다국어 자동번역 기술

Multilingual Machine Translation Technology

정보통신 미래기술 특집

최승권 (S.K. Choi) 언어처리연구팀 책임연구원
홍문표 (M.P. Hong) 언어처리연구팀 선임연구원
박상규 (S.K. Park) 언어처리연구팀 팀장

목 차

Ⅰ. 서론
Ⅱ. 다국어 자동번역 기술
Ⅲ. 연구 및 개발 현황
Ⅳ. 시장 현황
Ⅴ. 결론

W-CDMA 기술의 획기적인 발전과 보급은 향후 소규모 단말기를 통한 다국어 자동 통/번역에 대한 수요를 대폭 증가시킬 것으로 예상된다. 특히 북경올림픽을 기점으로 이와 같은 기술에 대한 수요는 매우 늘어날 것으로 예측되고 있다. 이에 대비하여 각국에서는 다국어 자동통역의 근간이 되는 다국어 자동번역 기술을 국가주도 하에 경쟁적으로 진행하고 있다. 독립 응용시스템에 적용되던 다국어 자동번역 기술은 인터넷의 발전 등과 더불어 이제는 문서에 관한 통합 프로세스를 담고 있는 응용시스템의 일부로서 적용되고 있는 추세이다. 본 논문에서는 다국어 자동번역 기술의 역사와 현황, 국내외 연구진들의 연구방향 등을 소개하고 향후 다국어 자동번역 기술 개발의 방향을 점검해보고자 한다.

Ⅰ. 서론

많은 사람들은 영화 '스타워즈'의 C3PO 로봇을 기억할 것이다. 이 로봇은 수백만 개의 은하계 언어를 서로 통/번역할 수 있는 거짓말 같은 능력을 지니고 있다. 인류 역사에 있어 언어장벽에 의한 의사소통 문제는 예전부터 인류가 넘어야 할 큰 과제로 여겨져 왔으며, 현재도 그러하며 영화에서 볼 수 있는 바와 같이 미래에도 그럴 것이다. 이러한 문제를 해결하기 위해 전세계적으로 활발히 연구 개발하고 있는 것이 바로 다국어 자동번역(multilingual machine translation) 기술이다.

2차 세계대전 이후 주로 미국과 러시아에 의해 군사적 목적에 의해 주도되어 왔던 다국어 자동번역 기술은 현재에는 군사적 목적 이외에도 웹 글로벌라이제이션(web globalization), 소프트웨어 및 문서 현지화(localization), 방송자막 및 채팅 번역과 같은 구어체 번역 등과 같은 목적을 위해서도 활발히 연구 개발되고 있다.

독립적인 응용시스템(stand-alone application)으로 사용되던 자동번역시스템은 최근에는 문서의 생성에서부터 번역, 검색 등을 하나의 시스템 내에서 가능하게 하는 통합시스템의 일부로서 사용되고 있다.

본 논문에서는 자연언어처리(natural language processing)의 가장 핵심적인 분야인 다국어 자동번역의 기술 개발 동향에 대해 논의한다. Ⅱ장에서는 다국어 자동번역시스템 개발을 위한 다양한 방법론을 소개한다. 또한 이러한 방법론을 통해 개발된 자동번역시스템의 다양한 응용분야에 대해서도 소개한다. Ⅲ장에서는 국내 및 국외의 자동번역 기술개발 현황에 대해 소개한다. Ⅳ장에서는 다국어 자동번역 시장 현황에 대해 살펴보고, 끝으로 Ⅴ장에서는 향후 다국어 자동번역 기술 개발의 방향을 가늠해보고자 한다.

Ⅱ. 다국어 자동번역 기술

1. 다국어 자동번역 기술 개발의 역사

다국어 자동번역 기술은 2차 대전이 끝난 후 미국과 소련에 의해 주로 군사적인 목적을 위해 본격적으로 개발되었다. 양국의 군사정보를 수집하기 위한 목적으로 영어와 러시아어간의 자동번역 기술이 개발되기 시작하였다. 주로 양국 국방성의 펀딩을 통해 개발되어온 자동번역 기술은 1966년의 ALPAC Report에 의해 큰 위기를 맞게 된다. 이 보고서는 "자동번역 기술은 결코 인간의 번역능력을 따라 잡을 수 없으며, 인간 번역가에 의해 결국 다시 재작업이 되어야 하므로, 궁극적으로는 인간 번역가에 의해 번역하는 것이 이득이 된다"라는 요지를 담고 있다. 미 국방성에 의한 이 보고서의 영향으로 한동안 자동번역 기술 개발은 큰 침체기를 맞게 된다.

다국어 자동번역 기술은 1980년대 중반부터 유럽과 일본에서 본격적으로 집중적인 연구개발이 재개되었다. 유럽연합 등의 이유로 일찍이 다국어 문화권을 이루던 유럽에서는 유럽언어들간의 자동번역에 대한 요구가 매우 높았다. 1982년부터 1993년까지 EC(유럽연맹)의 주도로 수행된 EUROTRA 프로젝트는 유럽 9개국 언어(영어, 독일어, 프랑스어, 스페인어, 포르투칼어, 이탈리아어, 덴마크어, 네덜란드어, 그리스어)간의 상호 자동번역을 위한 시스템 개발을 목적으로 했다.

이미 1980년대부터 다국어 문서 생성에 대한 수요가 엄청나던 일본의 경우, 주로 Toshiba, Fujitsu 등과 같은 기업들에 의해 일-영, 영-일 번역쌍을 중심으로 자동번역시스템 연구 개발이 추진되었다. 일본 문부성의 지원으로 오랜 기간 동안 끊임없이 자동 번역을 위한 기초기술 연구를 하고 있는 일본은 현재 자동 번역 분야에서 미국과 함께 세계적인 기술 수준을 보유하고 있다.

독일의 경우 1993년부터 독일정부의 지원으로 독어-영어-일어 간의 자동통역 시스템 개발을 목표로 하는 야심찬 Verbmobil 프로젝트를 추진했다. 약 10년간 독일정부의 지속적인 지원으로 진행된 이 프로젝트의 결과로 독일 연구진은 대화처리, 의미분석 및 구어체 자동번역 분야에서 세계적인 기술을 보유하게 되었다.

일본과 유럽에 비해 다국어 자동번역분야 기술개

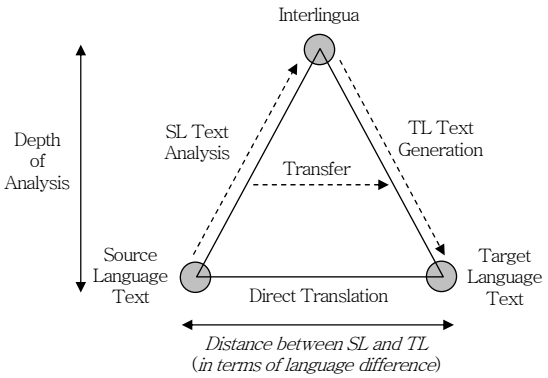
발에 대해 상대적으로 소극적이던 미국의 경우에는 주로 미 국방성의 연구비 지원으로 군사적 이해관계가 얽혀 있는 한국어, 아랍어, 러시아어 등에 대한 연구가 이루어졌다. 펜실베니아 대학의 한-영 자동번역시스템은 한국어 군사작전 문서의 번역을 위해 개발되기도 했다.

2. 다국어 자동번역 방법론

다국어간의 자동번역을 위한 방법론은 여러 가지로 분류가 가능하지만, 본 논문에서는 번역지식을 어디로부터 가져오느냐에 따라 합리주의적 방법론과 경험주의적 방법론으로 나누기로 한다. 합리주의적 방법론에서는 언어학자 혹은 번역가 등이 번역에 사용되는 지식을 직접 자신의 언어능력을 반영하여 작성한다. 이에 반해 경험주의적 방법론에서는 주관적일 수 있는 인간의 언어능력에 직접 의존하기보다는, 인간 세상에 존재하는 말뭉치, 즉 코퍼스(corpus)로부터 번역지식을 학습하는 방식을 택한다. 일반적으로 합리주의적 방법론을 따르는 기법에서는 단기간 내에 일정 수준의 번역 성능을 올릴 수 있는 반면에, 번역지식을 구축하기 위한 시간과 비용이 많이 드는 단점이 있다. 반면 경험주의적 방법론을 따르는 기법의 경우에는, 대개의 경우 기계 학습 등의 방법을 통해 적은 비용으로 번역지식을 얻어낼 수 있지만 학습 코퍼스의 분야에 매우 의존적인(domain-dependent) 문제점들을 가지고 있다. 다음의 절에서는 각 기법들에 대해 좀 더 상세히 알아보기로 한다.

가. 합리주의적 방법론(Rational Approach)

합리주의적 방법론은 인간의 언어능력, 즉 번역능력을 모델링하여 시뮬레이션하는 방식으로 번역을 시도한다. 일반적으로 번역의 대상이 되는 출발언어(source language)로부터 분석과정, 변환과정, 생성과정 등을 거쳐 목표언어(target language)를 출력하게 된다. 합리주의적 방법론은 그 분석의 깊이에 따라 직접번역방식, 간접변환방식, 중간언어방식 등으로 세분할 수 있다. 합리주의적 방법론은 주로 규칙(rule)을 통해 번역을 시도하므로 규칙기반 방법론(rule-based approach)으로 불리기도 한다. Hutchins&Somers(1992)에서 발췌한 (그림 1)은 분석 깊이에 따른 직접번역방식, 간접변환방식, 중간언어방식의 차이를 명확하게 보여준다.



(그림 1) 자동번역 방법론

1) 직접번역방식(Direct Translation Approach)

직접번역방식에서는 입력문을 형태소 분석(morphological processing), 태깅 등의 과정을 통해 매우 낮은 단계에서 분석을 마친 후, 변환 사전(bilingual transfer dictionary) 등을 참조해 대역문장을 생성해낸다. 입력문의 구조나 의미를 파악하는 단계를 거치지 않고, 단순히 입력문의 스트링을 사전참조 등을 통해 형태소 단위로 쪼갠 후, 각 형태소 단위를 대역어의 해당 형태소 단위로 변환한다. 이 과정을 거친 후 대역어의 문법적 특성을 반영한 간단한 어순 조작 규칙 등을 통하여 대역문을 생성해내게 된다. 이 기법은 초창기 기계번역 시스템에서 많이 사용되었으며, 최근에도 한국어와 일본어, 스페인어와 이탈리아어 등과 같이 언어학적으로 유사한 언어쌍에 대해 많이 사용되고 있다. 이 방법론은 깊은 분석 단계를 거치지 않으므로 처리 속도가 빠른 장점이 있으나, 어족이 다른 언어간에는 적용하기 어려운 단점을 가지고 있어, 영어와 한국어 등과 같이 서로 상이한 언어쌍에 대해서는 적용될 수 없다. 직접번역방식을 채택한 대표적인 시스템으로는

Georgetown 시스템과 초창기의 SYSTRAN 시스템을 들 수 있다[1],[2].

2) 간접변환방식(Indirect Transfer Approach)

직접번역방식이 입력문에 대해 형태소 분석 단계만을 거치는 반면, 간접변환방식에서는 형태소 분석을 거쳐 통사구조(syntactic structure), 의미구조(semantic structure)에 대한 분석을 더 거친 후 목표언어로의 변환(transfer)을 하며, 이 변환된 구조로부터 대역 문장을 생성하게 된다. 변환을 통사단계에서 하느냐 의미구조에서 하느냐에 따라 변환 규칙의 복잡성도 달라지게 된다. 일반적으로 통사구조는 사용하는 문법에 따라 다르게 표상화(representation)되는데, 주로 구구조문법(phrase structure grammar)이나 의존문법(dependency grammar) 등이 통사표상을 위해 사용된다. 의미구조는 통사적으로 분석된 문장의 의미구조를 나타내는 구조로서, 일반적으로 술어논리구조(predicate argument structure)가 표상으로 사용된다.

이 방식에서 번역지식은 변환규칙의 형태로 표현된다. 변환단계에서 입력문장의 인터페이스 구조는 목표문장의 인터페이스 구조로 변환되게 된다. 통사구조기반 변환방식은 의미구조기반 변환방식에 비해 분석을 위한 비용과 시간은 덜 소모되나, 변환규칙이 복잡해지는 단점이 있다. 이에 반해 의미구조기반 변환방식은 통사구조기반 변환방식에 비해 변환 규칙은 비교적 단순하나, 변환을 위한 인터페이스 구조를 도출해내기 까지의 비용과 노력이 많이 들 수 있으며, 그만큼 에러발생의 가능성도 높아지는 단점이 있다.

그러나 이 방식은 비교적 개발이 용이하고, 소수의 규칙만을 구축하더라도 비교적 높은 성능을 낼 수 있으므로, 현재 국내외에서 상용화되어 판매되고 있는 대부분의 다국어 자동번역시스템에 채택되고 있다. 간접변환방식을 채택한 대표적인 기계번역시스템으로는 캐나다 몬트리올 대학의 TAUM-시스템, 독일 자르브뤼켄 대학의 SUSY 시스템, 미국 텍사스 대학의 METAL 시스템 등을 들 수 있다.

이 방식의 대표적인 문제점으로 들 수 있는 것은 분석, 변환, 생성에서 많은 수의 규칙에 의존하므로 규칙 수가 많아질수록 규칙간의 충돌이 생기는 경우가 많다. 또한 많은 수의 규칙을 효과적으로 관리하는 문제도 발생할 수 있다.

또한 이 방식은 다국어 기계번역 시스템을 구축하기에 결정적인 단점을 지니고 있다. 즉, n개의 언어를 동시에 처리하는 다국어 기계번역 시스템을 구축하기 위해, n(n-1)개의 변환모듈이 필요하다. 한국어, 중국어, 영어 간의 번역을 수행하는 다국어 기계번역 시스템을 개발한다고 할 때, 한-중, 중-한, 한-영, 영-한, 중-영, 영-중, 총 6개의 변환모듈이 필요하게 된다. 물론 두 언어간의 양방향 번역이 가능하게 하는 변환규칙을 가지고 있다면 n(n-1)/2개의 변환규칙만이 필요하겠지만, 실제 기계번역 시스템을 개발할 때 양방향 모두에 적용될 수 있는 변환규칙을 개발하는 것은 매우 어렵다.

3) 중간언어방식(Interlingua Approach)

간접변환방식이 다국어 자동번역을 위해 다수 개의 변환모듈이 필요한 단점을 가지고 있는 반면, 중간언어방식을 도입하면 이러한 문제를 해결할 수 있다. 중간언어방식에서는 개별언어독립적인 의미표상(language-independent semantic representation)을 도입하고, 입력문을 분석단계를 거쳐 이 언어독립적인 의미표상으로 매핑한다. 따라서 n개의 언어를 번역하기 위해서는 단지 각 언어로부터 언어독립적인 의미표상(다른 말로 interlingua representation이라고도 부름)을 도출해 낼 수 있는 n개의 분석모듈과, 이 언어독립적인 의미표상으로부터 목표언어를 생성해 낼 수 있는 n개의 생성모듈만이 필요하다. 따라서 다수 개의 변환모듈이 필요한 간접변환방식과는 달리, 중간언어방식은 단지 개별언어로부터 중간언어로의 매핑을 위한 분석모듈, 중간언어로부터 목표언어를 생성하기 위한 생성모듈만이 필요하다. 따라서 이 방식은 다국어 자동번역에 적합하다고 할 수 있다.

중간언어방식에서 사용하는 의미표상은 자연언어에서 가능한 모든 의미를 표상할 수 있는 구조적인 장치와 어휘들을 가지고 있어야 한다. 이 방법론의 성공여부는 자연언어의 모든 의미들을 어떠한 방법으로 표상할 것인가에 달려 있다고 볼 수 있다. 일반적으로 의미를 표상하는 방법에는 다음과 같은 것들이 있다.

- 자연언어의 의미를 표현하기 위해 인공언어를 사용(예: TITUS-System)[3]
- 영어와 같은 자연언어를 차용하여 의미를 표현(예: UNL-System)[4]
- 자연언어의 의미를 더 이상 쪼갤 수 없는 단위(의미소)로 나누어 표현(예: UNITRAN-System)[5]

이 방식의 가장 큰 단점으로는 언어독립적인 의미표상을 정의하는 것이 매우 어렵다는 점이다. 예를 들어 한국어의 '김장'이라는 단어는 한국과 다른 문화권의 언어, 즉, 예를 들어 영어나 독일어 등과 같은 다른 외국어에서는 정확하게 대응되는 개념이 없다. 따라서 언어보편적인 의미표상을 구축하려고 할 때 이러한 어휘를 누락하게 되는 경우가 발생하는 문제가 있다.

나. 경험주의적 방법론(Empirical Approach)

1) 예제기반 방법론(Example-based Approach)

이 기계번역 방법론은 유추에 의한 번역(translation by analogy)이라고도 불리며, 1980년대 초반 일본 교토 대학의 나가오 교수(Prof. Nagao)에 의해 제안되었다. 이 방법론에서는 번역을 위한 지식을 번역가나 언어학자들의 언어능력을 통해 직접 코딩하는 것이 아니라, 대용량 말뭉치로부터 기계학습 등의 방법을 통해 번역지식을 반자동으로 학습하는 방식을 취한다.

이 방법론의 기본 아이디어는 수많은 번역쌍들을 데이터베이스에 저장한 후, 입력문이 들어 왔을 때 입력문과 가장 유사한 예문을 찾아, 예문의 번역을 참조하여 번역을 하는 것이다. 이 방법론의 장점은 대용량의 대역 코퍼스와 잘 정의된 시소러스가 있으면 어느 언어쌍에도 비교적 쉽게 적용할 수 있다는 점이다. 또한 예제 번역이 전문 번역가에 의해 이루어졌으면, 자동번역 결과도 비교적 높은 품질의 자연스러운 번역을 기대할 수 있다. 현재 마이크로소프트사에서는 이 방법론에 기반하여 영-중, 중-영 자동번역시스템을 개발하고 있다[6]. 그러나 이 방법론의 단점은 높은 성능을 내기 위해서는 대용량의 대역코퍼스가 필요한데, 많은 언어쌍의 경우 이것이 쉽지 않다는 점이다. 또 하나의 문제점은 대역코퍼스의 도메인에 따라 번역률이 많은 차이를 보인다는 점이다. 즉, 기계분야를 위해 사용된 대역코퍼스는 예를 들어 화학분야 문서의 적용에는 어렵다는 점이다.

2) 통계기반 방법론(Statistics-based Approach)

이 방법론은 1980년대 후반 미국 IBM Watson 연구소에서 처음 제안되었으며, 현재에도 여러 언어쌍에 대해 활발히 연구되고 있다. 이 방법론의 기본 아이디어는 단일어 코퍼스 또는 대역 코퍼스로부터 번역에 필요한 언어모델(language model)을 n-gram 방식 등을 통해 학습하는 데 있다. 이 번역 방법은 영어-프랑스어 등과 같이 유사한 언어간에는 어느 정도 성능을 보이나, 구조적으로 많은 차이를 보이는 한국어와 영어간의 경우에는 현재까지 개발된 기술 수준으로는 적용하기 어려운 점이 있다.

3. 응용분야

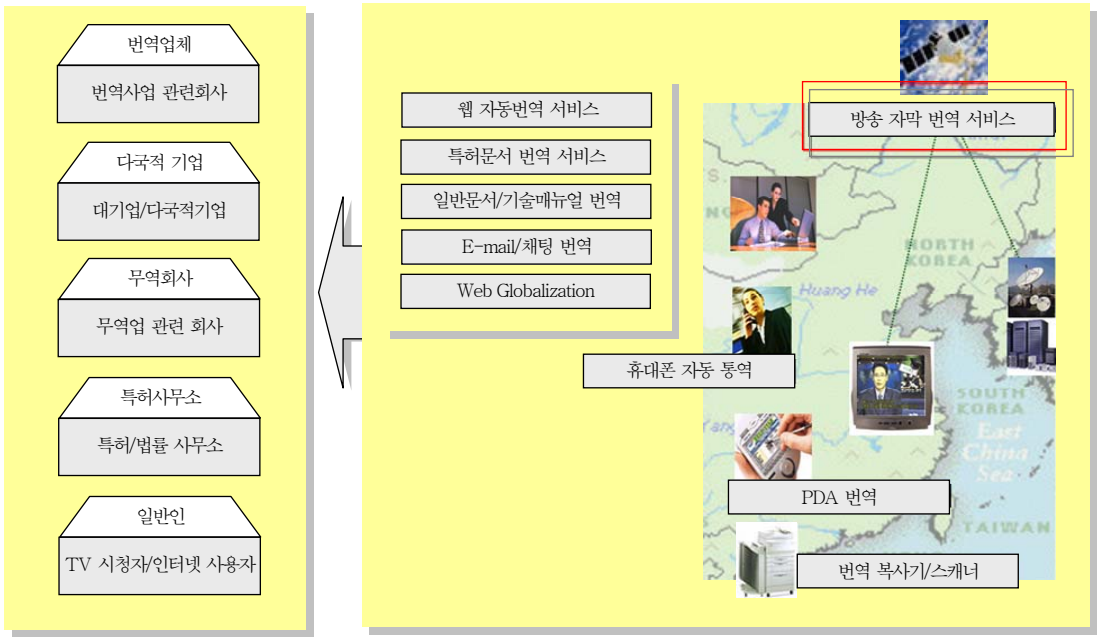
자동번역시스템은 개발초기에는 주로 독립적인 응용시스템으로 일반적인 목적을 위해 사용되었다. 그러나 일반 도메인에 적용할 경우 낮은 번역률로 인해 많은 사용자들로부터 외면되는 결과를 초래했다. 그러나 자동번역시스템은 한정된 분야에 적용할 경우 비교적 정확한 대역어 선택이 가능하므로, 상당히 유용한 번역결과를 생성해 낼 수 있다. 이러한 특성으로 인해 짧은 기간 내에 많은 양의 문서를 번역해야 하는 경우에 문서처리 프로세스의 한 부분으

로서 많이 사용되고 있다.

다국어 자동번역이 응용되는 분야를 요약해서 나타내면 (그림 2)와 같다.

일찍부터 다국어 번역에 대한 수요가 높았던 유럽의 경우 유럽 연합 내에서 공식적으로 발행되는 모든 문서에 대해 주요 회원국어로 번역되어야 하는 실정이다. 따라서 이러한 문서들의 번역만을 위한 기관이 따로 존재할 정도로 다국어 번역에 대한 필요성이 매우 높다. 유럽연합의 소속 기관인 유럽연방법원의 경우 매년 생성되는 문서가 33만 페이지인데, 이를 휴먼번역가가 번역을 할 경우 그 시간과 비용을 도저히 감당해 낼 수 없는 문제가 있다. 따라서 유럽연방법원은 이러한 문제를 해결하기 위해 Systran사에서 개발한 다국어 자동번역시스템을 모든 문서의 초벌 번역을 위해 사용하고 있다. 다국어 자동번역시스템에 의해 초벌 번역된 문서들은 전문가들에 의해 교정과정을 거치게 되나, 이러한 작업 프로세스의 비용과 시간이 전적으로 휴먼번역가들에 의해 번역을 하는 경우보다 월등히 유리하므로 자동번역시스템을 유용하게 사용하고 있다고 볼 수 있다. 이러한 상황은 다국어 사회인 유럽에만 국한되지는 않는다. 세계경제가 점차 블록화되는 양상을 보이는 가운데, 한국과 일본, 중국 특허청에서는 2006년부터 자국에서 출원되는 모든 특허문서에 대해 영어 번역을 제공함으로써 상호 특허 인정을 추진하고 있다. 이에 대비해 일본의 경우 일본 내에서 출원된 모든 특허문서를 외국인 특허 검색자들을 위해 영어로 제공하는 서비스를 이미 실시하고 있다. 특히 유럽과 미국의 특허 심사관들에 의해 호평받고 있는 일-영 특허번역시스템은 일본 도시바사의 엔진을 탑재하고 있는데 심사관들조차 만족할 정도의 상당한 번역수준을 보이고 있다. 한국 특허청에서도 현재 일본 특허청과 마찬가지로 한국에서 출원된 모든 특허문서들에 대한 영어번역제공 서비스를 준비하고 있다.

인터넷의 발달과 더불어 비즈니스의 영역에도 시간적, 공간적 제약이 없어졌다. 이러한 시대적 변화는 다국어 자동번역시스템의 응용분야에도 획기적인 변화를 초래하고 있다. 예를 들어 전세계를 상대로 비즈니스를 하는 기업의 경우 그 기업에서 생산



(그림 2) 다국어 자동번역 응용 분야

하는 모든 제품에 대한 카탈로그를 웹상에서 제공하게 될 것이다. 거의 무한정으로 쇄도하는 사용자들의 질의, 불만, 요청 및 제품 설명에 대한 업데이트 등 다국어로 제공해야 할 콘텐츠는 실로 무한하다고 볼 수 있다. 이러한 콘텐츠에 대한 다국어 서비스를 휴먼번역가에 의해 제공한다는 것은 경제적인 관점에서 볼 때 거의 불가능한 일이라고 할 수 있다. 따라서 이러한 web globalization에서의 다국어 자동번역서비스에 대한 수요 또한 매우 크다고 할 수 있겠다.

그 밖에 위성통신 및 인터넷 방송 등의 발달과 더불어 방송자막에 대한 실시간 자동번역 기술도 다국어 자동번역시스템의 주요 응용분야라 할 수 있겠다. 실제로 ETRI에서는 이미 2001년부터 한국어-중국어, 영어-한국어 뉴스방송자막 자동번역시스템을 개발하고 있다.

Ⅲ. 연구 및 개발 현황

본 장에서는 다국어 자동번역 기술과 관련한 국내외의 연구 및 개발 현황에 대해 기술한다.

1. 국외 현황

가. 텍스트 및 웹 자동번역

다국어 자동번역의 실용화라는 측면에서 다국어 텍스트 및 웹 자동번역 개발은 자동번역 기술 초기부터 가장 실용화에 근접하여 개발되었던 분야이다.

1) 미국

대표적인 프로젝트로서 CoGenTex 프로젝트가 있으며, 이 프로젝트는 미국방성과 Pennsylvania 대학, Systran Software, Inc.,의 합작 하에 추진되었던 프로젝트이다. 상용화와 관련해 미국에서 자동번역 핵심기술을 이용해 Marine Acoustics사는 영어-Urdu(아프카니스탄어)간의 500문장에 대한 자동번역 소프트웨어를 장착한 PDA “Phraselator”를 개발하여 아프칸 주둔 미군에 공급한 바 있다.

2) 일본

일본은 자국어 처리 및 자동번역 기술을 국책연구기관인 CRL, ATR에서 국책 프로젝트로서 연구개발하고 있다. 기업에서 개발한 다국어 자동번역시스템을 살펴보면, IBM Japan, ATR 등이 영일/일영, 중일/일중 자동번역시스템을 실용화 수준으로 개발하고 있고, NHK를 중심으로 방송자막 자동번역 기술을 개발하고 있다. 또한 NEC Corporation의 Bestiland는 한국어를 포함하여 10개 국어를 대상으로 다국어 자동번역 기술 개발중이며 NTT는 말레이시아어를 포함하는 다국어 기계번역시스템을 개발중이다. CICC 프로젝트에서는 일본어, 태국어, 말레이시아어, 인도네시아어, 중국어와 같은 동양권 언어에 대한 다국어 번역시스템 개발을 추진중에 있다.

3) 유럽

유럽은 유럽 공동체의 특수성으로 인해 가장 다국어 자동번역이 발달되어 있는 지역이다. 특히 프랑스에 본부를 두고 있는 Systran Software는 세계적으로 가장 잘 알려져 있는 다국어 자동번역시스템 개발회사이다. 본 절에서는 Systran Software에 대해 기술함으로써 유럽의 다국어 자동번역시스템 개발 현황을 대신하고자 한다.

- Systran Software

동종언어간 자동번역을 위해 7개 유럽언어(영어, 이탈리아어, 포르투칼어, 불어, 독일어, 스페인어, 러시아어)간의 다국어 자동번역을 중점적으로 시도하고 있으며 기타 언어(그리스어, 네덜란드어 등)를 포함하는 시도가 진행중이다. 이종언어간 자동번역을 위해 아시아권 언어 중 영중, 영일, 영한을 시도하고 있으나 언어유형의 차이로 인해 초벌번역을 목표로 하고 있다. 번역률은 영불 90.5%, 불영 86.3%, 영독 81.4%, 독영 80.0%, 영스페인어 73.8%, 스페인어영 68.9%, 러영 95.4% 정도이다[7].

나. 특허 문서 자동번역

각국 특허청은 특허출원의 급증으로 심사기간의 장기화, 심사의 질 저하, 신속한 권리확보 미흡 등의 문제에 직면하면서 지속적으로 증가하는 지재권 출원을 적정하게 처리할 수 없게 됨으로써 각국 특허청은 공동출원에 대한 '심사결과 상호인정제도'를 도입키로 1997년에 합의하고 자동번역시스템을 활용하기 시작하였다. 유럽에서는 1993년부터 유럽 특허청을 중심으로 특허 자동번역기를 활용하여 자동번역하고 있으며(예: "PaTrans"는 영어 특허 문서의 75% 정도를 덴마크어로 자동번역하고 있음), 이탈리아에서는 1987년부터 6개 국어(영어, 독어, 불어, 스페인어, 포르투칼어, 이탈리아어)간의 특허 자동번역 서비스를 하고 있다. 특히 일본 특허청에서는 2003년 3월부터 일영 자동번역 기술을 이용하여 인터넷상에서 일어 원문을 영문으로 서비스하고 있다.

다. CAT

CAT는 번역가의 번역을 지원하는 도구를 말한다. CAT의 출현은 번역 메모리(translation memory)의 출현과 밀접한 관련이 있다. 번역 메모리는 대상 구문과 목적 구문의 쌍을 데이터화 한 것을 말하는데, 번역 메모리 시스템은 실제 번역과정이 이루어지고 있는 문서 처리 상에서 번역 대상 구문의 문자열을 비교하여 유사 구문을 판정해내는 것이다. 일반적으로 문자열의 비교는 어휘단위의 문자열의 유사성과 연속성을 비교하는 퍼지 매칭(fuzzy matching) 알고리듬을 사용하며, 유사 구문으로 선택된 대상 구문은 자신과 쌍으로 있는 목적 구문과 함께 비교 대상 구문과의 문자열의 차이가 표시된다. 이를 통하여 번역가는 지금 번역할 구문이 기존에 번역된 유사한 구문과 어떻게 다른지, 기존의 번역 결과가 어떠했는지를 참고할 수 있게 되는 것이다. 이렇게 번역 메모리는 번역물의 자산화, 또는 번역 자산의 재활용을 통한 번역 생산성 향상을 목표로 개발되었다[8].

현재 Atril(Deja Vu), IBM(Translation Manager), SDL(SDLX), Star(Transit) and Trados(Translator's Workbench) 등의 응용 프로그램이 개발되어 있으며, 각 응용 프로그램들은 데이터 표준안 TMX에 대한 지원을 통하여 응용 프로그램간의 데이터 교환을 지원하고 있다. 현재는 네트워크 지원과 자동 번역 지원, XML을 이용한 데이터 처리 등 많은 시도들이 이루어지고 있어서 CAT의 개념이 확장되어 가고 있는 상태이다.

라. Web Globalization

Web globalization은 글로벌 커뮤니티에 더욱 수월하게 접근할 수 있도록 해당 웹사이트를 개량하고 유지하는 절차를 말한다. Web globalization이 인터넷에서 가장 큰 사업으로 떠오른 것은 인터넷으로 인하여 세계적 규모의 정보 공유가 가능해졌지만, 외국어에 능숙한 사람이라도 동일한 내용에 대해서는 외국 사이트보다는 자국어 사이트에 두 배 이상 머무르며, 구매할 확률은 3배 이상 높은 것으로 나타나기 때문에 국제 시장에 대응하기 위해 중앙 관리되는 다국어 웹 사이트를 구축해야 하는 엄청난 도전에 직면하게 되었기 때문이다.

2000년 이후 web globalization이 급격히 구현되어 유럽을 중심으로 활발히 이용되고 있고, 세부적으로 개선된 기술이 계속 개발되고 있다. 미국에서는 B2B 업체의 80%가 2004년에 웹 세계화를 추진하였으며, 브루킹스 연구소에 따르면, 미국 기업들은 인터넷의 효율성을 이용함으로써 지난 3년간 모두 1천550억 달러를 절약하여 이익이 4천439억 달러 늘어났으며, 2010년까지 계획된 인터넷 투자가 완료될 경우 기업들은 인터넷을 통한 추가적인 효율성 증대로 5천283억 달러를 절감할 것이며 이익은 1조5천500억 달러 급증할 것으로 예측했다.

마. 표준화

1) 번역메모리 표준화

현재 IBM, Netscape, Adobe, Cisco, 3Com 등의 출판, 소프트웨어, 하드웨어 및 로컬라이제이션

산업과 관련해 약 220 선도기업을 회원으로 보유한 단체인 LISA 산하 OSCAR에서 번역 메모리 표준화 작업을 추진하고 있다.

2) 말뭉치 표준화

말뭉치의 표준화는 연구 그룹별로 대규모 프로젝트를 형성하여 활발히 진행되고 있다. 말뭉치간의 호환성 문제에 대하여 말뭉치 작성 규범의 설정 활동(TEI)이 1987년부터 학회 수준에서 진행되고 있고, TEI에서는 크게 두 단계에 걸쳐 표준안이 도출되었는데, 1988년에서 1990년까지의 첫번째 개발 사이클 과정에서는 ACH, ACL, ALLC으로 구성된 조종위원회를 중심으로 프로젝트를 추진해 나가면서 첫번째 표준화 초안을 발표하였고, 두번째 개발 사이클은 1990년에서 1994년까지 진행되었는데 15여 개의 전문가 그룹을 중심으로 첫번째 표준화 초안에 대한 개선 작업에 착수하여 두번째 표준화 초안을 발표하였으며 이후 개선 작업은 계속되어 1994년 5월에 첫번째 공식적인 표준안을 공표하였다. EAGLES은 텍스트 말뭉치와 전자사전, 음성 DB와 같은 대규모 언어 자원의 표준화와 둘째, 이러한 언어 자원을 언어 포말리즘(formalism)이나 마크업 언어(markup language), 다양한 도구 등을 이용하여 관리하는 방법론에 대한 표준화, 셋째로 언어 자원, 도구, 응용 제품 등을 평가하는 방법론에 대한 표준화를 추진하고 있다.

3) 전자사전 표준화

전자사전의 표준화도 연구 그룹별로 대규모 프로젝트를 형성하여 활발히 진행되고 있다. GENELEX 프로젝트는 기존에 개발되어 있는 전자사전의 재사용을 가능하게 하기 위해 표준안을 정의하였고, GRAAL 프로젝트에서는 문법을 개발함에 있어서 여러 대학의 연구 결과를 수용하여 표준안을 정의하였다. 1992년 미국 내에서 처음 결성된 LDC는 공동 작업 기술을 적용시켜 서로 다른 문제 해결에 공통적으로 필요한 언어 자원을 공유하고 나아가 다른 나라의 연구자들에게도 배포함으로써 세계적인 컨소시엄으로 확대되고 있다. ISO TC37 등 전세계 25개국의 참여로 전문용어와 언어 리소스의 원리, 방법, 응용에 대한 표준화를 추진하고 있다.

4) 언어처리 응용 기반 기술 표준화

언어처리 응용 기반 기술과 관련하여 형태소 분석 API 기술(MorphOlympic), 구문 분석 API 기술(유럽의 프랑스, 이탈리아, 독일, 스위스, 영국 등)이 다국어 언어처리 평가 테스트 세트 마련을 위한 프로젝트인 TSNLP에서 국제 공동연구로 수행되었으며, 의미 분석 API 기술 평가 대회(SENSEVAL)가 개최되었다.

2. 국내 현황

가. 텍스트 및 웹 자동번역

개발 초기부터 현재까지 한국어와 관련해 가장 실용화에 근접한 대상 언어는 언어학적 유사성으로 인해 일본어이며, 일한/한일 자동번역시스템은 번역률 90% 이상으로 웹문서를 대상으로 상품화되어 활발히 사용되고 있다. 영어의 중요성으로 인해 개발된 영한/한영 자동번역 기술은 일반분야에 대해 이종언어간의 언어적인 차이로 인해 약 60~70% 정도의 번역률을 내고 있으나 사용자에 의해 적극적으로 사용되지 못하는 실정이며, 한중/중한은 중국 내에서의 한류 열풍으로 인해 2002년도부터 국내에서 개발되기 시작하였다. 텍스트 및 웹 자동번역의 개발은 그 자동번역의 핵심기술을 이용해 2002년도에 일한/한일 번역 업체를 중심으로 PDA 탑재 번역 제품(창신소프트의 「이지토키2002」, 시스메타의 「포켓트랜스위즈」 등)이 출시되었으며, 향후 영어 및 중국어 등에 대한 PDA 다국어 번역 서비스가 제공될 것으로 기대되고 있다.

나. 특허 문서 자동번역

국내에서의 특허 문서 자동번역은 2004년 5월에 특허청과 한국전자통신연구원(ETRI)과의 업무협약

조인식에 따라 특허문서에 특화된 한영 자동번역시스템을 개발하기 시작하였으며, 대량의 전문용어 DB 구축 및 전문 분야에 필요한 문형 처리 기술을 이룩한 후에, 특허청은 ETRI가 개발한 시스템을 이용하여 2005년도에 전기전자분야 특허정보에 대해 시범서비스를 실시할 예정이며 2006년에 전체 기술분야로 확대하여 특허정보 영문서비스를 본격적으로 시작할 계획이다. ETRI의 한영 특허 문서 자동번역시스템 개발과 더불어 일본의 영일/일영 특허 번역기를 한글화한 제품이 국내 시장에 2004년도에 출시되기도 하였다.

다. CAT

ETRI, 클릭큐, 삼성전자, 모비코 등의 연구소와 기업들이 공동으로 한/영/일/중 언어에 대한 TM 기반의 통합 CAT 시스템 개발을 한 바 있으며, 이 연구를 토대로 클릭큐에서는 기업의 매뉴얼을 대상으로 한 영어, 일본어, 중국어, 러시아어, 이탈리아어 5개 언어의 CAT를 시장에 출시한 상태이다.

라. Web Globalization

세계화로의 급속한 변화에 따라 국내에서는 중소기업 수준에서도 세계로 진출하여 글로벌화 하는 추세에 있다. 그러나, 앞서 언급한 바와 같이 국내의 자동번역 기술이 번역률 70% 정도에 머물러 기업 콘텐츠에 적용하기에 어려움이 있어, 현재까지는 기업의 글로벌 콘텐츠는 휴먼 번역이 주류를 이루고 있다. 따라서 온라인/실시간으로 발생하는 대량의 기업 콘텐츠를 글로벌화(다국어화) 하기 위해서는 고품질의 번역률을 보장하는 자동번역 기술의 개발 및 이를 이용한 web globalization 시스템 개발이 시급한 실정이다.

마. 표준화

1) 한국어 품사 태그 세트 표준화

한국어 품사 태그 세트 및 표기법과 관련하여 한국전자통신연구원을 중심으로 전산 언어학자, 국문학자들로 이루어진 관련 연구자들이 모여 1998년부터 1999년까지 2년 동안 품사 태그 세트 선정 표준화 작업을 한 바 있다.

2) 전자사전 표준화

전자사전의 종류, 종류별 구조 및 내용 표기 방법(전문용어 사전 포함)과 관련해서는 문화관광부와 국립국어연구원에서 추진중인 '세종계획'에서 영역별 전문용어에 대한 표제어 선정 기준 및 수록할 정보의 종류와 표현 양식의 표준안이 추진되고 있다. 또한 KORTERM에서 전문용어와 기타 언어 자원에 대한 표준화 그룹인 ISO/TC37 분과에서 한국어 언어자원에 대한 표준화가 추진되고 있다.

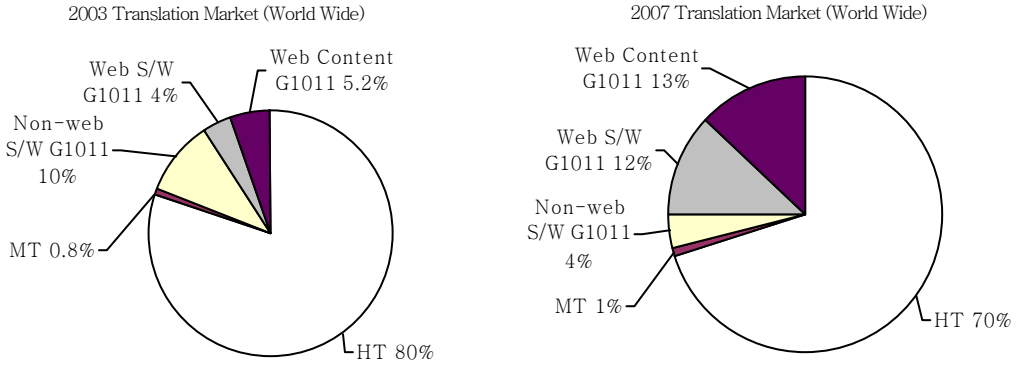
Ⅳ. 시장 현황

1. 다국어 번역 세계시장 규모

다국어 번역의 세계시장은 세계화로의 빠른 변화에 따라 연평균 10% 이상의 고성장을 보이고 있으며 2007년에는 120억 달러의 시장규모로 성장할 것으로 예측된다. 전체 시장규모와 관련해 HT가 차지하는 비율이 점차 줄어들고 소프트웨어를 이용한 자동번역이 늘어나는 추세이며, 패키지 소프트웨어 형태의 자동번역보다는 인터넷을 이용한 웹 세계화의 시장이 연평균 40%의 폭발적 증가세를 보이고 있다((그림 3) 참조).

2. 다국어 번역 국내시장 규모

다국어 번역의 국내시장은 세계화로의 빠른 변화에 따라 연평균 10% 이상의 고성장을 보이고 있으며 2007년에는 3,500억 원의 시장규모를 예측하고 있다. 세계시장과 동일하게 전체에서 HT이 차지하는 비율이 점차 줄어들고 소프트웨어를 이용한 자동번역이 늘어나는 추세이며, 패키지 소프트웨어 형태의 자동번역보다는 인터넷을 이용한 웹 세계화의 시

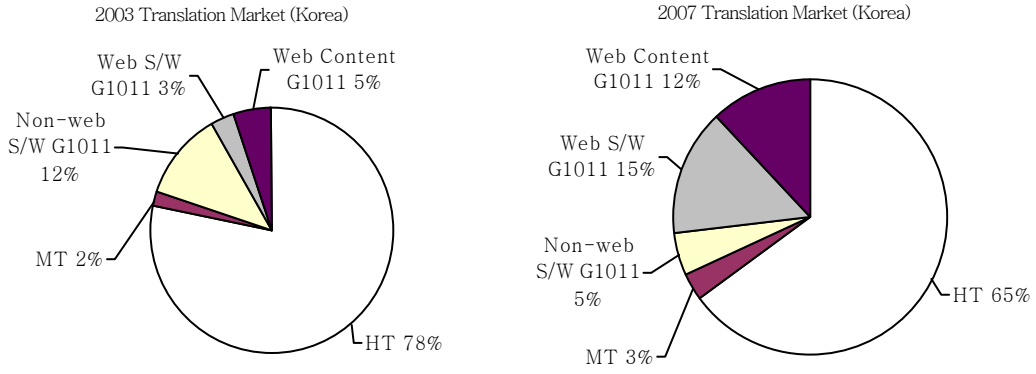


(단위: 십억 달러, %)

	Human Translation	Machine Translation Package	Software Globalization		Web Content Globalization	Total
			None-web Based	Web Based		
2003년	7.6(80)	0.076(0.8)	0.95(10)	0.38(4)	0.49(5.2)	9.5
2007년	9.1(70)	0.13(1)	0.52(4)	1.56(12)	1.69(13)	13

<자료>: ABI

(그림 3) 다국어 번역 세계시장 규모



(단위: 억 원, %)

	Human Translation	Machine Translation Package	Software Globalization		Web Content Localization	Total
			None-web Based	Web Based		
2003년	1872(78)	48(2)	288(12)	72(3)	120(5)	2400
2007년	2275(65)	105(3)	175(5)	525(15)	420(12)	3500

<자료>: 정보통신산업협회, 1994.

(그림 4) 다국어 번역 국내시장 규모

장이 연평균 50%의 폭발적 증가세를 보이고 있다 ((그림 4) 참조).

V. 결론

인간의 언어장벽을 컴퓨터를 이용해 해결하려는 노력은 1950년대 이후로 언 반세기가 넘도록 이어지고 있다. 자동번역이 인간의 언어장벽을 해결할 것이라는 희망이 인간의 인지능력이 내포되어 있는 언어의 난이성 때문에 좌절과 부흥을 맛보면서, 이제 자동번역 기술은 21세기에 들어 실용화에 초점을 두고 텍스트 및 웹 자동번역, 특허문서 자동번역, 전문번역가 지원도구(CAT), web globalization 그리고 방송자막 자동번역, 모바일 자동번역 등에서

그 결실을 보이고 있다.

특히 영어 및 중국어의 필요성이 더욱 증가 추세에 있는 한국의 입장에서는 언어 소외 계층을 더욱 양산하는 결과를 맺을 수 있으므로, 이를 해소하기 위해서는 다음과 같은 IT839와의 연계가 무엇보다 시급하다고 할 수 있다.

첫째로는 기존의 위성 및 지상파 DMB에서의 콘텐츠 관리 운영측면을 들 수 있다. 해외 콘텐츠를 관리하고 고객에게 제공하기 위해 콘텐츠 번역 및 세계화와 관련된 다국어 처리 기술이 뒷받침 되어야 한다.

둘째로는 W-CDMA 서비스에서의 휴대형 다국어 자동통역 서비스를 생각할 수 있다. W-CDMA 환경은 사용자의 소규모 단말을 통해 고속의 음성 및 데이터 통신을 가능케 한다. 향후 W-CDMA 서비스에서는 이러한 인프라에 기반해, 성능이 매우 향상된 고품질 다국어 자동통역 서버와 연동, 다국어 자동통역 서비스가 실현될 것이다.

마지막으로 소프트웨어 솔루션 및 디지털 콘텐츠에서의 디지털 콘텐츠 세계화 측면을 생각할 수 있다. 디지털 콘텐츠가 수출되기 위해서는 콘텐츠의 기획 구성의 단계로부터 세계화를 염두에 두어야 한다. 세계화를 위해서는 문화적 관점에서 뿐만 아니라, 언어처리 부분에서의 세계화가 핵심이 되며, 특히 번역을 포함한 다국어 콘텐츠 관리 운영을 위한 다국어 처리 기술의 발전을 필요로 한다.

약어 정리

ACH	Association for Computers and the Humanities
ACL	Association for Computational Linguistics
ALLC	Association for Literary and Linguistic Computing
ATR	Advanced Telecommunications Research
CAT	Computer-Aided Translation
CICC	Computer of the International Cooperation for Computerization
CRL	Communications Research Laboratory
DMB	Digital Multimedia Broadcasting
EAGLES	Expert Advisory Group on Language Engineering Standards
GENELEX	GENEric LEXion
GRAAL	Grammars which are Re-usable to Automatically Analyze Languages
HT	Human Translation
ISO	International Organization for Standardization
LDC	Linguistic Data Consortium
PDA	Personal Digital Assistance
TC	Terminology and other language resources
TEI	Text Encoding Initiative
TMX	Translation Memory eXchange
W-CDMA	Wideband Code Division Multiple Access

참 고 문 헌

[1] M. Kay, “Automatic Translation of Natural Languages,” Daedalus, Vol.102, No.3, 1973, pp.217-230.

[2] W.J. Hutchins and H.L. Somers, An Introduction to Machine Translation, Academic Press, London, 1992.

[3] W.J. Hutchins, Machine Translation - Past, Present, Puture, Ellis Horwood, Chichester/New York, 1986.

[4] M.P. Hong and O. Streiter, Overcoming the Language Barriers in the Web: The UNL-Approach, in Tagungsband der 11. Jahrestagung der Gesellschaft fuer Linguistische Datenverarbeitung(GLDV), Frankfurt am Main, Germany, 1999.

[5] B.J. Dorr, A View from the Lexicon, The MIT Press, Cambridge, Massachusetts, London, England, 1993.

[6] S. Richardson, W. Dolan, A. Menezes, and J. Pinkham, “Achieving Commercial-quality Translation with Example-based Methods,” *Proc. of MT Summit VIII*, Santiago De Compostela, Spain, 2001.

[7] Jane Mason, Adriane Rinsche, Translation Technology Products, Ovum 1995.

[8] 강명주, “번역메모리(Translation Memory) 기반 자동번역 및 번역지원(Computer Aided Translation) 시스템,” 정보처리학회지, 제11권 제2호, 2004, p.98-103.