중간언어에 기반한 기계 번역시스템의 설계

김상국, 박창호

시스템공학연구소 인공지능 연구부

Design of a Multilingual Translation System Based on Interlingual Approach

Sang-Kuk KIM, Chang-Ho PARK

Dept. of Artificial Intelligence

Systems Engineering Research Institute

요 약

다언어간 번역을 지향하는 기계번역시스템의 개발을 위해서는, 의미 이해기반의 해석기술과 언어에 독립적인 생성기술의 설계가 기본이므로 원시언어와 목표언어가 어느 한쪽의 언어지식에 의존하지 않고 언어형식화가 가능한 중간언어 구조를 설정하는 것이 중요하다. 따라서, 한국어를 중심으로 하는 다언어 번역의 설계에서는 비교적 문구조의 정형화가 이루어진 영어와는 달리 어순 배열의 자유도가 높고 조사의 격표시로 문장구조가 결정되는 한국어의 특성을 고려한 해석 및 생성 메카니즘이 필요하다. 본 논문에서는 문장에 내포된 심층의미의 중간 표현으로써, 단어의 의미를 개념화 시킨 개념소(Conceptual Primitive)간의 의미적 결합관계를 나타내는 개념 그래프(Conceptual Graph)를 채택하고 설계한 다언어 번역지향의 중간언어기반 번역시스템에 대하여 기술한다.

Ⅰ. 서 론

기계번역의 대표적 방식은 실용화된 대다수의 상용번역 시스템이 채택한 변환방식과 1980년대 후반부터 연구개발이 활성화되기 시작한 다언어번역 및 의미번역을 지향하고 있는 중간언어(Interlingua)방식으로 대별할 수 있다. 그러나, 구문해석에서 얻어지는 해석목의 변환에 비중을 두고 부분적인 의미 해석을 부가하는 기존의 변환방식 기계번역은 유사한 언어

권의 번역 지원시스템에서 대체적으로 실효를 거두었으나, 언어구조가 매우 상이한 어족간의 고품위 번역이라든지 다수의 언어쌍을 개별언어의 형식화된 언어지식을 공유함으로써 번역하고자 하는 다언어 번역에서는 이러한 변환 방식으로는 기술적 장벽 및 개발부담을 감소시키는 것이 한계에 도달했다고 볼 수 있다. 따라서 최근에는 기존의 변환방식의 번역기술에서 축적한 언어해석기술을 기반으로 하여 원시언어(Source Language : SL) 및 목표언어(Target Language :TL)에 의존하지 않는 중간언어의 개발과 이에 기반한 번역시스템의 개발이 미국(KBMT, JANUS 프로젝트), 일본(CICC, ATR, EDR 프로젝트), 유럽공동체(Eurolangue, Verbmobile 프로젝트)등에서 매우 활발히 진행되고 있으며 Corpus base, Example base 및 통계적 접근방법에 의한 번역기술이 추가되고 있다.

본 연구에서는 이와 같은 기계번역기술의 해외 발전 추세에 부응하고 선진 번역기술의 확보를 목적으로, 한국어와 영어간의 번역을 우선 대상으로 하여 Conceptual Graph를 중간언어로 하는 번역시스템의 설계를 목표로 하였으며, 이하에서는 그 성과를 소개한다.

II. 중간언어와 기계번역

1.중간언어(Interlingua)의 분류

기계번역기술의 질적 향상과 다언어간 번역을 실현하기 위한 중간언어는 SL의 해석결과 및 설계수단으로써 채택하고 있으며 모든 언어에 중립적인 Universal Language의 역할을 갖는다.

1) 해석결과로서의 중간언어 : 어떤 특정분야의 제한된 Task를 기계번역(예:예약 및 안내를 위한 대화 번역)하는 경우에는 그 분야에서 일상적으로 사용되는 한정된 정보가 있으며 해당언어의 특징과 거의 독립적으로 정의할 수 있다. 즉, 기계번역에 사용할 Task정보를 SL에서 추출함으로써 중간언어를 대신할 수 있다. 이러한 번역에서는 특정 Task를 설정하고 여기에 필요한 정보를 해당언어에 종속하지 않는 표현형식(일종의 중간언어 개념을 가짐)으로 나타낸다. 이 부류에 속하는 중간언어는, 카네기 멜론대학에서 KBMT에 채택한 Ontological KB구축에 사용하고있는 중간언어표현과 같이 적용영역에 매우 의존적인 반면 해당언어에서는 독립적인 영역 지식 의미표현이다.이러한 중간언어는 특정한 주제의 대화번역, 예를 들면 미국CMU의 JANUS 및 일본 ATR의 ASURA같은 영어-일어간 번역에서 고품질 번역의 가능성을 실증하였다.

2) 설계수단으로서의 중간언어 : 번역대상으로 하는 SL 및 TL의 속성에 따라 분류하며 적용영역에 영향을 받지않는 중간언어를 설정한다. 이러한 중간언어는 변환방식에서 n개 언어쌍을 상호 번역할 경우 필요한 n(n-1)개 언어 pair에 대한 언어모델 개발부담을 절대적으로 감소시킬 수 있다. 중간언어에 의한 n개 언어쌍의 다언어 번역기를 설계할 경우는 하나의

특정언어를 중심언어로 하고 n-1개 TL의 각 단어 의미를 중심언어의 개념표현에 대응하도록 정의할 수 있다. 이때, 중심언어 한국어의 단어는 최소의 개념소로 세분화하되, 세분화의 수준은 TL의 각 단어 의미가 상세화되는 수준으로 함으로써 n개 언어쌍 전체에 대한 중심언어의 특성을 갖게된다.

3) 중립중간언어(Neutral Interlingua) : 어떤 문장이 포함하고 있는 심층의미는 유일한 것이지만 그 표층구조는 다양하다. 따라서 SL과 TL이 갖는 다수의 표층구조가 상이해도 심층의미가 동일하면 SL과 TL의 의미적 결합관계를 특정언어의 언어적 속성에 의존하지 않는 중립언어로 (Neutral Language)로 설정, 표현할 수 있다. 이러한 중립언어 표현을 일반화할 수 있다면 가장 이상적인 중간언어로서의 역할을 갖는다. 일본의 EDR 및 CICC에서 이러한 중간언어의 개발을 진행하고 있다.

2. 중간언어기반 기계번역

변환방식의 기계번역에서는 목표언어의 생성과정에서 SL의 문장구조가 TL에 반영되는 구문구조적 제약을 충분히 제거할 수가 없다. 따라서 언어의 개별적 특성에 의존하지 않는 번역을 위해서는 문장의 해석 결과를 SL의 문장구조에 독립적인 형식으로 표현할 수 있어야 한다. 즉, 중간언어 기반의 번역은 문장의 구조적 특징에 의존하지 않는 개념모델을 중간매체로 설정하는 것이 필수적이다.

또한 다언어 번역의 관점에서는, 문구조나 어법등 다양한 언어현상을 논리적, 수리적으로 계산한 결과를 TL로 전환하는 변환방식은 대상언어간의 언어적 특징에 따라 변환 규칙과 이에 요구되는 지식 표현의 수준이 달라지게 되며 특정한 언어 Pair에 국한된 수준의 해석→변환→생성처리가 요구된다. 이 경우는 시스템 개발자가 SL과 TL에 관한 지식을 함께 갖는것을 전제로 한다. 그러나 하나의 SL로 부터 n개 TL을 다언어 번역하는 시스템을 설계할 경우는, SL의 해석 결과가 n개 언어쌍의 변환을 공유할 수 있는 수준의 언어이해표현이 되어야 한다. 따라서, 다언어 번역으로의 확장을 고려하는 시스템의 설계는 대상언어에 상호 독립적으로 해석 모듈과 생성모듈을 작성할 수 있는 중간언어방식의 기계번역이 효과적이다.

3. 중간언어와 개념Graph

기계번역의 궁극적 기술수준은 SL의 언어이해구조와 같은 TL을 생성하는데 있다. SL의 문장이해는 단어의 의미는 물론 문맥해석을 전제로 하고 있으며 인간과 동등한 수준의 언어해석 능력을 부여했을 때 가능하다. 따라서 언어적 지식의 총체인 해석 규칙과 기계사전에는 각 단어가 포함한 의미를 형상화 개념화하고 각 개념이 문법적 또는 의미적으로 어떠한 특성과

결합관계를 갖는지를 나타내야 한다. 그러나 정확한 의미해석을 위해서는 기계사전에 포함되어 있는 의미소성 및 개념소만으로는 부족하여 의미 개념간의 결합관계를 논리적으로 표현해야한다. 즉 개념소간의 의미적, 논리적 관계를 나타낼 수 있는 개념 Graph의 채택이 요구된다.

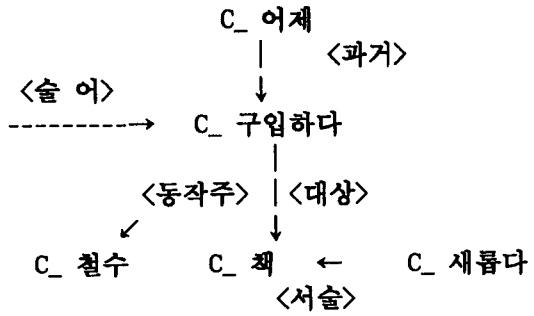
중간 언어 방식에서 SL의 해석은 기계사전의 정보와 해석 rule, 의미관계를 제한 또는 규정하는 의미해석 모델을 사용함으로서 실행되며, SL의 해석 결과는 중간언어로 채택한 개념 Graph를 생성한다. TL의 생성은 이 개념 Graph를 이용하여 TL의 기계사전과 생성규칙, 그리고 생성단어 및 어절간의 공기관계와 연접관계를 규정한 생성 Model에 의해 수행된다. 그러나 한국어와 영어의 기계번역에서와 같이 문구조 표현 및 어법의 차이에 따라 빈번히 발생할 수 있는 개념표현의 gap을 제거하기 위해서는 개념 Graph의 변형 rule을 적용하게 된다.

1)개념 Graph

문장을 구성하고 있는 자립성분인 단어가 어떤 개념을 갖고 있으며, 각 개념이 어떤 상관관계를 맺고 있는가를 나타내는 개념 Graph는 단어 또는 어절의 개념표현, 즉 개념소를 node로 하고 인접 node간의 제약조건 또는 의미관계를 arc로 갖는다. 따라서 본 시스템의 설계에서 중간언어로 채택한 개념 Graph는 Semantic Network의 node와 arc상에 개념소와 관계인자(relation)를 대칭시킨 2차원 구조를 갖는다.

[예문] 철수·는 어제 새로·운 책·을 구입하·였·다

〈술 어〉 → (구입하다) ∵ C_ 구입하다
(구입하다)라는 동작의 〈동작주〉 → (철 수) ∵ C_ 철수
(구입하다)라는 동작의 〈대 상〉 → (책) ∵ C_ 책
(책)이라는 대상의 〈서 술〉 → (새롭다) ∵ C_ 새롭다
(구입하다)라는 동작은 〈과 거〉 → (어 제) ∵ C_ 어제



[그림 1] 개념 Graph

이 개념 Graph의 node는 문장성분인 각 단어의 개념을 나타내며, arc는 행위자인 <동작주(agent)>, 행위의 대상인 <대상(object)>등을 나타내는 필수격 관계와 행위의 시점인 <과거(past)>과 같은 자유격 관계를 표시한다.

2) 개념 Graph의 변환 규칙

SL이 해석되면 이 문장이 가질 수 있는 의미구조는 개념화되어 이항관계로 작성된다

[SL]: 학생이 오토바이를 타고 달린다

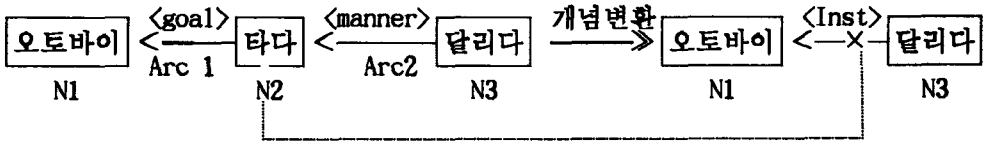
↓

[개념표현]:(C_ 학생, C_ 달린다,<동작주>), (C_ 탈것, C_ 달린다,<수단,도구>)

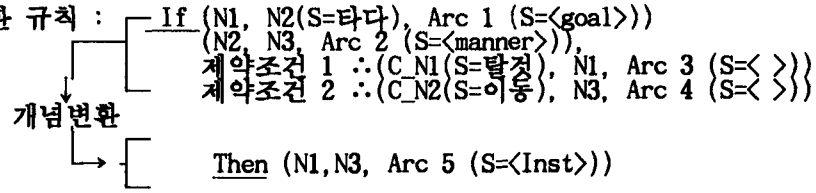
위의 예문에서와 같이 SL의 개념표현이 의미해석 모델(또는 세계 Model)에 포함되어 있으면 해석결과는 적납되지만 그렇지 못하면 다른 해석결과를 도출키 위해 개념변환 규칙에 의한 의미해석 모델을 추가 하거나 수정한다.

• 개념구조의 변환 규칙

① 개념표현



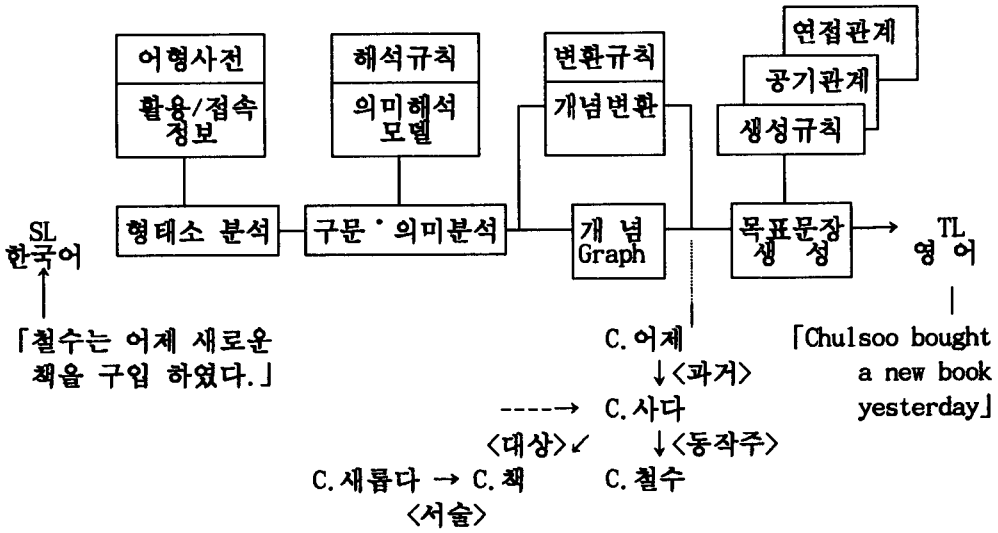
② 변환 규칙 :



중간언어 표현에서 개념간의 상관관계는 가장 일반적인 것으로 결정해야 하지만 각 언어는 여러 개의 독특한 개념을 갖고 있기 때문에 항상 1:1 대응관계가 성립하지 않을 수 있다. 따라서 중간언어의 어휘에는 각 언어에 유일한 개념도 각 언어에 공통적인 개념과 동등하게 취급되어 수록되어 있다. 이런 종류의 독특한 개념은 다른 언어의 개념구조와는 상이한 개념을 갖고 있을 수 있으며 이런 경우는 node상의 인접 개념소, 또는 개념 Graph와 개념소 사이의 대응관계를 적용하여 TL에 적합한 형태의 개념 Graph로 변환시킬 수 있다.

Ⅲ. 중간언어기반의 기계번역기 설계

개념 Graph를 중간언어로 하는 다언어번역 지향의 본 시스템은 해석규칙과 기계사전,의미해석 모델을 사용하여 SL을 해석하는 해석시스템과, 해석결과로 생성된 중간언어인 개념 Graph로 부터 생성규칙과 생성사전을 사용하여 TL을 생성하는 생성시스템으로 구성된다.



[그림 2]중간언어 '개념 Graph'기반의 번역 Module

1.해석 System

SL문장을 node List로 분리하는 형태소 분석기와 문구조적 의미해석을 수행하는 구문의미해석기의 2개 Subsystem으로 구성된다.

1)형태소 분석기 특징

① 좌우품사의 접속정보에 의한 종형탐색법 : 형태소 해석사전의 좌우접속정보와 인접어형간의 접속가능여부를 검사하기 위한 접속정보 matrix를 사용

② 최적 형태소 탐색을 위한 Backtracking : 형태소 해석결과 분리된 어형이 후속 형태소와 접속되지 못할 경우 backtracking 하여 차순위 후보 어형을 선택, 접속여부를 계속 검증.

③ 최적 형태소 추출 : 어형분석에서 도출된 형태소 List가 n개 경우, 최적 형태소 결정을 위하여 어형길이 및 각 어형의 사용빈도(통계정보)를 계산,

최대평가치를 갖는 형태소 해석 List로 결정

2) 구문/의미해석기

형태소 해석의 결과인 node list의 구문정보를 계산, 성분간의 관계를 분석하여 문법적 역할을 결정하고 의미해석을 행한다. 본 해석 module은 CFG의 해석 문법규칙에 의하여 수행되며 특히 SL(한국어)의 구문적 특성을 해석하기 위하여 격문법의 해석기법을 채택하였다. 격문법하에서 문장성분간의 상관관계를 나타낼 때는 Predicate 중심이므로 해석결과는 중심술어에 대한 Agent, Object, Goal, Cause등으로 출력되며 SL문장이 2개이상의 술어를 갖는 복수문절일 경우는 각 술어가 다른 중심술어와 어떤 관계인지를 해석한다. 이는 동사가 갖는 고유의 Case Pattern에 의해서 필수격 지배 형태를 검색하여 구문 및 의미적 결합 관계를 결정할 수 있다.

① 해석 Mechanism : 해석의 기본 프레임은 인접 노드 A,B가 갖는 각각의 문법 정보와 개념정보를 이용, Pattern matching 규칙에 따라 하나의 결합규칙으로 Reduction하는 기능에 의한다. 이 과정은 해석규칙의 L→R parsing을 계속함으로써 해석의 종료시에는 중심개념을 갖는 주술어 (Main predicate)기반의 개념 Graph형태로 해석결과를 도출한다.

<해석규칙 R1> : = < Condition 1 > < 문법속성 A (g1, g2, …… gn)>
< 문법속성 B (g1, g2, …… gn)>
→ <문법속성 C ()> < 처리 Type > < arc ($\overline{A}$, $\overline{B}$)>
< Action > < 적용 우선 순위>

A, B : node A,B의 문법속성
$\overline{A}$, $\overline{B}$: node A,B의 개념정보

② 해석 flow

[1] 해석 Window가 입력 String의 최초 node A 및 차순 node B에 위치, 양 node와 상태 Stack를 참조하고 적용규칙을 검색.

[2] 규칙 제어부는 <Condition 1> field의 문법속성이 상태 Stack에 놓여 있는지를 확인, 존재하는 경우 문법속성 A,B영역의 모든 속성이 해석 Window 좌우에 있는지 검색

[3] 해석규칙 R1이 매칭조건과 일치하면 문법속성 C의 형태로 Rewrite하고 해당규칙이 2개이상이면 < Priority>에 의해 순차적용

[4] 해석규칙이 매칭되면 새로운 node C는 부분 tree를 생성, 하나의 head node로 되고, 좌우 Window의 2개 node는 root node로 치환

[5] 해석 Window는 1 stack 우로 이동, 해석 tree를 얻을 때까지 해석규칙의 매칭을 계속

2. 개념 Graph 생성과 변형

1) 개념 Graph 생성 : 해석규칙의 적용으로 구문 Tree가 생성되면 개념소가 부가된 개념 tree가 규칙에 따라 도출된다. 최종 node까지 해석이 수행되면 각 node에는 개념소가, 이웃 node들간의 arc에는 개념간의 관계자(relation)가 결정되어 TL의 생성에 중립적인 중간언어로서의 개념 Graph가 작성된다. 이때 개념 Graph가 일반화된 의미해석모델에 포함되는지를 검증하려면 적용해석규칙의 적합성을 평가하면 된다.

2) 개념 Graph 변형 : SL의 해석결과로 얻어진 개념 Graph가 TL의 개념구조와 다른 때에는 SL과 TL간 개념구조의 gap을 제거하기 위하여 개념 Graph를 변형할 수 있는 변형 Rule을 적용한다. 예를 들면, 한국어 및 영어간 번역의 경우에 개념상의 차이를 갖는 BE(-이다.)형 언어구조인 한국어를 DO형 영어구조로 개념구조를 변형할 필요가 있는데, 이런 경우에는 개념구조 변형 Rule을 사용함으로써 SL과 TL간의 언어적 차이를 제거할 수 있다.

* 개념 변환 규칙의 Syntax

① < 변형문법명 > : = if 부분 NET 1 [제약조건] then ┌부분 NET 2
                                                    └(Rule)
                                              else ┌부분 NET 2
                                                    └(Rule)

② < 부분 NET 1> : = (n1,n2,arc) ∨ net (n)
n = ┌ name(ss=개념기호), sup=상위개념, SG=SL속성, TG=TL속성
    └ NIL ; Nil node를 표시

③ < 부분 NET 2> : = (n1,n2,arc) ∨ net (n')
n = ┌ name(ss=개념기호)
    └ NIL ; Nil node를 표시

변형규칙은 부분 NET1에 대응하는 부분 CG가 제약조건을 만족시키면 부분 NET 2로 변환된다.

3. 생성 System

SL의 해석결과인 개념 Graph는 TL의 생성 시스템의 생성규칙과 기계사전에 결합된다. TL의 생성시스템은 기본적으로 2차원적 개념 Graph를 1차원의 문자 String으로 변환시키는 규칙, 공기관계, 연접관계 사전으로 구성되어 있으며 개념구조를 node의 순서대로 탐색, 생성규칙에 따라 TL의 형태소를 생성한다. node의 탐색 순서는 생성규칙의 통제를 받도록 설계하였으며 TL의 표층문장 생성은 단어간 공기사전 및 연접관계사전을 참조하면서 수행된다.
생성시스템은 구문구조와 형태소 생성을 동시에 수행할 수 있도록 설계되며 언어에 독립적인 생성기능을 갖는다.

1) 생성 mechanism

생성시스템의 실행모듈은 생성 Window, 변형규칙 interpreter(RI),출력 list로 구성되어 있다. 생성 Window는 node와 arc를 참조할 수 있도록 이동하며 출력list는 각 형태소를 TL의 생성순서대로 격납시킨다.

① 생성시스템에 입력되는 개념 Graph의 node는 node명/basket/단어 list로 구성되어 있으며,basket으로서 자신의 node 또는 타 node로 부터의 생성정보를 받는다. 단어 list는 각 node의 개념을 갖는 단어들을 수록한다.

② Arc는 단어 list를 가지며, Arc명은 node간의 관계를 수록한다.

③ node명과 arc명은 TL의 단어사전을 검색할 때 key가 되며, 단어사전의 표제어에 생성기호를 수록하게 함으로써 생성규칙 group의 key로 사용한다.

④ RI는 생성규칙을 해석하고 생성 Window를 각 node에 이동시켜 node 및 Arc의 단어를 공기관계 및 접속관계사전을 참조함으로써 TL을 생성시킨다.

- 공기관계사전 : 2개 단어가 어떤 관계를 갖고서 문장에 사용되는지를 나타내는 정보를 수록하고 있다. 그러나. 개념소는 일반적으로 여러개 단어로 표현되므로 공기관계를 계산하여 적합한 역어를 선택한다.
- 접속관계사전 : TL의 문장 생성시에 단어가 서로 연접될 수 있는지를 나타낸 것으로서, 유연한 TL의 표층구조 생성에 적합한 형태소를 선택하는데 사용.

2) 생성규칙

TL의 생성규칙은 순차적 구조를 가지며 TL생성시의 적용순서 및 출력문장의 어순을 결정한다.

생성규칙 Syntax : If < condition 1 >
then < Arc - name > < Action > < Message >
else goto Next - Rule

< Arc - Name>은 생성규칙이 적용되는 Arc명을 나타내며, < Action >은 처리 Type를 표시하고, 다음 4종류의 기본 Type에 의해 생성을 진행한다.

① node 생성규칙 : 노드상의 개념소에 상응하는 단어생성
② Out - Arc 생성규칙 : 시점 Arc의 하위 net에 대한 어절을 생성
③ In - Arc 생성규칙 : In - Arc의 하위 net에 대한 어절을 생성
④ 단어 생성규칙 : 단일 개념소일 때 TL단어를 생성

참 고 문 헌

1) Brachman R., Schmolze : An overview of the KL-one Knowledge Representation Language. Cognitive Science 9, pp 171-216, 1985

2) Hiroshi UCHIDA. Meiying ZHU : An interlingua for Multilingual machine Translation. 89-NL-72-9, information Processing Society of Japan. 1989.

3) Muraki K : PIVOT, Two-phase Machine Translation System, MT summit, pp 113-115, 1989

4) Hiroshi UCHIDA : ATLASII : A Machine Translation System Using Conceptual Structure as an Interlingua. Machine Translation Summit. pp. 85-92 1989.

5) Center for the International Cooperation for computerization(CICC) : Interlingua, final edition, 1993

6) Japan Electronic Dictionary research Institute. Ltd. : Concept Dictionary, TR-27. 1990.

7) Japan Electronic Dictionary research Institute. Ltd. : EDR Electronic Dictionary, Technical Guide, Tr-042, 1993

8) Kaji H : HICATS/JE, A japanese-to-English Machine Translation System Based on Semantics, Proc. of MT Summit, pp 101-106, 1987

9) Sergei NIRENBURG, Lori LEVIN. : Knowledge Representation Support, MACHINE TRANSLATION. vol. 4. NO. 1 1989.

10) Sergei NIRENBURG. : Knowledge-Based Machine Translation. MACHINE TRANSLATION. vol.4. NO.1 1989.

11) Fass D : Collative - Semantics, A semantics for Natural Language processing, Technical Report, NEW MEXICO State UNIV., 1988

12) Goldman N : Conceptual Generation, Conceptual Information Processing, pp289-372, 1975

13) A. Tucker, S.Nirenburg : The Structure of Interlingua in TRANSLATOR, Machine Translation, proc of TMI, pp 90-113, 1987