한국어 문장으로부터 개념단위의 추출과 지식베이스의 구축

The Conceptual Unit Extraction and Knowledge Base Construction from Korean Sentence

한 광 록 이 주 근
인하대학교 전자공학과

K. R. Han J. K. Lee
Dept. of Electronic Engineering, Inha University

요 약

본 논문은 한국어를 대상으로 하는 자연언어 처리 시스템을 개발하는데 있어서 기초가 되는 지식베이스의 구축에 대하여 논한다.

한국어의 일반문에서 단문을 분리해 내기 위하여 형태소 해석의 결과로 부터 도출한 구단위를 한-일 기계번역 시스템의 구문, 의미 해석기 (VCPN)을 적용하여 절단위로 결합한다. 그리고 이들 단위절에 대하여 대명사의 조응관계, 생략어의 재생을 위한 추론, 부정어, 시제일치 등을 처리하여 논리적 지식베이스를 구성하는 방법을 제안한다.

본 논문은 입력문장에 제한을 두지 않고 단문으로 부터 장문에 이르기까지 광범위한 일반문을 대상으로 하여 Horn Clause 이론을 확장한다.

I. 서론

컴퓨터에 자연언어를 이해시키는 문제는 인간의 지적능력을 기계에 이식시키는 것이다. 그러나 자연어는 지식표현의 수단으로서 지식의 미세한 의미까지도 표현할 수 있는 특징이 있지만, 그 양이 지극히 방대하고 구조가 다양하게 변하므로 컴퓨터로 처리하기는 매우 어려운 문제이다.

따라서 자연어를 처리하는 인공지능 시스템을 개발하기 위해서는 다양하고 복잡하게 표현되는 일반 문장을 컴퓨터가 처리할 수 있는 형태로 지식베이스를 구성하는 작업이 선행되어야 한다. 이를 위하여 제한된 영역의 단문을 대상으로 Horn Clause 이론을 적용하여 지식베이스를 구성하는 방법이 연구되어 왔지만, 여러개의 문장이 내포된 복합문에 대해서는 분석과정의 복잡성 때문에 실현에 많은 어려움이 있다. 특히 한국어는 용언과 체언이 결합된 단위문들이 관형적 수식을 하거나 보문구조 등으로 결합하여 복문을 구성하고 있다[5]. 따라서 전문가시스템, 질문응답시스템, 추론머신 등의 자연언어 처리 시스템을 설계하기 위해서는 복문으로 부터 단위개념의 단문을 추출하고, 이들을 이용가능한 지식베이스 형태로 표현해야 한다[1,2]. 이를 위하여 본 논문에서는 한국어의 문장에서 체언과 용언의 결합관계를 분석하고, 형태소 해석의 결과로 부터 도출한 구단위의 문들을 저자들이 한-일 기계번역 시스템에서 사용한 구문, 의미 해석기인 VCPN (Variable pattern Net)에 의하여[3,4] 절단위로 결합하여 단문을 추출한다. 이때에 사용하는 VCPN은 문장을 용언 중심으로 생각하고, 이 용언과 결합관계가 있는 조사를 수반한 명사와 이 명사의 의미소성의 분포로서 기술한다.

II. 단위문의 분리

종래의 인공지능 시스템은 극히 제한된 단문을 대상으로 Horn Clause 이론을 도입하였기 때문에 적용범위가 한정되었다. 그러나 본 논문은 Horn Clause 이론을 확장하여 제한이 없는 한국어의 일반문을 대상으로 하고 있기 때문에 먼저 관형적 수식이나 보문구조 등으로 여러개의 단문이 결합

된 복문으로 부터 개념단위의 단문을 분리해 내는 과정이 필요하다.

(S1) (1) 깨끗한 집 앞에 행복이 깃든다.

(2) 씩씩한 군인들이 넓은 연병장에서 훈련을 하고 있다.

문장 (S1)은 모두 두개 이상의 단문이 내포된 복문으로서 (S1-1)은 「깨끗하다」가 관형형으로 전성되어 핵심명사 「집」을 수식하는 한정적 구조의 복문이고, (S1-2)는 「씩씩하다」, 「넓다」가 각각의 핵심명사를 수식하는 구조이다. 따라서 (S1)에 Horn Clause 이론을 확장하기 위해서는 문장의 용언을 중심으로 헤드명사와 선행하는 명사들 사이의 의미적 결합관계를 파악하여 수식범위를 해석함으로서 단문을 분리해 낸다. 즉 문장 (S1-1)은 두개의 용언 「깨끗하다」와 「깃들다」에 의하여 (S2)와 같이 (C1), (C2)의 두개의 단문으로 분리되며, 문장 (S1-2)는 세개의 용언 「씩씩하다」, 「넓다」, 「하다」에 의한 (C3), (C4), (C5)의 단문으로 분리할 수 있으며 이들은 VCPN에 의하여 실현된다.

(S2) (C1) 집 앞이 깨끗하다.

(C2) 집 앞에 행복이 깃든다.

(C3) 군인들이 씩씩하다.

(C4) 연병장이 넓다.

(C5) 군인들이 훈련을 하고 있다.

(S1)과 같은 복문을 (S2)의 단문으로 분리한 후에 효과적인 기계처리를 위하여 이들 단문들 사이의 논리적 관계를 해석하여 지식베이스를 구성한다. 이때 분리된 단문들 사이의 논리관계는 한국어에서 잘 발달된 용언의 어미정보를 이용하며[6], 다음과 같은 규칙에 의하여 실현한다.

(1) 하나의 용언에 대한 논리적 관계는 긍정과 부정으로 한다.

(2) 두 단문 사이의 논리적 결합관계는 AND관계, OR관계, IMPLIES관계로 한다.

(3) 전성어미에 의해서 삽입되는 하위문은 AND관계로 한다.

(S3) (1) 봄이 가고 여름이 왔다.

(2) 너는 그림을 그리거나 책을 읽어라.

문장 (S3-1)의 「봄이 가고」와 「여름이 왔다」의 두 단문 사이에는 AND 관계가 존재하고, (S3-2)의 「너는 그림을 그리거나」와 「책을 읽어라」의 사이에는 OR 관계가 존재하며, (S3-3)의 「내일 비가 오면」은 IMPLIES 관계로 연결되어 있으며, 「가지 않겠다」는 부정의 표현이다. 그림1은 (S3) 문장에 대한 논리관계를 나타낸 것으로 ∧는 AND, ∨는 OR, ->는 IMPLIES, NEG는 부정을 의미한다.

C1 (봄이 가고) ∧ C2 (여름이 오다)

C1 (너는 그림을 그리거나) ∨ C2 (책을 읽어라)

C1 (내일 비가 오면) -> C2 (나는 학교에 가다 (NEG))

그림1 단위문의 논리관계
Fig.1 The logical relation of unit sentence

III 개념단위의 지식표현

한국어 문장으로 부터 추출된 단문들을 구조적인 데이타베이스 형태로 표현하기 위해서는 대명사의 조응관계 (anaphoric relation)의 처리, 생략어의 재생처리, 부정의 처리, 시제의 일치 등을 해석한다[7,8,9].

1. 대명사의 조응관계

일반 언어에서 의미전달에 차질이 오지 않는 범위내에서 문장을 간단한 표현이나 대명사로 대치하거나, 주어진 정보를 되풀이하여 반복하지 않으려고 문장 성분의 일부 또는 전체를 생략하여 표층 상에 나타나지 않는 형태로 표현하는 경우가 많다.

대명사에 의한 대용이나 생략어에 의한 축소는 언제든지 복원될 수 있는 가능성 (recoverablity)을 가지고 있는 조건 하에서 이루어진다.

한국어에서 어떤 단순한 명사가 대명사화 하는 것이 아니고, 문장 전체 혹은 두 개의 연속적인 문장에서 앞 문장의 명사와 동일한 명사가 다음 문장에 올 때 두번째 명사를 대명사화 하는 경우가 많다.

(S4) (1) 철수는 정직하고 성실하고 명랑하다.

(2) 그는 모든 사람에게 칭찬을 받는다.

(3) 교통사고로 죽는 사람이 병으로 죽는 사람보다 많다.

(4) 그것은 교통이 무질서하기 때문이다.

문장 (S4-1)에서 「철수」가 「그」라는 대명사로 바뀌어 (S4-2)로 되었으며, (S4-3)에서는 문장 전체가

「그것」이라는 대명사로 바뀌어 (S4-4)가 되었다. 이와 같이 한국어의 대명사 처리를 위하여 문장을 최종 단위로 하는 것보다 담화(discourse)와 같은 큰 문장의 단위에서 처리하여야 할 것이다. 이것은 앞으로의 연구과제로 남기고, 본 논문에서는 일반 대명사만을 다음과 같은 규칙에 의하여 처리한다.

[일반 대명사의 처리 규칙]

(1) 전방향으로 가장 가까운 명사를 선행사로 한다.

(2) 선행사를 대입하여 의미소성이 일치해야 한다.

(3) 대용되는 명사구는 문장의 핵심어이어야 한다.

2. 생략어의 재생

한국어에서는 생략이 아주 심하며, 특히 주어의 탈락이 빈번하다. 그러나 언어활동이 일어나는 상황을 서로가 공동으로 인식하고 있기 때문에 의사전달이 잘 이루어진다고 볼 수 있다.

문장에서 생략어를 찾기 위해서는 생략이 일어나는 방향을 알 필요가 있다. 한국어에서는 그 방향이 생략되는 단어의 품사에 따라 두가지 방향으로 나타난다. 연속되는 문장에서 같은 명사구들이 나열될 때는 순방향으로 뒤에 오는 명사구가 생략되며, 같은 동사들을 갖는 문장들이 접속될 때는 역방향으로 앞에 오는 동일 동사구가 생략된다.

따라서 문장으로 부터 개념을 쉽게 추출하기 위하여 생략어를 재생시켜 놓고 처리할 필요가 있기 때문에 다음과 같은 생략어 재생규칙을 정한다.

[생략어의 재생규칙]

(1) 생략어가 명사이면 한 문장 내에서 전방향으로 하여 VCPN에 의하여 의미소성이 같은 명사를 우선하여 재생한다.

(2) 생략어가 용언이면 한 문장 내에서 후방향으로 나타나는 용언의 의미패턴과 절로 결합되지 못한 구의 의미소성들이 일치하는가에 따라 생략된 용언을 결정한다.

이 규칙을 적용하여 각 문장에서 생략된 명사와 동사를 재생해 내는 과정을 그림2에 나타낸다.

그림2에서 V5Q6, V2Q4는 한 문장에서 용언이 수반할 수 있는 명사의 의미소성과 조사가 짝을 이루는 패턴네트 상의 의미기호이고, HUM, EAT, PRO 등은 각각 명사의 의미소성이다. 따라서 그림2의 a)에서는 VCPN 상에서 의미기호인 V5Q6의 구성요소 중에서 의미소성이 같은 「자전거(PRO)」가 생략되었음을 알 수 있고, b)에서는 각 단문 C1, C2, C3의 명사구의 의미소성과 조사가 같은 짝을 이루는 패턴이기 때문에 용언도 동일한 의미기호 V2Q4를 갖는 「먹다」의 생략을 간단히 추론하여 재생할 수 있다.

3. 부정의 처리

부정문은 긍정문에 대립되는 문장으로서 어떤 사물이나 명제에 대하여 부정하는 의미를 가지는 문장이다.

한국어의 부정문은 제 1 부정문 (pre-stem negation)과 제 2 부정형 (post-stem negation)으로 나타내어, 부정문의 표면구조는 다르지만 본 논문에서는 기계처리를 쉽게 하기 위하여 문장에서 부정의 형태가 검출되면 긍정문의 형태로 변환한 후 부정을 표시하는 개념정보 「NEG」를 삽입한다.

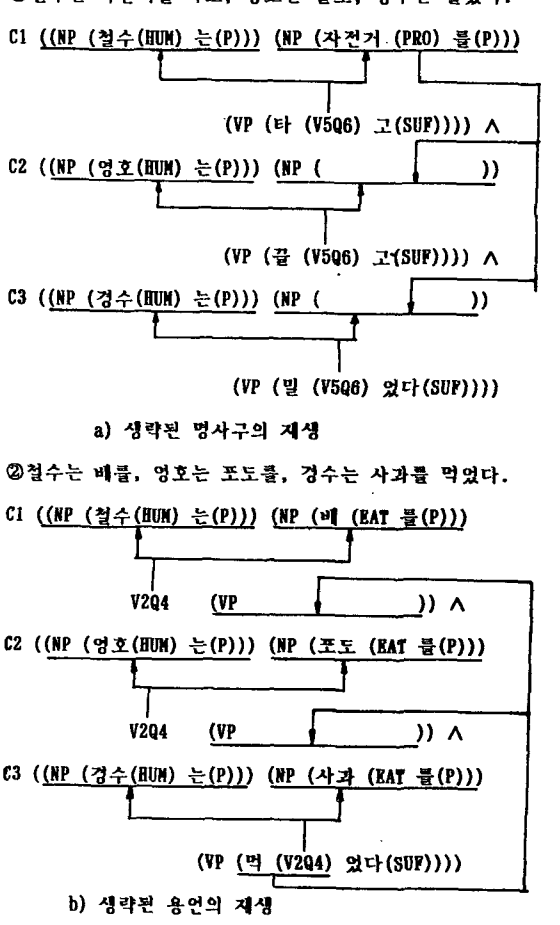

그림2 생략어의 재생 처리 과정

Fig.2 The regenerating process of abbreviated word

4. 시제의 일치

일반 문장이 관형절 또는 보문으로 표현된 복합문이면 이들로 부터 추출된 절 사이에는 시제가 일치되도록 처리해야 한다. 관형화된 표현에서 추출된 절의 시제는 다음과 같은 규칙에 의하여 관형어미를 시제 보조어간으로 치환함으로서 처리한다.

[시제 관형화의 처리 규칙]

V — [는 / 은/ㄴ / 을/ㄹ] — N ⟶ N(Cm) — V — [는 / 었 / 겠] — 다

핵심명사 N을 동사 앞으로 이동하고 VCPN에 의하여 문장의 성분을 결정하여 격표시 (case marker)인 Cm을 삽입하고, 종지형 어미 「다」를 붙인다.

(S5) (1) 그녀가 즐겨듣는 음악을 나도 좋아한다.

(2) 어제 내가 들은 음악은 고전 음악이다.

(3) 나는 내일 현대 음악을 들을 계획이다.

문장 (S5)의 관형절로 부터 추출된 절의 시제는 문장 (S6)과 같다.

(S6) (1) 그녀가 음악을 즐겨듣는다.

(2) 어제 내가 음악을 들었다.

(3) 나는 내일 현대 음악을 듣겠다.

VI. 논리적 지식베이스의 표현

입력문이 형태소 해석되고, 구구조의 규칙에 의하여 구단위가 분리되며, VCPN 해석기에 의하여 구문 및 의미해석 과정을 거치면 개념단위의 단위절이 도출된다. 또한 3장과 같은 후처리를 한 뒤에는 논리적 형태의 지식 데이타 베이스를 구축하는데 이때 표현되는 데이타베이스의 기본형식은 그림3과 같다.

그림3에서 N은 명사, Det는 관형사, Adv는 부사, A는 주어, C는 보어, O는 목적어, Xi는 변수, Ci, Cj는 단문의 번호이고, Sm은 의미표시, Agt, Cmp, Obj는 각각 주어표시, 보어표시, 목적어표시이다. 또한 Sifm은 구문정보로서 T는 시제표시, L은 논리표시, N은 부정표시이다. 보통명사는 총칭의 개념을 갖고 있는 것으로 하여 변수 Xi를 도입하여 ISA 형의 절로 표현한다. 의미네트워크에서 ISA Link는 속성을 계승하는 특성과 포함관계가 성립되는 성질을 가지고 있다.

(1) (ISA (Xi N (Det)))

(2) ((Ci) (V (Sinf) (A {Cj} (Sm Agt)) (C {Cj} (Sm Cmp)) (O {Cj} (Sm Obj)) (Adv))))

Sifm = (T L N)

T(Tense) = T1 :현재
T2 :과거
T3 :미래

L(Logic) = (AND | OR | IMPLIES)

N(Negation) = (NEG | nil)

그림3 지식베이스의 기본형식
Fig.3 The basic form of knowledge base

하나의 문장이 단위 개념의 절 형태로 표현되어 지식베이스를 구축하는 예를 그림4에 나타낸다.

(S7) 그 여자는 맴을 아주 싫어한다.

(ISA (X1 여자 (HUM (그))))

(ISA (X2 맴 (MAC)))

((C1) (싫어하다 (ㄴ다 (T1))
(X1 (AGT 는) X2 (OBJ 을) (아주))))

(S8) 영수는 어려운 프로그램을 빨리 작성한다.

(C2) 프로그램이 어렵다 (ㄴ (T1 AND))

(C3) 영수는 프로그램(C2)을 빨리 작성하다 (ㄴ다 (T1))

(ISA (X3 프로그램 (MCC)))

((C2) (어렵다 (ㄴ (T1 AND))
(X3 (AGT 이))))

((C3) (작성하다 (ㄴ다 (T1))
(영수 (HUM Agt 는) X3 (C2 Obj 을) (빨리))))

그림4 절형식의 논리표현
Fig.4 The logical reopresentation of clausal form

본 연구에서는 이상의 과정들을 실현시키기 위하여 LISP 언어를 사용하여 프로그램을 작성하였으며, 프로그램의 일부분을 그림5에 나타낸다. 그림5의 MAKE_ISA 함수는 ISA Link를 구성하는 것이고, DETER_TEST는 수식관계를 해석하여 단문을 추출하는 부분적인 함수이다.

```
;
(DEFUN MAKE_ISA (LAMBDA (ARG)
   (SETQ NUM* (PLUS NUM* 1))
   (SETQ NUBM_CONT (PACK (LIST 'X NUM*)))
   (SETQ ISABUF* (APPEND ISABUF* (LIST
     (LIST 'ISA (LIST NUBM_CONT
    (LIST ARG) ]
;
;
(DEFUN DETER_TEST (LAMBDA  (ARG)
    (SETQ SSS* ARG)
    ((MEMBER (CAR SSS*) ENDLIS) T)
    ((SEARCH (CAR ARG) KNTBL)
        (SETQ NOUNMD* (APPEND NOUNMD*
                  (LIST (CAR SSS*))))
        (DETER_TEST (CDR SSS*)))
    ((EQUAL 'EUI PEUI)
        (SETQ ASS1 (SUBSTRING (CAR SSS*) 1
                   (SUB1 LEUI)))
        (SEARCH ASS1 KNTBL)
        (SETQ NOUNMD* (APPEND NOUNMD*
    (LIST ASS1 (CDR (REVERSE KT)) PEUI)))
        (DETER_TEST (CDR SSS*)))
    (NOUNMODI (CAR SSS*) AAA BBB) ]
```

그림5. 처리 프로그램의 예
Fig.5 The example of processing program

V. 결론

본 논문는 체언과 용언으로 이루어진 한국어의 문에서 지식베이스를 구성하기 위하여 개념단위의 절을 추출하는 방법에 대하여 논했다.

여러개의 문장으로 구성된 복합문에서 용언의 지배를 받는 개념단위들의 단문들을 도출하여 그들의 논리적 결합 관계를 해석하였으며, 추출한 절 형식의 단문에 대하여 대명사의 조응관계, 생략어의 처리, 부정문의 처리, 시제일치 등을 처리하여 논리적 지식베이스를 구성하였다.

참 고 문 헌

[1] F. Hayes-Rothe, The Knowledge Based Expert System: A Tutorial, Computer, Vol.17, pp.11-28, Sep., 1984.

[2] J. A. Robinson, A Machine-Oriented Logic Based on the Resolution Principle, JACM, Vol.12, pp.23-41, Jan., 1965.

[3] S.H. Lee, K.R. Han, J.K. Lee, A Bidirectional MT System between Korean and Japanese Based on Pattern Net, Proc. of SITA'87, Japan, PP.405-410, Nov., 1987.

[4] 한광록, 이주근, 한-일 양방향 번역시스템의 기본설계, 전기전자공학회 학술대회 논문집(II), pp.1033-1037, 1987.

[5] 최창열, 국어 의미구조 연구, 한신문화사, 1980.

[6] 남기심, 국어 연결어미의 화용론적 기능, 연세논총15, 1978.

[7] 성광수, 국어 부정법 연구에 관하여, 문법연구1, 1974.

[8] 한재현, 생략규칙, 어학연구 16-2, 1980.

[9] 양동희, 국어 관형절의 시제, 한글 162, 1978.