실시간 영/한 번역을 위한 트랜스퍼 어휘 사전

이 대진

A Design of Transfer Lexicon Dictionary For English to Korean Translation

Daijin Lee

Department of Electronics Science, Dongguk University

요약

본 연구는 자연어의 실시간 기계 번역의 설계에 관한 것으로 반도체 메모리에 트랜스퍼 방식의 어휘 사전을 구성하고 어휘 정보 문법을 기본으로한다.

1 .서론

지역과 민족에 따른 언어의 차이를 해결하고 문화의 교류와 통합을 위하여 인간이 사용하는 자연언어(natural language:human language)의 상호번역이 필요하다. 번역의 방법에는 전문인력에의한 것과 기계적 전문가 시스템이 있다. 전문가에 의한 언어의 상호 번역은 낮은 효율성과 전문가가아닌 일반 대중이 보다 직접적이고 능동적으로 대처할 수 없다는 점에서 사람이 아닌 컴퓨터에 의한 자연언어의 번역은 유연성(flexibility)의 부족, 처리 속도의 저하, 그리고 오역 전달의 높은 가능성이라는 측

면에서 두가지 모두 충분하지 못하였다.

위와 같은 연유로 인하여 숙달된 전문가의 높은 번역 품질과 기계적 TOOL의 고속 · 고정도 처리를 함께하는 진보된 실시간의 자연어 번역 · 통역 시스템이 필요하다.

2 .TRANSFER 번역 방식

자연언어의 번역 방식에는 「다이렉트(direct) 방식」, 「트랜스퍼(transfer) 방식」, 「피보트(pivot) 방식」의 세가지가 있다.

「다이렉트방식」은 중간언어를 가지지 않고 원언어의 문을 직접 목표언어의 문장으로 변환하는 것이고, 「트랜스퍼」는 번역의 대상이 되는 언어마다 중간표현을 설치하여 그 중간표현 사이에서 변환을 하는 방식이며, 모든 언어에 공통인 보편적 의미표현을 설치하여 변환을 전혀 하지않는 것이 「피보트방식」이다.

그림 2-1은 번역 방식간의 관계를 도시한 것이다.

다이렉트방식은 해석의 깊이에 따라서 문장다이렉트방식, 구문다이렉트방식, 의미다이렉트방식의 3가지로 나눌 수 있으며 번역품질이 낮고 문장의자유도가 높아지면 동작하지 않는 결점이 있어 저급기(low-level machine)에 한하여 사용되고 있다.

트랜스퍼방식은 구문트랜스퍼방식과 의미트랜스퍼방식의 두가지로 나눌수 있는데, 구문트랜스퍼방식은 중간표현으로 나무구조(tree structure)화된 구문구조로 나타낸 표현형식을 갖는다. 이 중간표현 가운데 단어에 해당하는 부분을 대역사전(영어 → 한글 번역의 경우 영한사전)을 사용하여 목표언어의 단어로 치환하고 원언어의 문장구조를 그것에 대응한 목표언어의 문장으로 생성하는 것이다.

의미트랜스퍼방식은 중간표현에 의미 의존적인 의미 네트 워크(semantic network) 등의 의미표현형식을 채용하여 원언어문장의 해석결과가 이러한 의미적 개념 표현으로서 표시되도록 하는 번역 방식이다.

의미트랜스퍼방식을 극한까지 진행한 것이 피보트 방식이며, 중간표현이 모든 언어

에 의존하지 않는 단지 보편적인 개념으로 나타내어진다. 피보트방식은 다언어 대응의 번역에 최적이다. 그러나 현실적으로 모든 언어의 보편적인 개념집합(conceptual set)의 존재 가능성과 그 실현방법에 대한 의문점은 향후의 연구 과제이며 현재에는 의미 의존적 트랜스퍼방식의 실용화에도 어려운 점이 많다. 여기서는 의미 의존적, 구문 중심적 트랜스퍼방식을 택하였다.

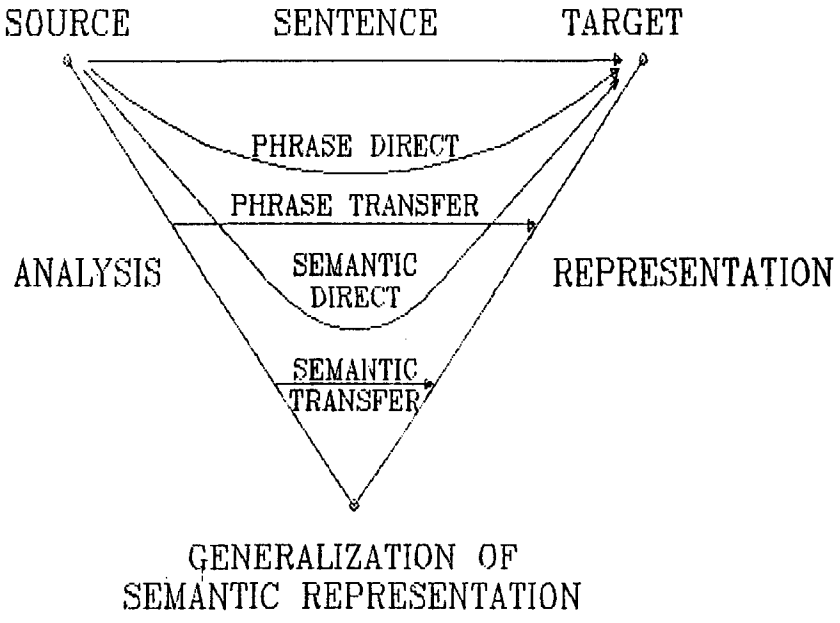

그림 2-1. 세가지 번역 방식간의 관계 .

Fig. 2-1. Relation of Three Translation Method .

3 .어휘정보 문법 (Lexicon-driven Information based Grammar : LIG)

원천언어를 대상언어로 변환하기 위해서는 양쪽의 문법적인 사항이 중요시되는데 LIG은 어휘부가 주축이 되는 정보기반의 문법으로 세밀한 어휘구조의 분석과 이에 의한 자질(feature)의 효과적인 나열에 의한 병합과 제범주들의 도출, 그리고 그 범주들

간의 방정식 풀이에 의하여 의미 · 구문 구조를 나타내는데 트랜스퍼 방식과 그 활용의 맥을 같이한다.

트랜스퍼방식에 사용된 LIG의 특성은 다음과 같다.

(1) 사전의 어휘가 자질구조 (feature structure)로 표시되어 정보내용의 처리에 병합 단조성(harmonic monotonicity)이 유지되어 계산처리가 용이하다.

(2) 원천언어의 분석과정과 목표언어로의 생성과정에 준동형성(homomorphism)이 있다.

(3) 프로그래밍 언어로 표현하기에 용이하며 사전자질(lexical feature)들의 규칙적인 분포로 인하여 불등장의 정연한 코우드(code)를 생성할 수 있다.

도식적인 영어문장의 LIG 모형은 사람들의 추상적 분석을 위해서는 커다란 도움이 되지만 준서술적인 이유로 기계가 이해하기는 쉽지않다. 그러므로 LIG의 모형은 기계에 합치될 수 있는 형태의 낮은 급(low level)으로 재편성되어야 한다.

4 .어휘 사전의 설계조건

대부분의 영 · 한 번역 시스템과는 달리 실시간 동작을 이루어내기 위해서는 입력으로부터 초당 수천개이상의 원천언어(영어) 자소(alphabat)를 데이타의 스트림(stream) 형식으로 받아들이고 처리하여야만 한다. 위에 대응하기 위하여는 필연적인 처리속도의 고속화라는 제 1의 설계조건이 발생하는데, 이외에도 속도 저하방지의 반대 급부작용인 비용효과(cost effective)의 후퇴, 회로의 복잡도 증가, 제조 공정의 복잡화와의 가치교환(trade off)도 그외의 설계 요건들이다.

이들 사이의 적합화된 목적구현 제한 요소들로부터 적절한 설계요인을 도출하면 다음과 같다.

(1) 어휘사전의 내용은 불 휘발성 반도체 기억장치 (Read Only Memory: ROM)에 기억되어 사전 탐색속도의 저하를 방지한다.

(2) 어휘사전에 입력된 단어, 숙어 그리고 관용어의 방대한 양을 감안할 때 탐색의 대상이 되는 어휘 항목명(entry name)은 기억 장치의 절감을 위하여 각 어휘의 길이에 따라 신축성이 있는 가변 불등장 코우드(code)를 채택한다.

(3) 어휘의 항목명에는 부속으로 여러 자질(feature)의 속성(attribute)들이 따르게 되는데, 입력 데이타 스트림(stream)을 사전의 어휘 항목명과 비교하는 탐색 모드(searching mode)에서 동작하도록 하고 이 때 2진 탐색법(binary searching)을 사용한다.

위와 같이하여 설계할 사전의 구비 조건을 요약하면 다음과 같다.

『사전의 충족 조건』

a. 자질(feature)값의 적당한 크기

b. 사전의 적절한 크기

c. 사전 정보 추출 양식의 다양화

d. 어휘 항목의 체계적 배열

e. 명확한 기술 형식

본문에서의 어휘사전 구축방식은 트랜스퍼, 더불어 LIG를 사용하고 있는데, LIG의 기계적인 사전 조건 충족방법으로는 사전자체가 언어 생성 체계 (generator)또는 언어 분석(parsing)을 위한 기반으로 보아야 한다는 견해이다. LIG어휘사전의 관계적이고, 합리적인 구성체계는 어휘사전이 단지 문법과 의미 체계의 발현 기반이라는 수동적인 데이타베이스(database)관을 과감히 탈피하고 생성력있는 어휘부를 기초로한 관계적 데이타베이스(relational data-base)임을 보여준다고 할 수 있다.

즉, 지식기반적(knowledge-based) 구현시스템의 장래는 입력(원인)에 대응하는 목적(결과)의 탐색처리를 가능한 많은 입력상황을 설정하고 동떨어져 설치된 지식기반 지원 기구의 단순한 탐색에 의한 수행(processing) 위주의 지식구조 기술이라는 2분적인 원리가 아니라, 잘 정돈되고 간단한 표현에 의해 생성력을 갖춘 지식기반 지원 기구와 이에 결합하는 전처리 발현 지원 기구가 시간의 낭비가 없는 실시간적 요소를 보다 많이 가질 것이다.

5 .어휘사전의 구성

어휘사전은 영어 자소의 데이타스트림 입력으로 부터 그 어휘가 갖고 있는 문법과 의미의 자질(feature) 그리고 그것들을 지시할 수 있는 모든 중간 매개들의 질서정연한 집합체이며 또한 발현자이다.

5-1 .지식 기반 어휘사전

맨 머신 인터페이스(man-machine interface)를 향한 자연어 어휘사전은

① 문법 수행 위주 어휘사전
② 지식기반 어휘사전

과 같이 구분된다. "He is able to learn"이라는 예의 항목에서 ①의 문법기준 처리형은 Noam Chomsky의 변형문법(transformational syntax)과 같이 문(文)의 근저에는 그 기본적 뿌리가 있어 기본적 뿌리의 동적 조성의 연합이 문의 모든 특성 성분을 나타낼 수 있으므로 그 기본적 뿌리의 문법적 특성과 특성 성분들간의 결합, 이해관계에 의해서 문의 성질이 표현될 수 있음을 보여준다고 하겠다. 축차적인 처리와 2치 논리 관계만이 활용성이 있었던 초기의 컴퓨터에 의한 자연언어의 처리는 처리의 간단화라는 면에서 이와 같은 방식이 도입되었다. 그러나 여기에는 커다란 결점이 있어서, 예를들어 "Diamonds are beautiful"이나 "The diamond is very large"처럼 동일한 문법 특성과 동적인 이해관계를 갖는문에서 'Diamond'가 갖는 양갈래의 의미(보석과 야구장의 내야) 여부의 판단이라는 면에서 지극히 낮은 인식률을 보이고 있다고 하겠다. 그러나 발전된 형태의 자연어 처리기들도 이러한 처리방식의 변형과 발전 선상에서 응용

이 이루어지기 때문에 그 기술적 가치는 충분한 인정을 받고 있다.

②항의 지식기반 처리형의 경우에는 단어사전과 병행하여 강력한 숙어 및 관용어의 사전을 가지고 "is able to"라는 어형이 'is', 'able' 그리고 'to'란 개별적인 단어 해석에 우선하여 숙어(idiom)란 점의 또다른 활용 지식을 중요시하여 필요한 해석 기준을 추출토록 하는 방식이다. ②는 이해하고자 하는 원천언어의 문법적, 의미적, 어형적, 그리고 활용적인 대부분의 데이타(data)를 지식 기반화하고 있어야 하기 때문에 보통 대용량의 사전 기억장치와 경우에 따라서는 치명적인 백트랙킹(back tracking)을 하여야 한다는 단점을 가지고 있다.

②의 지식기반 처리형의 어휘사전에서는 "is able to"가 'is'와 'able', 그리고 'to'의 단어 항목보다 높은 층에 있는 숙어이고 또 그 문법적인발현중심이 단어보다는 숙어에 있는것과 같이, ①의 "Diamond"라는 단어는 병렬적인 여러가지의 의미를 그 의미와 전 항목의 의미, 그리고 다음 항목의 의미와의 관계에 의하여 지식기반 사전에 저장된 의미자질(semantic feature)들의 비교하여 적절한 한가지 의미로의 출력이 가능하다. 5장에서는 트랜스퍼 어휘 사전의 지식 기반적인 설계 사상을 저변에 깔아 놓고 있다.

5-2 .EPROM 어휘사전

4의 설계조건 「1」항을 만족시키기 위한 어휘사전의 기능적 다이어그램을 나타내면 그림 5-1와 같다.

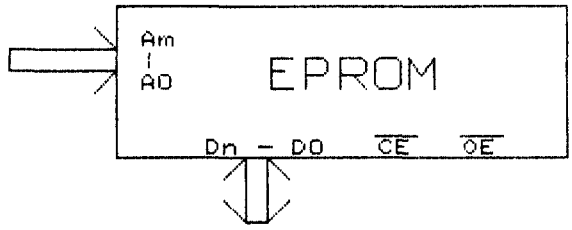

그림 5-1. EPROM에 의한 어휘사전의 기능도 .

Fig. 5-1. Functional diagram of EPROM Lexicon Dictionary

EPROM에는 입력 데이타 스트림과의 비교후 탐색(searching)되어진 특정어휘의 모든 자질(feature)과 그 속성(attribute)들이 들어있다.

이는 하나의 어휘사전이 설계조건에서 전술한 어휘 항목명의 불등장화와 2진탐색(binary searching)을 만족할 수 있는 구조를 위하여 탐색되어진 어휘의 자질과 속성들은 수동 어휘 집합(vocabulary set)에 위치하고 어휘사전의 주변요소들은 단지 수동 어휘 집합을 지적(pointing) 할 수 있는 이중적인 구조를 갖도록 변경 설계되어질 수 있는데 그림 5-2에 보여진다.

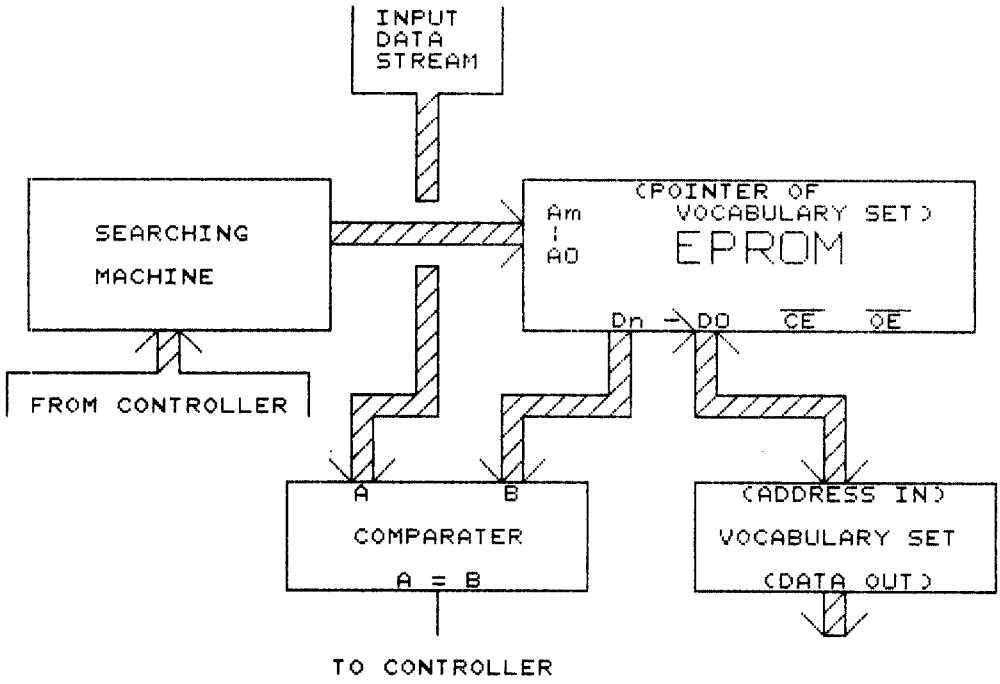

그림 5-2. 어휘사전과 그 주변부의 기능도 .

Fig. 5-2. Functional Diagram of Lexicon Dictionary and its Peripherals .

5-3 .어휘 항목명 (entry name)

다음 세가지 영어 어휘의 구성 예를 보기로 한다.

① I

② ambassador

③ be ready to

①, ②, ③ 전부 독립된 하나의 형태소(形態素)로서 어휘를 구성할 모든 조건을 갖춘다. 각각의 자소(alphabet)에 등장의 특정 코우드(code)를 부여할 때 ①은 1×n bit, ②는 10×n bit, ③은 11×n bit의 구성을 가진다. m개의 자소로 이루어진 원천언어(여기서는 영어) 어휘와 그것이 가질 수 있는 2진 공간은 m×n bit로 나타내어지며 원천언어의 사전에서 항목(entry)으로 가질 수 있는 최대 자소가 바로 하나의 어휘의 기억을 담당할 공간이 되어버린다. 즉 100,000개의 원천언어 어휘를 기억하기 위해서는 100,000×m×n bit라는 엄청난 크기의 메모리가 단지 어휘항목명을 위해서만 배당 되어져야 하며 이의 실제적 구현은 반도체 메모리의 현 시점에서의 집적도의 기술로는 요원하다.

그러므로 항목명(entry name)이 어휘사전 속에서 차지는하는 공간은 각자의 크기와 동일한 불등장이어야 한다는 것 즉, ①과 ②와 ③을 저장하기 위한 어휘사전의 길이는 등장방식의 3×11×n bit에 비하여 1×n bit + 10×n bit + 11×n bit로 줄어든다.

그림 5-3은 그림 5-2의 EPROM에 어휘집합 지시기(pointer of vocabulary set)의 기능적 도시를 어휘 항목명이 불등장으로 고려된 어휘사전의 내용으로 확장 시켜 나타내었다.

플래그(flag)=0일 때	자소코우드	다음 자소의 탐색대상 범위 번지
플래그(flag)=1일 때	자소코우드	자질값 저장기의 번지 포인터

그림 5-3. EPROM 어휘사전의 내용 .

Fig. 5-3. Contents of EPROM Lexicon Dictionary .

위에서 flag는 어휘 항목명의 불등장화를 지지하기 위하여 설치되었다. 별도의 마이크로 제어기는 플래그의 내용을 조사해서 0이면 이전까지의 자소의 조합구성으로 이루어진 어휘가 없기 때문에 다음번의 입력 스트림중 자소 데이타를 확인하여야 한다는 뜻이고 플래그가 1이되면 어떤 하나의 완벽한 어휘의 확인 작업이 이루어진 것으로 그 찾아진 어휘의 모든 속성들이 들어 있는 어휘사전의 자질값 저장기의 번지값이 들어있게 된다. 이는 두번째의 설계조건을 위해 설정한 것이다.

5-4 .2진 탐색기 (binary searching machine)

등간격의 가중치를 정수로 표시하되 2는 3보다 1만큼 작고 10은 2보다 8만큼 크므로 0부터 65535까지의 등간격을 갖는 수의 배열(array)이 있을때 임의의 수를 찾는 방법은 2진 탐색(binary searching)으로 가능하다. 특정한 수치 a를 탐색하는데 있어서 0과 65535의 산술평균치를 b라고 하면 a와 b를 비교하여 a가 b보다 크면, b와 65535와의 산술평균 c를 구하고 c와 a를 다시 비교한다. 이와같은 방식을 반복적으로 적용하면 2^x=65536에서 x는 단지 16번의 탐색처리 수행만으로 a값을 구할 수 있다. 번역의 대상이 되는 원천언어의 자소를 정의하고 그 자소에 가중치를 갖는 부호화(coding)를 하였을때 임의자소의 탐색방법은 수의 나열과 같은 2진 탐색(binary searching)으로

족하다.

5-5. 어휘 사전과 주변부의 전체적 구성

지식기반 처리형의 영 · 한 번역을 위한 어휘 사전에서는 모든 정보(information)가 기억요소(memory element)와 그 주변지원 요소에 다양한 형태를 가지고 퍼져있다. "EPROM 5" - "EPROM 0"는 어휘 집합(vocabulary set)의 지시기(pointer)로 작동을 하며, 퍼져 드리워진 전체적인 기억공간이 2^{23}의 한도 이내에 들것을 요구한다 (기억공간의 번지 지시기인 "23 BIT LO, CO"의 길이와 일치).

빠른 속도의 입력 데이타 스트림(input data stream)은 별도의 마이크로 제어기와 핸드 쉐이킹(hand-shaking) 신호의 수수에 의하여 "4 REG FILES"에 저장된다. "4 REG FILES"는 실시간 번역기의 어휘사전 입력 데이타 스트림이 주로 기계적 이미지 스캐너에 의한 문자인식의 처리결과, 또는 마이크와 음성분석 장치에 의한 음성인식 처리 결과가 대응이 될 것이기 때문에 인식의 불완전성을 가정하여 한가지 입력자소에 최대 4개까지의 오차가 포함된 입력을 허용한다. 오차가 포함된 입력 데이타 스트림은 후에 중간처리 과정을 거쳐 사전탐색에 의한 종합적인 진위(true or not)파악을 할 수 있는 프라이오리티 컨트롤러(priority controller)의 작동을 거쳐 가능한 상태의 다단적인 점수 자질(feature)을 갖고 최종적으로 가장 높은 점수를 갖는 입력의 어휘 처리결과가 형태소 해석되어 출력된다. 그림 5-4에 최종 설계된 어휘사전의 기능적 블럭 다이어그램을 나타낸다.

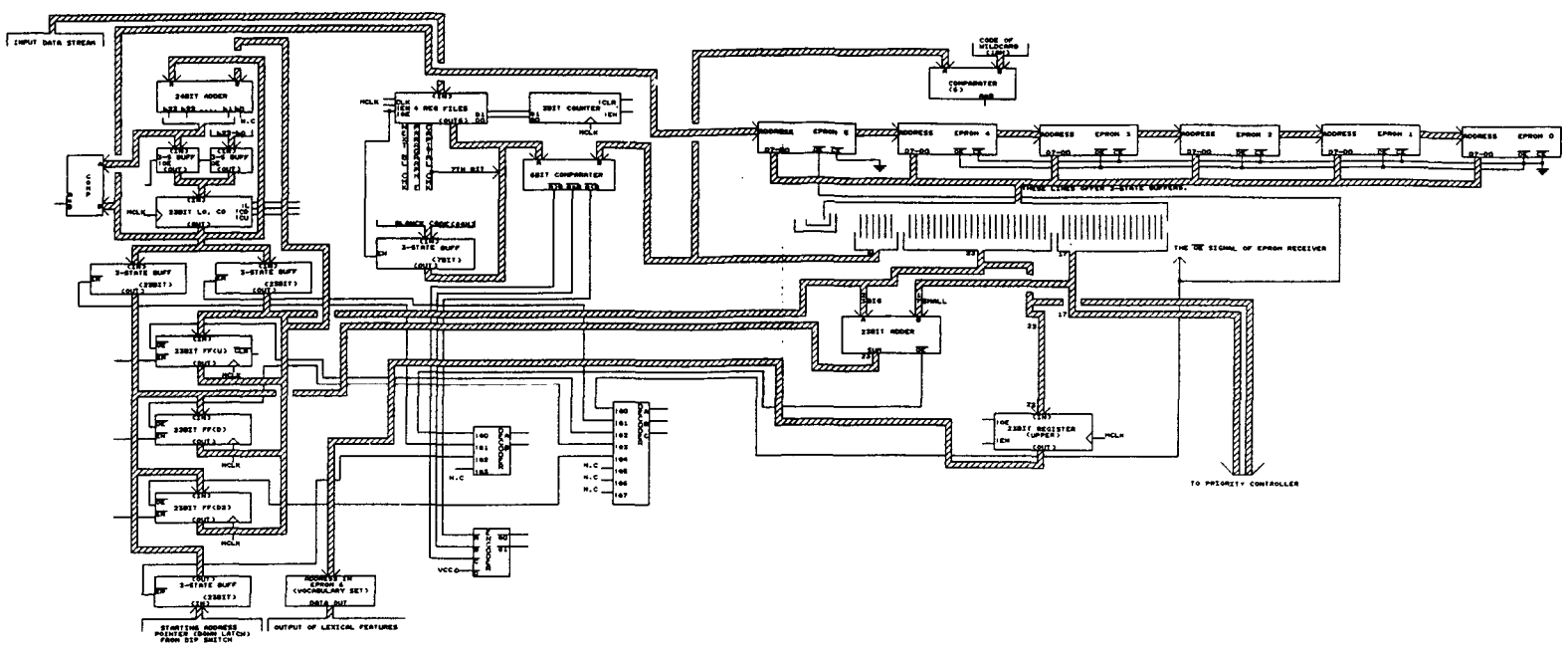

그림 5-4. 어휘사전의 기능적 블럭 다이어그램

Fig.5 - 4 The functional block - diagram of lexicon dictionary

왼쪽의 11개의 블럭은 2진탐색을 구현할 수 있는 회로이며 "6 BIT COMPARATER"는 입력자소와 사전자소와의 비교를 통한 데이타의 탐색유무를 판단하여 준다. "DECODER"라고 이름 붙여진 블럭은 별도의 마이크로 프로그램 컨트롤러로부터의 단지 피일드 디코더(field decoder)의 역할만을 수행하고 블랭크 코우드(BLANK CODE)를 참조할 수 있는 7 bit 의 3가상태의 버퍼는 단어사이의 구별을 하여 준다. 그리고, 가운데 최상단의 컴퍼레이터(comparater)는 "in one's opinion"의 형식을 갖는 숙어를 "in *'s opinion"으로 하여 와일드 카드(wild-card : *) 처리가 가능하도록 하여준다. 48bit의 구성을 갖는 어휘 집합의 지시기인 "EPROM 5" - "EPROM 0"은 2 bit의 플래그(flag), 6 bit의 영어자소의 참조를 대응하고 탐색되지 않은 자소가 퍼져있는 다음번의 번지를 지시하기 위하여 23bit를 사용하며, 만약에 특정어휘가 탐색 되어지면 일단 23bit가 오른쪽 아래의 레지스터에 저장이 되어졌다가 EPROM 6의 어휘집합(vocabulary set)을 지정하여 모든 어휘자질(lexical feature)들이 최종 출력되게 된다.

그림 5-3은 어휘집합 지시기("EPROM 5" - "EPROM 0")의 내부적 데이타 포맷(data format)을 보여주며 아래에 어휘 집합(vocabulary set)속의 여러가지 어휘적 자질(lexical feature)들의 예를 나타내었다.

1. 항목번호	2. 항목명	3. 범주
4. 문법기능	5. 하위범주	6. 부가어
7. 최대투영의 기능	8. 시제	9. 형태
10. 의미내용	11. 의미역할	12. 음성표현

6 .결론

어휘 정보 문법 (Lexicon-driven Information based Grammer : LIG)에 근거한 트랜

스퍼(transfer) 방식은 자연어 번역 처리의 유연하고도 강력한 해석 기준을 제공하며, 또한 잘 정돈된 어휘 사전은 반복적 수행에 의한 처리 시간의 감소 효과와 효율적 정보량의 압축(compression) 효과를 가져온다.

더불어 국부적(global) LIG 사전과 사전 전용의 마이크로 프로그래머블한 제어기(micro-programmable controller)를 사용한 병렬 통합적 하드웨어 구조는 실시간 자연어 번역과 더불어 실시간 자연어 통역 시스템의 도입가능성으로의 발전을 예견할수 있다.

참 고 문 헌

1. 이기용, 1989, "영한 기계번역체계 구축을 위한 소고", 89년 12월 정보과학회지.
2. 정희성, 1987, "한글 구 구조 문법", 제 1 회 자연언어 처리 워크숍 발표 논문집.
3. 이하규, 윤덕호, 김영택, 1989, "기계번역에서 트랜스퍼 방법에 관한 연구", 제 1 회 기계 번역 워크숍 발표 논문집, 시스템 공학 센터.
4. 1989, "영한 변환에 관한 기초연구 및 기계사전의 개발", 시스템 공학센터보고서.
5. Karttunen, L, 1986, "D-DATR : A Development Environment for Unification-based Grammer", CSLI,Stanford Univ.
6. A. S. Hornby , 1975, "Guide to pattern and Usage in English",

Oxford Univ. press.

7. S. M. Shieber, 1986, "An Introduction To Unification-based approach to grammer", CSLI.
8. Mary Dee Mattis, 1985, :Introduction to Natural Language Processing", Reston.
9. G. Gazdar, 1985, "Generalized phrase structure Grammer", Oxford, basial Blackwell.
10. Hutchins. W. J,1986, "Machine Translation-Past, Present, Future", Ellis Horwood Limited.
11. Malvino. A. P,1983, "Digital Computer Electronics", Mcgraw-hill.