언어정보처리산업 전망과 정책방향

정보통신부 임차식

1. 언어정보처리산업의 개요

언어정보산업이란 컴퓨터가 일상적인 언어를 이해하고 생성하여 인간의 지적 노동의 보조자 및 지원 도구로 활용될 수 있는 산업을 말한다. 현재 전형적인 언어정보산업이라 할 수 있는 분야는 자동번역, 정보검색 등이 있으나 국내외 기술 수준이 시장요구에 부응하지 못하고 있어 그 잠재시장을 형성하는데 큰 걸림돌이 되고 있다. 그러나 언어정보처리 기술이 시장에서 요구하는 수준에 오른다면 이 분야의 기술을 이용한 직접적인 응용 제품 뿐 아니라 소프트웨어 전반에 걸친 엄청난 파급효과를 가져올 수 있을 것으로 예상되고 있다. 실지로 인터넷의 발전으로 인한 정보의 폭증으로 언어정보 기술의 산업화가 급격하게 촉진되고 있어 Technology Review 최근호에서는 미래를 이끄는 혁신적인 기술 10가지 가운데 언어정보 산업 관련 기술을 선정하기도 하였다.

언어정보산업은 그 자체가 매우 급성장할 가능성이 있는 산업이지만 오히려 소프트웨어 전반에 걸친 파급효과 측면에서는 21세기 지식기반 사회로 가는데 있어 핵심전략 산업이라 할 수 있을 것이다. 그 이유로 첫째 한 나라의 언어를 다른 나라의 언어로 변환하고 생성할 수 있는 언어처리 기술이 발달하면 현재 부분적으로 적용되고 있는 자동번역 기술이 다국어 자동번역 시스템, 자동통역 시스템 그리고 다국어 검색 시스템 등에 확대 응용되어 언어장벽 해소 및 글로벌 시대를 앞당길 것이다.

둘째 대화처리 기술과 에이전트 기술이 발달하여 인간의 언어로 컴퓨터와 대화가 가능해진다면 모든 소프트웨어의 인터페이스가 달라질 것이다. 이는 또한 DOS의 문자기반 컴퓨터 운영체제가 Microsoft Window에 의해 시각적인 컴퓨터 운영체제로 바꾸었듯이 대화처리 인터페이스는 지능적인 컴퓨터 운영체제를 가져올 것이다.

셋째 언어처리 기술을 이용하여 지식의 획득, 생성 및 관리를 보다 지능적으로 함으로써 현재 각광받고 있는 기존 소프트웨어 산업인 지식관리시스템(KMS), 고객관계관리시스템(eCRM), 데이터마이닝 등의 기능 및 성능향상에 이용되어 그 부가가치를 크게 향상시킬 수 있을 것이다.

2. 언어정보처리 국내외 기술 동향

2.1 자연어처리 기술

자연어처리 기술은 대상언어 및 적용 방법에 따라 많은 차이가 있어 외국의 언어처리기술을 그대로 한국어 언어처리에 사용하는 것은 여러 가지 어려움이 있다. 현재 형태소 분석, 태깅 등에 대한 국내기술은 선진 외국기술과 비교하여 근접한 수준에 이르고 있으나 구문분석, 의미분석 등은 선진국에 비해 낙후된 실정이다.

2.2 대화처리 기술

대화처리기술은 미국 등 선진국을 중심으로 기술집약적 벤처기업에서 연구되고 있으며 그 형태는 데이터 형태에 따라 다양하게 나타나고 있다. Answers.com, Askit.com, YY software, Iaskweb.com 등의 회사들은 주로 미리 구축된 지식베이스인 FAQ 리스트를 대상으로 고객의 질문에 자동 응답하는 솔루션을 출시하고 있으며, CRM을 위한 email 자동 응답이나 웹 보드 자동 응답 등의 제품이 출시되고 있다.

국내에서는 다이퀘스트닷컴, 서치캐스트 등이 웹

사이트, FAQ, 데이터베이스 통합 질의응답 솔루션 등을 출시하고 있으며 Natural Approach, Answerer, 다음소프트같은 회사들이 데이터베이스 질의응답과 FAQ응답을 각각 시도하고 있는 상황이다. 이상으로 볼 때 선진국은 이미 자연어 질의응답을 이용한 지식관리, EP(enterprise portal), eCRM시장의 초기 형성 단계라고 볼 수 있으며 국내는 아직 자연어 질문의 의미/의도분석 기술의 초기개발 단계라고 볼 수 있다.

2.3 자연어 검색 기술

현재 국내의 언어정보처리 기술은 형태소분석을 중심으로 오래 전부터 발전되어 왔으나 텍스트에서 키워드 추출 정도의 수준이며 자연어처리를 기반으로 하는 연구는 초기단계인 상황이다.

그러나 현재 미국 등 선진국에서는 막연한 관련 정보가 아닌 사용자가 원하는 정확한 문서를 제공하는 보다 고급의 정보검색 시스템이 선보이고 있으며 사용자의 검색 만족도를 높이는데 주력하고 있다. 그 방법으로는 Askjeeves와 같이 방대한 양의 질의문장의 패턴을 수작업으로 미리 만들어 놓고 사용자의 질의와 비교하여 미리 정의된 문서의 링크를 제시하는 기법과 형태소 분석 뿐 아니라 더욱 정확한 검색을 위해서 구문분석이나 의미분석에 의한 정보검색 시스템 개발을 추진 중에 있다.

2.4 텍스트 마이닝 기술

사용자의 고급 검색 부가 기능에 대한 욕구가 높아짐에 따라 정보 추출, 요약, 분류와 같은 기술 개발에 많은 투자를 하고 있다. 국내에서 상품화된 요약기로는 MicroSoft 워드에 포함된 것과 미국의 Semio사의 자동분류기술을 한글화한 3Soft 제품이 시장에 출시하고 있으나, 영어와 한글의 언어적 특성의 차이를 극복하는데 많은 어려움을 겪고 있는 중이다.

이에 비해 미국의 경우 정보검색 및 필터링에 대한 성능 평가를 위한 TREC(Text REtrieval Conference) 프로젝트가 미국 정부의 지원 아래 이루어지고 있으며, 정보 요약에 대한 평가를 위해 SUMMAC 프로젝트가 진행되고 있다. 뿐만 아니라 미국 NIST에서는 국방성 주도아래 지난 10년간 언어 처리 기술 연구에 많은 노력을 기울여 왔으며 매년 TREC이라는 정보검색 경연대회를 열어서 기술평가를 주도하고 있다.

2.5 자동 번역 기술

국내에서는 자동번역을 위해 현재까지 상용화되었던 영한 번역 및 일한 번역 소프트웨어는 다양한 제품이 출시되었다. 그 대표적인 것으로는 영한 번역의 경우 IBM의 TransMate, 엘앤텍의 E-Tran, L&I Soft의 EnGuide, 언어와 컴퓨터의 iTran, ClickQ의 ClickQ, 트래니2000 등이 있으며, 일한 번역의 경우에는 유니소프트의 바벨, 디코시스템의 i-Seoul/JK, 창신컴퓨터의 EZ Trans 등이 있다.

한국에서의 번역 제품이 단일언어에서 단일언어로 번역하는데 비해 선진국에서는 다국어에 대한 번역을 시도하고 있다. 그 예로 SystranSoft에서는 일반 문서 번역 시스템인 SYSTRAN Personal, 웹을 통한 영어, 프랑스어, 독일어, 이태리어, 포루투칼어, 스페인어를 번역하는 SYSTRANET 및 SYSTRAN PROfessional Standard, SYSTRAN PROfessional Premium, SYSTRAN Enterprise을 개발하였고, IBM에서는 8개 언어 번역 서버를 개발하고, 본사의 언어처리 기반 기술을 이용하여 해외 지사에서 각국 언어처리 제품을 개발하였다.

3. 국내외 시장 동향

현재 언어정보처리산업에서 언어처리 기술을 이용한 정보검색, 자동번역 등과 같은 직접시장과 언어처리기술이 솔루션으로 들어간 지식관리시스템(KMS) 등의 간접적인 응용 시장으로 나누어 볼 수 있다. 간접적인 응용시장을 포함한 언어정보 관련 세계시장은 2001년 430억불 규모에서 2004년 1,545억불 규모로 연평균 50% 이상의 고성장이 예상된다.

(단위: 백만 달러)

구 분	2001년	2002년	2003년	2004년
지식처리	13,579	18,102	24,326	33,405
다국어처리	11,676	14,980	20,050	24,080
사용자 인터페이스	13,294	24,593	39,991	77,500
컨텐츠 처리	4,540	7,824	12,328	19,540
합 계	43,089	65,499	96,695	154,525

(출처: IDC) ※ 순수 언어정보산업은 언어정보 관련 시장의 10%를 점유할 것으로 예상

직접적인 언어정보처리 시장규모를 살펴보면 국내 검색엔진 시장 규모는 2001년 약 600억원으로 매년 100%의 성장률을 나타내고 있다. 언어정보산업의 또 다른 중요한 시장 중의 하나는 자동번역 시장규모는 약 2000억원으로 매년 40%의 성장률을 나타내고 있다.

4. 언어정보처리산업 육성 정책방향

언어정보처리산업은 현재로서는 기술개발의 미흡 등으로 시장이 크지 않지만 지능형 정보검색, 지식정보관리 등 응용분야는 상당히 광범위한 분야이다

정부는 이처럼 성장초기에 있는 언어정보산업을 다음과 같이 전략적으로 육성함으로써 관련 핵심기술 확보 및 업체의 경쟁력 강화를 유도해 나아갈 방침이다.

4.1 중소 벤처기업 육성

첫째 산업체의 응용기술개발을 지속적으로 지원하여 경쟁력 있는 제품을 조기에 개발하고 시장 점유율을 확대해 나갈 계획이다. 이를 위해 음성 및 언어분야에 약 35억원의 기술개발 자금을 지정공모사업으로 지원할 계획이며, 이미 협의회(음성정보처리산업협의회, 언어정보처리산업협의회)를 통하여 기술개발 과제를 도출한 바 있다.

또한 음성 및 언어분야의 융자지원 비중을 지속적으로 확대하고 S/W지원센터에 음성 및 언어정보처리업체의 입주비율을 확대해 나갈 계획이며 아울러 정보통신유망중소기업 지원분야로 음성·언어정보처리분야를 추가적으로 지정하여 해당기업이 유망중소기업으로 지정되어 기업가치를 향상시킬 수 있도록 지원할 계획이다.

정부는 앞으로 관련 협의회를 통하여 업체의 필요한 기술수요를 지속적으로 발굴할 것이며, 기타 언어정보처리업체의 애로사항을 수렴·정책에 반영할 것이다. 음성 및 언어정보처리산업협의회는 정부와의 협력업무뿐만 아니라 업체정보의 공유, 국내외 기술개발 동향 등의 정보제공, 언어정보처리업체의 애로사항 협의 등 다각적인 업체지원 방안을 모색해야 할 것이다.

4.2 핵심 기반기술 및 기초기술 개발

둘째, 한국전자통신연구원 등을 통하여 언어정보처리 핵심기술을 개발하고 아울러 학계의 기초기술 개발도 지원할 계획이다. 이미 단어인식 기술 및 한일·일한 번역, 한영·영한 번역 등은 업체들의 상용기술이 많이 개발중이나, 숫자, 대화체 인식, 한중·중한 번역·통역 등은 업체에서 개발하기에는 이른 단계로 국책연구소에서 개발할 필요가 있으며, 데이터 마이닝, 지식정보관리 등에 대해서도 국가적 차원에서 집중 연구가 필요한 분야이다.

또한 음성·언어정보처리산업은 전자공학, 전산학 및 언어학의 협력이 매우 중요하므로 대학정보통신연구센터(ITRC) 지원사업, 정보통신학술연구지원사업 등을 통하여 이들 분야의 기초 연구가 자연스럽게 접목되어 시너지 효과가 창출될 수 있도록 유도할 계획이다. 이미 ITRC 지정분야로 음성 분야가 선정되어 금년중에 지원이 이루어질 계획이며, 향후 동 분야의 기초 연구가 지속적으로 이루어질 수 있도록 ITRC를 확대해 나갈 계획이다.

4.3 음성·언어 DB 구축

셋째, 음성·언어 정보처리시스템의 개발기간을 단축하고 제품의 질적 향상을 도모할 수 있도록 공용 음성·언어 DB를 구축할 계획이다. 이를 위하여 음성·언어자원을 구축·관리하는 전담기관을 금년 중에 설립하고, 매년 20억 이상의 자금을 투여하여 음성·언어 DB를 구축할 것이다.

또한 문화관광부에서 추진중인 21세기 세종 계획과 연계하여 언어정보처리 업체에 필요한 언어자원을 구축할 것이다.

아울러 LDC(Linguistic Data Consotium), ELRA (Europcan Language Resourcc Association) 등 외국의 관련 연구기관과 연계·협력을 통하여 대규모의 국내외 음성·언어자원이 구축될 수 있도록 할 계획이다.

임 차 식

1982. 2 한국항공대학교 전자공학과 졸업
1988. 6 미국 조지아공대 대학원 졸업
1981. 12 제17회 기술고시 합격
1991. 1 정보통신국 정보통신기술과
1993. 10 체신부 연구기술과장
1998. 6 정보통신부 정보통신정책실 산업지원과장
1999. 10~현재 정보통신부 정보통신정책국 소프트웨어진흥과장
E mail:csleem@mic.go.kr