技術解說

자동번역기술의 현황과 전망(Ⅰ)

朴 昌 浩 · 李 基 式
(KAIST 시스템공학센터 선임연구원 · 한국전산원 연구위원)

■차 례■

1. 서 론
2. 기계 번역 연구의 발자취
3. 기계 번역 기술의 현황
 가. 기계 번역의 방식
  (1) Direct 방식
  (2) Transfer 방식
  (3) Pivot 방식
 나. 기계번역의 요소기술
  (1) 문장의 해석
  (2) 변환및 생성
  (3) 기계사전
4. 기계번역 시스템의 현황
 가. 시스템 기술
  (1) 격해석 Transfer 기반의 시스템
  (2) 개념구조 Transfer기반의 시스템
 나. 시스템의 현상
  (1) 기계번역의 특질

1 서 론

오늘날의 국제사회는 문명의 급속한 발달에 따라 새로운 각종정보가 대량으로 창출되고 있을 뿐 아니라, 국가간의 교역량이증가함에 따라서 정보의 소통량과 전달의 폭에 있어서도 커다란 변화를 겪고 있다.

이러한 각종정보는 진보된 정보통신기기와 세계각처를 연결하는 통신망을 통하여 서로 다른 언어권국가에도 대량으로 고속전달되어 교역과 공동발전을 크게 촉진시켜 나가고 있다. 그러나, 근래에 이르러 급격히 증대되고 있는 정보의 소통량은 언어의 장벽에 의한 의사소통및 정보전달의 문제를 더욱 가중시킴에 따라 기존의 인간에 의한 번역에만 의존할 수 없는 새로운 상황에 직면케 되었다. 따라서 인간은 機械(컴퓨터)에 의한 자동번역기술의 개발을 더욱 강화하지 않을 수 없게 되었다. 특히 유럽, 캐나다와 같이 多言語를 共用하거나 우리나라, 중국, 일본등과 같이 고립어인 단일언어를 사용하는 국가에서는 번역이 절실한 과제로서 부각되고 있으며 번역수요 또한 해마다 크게 증가하고 있는 추세이다. 또한 high-quality의 고속정보 처리를 요구하는 향후의 정보화사회에서도 언어에 의한 각종정보의 원활한 소통이라는 관점에서 번역의 역할이 크게 증대될 전망이어서 세계각국은 자동번역에 대한 관심을 집중시키고 연구개발을 강화하고 있다.

따라서, 유럽, 캐나다, 일본, 미국등 세계의 주요국가는 대량의 정보를 보다 신속히 번역할 수 있는 새로운 지원시스템으로서 機械翻譯(Machine Translation)에 대한 연구개발을 더욱 적극적으로 추진하고 있는데, 1970년대에 부분적

으로 실용화되기 시작한 인도·유럽어족(Indo-European)중심의 동일어족(語族)간의 기계번역 기술이 최근에는 컴퓨터의 H/W및 S/W 성능향상, 언어학의 연구성과를 바탕으로 한自然言語處理(Natural Language Processing)기술의 발전에 힘입어 영어-일본어 기계번역시스템과 같은 전혀 다른 語族間의 번역기술도 實用化되고 있는 수준에까지 도달하였다. 또한, 매우 한정적인 번역기능을 가졌던 기존의 기계번역 시스템을 개량하려는 적극적인 시도가 꾸준히 이루어지고 있으며 특히, 1980년대에 미국, 일본, 유럽국가가 참여하기 시작한 人工知能型의 제5세대 컴퓨터 개발계획에서도 知能機械의 핵심기능으로서 知能的인 번역기술을 채택함으로써 자연언어의 의미/문맥적인 처리가 가능한 人工知能的인 접근방법에 대한 연구도 강화되고 있다. 한편, 언어장벽을 절실하게 겪고 있는 일본은 1985년부터 자동통역전화시스템을 목표로 하는 15년계획의 대단위 프로젝트를 일본 우정성(체신부)의 주관하에 착수하여 그들이 필요로하는 특정언어 pair간의 음성번역 시스템의 연구개발을 진행시키고 있다.

2 기계번역연구의 발자취

기계(컴퓨터)에 의해 인간의 번역작업을 대행할 수 있도록 하려는 기계번역 기술의 연구가 착수된지도 30여년이 지났다. 컴퓨터의 처리능력을 지나치게 믿었던 초기의 연구자들은 자연언어가 가지는 독특한 문장구조, 의미, 문맥등에 관한 갖가지 문제가 단순히 산술적인 기호의 연산으로 해결될 수 없다는 것과 당시의 컴퓨터처리능력으로서는 무한한 언어 data를 해석하고 처리하는 것이 거의 불가능하다는 점을 인식하게 되었다.

기계번역에 대한 당시의 이러한 비관적인 견해는 미국 MIT의 기계번역 연구자 Bar Hillel이 언어번역에 관하여 1960년에 발표한 그의 보고서에서 밝혔듯이 고품질의 자동번역은 불가능하다고 지적한 것과 1966년의 ALPAC보고서가 기계번역의 실용성을 부정하였다는 점에서도잘 나타나고 있다. 따라서 기계번역의 연구는 일시적으로 크게 퇴조하게 되었다. 그러나 多言語社會인 유럽과 캐나다에서는 기계번역기술이 매우 중요시되어 특히 영어-불어간의 기계번역에 대한 끊임없는 연구가 이루어 졌으며 미국에서도 George-town大學이 러시아어를 영어로 번역하는 기계번역기술의 연구를 계속하였다. 이러한 초기의 발달과정을 거친 기계번역의 연구는 컴퓨터처리능력이 비약적으로 향상되고 언어이론이 발달하였을 뿐 아니라 사회적으로도 번역수요가 급격히 증가함에 따라서 기계번역의 연구개발이 새로운 변혁을 맞이하여 1970년 대에는 SYSTRAN, WEIDNER, ALPS, TITUS와 같이 실용화된 시스템이 보급되었으며, 이를 계기로 기계번역의 연구개발이 가속화되어 현재에 이르고 있다. 한편, 기계번역의 바탕기술인 자연어처리에 대한 연구는 1960년대에는 주로 構文論(Syntax)을 중심으로 수행되었으며 언어학적인 배경도 N. Chomsky의 변형문법(Transformational Grammar)이 주류를 이루었다. 그러나 ALPAC보고서의 충격이후 1970년대부터는 자연어처리의 연구테마가 구문론에서 意味論(Semantics)으로 이동하여 解析意味論과 生成意味論에 관한 연구가 수행되었다. 최근에는 文章 또는 문단간의 의미추출및 문맥처리가 가능하도록 상황 의미론의 연구를 비롯하여 선택의미론(preference semantics), Scripts이론등을 기계번역기술에 적용하려는 연구가 진행되고 있다. 또한 언어가 포함하고 있는 보편적 의미표현(pivot)을 매체로 하여 개념간의 의미전달을 목표로 하는 PIVOT개념의 번역방식이 연구되고 있으며, 이러한 연구를 지원하는 software공학의 측면에서는 언어/비언어적 지식의 표현기술(knowledge Representation)과 언어이해를 위한 추론메카니즘(Inference mechanism)이 인공지능의 영역에서 다루어지고 있다.

3 기계번역 기술의 현황

N. Chomsky의 變形文法理論을 토대로 하여

1960년대에 미국을 중심으로 매우 활발히 수행되었던 기계번역의 연구는 構文的인 모호성과 多義性을 갖는 어휘의 문제등으로 인하여 句構造規則으로 입력문의 문장구조를 분석할 수 없었기 때문에 초기의 변형문법에 기반한 기계번역은 결국 실패하고 말았다. 그러나 1970년대에 접어들면서 컴퓨터의 연산속도및 처리용량의 비약적인 향상, 소프트웨어기술 및 자연언어처리를 위한 언어해석이론의 진보, 지능기계를 향한 인공지능의 연구성과가 기계번역기술의 연구개발에 새로운 변혁을 불러 일으켰다.

• 컴퓨터하드웨어 및 소프트웨어 기술의 발전은, 자연언어의 어학적인 제현상을 분석하고 해석하는데 필요한 수많은 linguistic rules (언어적 규칙)과 언어data의 처리속도를 향상시켰으며 언어data의 처리용량도 비약적으로 향상시켜 놓았다. 또한 이론적으로 가능했지만 컴퓨터상에서 실험할 수 없었던 갖가지 자연언어처리에 관련한 요소기술의 실현이 가능하게 되었다. 예를 들면, 1960년대의 구문해석은 비교적 단순한 몇개의 언어정보(형태소, 품사등)만으로 수행되었던 것에 비하여 1970년대 이래의 구문해석은 더많은 언어정보를 사용하여 복잡하고 세밀한 解析規則을 작성함으로써 精度높은 解析을 가능하게 하였다.

• 자연언어처리를 위한 언어해석이론의 진보는, 인간이 사용하는 무수한 문장을 분석하고 의미를 추출하여 정확하고 유연한 번역을 할 수 있도록 morphology, syntax, semanties등에 대하여 전반적인 언어현상을 규명하게 되었으며, 이를 컴퓨터상에 실현할 수 있도록 Formal Grammar로 작성할 수 있는 methodology 및 언어해석model을 제시해 주게 되었다. 1970년대이후에는 특히 변형문법을 탈피하여 언어생성의 관점보다는 언어의 심층구조분석에 의한 언어의 인식을 중시하게 되었다. 따라서 언어해석이론에 대한 연구는 의미해석(Semantic Analysis)의 연구에 더욱 주력할 수 있게 되었으며 언

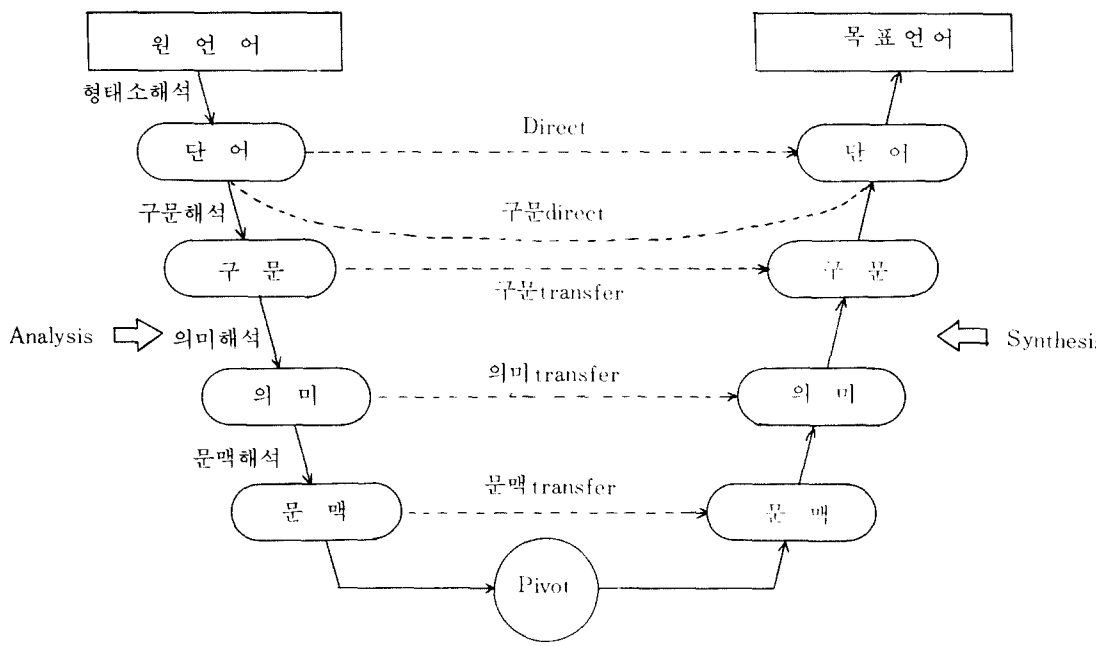


그림 1　기계번역의 처리과정

어이해level의 계산언어모델(Computational Linguistic Model)이 구축될 수 있도록 의미론의 연구를 활성화시켰다.

• 인공지능연구의 진보는, 자연언어가 포함하고 있는 의미구조를 분석하고 심층의 의미를 추출할 수 있도록 文의 意味表現및 知識表現(Knowledge Representation)기술을 제공하게 되었으며, 이들 의미/지식표현으로 작성된 Knowledge Base를 바탕으로 하는 推論메카니즘(Inference mechanism)을 지능적인 기계번역 기술에 도입할 수 있게 하였다.

다음은 각종 기계번역시스템의 개발에서 채택하고 있는 번역방식과 시스템을 실현하는 기계번역 요소기술을 소개한다.

가. 기계번역의 방식

현재 세계적으로 연구개발되었거나 진행중인 기계번역시스템의 번역방식(Translation methodology)을 엄밀히 분류하면 그 시스템의 수효만큼 존재한다고 볼 수 있다. 그러나 이것을 특정한 관점에서 본다면 그 번역방식을 분류할 수 있다. 가장 대표적인 분류방식은 언어의 해석방법과 각 언어의 의존성에 따라 분류하는 것으로써 'Direct방식', 'Transfer방식', 'Pivot방식'으로 대분류된다.

(1) Direct 방식

direct방식은 원언어(번역대상문)와 목표언어(번역결과문) 간에 구문적인 변환을 거치지 않고 직접 번역어를 선택하는 방식이다. 즉 단어의 결합관계나 구문정보등을 기계사전에 직접적으로 記述하고 이것을 번역시에 참조하여 목표언어를 출력하는 것으로써 中間表現은 사용하지 않는다. 원언어(Source language)의 分析정도에 따라서 단어, 구문, 의미direct로 분류된다. 목표언어(target language)의 결정은 기계사전에 등록된 것으로 한정하고 있으며 語順의 변경이나 구문적인 相異를 해결하기 위해서는 많은 번역규칙을 작성해야 한다. 따라서 이 번역 방식은 유연한 구문의 표현이나 역어의 선정에 상당한 제약을 갖고 있으며 영어-불어 또는 한국어-일본어와 같이 유사한 구문구조를 갖는 語族을 대상으로 하는 Machine-aided Translation 시스템에 적용할 수 있다.

(2) Transfer 방식

현재 실용화되어 있는 대부분의 번역시스템은 이 번역방식을 채택하고 있다. transfer 방식은 원언어와 목표언어사이에 중간표현을 두어 구문적 또는 의미적 변환을 행하는 방식으로서, 입력된 原文을 중간표현으로 변환하고 이것을 목표언어의 중간표현으로 재변환하여 譯文을 생성한다. 입력원문은 매우 다양한 構文을 무수히 갖고 있지만 문장을 分析하여 그 文章의 중간표현을 추출하면 변환rule은 크게 감소시킬 수 있다. 현재 채택되어 있는 transfer 방식은 解析level에 따라 구문transfer와 의미transfer가 있으며 앞으로는 문단간의 의미변환을 가능케 하는 문맥transfer의 적용도 예상된다.

① 구문transfer : 입력문을 분석하여 문장의 구성요소들(주어, 술어, 접속사등)간의 구문적 결합관계(Syntactic relation)를 출력문의 구문tree로 변환하는 방식이다. 영한번역의 경우, 'S+V+O'의 문구조를 가진 입력영문을 'S+O+V'의 한국어 문구조로 변환시키는 구문규칙을 작성하여야 한다. 여러개의 구문변환규칙은 부분적인 의미변환을 수행할 수 있으나 변환규칙의 수효가 크게 증가되기 때문에 의미분석에는 적합하지 않다.

② 의미transfer : 다수의 구문변환rule을 적용해야만 의미변환을 할 수 있는 언어 현상은, 의미적인 결합관계를 기술할 수 있는 Semantic network등을 중간표현으로 기술하는 이 방식을 사용함으로써 번역의 효율을 높일 수 있다.

이 방식은 文構造는 다르지만 意味가 같은 문장, 즉 ① I gave her a gift., ② I gave a gift to her., ③ A gift was given to her by me. 와 같은 경우에 적용하면 의미적으로 동일한 중간표현을 얻게 되므로 의미변환rule을 감소시킬 수 있다. 그러나 기계사전에 등록하는 단어에 의미속성(Semantic features)을 부여하고, 문장의 구성요소들간의 의미적인 결합관계를 구

축해야 하는등 기계사전과 변환규칙의 작성이 매우 복잡해진다. 따라서 번역의 質은 향상되는 반면에 문장의 해석과 생성과정이 길어진다.

(3) Pivot방식

이 번역방식은 원언어와 목표언어 사이의 transfer과정을 거치지 않고서 문장이 갖고 있는 심층적인 意味表現인 Pivot을 경유하여 개념적인 번역을 행하는 것이다. 즉, pivot방식은 의미변환방식을 최대한 확대하여 적용하면 모든 언어에 공통적인 개념구조를 추출할 수 있다는 착상에 근거하고 있는데, 목표언어는 단지 pivot에 의한 원언어의 개념구조분석결과에 따라서 생성할 수 있다는 것이다. pivot방식을 사용하면 이론적으로 볼 때 언어의 종류에 관계없는 多言語間 機械翻譯(Multilingual Machine Translation)도 가능하지만, 이러한 완전한 pivot이 존재하는가는 현재까지 증명되지 않고 있다.

그러나, 이러한 번역방식들중에서 실제로 시스템化된 것은 transfer방식이 대부분이며 일부의 구문direct방식이 있을 뿐이다. 또한, 이 2가지 번역방식에 의해 개발되었거나 진행중인 현재의 시스템들은 채택된 번역방식의 level에 따라서 번역의 속도나 시스템의 확장성, 기계사전및 번역규칙의 복잡성에 커다란 차이가 발생하고 있다. 이 차이는 채택된 번역방식이 cost/effectiveness의 면에서도 상당한 영향을 미치고 있다는 점을 나타내는 것이기 때문에 번역시스템의 이용자에게는 커다란 의미가 있다. 다만, 어떠한 번역방식을 채택하든지간에 이론적으로는 번역품질이 같을 수 있는데, 동일한 번역품질의 譯文을 生成하기 위해서는 기계사전과 해석rule의 구축범위가 매우 달라진다는 점이다.

나. 기계번역의 요소기술

인간이 사용하는 언어를 기계적으로 '좋은' 번역을 행하려면 자연언어처리를 위한 각 요소기술을 개량하고 그로 인해서 번역방식level을 향상시켜야 한다.

즉, 기계번역은 자연언어처리를 위한 요소기술인 解析(分析과 理解)－言語變換－生成의 과정을 거치게 되므로 각 처리단계의 요소기술이 균형있게 연구되어 있지 않으면 고품질의 실용적인 번역시스템을 기대할 수 없다.

따라서 본 절에서는 번역시스템을 구축하는 요소기술인 文章의 解析, 변환기술 및 生成기술과 기계번역의 品質에 매우 큰 영향을 미치는 기계사전에 대하여 소개한다.

(1) 文章의 解析

① 형태소분석(Morphological Analysis)

기계번역을 위한 제1단계 언어 분석으로써 입력된 문장을 구성하는 단어string으로부터 文의 최소단위인 형태소(morpheme)로 분리하는 것이다. 형태소분석은 上位解析level인 구문분석의 처리범위를 어디까지하느냐에 따라서 그의 분석결과가 달라지게 되는데, 분석된 형태소string을 나열하여 단순 출력하는 것과 단어 또는 句節단위로 분리하여 형태소정보를 동시에 출력하는 2가지 방법이 있다. 형태소분석은 형태소단위로 분리하기 위하여 참조되는 형태소사전만으로 해결되는 것이 아닌데, 한국어나 일본어같은 교착어는 활용 및 첨가에 의한 어형변화가 무수히 발생하여 이들을 형태소분석하는 것이 매우 난해하며 영어나 불어등에서도 活用語, 복합어, 파생어의 문제등 대형의 기계사전에 등록하지 않는 한 완전한 분석이 어려운 경우가 적지 않다. 한국어의 형태소분석은 명사, 대명사, 지시어등으로 구성된 非活用語辭書와 동사·형용사등으로 구성된 活用語辭書 및 활용어미변화table을 구축하고 각 형태소간의 결합가능성을 표시하는 접속table을 사용하여 수행하고 있으나 완전한 分析은 어렵다. 또한, 단어의 띄어쓰기가 확립되어 있다고 하더라도 작문자의 판단이 정확하지 못하여 띄어쓰기Error가 발생하는등 형태소분석프로그램의 self-function만으로는 정확한 분석이 어려운 경우가 있는데 이것은 higher level인 構文分析技能을 도입하거나 構文分析이상의 分析level에서 처리되도록 하고 있다.

②구문분석(Syntactic Analysis)

문장을 구성하고 있는 單語, 句, 節등이 구문적으로 서로 어떤결합을 하고 있으며 각 구문요소(syntactic element)들이 문장내에서 어떤 역할을 하고 있는지를 分析하는 것이다. 즉, 구문해석은 문장의 구조를 해석하는 것으로써 분석에 필요한 Grammar rules과 machine dictionary가 구축되어 있어야 한다. 그러나, 자연언어의 문구조는 많은 ambiguity를 포함하고 있어서 分析方法에 따라서는 한 문장에 수많은 구문해석이 가능하기 때문에 번역규칙과 기계사전의 記述방식이 매우 중요하다. 따라서 수개에 불과한 정확한 분석결과를 얻기 위하여 형태소분석으로부터 더 많고 정확한 분석결과를 제공받아 이용하며 가능한 구문분석tree의 수효를 최소화하기 위하여 의미해석기능을 부가하기도 한다. 구문분석에 사용된 문법규칙이 Context Free Grammar(CFG)에 의한 것인 경우는 보다 정밀한 구문분석을 위하여 분석rule을 증가시킬뿐 아니라 기계사전내에 의미속성을 부여하고 있다. 또한 tree-structure를 대상으로 한 변환규칙에 의해 매우 정밀한 분석과 문맥의존을 실현하고, 部分文法의 개념을 도입하여 문법규칙간에 제어를 용이하게 함으로써 分析規則의 수효가 증가되지 않도록 하고 있다. 한편, 기계사전의 측면에서는 단어자체가 갖고 있는 언어적인 의존성을 규칙화하여 사전에 기술함으로써 구문분석의 精度를 향상시키고 있다. 구문분석에 적용하는 문법이론에는 여러가지가 있는데 句構造文法(phrase structure Grammar), ATN(Augmented Transition Network), 격문법(Case Grammar), 변형문법(Transformational Grammar), Montague Grammar, Lexical Functional Grammar, 일반구구조문법(GPSG)등이 대표적이다. 이 문법에 의해 구문분석된 문장은 tree-structure로 나타낼 수 있는데, parsing하는 방법은 bottom-up, top-down 등 여러가지가 있다.

③의미해석(Semantic Analysis)

의미해석은 단위문장이 갖는 의미를 분석하는 것으로써, 문장내의 多義性를 제한하거나 의미적으로 성립되지 않는 문구조를 삭제시키는 역할을 한다. 또한 구문분석에서 산출된 여러개의 분석결과를 해석하여 정확한 문구조를 나타내는 구문tree를 선정한다. 의미해석기술은 1970년대부터 활발히 수행된 인공지능의 연구성과를 바탕으로 하여 최근에 많은 연구가 진척되고 있는데, 일반적으로 의미해석은 의미의 파악에 사용되는 방대한 각종 지식을 필요로 하고 있기 때문에 단어가 갖고 있는 의미속성도 매우 정밀히 분류하고 이것을 계층구조화하여 의미해석규칙에 기술하여야 하며, 의미해석에 사용될 의미표현에는 언어적인 지식은 물론 인과관계, 자연현상의 원리, 인간의 행동양식등 言語外的인 지식도 포함되어야 한다. 따라서 의미해석규칙은 방대한 지식을 기반으로 작성되는 것으로써 Frame, Semantic network, Predicate Logic등의 의미표현을 production system과같이 pattern matching시킬 수 있는 해석규칙을 사용하여 의미를 분석한다.

④문맥해석(Pragmatic Analysis)

의미해석이 단위문장을 대상으로 하는데 반하여 문맥해석은 문장 또는 문단간에 연결된 언어현상을 분석하는 것이다. 문맥해석은 기계번역은 물론 자연언어처리기술에 바탕을 두고 있는 모든 해석기술중에서 가장 연구가 미흡한 부분으로서 문장사이에 내포된 의미, 대용및 생략된 어법의 처리, 뉘앙스등 담화이해(Dialogue)에 직결된 가장 난해한 요소기술이며 한편으로는 가장 知能的인 언어처리기술이다. 특히 Voice recognition에 의한 自動通譯機의 경우에는 話者의 뉘앙스나 지시어 및 생략어의 번역이 매우 중요한 비중을 차지하게 되는데 문맥해석기술이 구축되어 있지 않으면 처리가 불가능하다. 특히 한국어의 경우에는 주어를 생략하거나 인간의 종적인 관계에 의한 경어법의 사용이 매우 발달되어 있는 반면, 복수나 性의 구분이 뚜렷하지 않기 때문에 기계번역의 관점에서는 문맥처리가 필요하다. 따라서 이를 해결하기 위해서는 단위문장을 의미해석할 수 있어야 하고, 각 단위문장의 의미를 문장 나아가 문단간의 의미(Contextual meaning) level로 결합하여 문맥

적 관계를 추론할 수 있어야 한다. 이것은 Inference메카니즘으로 problem solving하는 것을 의미하는데 이를 가능케 하기 위해서는 문맥처리가 가능한 추론기계(Inference Machine)를 제작하고, 추론기계의 입력data가 될 언어적지식과 비언어적지식(상황정보, 상식, 경험적지식등)의 상호작용관계를 Model화 할 수 있어야 하며, 계산모델화된 지식들은 知識表現시스템에 의해 Knowledge base로 구축하여 추론기계가 사용할 수 있도록 해야 한다. 현재는 이에 어느정도 대응해 보려는 초기의 노력으로서 인지과학적인 관점의 연구가 시도되고 있으며 상황의미론과 심리언어학에 대한 연구도 진행되고 있다. 또한 컴퓨터에 의해 실현가능성이 확인된 것으로는 Yale대학의 Schank교수가 제안한 Scripts이론과 New Mexico대학의 Y. Wilks 가 연구한 선택의미론등이 있는데, 이들은 모두 현실세계의 폭넓은 지식을 어떻게 처리하는가하는 문제를 남겨두고 있다. 따라서 이를 발전시키기 위해서는 machine learning, knowledge acquisition, knowledge representation과 추론mechanism은 물론 컴퓨터의 parallel architecture에 대한 연구도 병행되어야 한다.

(2) 변환 및 생성

(1) 변환(Transfer)

기계번역에 있어서의 言語間 transfer는 대상이 되는 양언어가 갖고 있는 언어적인 gap을 제거하기 위한 처리과정이다. 언어적인 gap은 각 언어를 사용하는 인간마다 서로 다른 언어생활양식과 문화적배경이 있기 때문에 발생한다. 이 gap은 표층적으로는 단순히 언어의 표기방법이나 문구조가 다르다는 것이지만 심층적으로는 사고방법과 개념의 발상이 다르기 때문에 생기는 것이다. Transfer는 구문level의 중간표현으로 행해지는 것과 의미level의 중간표현으로 행해지는 것이 있다. 구문transfer의 특징으로서는, 변환단계에서 번역어를 결정하는 lexical transfer과정과 문구조를 변동시키는 structural transfer과정을 동시에 행하는 점을 들 수 있다. 따라서 이 단계에서 번역어가 결정되기 때문에 譯語선택의 자유도가 떨어져 유연한 對譯語를 생성하는 것이 어렵다. 또한 의미transfer의 특징으로서는, 원언어가 갖는 개념에 의해서 구축된 의미표현을 목표언어의 개념에 의해서 재구성된 여러가지 의미표현으로 변환함으로써 譯語의 선택이 자유롭게 되어 유연한 의미전달이 가능한 점을 들 수 있다. 그러나 이렇게 의미변환을 시키기 위해서는 Analysis와 Generation처리단계에 더 많은 load를 주게 되지만 對譯語의 선택이 문장의 생성단계에서 행하여지므로 유연한 번역이 가능하다.

(2) 생성(Generation)

변환과정을 거쳐 얻어진 목표언어의 중간표현으로부터 번역문을 작성해내는 기계번역의 최종단계이다. 그러나 문장의 생성방법은 채택된 번역방식의 종류에 따라 크게 차이가 있다. 구문transfer방식을 사용한 번역에서는, 목표언어의 중간표현이 원언어보다는 목표언어에 가깝게 되므로 對譯語도 제한된 범위내에서 선정된다. 따라서 단순한 語順의 변동이나 活用形 처리로써 번역문장을 생성한다. 또한 의미transfer 방식을 채택한 번역에서는, 목표언어의 중간표현이 의미표현이므로 이러한 의미를 적절하게 표현하는 문장의 생성규칙을 이용하여 인간의 發話과정처럼 기계적으로 文生成을 할 수 있다. 생성규칙에 의한 譯語의 선택은 의미적으로 가장 적합한 것으로 할 수 있는데, 문장을생성할 때 對譯語辭書를 검색하여 부여된 의미를 가장 적절히 나타내는 譯語를 선택한다. 또한, 번역어를 최종적으로 결정할 때는 연결되는 단어나 句가 語用論적으로 적합한지를 나타내는 word-concurrence관계나, 인접하는 것이 옳은지를 표시하는 인접관계(correlation)를 사용한다.

(3) 기계사전

기계번역을 위해서는 각 단어가 갖고 있는 여러가지 언어적인 특질과 譯語情報를 제공하는 기계사전이 구축되어 있어야 한다. 즉 기계사전은 해석-변환-생성에 관계하는 번역rule의 각 element를 제공하는 언어정보의 집합체로서 번역의 質에 가장 영향을 크게 미치는 요체의 하

나이다. 사전에 기술된 정보량에 따라서는 기계번역의 각 처리과정에서 발생하는 모호성을 제거할 수도 있다. 실용화를 목표로 하는 시스템에서는 대규모의 기계사전을 작성하는 것이 중요한데, 많은 언어정보를 어떻게 균일하게 작성하여 기계번역의 처리과정에서 용이하게 참조할 수 있도록 하는가에 따라서 기계번역의 실용성이 크게 좌우된다. 이러한 기계사전은 채택한 번역방식에 의해서도 달라지게 된다. 즉 구문transfer방식인 경우, 원언어가 목표언어의 중간표현으로 변환되는 과정에서 譯語가 결정되므로 기계사전이 직접 참조되지만, 의미transfer방식의 경우에는 단어와 意味素의 관계가 記述된다. 따라서 구문변환에 사용되는 기계사전은 원언어에 1대1로 대응하도록 역어선정이 한정되는 결점이 있으며, 의미변환에 사용되는 기계사전은 원언어에 의미적으로 대응할 수 있는 意味素를 어떻게 설정하는가 하는 난점을 갖고 있다. 현재의 번역시스템에서 채택하고 있는 기계사전은 단어의 사용기준에 따라, 분야에 의존하지 않는 常用語및 非常用語사전과 전문분야에 의존하는 각 분야별 전문어사전이 있으며, 사용자만이 사용하는 개인용어사전도 이용되고 있다.

4 기계번역시스템의 현황

가. 시스템기술

번역시스템의 연구는 1980년대에 접어 들면서, 종래의 Syntax-base의 Machine Aided T-

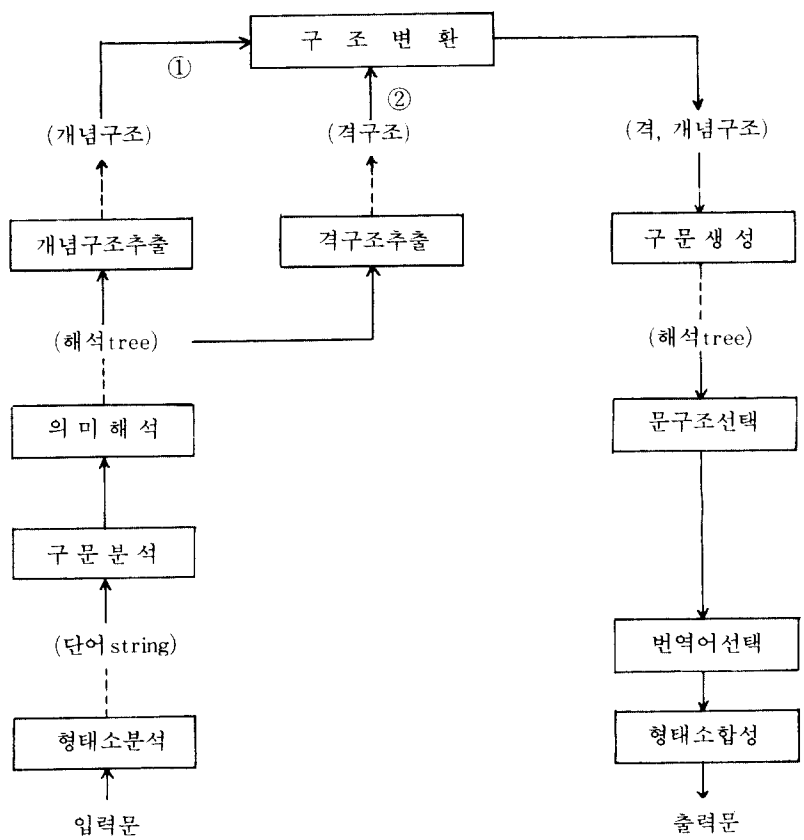


그림 2 Transfer 방식의 번역과정

ranslation으로부터 점차 문장의 유연한 구문처리와 의미해석을 중시하는 기술개발로 발전되어 가고 있다. 따라서 実用化를 전제로 하는 최근의 기계번역에서 채택하고 있는 번역방식은 Transfer方式이 일반화되고 있다. 이 방식에 의한 시스템기술이 현재 EC의 EUROTRA 프로젝트, 프랑스의 GETA프로젝트, 미국과 일본의 연구기관 및 기업체들이 진행하고 있는 각종 번역프로젝트등에 의해 구축되고 있다. Transfer방식의 번역처리과정은 input sentence 의 解析을 통하여 문장을 구성하는 각 단어성분간의 Structural/semantic relation을 나타내는 中間表現을 산출하고, 이 중간표현을 목표언어의 중간표현으로 변환하여 문장의 生成規則에 의해 번역문을 生成한다.

시스템의 처리기술로 실현되고 있는 Transfer방식은 각 시스템의 설계방향에 따라 차이가 있지만 크게 분류하면 3 종류가 된다.

• 첫째는, 입력문의 해석단계에서 Case Grammar에 의한 格解析을 이용하는 방식이다. 이것은 원문의 분석시 술어와의 관계인 格을 추출하여 입력문의 中間表現을 "格構造" 또는 이와 유사한 "依存構造"를 채택하도록 한다.

• 둘째는, 格을 사용하여 분석하지만 중간표현으로는 "개념구조"를 채택하는 개념구조transfer방식이다. 개념구조는 단어간의 의미관계를 추상화하여 개념간의 관계로서 문장의 의미적인 구조를 표현하는 것이다.

• 셋째는, 格을 사용하고 중간표현을 구문분석tree로서 표현하는 방식이다.

이러한 transfer방식중에서 최근에는 격해석이나 이와 유사한 개념구조방식을 적용하는 시스템이 강해지고 있다. 특히 語順의 자유도가 매우 높은 교착어(한국어, 일본어등)의 기계번역에는 이 방식이 유효한 처리방식인 것으로 판단되는데, 이 방식을 적용하면 구문적인 자유도가 높은 단어간의 의미적인 관계를 파악할 수 있기 때문이다.

(1) 격해석 Transfer 기반의 시스템

이 방식의 번역시스템은 원언어의 分析tree를 격구조로 하고 의존구조에 Case marker 를 부착하여 변환한다. 이 시스템에서 채택하는 구문분석의 문법으로는 의존구조문법과 구구조문법등이 있다. 의존구조문법은 단어간의 구문적인 의존관계를 표현하면서 해석하며, 구구조문법은 인정하는 여러개의 단어를 명사구, 부사구등 phrase단위로 구분하는데, 이들 문법으로 구문해석한 결과를 "依存構造", "句構造" 라고 한다. 의존구조는 그림 3 과 같이 文節의 상관

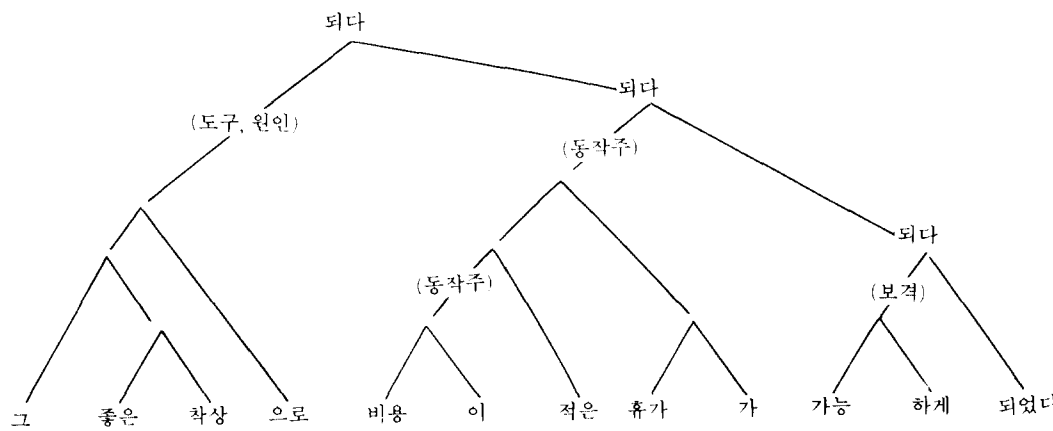


그림 3 의존구조의 해석 tree

관계를 표현하는데 적합하며, 구구조는 그림 4와 같이 문장이 어떠한 phrase들로 구성되어 있는지를 나타내는데 적합하다.

그림 3과 같이 의존구조문법에 의해 생성된 해석tree는 '비용이→적다', '휴가가→가능하다', '착상으로→되다'처럼 각 단어간의 의미관계를 잘 표현하고 있는데, 한국어와 같은 교착어는 어순이 자유롭기 때문에 이러한 의존구조표현이 적합하다. 한편 구구조문법은 문법적으로 어순이 고정되어 있는 언어의 구문해석에 적합하므로 영어를 구문분석하는데 효과적이다. 그림 4와같이 구구조문법을 사용하여 생성된 영어의 해석tree는 인접하는 단어를 NP(명사구), VP(동사구), PP(전치사구)등으로 결합하여 구문관계를 잘 나타내고 있다.

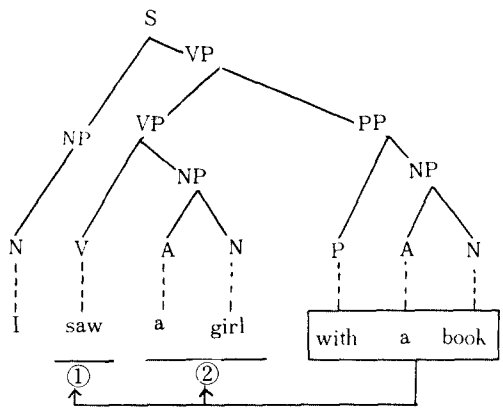


그림 4 구구조의 해석tree

구문분석이 종료되면 대개는 몇개의 해석tree가 얻어진다. 이것은 구문적인 정보만으로는 연결관계가 성립되는 것이 많기때문에 당연한 것인데, 복수의 해석tree로부터 정확한 한개의 해석tree를 결정하기 위하여는 의미적으로 결합가능한지를 판단해야 한다. 따라서 구문분석에 이어 의미해석을 행한다. 그림. 4의 해석tree에서 'with a book'은 구문적으로 ①Saw와 ②a girl에 모두 연결될 수 있으나 의미적으로 파악하면 'with a book'이 'Saw'에 연결되는 해석tree는 제거된다. 일반적으로는 의미해석은 구문해석과 동시에 행해지며 그 결과, 출력되는 해석tree는 1개이다. 그러나 의미적으로 모호한 경우에는 2개이상의 해석tree가 출력될 수도 있다.

이상의 분석과정이 종료되면 마지막 단계로써 해석tree를 格解析(Case analysis)한다. 격해석에 사용되는 格은 행위자(agent), 대상(object), 장소(location)등 格의 분류정도에 따라 세밀히 할 수 있으며 격해석을 하기 위해서는 해석tree를 격구조로 표현한다.

• 격구조 : 격구조는 述語를 중심으로 하여 다른 단어와의 의미관계를 格을 사용하여 나타낸 文level의 의미표현이며, 文의 의미는 格情報와 法情報로 기술한다.

격정보는 文의 의미구조를 구성하는 부분으로써 격요소로 이루어지며, 격구조에서 사용하는

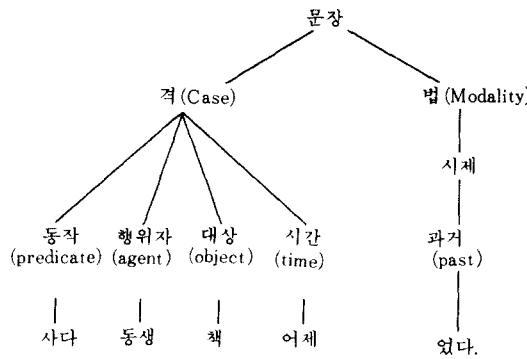


그림 5 격구조의 의미표현

격요소는 표층적인 category로 분류한 것이 아니고 용언과 格사이의 심층적(의미적)인 격관계를 분석한 것이다.

또한, 법정보는 술어가 갖고 있는 양태적인 의미(Modality)를 부여하는 것으로써 시제(tense), 상(aspect), 태(voice), 판단, 태도등의 정보가 포함된다. 격해석을 위한 격구조표현에서는 구문적으로 같은 문구조를 갖고 있더라도 의미적 격관계가 다르면 명확히 구별된다. 그러나 격구조에 의한 의미표현이 갖는 문제점은 격관계의 종류를 분류하는 것이 곤란하여 格시스템의 설정이 어렵다는 것이다.

격문법의 제창자인 Fillmore가 설정한 격시스템은 6종류의 격관계를 갖고 있었지만 정밀한 의미표현을 위해서는 다수의 격관계를 분류해야 한다. 많은 격관계를 사용한 번역시스템에서는 100여개 가까운 Case marker를 설정하고 있다.

〔格System〕:

格関係 - 対象関係群|方法関係群|方向関係群|時空関係群|付帯関係群|形容関係群
対象関係群 - 対象|対象2|比較対象|並列対象|主題|陳述対象|動作主|経験者
方法関係群 - 道具|手段・方法|原料・材料|構成素|原因・理由
方向関係群 - 源泉・起点|着点・方向|目的・目標|結果|与手|受手
時空関係群 - 場所|時
付帯関係群 - 場合|内容|役割|対照・従属|範囲
形容関係群 - 回数|割合|程度|数量|強調|真偽|様態

〔格Frame〕:

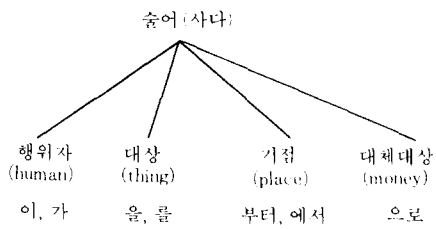


그림 6 격시스템과 격Frame

이와같은 격구조에 의해 작성된 의미 구조표현은 구문적/의미적인 제약을 가하기 위하여 格Frame으로 기술되는데, 이 Frame은 격요소가 실제로 어떤 격관계를 갖고 있는가를 해석할 수 있게 한다. 또한 格Frame은 구문분석에서 산출되는 임의격과 필수격중에서 각 술어에 특징적으로 나타나는 필수격만을 이용하는 것이 대부분이다.

따라서 원언어는 격구조의 변환rule에 의해 목표언어의 중간표현으로 변환되고 구문생성 및 형태소합성을 거쳐 번역문장을 출력한다. 현재, 격해석방식을 채택한 번역시스템으로는 LUTE, ROSETTA, VALANTINE등이 있다.

(2) 개념구조 Transfer 기반의 시스템

이 방식은 본질적으로는 前述한 격구조transfer 방식과 크게 다른 점은 없다. 개념구조transfer에서는 해석tree로부터 "개념"관계를 추출하여 개념구조를 작성하고, 이것을 변환한다. Transfer 부분은 목표언어의 중간표현으로 변환한 후 구문을 생성하여 목표언어의 해석tree

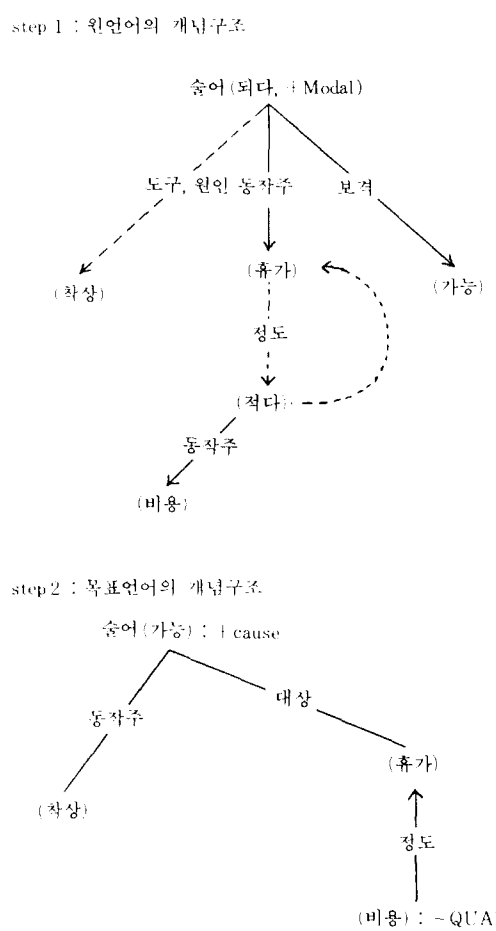


그림 7 개념구조transfer와 구구조변환과정

를 작성하며, 형태소의 합성규칙은 이 해석tree를 문장으로 생성한다. 개념구조는 개념과 개념간의 관계를 격관계등을 사용하여 표시하는데, 개념구조의 각 개념 node는 격구조의 각단어에 대응하는 것으로 볼 수 있다. 구체적인 개념구조transfer의 처리를 韓英번역의 경우로 하여 그림. 7에 나타낸다.

step 1은 '그 착상으로 비용이 적은 휴가가 가능하게 되었다'라는 문장의 한국어 개념구조를 나타낸 것이다. 개념을 network으로 나타낸 이 그림에서는 文中의 단어가 하나의 개념(primitive)으로서 node가 되어 있다. 각 primitive간의 관계는 link로 표시되어 있으며 link 상에는 각node간의 의존관계(또는 격관계)를 기술한다. 원언어(한국어)의 개념구조는 transfer에 의해 step 2의 목표언어(영어)개념구조로 변화된다. 이 구조변환에 의해서 영어다운 표현이 된다. 한국어에서 'A에 의해 B가 가능하게 된다'의 문형이 영어에서는 'A는 B를 가능하게 한다'의 문형으로 변환될 수 있다. 이 변환규칙은 구조변환 규칙에 기술하면 되는데, step 2에서와 같이 영어의 개념구조의 술어는 '가능+CAUSE'로 변화되고 도구격이었던 '착상'이 動作主格으로 바뀐다. 또한 동작주격이었던 '휴가'는 대상격으로 된다. 다음에 시스템은 영어의 개념 구조로부터 英譯文을 생성한다. 이것에는 각 개념node를 영어단어로 치환하는 과정과 영어의 구문을 결정하는 과정이 필요하다. 그 그리고 생성용의 기계사서에서 번역어를 선정하고 step 3의 영어구구조에 따라 譯文 'The idea enabled inexpensive vacation.'을 출력한다.

• 지식base의 적용 : 개념구조를 사용하는 방식의 번역시스템중에서 일부는 번역의 質을 향상시키기 위하여 특정한 언어적 지식으로 작성된 知識base를 사용하기도 한다. 이러한 시스템에서는 개념구조를 중간표현으로 사용하고 개념간의 관계를 해석할 때 현실세계에서 실제로 성립하는 관계를 정의한 지식base를 이용하여 목표언어의 개념구조를 추출한다. 시스템은 입력문을 해석하여 얻은 개념구조를 지식base 와 참조하여, 옳으면 이 처리를 계속하고 그렇지 않으면 새로운 개념구조를 작성한다. 지식base에 기술되지 않은 개념관계는 모두 옳은 것으로 써 처리한다. 비교적 단순한 처리방법으로서는 개념간의 관계를 共起관계로 나타내어 단어간의 결합이 의미적으로 옳은지를 판단할 수 있도

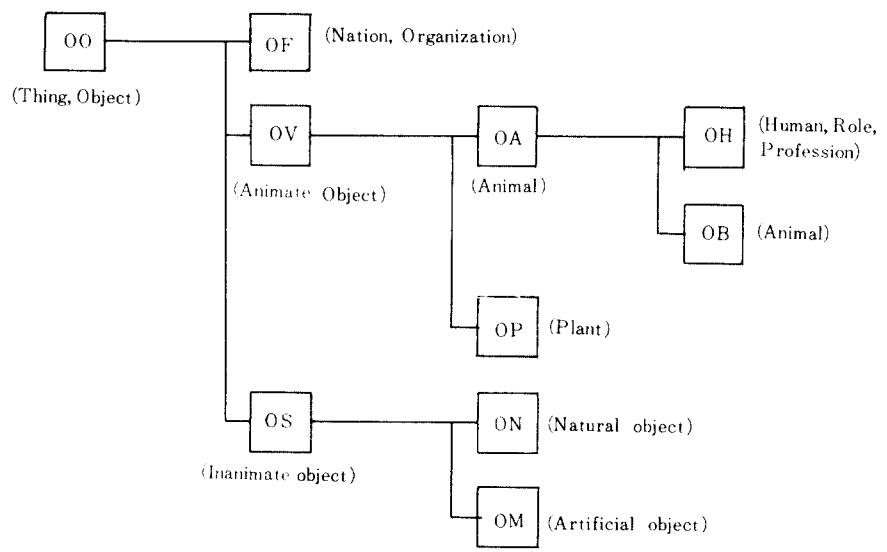


그림 8 의미소의 계층구조

록 하는 것이 있다. 즉 '동물은 먹는다'라고 하는 옳은 지식을 지식base로 구축하면 동물에 속하는 모든 생명체와 '먹는 행위'도 역시 共起관계를 갖고 있어 의미적으로 결합이 가능한가를 쉽게 판단할 수 있다. 이를 지식base로 표시하면 다음과 같이 되는데,

(ANIMAL, EAT+MOVE, (AGENT))
→TRUE

모든 단어에 대하여 共起관계를 이와같이 기술하면 지식base는 엄청나게 커진다. 따라서 지식base의 규모를 최소화하면서도 단어간의 共起관계를 처리할 수 있도록 각 단어의 개념소(primitive) 또는 意味素(Semantic marker)를 계층구조화한다.

또 다른 개념구조에서는 개념간의 관계를 격관계나 상위/하위, 전체/부분, 이유/원인과 같은 관계를 사용하여 기술하기도 한다. 그러나, 실용적인 기계번역시스템을 제작하기 위해서는 이러한 수많은 共起관계를 지식base로 구축하는 것이 요구되며 따라서 시스템제작상 가장 해결이 어려운 과제의 하나로 인식되어 있다.

나. 시스템의 현상

최근까지 실용화되어 있는 시스템은 20여개에 달하지만 이들이 갖고 있는 번역능력은 많은 제약을 받고 있다. 즉 어떤 시스템도 적용분야를 한정하고 있으며 전편집(pre-editing)과 후편집(post-editing)을 가하여 이용되고 있다. 따라서 엄밀히 말하면 自動翻譯이라기 보다는 번역지원시스템으로서의 역할을 수행하고있다고 볼수 있다. 이하에서는 현재의 번역시스템이 더 높은 실용성을 갖기 위하여 고려해야 될 번역의 特質과 현재의 시스템이 갖고 있는 번역기술의 한계에 대하여 요약한다.

(1) 기계번역의 특질

현재 제품화되었거나 개발중인 기계번역시스템의 성능을 객관적으로 평가하는 것은 어렵지만 통상의 hardware/software 기술과 그 성능에 비교하면 상대적인 비교가 가능하다. 예를 들면 번역의 外的인 기술로서 사용되고 있는 wordprocessor의 기술과 성능은 국내에서도 불과 5년만에 사용자의 욕구를 충족시키는 갖가지 기능이 보강되어 제품으로서 확고한 위치를 차지하고 있다. 그러나 기계번역의 기술은 근본적으로는 인간의 지능에 깊게 관여되어 있는 고도의 지능 소프트웨어기술이라는 점에서 현행의 기술로서는 상당한 한계가 있으며 번역품질의 평가에서도 그 기준이 분명치 않다. 예를 들면,

① native speaker가 보는 완전한 번역

② 부자연스러운 부분이 있지만 이해가 쉬운 번역

③ 이해가 곤란한 부분이 다소 있지만 이해할 수 있는 번역

④ 이해는 곤란하지만 전체적인 내용은 파악할 수 있는 번역

⑤ 전혀 이해가 불가능한 번역

와 같은 주관적인 평가가 가능할 뿐이다. 또한 원언어가 어떤 분야를 대상으로 한 것이며 text의 종류도 기술자료인가 소설인가 등에 따라서도 번역의 어려움이 크게 달라진다. 소설과 컴퓨터기술자료의 번역결과를 같은 기준으로 비교하면 의미가 없다. 결국, 시스템의 평가는 입력문장의 난해도와 번역문의 품질과의 상대적인 평가이다. 따라서 일반적인 관찰결과로서 현재의 번역기술에 대하여 다음과 같은 것을 지적할 수 있다.

• 한일/영한과 같이 어족이 다른 언어간의 번역에 있어서는 원언어에서의 단어, 句, 節이 목표언어에서도 각각 대응하지는 않는다. Indo-European어족간의 번역에서는 국소적인 명사구, 동사구등의 구조를 처리하는 것으로도 대응되는 번역문을 생성할 수 있지만 한국어및 영어간의 경우에는 대응처리만으로는 자연스러운 번역문을 생성할 수 없다. 적어도 單位文內의 문구조를 分析하여 새로운 문구조의 변환과 생성을 해야 한다. 또한 한국어와 영어가 각각 갖고 있는 독특한 tense, Aspect, Modal 표현의 처리 등, 언어고유의 문제도 있다. 이러한 언어적인 여건

때문에 인간이 개입하지 않는 실용적인 번역은 아직도 많은 문제점을 간직하고 있다.

• 번역은 분야의 의존도가 매우 높은 작업이다. 따라서, 현재의 시스템은 인간 번역가보다도 훨씬 분야의 의존성이 강하다. 이러한 번역의 의존성은 다음과 같은 몇가지 유형으로 나타난다.

① 적용분야의 지식 : 번역하고자 하는 분야에 관계되는 최소한의 지식과 전문적인 용어등이 필요하다. 이러한 지식은 대개 전문용어사전에 등록하여 제공한다.

② 분야에 독특한 文体와 表現 : 인간이 계약서나 법률문 또는 신문기사, 논문을 작성할 경우에는 표현하려는 내용을 파악하고 있으면서도 그 분야의 문장작성법이 갖는 독특한 표현이나 문체가 있기 때문에 문장의 작성에 많은 시간을 소비한다. 이것은 기계번역의 관점에서도 원언어의 문체나 표현이 번역문에 반영되어야 하는 것을 의미한다. 따라서 언어사용에 관한 지식은 번역규칙이나 기계사전의 형식으로 반영시켜야 한다.

③ 텍스트의 작성의도 : 원언어의 용도가 会話文이면 전후의 문맥에 따라 지시어나 생략표현 등이 많이 발생하는데 이러한 회화문을 기계번역하는 경우에는 문맥처리를 해야 한다. 또한, 논문의 초록문인 경우에는 문장을 축약시키기 위하여 명사의 연속적인 사용에 의한 복합명사, 각종의 並列句표현이 많이 사용되는데 이러한 각종의 언어현상을 취급하지 않으면 옳은 기계번역을 할 수 없다.

이와같은 시스템의 분야의존성은 기계번역시스템의 고유한 문제는 아니지만 기계번역의 質에 큰 영향을 미치는 요소이다. 따라서 분야의 존성의 지식을 번역규칙과 기계사전(또는 지식 base)에 반영하여 번역의 품질을 향상시킬 수 있어야 한다.

그러나, 知能的인 번역기술이 정착되지 못한 현재로서는 기계번역의 実用性을 향상시키기 위하여 특히 번역의 적용대상영역을 한정시키는 방법이 가장 유효하다. 기계번역의 실용성과 적용영역의 폭간의 상관관계를 도식하면 그림 9 와 같다.

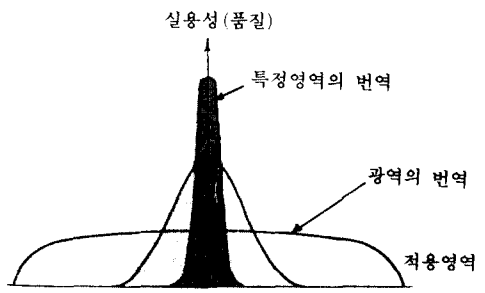
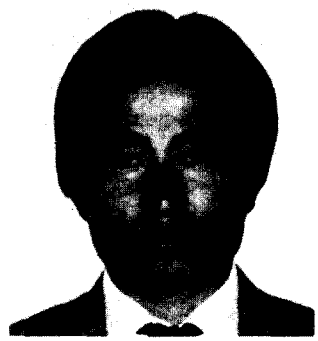


그림 9 적용영역과 번역품질의 상관관계

다음호에 계속(편집자註)



朴 昌 浩

저자약력

- 1953년 7월 22일
- 1972. 3～1980. 2 : 인하대학교 공과대학 전자 공학과
- 1980. 3～1 81. 2 : 인하대학교 공과대학원 전자 공학과
- 1987. ～8 현재 : 프랑스 Grenoble대학교 정보공학 대학원(ENSIMAG)
- 1980. 6～현재 : 한국과학기술원 시스템공학 센터 선임연구원
- 1987. 3～1987. 4 : 프랑스 국립과학기술연구센터 GETA연구소 파견 연구원

李 基 式

저자약력

- 1945년 3월 20일생
- 1980. 3. 3～1983. 2. 21 : 연세대학교 공과대학전자공학과(공학박사)
- 1975. 4. 1～1977. 3. 26 : 일본 농경공업대학교대학원 정보과학사(석사)
- 1965. 3. 13～1971. 3. 31 : 한양대학교 공과대학전기공학과(학사)
- 1987. 1. 1～현재 : 한국전산원 연구위원
- 1971. 8. 1～1986. 12. 31 : 한국과학기술원 시스템공학센터 책임연구원

♧용 어 해 설♧

● **부식 방지 교환기**(anti corrosion switchboard) : 부식성 가스가 발생하는 온천 지역 등에 설치하기 위해 사용 부품에 부식 방지 처리를 한 교환기.

● **보우스터**(booster) : 회로와 직렬로 접속하여 그 전압을 정·부의 임의의 방향으로 가감함으로써 회로의 전압을 일정하게 유지시키는 승압(昇壓) 장치.

● **보우스터 국**(booster station) : TV 방송에 있어서 주국의 전파를 수신하여 동일 주파수로 자동적으로 재방송하는 보조 중계국을 말한다. 부우스터 국은 주국과 동일 주파수로 방송하기 때문에 주국 전파와의 혼신이 생기므로 주국과 편파면을 달리하여 수직편파로 재방송한다.