자동번역 기술 동향 및 응용 사례

The Trends of Machine Translation Technology and Case Study

임베디드 S/W 기술 동향 특집

김 운 (Yun Jin) 언어처리연구팀 Post-Doc.
최승권 (S.K. Choi) 언어처리연구팀 책임연구원
김창현 (C.H. Kim) 언어처리연구팀 선임연구원
황영숙 (Y.S. Hwang) 언어처리연구팀 선임연구원
서영애 (Y.A. Seo) 언어처리연구팀 선임연구원
권오욱 (O.W. Kwon) 언어처리연구팀 선임연구원
김영길 (Y.K. Kim) 언어처리연구팀 팀장

목 차

최근 들어 인터넷 보급과 확산, 그리고 국제 교류가 심화됨에 따라 사람들이 단일언어의 장벽을 뛰어넘어 다른 언어 정보에 대한 수요가 급증하고 있는 추세이다. 또한 의사소통 방법도 기존의 면대면 대화나 편지 등으로부터 메신저, 이메일, 핸드폰 등으로 다양화되면서 자동 통/번역 기술을 통한 의사소통 및 타 언어 정보의 획득이 가능해지고 있으며 자동 통/번역 기술 또한 크게 각광받고 있다. 이에 대비하여 각국에서는 자동번역 기술을 국가주도 하에 경쟁적으로 진행하고 있으며, 그 응용도 웹 문서, 특허문서, 구어체 등으로 다양화되고 있다. 본 고에서는 자동번역 기술의 동향 및 국내외 응용사례를 소개하고 향후 자동번역 기술 개발의 방향을 점검해보고자 한다.

Ⅰ. 서론

자동번역(machine translation) 기술이란 언어장벽에 의한 의사소통 문제를 해결하기 위해 자연어처리기법을 이용하여 한 언어로부터 다른 언어로 변환해주는 기술을 말한다.

2차 대전 후 미국과 옛 소련에 의해 군사적인 목적으로 개발된 자동번역 기술은 1980년대 중반부터 유럽과 일본에 의해 다시 본격적인 연구가 시작되었다. 유럽은 다국어 문화권이지만 유럽 연합(EU)이라는 공동체의 특성상 언어장벽을 해소해야 하는 필요성으로 자동번역에 대한 수요가 늘었으며, 일본은 Toshiba, Fujitsu 등 기업들에 의해 자동번역시스템 연구개발을 활발히 추진해 왔다.

기존 자동번역시스템이 낮은 번역률로 인해 많은 사용자들로부터 외면당하였으나, 최근 들어 자동번역기술이 많이 향상되면서 Google, IBM, ISI 등의 자동번역시스템은 통합시스템의 일부로 실용화 수준까지 이르고 있다. 더욱이 특화된 분야에 적용하였을 경우, 그 성능이 일반 도메인에 적용했을 때보다 높아 텍스트 및 웹 문서 자동번역뿐만 아니라, 특허자동번역, 방송자막 및 채팅 번역과 같은 구어체 번역 등 다양한 분야의 많은 응용사례들이 속출하고 있다.

본 논문에서는 자동번역 기술동향과 그 응용사례에 대하여 자세히 알아본다. Ⅱ장에서는 자동번역 기술 및 그 응용시스템 개발에 사용되는 다양한 방법론에 대하여 알아보며, Ⅲ장에서는 웹 문서, 기술 문서, 구어체 등 분야별 국내외 자동번역시스템 응용사례에 대하여 알아본다. 끝으로 Ⅳ장에서는 향후 자동번역 기술 개발 방향을 가늠해보도록 한다.

Ⅱ. 자동번역 기술 동향

1. 개요

자동번역 기술은 자동번역이 이루어지는 방법에 따라 크게 2가지로 분류할 수 있다. 규칙기반 방법(rule-based approach)과 말뭉치기반 방법(corpus-based approach)이다. 규칙기반 방법은 언어학자 혹은 번역가 등이 번역에 사용되는 지식을 직접 자신의 언어능력을 반영하여 구축하는 언어학적 규칙(예: 구조분석 규칙, 변환 규칙 등)에 의해 자동번역이 이루어지는 반면, 말뭉치기반 방법은 주관적일 수 있는 인간의 언어능력에 직접 의존하기 보다는 인간 세상에 존재하는 말뭉치(예: 대량의 단일어, 대역어 문장)로부터 번역지식을 학습하여 자동번역이 이루어진다. 말뭉치기반 방법은 다시 예제기반 방법(example-based approach)과 통계기반 방법(statistics-based approach)으로 나누어 볼 수 있다.

각 방법은 그것의 장점을 수용하고 단점을 보완하기 위해 독자적 또는 서로 혼합되어 계속 진화되어 왔다. 각 방법의 동향은 그 방법이 나타난 시기로부터 볼 때, 1980년대까지는 규칙기반 방법이 주요 방법이었고, 1990년대에는 말뭉치기반 방법이 대세를 이루었으며, 2000년대에는 규칙기반과 말뭉치기반 방법이 독립 또는 공존하는 시기로 구분할 수 있다[1].

2. 규칙기반 자동번역

규칙기반 방법은 그 분석의 깊이에 따라 직접번역방식(direct translation approach), 간접변환방식(indirect transfer approach), 중간언어방식(interlingua approach) 등으로 세분할 수 있다.

직접번역방식에서는 입력문을 형태소 분석, 태깅(tagging) 등의 과정을 통해 매우 낮은 단계에서 분석을 마친 후, 변환 사전(bilingual transfer dictionary) 등을 참조해 대역문장을 생성해낸다. 이 기법은 초창기 자동번역 시스템에서 많이 사용되었으며, 최근에도 한국어와 일본어, 스페인어와 이탈리아어 등과 같이 언어학적으로 유사한 언어 쌍에 대해 많이 사용되고 있다.

간접변환방식에서는 형태소 분석을 거쳐 통사구조(syntactic structure), 의미구조(semantic structure)에 대한 분석을 더 거친 후 목표언어로의 변환

을 하며, 이 변환된 구조로부터 대역 문장을 생성하게 된다. 이 방식은 비교적 개발이 용이하고, 소수의 규칙만을 구축하더라도 비교적 높은 성능을 낼 수 있으므로, 현재 국내외에서 상용화되어 판매되고 있는 대부분의 자동번역 시스템에 채택되고 있다.

중간언어방식에서는 개별 언어 독립적인 의미표상(language-independent semantic representation)을 도입하고, 입력문을 분석 단계를 거쳐 이 언어독립적인 의미표상으로 매핑한다. 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.

3. 예제기반 자동번역

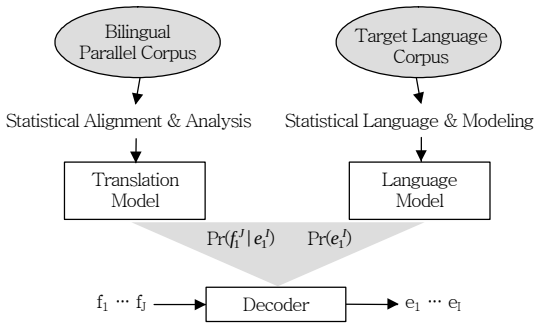
예제기반 방법은 유추에 의한 번역(translation by analogy)이라고도 불리며, 1980년대 초반 일본 교토 대학의 나가오 교수(Prof. Nagao)에 의해 제안되었다. 이 방법의 기본 아이디어는 수많은 번역쌍들을 데이터베이스에 저장한 후, 입력문이 들어왔을 때 입력문과 가장 유사한 예문을 찾아, 예문의 번역을 참조하여 번역을 하는 것이다. 이 방법의 장점은 대용량의 대역 코퍼스와 잘 정의된 시소러스가 있으면 어느 언어 쌍에도 비교적 쉽게 적용할 수 있다는 점이다. 그러나 이 방법의 단점은 높은 성능을 내기 위해서는 대용량의 대역코퍼스가 필요한데, 많은 언어 쌍의 경우 이것이 쉽지 않다는 점이다. 또 하나의 문제점은 대역 코퍼스의 도메인에 따라 번역률이 많은 차이를 보인다는 점이다. 즉, 기계분야를 위해 사용된 대역코퍼스는, 예를 들어 화학분야 문서의 적용에는 어렵다는 점이다.

4. 통계기반 자동번역

통계기반 자동번역(SMT) 기술은 이중언어 말뭉치로부터 통계적 분석을 통해 모델의 파라미터를 학습하고 그 모델에 근거하여 입력된 문장을 번역하는 기술이다. SMT 기술은 1949년 Warren Weaver [2]에 의해 소개된 이후, 1991년 IBM의 Thomas J. Watson 연구소 연구원들에 의해 다시 소개되면서부터 연구가 부활하여[3], 현재 가장 활발하게 연구되는 기계번역 기술이다.

SMT 기술이 활발히 연구되는 배경에는 다음과 같은 요인이 작용하고 있다. 1) 모델 파라미터를 학습할 수 있는 대용량의 가용 말뭉치가 구축되고 있다. 2) 특정 언어 쌍에 제한 받지 않고 모델을 자동으로 학습해 낼 수 있다. 3) 규칙기반/패턴기반 기계번역이 번역지식을 구축하는 데 상당한 비용을 요구하고, 다른 언어들에 일반화시켜 적용하기 어렵다는 문제가 있다.

SMT의 기본 요소는 통계적 번역모델(translation model)과 언어모델(language model), 이중언어 말뭉치(bilingual parallel corpus)로부터 은닉된 번역지식 파라미터를 찾아내는 학습 알고리듬, 그리고 학습된 번역모델에 기반하여 최적의 번역결과를 탐색하는 디코딩 알고리듬(decoder)으로 구성된다((그림 1) 참조). 기본적으로 원시언어 문장 f를 목적언어 문장 e로 번역하는 확률모델은 p(e|f)인데, 자연스러운 번역결과를 얻기 위해 Bayes Theorem을 적용하여 번역모델 p(f|e)과 언어모델 p(e)의 결합으로 유도된 생성모델을 만든다(식 (1) 참조).



(그림 1) SMT 기본 구성도

$$\tilde{e} = \arg\max_e P(e_1^m \mid f_1^l) \propto P(f_1^l \mid e_1^m) P(e_1^m) \quad (1)$$

그리고, 자료부족 문제를 완화하기 위해 언어모델은 n-gram 모델로, 번역모델은 단어/구문 대역모델, 어순재배열 모델 등 다수의 서브 모델들의 결

합으로 재구성하고, 모델 파라미터를 학습한다. 그리고 학습된 모델에 기반해서 가능한 모든 가설공간(hypothesis space)을 탐색하여 최적의 번역결과를 얻을 수 있다. 그러나 SMT 기본 모델은 단어단위 모델(IBM model 1-5)[3]로 문장 길이에 따라 어순 재배열의 계산 복잡도가 너무 높기 때문에 모든 가설공간을 탐색하는 것이 불가능하다. 이에 SMT 연구자들은 계산 복잡도를 낮추기 위해 새로운 번역 모델을 설계, 탐색공간 제한 알고리듬 및 휴리스틱 탐색 알고리듬들을 시도한다.

특히 2003년 어순 재배열에 따른 계산복잡도 감소 및 번역 효율을 고려한 구문단위 번역 모델[4]이 소개되면서 기술이 급격히 발전하여 현재 state-of-the-art를 이루었다. 그리고 최근에는 통사적 언어구조를 모델에 접합시키기 위해 언어학적 형태-통사적 구조(morphosyntactic structure)를 모델에서 직접 사용하거나[5], 또는 의사-통사적 구조(quasi-syntactic structure)를 도입하여 은닉된 통사적 구조정보를 자동으로 획득하여 모델을 학습하는[6] 등 통계 모델이 더욱 고등화되어 가는 추세를 보이고 있다. 더 나아가 자료부족에 따른 미등록어 문제 및 관용적 숙어 표현의 효과적 처리를 위해 paraphrasing 기법 또는 번역 지식 일반화 등의 기술에 도전하고 있다[7],[8].

아울러, SMT 분야에서는 time-to-market 및 시스템 평가에 따른 비용 절감을 겨냥, BLEU[9], NIST[10], METEOR[11],[12]와 같은 평가 알고리듬을 개발하여 번역 시스템 개발 과정에 효과적으로 활용하고 있다. 그리고, 더욱 인간의 평가에 근접한 지능적 자동평가 방법을 개발하기 위해 끊임없는 노력을 기울이고 있다.

현재 SMT 기술은 중국어-영어, 아랍어-영어 뉴스 자동번역 분야에 적용한 결과, Google, IBM, ISI의 자동번역시스템들은 실용적 수준에까지 이르고 있는 것으로 보고되고 있으며[13], 유럽과 아시아권에서는 자동통역의 번역 기술로 SMT 기술을 채택 각국의 기업, 연구소들이 컨소시엄을 구성하여 SMT 연구 개발에 박차를 가하고 있는 추세이다.

Ⅲ. 자동번역 기술 응용 사례

1. 웹 문서 자동번역

현재 웹에는 다양한 언어로 표현된 웹 문서들이 개인 흥미거리뿐만 아니라 사업적이고 전문적인 자료까지 포함하고 있어 자동번역을 적용하기 가장 적합한 분야라 할 수 있다.

웹 문서 자동번역이란 HTML/XML과 같은 마크업 언어(markup language)로 표현된 웹 상에 존재하는 다양한 문서들을 번역해주는 기술을 의미한다. 전통적인 텍스트 문서 기반의 자동번역과의 차이점은 자동번역된 결과 역시 원본 파일과 같은 웹 문서 형식을 유지하는 것이다.

웹 문서 자동번역시스템 및 서비스는 1990년 후반부터 개발되어 상용화되어 왔다. 대표적인 외국 사례로는 AltaVista의 Babel Fish[1]를 들 수 있다. 현재는 AltaVista가 Yahoo의 한 파트로 되면서, Yahoo의 Babel Fish로 더욱 알려져 있다. Babel Fish의 모든 자동번역 기술은 Systran에서 제공하고 있다. 현재 Yahoo의 Babel Fish에서는 16개의 언어 쌍에 대한 양방향 자동번역을 일반 텍스트와 웹 문서에 대해 서비스하고 있다.

Google 역시 몇 년 전부터 Systran에서 제공하는 Babel Fish를 이용하여 웹 문서 자동번역 서비스를 제공하고 있다. 최근에 Google에서 자체적으로 통계적 자동번역 방법으로 자동번역의 성능을 향상하기 위한 노력을 표방하고 이에 주력하고 있다. 그 과정으로 전문번역가들에 의해 번역된 200억 단어 규모의 병렬코퍼스(parallel corpus)를 확보하여 통계적 자동방법으로 Google에서 직접 영어-아랍어, 영어-중국어 간체, 영어-중국어 번체, 영어-러시아어, 중국어 간체-중국어 번체 등 5개 언어 쌍에 대한 양방향 자동번역시스템을 개발하였다[14],[15].

1) 더글라스 아담스(Douglas Adams)의 소설 “은하수를 여행하는 히치하이커를 위한 안내서(The Hitchhiker’s Guide to the Galaxy)”에 나오는 동시 통역을 위해 사용된 언어번역 기능을 가진 가상동물

또한 Google은 약 100만 책자 분량의 병렬코퍼스와 같은 자원을 이용하여 현재 서비스중인 자동번역 프로그램을 향상하여 Systran에서 제공하는 Babel Fish 자동번역과 차별화를 두고 있다[15].

국내의 대표적인 웹 문서 자동번역은 한/일 양방향 자동번역을 예로 들 수 있다. 한국어와 일본어의 언어적 유사성이 높은 특징에 의해 높은 번역률을 가지므로 초기부터 국내 포털서비스 업체에서 웹 번역 및 텍스트 번역 서비스를 해왔다. 또한, 2002년 한일 월드컵과 독도문제 같은 사회적인 이슈가 있을 때마다 한국과 일본 네티즌이 한 게시판에서 자국어로 논쟁을 벌이고 상대국 네티즌의 논쟁을 자국어로 이해하는 토론의 장을 마련하기도 하였다.

또한, 한국전자통신연구원에서는 특정 분야의 문서들에 대한 특징을 이용하여 일반 목적의 자동번역 시스템을 특화하여 그 분야에서만은 사용자가 번역 원본을 보지 않고도 이해할만한 고품질의 자동번역 시스템을 개발하여 왔다. 이러한 응용특화 방법을 이용하여 웹 문서 자동번역에서는 우선 신문기사에 특화한 중한 웹 신문 자동번역시스템을 개발중에 있다.

2. 기술문서 자동번역

기술문서란 특정 기술(technology)과 관련된 내용을 기술한 문서로써, 특허, 논문, 매뉴얼 등이 이에 해당한다. 자동번역의 대상 도메인을 특정 분야로 한정할 경우에는 자동번역의 성능이 향상될 수 있으며, 상용화 가능 수준까지도 도달할 수 있다. 본 절에서는 기술문서 가운데 특허 및 기술논문에 대한 자동번역 기술 개발 동향을 살펴보고자 한다.

가. 특허문서 자동번역

특허문서 자동번역은 특허출원의 급증으로 심사기간의 장기화, 신속한 권리확보 미흡 등의 문제에 직면하면서 지속적으로 증가하는 지재권 출원을 적절하게 처리할 수 없게 됨으로써 각국 특허청이 공동출원에 대한 '심사결과 상호인정제도'를 도입하기로 1997년 합의하고 활용하기 시작하였다.

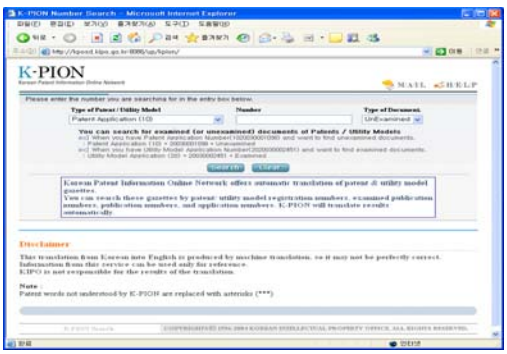
유럽에서는 1993년부터 유럽 특허청을 중심으로 특허 자동번역기를 활용하여 자동번역하고 있으며(예: PaTrans는 영어 특허 문서의 75% 정도를 덴마크어로 자동번역하고 있음), 이탈리아에서는 1987년부터 6개국어(영어, 독어, 불어, 스페인어, 포르투갈어, 이탈리아어)간의 특허 자동번역 서비스를 하고 있다. 특히 일본 특허청에서는 Toshiba 자동번역 엔진을 이용하여 2000년 3월부터 일영 자동번역 기술을 이용하여 인터넷상에서 일어 원문을 영문으로 서비스하고 있다.

국내에서의 특허문서 자동번역은 2004년 5월에 특허청과 ETRI와의 업무협약 조인식에 따라 특허문서에 특화된 한영 자동번역시스템이 개발되기 시작하였다.

특허청에서는 ETRI에서 개발한 한영 특허문서 자동번역시스템을 채용하여 2005년 11월에 특허문서 한영 자동번역 시범서비스를 실시하였다((그림 2) 참조). 또한 ETRI의 영한 특허문서 자동번역 시스템이 2006년도부터 산업자원부 특허지원센터 영한 특허문서 자동번역 서비스에 채택되고 있다.

영역 특화를 통한 자동번역 성능의 향상을 위해서는 특화시킬 영역의 특성분석이 필수적이다. 이러한 특성분석을 통해 기존의 특화 이전의 자동번역시의 문제점을 해결하거나 완화시킬 수 있는 방안을 찾아내야 하기 때문이다. 특허문서 자동번역 시스템을 개발하기 위해 수행한 특허문서의 특성들을 살펴보면 다음과 같은 것들이 있다.

우선, 용어적인 특성으로는 특허문서 중 전문용



(그림 2) ETRI 한영 특허문서 자동번역시스템

어가 광범위하게 발생한다는 것이다. 따라서, 이러한 전문용어를 구축하고 이들에 대한 대역어를 부착하는 것이 특허문서 특화 작업의 최우선 작업 중 하나이다. 이때, 일반적으로 많이 사용되는 용어들과 동일한 형태를 가지지만 전혀 다른 의미를 가지는 용어들에 대해서는 특별히 주의를 기울여야 한다. 예를 들어, '조사하다'라는 용어에 대해 일반적으로 사용되는 의미는 'investigate'이지만, 광학분야나 반도체 분야에서는 'irradiate'라는 의미로 사용된다. 이렇게 일반분야와 특허분야간 뿐만 아니라, 특허문서의 세부 영역(예: 전기전자, 의료, 화학 등)간에도 이러한 문제가 발생한다. 그러나, 대부분의 특허 전문용어들은 이러한 문제가 발생하지 않으며 평균적으로는 일반적인 분야에 비해 어휘의미 중의성의 비율이 지극히 낮아 번역 성능을 높이는 역할을 한다. 그러나, 이러한 고품질의 언어자원 구축은 오랜 시간과 많은 비용을 필요로 한다.

용어 특성 이외에 특허문서에 대한 언어학적 특성들을 영어 특허 문서로부터 살펴보면, 형태소적인 특성으로는 전문용어로 사용된 단어들로 인해 품사 애매성이 감소되고, 특허 문장의 특성에 따라 특정하게 나타나는 품사 패턴이 존재한다. 또한 기호단어(수식, 함수명 등), 약어, 하이픈 복합어 등이 다수 출현한다. 그러나, 현재형 문장이 사용되므로 동사/명사 중의성이 있는 단어에 대해서는 품사 애매성이 더 크게 나타난다. 구문적 특성으로는 일반 분야에 비해 긴 장문, 상투적 표현, 복잡한 병렬구조, 동사구 분사 등을 빈번하게 사용한다. 반면, 생략이나 시간부사구 등이 거의 사용되지 않고 비교적 정형적인 문장들이 사용되는 특성이 있다. 또한 특허 문서를 구성하는 각 필드별로 자주 쓰이는 문형들이 존재하며, DRAWINGS와 CLAIMS 필드 같은 경우에는 특히 사용되는 동사의 종류가 한정적인 특성을 보인다.

나. 기술논문 자동번역

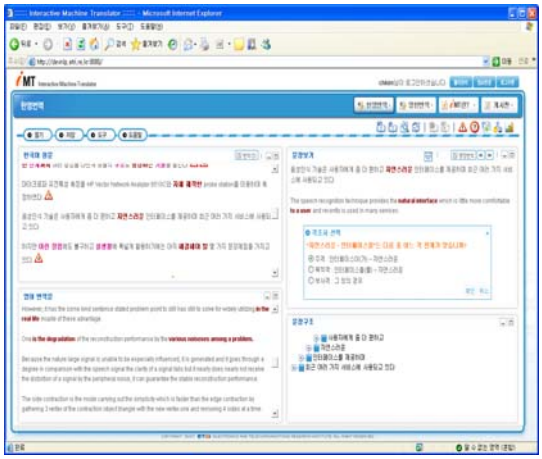
기술논문 자동번역의 필요성은 크게 2가지로 나눌 수 있다. 하나는, 외국어로 된 최신 논문으로부터 최신 연구동향에 대한 정보습득을 위한 것이고 다른 하나는, 외국 학회에 논문투고를 위해 모국어로 작성된 논문을 외국어로 번역하기 위함이다.

일본의 경우, IT 분야 정부출연연구소 NICT에서는 2006년부터 5년 계획으로 예제기반 자동번역방식으로 일-중 논문 자동번역시스템을 연구 개발하고 있다. 유럽은 다국어 사회라는 지역적 특색으로 인해 Trados의 "Translators Workbench", Star의 "Transit", Eurolang의 "Optimizer"와 같은 번역가 지원 도구가 활발히 개발되어 사용되고 있다.

번역 지원도구로써 자동번역에서 가장 중요한 것 중의 하나가 원문이 자동번역 시스템이 처리하기에 적절하게 작성되어야 한다. 이는 원문이 잘 쓰여진 경우에는 번역이 잘 될 수 있으나, 그렇지 않을 경우 번역기의 성능이 아무리 뛰어나더라도 번역성능이 좋을 수 없기 때문이다. 이를 위해 미국, 유럽 등에서는 이미 AECMA Simplified English를 비롯한 영어, 독일어, 프랑스어 등과 같은 언어에 대한 통제언어(controlled language) 및 이를 적용하기 위한 도구를 연구 개발하고 있으며, 이를 통해 원문의 완전성을 높이고자 하는 연구들이 진행되고 있다.

유럽항공산업 연합(AECMA)에서는 이미 1980년대 초반부터 통제언어에 대한 개발을 해왔으며, 1990년대 후반부터는 Xerox, IBM, Siemens 등과 같은 연구소 등에서 통제언어의 적용을 위한 도구 등을 개발하고 있다. CMU에서는 지식기반 자동번역시스템 KANT에 적용하기 위한 통제언어인 KANT Controlled English를 개발하였으며, 이를 자동번역시스템에 적용하기 위한 KANT Simplified English Diagnostifier를 개발하였다. 현재 통제언어 관련 설계는 국제적으로 통용되는 표준은 수립되어 있지 않으나, AECMA Controlled English가 영어에 관해서는 실질적인 표준 언어로 사용되고 있는 상황이며, 유럽에서는 TEKOM을 중심으로 유럽어들에 대해 통제언어에 대한 표준화 작업이 진행되고 있다.

국내의 경우, 현재 ETRI에서 연구되고 있는 논문 자동번역에서는 기존의 외국의 연구 방향과는 조금 다른 방향의 연구가 진행되고 있다. 즉, 통제언어의



(그림 3) ETRI 한영/영한 기술논문 자동번역시스템

연구를 통해 원문의 완성도를 높이는 동시에, 사용자는 원문의 작성뿐 아니라 자동번역기의 번역 과정에도 참여함으로써 자동번역 엔진과 사용자의 상호작용을 통해 보다 완성도 높은 번역 결과를 제공하려는 시도를 하고 있다. (그림 3)은 ETRI에서 연구하고 있는 논문 자동번역 시스템의 일부를 보이고 있다.

ETRI에서 제공하는 논문자동번역 시스템은 크게 3가지 기능을 제공한다. 첫번째는 통제언어 연구를 통한 원문 수정 기능이고, 두번째는 엔진 분석 결과에 대해 사용자의 의견을 반영하는 엔진-사용자 상호작용 기능이며, 세번째는 번역 결과에 대해 오류 가능성을 알려주고 이에 대한 대안을 제시하는 후처리 기능이다. 이러한 사용자와의 보다 폭넓은 상호작용 및 이러한 상호작용들을 사용자별로 저장함으로 인해 시간이 지남에 따라 더 나은 번역 성능을 얻을 수 있다.

3. 자동통역

자동통역 기술이란 모국어를 사용하여 외국인과 자유롭게 대화할 수 있도록 하는 기술을 말하며 대화체 음성인식, 대화체 언어번역, 음성합성 등의 요소 기술이 어우러진 기술을 말한다. 따라서, 자동통역은 자동번역 기술의 또 하나의 응용이라 할 수 있다.

자동통역 기술의 완성도는 세계적으로 아직 높지 않으나, 현재의 기술력으로도 작업의 내용이 명확히 정의된 여러 제한된 응용분야에 대해서는 가까운 장래에 자동통역시스템의 구현이 가능하다.

미국의 경우 DARPA의 지원으로 수행하는 GALE 프로젝트는 최근 중국의 부상과 이라크 전쟁 상황에 직면하여 언어음성 연구에 대한 지속적 필요성을 절감하고, IBM, BBN, SRI를 주축으로 중국어 및 아랍어를 영어로 변환하는 자동통역 연구를 대규모로 지원하고 있다. 대표적인 사례로는 미 국방부가 이라크와 아프카니스탄에 파견된 미군부대에서 사용중인 총 2천 대 가량의 휴대용 자동통역기인 Phraselator이다. 또한, 2003년에는 CMU에서 개발된 PDA용 자동번역/통역 시스템 Speechlator는 iPaq PDA에서 작동하며, 번역 정확도 약 80% 수준으로 의료정보를 아랍어와 영어로 상호 통역하고 있다. 2006년 캘리포니아주 세인트 마틴 메디컬 센터 등 3개 병원은 최근 영어를 잘 못하는 이민자 환자를 위해 진료실과 통역실을 연결하는 TV 회의용 자동통역 시스템을 설치하는 등 서비스를 강화했으며 이 시스템은 스페인어, 캄보디아어, 힌두어 등 26개 국어를 통역할 수 있다.

EU 의회에서 행해지는 모든 연설을 포함하여, 11개 공식 언어로 문서화하는 작업에 연간 5억 4천 9백만 유로를 지출하고 있으며 회원국 간의 교류에 있어서 언어장벽이 가장 심각한 문제로 대두됨에 따라, EU 제6차 연구지원사업 프레임워크에서는 의회 연설문의 자동전사 및 통역/번역 기술 개발을 목표로 하는 TC-STAR 프로젝트를 지원하고 있다.

일본 NEC의 경우 여행 도메인을 대상으로 하는 PDA 실시간 일-영/영-일, 일-중/중-일 통역 시스템을 개발하여 2005년부터 나리타 공항에서 시범서비스를 실시하였으며, ATR 연구소에서는 자동통역기술의 상용화를 목표로 연간 200억 엔의 연구비를 투입하여 일어-영어, 일어-중국어 간 양방향 대화체 자동통역시스템을 개발하고 있다.

국내의 경우, 1999년 ETRI가 C-STAR의 핵심 회원 기관으로서 고객이 여행사 직원과 여행계획을

상담하는 대화를 대상으로 5,000단어를 처리하는 한/영/일/프랑스 4개국 간 실시간 음성언어번역 국제시연을 실시하였다. 또한 2001년 삼성과 히타치 연구소가 한/일 월드컵 국제행사를 통해 제한 문장에서 휴대전화로 이용할 수 있는 한/일 자동통역 시범 서비스를 실시한 바 있다.

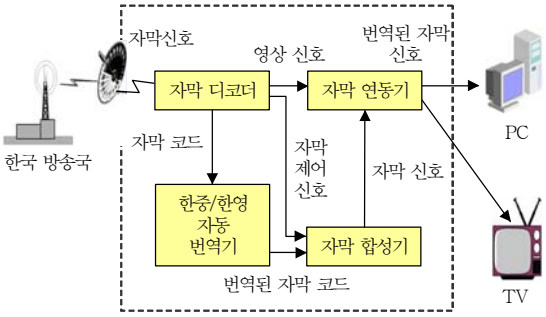
4. 방송 자막 자동번역

자동번역의 또 다른 응용분야가 바로 방송자막과 같은 구어체 자동번역이다. 방송자막 자동번역시스템은 아직까지 특허번역과 같은 기술문서 자동번역 시스템보다 번역 성능이 낮은데 그 주요한 이유는 방송자막 문장을 이루는 구어체의 경우, 비문이거나 불완전한 문장이 많고 사투리, 신조어, 은어 등의 사용이 빈번하고, 문장 성분의 생략 및 오용이 자주 발생하며, 방송자막과 같은 구어체는 또한 장면의 빠른 전환, 배경 지식의 생략, 짧은 발화 등으로 인해 문장의 의미 중의성이 높다. 그럼에도 불구하고 방송자막 자동번역 분야는 방송 화면 등과 같은 다른 정보 전달 매개체가 존재함으로 인해 기계에 의한 80% 정도의 번역률만으로도 실용화가 가능한 자동번역의 블루오션 분야이다.

외국의 경우, 미래 기술 연구의 대표 주자인 미국의 왓슨 연구소에서는 2006년 9월에 방송자막 자동번역을 위한 탤스 시스템(TALES)을 개발하여 발표하였다. 이 시스템은 TV 방송국이나 라디오 방송국으로부터 음성신호를 받아 들여 이를 텍스트로 변환한 후, 자동번역을 수행하고 번역 결과를 모니터 화면 아래에 자막으로 보여준다. 현재 아랍어, 스페인어, 중국어 등을 실시간으로 번역 가능하다. 자동번역 대상 언어 수를 늘리기 위해 Stanford 대학 등 6개의 대학과 제휴해 연구 개발중에 있으며, 시스템을 상용화하기 위해 미국 주요 방송사들과 협의중에 있다.

우리 나라의 경우, 현재 지상파 3사의 한국어 뉴스 방송을 청각 장애인을 위해 자막방송을 실시하고 있다. 즉 방송 신호를 송출할 때, 영상신호 및 음성신호에 더하여 자막신호를 함께 제공하는 것이다. 이러한 자막방송은 드라마 등으로 계속적으로 확대되고 있는 추세이다.

현재 ETRI의 한영/한중 방송자막 자동번역시스템[16]은 방송자막신호로부터 한국어 문장을 추출하고, 이를 한영/한중 자동번역시스템을 이용하여 실시간으로 영어나 중국어로 자동번역한 후, 번역된 영어 또는 중국어 문장을 다시 영상신호 등과 결합하여 방송중인 화면에 보여 준다. (그림 4)는 한영/한중 방송자막 자동번역이 처리되는 흐름도를 보여준다.



(그림 4) 한영/한중 방송자막 자동번역 흐름도

이 시스템은 한국어를 모르거나 서투른 영어권 또는 중국어권 외국인들도 한국 방송의 내용을 이해하면서 시청할 수 있도록 하는 데 그 목적이 있으며, 휴대의 편의성을 위해 셋톱박스의 형태로 개발하였다. (그림 5)는 사용자가 한중 방송자막 번역시스템을 이용해 실시간으로 한국어 뉴스 방송을 시청하는 경우에 대한 TV 화면의 실시 예이다.

특히 한중 방송자막 자동번역시스템의 경우는, 우리별 인공위성의 성공적 발사로 인해 중국을 비롯

(그림 5) 한중 방송자막 자동번역시스템의 실시 예

한 동남아 지역이 한국방송 시청권으로 포함되고, 한류열풍이 확산됨에 따라 자국어 자막을 통한 한국 방송의 실시간 시청에 대한 요구가 증대되고 있어 뉴스, 드라마 등의 한국 방송에서 제공되는 자막방송에 대한 중국어와 영어로의 자동번역 기능을 탑재한 위성 셋톱박스를 개발하여 상용화에 성공할 경우, 이 장치의 중국 및 동남아 시장에서의 수출 증대 효과, 타 지역으로의 한류열풍의 확산을 기대할 수 있다.

Ⅳ. 결론

21세기에는 세계화의 가속화로 인한 국가간 인적, 물적 교류가 잦아질 것이며, 따라서 언어간 장벽을 허무는 자동번역 기술의 확보는 무한 경쟁 시대에서 국가 경쟁력 제고와 직결된다. 이미 미국, 유럽, 일본 등 여러 선진국가에서는 자동번역 시장의 중요성을 인식하고 국책과제를 통해 장기적인 연구 지원을 하고 있다.

자동번역은 이미 언어적 유사성이 많은 언어들 간에는 웹 문서 등에 실용화 서비스(Google, Yahoo 등)까지 하고 있는 단계에 이르고 있으며, 특허문서와 같이 특화된 자동번역시스템은 그 성능이 80%를 넘어 상용화 서비스(ETRI 특허문서 자동번역시스템)까지 하고 있는 단계까지 도달했다. 이러한 성공 사례로 인해 영역 특화를 통한 자동번역 기술의 상용화라는 흐름이 점차 가속화될 것으로 예상된다.

자동번역은 또한 자동통역 등의 분야로 응용을 확대함으로써 자동번역/통역 기술은 향후 사회전반에 역할을 확대해가고 있다.

● 용어해설 ●

형태소 태깅: 문자열을 분석하여 자연언어 분석의 기본 단위인 형태소로 분해, 해당되는 정확한 문법 정보를 제시하는 것을 말한다. 예를 들어, 'love'라는 단어는 사전에 'love: 동사, 사랑하다' 또는 'love: 명사, 사랑'으로 등록되어 있는데, 'I love you'라는 문장에서 'love'는 명사가 아니라 동사로 컴퓨터가 분석하는 기능을 말한다.

특화 작업: 특화라는 말은 '사용자의 사정에 맞추다'라는 뜻으로, 일반적인 도메인을 대상으로 만들어진 번역 시스템을 특허 도메인에 맞게 대용량의 전문용어를 추가 및 수정하고, 특허문서에 특수한 패턴을 추가하고 일반 시스템을 특허 도메인으로 수정하는 작업을 말한다.

약 어 정 리

BBN	Bolt Beranek and Newman
BLEU	BiLingual Evaluation Understudy
CAT	Computer-Aided Translation
CMU	Carnegie Mellon University
DARPA	Defense Advanced Research Projects Agency
GALE	Global Autonomous Language Exploitation
HTML	Hyper Text Markup Language
ISI	International Statistical Institute
KANT	Knowledge-based, Accurate Natural-language Translation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
NIST	National Institute of Standards and Technology
SMT	Statistical Machine Translation
TALES	Translingual Automatic Language Exploitation System
XML	eXtended Markup Language

참 고 문 헌

[1] 최승권, 홍문표, 박상규, "다국어 자동번역 기술," 전자통신동향분석, 통권 95호, 제20권 제5호, 2005. 10., pp.16-27.

[2] W. Weaver(1955) and Translation(1949), In: Machine Translation of Languages, MIT Press, Cambridge, MA.

[3] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, Vol.19, No.2, 1991, pp.263-311.

[4] P. Koehn, F.J. Och, and D. Marcu, "Statistical Phrase Based Translation," *In Proc. of the HLT/NAACL*, 2003.

[5] Y.S. Hwang, A. Finch, and Y. Sasaki, "Improving Statistical Machine Translation Using Shallow Linguistic Knowledge," *Computer Speech and Language*, Vol.21, No.2, 2007.

[6] D. Chiang, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," *In Proc. of ACL'05*, 2005.

[7] C. Bannard and C.B. Callison, "Paraphrasing with Bilingual Parallel Corpora," *In Proc. of ACL'05*, 2005.

[8] Y.S. Hwang, Y.K. Kim, and S.K. Park, "Paraphrasing Depending on Bilingual Context Toward Generalization of Translation Knowledge," *Proc. of the Third Int'l Joint Conf. on Natural Language Processing*, 2008.

[9] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, Bleu: A Method for Automatic Evaluation of Machine Translation, IBM Research Report, RC22176, 2001.

[10] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurence Statistics," *Proc. of the HLT Conf.*, 2002.

[11] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proc. of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the ACL'05*, 2005.

[12] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," *Proc. of Workshop on Statistical Machine Translation at the ACL-'07*, 2007.

[13] http://www.nist.gov/speech/tests/mt/doc/mt06eval_official_results.html

[14] Google 자동번역 FAQ, http://www.google.com/intl/ko/help/faq_translation.html#statmt

[15] Google Machine Translation System, http://adjuster.blogspot.com/2006/06/google-machine-translation-system.html

[16] Hwe-Mo Kim and Kyong-Ho Lee, "Device-independent Web Browsing Based on CC/PP and Annotation," *Interacting with Computers*, Vol.18, 2006, pp.283-303.