정보거래 자동 중개 시스템을 위한 한국어 문형 표준안◊

정의석 *김기태 임수종 차건희 박제득 윤보현 강현규
언어이해연구팀, 언어공학연구부 한국전자통신연구원, *국어국문학과 충남대학교

{eschung, isj, chakh, jdpark, ybh, hkkang }@etri.re.kr, *kimkitae@yahoo.com

Controlled Korean Phrase-Structure Standard Spec. for the Automatic Information Trading Mediator System

Euisok Chung, *Kitae Kim, Soojong Lim, Gunhae Cha, Jae-Deuk Park, Bo-Hyun Yoon, Hyun-Kyu Kang
Linguistics Engineering Dept. ETRI. *Korean Dept. Chungnam Univ

요 약

본 논문은 정보거래 자동 중개 시스템을 위한 한국어 문형 표준안에 대하여 기술한다. 정보거래 자동 중개 시스템은 인터넷상에서 지식정보자산의 공급자와 수요자를 자동으로 연결해주는 시스템으로서 텍스트로 기술되는 수요자의 의도와 공급자의 지식정보 내용을 정확히 연결할 수 있는 신뢰성을 보장한 고품질의 정보검색 기술이 필수적이다. 그러나 자연어의 복잡성과 불규칙성은 정확한 언어처리 기술이 필수적인 고품질의 정보검색을 보장할 수 없다. 따라서 본 논문은 한국어 문장 표현 방식을 표준화하여 언어처리 기술 적용의 한계를 극복해보자는 데 그 목적이 있다. 또한 일반 사용자의 언어 표현을 문형 표준안으로 유도하는 방법에 대하여 기술한다. 문형 표준안의 구성은 표준 문형, 표준 문형 유도 방법, 어휘부로 구성되어 있다.

1. 서론

정보거래 자동 중개 시스템에서 정보거래란 노하우, 노웨어, 전문, 전문지식, 아이디어등의 지식정보자산을 일반 유저들이 누구나 쉽게 상품처럼 사고 팔거나 무료로 제공하는 거래를 말하며, 자동 중개란 인터넷을 통하여 수백만에서 수천만 이상의 지식정보 자산을 동시 다발적으로 적합한 공급자와 수요자를 정확하게 자동 연결하여 거래가 이루어지게 지원하는 체제를 말한다.

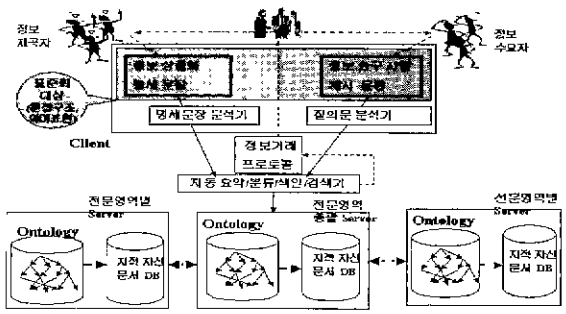

그림 1 정보거래 자동 중개 시스템 개념도

그림 1 은 정보거래 자동 중개 시스템의 개념도이다. 정보제공자와 정보수요자는 각각 제공할 정보와 원하는 정보에 대한 메타 데이터와 해당 정보를 문형-의미 표준화를 적용한 명세문장 분석기와 질의문 분석기를 통해 자동 요약/분류/색인/검색기에 등록한다. 그러면 정보거래 프로토콜에 의해 상호 적합한 공급-수요자를 연결해 주는 것이 본 시스템의 주된 기능이다. 정보거래 프로토콜은 정보의 공급가격, 수요가격과 정보의 신뢰성 및 부가가치성에 따라 공급-수요자를 결정하는 역할을 한다. 그리고 정보표현을 위한 어휘 개념을 구성하는 온톨로지 정보는 개념 기반 온톨로지와 특정 전문 영역을 위한 지역적 온톨로지들로 분리되어 구성된다

신뢰성 있는 자동 중개의 핵심은 요소기술중의 하나인 정보검색의 신뢰성이다. 현재의 인터넷 정보검색의 가장 큰 문제점은 정보검색 품질 및 신뢰도 저조이다. 검색 결과 중 원하는 정보의 빈약 및 누락, 원하지 않는 정보의 과다 제시 문제가 그 예가 될 수 있다. 이의 원인으로 유용한 정보의 웹 페이지 등재부족, 검색 기술 수준의 저조, 무절제한 문장 생성으

◊ 본 논문은 2000년도 정보통신부의 지원을 받아 수행된 "정보거래 자동중개용 문장구조 및 의미표현 표준화"의 일환으로 이루어졌다

로 인한 문장 분석의 어려움을 지적한다면, 신뢰성 있는 정보 검색을 위해서는 정돈된 문형 및 대응 의미표현의 표준화에 기반을 둔 의미기반 정보검색 기법이 필수적이라 하겠다.

2. 문제 정의

한국이 문장은 그 쓰임에 있어 생략, 도치, 장문의 복합 명사 사용이 빈번하다. 특히 뉴스그룹이나 경매 시스템, 기존의 지식 정보 거래 시스템의 경우 통신 특유의 문장 형식들로 인해 의미 기반 정보검색 기법을 적용하기엔 큰 문제가 있다. 다음 예문의 경우 국내 특정 지식 정보 사이트의 질의문 중 한 예이다. 이러한 문장의 경우 언어처리를 통한 자동적 정보 중개는 불가능하다고 본다. 따라서 사용자의 언어 기술을 제약해야만 할 방법이 제시되어야 한다.

(1) 현재 가장 좋아하게된 엘리자베트 ..우리나라에서 도 공연이 된지가 있나요???
(2) 메이비록스 민나델 앨범 언제 나오나여?

본 논문은 정보 거래 자동 중개 시스템을 위한 한국이 문형 표준안에 한하여 기술한다. 문형 표준안은 한국어 문장 표현을 제약하여 한국이의 복잡성과 불규칙성으로 인한 언어처리 기술 적용의 한계를 극복해 보고자 함을 목적으로 한다. 그리고 일반 사용자들의 언어 표현을 문형 표준안으로 유도하기 위한 다양한 방법-바꾸이 말하기(paraphrase), 분석 확인 질의등의 대화형 구문분석 기법-의 기반을 제공한다. 특히 표준안 유도 방법은 신뢰성 있는 정보 기술의 생산성에 있어 중요한 변수로 볼 수 있겠다

문형 표준안의 구성은 표준문형, 표준문형의 유도와 어휘부로 구성된다. 표준문형은 다시 기본문의 구성과 확대로 이루어졌으며 이에 대한 해석과 제한은 표준문형 유도에서 기술한다. 또한 어휘부에서는 표준화된 어휘 범주 분류에 대해 기술한다. 문형 표준안은 서술어에 따른 논항들의 위치를 고정하여 자유어순 문제에 접근하였고, 표준문형 유도에서 추정패턴을 이용하여 생략현상에 대한 접근방법을 제시한다 그리고 관형이의 피수식어를 최소 거리로 제한하여 수식어에 의해 발생하는 구문 모호성을 제거하였다.

3. 관련 연구

제한된 언어를 이용한 작업은 영어권에서 지속적으로 진행되어 온 작업이다 응용 영역별로 분류하자면 기술 문서 작성[3], 지식 기반 기계번역[4], 데이터 베이스 질의 언어[5] 그리고 가장 최근에 진행된 요구사항 명세서를 위한 언어 표준화 작업-ACE(Attempto Controlled English)-이 있다[1][2]. 특히 ACE 는 표준화된 문장을 first-order logic 으로 변환하여 지식 베이스를 구축 가능하게 접근하였다 표준안 유도 방법으로 바꾸어 쓰기(Paraphrase)를 이용하여 사용자 입력 텍스트에 대한 해석을 보여주어 반복적인 입력을 유도하여 표준안 문형으로 유도하였다.

4. 표준 문형

4.1 기본문의 개념

기본문은 기본적인 문장패턴에 대한 규정으로서 필수 논항과 서술어로 구성되어 있다. 기본문의 구성체인 논항과 서술어는 일정한 순서를 지니고 있으며, 예외적인 경우 단서조항을 둔다. 문장은 최소한 하나의 주어와 서술어를 지녀야 한다. 다만, 문장이 문장 안에 내포된 경우나 연결어미를 이용하여 문이 접속될 경우 논항의 일부가 생략될 수 있다. 모든 서술어는 일정수의 논항을 요구하며, 동사, 형용사, 혹은 '명사, 대명사, 수사+이다'로 구성된다. 서술어는 부사어, 부사구, 부사절에 의해 수식을 받는다.

4.2 기본문의 분류

기본문은 서술어가 요구하는 논항의 수에 따라 1 항술어, 2 항술어, 3 항술어로 나뉜다. 각 술어에 의해 요구된 논항은 그 위치가 각각 정해져 있으며, NP1 은 주어 자리, NP2 는 목적어 혹은 보어 자리, NP3 는 필수 부사어 자리이다.

논항 1	NP1 (주어)	- NP+주격조사
논항 2	. NP2 (목적어/보이)	- NP+목격격조사/보격조사
논항 3	. NP3 (필수부사어)	- NP+부사격조사

4.2.1 1 항술어

(문형 1)	S → NP1 VP	- 1 항술어
	컴퓨터가 확인한다	(1)[1]
	컴퓨터는 편리하다	(2)
	컴퓨터는 기계이다	(3)
	*확인한다	(4)
	*확인한다 컴퓨터가	(5)

[1] 본 논문의 예문들은 해당규칙 기술에 한정되어 있다

예문 (1)은 동사를 서술어로 하며, 예문 (2)는 형용사, 예문(3)은 명사에 지정사 '이'가 결합되어 서술어 역할을 하는 예이다. 그리고 예문 (4)는 필수격이 생략되어 비문이며, 예문 (5)는 고정순서 위반으로 비문이다.

4.2.2 2 항술어

(문형 2) S → NP1 NP2 VP - 2 항술어
(문형 3) S → NP1 NP3 VP - 2 항술어

xml 은 문서 구조를 나타낸다.(NP2)	(1)
xml 이 희망이 되었다 (NP2)	(2)
*유니코드를 xml 은 채택한다	(3)
xml 은 html 과 비슷하다.(NP3)	(4)
html 과 xml 은 비슷하다	(5)

예문 (3)은 고정순서 위반으로 비문이다. 그리고 예문 (4)가 2 항 술어인 반면, 예문 (5)는 'html 과 xml 은' NP1 이 되는 1 항술어로 본다.

4.2.3 3 항술어

(문형 4) S → NP1 NP2 NP3 VP - 3 항술어
(문형 5) S → NP1 NP3 NP2 VP - 3 항술어

자연어처리는 언어학을 모태로 여긴다	(1)
자연어처리는 검색에 편리성을 준다.(수여동사)	(2)
자연어처리는 언어학을 전산학적으로 말한다 (발화동사)	(3)
*자연어처리는 언어학으로 모태를 여긴다	(4)

전술한 바와 같이 VP 가 수여동사거나 발화동사[2]일 경우 문형 4,5 모두 적용될 수 있다 그러나 예문 (4)와 같이 수여동사, 발화동사 이외의 동사가 문형 5 로 쓰였을 경우 비문으로 간주한다.

4.3 논항 확대

NP1, NP2, NP3 는 앞에 관형어를 선행시키거나 다른 NP 와 결합하여 논항을 확대할 수 있다. 관형어는 관형사, NP+의(관형격 조사) 그리고 관형절을 포함한다 관형절에 의한 논항의 확대는 절에 의한 확대이므로 문의 내포에서 다룬다

또 다른 확대의 방법은 접속 조사를 이용한 병렬구성, 접속부사를 이용한 나열식 구성 그리고 병렬구성으로 확대된 논항이 접속 부사로 결합된 복합 구성이 있다. 결국 각각의 논항 1, 논항 2, 논항 3 은 아래와 같은 방법으로 확대될 수 있다.

새 버전	(1)
새 버전의 개발자	(2)
마이크로소프트와 IBM 이 공동으로 개발하였다.	(3)
basic, C, COBOL 그리고 Pascal 은 프로그램 언어다.	(4)
*나는 IBM 한국지사 서비스 사업본부에 문의하였다.	(5)

위의 예문에서 (1)은 관형사에 의한 논항확대이고, (2)은 관형격 조사, (3)와 (4)는 각각 접속 조사와 접속 부사를 이용한 논항 확대의 예제이다 (5)의 경우는 다음의 NP 구성 규칙에 의해 비문으로 처리된다.

(규칙 1) $NP \rightarrow (nc|nn)^{+\leq 3} \mid nb \mid np$
(규칙 2) $NP' \rightarrow NP$

규칙 1 은 NP 구성 규칙으로써 3 개 이하로 제약된 복합명사를 이루는 연속된 nc(일반명사)나 nn(수사), 또는 nb(의존명사), np(대명사)의 생성을 표현한다. 이후 NP 는 규칙 1 의 정의를 따른다. 아래의 예문에서 (1)-(4)는 정문이 되며 (5)는 비문으로 간주한다.

넷스케이프 버전	(1)
인터넷 익스플로러 설치	(2)
칠천 팔백 오십라인	(3)
에플릿 하나	(4)
*인터넷 익스플로러 설치 방법	(5)

규칙 2 는 NP 를 논항 확대된 NP'로 간주함을 나타내며, 이후 NP'는 논항확대를 표시한다.

4.3.1 관형사의 논항확대

각각의 논항은 관형사를 취할 수 있다. 이때 관형사 지시관형사, 수관형사, 성상관형사의 순서로 연속되어 나타날 수 있으며, 각각 관형사는 중복될 수 없다.

(규칙 3) $NP' \rightarrow mm^{+\leq 2}\ NP$
제약 . - $mm^{+\leq 2}$ 는 {지시관형사 < 수관형사 그리고 지시관형사 < 성상 관형사}의 패턴에 한정한다.
- $v \in nn$ 이고 $v \notin nb$ 이다 $(v \in NP)$

아래의 예문중 (1)-(5)는 정문으로 취급된다. (6)의 경우 관형사의 제약에 위배되어 비문으로 간주한다.

새 노트북	-성상관형사	(1)
다섯 카피	-수관형사	(2)
이 노트북	-지시관형사	(3)
이 세 버전	-지시 < 성상관형사	(4)
저 두 개발자	-지시 < 수관형사	(5)
*저 한 새 노트북	-지시 < 수 < 성상관형사	(6)

4.3.2 접속 조사, 접속부사의 논항확대

논항은 접속조사와 접속부사를 이용하여 확대시킬 수 있다.

[2] 발화동사는 말하다로 대표되는 상위용언을 말한다[6]

병렬 구성(규칙 4)의 경우 접속조사를 사용하고, 나열 구성(규칙 5)의 경우 접속부사의 사용을 원칙으로 한다. 복합구성(규칙 6)은 나열식 구성에 병렬식 구성이 혼합된 형식이며, 이에 대해서는 아래와 같은 규칙을 사용한다. 논항의 확대에 사용되는 접속조사와 접속부사는 어휘부에서 따로 정의한다.

(규칙 4) NP′ → NP+jj NP

(규칙 5) NP′ → (NP,)$^{+}$| NP maj NP

(규칙 6) NP′ → NP_1′ maj NP_1′
제약 - NP_1′은 규칙 4 에 의한 논항 확대에 한한다.

아래의 예문 중 (1)은 규칙 4 에 의한 논항확대, (2)는 규칙 5 에 의한 논항확대, 그리고 (3)과(4)는 규칙 6 에 의한 논항확대를 나타낸다. 예문 (5)-(6)은 위의 규칙들의 형식에 위배된 문장들이다. 그리고 예문 (7), (8)로부터 접속 조사나 접속부사를 사용하지 않는 병렬 표현을 비문으로 간주함을 알 수 있다.

야후와 엠파스는 경쟁한다 (1)
다음, 엠파스, 구글 그리고 야후는 검색 사이트다. (2)
유닉스와 리눅스 그리고 오라클과 DB2를 설명한다. (3)
유닉스와 리눅스 또는 오라클과 DB2 를 설명한다 (4)
*유닉스, 리눅스와 오라클, DB2 를 설명한다 (5)
*유닉스, 리눅스 그리고 오라클 DB2 를 설명한다 (6)
*여기서는 유닉스이고 리눅스를 설명한다. (7)
*여기서는 유닉스이건 리눅스를 설명한다 (8)

4.3.3 속격 조사의 논항 확대

논항은 속격 조사를 사용하여 논항을 확대할 수 있다.

(규칙 7) NP′ → NP′+jm NP_1′
제약 - NP_1′은 단일 NP 만을 갖는다

아래의 예문은 규칙 7 에 의해 기술 가능한 표현들의 예이다. 예문(2)의 경우 규칙 7 의 제약 조건에 의해 '버전의 장점'이 NP′가 된다. 그러나 규칙 4 에 의해 접속조사 jj 는 논항 확대를 거치지 않은 NP 와 결합하므로 예문(2)는 비문이 된다.

그 버전의 특징 (1)
*버전의 장점과 단점 (2)

4.4 서술어 확대

서술어의 확대는 부사를 첨가하는 방법과 보조적 연결어미를 이용하는 방법, 그리고 부사절과 서술절을 이용한 확대 방법이 있다. 기본문의 서술어 확대는 부사 첨가와 보조적 연결어미에 한하여 기술하고 부사절과 서술절을 이용한 확대 방법은 문의 내포에서 기술한다.

서술어는 동사, 형용사, 혹은 '명사, 대명사, 수사+이다'로 구성된다. 아래의 규칙으로 확대 이전의 서술어를 기술한다. 이후 서술어는 VP 로, 서술어 확대는 VP′로 나타낸다.

(규칙 8) VP → (pv | pa | NP+co)+[ep]+[ef]

(규칙 9) VP′ → VP

4.4.1 부사의 서술어 확대

부사를 첨가하여 서술어를 확대할 수 있다. 부사는 어휘부에서 일반부사 mag 와 접속부사 maj 로 분류되어 있다. 규칙 10 은 일반부사의 쓰임을 표현하고 규칙 11 은 접속부사의 문장부사 역할을 기술한다. 문장부사 역할은 본 문서에서 서술어 확대로 분류한다.

(규칙 10) VP′ → . .mag$^{+\leq 2}$ VP′
제약 - 부사 mag 는 VP 를 포함하는 문장내의 위치 제약이 없다.
- VP′는 단일 VP 만을 포함한다

(규칙 11) VP′ → maj VP′
제약 - 접속부사 maj 는 문두에 위치하여 서술어를 확대한다.

아래의 예문중 (1)은 부사 '바로'에 의한 서술어 확대의 전형을 보여주고 있다. 그리고 (2)는 문장내 위치 제약을 받지 않는 부사의 쓰임을, 그리고 (3)은 복합 부사의 쓰임을 보여준다. 예문 4 의 경우 접속부사 maj 가 문장부사의 역할로서 서술이 '소개한다'를 확대한다고 본다.

발전의 원동력은 바로 객체지향 언어이다 (1)
나는 오늘 자바를 소개한다. (2)
이 책은 개념적인 배경지식부터 매우 자세히 소개한다. (3)
그리고 응용할 수 있는 예제를 소개한다. (4)

4.4.2 보조 용언의 서술어 확대

보조 용언를 이용하여 서술어를 확대할 수 있다.

(규칙 12) VP′ → VP′ px$^{+\leq 2}$

예문 (1)은 단일 보조용언, (2)는 복합 보조용언의 쓰임을 보여준다 예문 (3)의 경우 '출력해 놓다'와 '저장하다'가 연결어미 '-고'로 연결된 이어진 문장으로 본다. 이는 문의 접속에서

다룬다.

Redirection 에 관한 내용을 올리기 위해 시작했다 (1)
스크립트를 출력해 놓고 보면 에러가 발생한다. (2)
파일을 출력해 놓고 저장한다. (3)

4.5 문의 내포

문장은 문장 내에 다시 문장을 취할 수 있는데 이를 절이라 한다. 절은 주어와 서술어를 필수요건으로 하며, 각각의 기능에 따라 관형절, 명사절, 부사절, 인용절, 서술절로 나뉜다. 부사절은 문장내의 내포를 원칙으로 하되, 예외를 두어 문장 밖에 사용도 허용한다. 절은 한 문장 안에서 반복적으로 사용될 수 있다.

4.5.1 관형절

관형절은 서술어 어간에 관형형 전성어미(etm)를 취하여 후행하는 NP 를 수식한다. 규칙 13 은 관형절 ETMS 을 표현하고, 규칙 14 는 ETMS 에 의한 논항 확대를 나타낸다.

(규칙 13) ETMS → S+etm

(규칙 14) NP′ → ETMS NP_1′
제약 : - NP_1′은 단일 NP 만을 내포한다.

다음 예문은 관형절에 의한 논항 확대의 예제들이다. 예문들을 살펴보면 관형절이 기본문의 논항들을 만족시키지 못하고 있음을 알 수 있다. 그러나 구성 기본문의 논항들은 피수식어나 안은문의 논항들로 대치될 수 있다. 이에 대한 내용은 표준 문형 유도에서 다룬다

{앞서 말한 특징}이 자바의 손꼽히는 장점이다. (1)
{인간이 쓰는 말}이 자연어이다 (2)
{무한 루프인 알고리즘}은 위험하다. (3)

4.5.2 명사절

명사절은 문장 S 와 명사형 전성어미(etn)이나 '-etm 것'을 결합하여 구성된다. 규칙 15, 16 은 기본문 S 에 -etn 과 '-etm 것'을 결합한 명사절 ETNS 를 표현하고, 규칙 17 은 명사절을 이용한 논항 확대를 말한다

(규칙 15) ETNS → S+etn

(규칙 16) ETNS → S+etm 것

(규칙 17) NP′ → ETNS

예문(1)-(3)은 ETNS 에 의한 논항 확대의 예제이다.

그래픽 뷰어가 대안이기를 원한다. -기 (1)
그래픽 뷰어가 대안임을 주장한다 - ㅁ.음 (2)
그래픽 뷰어가 대안인 것이 다행이다 - ㄴ 것 (3)

4.5.3 인용절

문장 내에 다른 말을 인용할 수 있다. 이때 인용되는 부분은 하나의 절 단위로 처리하며 이를 인용절이라 한다. 인용절은 인용부호를 이용할 수도 있고 아닐 수도 있다. 규칙 18 은 인용절 JQTS 에 대한 기술이다. 규칙 19 는 JQTS 에 의한 서술어 확대의 기술이다.

(규칙 18) JQTS → (S| "S")+ (-라고|-고|-하고)

(규칙 19) VP′ → JQTS VP′

4.5.4 부사절

부사절은 서술성 부사 '같이, 없이, 달리'와 연결어미 '-게, -도록, -듯이'에 의해 생성되며 후행하는 서술어를 수식한다. 규칙 20 은 서술성 부사를 이용한 부사절 생성이고 규칙 21 은 연결어미를 이용한 부사절 생성을 보여준다 규칙 22 는 부사절 ADVS 가 문장내 위치제약이 없으며 서술어 확대에 적용됨을 나타낸다.

(규칙 20) ADVS → NP1 [NP2] [NP3] (같이|없이|달리)
제약 : - '같이 없이 달리'는 서술성 부사로써 부사절을 생성

(규칙 21) ADVS → NP1 [NP2] [NP3] (palpv)+[ec]+(-게|-듯이|-도록)
제약 : - '-게, -듯이, -도록'에 한하여 부사절 생성

(규칙 22) VP′ → ADVS VP′
제약 - ADVS 는 문장내 위치 제약이 없다.

예문 (1), (2)는 서술성 부사의 부사절 생성을 보여주는 예문이다. 예문 (3),(4)은 연결어미 '-게'에 의한 부사절 생성이고, 부사절의 문장내 위치가 자유로움을 보여준다 예문(5)는 부사절이 필수논항을 충족시키지 못하고 있다. 그러나 필수 논항에 대한 추정을 안은문의 필수논항으로 대치하여 정문으로 간주할 수 있다. 이에 대한 논의는 표준 문형 해석에서 다룬다

헤커는 흔적도 없이 침입한다 (1)
넷스케이프는 우리가 예상한 것과 달리 변화가 없다 (2)
우리가 예상한 것과 다르게 넷스케이프는 변화가 없다 (3)
넷스케이프는 우리가 예상한 것과 달리 변화가 없다 (4)
시스템이 빠르게 작동한다. (5)

4.5.5 서술절

서술절은 이중주어 문제를 다룬다. 중복된 NP1 이 문장에서 발생할 경우 서술절과 근접한 NP1 과 서술절을 VP'로 본다.

(규칙 23) $VP' \rightarrow NP1\ VP_1'$
제약 : - 서술절은 순환적으로 적용될 수 없다.

프로그램이 실행속도가 느리다. (1)
시스템이 용량이 크다. (2)
*그 회사는 시스템이 용량이 크다. (3)

4.6 문의 접속

문의 접속은 연결어미를 사용하며 아래의 형식을 유지한다. '게, 도록, 듯이'는 문의 접속이 아닌 부사절로 처리한다. 규칙 24 는 문의 접속을 기술하며, 순환적 적용이 허용되지 않음을 알 수 있다.

(규칙 24) $S' \rightarrow S_1$+ec S_2

예문 (1),(2)는 문의 접속의 예제이다. 예문 (3)는 S_2 의 NP1 이 S_1 앞으로 위치하여 규칙 24 에 위배되어 비문이다.

이 기법은 에이젠트를 실행하며 문서를 요약하며 DB 에 저장할 수 있다. (1)
HTML 로 수평선의 길이를 조절하거나 굵기를 바꿀 수 있다 (2)
*화질은/NP1 (컴퓨터를 사용하여 많은 모니터를 연결할수록) 나빠진다 (3)

다음 예문(4),(5)은 이어진 문장으로 보지 않고 부사절로 본다.

포맷을 한듯이 아무 짓도 없다 (4)
하드웨어 특성을 타지 않도록 호환성 여부를 doc 문서를 통해 확인한다. (5)

5. 표준문형 유도

본 절은 사용자의 비정규적인 언어 표현을 표준문형으로 유도할 수 있는 접근 방법에 대하여 기술한다. 사용자가 표준 문형에 따른 텍스트를 생성할 때 그 절차의 복잡도는 응용 시스템에 있어 중요한 요소로 작용할 수 있다. 따라서 표준문형과 같은 제한된 언어표준을 따르는 응용 시스템의 텍스트 입력은 사용자에게 표준문형안에 대한 특별한 지식을 요구하지 말아야 하며, 일상적 문장 표현과의 차이점을 줄일 수 있어야 한다.

ACE 를 이용하는 Attempto 의 경우 사용자에게 제한된 언어 표준에 대한 학습을 가정하고 있으며, 입력 문장을 바꾸어 쓰기(paraphrase) 형식으로 사용자에게 언어 기술 표준에 적합한 문장 구사를 유도하는 방법을 취한다[1][2].

문형 표준안의 관점에서 일상적 한국어 문장 기술에서 고려될 사항들로는 중요성분 생략현상, 자유 어순, 구조적 모호성을 발생시키는 관형어의 피수식어 선택 문제들이 있다. 이에 대한 접근 방법으로는 실시간 문법 검사 방법과 바꾸어 쓰기(paraphrase)를 통한 표준 문형 유도 방법이 타당하겠다고 본다. 따라서 본 절에서는 바꾸어 쓰기(paraphrase)를 위한 기본문의 해석, 관형어 표현의 해석, 생략 추정 패턴과 추가적인 몇가지 해석 규칙을 제시하고 이를 이용한 실시간 표준 문형 검사의 예를 보여준다.

5.1 기본문 해석

기본문은 서술어의 필수 논항 개수에 따라 1 항 술어, 2 항 술어, 3 항 술어로 구분된다. 논항은 격표지에 의해 논항 1, 논항 2, 논항 3 으로 구성된다. 문형 표준은 서술어와 논항의 개수가 일치해야 함을 말하고 있으며 그렇지 않은 경우 비문으로 간주한다 논항에 대한 확인은 격표지와 대상 어휘의 위치에 따라 결정한다. 즉 보조사 '은, 는'이 주어 자리에 위치하며 NP1 으로 결정한다.

5.2 관형어 해석

관형이에 의한 논항 확대의 경우 피수식어의 제약 조건으로 단일 NP 만을 내포한 NP'로 한정하였다. 규칙 7 과 규칙 14 에 의해 기술된 제약된 관형어의 표현은 피수식어 선택의 모호성을 제약하기 위함에 그 목적이 있다.

5.3 기본문 논항 생략

내포절이나 문의 접속에서 기본문의 필수 논항이 생략되었을 경우, 필수 논항에 대한 추정에 의해 해석을 진행할 수 있다 4.3.1 과 4.3.2 는 각각 생략현상에 대한 문형 표준안의 접근

방법에 대한 기술이다. 문장의 생략된 논항에 대한 추정 패턴과 각 입력문장에 대한 바꾸어 말하기(paraphrase)의 예를 보여주고 있다.

5.3.1 내포절의 논항 생략

피수식어=>**NP1**
(1) 불편한 OS 는 없어진다. :[{OS 가) 불편한] OS 는 없어진다.

피수식어=>**NP2**
(2) 바이러스가 침투한 시스템은 포맷이 낫다. [바이러스가 (시스템을) 침투한] 시스템은 포맷이 낫다

피수식어=>**NP3**
(3) 사용자 중에는 인터넷에 필요한 통신장비를 사지 않는 사람이 많다 '사용자 중에는 [인터넷에 (통신장비가) 필수적인] 통신장비를 사지 않는 사람이 많다.

안은문 주어=> **NP1** ^ 피수식어=>**NP2**
(4) 맥킨토시가 인식하는 디바이스가 따로 있다 맥킨토시가 [(맥킨토시가 디바이스를) 인식하는] 디바이스가 따로 있다.

안은문 주어=> **NP1**
(5) 서버가 해킹을 당한 사실을 스스로 인식한다. 서버가 [(서버가) 해킹을 당한] 사실을 스스로 인식한다.

NP+jm ETNS 에서 **NP** 를 **ETNS** 의 **NP** 으로 추정
(7) 프로그래머는 보여준 사실을 몰랐다. ?프로그래미는 [(프로그래머가) 보여준] 사실을 몰랐다

5.3.2 문의 접속의 논항 생략

후행절의 논항 추정
(1)1 번 프로세스가 프로그램을 설치하면 관리를 하여준다. : 1 번 프로세스가 프로그램을 설치하면 [1 번 프로세스가] 관리를 하여준다

선행절의 논항 추정
(2) 버그가 존재하여 프로그램을 수정하였다. : 버그가 [프로그램에] 존재하여 프로그램을 수정하였다

5.4 일상적 문장 표현에 대한 해석 규칙

일상적으로 빈번히 사용되는 한국어 문장 표현들은 추가적 해석 규칙으로 바꾸어쓰기(paraphrase)에 적용하는 것이 타당하리라 본다. 본 절에서는 "명사구 해석 규칙"과 "명사(조사생략)+서술어" 표현에 대한 해석 규칙을 제시한다.

5.4.1 명사구 해석 규칙

(1) N + NV => N{논항} NV{서술어}
(2) NV + N => NV{관형어} N
(3) NV + NV => 비문(명사구를 고유명사로 인식)
(4) N + N => N{관형어} N

* 비서술성 명사 N, 서술성 명사 NV

예) 다이얼로그 박스의 입력값 유지 방법
. (입력값/{논항} 유지/{서술어})/{관형어} 방법

입력값/N 유지/NV : 해석규칙(1) 적용
(입력값 유지)/N 방법/N . 해석규칙(4) 적용

5.4.2 명사(조사생략) + 서술어 해석 규칙

조사 생략된 명사는 서술어 논항 항목중 인접 항목으로 결정

(1) NP1 서술어 => 명사(조사생략)/NP1
(2) NP1 NP2 서술어 => 명사(조사생략)/NP2
(3) NP1 NP2 NP3 서술어 => 명사(조사생략)/NP3

예) WM_MESSAGE 를 이용하여 다른 응용프로그램 제어하는 방법
. 응용프로그램/NP2 제어하다{2 항술어} : 해석규칙(2) 적용

5.5 실시간 표준 문형 검사

다음은 표준 문형안의 실시간 입력문장 검사 과정을 보여준다. (2)는 NP2 인 "클래스를"에 의한 NP1 추정슬롯 생성을 (4)는 NP3 로 추정된 '파일로부터"에 의한 NP1 NP2 추정 슬롯 생성을 보여준다. (5)는 NP2 추정슬롯의 삽입을 보여준다. 최종 검사된 (7)은 이어진 문장으로 입력문장을 해석하고 미결된 슬롯 $NP1^1$ $NP1^2$ 를 제시한다. 이 경우 사용자의 선택에 따라 슬롯 삽입이 진행될 수도 있고 변수로 진행될 수도 있다고 본다.

예문) CstdioFile 클래스를 이용하여 파일로부터 라인을 읽다.

(1) **CstdioFile**
(2) NP1 CstdioFile 클래스를
: NP1 추정슬롯 삽입

(3) NP1 CstdioFile 클래스를 이용하여
· 기본문 확인 "[NP1] CstdioFile 클래스를/NP2 이용하-"

(4) $NP1^1$ CstdioFile 클래스를 이용하여 $NP1^2$ NP2 파일로부터
. "파일로부터"- NP3 추정 - $NP1^2$ NP2 추정슬롯 삽입

(5) $NP1^1$ CstdioFile 클래스를 이용하여 $NP1^2$ NP2 파일로부터 라인을
. "라인을" - NP2 추정 - 라인을 ^{NP2}

(6) $NP1^1$ CstdioFile 클래스를 이용하여 $NP1^2$ [라인을]^{NP2} 파일로부터 라인을 읽다.
기본문 확인 "[NP1} 라인을/NP2 파일로부터/NP3 읽다"

(7) {$NP1^1$ CstdioFile 클래스를 이용하-}**S1**-여 {$NP1^2$ [라인을]^{NP2} 파일로부터 라인을 읽다}**S2**
S1 + ec S2 연결문 확인
슬롯 $[NP1]^1$, $[NP1]^2$ 미결

6. 어휘부

고정된 어휘를 제공하고 신규 어휘에 따른 어휘부 확장 방법에 대한 기술이 일반적인 제한 언어의 접근 방법이다. 이는 정보거래 자동 중개 시스템에 있어 필수적인 요소이다. 문형 표준안의 어휘부는 어휘 범주별 고정된 어휘 정보가 있다고 가정하며, 특히 서술어의 경우 서술어-논항구조에 적합한 단일 하위범주화 정보를 갖고 있다고 가정한다. 표 1 은 본 문형 표준안이 따르고 있는 어휘 범주에 대한 기술이다. (1)은 ETRI 표준 어휘 범주 분류이고 (2)는 이중 문형 표준안에서 기술되고 있는 추가적인 세부 어휘 범주를 기술한 것이다. 그리고 (3)은 문형 규칙에 사용된 어휘들의 목록이다.

(1) f(외국어), nc(자립명사), nb(의존명사), np(대명사), nn(수사), pv(동사), pa(형용사), px(보조용언), co(지정사), mag(일반부사), maj(접속부사), mm(관형사), ii(감탄사), xp(접두사), xsn(명사 파생 접미사), xsv(동사 파생 접미사), xsm(형용사 파생 접미사), jc(격조사), jx(보조사), jj(접속조사), jm(속격조사), ep(선어말어미), ef(종결어미), ec(연결어미), etn(명사형어미), etm(관형형어미)
(2) 발화동사, 수여동사, 주격조사, 목적격조사, 보격조사, 부사격조사, 지시관형사, 수관형사, 성상관형사, 서술성부사
(3) 것, -라고, -하고, -고

표 1 어휘범주 목록

7. 결론 및 향후 연구 방향

본 논문은 정보거래 자동 중개 시스템의 소개와 요소 기술인 의미 기반 정보검색을 위한 한국어 문형 표준안의 필요성을 제시하였다. 그리고 문형 표준안의 관점에서 일상적 한국어 문장 기술에서 고려될 사항들로 중요성분 생략현상, 자유 어순, 구조적 모호성을 발생시키는 피수식어 선택 문제에 대한 접근 방법으로 실시간 문법 검사 방법과 바꾸어 쓰기(paraphrase)를 통한 표준 문형 유도 방법을 소개하였다.

그림 2 는 정보거래 자동 중개용 문서 구조를 보여준다. 문서는 문형-의미 표준화를 거친 요약문장으로 구성되는 메타 정보와 내용 정보로 구성된다. 메타 정보는 문형 표준화 과정 이후의 단계로 의미 표준화 과정을 필요로 한다. 문형-의미 표준화 과정은 대화형 중의성 해소 방식을 이용한다. 현재 진행되고 있는 의미 표준안은 문장을 CG(Conceptual Graph) 형식으로 표현하는 것이다. 따라서 어휘간 개념관계를 파악할 수 있는 온톨로지 정보 구축과 CG 기반 색인 및 검색 기술이 향후 진행될 것이다. 또한 일반 사용자들이 생성하는 텍스트를 자연스럽고 유연하게 문형-의미 표준화에 적합하도록 유도할 수 있는 방법론과 지원도구-시각화기술(visualization)을 적용한 문서작성지원도구-가 개발될 것이다.

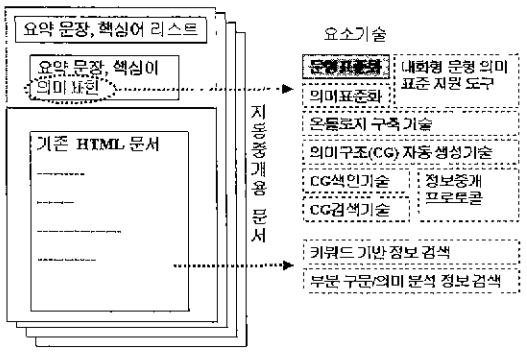

그림 2 자동 중개용 문서 구조와 관련 요소기술

참고 문헌

[1] N. E. Fuchs, R. Schwitter. "Attempto Controlled English (ACE)", CLAW 96, March 1996

[2] N E. Fuchs, U. Schwertel, R Schwitter, "Attempto Controlled English (ACE) Language Manual Version 3.0", Dept of Computer Science, Univ. of Zurich, March 1999

[3] G. Adriaens, D. Schreurs, "From Cogram to Alcogram Towards a Controlled English Grammar Checker", Proceedings COLING 92, pp 595-601, 1992

[4] T. Mitamura, E.H. Nyberg, 3rd, "Controlled English for Knowledge-Based MT: Experience with the KANT System", Center for Machine Translation, Carnegie Mellon University, Pittsburgh, 1995

[5] I. Androutsopoulos, G. D. Ritchie, P. Thanisch, "Natural Language Interfaces to Databases – An Introduction", Journal of Natural Language Engineering, vol.1, no 1, Cambridge University Press, 1995

[6] 서정수, 국어 문법, 한양대학교 출판원, pp 1344, 1996