자동 번역과 CAT의 현황과 전망

(주)클릭큐 부설 기술연구소 박주형 · 이창우 · 강명주

1. 서 론

자동 번역에 대한 연구는 Warren Weaver와 Andraw Booth에 의하여 1940년대부터 시작되었다. 이 때의 연구는 일반 사람들에게는 거의 알려져 있지 않았으며, 본격적인 연구는 컴퓨터가 나오게 되는 1950년대부터 시작되었다. 이들이 생각했던 자동 번역은 두 언어간에 단어들을 교환하여 대치시키고 단어 쌍으로 이루어진 전자사전을 컴퓨터에 입력하여 번역하는 형태였다. 이 시기에 사용한 언어 쌍은 영어-러시아어였으며 주로 군사적인 목적으로 미국과 구 소련에서 연구되었다[1, 2].

1950년과 60년대의 연구는 주로 우주항공과 군사 분야에서 이루어졌다. 1954년에 IBM과 Georgetown 대학이 핵물리학에 관한 러시아어 문서를 영어로 번역하는 시스템을 공동으로 개발하였다. 이 시기에 직접 번역 방법(Direct Translation Method)과 중간언어 번역 방법(Interlingual Translation Method)이 연구되었다[3]. 그러나 느린 발전 속도와 수준 낮은 번역 결과는 일반인은 물론 자동 번역을 연구하는 사람들까지도 실망시켰다. 특히 ALPAC(Automatic Language Processing Advisory Committee) 보고서에서 자동 번역은 실질적인 면에서 실현성이 없다는 결론을 내린 후 컴퓨터가 번역가를 지원할 수 있는 시스템, CAT(Computer Aided Translation)에 대한 필요성이 제기되었다[4].

1970년대에는 언어학에서 개발된 변형생성(Generative-transformational) 문법이나 W. Woods에 의하여 연구된 ATN(Augmented Transition Networks) 문법 그리고 C. Fillmore의 격(Case) 문법 등 자연언어처리에 효과적으로 적용될 수 있는 방법들이 제시되었다. 1970년에는 러시아어-영어 자동 번역 시스템인 SYSTRAN이 개발되었다[5]. 이 시스템은 유럽공동체에서 지원을 받아 유럽권 언어를 번역할 수 있는 시스템으로 발전하였다[6].

1980년대에 ALPAC 보고서 이후 침체되었던 연구는 인공 지능 기반 방법(AI-based Method), 지식 기반 방법(Knowledge-based Method) 또는 규칙 기반 방법(Rule-based Method)의 제시와 컴퓨터의 성능 발달에 힘입어 전세계적으로 활발하게 진행되었다[7]. 규칙 기반의 방법은 자연어가 사용될 때 적용되는 일반적인 규칙을 찾아내고 그 규칙을 이용하여 자동 번역에 사용하는 방법이다. 그러나 이 방법은 복잡한 언어 현상들로부터 규칙을 추출하기가 매우 어렵고, 다른 언어 영역으로 확장하기가 어렵다는 단점이 있다. 그럼에도 불구하고 이 시기에 일본에서는 기업(Sharp, NEC, Oki, IBM, Mitsubishi, Sanyo, Nova)에서의 연구가 활발하여 시장에서 판매될 수 있는 제품들이 개발되었고, 한국에서도 한국어를 중심으로 영어와 일본어에 대한 자동 번역 시스템이 연구 및 개발되기 시작하였다. 또한 자동 번역의 질을 높이기 위하여 전 편집과 후 편집 기능이 나오게 되었으며 번역가를 위한 CAT에 대한 연구가 활발하여 용어 관리 도구(Terminology Management Tool), 정렬 도구(Alignment Tool)와 같은 소프트웨어가 개발되고, 초보적인 번역 메모리(TM, Translation Memory)가 연구되었다.

1990년대에는 규칙 기반의 방법을 극복하기 위해 IBM Candide 프로젝트에서 처음 소개된 코퍼스 기반 방법(Corpus-based Method), 예제 기반 방법(Example-based Method), 통계 기반 방법(Statistics-based Method)이 연구되었다[8,9,10,11].

통계 기반 방법은 사람들이 실제로 사용하는 많은 언어 데이터로부터 확률 정보 및 통계 정보를 추출하고, 이를 이용하여 여러 가지 언어 현상을 규정하여 자동 번역에 이용하고자 하는 방법이다. 그러나 통계 기반 방법은 실제로 사용하는 대용량 코퍼스를 모으기가 어렵고, 언어적 활용이 많아 다양한 문장 구성의 연구가 필요하고, 대용량 데이터를 처리할 수 있는 컴퓨터가 필요하다는 단점이 있다. 한편 이 시기에 개인용 컴퓨터가 보편화되고 인터넷이 활성화되어 월드 와이드 웹(WWW, World Wide Web)이 등장함에 따라 일반 사람이나 번역가들도 쉽게 컴퓨터를 사용할 수 있는 환경이 되었다. 이런 시대적인 환경에 따라 번역가를 위한 도구인 Trados Translator Workbench, IBM TranslationManager/2, STAR Transit, Eurolang Optimizer와 같은 CAT 소프트웨어들이 시장에 출시되었고 번역 메모리, 용어 관리 도구, 정렬 도구, 워드 프로세서, 탁상용 출판 소프트웨어 등 번역 작업을 지원하는 도구들이 등장하였다. 1990년대 후반에는 기업들을 중심으로 규칙 기반 방법과 통계 기반 방법을 혼합하는 방법도 제시되었고, 자동 번역과 번역 메모리를 혼합한 시스템도 등장하였다.

앞으로 웹 환경이 활성화 됨에 따라 많은 지식들이 공유되고 그 지식에는 외국어로 된 지식들이 더욱 많아져 실시간(Real-time) 자동 번역에 대한 요구가 많아질 것이다. 일부 기업에서는 이미 웹 환경에서 실시간 자동 번역 서비스를 제공하고 있으며, 일본에서는 자동 번역기가 포함된 워드 프로세서가 나오고 있다.

본 논문에서는 자동 번역과 CAT의 역사 그리고 현황을 중심으로 각 방법론과 향후 발전 방향에 대하여 논의하고자 한다.

2. 역사 및 현황

2.1 1956년에서 1966년까지

이 시기의 자동 번역과 관련된 연구들은 크게 두 가지로 나뉜다. 그 하나는 컴퓨터에 적용할 수 있는 문법과 어휘에 대한 규칙을 확률적으로 추출하는 방법이고, 나머지 하나는 언어학적인 관점에서 언어의 기본 특성들을 고려하여 이론적으로 접근하는 방법이다. 이론 방법들을 통해 직접 번역 방법과 중간언어 번역 방법 등의 방법론과 변형생성 문법, 의존(Dependency) 문법, 성층(Stratificational) 문법 등의 이론적인 접근이 이루어졌다[12].

미국을 중심으로 하는 초기의 자동 번역에 대한 연구는 하드웨어의 제약으로 인하여, 그 방법론들이 주로 이론적으로만 확립되었다. 이론들에 대한 검증은 제한된 수의 원문(Source Segment)과 대역문(Target Segment)의 쌍(Pair)에 적용해 봄으로써 이루어졌으며, 이는 자동 번역 기술에 대한 큰 기대감을 가지게 하였다. 1960년대에 들어서면서 초기 자동 번역 기술들이 보여 주었던 검증 결과는 새로운 연구의 결과에 부담을 주어 자동 번역 기술의 발전에 대한 회의적인 시각이 나오기 시작하였다. 1966년 ALPAC의 보고서는 이러한 회의적인 시각들을 표면화 하였고, 이는 곧바로 자동 번역에 대한 연구들의 중단으로 이어졌다.

2.2 1967년에서 1976년까지

미국의 자동 번역 연구는 냉전 체제하의 정치적, 군사적 목적에서 진행되었지만, 캐나다나 유럽의 연구는 좀 더 현실적인 문제에 대한 고민에서 출발하였다. 따라서 ALPAC 보고서의 영향으로 미국 내 자동 번역에 대한 연구는 중단되었지만, 캐나다와 유럽쪽의 연구는 실현 가능한 제한된 분야에 적용하기 위한 방향으로 진행되었다. 이러한 필요성에 기반한 자동 번역에 대한 연구는 몇 개의 성공적인 도입 사례를 남겼다. 유럽의 Systran이 미 공군과 EC(European Community)에서 대량 문서의 처리를 위해서 도입되었고, 캐나다에서는 Meteo가 성공적인 결과를 보여주었다. 또한 일본에서는 국제 무역 증가에 따른 번역 대상 문서의 양이 기존의 번역 방식으로 처리할 수 있는 양을 넘어서 일본 내 자동 번역에 대한 연구를 촉진하였다.

이 시기 대부분의 연구들은 중간언어 번역 방법의 자동 번역을 채택하고 있었다. 따라서 피봇 언어(Pivot Language) 또는 정규 형식(Normal Form)으로 표현되는 중간언어들을 이용한 자동 번역 방법들이 등장하였다. 그러나 분석 정확도의 한계와 중간언어를 사용함으로써 발생할 수 있는 원문 정보의 손실 등의 이유로 중간언어 번역 방법의 자동 번역에 대한 연구들은 그 한계점을 보이고 있었고, 변환 방법을 이용한 자동 번역 기술의 가능성이 더 부각되었다.

2.3 1976년에서 1989년까지

1980년대 자동 번역에 대한 연구는 여러 종류의 언어학적 규칙들을 만들고, 그것을 이용하는 시스템을 만드는 것에 중점을 두었다. 변환 방법의 자동 번역 시스템들에서는 당연히 규칙 기반(Rule Based)의 처리가 이루어졌다. 그러나 언어학 기반, 또는 지식 기반의 중간언어 번역 방법의 시스템들에서도 처리 과정에서는 규칙 기반의 접근이 이루어졌다. 또한 이 시기에 촘스키(Chomsky)의 지배-결속(GB:Government-Binding) 이론이 나오면서 이를 바탕으로 어휘 기능 문법(LFG:Lexical-Functional Grammar), 핵 중심 구 구조 문법(HPSG:Head-driven Phrase Structure Grammar), 일반 문법(UG:Universal Grammar) 등이 제안되었다[16].

중간언어 번역 방법에 대한 한계성의 인식은 언어학 기반의 2세대 변환 방법으로 불리는 자동 번역 시스템에 대한 연구를 시작하게 하였다. GETA(Groupe d'Etudes pour la Traduction Automatiquc)의 Ariane 시스템은 모듈화를 통한 유연성과 트리 구조 변환에 대한 알고리즘 및 정적, 동적 문법 규칙에 대한 개념들을 포함한 2세대 변환 방법 자동 번역 시스템이었다[17]. 비록 제품 개발 단계까지는 진행되지 못하였지만, 이 시스템은 1980년대 자동 번역에 대한 연구의 기반이 되었다.

EC에서는 1976년에 Eurotra 프로젝트가 시작되었다[18]. 이 프로젝트에서 EC에 속한 국가들의 언어간 자동 번역 시스템에 대한 연구가 이루어졌다. 처음에는 EC 회원국 영국, 프랑스, 독일, 이탈리아, 덴마크, 네덜란드, 그리스 등 7개국이 대상이었으나, 이 후 스페인과 포르투갈이 추가 되어 9개국어 자동 번역 시스템이 연구되었다. 이 프로젝트는 어휘, 구문, 의미론적 정보의 다단계 계층 구조를 이용한 저장 및 변환 방법에 대한 많은 연구 결과를 얻어 내었다. 더욱이 향후 자동 번역 시스템에 대한 국가간 협력 프로젝트의 좋은 선례를 남기었다.

1980년대 후반에는 인공 지능과 인지 언어학에 대한 연구의 영향으로 중간언어 번역 방법에 대한 접근이 다시 한번 이루어졌다. 네덜란드의 BSO사에서 개발한 DLT(Distributed Language Translation) 시스템은 에스페란토어를 변형한 중간언어를 사용한 자동 번역 시스템이었다[19]. 이 프로젝트는 대용량 어휘 데이터베이스 구축에 관해 많은 연구를 하였고, 그 결과로 인간이 번역한 코퍼스를 이용한 Bilingual Knowledge Bank를 만들 것을 제안하였는데, 이는 향후 등장하는 예제 기반 자동 번역(EBMT:Example Based Machine Translation)의 기본 아이디어가 되었다.

1980년대 말에 카네기 멜론 대학에서는 KANT 시스템을 완성하였다[20]. 이 시스템은 개념 어휘(concept lexicon)로부터 얻어낸 지식을 사용하여 오그멘터(Augmentor)가 모호성을 해소하고, 의미적 분석을 함으로써 얻어낸 명제들을 네트워크로 구성하여 문장을 중간언어로써 표현하였다. KANT 시스템은 향후 지식 기반(Knowledge based) 자동 번역의 가능성을 열었다.

일본의 자동 번역에 대한 연구는 중간언어를 이용한 NEC의 PIVOT 시스템과 기초적인 지식 기반(knowledge based) 방법을 적용한 NTT의 LUTE 프로젝트가 진행되었다[21]. 국내의 자동 번역에 대한 연구는 1980년대 초반 일본어-한국어에 대한 연구를 시작으로 1980년대 중반부터 서울대와 한국과학기술원에서 연구가 시작되었다. 초기에는 외국 시스템에서 변환과 생성과정을 한국어에 맞게 바꾸는 방식으로의 연구가 진행되었으나, 그 후 대학과 연구소들을 중심으로 한국어 관련 자동 번역과 관련된 프로젝트들이 생겨났다.

Systran의 자동 번역 시스템이 미 공군(1970년)과 EC(1976년)에 도입되면서 시작된 자동 번역 시스템의 상용화는 이 시기에 Systran과 Logos를 필두로 하여 전세계에서 이루어졌다. Systran의 직접 번역 방법의 자동 번역 시스템은 모듈화, 분석의 호환성 등에서 많은 발전을 이루어냈고, 이는 다른 언어간의 자동 번역을 위한 자동 번역 시스템 개발에 드는 시간과 비용을 줄였다. Logos Corporation은 영어-베트남어 항공기 매뉴얼 번역을 위해 자동 번역 시스템을 만들었다[22]. 이 연구 결과를 이용하여 독일어-영어 자동 번역 시스템이 개발되었고, 그 후 다른 여러 언어로 확장되었다. Systran, Logos 등의 시스템들은 범용으로 쓰일 수 있는 자동 번역 기술을 이용하여 만들어졌으나, 실제 사용에서는 정해진 특정 영역을 위해 사전을 재구성하였다. 이런 특정 영역의 번역을 위한 자동 번역 시스템은 Xerox나 Smart Corporation의 시스템에서 제한 언어로 발전되었다[23]. 제한 언어

자동 번역 시스템은 입력되는 원문에 대하여 구문적, 어휘적 제한을 두어 자동 번역 시스템의 결과에 대한 후편집(Post-editing) 과정을 최소화 하는데 그 목적을 두었다.

비록 1950년대부터 자동 번역이 개발되어 왔지만, 번역가들은 1980년대 후반이 되어서야 비로소 자동 번역이 자신의 업무에 별 도움을 주지 못한다는 사실을 알게 되었다[24]. 이는 PC가 1980년대에 보급되었기 때문인데, 번역가들은 당시 상용화된 몇몇 자동 번역 프로그램을 접하게 되었고, 그 프로그램들의 불확실성은 오히려 번역 작업 실무를 번거롭게 할 뿐이었다. 비록 PC의 보급이 당시 번역가들이 가지고 있던 자동 번역에 대한 과도한 기대와 그에 따른 회의스러움을 파급시켰지만, 문서 처리(Text-processing)에 대해서는 새로운 장을 열게 하였다. 문서 처리의 지원에 힘입어 번역가들은 자동 번역에 대한 요구보다는 전문 용어 데이터의 처리 및 관리, 번역 생산물의 관리 및 수용을 해결할 수 있는 기계적인 대안(소프트웨어와 하드웨어)을 요구하게 되었다. 이러한 요구는 문서 처리와 데이터베이스가 접목된 번역 워크스테이션(Translation Workstation)의 개발을 가져오게 하였다. 여기서 등장한 것이 번역 메모리 시스템이었다[25].

번역 메모리 시스템은 실제 번역과정이 이루어지고 있는 문서 처리 상에서 대상 구문과 목적 구문의 쌍을 데이터화 하여, 이후 번역 대상 구문의 문자열을 비교하여 유사 구문을 판정해내는 것이다. 일반적으로 문자열의 비교는 어휘단위의 문자열의 유사성과 연속성을 비교하는 퍼즈 매칭(Fuzz Matching) 알고리즘을 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.

번역 워크스테이션은 탁상 출판과 같은 고품질의 문서 처리 도구와 번역 메모리와 같은 데이터뱅킹 도구들을 번역가가 함께 사용할 수 있는 하나의 워크스테이션으로 구체화되었다. PC의 보급과 더불어 번역가 워크스테이션은 클라이언트 독립 프로그램의 형태로 제품화되었는데 그 응용 프로그램이 1990년대 이후에 등장하는 CAT이다. 그러므로 1980년대 말까지는 기본적인 번역 메모리 시스템과 전문 용어 관리를 지원하는 데이터 뱅크 프로그램이 워크스테이션에서 구동하는 형태를 취하고 있었다.

2.4 1990년에서 현재까지

1990년대에 들어서면서 통계 기반 방법과 코퍼스 기반 방법이 많이 연구되었다. 음성 인식 분야에서 성공적인 결과를 얻을 수 있었던, 통계 기반의 언어 모델을 사용하여 IBM에서 통계 기반 자동 번역 시스템을 디자인 하였다. 이들은 영어와 불어 코퍼스를 사용하여, 어떠한 언어학적인 규칙도 적용하지 않은 상태에서 단순히 원문장의 각각의 단어에 대해 대응되는 번역문장의 단어들이 출현하는 빈도를 통계적 지식으로 활용하여 번역하는 방법을 이용했다. 이러한 방법은 놀랍게도 좋은 결과를 보여주었다.

코퍼스 기반의 방법은 예제 기반 또는 메모리 기반(Memory Based)이라 불리는 방법이다. 번역 작업에는 이전에 번역한 결과를 참조하거나 그것을 그대로 사용하는 경우가 많다는 것에 착안한 이 방법은 이미 번역된 예제 문장들을 데이터베이스에 넣어두고 번역하고자 하는 문장과 가장 유사한 문장을 찾아 대응되는 번역문을 보여주었다. 유사한 문장을 찾아내는 방법으로는 용어에 대한 의미론적인 시소러스에 의한 계층 구조화 등을 통한 방법과 어휘 빈도에 대한 통계적 자료를 이용하는 방법 등이 제안되었다. 코퍼스 기반 방법의 장점은 이미 번역되어 있는 번역 결과를 이용하므로 숙어구나 관용적인 표현에 대한 번역 결과가 정확하다는 것이었다.

이 시기에 대화 음성 번역에 대한 연구가 시작되었고, 이에 필요한 음성 인식, 음성 합성 그리고 규칙 기반 또는 코퍼스 기반의 자동 번역 모듈에 대한 연구가 함께 진행되었다. 대표적인 프로젝트로는 일본의 ATR에서 이루어진 JANUS와 독일의 Verbmobil 등이 시작되었다[26, 27]. 또한, 전형적인 규칙 기반 방법에 대한 연구도 카네기 멜론 대학의 Catalyst 프로젝트 등에서 계속 이루어졌다.

이 시기에 개인용 컴퓨터 기반의 자동 번역 소프트웨어들이 전세계적으로 많이 출시되었고, 국내에

서도 일반 문서 번역을 위한 많은 제품들이 상용화 되었다. 영한 자동 번역 소프트웨어로는 IBM에서 출시된 앙꼬르를 비롯하여 E-tran, EnGuide, EZReader, 트래니, 워드체인지, ClickQ 등이 출시 되었다. 일한 자동 번역 소프트웨어로는 오경박사, J-Seoul/JK, 한글가나 등의 제품이 출시되었다.

1990년대 초반부터 활발하게 등장한 CAT 시스템의 아이디어는 1980년대의 번역가 워크스테이션과 동일한 것이지만, 각 도구가 처리하는 데이터들을 보다 통합적으로 관리, 연동할 수 있게 되었으며, DTP에 대한 지원도 보다 복잡한 형태를 띠게 되었다. 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 TMX(Translation Memory exchange)에 대한 지원을 통하여 응용 프로그램간의 데이터 교환을 지원한다. 현재 네트워크 지원과 자동 번역 지원, XML을 이용한 데이터 처리 등 많은 시도들이 이루어지고 있어서 여전히 많은 부분 CAT의 개념이 확장되어 가고 있는 상태이지만, 일반적으로는 번역 메모리 시스템을 포함하여 아래와 같은 공통된 요소들로 구성되어 있다[28].

2.4.1 정렬 도구

이미 번역된 전자문서를 TM데이터로 만들 수 있게 해주는 도구이다. 일반적으로 사용자에 의해 정의 가능한 세그멘테이션 규칙을 통해 분할된 구문단위 원문과 대역문을 데이터화 하게 한다[29]. 대부분이 GUI형태의 인터페이스를 띠고 있으며, 구두점을 이용한 문장 경계 구분(SBD, Sentence Boundary Disambiguation) 외에도 RTF(Rich Text Format)의 파일 형식 정보를 이용하여 보다 편집자의 의도에 근거한 데이터 추출을 하고 있으며, 동족어 정보나 약어 사전을 이용한 정렬 방식을 사용하기도 한다.

2.4.2 데이터 관리 도구(Data Management Tool)

대부분의 번역공정은 프로젝트 단위로 구성된다. 하나의 프로젝트에는 다양한 문서가 다양한 형식으로 존재하게 되는데 예를 들어 소프트웨어 L10N (Localization)의 경우 하나의 프로젝트는 텍스트, html, 프로그램 소스 전반을 포함한다. 하나의 프로젝트에는 번역 메모리뿐만 아니라 용어 데이터와 품질관리 및 일관성 유지를 위해서 별도의 History 데이터가 관리되어야 용어의 재정의 및 변환(Terminology Change Tracking) 일괄작업이 가능해 진다. 또 이렇게 축적된 프로젝트별 데이터를 효과적으로 검색, 적용하기 위해서는 각 데이터의 주제별 관리가 가능해야 하기 때문에 별도의 데이터관리 프로그램이 프로젝트 관리자(Project Manager)의 효과적인 지식구축을 가능하게 하는 것이다. 최근 대규모 프로젝트는 C/S(Client/Server)환경에서 진행되는 경향을 보이고 있어, 데이터의 관리가 보다 입체화되고 있는 추세이며, 표준안에 대한 지원 역시 발 빠르게 진행되고 있다. 웹 페이지의 L10N이나 G11N(Globalization)의 경우에는 컨텐츠 데이터가 평면적인 구성에서 입체적인 구성으로 변화함에 따라, 번역 공정과 컨텐츠의 생산을 종합적으로 관리할 수 있는 관리 도구와 컨텐츠의 주기적인 관리를 가능하게 하는 컨텐츠 관리 도구(Contents Management Tool)가 함께 사용되기도 한다.

2.4.3 용어 관리 도구

사용자가 번역 프로젝트에 필요한 용어를 정의하고 번역 작업 환경에서 참조할 수 있게 하는 프로그램이다. 대규모 프로젝트를 빠르면서도 일관성 있게 처리할 것을 요구하는 최근의 추세에 맞춰 빠른 데이터 검색처리와 다른 프로그램에서도 데이터를 호환할 수 있는 표준화의 지원(TBX)이 요구되었다. 대부분은 퍼즈 매칭 알고리즘을 사용하고 있으며, 가상 작업 공간의 발전에 따라 C/S환경에 적합한 네트웍기능을 지원하기도 하였으나 대부분은 제한된 범위에서였다.

2.4.4 문서 처리 도구

문서 처리를 한다는 점에서는 워드프로세서나 탁상 출판 프로그램과 비슷한 기능을 가지고 있지만, 보다 번역과정을 용이하게 만들어 주는 기능들을 가지고 있다. 제품에 따라서는 독립적인 문서처리 도구를 가지고 있는 경우와 다른 문서 처리 프로그램에 플러그인(plug-in)되는 형태를 띠고 있는데, 최근은 다른 도구들과 통합되는 형태의 UI가 일반적인 형태이고 ASP(Application Service Provider)의 요구에 따라 웹 환경에서 동작하는 응용 프로그램도 등장하였다. 문서 처리 이외에도 맞

춤법 및 철자교정기능을 가지고 있고 번역 메모리 시스템과의 연동을 최우선으로 지원하였다. 독립적인 문서 처리 도구를 사용하는 경우에는 다른 탁상출판 프로그램 및 문서 처리 프로그램의 파일을 편집할 수 있도록 파일 형식을 분석하는 필터(Format Filter)를 가지고 다양한 형태로 가져오기/내보내기(import/export)를 할 수 있도록 지원하였다.

2.5 현재 국내의 CAT 연구

현재 국내의 자동 번역 기술은 영어, 일어 중심의 웹 문서 번역, 채팅 번역 등에 관한 연구가 주로 진행되고 있으며 주로 대상 언어의 문법 규칙에 의존하여 지속적인 번역 품질의 향상이 어렵다. 기존의 번역 시스템에서 번역률의 한계를 극복하기 위하여 데이터 중심의 번역 방식을 시도한 바 있지만 고품질의 전문 번역을 위한 전문 분야별 데이터 중심의 CAT 기술 및 번역 연동 기술은 확보하고 있지 못하다. 현재 ETRI, 클릭큐, 삼성전자, 모비코 등의 연구소와 기업들이 공동으로 동양 4국 언어에 대한 TM 기반의 통합 CAT 시스템 개발을 하고 있다. 이 연구는 앞으로 많은 수요가 예측되는 한중 자동 번역 기술을 개발하고, 영어, 일어, 중국어의 고품질 자동 번역을 위한 TM 기술을 개발하여 동양 4국 TM 기반 통합 CAT 시스템을 구축하는 연구로써 국내 업체들의 수출 및 요소 기술 수입에 따른 전문 매뉴얼 문서 번역 기술로 활용될 것이다.

3. 향후 전망 및 시장 규모

궁극적으로 자동 번역의 연구 동기는 커뮤니케이션에 기반하고 있다. 지난 50년간의 정치와 산업의 변화가 곧 자동 번역과 CAT의 성장 과정과 긴밀한 관련을 가지게 된 것은 정치와 산업이 가졌던 다국적 커뮤니케이션(Global Communication)의 요구 때문이라고 할 수 있다. 마찬가지로 주춤했던 자동 번역 연구가 다시 활성화 된 배경 역시 인터넷의 급속한 성장이다. 인터넷은 접근 가능(Universal Access)을 전제로 한다. 최근 인터넷의 상업화 때문에 일정 부분 이 전제가 축소되기는 하였지만 여전히 인터넷의 정보는 우리가 담아둘 수 없을 정도로 홀러 넘치고 있다. 하지만 정보의 홍수는 여전히 양적인 홍수일 뿐 질적인 급류가 되어 밀려오지는 않고 있다. 즉, 컨텐츠와 사용자간의, 사용자와 사용자간의, 컨텐츠와 컨텐츠 간의 커뮤니케이션에 문제가 있다는 것이다. 이는 커뮤니케이션을 가능하게 해주는 기존의 번역 프로세스의 변화가 외부의 커뮤니케이션을 위한 채널로서의 테크놀로지의 발전 속도에 부합하지 못한 결과이다. 그렇기 때문에 산업과 정보 전반에 걸친 기존의 번역 프로세스의 부하를 해결하는 방안과 기존의 프로세스로 해결할 수 없는 것에 대한 기술적인 대안에 대한 수요가 폭증하게 되었다. 이러한 수요는 90년대 후반의 자동 번역기술이 적용 방향과 CAT의 변화에 대한 요구의 촉매라고 하겠다.

인터넷의 신화에도 불구하고 많은 산업 기반이 다국적 산업의 한계를 갖게 된 이유 중 가장 큰 것은 언어의 문제이다. 다국적 산업 환경 하에서 언어의 문제는 사업 모델 자체를 바꿀 수 있는 정도로 중요한 문제가 되었다. 그렇기 때문에 향후 관련업계의 가장 큰 이슈는 인터넷을 통한 다국적 산업 환경을 지원하는 커뮤니케이션 솔루션이 될 것이다. 위에서 언급한 양적인 홍수와 질적인 급류의 구분이 명확해 질 것이며, 필요에 의해서는 소스 자체가 조절될 수도 있을 것이다. 양적인 홍수에 대해서는 번역 품질의 상하 개념이 명확해져 이해 가능한 번역에 대한 요구를 자동 번역이 해결해야 할 것이며, 자동 번역은 기존의 범용적인 언어의 연구에서 보다 특정한 분야에 적합한 형태의 연구가 진행될 것이다. 이는 개념적으로는 전통적인 방식에 대한 역방향적 접근이라고 할 수 있으나, 기술적으로는 전통적인 연구 결과의 계승이라고 하겠다. 번역의 질적인 급류는 CAT 시스템이 담당해야 할 것이다. CAT 시스템은 기존의 번역 프로세스의 부하를 해결하는 방향으로 연구가 진행될 것이다. 대부분의 시스템이 C/S환경에서 동작하여 일련의 프로세스를 구조화 시키고 프로세스 전반을 효과적으로 관리함으로써 생산성의 극대화를 꾀할 것이며, 대용량의 데이터를 동시 다발적으로 생산 및 관리할 수 있는 연구가 진행될 것이다.

한국어를 중심으로 하는 자동 번역 시장은 2000년에는 약 300억원 정도 되었고, 자동 번역 시스템 및 CAT 시스템이 나옴에 따라 2001년에 약 500억원, 2002년에 1,200억원, 2003년 2000억원, 2004년에는 약 3000억원 정도로 확대되리라고 전자신문은 예상하고 있다. 또한 한국소프트웨어산업협회 산하 언어정보산업협의회는 세계 언어정보 관련 시

장 규모가 2001년 430억 달러에서 2004년 1545억 달러로 연평균 50% 이상의 고성장이 되리라고 예상하고 있다.

4. 결 론

지금까지 자동 번역과 CAT의 현황 그리고 앞으로의 방향에 대하여 살펴보았다. 자동 번역은 아직까지 사람이 번역한 것보다는 질적인 면에서 미흡하다. 지금까지 이런 차이를 극복하기 위해 연구된 많은 방법들을 살펴보았으며, 완전 자동 번역이 아닌 번역가의 번역 작업을 지원하는 도구인 CAT를 살펴보았다.

앞으로 컴퓨터가 인터넷과 웹 환경에서 사용됨에 따라 컨텐츠(Contents)와 지식의 공유가 더욱 활성화되고 그 지식에는 영어, 일어, 중국어 등 외국어로 된 지식들이 더욱 많아져 실시간 자동 번역에 대한 요구가 증가될 것으로 예상된다. 또 음성 인식, 음성 합성 기술과 접목되어 실시간 통역 서비스가 제공될 것이다. 기업이 자사 제품을 여러 나라에 출시하기 위한 방법론이 L10N에서 G11N으로 변화함에 따라 기업의 제품이 출시되면서 여러 나라 언어로 된 제품과 매뉴얼 및 기술 문서들이 동시에 나와야 하는 환경으로 변하고 있다. 이런 변화에 따라 자동 번역 시장에서 질 높은 자동 번역과 CAT 도구에 대한 수요는 더욱 절실해 지고 있다. 자동 번역의 질을 높이고 매체의 다양성을 추구하기 위해 대학, 연구소 그리고 기업간의 공동 연구도 활발히 이루어지리라고 전망된다.

참고문헌

[1] Barr, A., Feigenbaum, E.A., The Handbook of Artificial Intelligence, Vol. 1, William Kaufmann Inc., 1981.

[2] John Hutchins, "Retrospect and Prospect in Computer-based Translation.," MT Summit, 1999.

[3] Bar-Hillel, Y., "The present status of automatic translation of languages.," Advances in Computers 1, pp. 91-163, 1960.

[4] Languages and Machines, Computers in Translation and Linguistics, ALPAC-Report(Automatic Language Processing Advisory Committee Report), Washington, D.C., 1966.

[5] Bostad, D., "Machine translation in the USAF.," Terminologie et Traduction no.1, pp. 68-72, 1986.

[6] Pigott, I.M., "Systran machine translation at the EC Commission : present status and history.," Luxembourg CEC, January 1988.

[7] Nirenburg, S. et al. "Machine translation: a knowledge-based approach." San Mateo, Ca., Morgan Kaufmann, 1992.

[8] Rimon, M. et al., "Advances in machine translation research in IBM.," MT Summit 3, pp. 11-18, 1991.

[9] Church, K.W., Miiiiiiercer, R. L., "Introduction to the Special Issue on Computational Linguistics Using Large Corpora.," Computational Linguistics, Vol. 19, No. 1, pp. 1-24, 1993.

[10] Sumita, E. et al., "Translating with examples: a new approach to machine translation.," TMI-90, pp. 203-212, 1990.

[11] Brown, P. et al., "A statistical approach to language translation.," Computational Linguistics 16, pp. 79-85, 1990.

[12] Hutchins, W.J., "Machine translation: past, present, future.," Chichester (UK): Ellis Horwood, 1986.

[13] Harris, Z.S. 1954. "Transfer grammar.," International Journal of American Linguistics 20, pp. 259-270, 1954.

[14] Bourbeau, L. et al., "Bilingual generation of weather forecasts in an operations environment.," COLING 90 (1), pp. 90-92, 1990.

[15] Isabelle, P. and Bourbeau, L., "TAUM-AVIATION: its technical features and some experimental results.," Computational Linguistics 11(1), pp. 18-27, 1985.

[16] Kaplan, R.M. and Bresnan, J., "Lexical-functional grammar: a formal system for grammatical representations.,"

Bresnan, J. (ed.), The mental representation of grammatical relations. Cambridge: MIT Press, 1983.

[17] Boitet, C., "Current state and future outlook of the research at GETA.," MT Summit 1987, pp. 26-35, 1987.

[18] Arnold, D.J., "Eurotra: a European perspective on MT.," Proceedings of the IEEE 74(7), pp. 979-992, 1986.

[19] Sadler, V., Working with analogical semantics: disambiguation techniques in DLT., Dordrecht: Foris, 1989.

[20] Teruko M., The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains, COLING-92, 1992.

[21] Muraki, K., "PIVOT: two-phase machine translation system.," MT Summit 1987, pp. 81-83, 1987.

[22] Wheeler, P., "LOGOS.," Sprache und Datenverarbeitung 9(1), pp. 11-21, 1985.

[23] Mann, J.S., "Get Smart! Industrial strength language processing from Smart Communications.," Language Technology 3, pp. 12-15, September/October 1987.

[24] Lynn E. W., Advantages and Disadvantages of Translation Memory : Cost/Benefit Analysis, Einfhrung in die Informatik II, Vol 6, 1998

[25] Hutchins W.J., "Machine translation and human translation: in competition or in complementation?," International Journal of Translation Vol.13, no.1-2, pp. 5-20, Jan-Dec 2001.

[26] Woszczyna, M. et al., "Recent advances in JANUS: a speech translation system.," TMI-93, pp. 195-200, 1993.

[27] Wahlster, W., "Verbmobil: translation of face-to-face dialogs.," MT Summit 4, pp. 127-135, 1993.

[28] Gerald D., "Translation Memory:Concept, products, impact and prospects," Major project report South Bank University School of Electrical, Electronic and Information Engineering, 1995.

[29] Achim B., "Workflow using linguistic technology at the Translation Service of the European Commission," EAMT Workshop, Geneva, pp. 7-18, April 1998.

박 주 형

1996 한국과학기술원 전산학과(학사)
1994~1999 (주)코아기술 연구원
1999~현재 (주)클릭큐 부설 기술연구소 소장
관심분야:자연어처리, 자동 번역, 번역 메모리, CAT
E-mail:norther@clickq.com

이 창 우

1999 한국외국어대학교 영어학과(학사)
1999~2000 The English Hunters
2000~현재 (주)클릭큐 부설 기술연구소 책임연구원
관심분야:Web Globalization Architectre
E-mail:hardtype@clickq.com

강 명 주

1988 동국대학교 컴퓨터공학과(학사)
1991 동국대학교 컴퓨터공학과(석사)
1991~1993 아이비 컴퓨터 시스템
1993~1997 제주산업정보대학 컴퓨터 정보계열 전임강사
1997~2000 Tyco/Fire & Security/ 동방전자산업
2000~현재 (주)클릭큐 부설 기술연구소 책임연구원
관심분야:자연어처리, 데이터베이스, 정보 검색, 실시간 시스템
E-mail:rnjkang@clickq.com