자연언어처리의 이론과 실제

최 기 선

1. 서 론

컴퓨터 프그로램은 데이타와 그 처리(혹은 제어)로 이루어진다. 이를 합하여 계산(computation)이라고 한다. 프로그램의 목적은, 어느 입력을 받아 계산을 하여 출력을 내는 것이다. 전산학의 모든 분야가 그렇듯이, 자연언어처리도 궁극적으로 컴퓨터 프로그램으로 구현됨을 목적으로 한다. 즉, 자연언어처리의 가시적 구현은 프로그램이다. 여기서, 우선 자연언어처리의 입력은 무엇이고, 출력은 무엇이 되어야 하는지를 밝혀야 한다. 또, 자연언어처리의 계산, 즉 데이타와 그 처리란 무엇을 뜻하는 것인가를 고려해 볼 필요가 있다. 2절에서는 위의 상황을 고려하여, 자연언어처리분야의 정의를 내리고자 한다.

직관적으로, 자연언어처리의 데이타라 함은, 자연언어처리를 위한 지식을 뜻한다. 처리라 함은, 계산의 궁극적 목적인 입력으로부터 출력을 만들어내는 과정의 기술이며, 그 처리 과정에서 데이타(즉, 자연언어처리를 위한 지식)를 참고로 한다.

그러면, 자연언어처리를 위한 지식에는 어떠한 것이 있는가? 언어 자체의 지식, 실세계에 대한 지식 등이 있음은 자명하다. 처리란 무엇인가? 입력이 자연언어 텍스트이고 출력은 그에 대한 어떤 의미구조라면, 처리는 파싱(parsing) 알고리즘에 해당한다.

언어 자체의 지식이 자연언어처리에는 필요불가결한 데이타가 되므로, 언어학 또는 계산언어학, 도서관 등에서 정리한 데이타가 자연언어처리에 쓰인다. 문제는 이러한 데이타가 그대로 쓰여질 수 있는가, 또 언어학의 언어모델이 그대로 자연언어처리에 적용될 수 있는가를 숙고할 필요가 있다. 한편, 처리는 컴퓨터에 의한 처리를 뜻하나, 인간의 언어처리와 어떠한 관계를 맺어야 하는가를 고려하고자 한다. 인간의 언어처리는 인지심리학 등에서 다루고 있다. 다시 말하여, 자연언어처리의 인접분야로서는 전산학은 물론, 언어학, 인지심리학 등을 들 수 있다. 자연언어처리 분야와 인접분야를 비교함으로써, 자연언어처리의 정의에서 언급한 내용을 재확인하고, 이 분야의 연구목표를 정확히 인식할 수 있다. 이러한 내용을 2.1절에서 논한다.

그러나, 자연언어처리의 정의의 한도 내에서, 자연언어처리의 분류는 크게 둘로 나눌 수 있다. 첫째는, 언어학의 모델의 적극적 수용, 또는 심리학적 처리방법의 프로그램화에 역점을 두고자 하여 인접분야의 이론을 자연언어처리의 이론으로 적극 수용하려 하는 연구이다. 반면, 그와 반대로, 응용에 맞는 프로그램에 역점을 두어 입력에 맞는 출력만을 내려 하는 연구개발을 들 수 있다. 이러한 자연언어처리의 극단적 분류를 2.2절에서 언급한다.

후자의 응용적 관점에 역점을 두어, 우리의 당면한 현실에서 자연언어처리에서 해야 할 실제적 과제를 3절에서 고려한다. 특히, 한국어 처리의 문제점을 예를 들어 4절에서 생각하여 본다. 마지막으로 5절에서 자연언어처리의 이론에 대하여 언급한다.

2. 자연언어처리의 정의

자연언어처리의 정의를 내리자면 다음과 같다.

> '자연언어'로 '의사소통'을 하기 위한 '계산효율성' 있는 장치의 형성과 탐구.

여기서 '자연언어'라 함은, 인공언어에 반하는 용어로서, 한국어, 영어와 같은 일상생활에서 사용하는 언어를 뜻한다. '의사소통'이라 함은, 인간과 기계 간의 의사소통(communication)을 의미한다. 의사소통의 방법으로는 대화, 번역 등을 열거할 수 있다. 대화라 하더라도 질문과 응답이라는 일방적 대화 만을 고려한다면, 질문을 자연언어로 하고 응답은 컴퓨터 데이타베이스의 내용을 출력하여 주는 상황을 생각할 수 있다. '계산효율성' 있는 장치(mechanism)라 함은, 추상적인 모델만을 생각하는 것이 아니라, 실제적인 컴퓨터 프로그램으로 실현가능한 장치를 의미한다.

다음 절에서는, 자연언어처리의 인접분야를 살펴 봄으로써, 자연언어처리가 인접분야인 언어학이나 심리학으로부터 무엇을 얻을 수 있고, 무엇을 줄 수 있는가를 고려한다.

2.1. 자연언어처리의 인접분야

이 절에서는, 언어학과 인지심리학 분야의 연구초점, 목표와 앞에서 언급한 자연언어처리의 정의와의 대비를 시도한다.

2.1.1. 언어학과 자연언어처리

언어학은 연구초점을 자연언어의 형식적, 일반적, 구조적 모델을 구축하는 데 둔다. 따라서, 그 목표는 언어의 형식 모델을 세우는 데 있으며, 그 모

델은 언어의 규칙성(regularity)를 포착하기 위한 것이어야 하며, 모든 언어의 보편성을 설명할 수 있는 모델이어야 한다. 따라서, 언어학의 관점은 언어를 처리하기 위한 장치와는 거리가 있다는 것을 알 수 있다.

그러나, 최신 언어학의 경향은, 전산학의 영향을 받아, '單一化(unification) 문법'의 경향을 띄고 있다. 언어학의 모델이 바로 자연언어처리의 모델이 되게 하고자 하는 경향이다.

2.1.2. 인지심리학과 자연언어처리

인지심리학은 연구의 목표를, 언어학과 같은 언어구조의 모델이 아니라 심리학적 타당성 있는 언어처리의 모델에 둔다. 심리학적 타당성 있다는 말은, 인간과 같은 방법을 의미한다. 그러나, 인간의 처리방법이 컴퓨터의 처리방법과는 다르므로, 인지심리학이 처리 모델을 추구한다 하더라도 '계산적 효율성'에는 중점을 두지 않는다.

따라서, 그 초점이 자연언어처리 그 자체에는 없고, 인간의 인지와 기억구조의 일반적 측면을 다루는 데에 있다.

2.2. 자연언어처리의 분류

자연언어처리 분야를 극단적으로 나누면 '일반' 자연언어처리와 '응용' 자연언어처리로 다음과 같이 나눌 수 있다.

자연언어의 인접분야인 언어학이나 인지심리학의 방법론을 적극 수용하는 반면, 자연언어처리의 특징인 계산효율성을 고려하는 쪽을, '일반자연언어처리'라 한다. 반면에, 언어학의 방법론이나 인지심리학의 모델과는 관계없이, 입력과 출력과의 관계만을 생각하는 쪽을 '응용자연언어처리'라 한다.

2.2.1. 일반자연언어처리

일반자연언어처리는 인지심리학의 이론에 강한 영향을 받으므로, 그 연구목표는 인간의 언어사용에 대한 모델을 추구하지만, 자연언어처리의 본질적 문제인 계산효율성 측면을 생각한다. 그 연구대상으로서는 일반적 스토리 이해나 대화 모델링을 들 수 있다. 전자의 경우로는 Northwestern 대학교의 Schank 교수팀이 유명하다.

그러나, 이 연구의 문제점으로서는, 현실적으로 방대한 양의 실세계 지식이 필요하다는 데에 있다. 각 개인이 갖고 있는 도메인에 따른 지식의 갯수를 천 개라 할 때, 과연 그것을 모두 컴퓨터에 가공하여 넣을 수 있는 가에 대한 근본적 문제가 대두된다. 또, 모든 도메인의 지식을 같은 지식표현으로 나타낼 수 있는가에 관한 문제도 도사리고 있다. 한 가지의 전문적 도

메인에 대한 전문가 시스템을 꾸미는데 걸리는 시간이 반년 내지 2년, 길게는 5년이 걸린다는 실질적 경험으로 미루어 보아, 천 개의 도메인에 대한 지식의 가공에는 어느 정도의 시간이 걸릴지를 미루어 상상할 수 있다.

따라서, 연구의 초점이 실세계 지식의 표현방법에 두어지며, 그 실험은 자연언어 입력을 이해하기 위한 응용시스템을 만드는 데 맞추어진다. 이러한 시스템은 실용적 시스템과는 거리가 있는 장난감 시스템에 불과하다. 그 프로그램의 의의는, 개념, 방법론의 타당성을 보여주기 위한 시스템이며, 그 입력은 주의깊게 선택된 몇 개의 예에 대하여만 작동한다. 예를 들면, 식당에서의 일, 이혼 성립에 대한 도메인, 새로 닥친 일의 해결 방법 등을 열거할 수 있다. 결론적으로, 일반자연언어처리 분야에서는 아직 실용적 시스템을 구축하기에는 전체적 인공지능의 수준이 미흡한 편이다.

2.2.2. 응용자연언어처리

응용자연언어처리의 목표는 인지심리학의 이론을 반영하는 것이 아니므로, 인지의 시뮬레이션에는 관심이 없고, 인간이 자연언어로 기계와 의사만 소통하도록 프로그램을 만드는 데에 있다. 따라서, 실용적인 면이 강조되어, 기계의 자연언어입력의 이해과정에서의 '인지적 타당성'은 고려를 하지 않는다. 프로그램 구성요소 중 입력과 출력만을 중시하고 그 과정은 어떻게 되든 상관이 없으므로, 자연언어 입력에 대한 반응에 중점을 둔다. 결국, 이 분야의 시스템의 평가는, 출력된 반응이 사용자에게 도움이 되는가, 또는 사용자의 입력 의도에 상응하는 출력을 내고 있는가에 둔다.

응용분야로서는 데이타베이스에 대한 인터페이스, 전문가 시스템을 위한 인터페이스, 기계번역시스템이 이에 속한다.

실용적 시스템을 구축하기 위한 것인 만큼, 자연언어처리의 입력방법이 선결되지 않으면 안된다. 사실 자연언어를 키보드에서 직접 입력한다는 일은, 컴퓨터 명령어나 메뉴를 선택하는 것보다는, 사용자가 입력해야 하는 글자의 갯수가 상당히 늘어날 것이므로, 그 입력과정에서 오류가 일어날 가능성도 그만큼 높다. 따라서, 사용자 실수의 파악과 그 해결을 제시해주는 preprocessor가 필요하며 이러한 것을 robust 시스템이라고 부른다.

3. 자연언어처리의 응용과제

실용적 시스템 중 자연언어처리가 응용되어야 할 곳은 다음과 같다.

○고급 워드프로세서
○사용자 인터페이스

○기계번역

3.1. 고급 워드프로세서

고급 워드프로세서로서 갖추어야 할 기능은 무엇인가? 다시 바꾸어 말하면, 현존하는 한글 워드프로세서가 영어 워드프로세서보다 못하고 있는 기능이 무엇인가를 열거하면 다음과 같다.

○한글 자동 철자, 맞춤법 검사
○한글 자동 띄어쓰기 검사
○한글 KWIC(Key Word In Context), 즉 indexing

3.2. 사용자 인터페이스

사용자 인터페이스의 문제는, 자연언어의 제한도와 입력방법에 대한 두 가지 문제로 나누어진다. 자연언어의 제한도의 문제로서는, 기존의 최상의 입력방법으로는 필요한 정보를 단말기에 나타난 메뉴에 의하여 선택하게 하는 것이다. 자연언어를 메뉴에 대신하는 입력으로 채용하려면, 응용 도메인에 따른 그 타당성이 먼저 입증되어야 할 것이다. 또는 메뉴와 자연언어입력이 공존하는 시스템을 생각할 수도 있다. 예를 들어, 한국인이 쓰고 있는 모든 한국어를 입력으로 허용하는 시스템을 구축한다는 것은 무리가 있으므로, 메뉴에 의하여 제한된 한국어를 입력수단으로 사용자를 유도하도록 하는 방법도 생각할 수 있다.

입력문제로서는 앞 절에서 지적한 바와 같이, 키보드에서의 사용자의 실수가 예상되므로, 그 실수를 복구하여 주는 기능이 필요하며, 한걸음 더 나아가서, 음성 및 문자 인식, 또는 그 생성과의 결부가 이루어진다면 더욱 편리한 사용자 인터페이스가 만들어질 수 있다.

3.3. 기계번역

기계번역에는 수십만 단어의 사전을 운용하여야 한다. 우선 그러한 사전을 편집하고 관리하여 주는 사전편집기가 마련되어야 한다.

그 사전편집기와 함께, 사전의 속성값의 일관성을 관리하기 위한 시스템이 필요하다. 만일 어떤 사람이 속성명으로서 ANIMATE를 쓰고, 다른 사람은 ANIM과 같이 약자를 쓰고 있다면, 컴퓨터가 그 두 가지를 같은 것이라고 여기기에는 상당한 노력이 필요할 것이다. 혼자 사전 속성값의 입력을 한다고 하더라도 혼동할 염려는 얼마든지 일어날 수 있다.

어느 속성이 필요하고, 그 속성이 어떤 값을 가져야만 하는가에 대한 조사도 응용분야에 따라 이루어져야 하며, 혹은 모든 텍스트에서 필요로 하는

것이 무엇인가를 조사하여야 한다. 그 대표적인 예로서, 각 용언의 subcategorization, 다시 말하면, 입력문의 분석과정에서, 격인식에 필요한 정보는 무엇인가, 표층격조사와 심층격과의 대응에 필요한 정보는 무엇인가, 각 용언에 필수적인 필수격과 그렇지 않은 임의격은 어떤 기준에 의하여 분류하여야 하는가와 같은 문제가 있다.

3.4. 문제의 복잡도의 비교

이 절에서는, 위에서 열거한 세 가지 응용분야 : 고급 워드프로세서, 기계번역, 사용자 인터페이스의 문제의 복잡도를 비교하여 보고자 한다. 그 비교는 실증적인 것은 아니나, 도메인과 사용자의 불특정, 특정에 따른 표면적인 비교를 하여 그림 1에 보인다.

	고급 워드프로세서	기계번역	사용자 인터페이스
도메인	불특정	특정	특정
사용자	불특정	불특정/특정	불특정/특정

그림 1. 응용시스템의 도메인, 사용자의 다양성의 비교

워드프로세서는 어린이로부터 노인까지 모든 계층에서 쓰여지도록 만들어진 것이다. 따라서 사용자는 불특정이다. 또, 쓰여지는 텍스트의 내용도 일정치 않으며, 모든 분야에 걸쳐 있으므로, 그 도메인도 불특정이다.

반면에, 기계번역이나 사용자 인터페이스는 그 쓰여지는 텍스트의 도메인이 특정화되어 있다. 예를 들어, 컴퓨터 메뉴얼의 번역을 위한 기계번역 시스템이나, 주식정보를 얻기 위한 사용자 인터페이스의 경우, 각각 쓰여지는 도메인은 컴퓨터분야, 주식분야로 특정화되어 있다. 사용자는 대부분 그 도메인에 익숙한 사람이 쓰지만, 그렇지 않은 경우도 있다고 생각하여, 불특정/특정이라고 하였다.

그림 1과 같은 분류를 통하여 볼 때, 고급 워드프로세서의 경우가 도메인이나 사용자가 모두 불특정으로 되어 있어서, 가장 어려운 과제로 보인다. 사실 철자법 교정 시스템을 만들 경우, 사전에 모든 도메인에 걸친 단어를 모두 등록하거나, 철자법 교정을 위한 추론 정보가 파악이 되어 있어야 한다. 이러한 것은 기계번역의 초기 단계에서도 필요한 단계이다. 띄어쓰기 교정 시스템의 경우, 기계번역에서는 대개 입력텍스트가 올바른 띄어쓰기를 하였다고 가정하고 있고, 사용자 인터페이스에서는 능동적으로 잘못된 띄어쓰기를 극복한다는 것을 가정하고 있지만, 워드프로세서에서의 띄어쓰기 교정 시스템은 한 차원 높은 처리를 하여야 한다. 이와 같은 사실로 미루어 볼

때, 왜 아직까지 한글 워드프로세서에 철자법 교정, 띄어쓰기 교정 등의 기능이 결여되어 있나를 알 수 있다. 자동 인덱스를 만들어 주는 기능도 위의 기능이 선결된다면 따라서 구현될 수 있는 것이다.

4. 한국어처리의 과제

자연언어이해 시스템의 흐름을 형태소해석, 구문해석, 의미해석으로 대체로 나누고 있다. 그것이 순차적으로 이루어지는가, 혹은 서로 호응하여 이루어지는가는 각각의 장단점이 있다. 물론 인간의 언어이해는 위 삼단계의 구분없이 이루어질 것이나, 컴퓨터 응용시스템을 위한 프로그램을 위하여는 순차적으로 하는 것이 더 간단할 것이기 때문이다.

4.1. 형태소해석

4.1.1. 단어식별의 의미해석 단계의 필요성

우선 형태소해석을 살펴보자. 이 단계의 목표는 입력문을 단어 단위(혹은 형태소 단위)로 쪼개는 것이다. 그것이 사전 혹은 다른 테이블에 등록되어 있는 형태소 정보만으로 가능하다면 매우 좋을 것이다. 왜냐하면 컴퓨터에서의 프로그램이 훨씬 쉬어질 것이기 때문이다. 영어의 경우, 이것이 어느 정도 가능해 보인다. 언제나 단어와 단어 사이에 빈 칸이 있기 때문이다. 그러나, 한국어의 경우, 명사+조사, 혹은 용언+어미 등이 부착된 단위로 띄어쓰기를 하므로, 형태소 정보만으로는 해결하기 어려운 예는 얼마든지 있다. 다음의 예를 보자

(1) '감기는' :
  ㄱ. 감기는 걸렸지만 견딜만 하다. (감기+는 : 명사+조사)
  ㄴ. 머리를 감기는 했지만 아직도 가렵다. (감+기+는 : 동사어간+명사형 전성어미+조사)
  ㄷ. 실을 감기는 했는데 또 헝클어졌다. (감+기+는 : 동사어간+명사형 전성어미+조사)

(2) '차는' :
  ㄱ. 차는 설록차가 좋다. (茶+는 : 명사+조사)
  ㄴ. 차는 벤쯔가 좋다. (車+는 : 명사+조사)
  ㄷ. (던지는 볼이 아니라) 차는 볼이 좋다. (차+는 : 동사어간+관형형어미)

형태소 단계에서는, (1ㄱ)과 (1ㄴ, ㄷ)의 구별, (2ㄱ, ㄴ)과 (2ㄷ)의 구별은 하여야 한다. 그러나, 형태소 정보만으로는 파악할 수 없다는 것을 알

수 있다.

4.1.2. 단어의 사전 등록

사전에 등록할 단어의 단위는 어떻게 해야 할 것인가? 한영기계번역의 예를 든다면, 다음과 같은 대역사전이 필요할 것이다.

'—에 대하여' ⇒ 'for'
'—고 싶다' ⇒ 'want to'
'—고 있다' ⇒ 'ing'
'—ㄹ 수 있다' ⇒ 'can'

사전의 등록어는 용언의 경우, 그 원형도 중요하지만, '—에 대하여'의 경우 처럼 조사 상당어구의 등록이 더욱 절실하다. 사전에 '대하여'의 원형 '대하다'만 있어서는 위와 같은 대역은 시도조차 하지 못할 것이다.

4.2. 구문 의미해석

4.2.1. 용언의 하위범주

구문의미해석에서 필수적인 요소는 용언의 격패턴, 혹은 하위범주화이다. 이는, 패턴의 형태로 테이블화하여 표현할 수 있다. 해석의 과정에서, 표층격조사로부터 심층격(혹은 θ-role)으로의 해석을 내려주는 역할을 한다. 우선 한국어 용언에 나타나는 표층격조사를 대표하는 조사는 어느 것이 있는가를 조사하고, 한국어에는 어떠한 심층격이 필요한가에 대하여 실제의 방대한 텍스트로부터 조사가 이루어져야 할 것이다.

4.2.2. 임의격의 인식

임의격(혹은 자유격 : optional case)의 경우, 용언에 의존하지 않고, 앞뒤의 단어에 의하여 그 의미가 결정된다. 필수격과 마찬가지로, 표층격조사로부터 심층격을 얻어내기 위한 휴리스틱(heuristic)한 규칙을 많은 텍스트 데이타로부터 조사하여 만들어내야 할 것이다. 임의격의 한 예를 들면, '—으로'에 대하여 다음과 같이 한국어와 영어를 대조할 수 있다.

인플레로 땅값이 올라가다. (원인 : due to)
공인으로 말하다. (자격 : as)
시속 100km/h로 달리다. (상태 : at)
앞으로 전진하다. (목표, 방향 : toward)
회의가 3시로 빨라지다. (목표, 시간 : at)

위의 예에서 보는 바와 같이, 기계번역의 경우에도 심층격의 파악 없이는 번역이 될 수 없다는 것을 알 수 있다. 우선 한국어의 경우에도 위와 같은 방대한 텍스트 데이타로부터 분류작업이 선결되어야 하겠다.

4.3. 생성단계 : 수량표현의 번역

기계번역에서 해석을 한 다음, 적당한 중간구조로부터 목표언어로 생성을 하게 된다. 한영 기계번역의 경우, 그 도메인이 기술관계문서일 경우, 상당한 분량의 문장이 수량표현을 포함하고 있다. 그 한 예를 보자.

(3) ㄱ. 그것의 길이는 3cm이다.
ㄴ. It is 3cm long.
ㄷ. ?Its length is 3cm.

(3ㄱ)의 한국어문에서 영어로의 번역문이 (ㄷ)이 아니라, (ㄴ)으로 번역되어야 한다는 것은 생성단계 중, 어순 결정 단계에서 배열 정보를 요구한다. 이러한 수량표현의 대역도 어디까지나 이론적인 것인 아니라, 기술문서의 실질적 조사에 의하여 파악되어야 한다.

5. 자연언어처리의 이론

끝으로 자연언어처리 분야의 이론에 대하여 언급하고자 한다. 앞에서 논한 바와 같이, 자연언어처리는 지식과 그 처리로 양분할 수 있다.

해석을 위한 처리 방법으로는, 기본적인 방법으로서 Cock-Kasami-Younger 알고리즘(Aho & Ullman 1972), Earley 알고리즘(Earley 1970), chart 방법(Kay 1973, Thompson 1983) 등이 있다. 좀더 응용적 방법으로서는, 확장 LR 법(Tomita 1987), PARSIFAL(Marcus 1980) 등이 열거된다. 그 외에 혼합적 방법으로서, LINGOL(Pratt 1973), ATN (Woods 1970)이 있다. 논리해석에 대하여는, DCG(Pereira & Warren 1980), GPSG(Gazdar, et al. 1985) 등이 있으며, 최근 발전하고 있는 단일화문법(unification analysis)으로서는 언어학적 모델을 거의 그대로 채용하여, LFG(Bresnan 1983), PATR-II (Shieber 1986), CUG(Uszkoreit 1986), HPSG(Pollard & Sag 1987)가 연구되고 있다. 이 외에 극히 지식을 바탕으로 하는 방법으로서는, 어휘해석(lexical-based analysis)으로서, 격해석(case analysis)(Shimizu, et al. 1987)이나, lexicase analysis(extended dependency analysis)(Starosta & Nomura 1986)를 열거할 수 있다.

이 중에서, 가장 대표적으로 실용적 시스템에 이용되고 있는 것은 chart 파싱으로서, 어느 응용에라도 쓰여진다. 그러나, ATN은 영어에는 많이 적

용되어 왔지만, 우리나라와 같은 어순을 가진 언어에는 잘 맞지 않는다는 보고가 있다. 최근에는 LFG나 HPSG와 같은 단일화문법이 많이 연구되고 있다.

6. 결 론

실용적 시스템을 위하여는, 언어모델이나 파싱방법보다는, 어느 정도 충실한 지식이 마련되어 있는가가 성공의 관건이라고 생각한다. 아무리 처리방법이 좋아도, 방대한 양의 언어지식이 파악이 되어 있지 않다면 장난감이 될 수 밖에 없다.

자연언어처리 연구자는 언어지식의 조사와 정리를 위한 도구적 시스템을 만들어 보급하여야 하며, 언어학계에서는 기초적 언어지식의 정리, 특히 한국어, 또는 한국어와 영어 등의 대조분석에 적극적으로 참여하여야 하겠다. 실용적 자연언어처리 시스템이 구축될 수 있도록 전산학분야, 언어학분야의 연구자는 깊이 협조하여 충실한 공동연구가 되도록 힘써야 할 것이라고 생각한다.

앞으로 10년후에는, 세계에 대한 수출품은 한국어와 관련된 소프트웨어가 크게 늘어날 것이며, 더구나 한국어의 내실화, 세계화를 위하여, 우리나라의 정보산업국으로의 변신을 위하여서도 한글정보처리에 대한 작업은 서둘러야 한다.

참 고 문 헌

Aho, A.V. and J.D. Ullman (1972) *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing*, Prentice-Hall.

Bresnan, J. (1983) *The Mental Representation of Grammatical Relations*, MIT Press.

Earley, J. (1970) 'An Efficient Context-Free Parsing Algorithm,' *Commun. ACM*, 13, 2.

Gazdar, G., E. Klein, G.K. Pullum & I.A. Sag (1985) *Generalized Phrase Structure Grammar*, Basil Blackwell.

Kay, M. (1973) 'The MIND System,' in R. Rustin, ed., *Natural Language Processing*, Algorithmics Press.

Marcus, M.P. (1980) *A Theory of Syntactic Recognition for Natural Language*, MIT Press.

Pereira, F.C.N. & D.H.D. Warren (1980) 'Definite Clause Grammar for Language Analysis—A Survey of the Formalism and a Comparison with Augmented Transition Networks,' *Artificial Intelligence* 13.

Pollard, C. & I.A. Sag (1987) *Information Based Syntax and Semantics, Vol. 1 Fundamentals*, CSLI Lecture Notes, 13.

Pratt, V.R. (1973) 'A Linguistics Oriented Programming Language,' *Proc. of the International Joint Conference on Artificial Intelligence.*

Shieber, S.M. (1986) *An Introduction to Unification-Based Approaches to Grammar*, CSLI Lecture Notes, 4.

Shimizu, A., S. Naito & H. Nomura (1987) 'Semantic Structure Analysis of Japanese Noun Phrases with Adnominal Particles,' *Proc. of the 1987 Conference of Association for Computational Linguistics.*

Starosta, S. & H. Nomura (1986) 'Lexicase Parsing: A Lexicon-driven Approach to Syntactic Analysis,' *Proc. of the International Conference on Computational Linguistics.*

Thompson, H.S. (1983) 'MCHART: A Flexible, Modular Chart Parsing System,' *Proc. of the Third Annual Meeting of the American Association for Artificial Intelligence.*

Tomita, M. (1987) 'An Efficient Augmented-Context-Free Parsing Algorithm,' *Computational Linguistics* 13, 1-2.

Uszkoreit, H. (1986) 'Categorial Unification Grammars,' *Proc. of International Conference on Computational Linguistics.*

Woods, W.A. (1970) 'Transition Network Grammars for Natural Language Analysis,' *Commun. ACM*, 13, 10.

ABSTRACT

Practice and Theory of Natural Language Processing

Key-Sun Choi

Natural language processing (NLP) is defined as 'formation and investigation of computationally effective mechanism for communicating by natural languages.' Practice and theory of NLP is projected by following

this definition. The goal of NLP is clarified by comparing with its related fields such as linguistics and cognitive psychology. The spectrum of research fields in NLP is broad from scientific to engineering-oriented.

Focusing at Korean language processing, it is claimed that several basic problems should be solved, for example, Korean dictionary and bilingual dictionaries for practical use of engineering. Three R&D projects are discussed: Korean intelligent wordprocessing, machine translation, natural language in user interface, for they should be done in the near future. Their problems and complexities are discussed.

Next, theoretical trends of computational linguistics are provided. In conclusion, it is claimed that linguistic knowledge should be studied, re-arranged, and integrated in the new engineering viewpoints for processing Korean language. Such basic R&D is believed that it really bridges the gap between practices and theories of NLP.

130-650
서울시 동대문구 청량사서함 150호
한국과학기술원 전산학과