주식 투자 영역에서의 한글 자연어 처리 시스템의 설계 및 구현

신 성우, 이 일병
연세 대학교 전산과학과

A Design and Implementation of Hangul Interface System for Stock Domain

Sungwoo Shin and Yillbyung Lee
Department of Computer Science, Yonsei University

요약

본 연구는 주식투자 영역에서의 한글 인터페이스 시스템의 구현에 관한 것으로 영역에 대한 지식을 기반으로 하위 계층 지식을 설계하고, 의미 분석을 중심으로 효율적인 중간 표현을 생성하여 지능적인 응답문의 생성을 목표로 시스템을 구현하였다. 또한 제한된 문맥적 상황에 대한 인식 능력을 보유하여 생략어구와 대용어구의 처리가 가능하도록 하였다.

Ⅰ. 서론

자연어 분야에 대한 연구의 관심은 초기에 기계번역 분야가 많이 연구 되었으나 주로 단어에 대한 변환 방식으로, 이러한 변환 방식은 자연어의 내면 구조에 대한 연구없이 곧바로 표상적인 특징만을 갖고 번역을 시도함으로써 만족할 만한 결과를 제공하지 못했다.
그러나 1970년대에 들어서면서부터 언어학에서의 구문론과 의미론에 대한 복합적 연구와 컴퓨터 기술의 비약적인 발전에 힘입어, 현재 대부분의 자연어 처리 시스템은 다양한 지식 표현 기법을 사용한 지식 기반 시스템들로서 몇몇 분야에서는 상업화된 시스템도 발표되어 주목할 만한 성공을 거두고 있다. 그러한 상업화된 시스템의 대부분이 데이타베이스에서의 자연어 인터페이스이며, 컴퓨터에 생소한 일반 사용자들에게는 간단한 형식의 정형 질의어라도 완전히 이해하고 숙지하여 사용하는데 상당한 부담이 따르므로 한글을 사용하여 원하는 정보를 검색할 수 있는 환경을 제공하는 것은 상당히 의의있는 일이라 할 수 있다.

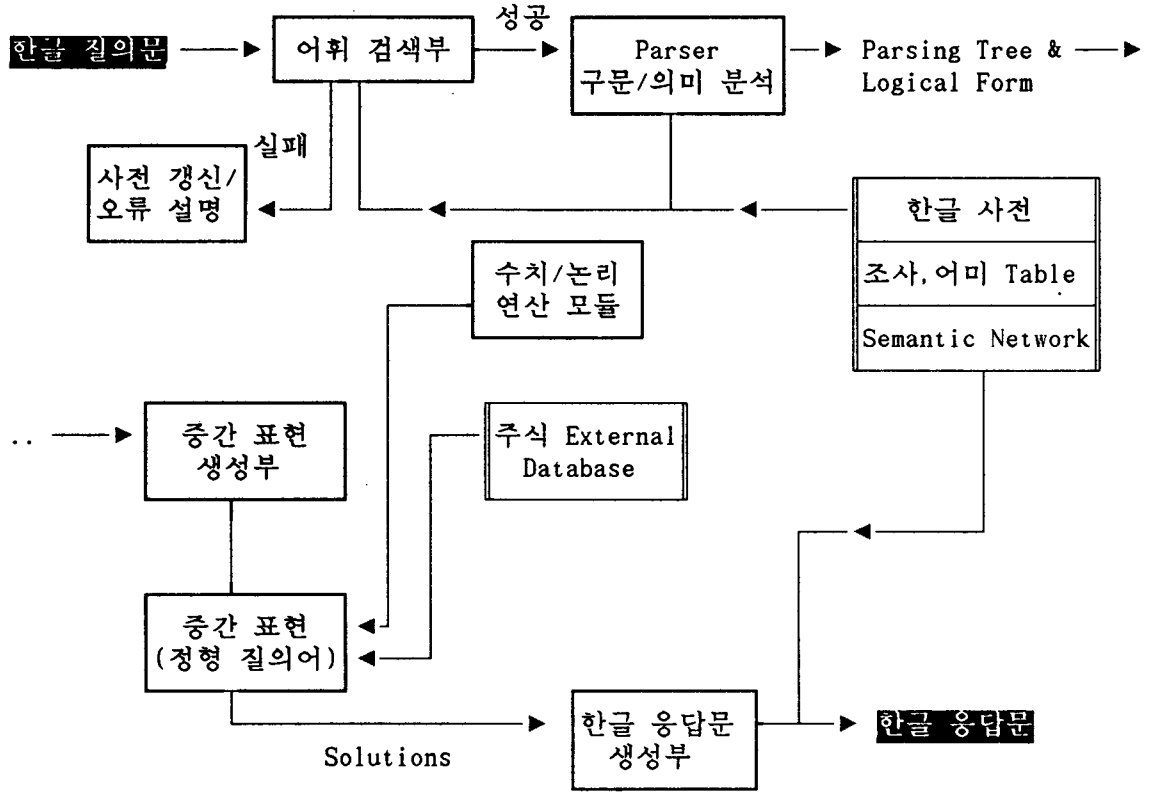
본 연구는 Prolog External Database로 구성된 주식/증권 영역에 대한 한글 질의문을 사용할 수 있게하는 한글 인터페이스 시스템의 개발에 관한 것이다. Semantic test와 Extra condition을 첨가한 Logic grammar로써 질의문을 정의하고 분석하였으며 모호한 구문의 심층 의미분석을 위하여 Semantic grammar를 적용하여, 질의문 처리를 위해 필요한 지식을 추출하고 이를 이용하여 한글 응답문을 생성하는 기법을 연구 하였다.

또한 기존의 시스템들 대부분이 대화에 대한 상황(conversational context)의 처리 없이 데이타베이스에 대한 단순하고 편리한 질의 언어로서의 자연어 인터페이스를 제공한다.
그러므로 실제로 시스템의 기능은 입력된 자연어 질의문을 실제 데이타베이스에 대한 질의어로 변환하여 연산하는 일이라 할 수 있다. 이에 반하여 본 시스템은 질의문들에 대한 제

한된 문맥적 상황을 인식하여 [14] 생략문 처리나 대용어 처리등의 수행능력을 보유하며, Logic grammar와 Semantic grammar를 적절히 혼용하여 의미론적으로 모순된 구문분석을 초기에 제거하여 (semantic filtering) parsing의 효율을 향상 시켰다. 또한 한글 질의문에 대한 중간 표현으로, 테이타베이스 구조에 곧바로 적용할 수 있는 효율적 구조의 형식을 채택하여 질의문내 변수에 대한 binding time을 최소화 할 수 있도록 하였다.

Ⅱ.시스템 개요

본 시스템은 한글 질의문을 입력으로 받아 이를 Prolog의 정형 Query로 Mapping하여, Prolog External Database로 구성된 주식투자 영역에서의 산업과 기업지식을 검색하여 요구하는 해를 찾아 이를 한글 응답문으로 생성하는 시스템으로써 다음과 같은 구조를 지닌다.



Ⅲ.한국어 질의문에 대한 고찰과 테이타베이스의 구조

업종과 회사에 대한 두 가지 지식으로 구축되는 제한된 영역에서의 데이타 베이스에 대한 질의문은 어떠한 데이타나 대상에 대하여 요구되는 내용을 구체적으로 서술함으로써 시스템으로 하여금 요구 내용을 파악할 수 있게 하여야 한다. 한글 질의문을 사용할때에는 요구되는 대상을 서술키 위해서 관형구나 관형절을 통하여,혹은 대상간의 관계를 나타내는 용언을 사용하여 그 대상을 구체화한다. 예를 들어 "전자업종의 회사", "주가가 45000원 이상의 회사", "컴퓨터를 생산하는 회사"등에서는 회사를 수식하는 관형구, 절, 용언등이 그 대상을 구체화하고 있다. 즉, 한글 질의문은 요구되는 대상을 구체화하는 관형구나 관형절과 그 대상인 명사 , 그리고 각 대상간의 관계를 나타내는 용언으로 주로 구성되며 "X는 무엇인가?", "X는 Y인가?", "X를 ...하여라"등의 의문문과 명령문으로 구성된다.

한글 질의문은 그 의미가 누구에게나 인식될수 있어야하는 보편 타당성을 지녀야 한다. 그러나 타당한 올바른 의미의 질의문이라 할지라도 목적 데이타 베이스에서의 한글 질의문으로 항상 유효한 것은 아니며 [1,3] 다시말해 데이타 베이스의 구조와 일치하지 않는 의미를 지닌 질의문이나 저장되지 않은 내용에 관한 질의문은 올바른 중간 표현으로 변환되지 못하여 원하는 결과를 얻지 못할것이다 .위와 같은 데이타 베이스에서 예측 되는 한글 질의문을 몇 가지 유형으로 분리해 보면 다음과 같으며, 각 유형의 질의문은 복합되어 사용될 수 있으며 질의문의 최소 단위로 생각될 수 있다.

<유형1.가장 평범한 인가 의문문>
삼성전자의 자산(주가,PER,주요 제품,업종,소속 그룹..)은 얼마(무엇)인가?
<유형2.형용사 의문문>
1989년도 전자업종에서 어떤 회사의 주가가 가장 비싼가?
<유형3.동사 의문문>
대우전자는 무엇을 (어떤 제품을) 만드는가? (생산하는가?)
<유형4.대용어구와 생략어구가 포함된 의문문>
그것이 속한 업종의 1989년도 최고 거래량은 얼마인가?
금성통신은?(1989년의 최고 주가는?)
(바로 이전의 질의문이 "삼성전자의 종업원수는 얼마인가?" 일때)
<유형5.연산자가 사용된 OP_의문문>
1989년도 최고 PER이 1.5보다 많고 자본금이 500억원 이상의 회사는 무엇인가?
<유형6.내포문이 포함된 복합 의문문>
현대전자가 속한 업종의 업종지수와 거래량은 얼마인가?
<유형7.가부 결정을 묻는 판정 의문문>
삼성전자의 자본금은 전자업종에서 가장 많은가?

Ⅳ.시스템 설계 및 구현

1.시스템을 구성하는 지식 요소

1.1. 한글 사전

사전은 주어진 주식 데이타베이스 환경 하에서 정의되고, 쓰여지는 단어의 집합으로써 그 종류에는 일반명사(영역 명사), 고유명사(상수 명사), 의존 명사, 형용사, 동사, 부사, 관형사등과 시스템 명령으로 사용되는 키 워드로 구성되어진다. 명사는 단어 필드와 상수/영역의 구분 필드, 소속 영역, 조사정보의 네 개의 필드로 구성되며,상수/영역 구분 필드는 데이타베이스의 항목인가의 여부에 따라 구분되고, 조사정보는 완성형(7bit-2byte) 한글코드의 사용으로 인해 분석이 불가능한 종성의 유무를 표시하여 뒤에오는 조사의 형태(은/는, 이/가,을/를..)를 알 수 있게 하였다. 또한 "무엇", "얼마"등의 불특정 영역에 소속될 수 있는 명사는 영역명사로 따로 저장하였다.

```
(예: noun("삼성전자","constant","회사c","","type1""),
     noun("업종","domain","업종d","type2"),
     domain_noun("무엇","constant",["업종c","회사c","주제품c",...],"type2") )
```

동사는 기본형과 어간,가능한 하위 범주화 정보, 타동사/자동사의 구분 필드의 4 필드로

구성되며, 형용사는 동사와 마찬가지로 기본형과 어간의 필드와 하위범주화 정보, 소속 영역 으로 구성된다.

(예 : verb(생산하다 , 생산 , [Subj:회사,Obj:주제품],transitive)
adj(비싸다,[비싼,비쌌던],[주가_high])
adj(높다,[높은,높았던],[주가_high,배당율_high,PER_high..))

1.2.조사/어미 테이블

시스템이 허용하는 조사의 종류에는 주격, 목적격, 부사격, 관형격, 보격,서술격조사등이 있으며 어미의 종류는 의문형, 명령형의 종결어미와 관형형어미의 전성어미의 사용 만으로 제한한다. (청유형, 서술형 어미는 질의문에서 의문문과 명령문만의 사용으로 인해 배제한다.)

1.3 영역 계층의 의미망

각 영역표시 명사에 대한 분석내용(기업,산업,투자)과 단어에 대한 소속 영역의 계층 구조를 명시한다.

1.4.데이타 베이스

본 시스템에서의 데이타 베이스는 크게 업종에 대한 지식과 회사에 대한 세부 지식의 두 가지 지식과 주식용어, 금융상품에 대한 지식으로 구현되며,TURBO PROLOG의 EXTERNAL DATABASE를 사용하여 구현한다.

다음의 표는 데이타 베이스를 구성하는 다섯 종류의 항목이다.

1. 업종정보([업종명],[회사수],[업종지수],[거래량])

2. 회사개요(회사명,업종명,[설립일,상장일],종업원수,
결산기,소속계열,주거래은행,수권자본금)

3. 주가관련지표(회사명,주당순이익,주당순자산,[배당율(현금/주식)],
[주가(최고,최저)],[PER(최고,최저)])

4. 주요재무비율(회사명,매출액증가율,부채비유,유보율)

5. 증자&주제품(회사명,[증자사항],[주요제품])

6. 금융상품(상품명,설명)

7. 주식투자(용어,설명)

2.시스템 구현 단계

2.1.어휘 검색

한글 질의문은 사전에 기재된 단어와 수량, 금액등의 기재할 수 없는 단어로 구성된다. 어휘 검색 단계에서는 사용자의 오타나 사전에 기재되지 않은 단어의 포함여부를 검사하여 해당 단어에 대한 오타 여부와 새로운 지식을 요구하여 사용자 스스로 오류 수정과 사전 구축을 허용한다. 특히 사전 구축시에는 데이타 베이스 구조에 의존적인 새로운 단어의 추가

는 허용되지 아니하며, 기존의 용언등에 대한 동의어등 시스템의 골격을 이루는 지식을 침해하지 않는 범위에서 허용된다.

사전에 기재된 단어에 대해서는 조사/어미 표의 내용과 사전에 기재된 어휘와의 결합성을 검사하며, 미기재된 단어는 뒤에 나오는 단위 명사(원,%,주..)와 어휘 패턴[2]에 의한 검사로 그 단어에 대한 자질과 소속 영역을 결정한다.

(예:명사-> 모음,자음으로 끝나는 여부에 따른 조사의 결합성 검사

용언-> 사전의 어간과 결합가능한 모든 어미의 패턴과의 결합성 검사)

또한 질의문의 띄어쓰기는 새맞춤법의 띄어쓰기 원칙[2]의 41항부터 50항에 따름을 원칙으로 한다.

2.2.구문 분석 및 의미 분석

본 시스템에서는 한글 질의문을 Top down proof method의 Logic Grammar로 정의하고 분석하였으며, 제한된 영역에서의 가능한 질의문은 예측 가능하므로 몇몇의 경우에는 구문분석 초기의 애매성을 제거하기 위하여 각 범주에 대한 의미 정보를 활용한 Semantic Grammar를 적용하였다.또한 Prolog로 구현된 Logic Grammar에 각 단어간의 영역계층검사와 자질검사,용언에 대한 하위범주화 정보를 이용한 속성검사등의 의미 검사를 동시에 실시하며 질의문에 대한 Logical Form은 문법 규칙으로부터 자동 생성케 하여 나중에 데이타 베이스 검색에 직접 활용되는 중간 표현(정형 질의어) 생성에 용이하게 사용되게 하였다.

이상 위에서 설명했듯이 시스템에서는 구문분석과 의미분석을 동시에 수행하여 Logical Form을 생성하며, 이를 이용하여 사실(prolog fact)로 기술된 단위별 데이타 베이스에 대한 정형 질의 표현에서 다음과 같은 중간 표현들의 집합을 생성하며 중간 표현속에는 각종 연산 루틴이 포함되어 연산자나 수치 비교의 계산을 수행 한다.

예 1) 1989년도 최고 PER이 20보다 크고 자본금이 500억원 이상의 회사는 무엇인가?

```
(Retrieve ?goal (주가관련지표(_,_,_,_,_,?x);
                (Find_all ?y  (?x > 1.5));
                (주가관련지표(?sg1,_,_,_,_,?y);
                (회사개요(_,_,_,_,_,_,_,?x1));
                (Find_all ?y2 (?x1 >= 500));
                (회사개요(?sg2,_,_,_,_,_,_,?y2));
                (Find_all ?goal (Intersect ?goal (?sg1 ?sg2))))
```

예 2) A회사가 속한 업종의 업종지수와 거래량은 얼마인가?

```
(Retrieve ?g1 ?g2 (회사개요("A회사",?x,,_,_,_,_,_,_,_));
                  (업종정보(?x,_,?g1,?g2)) )
```

예 3) A회사의 자본금은 B업종에서 가장 많은가?

```
(Return yes/no (회사개요(_,"B업종",_,_,_,_,?x,_));
               (Find_max ?y in the list of ?x);
               (회사개요("A회사",_,_,_,_,_,?y,_)) )
```

2.3.특수 구문의 처리 (생략어구와 대용어구의 처리)

자연어에서는 의미전달에 차질이 오지 않는 범위 내에서 문장의 일부를 대명사로 대치하거나,이미 주어진 정보를 되풀이하여 반복하지 않으려고 문장 성분의 일부를 생략하여 사용하는 경우가 빈번하다 [8]. 본 시스템은 이러한 대용어의 처리와 생략어구의 재생 처리를 제공 함으로써, 기존의 시스템들에서 부족한 문맥에 대한 상황인식 능력을 보유하여 사용자

가 보다 편리하게 질의할 수 있게 하였다.

전제 조건

1. 시스템은 전 질의문의 수식부와 주부로 이루어지는 의미 표현을 항시 기억한다.
2. 전 질의문에 대한 해의 리스트를 기억한다.

A> 대용어구의 처리

본 시스템에서 허용하는 대용어구는 대명사와 관형사+명사의 두 형태로 구분된다. 대명사에는 "그것","그"(+의존명사) 등이 사용되고,"그러한","그"등의 관형사와 명사의 형태로 대용어구가 구성된다. 대용어구는 다음의 질의문에서 보듯이 한 질의문내(복문)에서 혹은 연속된 두 개의 질의문에서, 앞 문장의 명사(혹은 "해")와 동일한 명사가 다음 질의문에서 나타날때 사용된다.

```
질의문 A1) A업종에 소속해 있는 회사는 무엇인가?
질의문 A2) 그것 중에서 자본금이 B보다 많은 회사는?
          그것들의 (그러한 회사들의) 자본금과 1989년의 증자 사항을 보여라
질의문 B1) A회사의 주요 제품은 무엇인가?
질의문 B2) 그것(그 회사)의 종업원수는 무엇인가?
          그것들을 생산하는 다른 회사는 무엇인가?
```

위 질의문 A에서는 소속회사의 해(예:a1,a2,a3,a4,a5..)에 대하여 대용어구가 사용 되었으며(영역 명사:"회사"),질의문 B에서는 전 질의문 표층에 나타난 명사(상수 명사:"A회사")와 주요 제품의 해에 대하여 대명사를 사용하였다.

이상에서 살펴 본 바와 같은 대용어구의 사용 가능한 예를 바탕으로 본 시스템에서의 대용어구에 대한 처리 방법을 다음과 같이 두 가지로 제시한다.

(1.a) 대용어구가 수식하는 인접한 명사로부터 대용어구의 가능한 영역을 찾는다.
(영역 계층에 대한 의미망 이용)
(1.b) 선행 명사들을 차례로 대입하여 의미소성(영역)이 일치하는 명사를 찾는다.
(1.a),(1.b)의 과정에서 발견된 대용어가 상수 명사일 때에는(B2의 첫 질의문) 그 명사의 영역과 함께 새로운 문장 정보로써 추가되며, 영역 명사인 경우에는(질의문 A2) 버퍼에 있는 전 질의문의 해와 각 명사의 영역이 추가한다.
(2) 대용어구와 인접한 용언의 사전정보(하위범주화 정보)로부터 가능한 영역을 찾는다.
(B2의 두 번째 질의문)

B> 생략어구의 처리

보통 한글에서는 명사구 혹은 동사의 생략에 따라 뒤에 오는 명사구의 생략과 앞의 동사구가 생략되는 두 방향의 생략이 많이 쓰이게 된다. 본 시스템에서는 질의문의 특성상 동사의 생략은 배제하기로 하며,명사구의 생략은 앞,뒤의 두 방향 모두를 허용하며 하위범주화 되는 성분이 문맥상 분명한 경우로 제한한다. 또한 생략어구를 주부와 수식부의 생략으로 구분하여 처리하며,질의문의 내용을 제한하는 수식어구의 생략을 별도로 처리한다.

질의문 A) A업종에 있는 B회사의 종업원수는 얼마인가?
질의문 B.1) C회사는?
질의문 B.2) 1989년의 최고 주가는?
질의문 B.3) 1989년의 업종 지수와 거래량은?
질의문 C) A업종에서 1989년도에 제일 비쌌던 회사는?

형태 1) 주부의 생략

질의문 B.1는 목표부(주부:"종업원수")가 제거된 생략의문문으로 전 질의문의 제한요소 "A업종"를 이어받아 "A업종에 있는 C회사의 종업원수는 얼마인가?" 의 완전한 질의문으로 재생처리한다.

형태 2) 수식부의 생략

질의문 B.2는 조건부(수식부:"A업종에 있는 B회사")가 제거된 생략 의문문으로 "A업종에 있는 B회사의 1989년의 최고 주가는?" 의 완전한 질의문으로 재생한다. 또한 단순히 "자본금은?" 이라는 질의문은 수식부의 Gap으로 처리하나, 엄격한 의미에서 위 질의문은 관형형 수식부가 존재하여 수식부와 주부의 영역 포함관계를 검사하게 된다.즉, "1989년"과 "주가"는 서로 상이한 영역계층이 아닌 하나의 영역 "89주가"를 이루므로 "주가" 영역과 계층을 이루는 위 "B회사"의 "회사"영역이 생략어로 선정된다.

B.3의 질의문은 위와 같은 절차를 거쳐 "A업종"이라는 수식어를 생략된 명사로 찾아 "A업종의 1989년의 업종지수와 거래량은?"의 질의문으로 재생한다.

형태 3) 의미상 주어의 생략

위 질의문 C는 용언 "비싸다"에 대한 의미상 주어인 "주가"가 생략된 질의문으로 이러한 의미상 주어에 대한 생략은 문법 규칙에서 gap(영어의 filler_gap dependendy 현상에서의 gap과 유사하게)으로 처리하며 용언의 하위범주 정보를 이용하여 생략된 의미상 주어 "주가"를 재생한다.

2.4.응답문의 생성

응답문 생성은 파싱과정에서 얻어진 주 성분에 조사/어미 표의 조사 정보로부터 조사가 추가된 주어를 중심으로,추론 단계에서 구해진 응답정보를 이용하여 응답문을 생성한다. 즉,표층에 나타난 주어나 무엇에 해당하는 명사를 찾아 해와 접속하여 응답문을 구성한다. 또한 응답정보의 내용과 질의문과의 관계에 따라(특히 가부를 묻는 판정 의문문) 어휘항목의 사전정보와 더불어 몇 개의 법칙을 사용하여 응답문을 생성한다.

V.실험 및 문제점분석

본 시스템은 IBM PC-AT상에서 7bit-2byte 완성형 한글코드를 사용하여 Turbo Prolog로 구현하였다. 주식 투자 영역에 대한 데이타 베이스는 현대증권에서 발행한 상장기업분석에 근거하여 산업분석과 기업분석에 대한 지식으로 구성되었으며, 기업분석에 사용되는 지식은 일반 사용자에게 가장 빈도수 높게 문의된다고 생각되는 항목을 선정하여 구축하였다. 또한

질의문에 대한 문법 규칙은 가능하다고 생각되는 약 100개 정도의 질의문을 선정하여 이를 다시 7개의 단위 유형으로 구분하여 그 유형들의 가능한 접속을 고려하여 구성하였다.

현재의 시스템은 질의문의 분석단계에서 영역(domain) 의존적인 지식을 많이 사용하여 구문분석과 의미분석을 동시에 수행하므로 다른 영역에로의 전환에는 상당한 문제점을 지니고 있으며, 좀더 일반적이고 영역에 의존하지 않는 인터페이스 시스템을 구축하기 위해서는 상당히 확장된 규모의 문법을 중심으로 구문 분석이 이루어진 후에 영역 의존적인 의미분석이 수행되어야 할이다. 또한 사용자 편의의 측면에서 데이타베이스 구조나 항목의 단위, 범위 혹은 사전지식등에 대한 사용자의 의미상의 오류에 효과적으로 대처할 수 있는 능력이 요구되며, 질의문의 해에 대하여(특히 판정의문문의 경우) 해가 찾아지기까지의 추론과정을 효율적으로 보이기 위한 확장된 응답문의 생성이 요구된다.

앞으로 궁극적으로는 주식투자 전문가 시스템(ASIE)에서 실용화 될 수 있도록 사회 여건(국내/국외 동향), 기술적 분석에 필요한 제반 지식과 ASIE와의 인터페이스 모듈을 구축하여 좀 더 실용성 있는 시스템으로의 확장이 필요하다.

Ⅵ. 참고 문헌

[1] 김 성기, "자연 한글 질의어 처리를 위한 인터페이스의 설계및 구현", 서울 대학교 대학원 전자계산기 공학과 석사학위 논문, 1985.

[2] 미 승우, "새 맞춤법과 교정의 실제", 어문각, 1988

[3] 윤 덕호, "한국어 질의응답 시스템의 설계및 구현에 관한 연구," 서울 대학교 대학원 계산통계학과 석사학위 논문, 1987.

[4] 이 신영, "정형 질의어로부터 한국어의 생성," 서울 대학교 대학원 전자계산기 공학과 석사학위 논문, 1987.

[5] 이 찬영, "자연 질의어 인터페이스를 위한 지식의 습득 및 관리," 서울 대학교 대학원 전자계산기 공학과 석사학위 논문, 1987.

[6] 진 청희, "자연언어 모호성으로 인한 다중출력 나무구조의 최적 선택에 관한 연구," 서울 대학교 대학원 전자계산기 공학과 석사학위 논문, 1988.

[7] 조 규빈, "하이라이트 고교 문법," 지학사, 1989.

[8] 한 광록, 이 주근, "한국어 문장으로부터 개념단위의 추출과 지식베이스의 구축," 한글 및 한국어 정보처리, 한글날 기념 학술대회 발표논문집, 1989.

[9] Carl Pollard, Ivan A. Sag, "Information-Based Syntax And Semantics," vol 1 FUNDAMETALS, CSLI, 1987, pp. 81-113

[10] F.C.N. Pereira, S.M. Shieber, "Prolog and Natural-Language Analysis," CSLI, 1987.

[11] James Allen, "Natural Language Understanding," The Benjamin/CummingsPub, Inc. 1987, pp. 192-310.

[12] Margaret King, "Parsing Natural Language," Academic Press, 1983, pp. 197-pp. 246

[13] P. Velardi, M.T. Pazienza, M.D. Giovanetti, "Conceptual graphs for the analysis and generation of sentences," IBM J. RES. DEVELOP. VOL. 32 NO. 2 March 1988

[14] "The *COURTIER* System," Cognitive Systems Inc.