특집

한글과 한국어 처리 문제 및 동향

박 동 인
시스템공학연구소

I. 개 요

'언어는 사고의 옷이다'라는 말이 있다. 언어는 인간의 사고 과정과 그 결과를 외형적으로 표출하여 전달할 수 있는 인간 고유의 상징적인 도구이다. 따라서 언어라는 상징체계가 없었다면 인간 문화의 발생과 교류 및 계승 발전이라는 것이 매우 어려웠을 것이다. 우리는 우리 민족 고유의 자산인 사고의 옷, 즉 한글과 한국어를 가지고 있으며, 한글의 과학적 우수성이 세계 최고 수준임을 자랑스럽게 생각하고 있다. 하지만 우리는 그 우수한 우리말을 정보화 시대에 걸맞게 계승 발전 시키고 우월성을 보여주고 있는가? 본 고에서는 이러한 질문에 대해서 우리는 현재 어디까지 와 있고, 어떤 문제점들이 있으며 앞으로 어떻게 이러한 문제점들을 헤쳐나가야 할 것인가를 고찰해 보는 계기로 삼고자 한다.

한글과 한국어 정보처리 분야는 다른 정보처리 분야와 마찬가지로 그동안 많은 연구가 진행되었으며, 여러 가지 방법론과 기술들이 개발되어 왔다. 그동안 주로 개발되어 활발하게 활용되고 있는 기술들은 한글 코드와 자판, 한글 폰트, 한국어 형태소 분석 분야 등 한글 정보처리를 위한 기초 기술 위주였고 최근에 와서 첨단의 기술을 응용한 제품들이 출시되고 있다. 혹자는 워드프로세서 등에서 한글을 편집하고 입출력할 수 있으면 컴퓨터에서 한글 정보화는 끝난 것이 아니냐라고 반문하는 사람도 있을 수 있다. 그러나 한글과 한국어 정보처리 기술이 추구하는 영역은 이에 머무르지 않고 인간이 언어를 통하여 할 수 있는 모든 지적인 행위와 이를 위한 요소 기술들이 언어처리 응용 분야의 작업을 지원하는 것이다. 예를 들면, 문장이나 글을 보고 들으며 내용을 이해하고, 문장을 생성하고, 문장의 오류를 발견하고 수정하며, 원하는 정보를 자연언어로 요구하여 인터넷이나 데이터베이스에서 획득하고, 외국어로 된 문서를 한글로 번역하고, 어떤 영역의 주제에 대해서 사람과 대화하는 일들을 수행하는 것들을 들 수 있다.

그러나 이와 같은 모든 능력을 완전히 컴퓨터에

게 부여한다는 것은 매우 어려운 도전이어서 해결해야할 난제들이 무수히 쌓여 있다. 이러한 문제를 단시일에 해결할 수 없기 때문에 그야 말로 백년대계를 세워서 기초부터 열심히 다진다는 자세가 필요하다. 자연언어처리의 연구는 컴퓨터 과학의 다른 영역의 학문과 접근 방법이 사뭇 다르다. 자연언어처리에서는 방대한 양의 어휘와 그 어휘들이 가지는 현상에 대한 정보, 즉 사전적인 어휘정보들이 중요하므로, 알고리즘 뿐만 아니라 방대한 양의 정확한 언어 데이터의 구축이 무엇보다도 중요하게 취급되는 분야이다. 그러므로, 어떤 천재가 매우 우수한 알고리즘을 개발하여 단시일내에 문제가 해결될 수 있는 성질의 것이 아니라고 여겨진다.

이제 한글과 한국어 정보처리에 관련된 세부영역별로 고찰해보기로 한다. 이 특집의 다른 주제분야에서도 각 세부 응용 및 요소 기술별로 자세히 다루고 있으므로 여기서는 다른 분야와 관련된 공통으로 풀어야 할 미래 지향적인 문제들을 중심으로 언급하도록 하겠다.

II. 국어 정보처리 기반기술

한글 및 한국어에 대한 연구는 대상언어의 유형에 따라 세분류 될 수 있으며, 분류에 따라 접근 방법과 처리 기술이 다르다. 시대적 배경에 따라 중세, 근대, 현대어로 나눌 수 있으며, 표준어, 방언, 속어, 은어, 유행어 등으로도 나눌 수 있을 것이다. 문장 유형에 따라 구어체, 문어체, 또는 대화체, 서술체 등으로 분류할 수 있다. 언어데이타의 양상(modality)에 따라서 음성, 필기(인쇄체), 텍스트, 수화(제스춰) 인식으로 나눌 수 있다. 그리고 언어처리의 다루는 측면에 따라서 코드, 폰트, 문서교환 양식의 표준화와 같은 외형적 언어 처리 기술과 형태소 분석, 구문-의미 분석, 문장 생성 등과 같은 내용적 언어 처리 기술로 분류할 수도 있는데, 이러한 기술들은 각각 기술의 성숙도가 다르다. 초창기에는 외국의 언어처리 방법론이나 기술을 그대로 한국어에 적용해보는 시도들이 많이 있었으나, 최근에는 한국어의 특성을 반영하여 확장 또는 수정한 방법들이 많이 대두되고 있다.

과거의 언어처리 연구가 주로 제한된 어휘 수와 교과서적인 단순문 위주의 처리를 하는 프로토타입 시스템 개발에 치중했던 반면 현재는 대량의 어휘와 복잡한 문형을 다루는 실용적 시스템 개발을 목표로 연구가 진일보 발전하여 진행되고 있다. 그리고, 선험적 언어 직관에 기초한 규칙기반의 언어처리 기법에서, 대량의 언어데이터의 관찰에 기반한 코퍼스(말뭉치) 기반 언어처리기법(통계적 기법 포함)으로 언어처리 방법의 중심이 옮겨 가고 있다. 대규모 실용적인 언어처리 시스템을 구현하기 위해서는 대규모의 기초 언어 데이터인 코퍼스와 이에 근거하여 추출한 언어지식인 전자사전 또는 구문 규칙을 구축하고 이를 활용하는 언어처리 도구들을 개발하여야 한다. 이러한 언어처리 도구들은 다양한 언어처리 응용 시스템에서 공통적으로 활용할 수 있는 요소기술로서 이를 집중적으로 개발, 보급하여 중복 개발을 막고 주요 기술 개발을 가속화하여 할 것이다. 효율적이고 범용적인 한국어 정보처리를 위하여 필요한 기술들을 살펴보면, 코퍼스나 전자사전과 같은 기초정보베이스에 관련된 기술과 응용시스템을 만들기 위한 기본적 처리 도구인 언어처리도구 개발 기술로 나눌 수 있다.

1. 코퍼스(corpus)

최근 통계적 기법을 기반으로 한 언어처리에서 필요로 하는 대규모 말뭉치 구축 활동이 전 세계적으로 새로운 언어처리 방법으로 유행처럼 진행되고 있다. 또한, 언어현상적으로 균형잡힌 분포를 갖는 코퍼스(balanced corpus)를 구축하는 기법도 개발되고 있다. 영어를 중심으로 한 구미언어에 대해서는 여러 연구기관에서 이미 대규모 코퍼스를 구축한 바가 있고, 현재도 더 큰 규모의 다양한 영역의 말뭉치를 계속하여 구축하고 있다. 국내에서도 몇몇 학교와 연구소들을 중심으로 자체 연구용으로 구축 사용 중이나 그 규모는 비교적 적은 편이고, 아직 일반에게 공개한 경우는 별로 없기 때

문에 공유가 잘 되지 않고 이를 위한 표준화도 잘 이루어지지 않고 있다. 또한, 원시 말뭉치의 경우에는 원문 저자의 저작권법 문제가 있어서 일반에의 배포가 어려운 점도 있다[2].

2. 전자사전

한국어 정보처리를 위한 언어처리용 전자사전은 일반 사용자가 읽어 보고 단어의 용법과 뜻을 이해하기 쉽게 만든 일반 편찬 사전을 단순히 전자화한 것과는 그 내용과 구조가 판이하게 다르다. 일반 편찬 사전의 정보를 언어처리기에 사용하기 위해서는 여러 언어처리기의 사용 목적에 맞게 각종 언어 정보가 보다 일관되고, 구체적이고 효율적으로 표현되어야 한다. 한국어 각 어휘의 음운, 형태, 구문적 정보를 담고 있는 어휘정보 사전, 어휘의 개념적 정보를 담고 있는 개념사전, 각 전문 분야별 용어사전, 시사용어 사전, 관용어, 정형패턴, 연어(collocation) 사전, 동의어와 반의어 사전 등 다양한 종류와 기능의 사전이 요구된다. 현재 국내 각 연구팀들은 자체적으로 개발한 사전을 가지고 연구를 진행하고 있으며, 누구나 이용할 수 있는 공개된 사전이 아직 없기 때문에, 연구기관 간에 전자사전의 공유가 잘 되지 않고 있다. 전자 사전의 공유를 위해서는 품사 태그 세트 및 표현의 저장 방법에 대한 표준이 먼저 정립되어야 한다. 이러한 공유를 위한 사전의 표준적 표현 양식에 관한 문제에 대하여 최근에 KAIST와 시스템공학연구소가 공동으로 수행한 과기처 특정과제에서 SDML(Standard Dictionary Mark-Up Language)과 TDMS(Text & Dictionary Management System)를 개발하여 그 해결 방안을 제시한 바 있다[4].

이러한 대규모 전자사전의 구축을 위해서는 일관성과 정밀성 및 작업의 효율성을 위하여 각종 지원 도구를 활용하는 것이 필요하다. 코퍼스 문장의 철자 오류 등을 복구하고 정제하는 코퍼스 정련기법과 어휘에 대한 용례를 빠르고 정확하게 제공하기 위한 용례정보 추출기와 통계정보 추출기가 필요하고, 코퍼스로부터 자동 혹은 반자동으로 각종 언어정보를 획득할 수 있는 고기능 도구들의 개발이 필요하다.

3. 언어처리 도구 개발

한국어에 대한 언어처리 측면의 의미 있는 연산을 대분류하면 한국어 분석, 변환, 생성, 추출(요약) 등으로 나눌 수 있다. 여기서 변환은 한글-한자 변환 또는 번역 등을 의미한다. 한국어의 내용처리의 수준에 따라서는 한국어 형태소, 구문, 의미(개념), 담화/화용 등의 레벨로 나눌 수 있다. 따라서, 이들에 대응되는 한국어 형태소 분석기, 담화 분석기, 구문구조 생성기 등이 개발될 수 있다. 지금까지 대부분의 언어처리 도구는 분석 위주의 기술 개발이 이루어지고 있으며 비교적 생성에 대한 연구는 미약한 편이다. 변환에 해당하는 번역 기술은 다른 장에서 따로 다루기로 한다. 형태소 분석 기술은 이미 여러 연구기관에서 개발이 진행되어 왔으며, 현재 무료로 보급되거나 판매되고 있는 상황인데 반하여, 구문 분석 레벨 이상의 처리 도구는 아직 연구실에서 연구과제로 진행되고 있는 상태이다. 국어 정보 내용 처리 기술 수준을 한 차원 더 높이기 위해서는 구문-의미분석 및 담화 분석 기술의 개발에도 노력을 기울여야 할 것이다. 그러나 한국어 구문구조상의 특징인 부분자유어순 구조와 자유로운 생략현상 등으로 인하여 문장상에서 많은 구조적 모호성(ambiguity)이 유발되는 점과, 이들을 효과적이고 효율적으로 해결하는 기법과 이에 필요한 대규모의 정교한 언어지식((통계적)규칙, 사전정보)의 구축이 걸림돌이 되고 있다. 통계적 기법에서는 코퍼스의 각 형태소에 품사 태그(tag)를 부착한 품사부착 코퍼스나 문장의 구문 단위에 구문구조 태그(tag)를 부착한 구문구조 부착 코퍼스를 구축하여야 하며, 이것을 일정한 통계적 모델에 의하여 기계학습(learning)시키는 것이 필요하다. 이러한 학습을 위한 대규모의 기초 데이터의 구축과 검증작업 또한 만만치 않은 작업이다. 국내에서 의미분석은 개념 그래프(conceptual graph)와 QLF(quasi-logical form)를 출력구조로 생성하는 연구가 진행 중이며, 담화 분석은 호텔 예약, 자동차 구매 등과 같은 제한된 영역에서의 대화 모델링의 연구가 KAIST, 서강대,

시스템공학연구소 등에서 진행 중이다[3].

III. 응용 시스템

1. 정보 검색 및 가공 기술

정보검색 기술은 고도의 정보화 사회를 앞당기는 데에 필수 불가결한 요소로 보기 때문에, 선진국에서는 정보검색 관련 연구를 정보산업의 핵심 기술 중의 하나로 인식하고 대규모의 투자를 행하고 있다. 정보 검색 시스템의 성능과 평가의 기준들로는 신뢰도(정확도와 완성도), 속도, 대용량 데이터 처리를 위한 확장성, 사용자 편리성 등이 있으나 가장 중요한 요소는 사용자 언어의 특성과 문화에 매우 의존적이다. 현재 국내에서는 외국어를 위주로 만들어진 엔진 부분을 한글화하는 수준의 정보 검색 시스템이 상용화를 위한 개발과 더불어 몇몇 기업에서 독자적인 검색엔진 개발과 일부 상품화에 기여하고 있으나 국제 경쟁력을 지닌 고부가가치를 지닌 기술 개발의 필요성이 대두되고 있다. 그러므로 우리말과 문화의 특성에 기반한 정보검색 기술의 육성과 더불어 체계적이며 장기적이고 국가적인 육성책이 필요하다. 정보의 가공 기술은 정보 분류, 정보 여과, 정보 추출, 정보 요약, 하이퍼텍스트 자동 생성 기술 등으로 세분된다[5].

정보검색 및 가공 기술에서의 한국어 정보처리 기술은 아직 형태소 분석 레벨을 중심으로 명사 위주로 색인어를 추출하는 기술이 활용되고 있으며, 구문 분석 이상의 레벨의 기술은 관련 기술이 미성숙한 관계로 아직 적극적으로 활용되지 못하고 있으며 문서 내용의 이해 수준이 낮아서 의미 내용적 유사성에 기반한 검색은 거의 불가능하므로 현재로서는 키워드 스트링의 정확한 일치와 통계적 기법에 대부분 의존할 수 밖에 없는 실정이다. 그리고, 과다한 색인어 자동추출의 오버헤드와 복합명사에 대한 처리의 불완전성, 명사 이외의 키워드에 대한 색인 및 검색에 대한 제약 등이 문제로 대두되고 있다. 또한 한국어 어휘의 개념적 상-하위, 유사 및 기타 관계를 정의한 시소러스를 일부 기관에서 자체적 목적으로 구축하여 활용하고 있으나, 공유되거나 이를 위한 표준화가 거의 이루어지지 않은 것이 현실이다.

2. 기계번역 기술

일반 문서를 번역하는 시스템이 국내에서 다수의 제품들로 출시되어 있고, 최근 들어 인터넷의 확산이 본격화되면서 각종 외국어에 대한 번역 수요가 늘고 있다. 업체들은 인터넷과 번역 소프트웨어의 결합을 대세로 보고 기존 번역 소프트웨어들의 인터넷 버전이나 번역소프트웨어를 내장한 웹 브라우저를 출시하는 경향을 보이고 있다. 하지만 이러한 제품들은 기계번역의 본질적인 문제에 대한 적극적인 해결책을 가지고 출시된 것이 아니기 때문에, 번역의 품질에 있어서는 기대에 크게 못 미치고 있다. 그러나 이 분야의 시장형성에 크게 기여한 바는 인정하여야 할 것이다.

현재 활발히 연구되고 있는 번역기술로는 EBMT(Example-Based Machine Translation), CBMT(Corpus Based Machine Translation), SBMT(Statistically Based Machine Translation), KBMT(Knowledge Based Machine Translation), 그리고 Carnegie Mellon대학의 기계번역 센터에서 개발 중인 KANT와 같이 다양한 번역모델을 복합한 기술 등이 있으나, 상품화에 성공한 시스템들은 대부분 RBMT(Rule Based Machine Translation)엔진을 주축으로 구성되어 있다. 기계번역은 자연언어 처리의 언어 분석, 변환, 생성 등에서 발생하는 모호성을 포함한 여러가지 문제점을 공유하고 있으나, 좀 더 조건이 나은 측면이 있다. 번역의 결과는 어느 정도의 지적 능력을 갖춘 사람이 보기 때문에, 비록 역어 선택이나 문체가 매끄럽지 못하더라도 사람의 지적 능력으로 전후 문맥이나 이미 갖고 있는 영역지식 및 상황을 참조하여 오류를 복구할 수 있는 능력에 어느 정도 의존할 수 있는 측면이 있다. 따라서 기계번역 시스템을 완전한 번역 전문가로서 기대할 것이 아니라, 인간의 번역작업을 비록 서툴지라도 전혀 없는 것 보다는 훨씬 나은 보조자로 기대하고 취급하는 것이 단기적 목표로는 바람직하다고

본다[6].

자동통역은 음성인식, 기계번역, 음성합성 등의 기술이 복합된 기술 영역으로 현재 국내 기술은 한국통신(KT)과 전자통신연구원(ETRI)을 주축으로 장기적인 국제공동과제로 진행중이며 현재로는 제한된 영역(호텔 예약)에서의 시제품을 개발하는 수준이며, 만족할 만한 연구 성과는 2010년 이후에나 나올 것이라고 예측하고 있다.

3. 문서 퇴고 시스템

퇴고 시스템이란 문서의 작성 과정을 지원해 주고, 작성된 문서의 철자나 문체 오류를 검증해 주는 시스템이다. 퇴고 시스템은 1) 철자 및 문체 오류의 교정, 2) 동의어, 반의어 및 순화 용어의 제공, 3) 단어의 사용 빈도나 난이도에 대한 정보 제공, 4) 단어 및 문장의 사용 용례의 제공, 5) 한글-한자 상호 변환 기능 등이 있다[1].

국외에서는 구문 레벨의 오류를 탐지하고 교정을 지원하는 S/W가 출시되고 있고, 국내에서는 80년대와 90년대 초에 집중적인 연구가 진행되었다. 그 결과 단일 어절 형태소 분석에 기반한 철자 검사 및 교정 기능, 음절 및 단어 수준의 한글-한자 단순 변환 기능을 제공하는 한글과컴퓨터(주)의 아래아한글, 마아크로소프트사의 Word 등의 워드프로세스가 개발되었다. 한편 다수 어절을 고려한 부분적인 구문 분석 및 의미 분석 기법을 활용한 문서 퇴고 시스템의 시제품을 STEP2000과제를 통하여 부산대에서 개발하고 있다[5].

4. 에이전트

사용자의 지적인 작업을 대신해 주는 에이전트의 필요성은 인터넷 정보검색, 개인 비서 시스템, 전자 상거래 및 사용자 인터페이스 등 많은 분야에서 대두되고 있으며 특히 한국어를 처리할 수 있는 한국어 처리 에이전트의 필요성이 증가되고 있다. 국내의 경우는 KT와 ETRI 등에서 전자 비서 시스템을 개발 중이며, 포항공대 등에서 자연언어 인터페이스를 지원하는 인터넷 검색 기술(Air-Web)을 개발 중이며, 성균관대에서는 멀티에이전트 시스템인 ICOMA에 관해 연구중이다. 한국어 에이전트 기술은 단기적으로 홈페이지 검색 및 분류 기능을 가진 지능형 인터넷 검색 로봇이 실용화 추세에 접어들고 있으며, 사용자가 자주 검색하는 내용을 바탕으로 사용자의 의도를 분석하는 인터넷 consultant에 대한 연구가 활성화될 전망이다[8,9]. 장기적으로는 자연어 인터페이스에 대한 연구가 진행되어 인터넷 검색 뿐만 아니라 다른 응용 프로그램에도 적용될 것이며, 전문영역 분석에 의한 문서 여과 및 요약 기술이 발전하여 전자우편, 전자 뉴스 등을 여과/요약하여 사용자에게 정리/제시하는 지능형 비서가 개발될 전망이다.

IV. 한글 정보 처리의 문제

한글 정보처리와 관련된 기술에 대해 우리 나라와 선진국과의 수준을 비교하면, 각 분야에 따라 차이는 있지만 배분율로는 대략 50~60% 정도, 시간상으로는 대략 5내지 10년의 기술적 격차가 있는 것으로 파악된다. 선진국의 경우 이미 언어정보 베이스를 활용하여 연구 개발에 응용하고 있는 단계이며, 구문 분석을 통한 언어처리가 일반화되었다. 우리 나라의 경우는 연구실 수준의 언어정보 베이스를 구축하고 있으며, 단어 중심의 언어처리 수준이 주종을 이루고 있고 구절이나 문장 단위의 처리는 아직 그 수준이 비교적 낮다고 볼 수 있다. 컴퓨터 기술을 둘러싼 국제적인 경쟁이 점점 격화되고 있고, 유니코드의 제정 등으로 이제까지 우리의 고유 영역으로 인식되어온 한글처리 분야에 대한 선진국의 기술적 침투는 가속화될 것이 명백하다. 이에 대처하는 방법은 한글 및 한국어 처리 분야 만큼은 선진국들과 당당히 대결한다는 민족적 자존심을 걸고 기술 개발에 모든 노력을 바쳐야 할 것이다. 이러한 기술력을 배양하기 위해서는 한글 및 한국어처리 분야에 대한 집중적인 연구, 개발과 함께 장기적인 기획에 따른 지속적이고 체계적인 지원과 연구비 투자가 있어야 하며, 개발된 연구 결과들이 사장되지 않고 꾸준히 개선, 보완 및 활성화될 수 있는 체계가 마련되어

야 한다.

특히 언어학과 국어학을 하신 분들과 컴퓨터를 하신 분들과의 학제간 협동연구가 필요하다. 언어학이나 국어학을 하신 분들과 컴퓨터 분야에서 자연어처리를 하는 분들 사이에는 분명히 그 목표가 다르다. 그 현실을 양 분야에서 일하는 사람들 모두가 인식하여야 한다. 이제는 원론적인 말로만 컴퓨터 분야와 언어학 분야의 분들이 모여서는 아무런 소득이 없다. 많은 외국의 훌륭한 연구기관에서처럼 같은 방에서 같이 이마를 맞대고 일하여야 한다. 물론 아직도 우리나라에서는 어문학 계열에 계신 분들 중 자연어 처리에 관심을 가진 분들을 찾기가 그리 쉽지 않다는 것을 알 수 있다.

언어는 살아 있는 생물체와 같기 때문에 언어처리 연구는 앞으로 지속되어야 하며, 끊임없이 발생하는 새로운 언어현상에 대한 튜닝과의 전쟁에 임하여 신무기 '언어처리 기술' 개발과 적극적이고 끈기있는 연구 자세가 필요하다고 본다.

V. 향후 추진 방향과 전망

자연어처리 분야는 음성인식 분야와 마찬가지로 아주 오래 전부터 시작하여 왔으나 그 오랜 기간에 비해 결과물이 제대로 나오지 못한 것도 사실이다. 그 이유는 인간의 인식의 문제를 해결하여야 하기 때문이다. 그래서 많은 사람들에게 그 분야는 나이도 먹지 않느냐는 조롱을 받는 것도 사실이다. 그러나 이 분야의 기술이 어려운 것도 한 이유이나 이 분야에 종사하는 사람들의 상품화 노력도 부족하다고 할 수 있다. 모든 기술 분야가 그렇지만 특히 정보통신 분야는 연구 투자에 비해 상품효과가 큰 분야라 살아남기 쉽다. 왜냐하면 시장성 없는 결과물에 대해 지속적인 투자를 할 만큼 우리나라의 연구개발(R(D)분야가 성숙하지 못하기 때문이다.

이때까지 자연어처리 분야에서 나온 문법이나 알고리즘들을 보면 언어현상의 한 일면만의 처리를 위해 개발된 것이 거의 대부분이다. 잘 고쳐지지 않는 병에는 약 종류가 많다는 이야기가 있다. 이런 현상을 다른 측면으로 본다면 자연어의 언어현상은 생각보다 더 다양하고 복잡하여 한번에 해결될 수 없다는 것을 의미한다. 사실 각종 문법(grammar)에 열을 올리던 방법은 코퍼스(corpus)기반의 자연어처리 이전에는 무슨 유행처럼 번지던 풍조였으나 지금은 조금 조용한 편이다. 우리 나라말을 조사해보면 학교에서 배우거나 생각했던 것보다 조사의 개수도 훨씬 많고, 동사의 변형도 훨씬 많다. 또 다른 언어현상도 생각했던 것보다 훨씬 많다. 이러한 모든 언어현상을 처리할 수 있는 것이 코퍼스(corpus)를 기반으로 하는 방법밖에 없다는 맹신을 또 갖지 말아야 한다. 1억 어절이면 해결될 지 10억 어절은 되어야 할 지 아무도 아직 모른다. 그래도 분명한 것은 한국어 처리에 맞는 이론을 세우기 위해서는 한국어에 대한 자료가 축적되어야 한다.

현재 범국가적으로 한글 정보 처리 기술 확보의 필요성을 인식하여 여러 방면으로 노력하고 있다. 일차적으로 21세기의 국가 전략 산업으로 소프트웨어 산업이 지정되고 이 분야에 종사하는 중소기업 및 연구진에 대한 정책적, 재정적 지원 사업이 시작되었으며, 이차적으로 국책과제로 "국어정보처리 기술 개발" STEP2000 과제(1994 −2005)(과기처), "우리말 정보처리 S/W 기술 개발" (정통부) 등이 추진되고 있다.

특히 이들 국책과제의 추진을 통하여 지금까지의 산발적이며 비지속적인 연구과제에서 벗어나 장기적이며 계획적인 한글 정보 처리 연구가 가능하게 되었다. 그리고 연구 결과로 얻어진 정보베이스, 처리도구 및 첨단 기술들은 국내 연구진과 중소기업들에게 기술 전수의 형식으로 보급됨으로써 범국가적인 한글 정보 처리 기술의 확대와 축적이 가능하게 되었다. 다른 분야보다 외국 소프트웨어에 비해 상대적으로 경쟁력이 강한 한글 정보 처리 관련 소프트웨어를 개발하는 중소 기업도 늘고 있는 추세이다. 이러한 국가적인 지원과 기업체, 학교, 연구소의 집중적인 노력이 결집되어 21세기에는 한글 정보 처리 분야가 기술적으로 큰 발전을 이룰 것을 기대한다.

참 고 문 헌

〔1〕 권혁철, "한글 및 한국어 정보 처리의 현황", 정보과학회지 제 12권 제 8호, 1994.

〔2〕 노용균, 박동인, "Corpus Linguistics의 현황과 한국어 Corpus 구축 및 활용의 제문제", 정보과학회지 제 12권 제 8호, 1994.

〔3〕 시스템공학연구소, 국어정보처리기술개발 : 한글 언어처리 기반 기술(STEP2000 1단계 최종보고서, 과학기술처, 1997.

〔4〕 시스템공학연구소, 국어정보처리기술개발 : 국어정보베이스 구축(STEP2000 1단계 최종보고서, 과학기술처, 1997.

〔5〕 시스템공학연구소, 국어정보처리기술개발 : 지능형 처리기 개발(STEP2000 1단계 최종보고서, 과학기술처, 1997.

〔6〕 시스템공학연구소, 우리말 컴퓨터를 위한 개발 계획 수립, 과학기술처, 1995.

〔7〕 시스템공학연구소, 1997 국가 정보화 백서 : 국어정보처리, 한국전산원, 1997.

〔8〕 신봉기, 김영환, "웹 에이젼트", 정보과학회지 제 15권 제 2호, 1997.

〔9〕 최중민, "에이젼트의 개요와 연구방향",정보과학회지 제 15권 제 2호, 1997.

저 자 소 개

朴 東 仁

1953年 9月 7日生
1972年 2月 서울고등학교 졸업
1979年 2月 서강대학교 전자공학과 졸업(학사)

1979年 11月～현재 시스템공학연구소 자연어정보처리연그부 부장
1994年 4月～현재 공업진흥청 산업표준 심의회 위원
1995年 5月～현재 국어정보학회 이사
1995年 9月～현재 문화체육부 국어심의회(국어정보화분과) 위원
1996年 5月～현재 한국어정보처리연구회 위원장
1996年 8月～현재 한국정보과학회 평의원

주관심 분야 : 정보검색, 자연언어처리, 기계번역