분은 앞에서 지적한대로 이 논의의 핵심이 아니기 때문에 우선 사전에 기술된 대로 패턴매칭되어 위의 기술에서처럼 ';'로 분리된다. 이 다의어의 더 세밀한 분류는 다른 전자사전으로의 통합(4.3절 참조)에서 사전기술자에 의해 이루어진다.

실제로 자질부여는 2절에서 살펴 본 특질과 관련된 원칙들에 따라 이루어진다. 이 경우에 잘못된 표현을 교정한다. 특히 한 단어로 된 번역어가 있고 그 동의어로 구표현들이 있을 때 구표현 보다는 한단어로 된 표현을 채택한다. 이러한 교정작업은 위의 구조에 새로운 표시자를 도입하여 그 교정과정을 확인할 수 있게 된다.

(26) 가. $H 독립하다
$C Verb
$M become independent;stand on one's own legs;separated from;separate from;rely on oneself;help oneself;

10) 세종전자사전에서 제한하고 있는 기능동사로는 '하다, 되다, 시키다, 당하다, 주다, 받다, 있다, 없다' 등이다. 따라서 한영사전에서 표제어항목 밑에 '~하다/되다/시키다/당하다/주다/받다/있다/없다' 등의 구조는 자동적으로 표제어+기능동사의 결합으로 새로운 표제어로 구성한다.

11) 다음 절의 예 (26) 다. '여권' 참조.

```
    support oneself;
$I become independent<phv>;separate
    from<phv>;
$$
나. $H 얼빠지다
    $C Verb
    $M be absent minded;be stupefied;
    $I absent minded<pred>;stupefy<pass>;
    $$
다. $H 여권
    $C Noun
    $M passport;womens rights;woman
        suffrage;
    $I passport;women's right<nc plur>;
    woman suffrage<nc>;
$$
```

새롭게 자질부여된 그리고 정리된 번역어들이 $I에 기술된다. 나중에 사전이 데이터베이스화 될 때는 $M은 중간단계로 무시되고 실제로 입력되지 않는다.

이런 정리작업은 번역사전의 적격성을 보장하고 그 질을 높인다는 점에서 중요하다. 따라서 형식적인 많은 부분을 구문분석에 의해 정리하여 교정하는 사람의 노력을 최소화할 필요가 있다. 실제로 일차적으로 분석된 사전에서 여전히 남아 있는 구문상의 오류는 대략 40% 정도에 이른다. (위의 기술에서 $M이 처음으로 형성되는 수와 $I의 증가비율에 의해) 이 오류는 사전기술자의 정리작업에 의해 교정되는데 이 단계에서 다의어는 여전히 리스트로 기술되어 확실히 구분되지 않는다.

4.3 기존사전으로의 통합

이렇게 구축된 사전은 깊은 언어적 지식을 요구하지 않은 자연언어처리시스템에 그대로 사용될 수 있다. 실제로 단어 대 단어 번역시스템에 사용될 수 있는데 뉴멕시코 주립대학 CRL의 한-영 단어 대 단어 번역시스템 (glossary-based MT)에 적용되었다. 단어 대 단어 번역시스템은 두 언어의 깊은 통사, 의미적 분석없이 형태소 분석과 최소한의 구조적 차이 - 어순, 활용, 수식정보 - 만으로 구축된 번역시스템으로 간단한 요약, 정보검색/요약 시스템과 번역시스템이 결합된 통합시스템에서 빠른 속도의 그리고 번역의 질은 높지 않더라도 의미가 통할 수 있는 시스템 구축에 적용된다. 새롭게 정리된 번역사전을 채택하였을 때 기존의 정리안된 사전과의 번역의 질은 정확히 비교하기 어렵고, 또 다의어에 구분 및 다른 의미적 정보에 따른 역어의 선택이 이루어지지 않고 가장 앞에 있는 것, 그리고 짧은 것을 선택하는 문제가 있지만 형태적 정확성, 그리고 고유명사와 단일어로 된 대역어 선택 등에 의해 가독성은 높아지게 된다.

다의어의 구분은 이 사전을 다른 사전과 통합할 때 적절히 사전기술자에 의해 이루어질 수 있다. 이런 통합관점에서 세종사전에서의 역어정보로 통합가능성에 대해 살펴보자. 세종사전은 어휘문법(lexicon-grammar)에 기초하여 어휘정보를 구축하고 있는데 각 어휘항목들이 동형어, 다의어 관점에서 구별되어 기술되고 있다. 따라서 간략화된 세종사전의 '먹다' 기술은 다음과 같다.

(27) 먹다

```
<entry>
    <selRst>N0=연장(대패|톱)      N1=(나무)</selRst>
    <selRst>N0=(돈|비용|경비) N1≈추상적대상(일|작업)</selRst>
<entry>
    <selRst>N0=인간 N1=귀 N2=소리(폭발|다이나마이트|폭탄)</selRst>
<entry>
    <selRst>N0=(기름|물감|풀|먹) N1=(장판|옷|종이)</selRst>
    <selRst>N0=(화장|로션) N1=신체(얼굴|피부)</selRst>
    <selRst>N0=(벌레|좀) N1=사물(옷|나뭇잎)</selRst
    <selRst>N0=(장판|옷|종이) N1≈(물감|풀|먹)</selRst>
    <selRst>N0=신체(얼굴|피부) N1=(화장|로션)</selRst>
    <selRst>N0=(옷|과일|일)      N1=(벌레|좀)</selRst>
<entry>
    <selRst>N0=인간   N1=음식(음식|밥|국|물|빵)</selRst>
    <selRst>N0=인간 N1=식사(아침|점심|저녁|끼니)</selRst>
    <selRst>N0=동물   N1=먹이(모이|여물|풀)</selRst>
    <selRst>N0=인간 N1=(담배|술|커피)|(약|마약)</selRst>
<entry>
    <selRst>N0=인간|단체    N1=(돈|뇌물) N2=인간|단체</selRst>
<entry>
    <selRst>N0=인간|단체 N1=(골|주먹|집) N2=인간|단체</selRst>
<entry>
    <selRst>N0=인간 N1≈추상적대상(이익|돈|이자|집세|구전|재산)|사물
        </selRst>
    <selRst>N0=사물(자판기)     N1=(돈|동전)</selRst>
    <selRst>N0=인간 N1=(나이)</selRst>
    <selRst>N0=인간   N1=(등수|등급|점수|벌점)</selRst>
    <selRst>N0=사물(기계|차) N1=(연료|기름|전기)</selRst>
    <selRst>N0=인간|국가 N1=인간|국가|장소(지역|서울|산)</selRst>
    <selRst>N0=인간   N1=(구류|벌점)</selRst>
<entry>
```

<selRst>Npr1=(옥|편잔) N0=인간 N3=인간</selRst>
<selRst>Npr1=(마음) N0=인간</selRst>
<selRst>Npr1=(집) N0=인간 N2=인간</selRst>
<selRst>Npr1=(애) N0=인간</selRst>
<selRst>Npr1=(편|짝) N0=인간|단체N2=인간|단체</selRst>
<selRst>Npr1=편 N0>{Ni-와 Nj} Ni=인간|단체 Nj=인간|단체
</selRst>
<entry>
<selRst></selRst>

이 사전 기술에 의하면 '먹다'는 <entry>로 표시되는 9 가지의 동형어와 각 <entry>에서 <selRst>로 표시되는 다의어들로 표시되어 있다. 일차적으로 정리된 번역어는 각 다의어의 <trans>라는 항목에 기술된다. 이 부분의 정보를 채워넣기 위한 자동화방법을 생각해 볼 수 있으나 부정확성이 그대로 남아있고 현재 이 사전이 직접 사람의 손으로 기술되고 있기 때문에 각 항목 기술시 대역어 사전에서 추출되고 형태적 자질이 첨가된 정보를 그대로 가져다 넣는 것이 효율적이고 정확하다. 이런 과정에 의해 번역정보가 명시되는 몇 예를 살펴보면 다음과 같다.[12]

(28)

<selRst>N0=연장(대패|톱) N1=(나무)</selRst>
<trans>bite well</trans>
<selRst>N0=(옷|과일|일) N1=(벌레|좀)</selRst>
<trans>worm-eaten{pred}</trans>
<selRst>N0=인간 N1=(구류|벌점)</selRst>
<trans>given{pass}</trans>
<selRst>Npr1=(마음) N0=인간</selRst>
<eg>그 장관은 다시는 돈을 받지 않겠다고 마음을 굳게 먹었다.</eg>
<trans>make up {poss} mind{phv}</trans>

5. 결론

지금까지 glsossray에 기초한 시스템을 위한 빠르게 구축될 수 있는 한영 대역어 사전을 위해 기존의 인쇄사전을 구문분석하고 정리하여 필요한 자질을 부여하는 것에 대해 논의했다. 이런 작업은 영어의 생성 관점에서 정확한 형태를 도출하기 위해 필수적인 작업이나 많은 연구에서 간과되고 있는 부분이다. 기존의 사전기술에 근간을 두고 또 영어의 특성에 따라 필요한 자질을 크게 명사와 서술어 관점에서 설명하였다. 그러나 이런 연구의 일차적인 초점은 형태/통사적 적격성이기 때문에 의미적 모호성 해결은 다른 사전과의 통합, 특히, 다의어와 동형어와의 구분이 엄격한 세종전자사전과의 통합에서 이루어 진다. 이 사전은 뉴멕시코 주립대학 CRL에서 통합(unification)에 바탕을 둔 자질구조시스템에서 사용되기 때문에 여기서 기술되지 않은 많은 다른 자질들이 번역시스템에 필요한 자질구조로 정의되어 있으나 자세히 논의하지 않았다. 이런 자질은 자질구조시스템뿐만 아니라 다른 시스템에도 쉽게 적용될 수 있으며 생성관점에서 필요한 정보는 비단 이런 자질뿐만 아니라 다른 방법으로도 기술될 수 있다.

12) 여기서는 <trans>기술을 위해 몇 예만을 추출하였고 또 대역어의 자질기술로 쓰이는 <>이 xml tag기술과 겹치기 때문에 []로 바꾸었다. 마지막 entry는 관용적 표현 (까 먹다 등)이라 <SelRst>가 비어 있다.

참 고 문 헌

[1] 21세기 세종계획 전자사전 개발분과, 연구보고서, 문화관광부/국립국어연구소, 2002.
[2] Nirenburg Sergei, Ontology-based Cross Lingual Information Retrieval, Unpublished Proposal, Computing Research Laboratory, New Mexico State University, 1988.
[3] Shin Hyopil and Spencer Koehler, A Knowledge-Based Fact Database: Acquisition To Application. Proceedings of the International Conference KBCS2000, National Centre for Software Technology, India, 2000.
[4] 신효필, 지식기반(Knowledge-based) 질의응답시스템: 사실자료(Fact Database) 구축을 중심으로, 인지과학 13-1, 한국인지과학회, 2002
[5] 옥철영, 한-영 기계번역을 위한 구 단위 변환사전, 서울대학교 박사학위 논문, 1993.
[6] 엣센스 한영사전, 민중서림, 1993.
[7] Boguraev, B. and Levin, B., Models for lexical knowledge bases. Pustejovsky ed.: Semantics and the Lexicon, Kluwer Academic Publisher, 1993
[8] Boguraev, B. Special Issue on computational lexicons. International Jorunal of Computational Lexicography 4, 1991.
[9] Copestake, A, An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary, Proceedings of the First International Workshop on Inheritance in Natural Language Processing, 1990.
[10] 최병진, 어휘정보구축을 위한 사전텍스트, 언어와 정

보 6-2, 한국언어정보학회, 2002.
[11] 허정, 옥철영, 사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결시스템, 정보과학회 논문지: 소프트웨어 및 응용, 28-9, 2001.

접 수	2003년 1월 24일
게재승인	2003년 5월 21일