論文 88-25-10-7

PC를 이용한 일·한 번역시스템 ATOM의 개발에 관한 연구(Ⅱ)
-구문해석과 생성과정을 중심으로-

(Development of Japanese to Korean Machine Translation System ATOM Using Personal Computer Ⅱ -Syntactic/Semantic Analysis and Generation Process-)

金 榮 暹,* 金 漢 宇,* 崔 炳 旭*

(Young Sum Kim, Han Woo Kim and Byung Uk Choi)

要 約

구문 해석과정에서 동사가 갖는 필수격을 기준으로 격 프레임을 구성하여 격 구조를 생성하며, 형태소 해석 결과에 단문을 기준으로 한 부분 문법을 재귀적으로 적용함으로 해서 구문 의미 해석을 수행한다. 또한 역어 생성과정에서 일본어 조사처리의 중요성을 고려하여 중요 조사의 애매성 해소와 역어 분류를 위한 독립적인 프로시쥬어를 기술하여 효율을 제고한다. 그리고 일본어 종결구의 처리를 위해서 동사와 조동사의 복합 가능성을 고려한 생성 테이블을 작성하여 형태소와 구문 해석정보에 의해 일의적(一意的)인 결정을 행하여 보다 자연스런 역어의 생성과 생성과정의 간략화를 도모하였다.

Abstract

In this paper, we describe the syntactic and semantic parsing methods which use the case frames. The case structures based on obligatory cases of verbs. And, we use a small set of partial-garammar rules based on simple sentence to represent such case structures. Also, we enhance the efficiency by constructing independent procedure for particle classification and ambiguity resolution of major particle considering the importance of Japanese particle process in the generation. And we construct the generation table considering the combination possibility between the verbs and auxiliary verbs for processing the termination phrase. Therefore we can generate more natural translated sentence according to unique decision with information of syntactic analysis and simplify the generating process.

Ⅰ. 서 론

실용화를 목표로 하는 번역시스템의 구현에서 프로세스의 효율과 번역문의 품질을 결정하는 것은 구문 해석, 의미 해석만이 아니고 번역 사전의 구성과 관용구의 처리 방법등 시스템 전체의 밸런스이다. 즉

*正會員, 漢陽大學校 電子通信工學科
(Dept. of Elec. Comm. Eng., Hanyang Univ.)
接受日字 : 1988年 5月 31日

대상 언어에 대한 문법 기술의 기준이나 해석 알고리듬의 적용 방식등에 의하여 프로세스의 효율에는 커다란 차이가 있다. 또한 시스템적인 측면에서 형태소, 구문, 의미등 번역의 제 과정이 해석 알고리듬이나 처리 개념에 의해 명확하게 구분되는 것은 아니며, 언어 정보도 프로세스의 순차적인 흐름보다는 시스템 전체의 효율면에서 적용 시점이 고려되어야 한다.

본 논문에서는 다음과 같은 가정하에서 실용화를 목표로 하는 번역 시스템을 설계 한다. 일본어와 한국어와 같이 계통적, 형태적으로 동일 어군에 속하는 언어간의 번역은 심층의 의미 해석을 일부 유보하여도 구문의 의미에 근사한 적절한 역어를 생성한다면, 후편집의 개제를 전제로 하여 실용 시스템을 구현할 수 있다. 즉, 구문, 의미 해석등의 제 과정에 대한 부분적인 완성도보다는 시스템 전체의 효율과 미시적인 언어 현상의 기술에 중점을 둔 번역 과정을 구현한다.

최근 자연언어의 연구에서 격 구조(격 프레임) 를 도입한 다수의 구문 의미해석 시스템의 구축이 이루어지고 있다. 일본어, 한국어와 같은 자립어에 부속어가 접속하여 문법적인 관계를 표현하는 첨가어에 속하는 언어를 대상으로 하는 번역에 있어서 구문 의미 해석에 격 프레임을 이용하는 방식이 상당히 유력해 보인다.[12,13]

본 시스템에서는 형태적인 애매성을 해소하기 위하여 자립어와 조사간의 애매성이나 조사의 다의성 처리는 해당 어휘를 갖는 조사를 지표로 프로시쥬어를 부가한 독립적인 처리 루틴을 이용한다. 한편, 명사의 의미 표지(semantic marker)는 처리 효율을 위하여 3개로 제한하여 기술하며 부여된 의미표지에 빈도수에 관한 가중치를 부여하여 보조적인 선택 정보로 이용한다. 이때 의미표지의 선택 정보는 부분 명사구가 갖는 의미 표지의 대표치를 결정하며 격구조 할당 정보로 이용된다. 또한 의미표지에 부분-전체, 상위-하위등의 관계를 설정한 네트워크를 구성하여 명사구의 의미 조합과 병렬구의 의미 관계를 처리한다.

부분 문법에 의해 구성된 각각의 구문 해석 정보의 의미적 관계의 인식은 동사의 격 패턴과 구문 해석 결과에 표현된 격 관계의 정합성 판단과 격이 갖는 의미표지의 정합을 검색하며 격의 기준은 동사 사전에 기술된 필수격을 중심으로 행한다.

생성과정에서 구문 해석과 이에 준한 역어의 선택에서 일본어 조사의 역할은 매우 중요하다고 할 수 있다. 그러므로 동사의 표층격을 기준으로 하는 격구조 출력과 연계시켜 조사 역할의 애매성과 역어 선택의 다양성을 해소할 수 있는 독립적인 조사 처리 프로시쥬어를 기술하여 보다 적격한 번역 결과를 얻도록 하였다. 또한, 활용 용언과 조동사의 접속시 보다 고품질의 역문을 생성하기 위하여 형태소 정보를 지표로 하는 독립적인 생성 테이블을 한국어의 조사 테이블과 한국어 용언(보조용언을 부가한)의 활용 어미 테이블에 부가하여 구성하였다.

본 논문에서는 실용화를 목표로 구성된 일·한 번역 시스템 ATOM(jApanese To kOrean Machine translation)의 격 프레임을 이용한 구문 의미 해석 방식과 한국어의 생성 방식에 대하여 논한다. 시스템에서 구성한 사전의 규모는 약 2만어이며, 시스템의 검증에서 사용한 예문은 과학 기술 관계 문헌과 컴퓨터 관계 매뉴얼로 제한 하였다.

Ⅱ. 구문 의미 해석

1. 명사의 의미 표지 할당

번역 시스템의 구성 작업에서 명사의 의미표지의 부여는 상당한 난점을 갖는다. 의미표지의 할당이 연구자의 자의적인 부여 작업에 의존하기 때문에 사전의 크기에 상대적으로 일관성의 결여가 발생하기 쉬우며, 언어의 다의적인 특성에 의한 할당의 미비와 격 프레임과의 균형을 위해서 다량의 예문에 대한 검증 작업을 통한 정비가 요구되는 분야이다.[8,10]

본 시스템에서는 번역 대상 영역의 한정이라는 전제와 처리 효율의 제고에 중점을 두어 대상 명사의 의미표지를 3개 이내로 제한하여 기술한다. 의미표지 기술의 기준은 ICOT 분류에 준하며 각각의 대표 분류에 하위 분류가 속하는 것으로 하여 의미표지에 계층성을 부여하였다. 또한 명사에 부여된 의미표지는 각각에 사용 빈도를 고려한 임의적인 heuristic을 부가하여 순서적인 할당 가중치를 갖게 한다.

2. 부분 문법 규칙

일본어 구문은 일반적으로 구의 종단에 구문과 의미상의 대표 문절을 갖으며 또한 문두측의 대표 문절이 문미의 대표 문절에 의존하는 (係り受け) 관계를 갖으며, 두개의 구 접속에 의한 복합구의 생성은 의존 관계가 존재하지 않을 때에 가능하게 된다. 본 시스템에서는 일본어 입력문의 구문 해석을 단문을 기준으로 하여 문절간의 의존 관계를 중심으로 처리하며, 의존 관계 이외의 병렬, 동격 구조등의 해석은 독립된 프로시쥬어를 기술하여 부분 문법규칙상에 부가하여 행한다.

한편, 격문법에 기반을 둔 격 프레임을 이용한 구문 의미 해석은 동사의 표층격 구조의 정보를 중심

```
oo(thing, object) [49]
  of(nation, organization) [107]
  ov(living object) [6]
      oa(animal) [13]
          oh(human, role, profession) [87]
      op(plant) [9]
          ob(animal) [2]
  os(non-living substance) [145]
          on(natural) [64]
          om(artificial) [1672]

po(phenomenon) [229]
  pn(natural phenomenon) [143]
  pa(artificial phenomenon, experiment) [277]
  ps(social, phenomenon) [35]
          ph(event, happenings) [43]
          pe(politics, economics) [32]
          pc(custom, social convention) [21]
  pp(power, energy, physical objct) [663]

to(time, space) [25]
  ts(space, topography) [428]
  tt(time) [153]
          tp(time point) [9]
          td(time duration) [20]
          ta(time attribute) [22]
```

그림 1. 명사의 의미표지 분류와 본 시스템에서 구성한 사전상의 의미표지 빈도의 일례

Fig. 1. An example of semantic marker and its frequency.

으로 수행한다. 실제, 시스템의 구축 과정에서 명사구에 관계된 조사 표현을 동사의 표층격에 관계하는 격 프레임에 적절하게 할당시킬 수 있다면 부분 문법의 구조와 표층격 구조를 갖는 격조사 혹은 격조사 상당 표현에 의하여 격 프레임에 의한 구문 의미 해석이 수행될 수 있기 때문에 시스템이 상당히 간결하게 기술되었다고 보여진다. 그러나, 격조사를 포함하지 않은 명사구의 격할당 문제등이 우선적으로 고려되어야 하며, 격조사 상당 표현에 대한 표층격 부여가 문제로 된다.[2,3,4,13]

본 시스템에서는 구문 해석 과정에 단문을 기준으로한 부분 문법을 작성하고 이를 반복 적용하여 복문의 구문 의미 구조를 추출한다. 문법규칙의 일반형은 다음과 같다.

a→b, c d 〔: m〕 (: m은 optional로 특정 품사 정보나 활용정보가 부가된다.)

단어 또는 구 b, c가 d의 프로시쥬어의 수행에 의해서 결정되는 구문 의미 구조, 즉 a라는 상위 구조를 형성한다는 의미이며, 이 때 〔: m〕은 프로시쥬어에 선택적으로 부가되는 정보이다. 본 시스템에서 문법규칙은 처리 효율을 위해서 문맥 자유 문법의 형으로 구성되지만, 문법규칙에 구문 의미 관계를 결정하기 위한 프로시쥬어를 기술하여 격의 추출과 구문의 의미적 관계를 결정한다.[1,6,12] 문법규칙은 bottom-up 방식에 의해서 적용되며, 주 목적은 동사의 필수격 패턴을 기점으로 하는 명사구의 격구조 추출에 있다. 그림 3는 사전상에 기술된 동사 격 프레임의 일례이다.[2,13]

문장내에서 격 표시는 격조사에 의해서 할당되며 부조사, 계조사등의 비격조사는 주제의 표시와 한정사적으로 이용된다. 그러므로 격조사를 포함하지 않

```
nc→ np, の, resolution[5] → (1)
nc→ n, , resolution[0]
np→ n, , resolution[0], , 1    /* 1 : 대명사 */
np→ n, , resolution[0], , 2    /* 2 : 고유명사 */
np→ m, u, resolution[15]       /* m : 수사, u : 조수사 */
np→ n, n, resolution[1]
n  → a, n, resolution[23], , 3 /* a : 형용사, 3 : 형용사 연체형 */
```

그림 2. 부분 문법의 일례

Fig. 2. An example of partial grammar rule.

```
/* lexicon          case frame                 suffix   entry      a v s   origin          */
/* verb patt                                   1  2     3          p p t                   */
"あらわ"     represent   act obj    inst      が を     で         3 1 3   "あらわす"
"ぶんかい"   decompose   act obj1   obj2      が を     に         2 2 10  "ぶんかいする"
"ぶんぱい"   distribute  act obj               が を                2 1 10  "ぶんぱいする"
"はつびょう" express     act the               が を     について   2 1 10  "はつびょうする"
"はた"       achieve     act obj               が を                3 1 3   "はたす"
"はじめ"     start       act                   が を                3 1 9   "はじめる"
"はじま"     start       act                   が                   3 0 8   "はじまる"
"表わ"       represent   act obj    inst      が を     で         3 1 3   "表わす"
"分解"       decompose   act obj1   obj2      が を     に         2 2 10  "分解する"
"分配"       distribute  act obj               が を                2 1 10  "分配する"
"分散"       distribute  act obj               が                   2 0 10  "分散する"
```

그림 3. 동사의 격 프레임 구성의 일례

Fig. 3. An example of case frame for verb.

는 명사구와 술어의 격관계는 의미를 고려하지 않고서는 결정할 수 없는 경우가 많다. 이러한 문제의 해결을 위해서 비격조사의 대치격의 할당등 다수의 방법이 제시되어 있지만, 일·한 번역이라는 시스템적인 견지에서 볼 때 번역 대상 언어의 유사성등에 의해서 상당한 리던던시를 갖는다. 본 시스템에서는 효율적인 역어 생성을 고려하여 조사 자체를 기준으로 한 해석을 수행한다.[7,13]

그림 2는 명사구에 대한 부분 문법의 부분적인 예이며 본 시스템에서 작성한 문법 규칙은 107개이다. (1)규칙에 부가된 프로시쥬어는 조사の의 구문 의미적 역할을 결정한다. 일례로 명사구에서의 조사の의 용법은 다음과 같이 분류되며,

(1) "から", "まで", "について"등의 격조사 부조사와 함께 이용되어 연용수식어를 연체수식어화 한다. -からの, -についての
(2) "-上の", "-の上の"등과 같이 피수식 명사의 위치를 표현한다. -の上の, -内部の
(3) 피수식 명사가 수식 명사의 속성인 것을 표현한다. テイスクの記憶容量
(4) 수사의 후속에 위치하여 피수식 명사의 속성값을 표현한다. 32ビツシのレジスタ
(5) 수사의 후속에 위치하여 피수식 명사의 갯수를 표현한다. 2個の直列ポート
(6) 수사의 후속에 위치하여 피수식 명사의 량을 표현한다. 2バイトのテータ
(7) 피수식 명사가 さ변 명사 혹은 동사의 연용형이 명사화한 것이며 수식 명사가 이 명사의 의미상의 목적어인 것을 표현한다. 辭典の檢索
(8) 수식 명사가 동사의 연용형이 명사화한 것이며 피수식 명사와 동격 관계로 결합한 것을 표현한다. 讀み出しの場合
(9) 피수식 명사가 동사의 연용형이 명사화한 것이며 수식 명사가 이 명사의 의미상의 주어인 것을 표현한다. 方法の改善
(10) 대명사+"の" 및 "2の補數"등과 같은 특정의 표현에만 사용한다.
(11) 기타.

문법 규칙에 부가된 프로시쥬어는 조사の의 형태적인 접속 관계와 의미표지의 검색에 의해서 명사구내의 용법을 결정한다. 즉, 수식 명사가 수사일 경우 해당 조수사와 피수식 명사의 속성의 일치를 검사하여, 즉, 의미표지 mn(number), mu(unit) 등의 조합에 의해 (4),(5),(6)의 용법을 결정한다. 또한 수식 명사가 さ변 명사이면 (8)의 용법으로, 피수식 명사가 속성 명사인 경우에는 피수식 명사와 수식 명사의 속성의 일치를 검사하여 존재하면 (3)의 용법으로 한다. 피수식 명사가 さ변 명사인 경우 さ변 명사의 필수격에 목적격이 있으면 (7)의 용법, 목적어가 없고 주어가 있으면 (9)의 용법이다. 따라서 (9)의 용법보다 (7)의 용법을 우선적으로 처리한다. (1),(2) 의 용법에 대해서는 "からの, についての, 内部の, の上の" 등을 하나의 단어로 간주하여 사전에 등록하며, (10)의 경우도 특정의 용례를 사전상에 기술하여 처리한다. 한편, 이러한 부분 문법 상의 프로시쥬어의 출력 결과는 조사의 역어 결정 프로시쥬어에 직접 연계되어 있다. 조사 の의 처리 결과는 그림 6에 보이는 생성 과정의 역어 결정 프로시쥬어의 입력으로 된다.

일반적인 일본어의 구문 의미 해석을 수행할 때 병렬구의 처리는 상당한 난점을 갖는다. 그러나 역어 생성 과정을 고려하면, 한국어 구문과 일본어 구문과의 비교에 의해서 일본어 상에서 병렬구를 이루는 の와 と등의 조합은 병렬구의 범위 인식등의 복잡한 프로세스에 의하지 않고도 상기한 부분 문법적인 처리와 사전 기술에 의해 생성된 역어 자체~생성의 적격성에 문제로 되는 병렬성을 갖는 표현은 그림 4에 보이는 조사의 용례이며, 구문정보에 의해 적격한 출력을 얻기 위해 해당 조사의 사전과 프로시쥬어상에 병렬 정보의 해석 프로세스를 부가하였다. 이 때 해석된 병렬성의 조사는 역어의 생성을 기준으로 처리된다. 즉, 구문 구조의 상위에 관계없이 동일한

```
부조사か → 둘 중의 하나를 선택할 때
부조사だって → 대상이나 사실의 나열
부조사だの → 대상, 사실의 열거
부조사とか → 대상이나 사실의 열거 혹은 대비
부조사なり → 대상의 열거
계조사よ → 대상의 나열
부조사やら → 대상의 나열
격조사に → 대상의 나열
격조사の → 대상의 나열
격조사や → 대상의 나열
연어にしろ → 대상의 나열

か;ます/です//, 347802, 347801, k$ます;28)/k$です;28)/
    k$DEF;04)까/q$0;09)/q$10;18)/q$20;18)
だって;01,,,para[1]
とか;#いう/#つて/#言つて/#言う, 34780891,,q$0;31)든
    가/q$10;00)느니/q$20;18)
なり;125603, 47801,,p$10;19)/p$DFE;31)채로/q$,para[3]
に;う/よう, 125613, #なる/#なり, 01760, h$なる;01)/h$なり/
    01)/k$う;04)텐데/k$よう;04) 텐데/h$0;nvsjo_tab/h$'16;
    21)/p$11;00) 건데, para[2]
にしろ;,,,q$ALL0;00)든, para[7]
```

그림 4. 병렬성을 갖는 조사의 일례(と 제외)와 관계 생성사전

Fig. 4. An example of particle related to coordinate structure and related generation dictionary.

역어가 생성될 수 있으면 해석은 수행하지 않는다. 그림 4의 하단에 부가된 사전은 형태소 검색부가 부가되어 있는 점을 제외하면 그림 6의 생성 사전과 동일한 기능을 갖는다.

3. 문의접속 처리

일본어 문에서 문과 문의 접속은 격조사와 접속조사에 의해서 행하여 진다. 접속사는 문과 문, 어구와 어구등을 접속하는 것에 이용되지만 본 논문에서는 시스템상에서 구현한 문의 접속 처리를 중심으로 기술한다.[9,10,12]

문접속의 일반적인 형태는 다음의 3가지로 분류할 수 있다.

1) 연용중지+문
2) 연용중지+접속사+문
3) 동사구+접속조사+문

1)의 연용중지법은 병렬, 추이연속, 수단, 방법, 역접등 여러가지로 이용되지만 정확한 의미의 식별을 위해서는 문의 전후 관계를 고려하지 않으면 안된다. 그러나 실제 현상의 시스템에서는 상당히 곤란한 문제이며, 구문 관계의 인식에 의한 적절한 역어 선택이 최선의 과제라고 할 수 있다.

접속사는 절과 절을 대음 관계로 접속하는 등위 접속사와 부사절을 이루는 종속 접속사로 분류된다. 실제 일본어에서 양자의 구별은 명확하게 구분되지 않지만 본 시스템에서는 한국어 접속사의 분류와 역어의 대응에 의하여 분류한다. 그림 5는 일본어의 접속 조사와 접속사의 분류의 일례이다.

등위 접속사에 의한 문접속은 "문1+접속사+문2"의 형으로 표현할 수 있다. "문1"과 "문2"는 대등의 관계인 까닭에 해석 과정에서는 "문1"과 "문2"를 독립적으로 해석한다. 한편, 종속 접속사를 포함한 문의 경우 접속사에 의해서 도입되는 절은 부사절로서 주절중에 포함된 삽입문의 형태로 나타난다. 해석 과정에서는 먼저 종속절을 하나의 완전한 문으로 해석해서 이를 접속사와 접속하여 부사절로 하고, 다음에 부사절을 포함하는 주절을 해석하여 전체의 구문 해석을 얻는다. 전체 구문은 〔〔접속사+종속절〕+주절〕로 구성되며, 〔접속사+종속절〕의 부분은 문수식의 부사절을 형성해서 주절을 수식한다.

4. 포유문(삽입문)의 처리

문내에 또하나의 문이 삽입되어 있는 구조를 포유구조라 한다. 본 시스템에서는 일본어 문상에서 포유문의 가장 일반적인 형태인 연체수식의 역할을 하는 포유문의 처리에 주안을 두어 프로세스를 기술하였다.[9,10,12]

lexicon	class(mean)	exp.	conj. class /* conjunction */
また	parallel	and	coordinate
そのうえ	add	because	
さらに	add	because	
かつ	add	because	
しかも	add	because	
あるいは	selection	or	
または	selection	or	
もしくは	selection	or	
だから	order	so	
したがって	order	so	
それで	order	so	
ゆえに	order	so	
しかし	reverse	but	
そして	transition	aud	
それから	transition	and	
し	parallel	and	coordinate /* particle */
たり(だり)	parallel	and	
から	reason	because	subordinate
ので	reason	because	
けれども	reverse	but	coordinate
が	reverse	but	
のに	reverse	though	subordinate
ものの	reverse	though	
ながら	reverse	though	
ても(でも)	yield	even if	
とて	yield	even if	
て(で)	transition	and	coordinate
ば	condition	if	subordinate
と	condition	if	
ながら	synchronous	while	
つつ	synchronous	while	

그림 5. 접속사와 접속 조사의 의미 분류
Fig. 5. Example of conjunction and conjunctive particle meaning classification.

연체수식의 역할을 갖는 일본어 표유문은 크게 관계적 표현과 동격 표현으로 분류할 수 있다.

a : 관계적 표현 : 피수식 명사는 일반적으로 "が, を, に"의 표층격 표현을 내포한다.

1 : 포유문의 술부가 갖는 격 프레임 슬로트에 해당하는 격요소의 일부가 생략되어 있으며 생략된 격요소에 피수식 명사가 대응된다. 한국어 문장의 포유복문중에서 주어절 혹은 목적절 포유에 해당한다.
〔ex : UNIXがもつ互換性⇨もつ의 목적격이 생략되어 있고, 피수식 명사가 이에 대응한다.〕

2 : 피수식되는 명사가 포유문의 격 요소를 수식 속성으로 소유한다. 즉, 피수식 명사가 상태를 의미하는 술어에 의해 부분, 속성, 동작등의 의미 표지를 갖는 명사의 수식을 받을 때.
〔ex : 効率が好い算法⇨好い의 격요소는 생략되지 않고 算法의 効率이 좋다라는 의미 관계를 갖는다.〕

a' : 수동태의 술어를 갖는 문장에서는 표층의 が 격

을 포유하는 경우가 존재한다.

b : 동격적 표현

1 : 피수식 명사가 동격 관계를 구성하는 속성을 갖는 명사일 때

〔ex : AUTOEXEC. BATを實行する場合〕

2 : 피수식 명사가 형식 명사로 명사절을 구성할 때

〔ex : エラーメッセージが表示されること〕

동격 표현의 포유문을 구성하는 명사는 다음과 같이 특정의 보통명사, 형식 명사에 한정된다. 포유문의 인식과 처리를 위해서 해당 명사의 사전에 정보를 기술한다.

1 : 보통 명사 : 場合, 必要, 事實, とき, 目的, 關係, 能力, 狀態, …

2 : 형식 명사 : こと, の, よう, 點, 方, …

포유문의 해석은 a, b 두 개의 프로세스로 구성된다. a, b의 프로세스의 구분은 b의 명사를 기준으로 일의적(一意的)인 결정을 행하며, a의 1, 2 프로세스는 술어의 격 요소 생략여부를 기준으로 판단한다. 즉, 1의 프로세스가 2의 프로세스에 대해 우선적으로 수행된다. b의 1, 2는 사전상의 정보를 기준으로 결정한다.

실제, 포유문의 해석은 의미의 영역에 깊이 관련되어 있기 때문에 구문적인 접근 방법으로는 근본적인 제약을 갖는다고 보여지지만 본 시스템에서는 구문 정보를 세분화시켜 최선의 출력을 얻도록 하였다.

Ⅲ. 역어 생성

본 시스템에서 구성한 역어 생성은 한국어 활용 어미 테이블 참조와 조사 테이블 그리고 부가적으로 구성한 활용어와 조동사의 접속 테이블의 참조에 의하여 이루어진다. 구문 의미 해석 과정에서 결정된 문의 구문 의미 구조는 동사의 격 프레임을 중심으로 하는 연계 리스트 형으로 구성된다. 이때 격 프레임상에는 명사구 혹은 명사절의 표층격 표현을 기준으로 연계되어 있기 때문에 생성 과정의 단위는 명사구의 생성을 기준으로 수행한다.[14-17]

역어 생성과정에서 명사의 생성은 격 프레임에 할당된 해당 명사의 의미표지 정보에 의해서 이루어진다. 또한 동사등의 활용어 정보도 격 프레임 상에 할당된 의미표지를 기준으로 역어가 선택된다. 이때 활용어 어미의 결정은 형태소 해석 과정에서 추출된 일본어 활용어의 활용형을 기준으로 선택된다. 활용 테이블은 동사와 형용사의 두부분으로 구성되며, 각각의 테이블은 2차원 배열의 형으로 되어 있다.[18] 배열의 행에는 각각 기본형, 평서형, 관형형(현재, 미래), 명사형(정법, 미정법), 명령형, 가정형, 부사형등의 순서로 기술되어 있으며, 열에는 규칙, ㄷ불규칙, ㄹ불규칙, 러 불규칙등 활용형이 기술되어 있다. 한편 조사 테이블은 전치하는 명사의 받침 유무에 따라 2개의 배열로 구성된다. 그리고 부가적인 생성 테이블로 보다 자연스러운 역어의 생성을 위하여 종결구 처리에 중점을 둔 즉, "활용어+조동사+조동사"의 역어 생성 테이블을 구성하였다.

한국어와 일본어등의 첨가어의 역어 생성 과정에서 가장 중요한 부분은 조사의 생성과정이다. 본 시스템에서 조사의 역어 생성은 독립된 조사 생성 사전과 역어 선택 프로세스에 의하여 수행된다. 일반적인 조사의 역어 선택은 다음과 같은 4개의 선택적 프로세스로 구성된다.

1 : 접속 전후의 특정 단어의 접속정보

2 : 접속 전후의 품사 정보

3 : 접속 전후의 의미 소성

4 : 접속형의 격 구조 정보

즉, 조사의 역어 사전에는 각각의 생성정보에 부가된 프로세스의 결정 정보를 갖는다. 한편の, で, に, と 등에 대한 조사의 역어 결정은 조사가 갖는 역어의 다양성 때문에 대역적(global) 함수에 의한 역어 결정이 곤란하다. 그러므로 중요 조사에 대한 역어 생성은 시스템상에 개별적인 프로시쥬어를 기술하여

```
a:전치 명사의 の 관계tag 참조
  (명사 후속の의 해석여부 결정)
b:명사+の+활용어 연체형
  (전치 명사의 의미 소성→ov, of⇐주격 이, 가
                →以外 ⇨을, 를)
c:활용어 연체형+の+조동사→것
d:명사+の+특정 lexical こと→에 관한
e:활용어 연체형+の+특정 lexicalは, が→것
f:명사+の+は, が→것
  g:부사+の→일반부사⇨의, 접속부사⇨무해석

くせに;ps"の";00)인 주제에/ps DEF;03)주제에
くらい;qs"#なら";04)정도/qs DEF;00)만큼
けど;ks DEF;09)만/ks DEF;20)
くらい;qs"#なら";04)정도/qs DEF;00)만큼
かどうか;qs0;00)인지 아닌지/qs10;00)는지 없는지/qs20;00)
  없는지/qs20;00)지 어떤지/qs50;00)지나 않 않나
か;ks"ます";28)/ks"です";28)/ksDEF;04)까/qs0;09)/
  qs10;18)/qs20;18)
けども;ksDEF;09)만/ksDEF;20)
けれども;ksDEF;09)만/ksDEF;20)
けれど;ksDEF;09)만/ksDEF;20)
こととて;ps"0の";00)이므로/ps DEF;00)까닭에
```

그림 6. 조사の의 역어생성 프로시쥬어와 조사 사전의 일례

Fig. 6. Generation procedure for particle の and example of particle dictionary.

/*助動詞 生成 table work sheet */
(其一：終止形)

後接情報	活用 array 情報	後接情報	活用 array 情報
xx	たい	＋です,	⇨ です
	15	09	25 でしょ
			でし
xx	たがる	＋です,	⇨ です
	19	09	25 でしょ
			でし
xx	らしい	＋です,	⇨ です
	35	09	25 でしょ
			でし

그림 7. 활용어 복합 표현의 종결구 역어 생성테이블 [형태소 해석 정보로 부터 xx가 부여되면 역어 생성 테이블의 지표가 결정된다]

Fig. 7. Example of generation table for compound inflected word.

/* verb action→generation */

/* 0 */ "ㅂ니까", "", "ㄴ다", "는", "ㄹ", "기", "ㅁ", "면", "라", "지" "서", 면서", "게", "", "고", "든지", "는데", "ㅂ시다", "ㅂ니다", "시요", "자", ㅆ다", "ㄴ", "요", ㅆ다", "ㄴ" "ㅆ을", "ㅆ으면", "ㅆ고", ㅆ읍니다", "ㅆ었읍니까", "ㅆ어요"
/* 1 */ "ㅂ니까", "", "ㄴ다", "는", "ㄹ", "기", "ㅁ", "면", "아라", "지", 아서", "면서", "게", "아", "고", "든지", "는데", "ㅂ시다", ㅂ니다", "시요", "자", "았다", "ㄴ", "아요", "았다", "던", "았을", "았으면", "았고", "았읍니다"

/* adjective action → generation */

/* 0 */ "", "다", "ㄴ", "ㄹ", "기", "ㅁ", "면", "지", "면서", "게", "", "고", "든지", "ㅂ니까", "ㄴ", "던", ㅆ었읍니까"
/* 1 */ "", "다, "ㄴ", "ㄹ", "기", "ㅁ", "면", "지", "면서", "게", "", "고", "든지", "ㅂ니까", "ㄴ", "던", "었읍니까"

그림 8. 동사 & 형용사 어미테이블의 일례

Fig. 8. Example of korean verb and adjective inflection table.

행한다. 그림 6는 조사의에 대한 역어 생성 프로시쥬어를 보인 것이다.

/* 1 */
"가", "는", "와", "를", "로", "라고", "라는 & 것은", "든지", "인가", "에", "라는", "대로", "야말로", "며", "나미", "라도", "조차", "에서", "에게", "이고", "도", "나", "부터", "보다", "의", "밖에", "만", " 까지", "만큼", 정도", "라", "로부터", "예요",

/* 2 */
"이", "은", "과", "을", "으로", "이라고", "이라는 & 것은", 이든지", "인가", "에", "이라는", "이대로", "이야말로", "이며", "이며", "이나마", "이라도", "조차", "에서", "에게", "이고", "도", "이나", "부터", "보다", "의", "밖에", "만", "까지", "만큼", "정도", "이라", "으로부터", "이예요"

그림 9. 조사 생성 테이블의 부분례

Fig. 9. Example of korean particle generation table.

4. 시스템 구현 및 고찰

본 연구에서 구축한 실용화를 목표로 하는 일·한 번역 시스템 ATOM의 제과정은 IBM PC/AT 상에서 Microsoft C(V. 5. 0)을 이용하여 구현하였다. 시스템의 개략적인 규모는 번역 환경 지원 S/W로 개발된 interface가 약 400Kbyte, 그리고 사전을 제외한 주 프로그램이 약 100Kbyte 정도이며, 역어 생성 테이블이 75Kbyte이고, 사전은 대략 20,000단어 규모로 1.2Mbyte 정도의 크기로 구성되어 있다.

실제 시스템의 검증 과정에서 사용한 데이타는 문체의 다양성을 고려하여 임의로 선정한 일본어 컴퓨터 매뉴얼 10권에서 임의로(=1206 문장) 추출하여 사용하였다.

현재 시스템의 번역 속도는 20단어 기준의 문에서 대략 12초 평균으로 (시간당 5,000 단어 정도의 번역을 수행한다. 그러나 아직 역문의 질은 높은 수준에 있다고 볼 수 없으며, 좀 더 자연스러운 역문의 출력을 위해서 다의성을 갖는 명사의 의미 분류와, 동사등의 용언의 역어 선택에 대한 프로세스의 보완이 이루어져야 한다. 현재는 이를 위해서 동사와 명사의 의미 관계에 기초한 격 프레임 해소용 지식 베이스의 구축과 사전의 보완을 진행하고 있다.

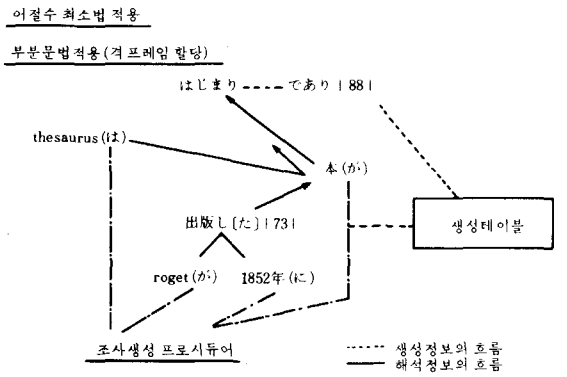
다음은 번역 과정의 일례이다.

example : 1
/* 해석 결과의 정보 : a ; 어휘항목(lexicon), b ; 어휘범주(lexical category),
c ; 접속형분류→ 1 ; 미정의 어, 2 ; 분류 기호,
d ; 접속형→ 0 ; 단어 분리, 1 ; 접속가능이나 분리하지 않음,
e ; 전 접속 정보
f ; 활용 테이블에서 리턴된 접속 정보의 열(최대 5 개) ,
g ; 활용 테이블에서 리턴된 활용형 정보,
h ; homonym weight, i ; generation entry No. */

a	b	c	d	e	f					g					h	i
⟨thesaurus	n	(0	0)	(1)	-	(1	0	0	0	0)	-	(0	0	0	0	0)	(1	1)
⟨は	q	(2	0)	(35)	-	(0	0	0	0	0)	-	(0	0	0	0	0)	(1	1)
⟨roget	n	(1	0)	(1)	-	(1	0	0	0	0)	-	(0	0	0	0	0)	(0	0)
⟨が	h	(2	0)	(27)	-	(0	0	0	0	0)	-	(0	0	0	0	0)	(1	4)
⟨1852	m	(1	0)	(2)	-	(6	0	0	0	0)	-	(0	0	0	0	0)	(0	0)
⟨年	u	(0	0)	(5)	-	(6	0	0	0	0)	-	(0	0	0	0	0)	(1	1)
⟨に	h	(0	0)	(28)	-	(24	0	0	0	0)	-	(0	0	0	0	0)	(1	5)
⟨出版し	y	(0	0)	(2)	-	(99	0	7	0	0)	-	(0	0	7	0	0)	(1	1)
⟨な	x	(0	0)	(15)	-	(66	0	10	0	0)	-	(0	0	3	0	0)	(1	2)
⟨本	n	(0	0)	(1)	-	(1	0	0	0	0)	-	(0	0	0	0	0)	(1	1)
⟨が	h	(0	0)	(27)	-	(0	0	0	0	0)	-	(0	0	0	0	0)	(1	4)
⟨はじまり	n	(0	0)	(2)	-	(99	0	8	0	0)	-	(0	0	8	0	0)	(1	2)
⟨であり	x	(2	0)	(17)	-	(66	8	0	0	0)	-	(0	8	0	0	0)	(1	2)
⟨,	n	(1	0)	(1)	-	(1	0	0	0	0)	-	(0	0	0	0	0)	(0	0)

The total time is 00 : 04 : 95

thesaurusは rogetが 1852年に 出版し た 本 が はじまり であり,
⟨00 : 04 : 95⟩



シーソラスは rogetが 1852年に 出版し な本がはじまりであり, すべての單語 を体係的に 分類しようとするものである. MS-DOSを起動したシステムデイスクに, config.sys ファイルが存在しいない場合, もしくは存在していてもその 内容に BUFFERS=xxがない 場合は, DIRのキーインのたびに毎回デイスクがアクセスされるはすです.

Teasaurusは roget이 1852년에 출판한 책이 시초이고,
모든 단어를 체계적으로 분류하려고 하는 것이다.
MS-DOS를 기동한 시스템 디스크에,
config.sys 화일이 존재하고 있지 않는 경우,
혹은 존재하고 있어도 그 내용에 BUFFERS=xx이 없는 경우는,
DIR 키-입력 때에 매번 디스크가 참조되는 것입니다.

The total time is 00 : 47 : 63

그림10. 번역 과정의 일례
Fig. 10. An example of translation process.

V. 결 론

본 논문에서는 실용화를 목표로 하는 일·한 번역 시스팀 ATOM에서 작성한 일본어 입력문의 구문 의미 해석과정과 생성 과정에 대하여 기술하였다.

구문 의미 해석 과정은 격 프레임을 도입하여 동사의 표층격을 중심으로 해석을 수행하고, 역어 생성 과정은 활용어미 테이블과 조사 생성 사전과 부가적인 조사 생성 프로세스에 의하여 수행하였다. 또한 구문 해석의 주요 문제인 포유문, 그리고 문접속의 처리를 구문 정보의 정밀화에 따라 해석하였으며, 종결구에 대한 보다 자연스러운 역어 생성을 얻기 위해서 자립 활용어와 조동사의 접속시의 역어 생성 테이블을 부가적으로 구성하였다.

본 논문에서 구현한 ATOM 시스팀은 번역대상 영역의 제한이라는 가정하에서는 상당히 만족할 만한 출력을 얻고 있다. 출력시간의 효율에 중점을 두어 구축하였으므로 사전의 정비와 의미 분류의 정밀화에 대한 연구를 진행하여, 역문의 질을 개선한다면 실용 영역에도의 사용도 가능하다고 생각된다.

參 考 文 獻

[1] Ullman, A.: "The Theory of Parsing, Translation and Compiling," Pretice Hall, 1975.
[2] Fillmore, C.: "The Case for Case," in Bach & Harms (eds.), Universals in Linguistic Theory, Holt Renehart&Winston, New York, 1968.

[3] 村木一至："知識Baseと，言語に獨立の中間表現とを用いた日英機械翻譯システム"，Nikkei Electronics, Dec. 17, 1984, pp. 195-220.
[4] 内田 裕士："言語に依存しない概念構造を中間表現の基本とし，常識を使う多言語向き機械翻譯システム," Nikkei Electronics, Dec. 17, 1984, pp. 221-240.
[5] 内田，增山："機械翻譯における概念變換について," 情報處理學會者然言語處理技術シンポジウム資料, Jun. 1983.
[6] 田中："自然言語處理のためのプログラミングシステム－擴張 LINGOLについて－"，電子通信學會論文誌D, vol. 60-D, no. 12, 1977.
[7] 長尾眞："計算機による日本語文章の解析に關する研究," 文部省科學研究費特定研究研究報告書, 1978.
[8] 長尾眞："國語辭書の記憶と日本語の自動分割," 情報處理, vol. 19, no. 6, 1978.
[9] Akira KOSAKA: "A Trial of the Japanese-English Machine Translation System via Logical Expression," Koyto Univ. 1980.
[10] Akira KOSAKA："A Trial of the Japanese-English Machine Translation System via Logical Expression," Koyto Univ., 1980.
[11] 長尾眞："言語の機械處理," 三省堂, 1984.
田村，田中："意味解碩に基づく並びく並列 名詞句の構造解析" 自然言語處理研究會報告, no. 59, 1987. 1.
[12] 長尾眞(eds.)："機械翻譯システムの調査研究," 日本電子工業振興協會, 1984.
[13] 平井，北橋："格の強度と述語の構文および意味屬性を用いた格構造の變換生成について，情報處理學會論文誌, no. 3, vol. 28, 1097.
[14] 金榮振："日本語 助詞 助動詞 活用 辭典,"정진출판사, 1980.
[15] 김승곤："한국어 통어론," 아세아 문화사, 1988.
[16] 고영근, 남기심："표준 국어 문법론," 탑출판사, 1988.
[17] 洪思滿："國語特殊助詞論," 學文社, 1983.
[18] "韓日，日韓 自動 翻譯 시스템의 開發에 관한 研究," 科學 技術處, 1986.
[19] 김영섬, 김한우, 최병욱："PC를 이용한 일·한 번역 시스템 ATOM의 개발에 관한 연구," 대한 전자공학회 논문지, vol. 25, no. 10, 1988. *

著 者 紹 介

金 榮 暹 (正會員)　第25卷 第10號 參照
현재 한양대학교 전자통신공학과 박사과정

●

金 漢 宇 (正會員)　第25卷 第10號 參照
현재 한양대학교 전산과 부교수

崔 炳 旭 (正會員)　第25卷 第10號 參照
현재 한양대학교 전자통신과 교수