한국어 정보처리 : 어제와 오늘

부산대학교 권혁철*

한국어 정보 처리는 컴퓨터와 관련한 연구분야 중에서 우리가 주도하고 있고, 또 주도해야만 하는 분야다. 크게 보면 한국어 정보 처리는 자연언어 처리의 한 부분이다. 하지만 한국어 정보처리 분야의 연구에는 우리의 문화와 역사에 바탕하여 독자적으로 발전한 우리의 언어 환경에 따른 우리만의 여러 문젯거리가 있다. 따라서 한국어 정보 처리 분야의 연구와 개발은 자연언어 처리의 세계적 흐름을 따르면서도 독자성을 가지고 발전해 왔다.

초기의 한국어 정보 처리 연구는 한글정보를 컴퓨터에 입출력하고 저장하여 처리하는 기본 기능을 제공하는 한글코드와 한글자판을 중심으로 수행했다. 한글코드에 대한 체계적 연구는 '70년대 말에 인하대학교 이주근 교수가 시작했다. 1982년에는 표준 한글코드의 필요에 따라 과기처가 박동인 등이 기초한 표준 한글코드를 발표했다. 이 한글코드는 ISO2022에 기반한 N-byte 코드와 이를 이용한 2byte 조합형(상용조합형이라 부름) 및 2byte 완성형 코드를 포함하는 체계다.

'80년대 중반에 이 코드체계의 문제점이 제기되면서 새로운 한글코드의 제정을 시도했다. 그 결과 발표된 한글코드가 완성형 한글코드인 KS5601이다. KS5601은 그 당시 컴퓨터에 한글을 입출력하는 기술이 일본에 종속된 상황에서 일본기술을 도입하던 기업체가 중심이 되어 만든 코드로서, 태생적 한계가 있었다. 따라서 KS5601은 자소에 기반한 한글의 특성을 반영하지 못할 뿐 아니라 우리가 쓰는 한글 음절마저 모두 표기할 수 없는 기형적 한글 코드가 되었다. 따라서 문화체육부(현 문화관광부), 한국정보과학회, 언어학자와 한글과컴퓨터사 등이 중심이 되어 KS5601에 대해 집중 비판을 하고 대안을 모색했다. 결국 KS5601을 기초한 개발자마저 KS5601의 문제점을 인정하였지만, 한 번 정해진 코드를 바꿀 수는 없었다.

한편 '90년대 초부터 미국을 중심으로 전세계 모든 문자를 표현할 수 있는 통합코드인 유니코드(uni-code)의 제정이 시도되었다. 그런데 잘못된 KS5601은 유니코드의 한글코드영역 배정도 어렵게 했다. 이에 따라 유니코드에서 한글 영역의 확보와 효과적인 한글 표현을 위해 국가 차원에서 집중 노력했으나, KS5601이 가진 독소는 기존 유니코드의 한글 표현을 체계성이 없고, 중복적인 형태가 되게 했다.

비슷한 문제가 한글 자판에서도 나타났다. 기계적 타자기가 쓰이던 때에는 글쇠가 엉키는 현상을 방지하려고 최대한 타자속도를 줄이는 방향으로 자판을 만들었다. 이런 기계적 타자기에서 유래한 한글자판이 컴퓨터용 자판으로는 부적합함이 널리 알려졌다. 또 공병우 선생이 중심이 되어 한글에는 1)세벌식 자판이 적합하다는 주장도 제기했다. 이에 따라 인체 공학적 관점에서 개발한 다양한 한글자판이 제시되었다. 하지만 이 노력도 역시 실패로 끝났다.

불행히도 '90년대 중반을 지나면서 현실의 한

* 종신회원

1) 두벌식 자판은 자음과 모음으로 구별하고 세벌식 자판은 초성, 중성과 종성으로 구별하여 자모를 배치한다

계로 한글코드와 한글자판에 대한 논의는 더 이상 일어나지 않고 있다. 이제 우리 문화의 가장 찬란하고 자랑스러운 한글을 컴퓨터에 입력하는 자판과 내부에 표현하는 코드가 미필적(?) 회피 속에 왜곡된 채 새로운 세기를 맞게 되었다.

한국어 정보 처리의 핵심 연구는 언어처리, 음성인식과 문자인식으로 나뉜다. 여기서는 언어처리 중심으로 지난 20년간 일어난 중요 사건에 따라 한국어 정보 처리에 대해 설명하고, 음성과 문자 인식은 지면 관계로 생략하겠다.

한국어 정보 처리에 대한 연구는 인하대학 이주근 교수가 '70년대 말에 시작했다. 그 후 '80년대 중반에 한국과학기술원 김길창 교수, 서울대학교 김영택 교수와 한양대학 최병우 교수가 기계번역 시스템을 개발하면서 본격적으로 한국어 정보 처리에 대한 연구가 시작되었다. '80년대 중반에 김길창 교수는 NEC의 지원으로 NEC에서 개발하던 영·일 기계번역에 기반하여 영·한 기계변역을 시작했다. 김영택 교수는 자체에서 개발한 한·영 기계번역 시스템(KEMTS)을 일본에서 발표하자 일본IBM이 연구비를 지원하겠다는 요청을 하여 영·한 기계번역 시스템을 개발하기 시작했다. 최병욱 교수는 일·한과 한·일 기계번역 시스템을 연구함으로써 한국어 정보 처리 분야에 한 흐름이 되었다.

거의 동시에 기계번역을 시작한 국내 연구진은 '60년대에 외국 기계번역 연구자가 가졌던 기대와 희망을 가졌다. 하지만 '80년대 후반에 들어서면서 외국 기계번역 연구자가 겪었던 실망을 하게 된다. 그러나 이 기간에 있었던 기계번역 연구는 국내 한국어 정보 처리의 씨앗을 뿌리는 중요한 계기가 된다. 특히 이기용, 장석진, 신수송과 임홍빈 교수로 대표되는 이론언어연구모임이 단일화기반 문법을 도입하여 매주 숭실대학에서 세미나를 진행함으로써 한국어 정보 처리 연구자가 언어학의 관점에서 한국어 처리를 바라보게 되는 계기가 된다. 또 동경대학에서 박사학위를 한 정희성 박사가 KPSG(Korea Phase Structure Grammar)와 국어공학이란 개념을 도입함으로써 한국어 정보 처리 분야에 신선한 충격을 주었다. 한편 '80년대 후반을 지나면서 김길창 교수는 최기선, 이종혁, 안동언, 김덕봉을 김영택 교수는 권혁철, 옥철영, 서명훈, 윤덕호, 강승식 등을 배출함으로써 국내 한국어 정보 처리 분야의 밑거름이 되게 했다. 또 외국에 학위를 마친 서정연, 임해창, 이근배, 정희성, 나동렬이 귀국함으로써 한국어 정보 처리 분야가 국내 컴퓨터 분야에서 독자성을 이룰 토대를 마련했다 이 젊은 연구자들은 그때까지 연구에 대한 반성을 바탕으로 감상적 낙관론을 비판하면서 굳건하고 실용화할 수 있는 기술개발이 필요함을 주장했다.

1989년부터 한국정보과학회, 한국인지과학회와 한국언어학회가 공동으로 '인간과 기계와 언어'라는 화두를 내걸고 매년 한글날을 기념하여 '한글 및 한국어 정보처리'에 관한 학술대회를 개최하여 한국어 정보 처리 연구자들이 서로 접근방법을 비교하고, 연구결과에 대한 기술적 비판을 함으로써 '80년대 접근 방법의 한계를 서서히 넘어서기 시작했다. 또 1990년 10월 26일에는 한국과학기술원의 김진형, 권용래와 전길남 교수의 지원으로 한국정보과학회 산하에 한국어 정보 처리연구회를 설립함으로써 한국어 정보처리 분야의 발전에 획기적 전기를 마련했다.

김진형과 권용래 교수는 1989년에 시작한 한글날 기념 학술대회를 만드는 산파역을 함으로써 당시에 중견급 연구자가 없어서 어려움을 겪던 이 분야의 후원자로서 역할을 충실히 했다. 곧이어 김영택 교수가 과학재단의 학문 분류에 '한국어 정보 처리 분야'를 추가함으로써 한국어 정보 처리가 컴퓨터 분야의 독립 분야로 설 수 있게 했다. 또 오길록 박사의 후원 하에 연구소에서는 박동인 씨와 박세영 박사가 한국어 정보 처리 연구의 독자적 영역을 구축하였다.

'90년대에 들어 한국어정보처리 연구자들은 단계적으로 확장할 수 있으면서 실생활에 쓸 수 있는 언어처리 응용시스템의 개발에 큰 관심을 보였다. 1991년에는 한글과컴퓨터사와 부산대학에서 국내 최초로 상용 언어처리 응용시스템인 한국어 맞춤법검사기를 발표하였지만 성능이 기대에 못 미쳤다. 따라서 기반 기술부터 다시 연구해야한다는 인식과 통계적 접근에 의한 언어처리 기술이 세계적으로 각광을 받던 시대 흐름에 따라 대용량 말뭉치 구축의 필요성이 제기되었다.

이런 흐름을 읽은 박동인이 과기부의 지원으로 언어처리 기반기술의 확보 차원에서 대용량 말뭉

치와 전자사전을 개발하기 위한 과제인 STEP2000(1994~2000)을 시작함으로써 한국어 정보 처리 분야의 체계적 연구환경을 구축하는 계기를 만든다. 박동인은 뛰어난 친화력으로 과기부와 문화부를 설득하고 한국어정보처리 분야 연구자를 모아서 1994년 11월 19일에 국어공학센터를 설립함으로써 한국어 정보 처리 연구의 역량을 결집할 수 있는 토대를 만든다. 또 오길록 소장의 적극적 지원에 힘입어 시스템공학센터 내에 자연언어처리 연구부를 만듦으로써 각 대학에서 배출한 한국어 정보 처리 전공자가 소신껏 연구할 수 있게 했다. '90년대 말에는 문화부를 설득하여 세종계획(10년 장기 계획)을 시작하게 하여 언어학자를 중심으로 한국어 정보처리를 위한 기초자료를 구축하게 했다.

1994년을 지나면서 사회적 요구에 따라 한국어정보처리 응용시스템 개발이 크게 진척되어 영·한 기계 번역기, 일·한 기계 번역기, 한국어 정보검색기와 한국어 맞춤법 검사기의 성능이 획기적으로 향상되고, 한국어 처리와 관련된 다수 회사가 설립되었다. 하지만 기술자 입장에서 느낀 획기적 기술 진보도 일반 사용자를 만족시키는 데는 한계를 보였다. 이에 따라 '90년대 중반을 지나면서 한국어 형태소 분석 기술부터 완성해보자는 의견이 대두되었다.

'90년대 후반에 오면서 형태소분석과 태깅기법에 대한 연구·개발은 규칙을 따르는 연구자와 통계와 기반한 학습을 따르는 연구자가 경쟁적으로 기술을 개발함으로써, 1998년에 이르러 한국어 형태소 분석 기술이 컴포넌트 수준에서 실용화에 도달하게 되었다 이런 기술적 성공에는 KAIST 말뭉치, 연세대 말뭉치와 고려대 말뭉치 등 대용량의 말뭉치가 큰 역할을 했다. 드디어 1999년 한글날기념 학술대회에서 지금까지 개발한 한국어 형태소 분석기의 성능을 비교하는 MATEC99가 17개 팀이 참여한 가운데 진행됐다. MATEC99를 준비하기 위해서 한국전자통신연구원의 박세영 부장과 박재득 박사가 1년간 여러 차례 전문가 회의를 거쳐서 표준 테그셋(tag set)을 설정했다. MATEC99는 이 표준에 기반하여 각 시스템을 비교하고 평가했다. MATEC99과 같은 체계적인 시스템 성능의 비교·평가는 국내에서는 처음이며, 외국에서도 드물다.

MATEC99는 한국어정보처리 분야의 기술현황과 발전방향을 설정하는데 중요한 이정표가 되었다. 더구나 250명이 넘는 인원이 한글날기념 학술행사에 참여함으로써 한국어 정보 처리에 대한 관심이 커지고 이 분야의 전문가가 크게 늘었음을 보인 것도 큰 성과다. MATEC99를 기점으로 한국어정보처리 연구자는 한국어 형태소 분석 기술에 대한 충분한 경험에 힘입어 이제 한국어 품사분석과 의미분석에 대한 연구를 시작할 준비가 시작되었다.

인터넷의 확산에 따라 기계번역과 정보검색 기술의 혁신적 발전이 절실히 필요한 현시점에서 한국어정보처리 기술은 우리 민족의 자존과 우리 문화의 정체성을 지키고 국민의 정보활용을 지원하는 핵심 기술이라고 자부한다. 하지만 한국어 정보 처리 분야의 종사자로서 다음 몇 가지에 대해서는 걱정이 앞선다.

먼저, 한국어 정보처리 기술은 하루아침에 완성될 수 없으며, 또 남이 개발해 줄 수 없는 만큼 기반 기술부터 체계적으로 개발해야 한다. 그러나 IMF 이후 실적 위주의 단기적 상품화 기술 개발에 치중하는 현 상황에 따라 기반 기술의 개발이 도외시될까 걱정이다. 한글코드나 자판처럼 한 번 잘못되면 다시 돌리기 어려운 만큼 기반기술부터 인내하며 연구를 해야만 국민이 요구하는 한국어 정보 처리 기술을 개발할 수 있다.

다음으로 일부 외국 기업이 한국어정보처리 기술을 국내시장 장악을 위한 기반기술로 보고 거금을 들여 국내 기업을 인수하고 박사급 인력을 확보하고 있지만, 우리 나라 기업과 정부가 한국어정보처리 기술을 소프트웨어 산업의 기반 기술로 인정하고 있는지 의문이다. 국내에서 사주지 않아 일본에 판 한글폰트가 10배의 가격으로 국내의 대기업에 역수입된 '90년대 초기의 사건이 새롭다. 또 외국에서 개발한 한국어 음성합성시스템의 성능이 한국에서 개발한 어떤 제품보다 우수하다는 현실이 정말 걱정스럽다.

마지막으로 기술개발은 시간과 인내를 요구한다. 본인의 경험에 따르면 한국어 맞춤법검사기를 처음 개발하는 데 6개월밖에 걸리지 않았지만, 그 후 7년의 연구·개발에도 아직 만족스러운 결과를 얻지 못하고 있다. 그렇다고 이 연구

를 그만둘 수는 없다. 최근에 몇몇 신문기사가 본인이 개발한 시스템을 사용하기 시작하는 것을 볼 때 이제 조그만 희망을 본다. 하지만 내 평생을 연구해도 한국어 맞춤법검사기의 끝을 볼 수 없으리라 생각한다. 한국어 정보 처리에 관여하는 연구자는 조바심을 내지 말고 소신에 따라 평생을 기술개발에 투자해야 하며, 정부와 기업도 국가백년을 바라보며 체계적으로 기반부터 한국어 정보 처리 연구에 투자해야 한다.

이제 새로운 세기를 맞는다. 기독교 달력에 따른 새로운 세기이지만, 그래도 새로이 바꾸어야 할 그리고 바꿀 계기로서 새 천년의 의미는 크다. 한국어 정보 처리 연구가 새 천년에는 새로운 발전을 통하여 백성을 이롭게 하고, 백성이 널리 정보를 쓰게 하는 기술이 되었으면 한다.

권 혁 철

1982 서울대학교 공과대학 전산학 학사
1984 서울대학교 공과대학 전산학과 석사
1987 서울대학교 공과대학 전산학 박사
1988~1992 부산대학교 자연과학대학 전자계산학과 조교수
1992~현재 부산대학교 자연과학대학 전자계산학과 교수
1992~993 미국 Stanford대학 CSLI연구소 연구원
1993 Xerox Palo Alto 연구소 자문
관심분야 한국어 정보처리, HCI, 정보검색
E-mail:hckwon@hyowon.cc.pusan.ac.kr