한⇒영 대화체 기계번역 시스템

서정연°*, 조정미*, 김길창*
한국과학기술원 전산학과*

A Korean to English Dialogue Machine Translation System

Jungyun Seo°*, Cho, Jeong Mi*, Kim, Gil Chang*
Department of Computer Science, KAIST*

요 약

대화체는 문어체와는 달리 생략과 대용현상이 빈번히 발생하고, 문장의 표면적 의미외에 화자가 전달하고자 하는 의도를 내포하고 있다. 그러므로 대화체 번역은 언어적 분석에 의한 단순한 번역이 아닌, 이해에 기반한 번역이어야 한다. 본 논문에서는 대화의 상황을 모델링한 대화모델을 이용하여 이해에 기반한 대화체 기계번역을 시도하였다. 또한 대화체 기계번역이 자동통역 등에 응용된다고 할 때, 실시간 번역과 불완전한 입력과 같은 예외 상황에 대한 적절한 대응이 보장되어야 한다. 이러한 점을 반영하기 위하여 지식 기반 모델과 확률 기반 모델을 결합한 해석, 생성 시스템을 구현하여 효율성과 견고성을 갖춘 이해에 기반한 대화체 기계번역 시스템을 연구하고자 한다. 이 연구는 한국통신으로부터 지원을 받아서 수행하고 있는 과제로써 현재 3,000단어 수준의 실제 대화를 대상으로 한⇒영 대화 번역에 대해 실험을 하고 있으며, 시스템의 확장성을 고려한 지식 베이스 - 사전, 문법 등 - 를 구축하였다.

1. 서론

대화체 기계번역은 기존의 문어체 기계번역과 많은 점에서 다르다. 첫째, 대화에서는 서로가 알고 있는 많은 사항들이 대화를 통하여 직접 표현되지 않고, 묵시적으로 대화자들 사이에서 오고 간다. 둘째, 대화는 문장의 표면적 의미외에 화자가 전달하고자 하는 의도를 내포하고 있다. 그래서, 대화체 기계번역은 대화의 언어적인 분석에 의한 변환 번역이 아니라 반드시 이해를 기반으로 이루어져야 한다[12, 10]. 이와 같은 이유로 대화체 기계번역은 대화의 상황이 모델링되어야 하며, 필요에 따라서는 생략현상과 대용어에 대해서도 처리되어야 한다.

특히, 대화체 기계번역에서는 불완전한 형태의 문장이 자주 입력된다. 따라서, 불완전한 형태의 입력과 같은 예외 상황에서도 대화의 지속성을 위해서 항상 어떤 형태로든지 출력을 내주어야 한다. 또한 대화체 기계번역의 주요 목표인 동시 통역기와 같은 시스템에서는 사후 교정(post-editing)이 불가능하기 때문에 반드시 올은 번역을 실시간에 출력하여야 한다는 점이 문어체 기계번역과 또 다른 면이다.

이러한 제약 조건을 맞추기 위해서는 번역 시스템이 화자와 발화에 담겨있는 의도를 실시간에 파악하여, 그 의도를 해당하는 상황에 맞게 외역하여야만 할 경우가 많기 때문에 영역지식이나 담화지식과 같이 다양한 형태의 지식이 필요하다. 그러나 많은 양의 지식은 얻기도 힘들 뿐만 아니라 기계번역에 적용할 때에는 여러가지 복잡한 문제를 가지고 있어서 확장하기에 어려운 점이 많이 발생한다. 또한 확률 모델을 이용한 방법들은 대량의 지식을 사용할 수 없음으로 해서 나타나는 애매성 해소에 여러 문제점들을 보완해 줄 것이다.

본 논문은 지식 기반 모델과 확률 기반 모델을 적절히 결합한 해석, 변환, 생성 시스템을 구현하여 효율성과 견고성을 갖춘 이해에 기반한 대화체 기계번역 시스템에 대해서 설명하고자 한다. 현재 3,000단어 수준의 실제 대화를 대상으로 한⇒영 대화체 번역 실험을 하고 있으며, 여러 지식 베이스(사전, 문법 등)을 구축해 나가고 있으며 위에서 언급한 여러 문제들 때문에 현재 대화의 영역을 호텔 예약에서 일어나는 대화로써 제한하여 시범 시스템[1]을 구축하고 있다.

2. 한⇒영 대화체 기계번역 시스템

한⇒영 대화체 기계번역 시스템의 구성도는 [그림 1]과 같다. 다음 절에서 각 부분을 자세히 설명한다.

2.1 형태소 분석기

형태소 분석은 주어진 문장을 의미 있는 가장 작은 단위인 형태소

[1] 본 연구는 한국통신의 장기기초 특제 "자동통역 전화 개발을 위한 대화체 기계번역에 관한 연구"의 지원을 받은 것입니다.

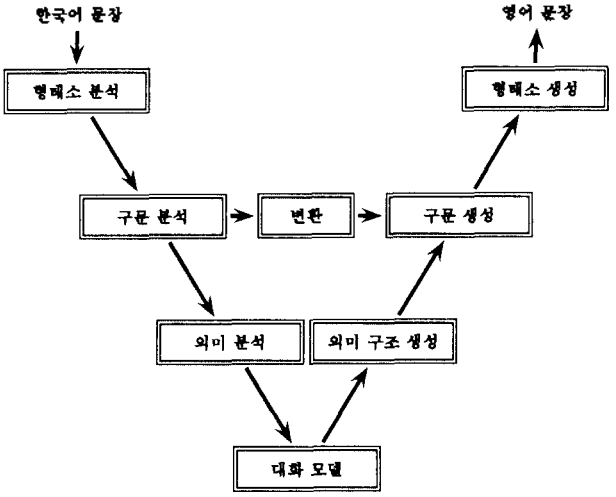


그림 1: 한⇒영 대화체 기계번역 시스템 구성도

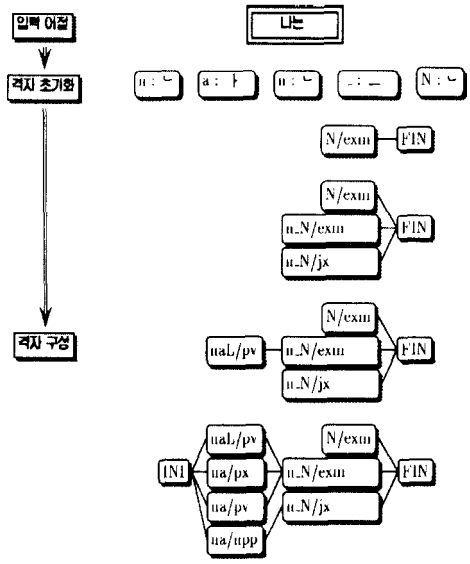


그림 2: '나는'의 형태소 격자 구성

로 나누는 것이다. 하나의 어절을 형태소 분석하면 형태소들이 격자 구조(lattice structure)를 이룬다. 이러한 격자 구조를 쉽게 구성하기 위해, 형태소 시작 가능 위치를 어절의 뒤에서 앞으로 움직이며 사전을 참조하고 이미 찾아진 형태소와의 접속을 조사하였다. 예를 들어 '나는'을 뒤에서 앞으로 형태소를 찾아 격자 구조를 만들어 가는 과정이 [그림 2]이다. 입력 어절을 '이성진 코드[6]'로 변환하고 형태소 분석의 기초 정보로 격자 초기화를 한다. 격자 구조의 에지(edge)를 구성하기 위한 정보, 즉 형태소와 형태소간의 접속 가능 여부는 태깅된 말뭉치에서 한국어 어절을 이루는 품사열을 추출한 뒤, 이를 토대로 접속 가능표(connectability matrix)를 만들어 해결하였다. 그리고 입력 어절의 불규칙 혹은 음운 현상 위치를 추정하는데, [1]의 불규칙 음절 정보를 이용하여 총 20여개의 불규칙 음운 현상으로 나누고 이것을 340개의 '이성진 코드 패턴'으로 변환하여 불규칙-음운 현상 위치를 형태소 분석 전에 미리 추정한다. 본 시스템에서 사용한 형태소 분석기에 대한 보다 자세한 내용은 [4]에 잘 설명되어 있다.

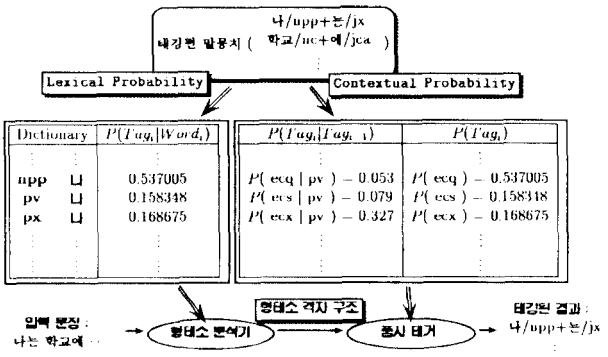


그림 3: 말뭉치와 형태소 분석기와 품사 태거

2.2 품사 태거

일반적으로 형태소 분석 결과는 하나 이상 나타나고 이것은 구문 분석의 효율 저하에 많은 영향을 미친다. 그러므로 형태소 분석 후 품사 모호성을 해소하기 위한 태깅을 하게 된다. 최적 품사열을 구하는 $T(w_{1,n})$는 다음과 같은 수식으로 전개된다[9].

$$T(w_{1,n}) = argmax_{t_{1,n}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \quad (1)$$

태깅 모델은 바이그램을 이용하며, 문맥 확률($P(t_i|t_{i-1})$)과 어휘 확률($P(w_i|t_i)$)은 약 6만 어절의 태깅된 말뭉치로부터 자동으로 구했다. 태깅된 말뭉치로부터 어휘 확률과 문맥 확률을 구하고, 이를 형태소 분석과 품사 태거에 적용하는 과정은 [그림 3]과 같다. 현재 본 시스템에서 사용하고 있는 품사 태깅을 위한 통계적 모델은 [5]에 자세히 설명되어 있다.

2.3 구문 분석 전처리기

구문 분석 전처리기는 구문 분석의 효율을 높이고 모호성을 줄이기 위해 형태소 분석 결과열을 구문 분석 단위인 구문 요소로 묶어 준다[3]. 구문 분석 전처리기에서는 다음과 같은 일을 한다: 1) 어절 내의 결합 (명사와 조사, 어간과 어미의 결합), 2) 숫자열, 날짜열, 시간열의 결합, 3) 본용언과 보조 용언의 결합, 4) 복합 명사의 결합. 구문 분석 전처리에서는 어절 내의 결합 뿐 아니라 입력 문장에서 연속적으로 발생하는 관련도가 높은 둘 이상의 어절들도 결

합한다. 현재는 품사 모호성 해소를 위한 통계적 처리 단계를 거침으로 유일한 하나의 해석열을 입력으로 받는다. 그러나 품사 모호성 해결이 100% 정확하게 된다고 할 수 없음으로, 하나 이상의 해석열의 처리에 대한 고려가 필요하다. 또한 둘 이상의 어절을 다루기 때문에 문맥 정보의 이용에 대한 고려, 어휘 의미 모호성이 발생할 경우에 대한 고려 등이 더 필요하다. 이 단계에서 문장 분석에 필요한 용언부의 격정보가 채워지게 된다.

2.4 구문 분석기

한국어는 문법적인 성격을 표현하는 방식으로서 기능어가 발달한 언어이다. 이들 기능어로 인해 문장을 이루는 문장 성분들 간의 위치적인 정보는 영어에 비해서 상대적으로 중요한 의미를 갖지 않으며, 수식 관계를 갖는 두 성분간의 관계가 문장 구조에 중요한 역할을 하게 된다. 이러한 한국어의 특징은 의존 문법와 특성과 잘 부합됨으로, 본 연구에서는 한국어 분석 방법으로 의존 문법을 이용하고 있다.

의존 문법과 함께 한국어의 '지배소 후위의 원칙'과 구성 성분들 간의 수식 관계를 제한하는 'No Crossing 조건'을 이용해 오른쪽 우선 분석을 한다[2]. 오른쪽 우선 분석은 의존 문법을 이용한 구문 분석시, '지배소 후위의 원칙'에 의해 정의되는 지배 가능 경로를 이용하게 된다. 구문 관계의 검사에 있어 지배 가능 경로는 효율적인 경로를 제공하며, 이때 생성되는 의존 트리 또한 지배소 후위의 특성에 의해 감소하게 됨으로써 전체적으로 효율적인 해석 방법을 제공하게 된다. 이와 같이 의존 문법을 이용한 오른쪽 우선 분석을 할 때, 구문 분석의 모호성을 어떻게 처리할 것인가에 대해 아직 개선의 여지가 남아 있다.

2.5 의미 분석기

의미 분석기는 예를 기반으로 하는 격틀을 이용하여 문장의 의미를 찾아 낸다[13]. 예에 기반한 격틀이란, 동사 격틀의 격 슬롯을 특정 의미 표지(semantic marker)로 표현하는 것이 아니라, 직접적인 예들로 표현하는 것이다. 예에 기반한 격틀의 예는 다음과 같다.

```
[나가다        [[손님,학생,우리],가,AGENT]
               [[방,학교,집],에서,LOC_FROM]
               [[마당,시내,밖],으로,LOC_TO]
               [[작년,세시,오후],에,TIM_AT]]
```

위와 같은 격틀을 동사 격틀 사전에 저장해두고, 입력문과 가장 유사한 격틀을 선택하여 이를 토대로 격구조를 생성한다. 입력문과의 유사도 측정은 격틀의 구조적 유사도 측정과 대상 단어간의 의미 유사도 측정, 두가지로 이루어진다. 격틀의 구조적 유사도 측정은 다시, 격조사의 유사도와 어절간의 위치 유사도로 측정되며, 대상 단어간의 의미 유사도는 단어가 가지는 의미간의 유사도를 측정

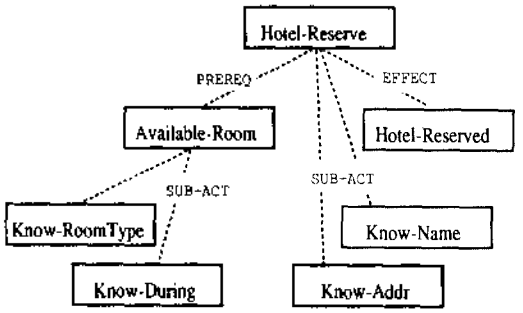


그림 4: 영역 지식의 일부

함으로써 알 수 있다. 이와 같이 유사도 측정을 하여 유사도가 가장 큰 격틀을 선택하고, 선택한 격틀을 이용하여 입력문의 격구조를 생성한다.

2.6 대화 모델

대화체 기계번역에서 가장 중요한 점은 원시 언어로 표현된 화자의 의도를 정확히 목적 언어로 번역하는 것이다. 대화체 문장은 다음과 같은 특징이 있다. 첫째, 생략과 대용어의 사용이 빈번하다. 둘째, 화자의 의도는 다양한 표현으로 나타날 수 있으며 또한 같은 표층 표현일지라도 다른 의미를 나타낼 수 있다. 예를 들면, 한국어의 '예'는 그 사용의 상황에 따라 'Yes','O.K.','Pardon me?' 등으로 번역이 된다. 따라서 화자의 의도에 대한 파악 없이는 정확한 번역이 어렵고, 화자의 의도를 파악하기 위해서는 대화가 진행되는 상황과 문맥, 대화 자체에 대한 언어적인 지식 등이 고려되어야 한다.

화자의 의도는 담화 계획(discourse plan)과 영역 계획(domain plan)을 사용하여 추론할 수 있다. 대화는 기본적으로 질문이나 어떤 요구 유형의 발화와 이에 대한 반응 발화에 의해 진행된다. 이러한 발화를 하나의 행위로 볼 때 이는 계획(plan)으로 표현될 수 있고 이를 담화 계획이라고 정의한다. 영역 계획이란 영역 지식을 계획 기법을 이용하여 구성한 것이다[11, 8].

계획은 인공 지능 문제해결 기법의 하나로, 하나의 문제를 해결하기 위해서는 그 문제를 풀기 위한 선행 조건(precondition)들을 만족시켜야 한다. 그 선행 조건을 만족시키는 문제 역시 그 자체가 다른 문제로써 그것을 해결하기 위한 계획을 찾아내어 그 계획을 순서대로 수행해 나가야 한다. 일단 선행 조건이 만족되면 그 문제를 세부과제로 나누어 하나씩 풀어나간다. 이러한 계획 기법을 이용하면 행동(action)이나 사건(event), 상태(state)를 나타내는 개념들 사이의 관계를 세부 행동(sub_action), 결과(effect), 선행 조건(precondition)등의 관계를 이용하여 표시할 수 있다. [그림 4]는 호텔 예약에 대한 영역 지식의 일부를 나타낸 것이다. [그림 5]는 아래과 같은 대화에서 담화 계획과 영역 계획을 사용하여 화

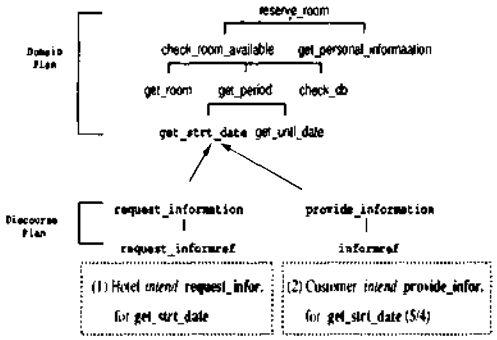


그림 5: 화자의 의도

자의 의도를 기술한 것이다.

1) 호텔 : 언제 예약하실 겁니까?
2) 고객 : 5월 4일날요.

담화 계획과 영역 계획의 추론은 인공지능 연구 분야에서 일반적으로 사용되는 계획 추론 규칙을 이용한다. 본 시스템에서 사용하고 있는 대화모델 및 관련 지식에 대한 것은 [7]에 자세히 설명되어 있다.

2.7 변환

변환은 대화모델을 통한 해석이 실패했을 경우 이를 복구하는 것으로, 견고한 기계번역 시스템을 위해서는 필수적인 것이다. 변환 방법은 원시 언어와 구문 구조로부터 목적 언어의 구문 구조를 바로 생성하는 예문(example)에 의한 변환이다[15, 16].

예문에 기반한 변환은 예문 데이타베이스로부터 입력문과 유사한 예문을 선택하여 번역을 하는 것이다. 번역과정은 첫째, 주어진 입력문에 해당하는 적절한 예문들을 선택한다. 입력문을 번역하기 위해서 이 입력문을 구성하는 서브트리를 찾아내서 이들을 예문 데이타베이스에서 검색한다. 적절한 예문을 선택하기 위한 방법으로는 예문의 서브트리와 입력문으로부터 추출한 서브트리와의 근사도를 계산한다. 둘째, 이렇게 선택된 예문들을 결합하여 번역문의 의존 트리를 생성한다.

예를 들어, [그림 6]과 같은 입력 문장이 들어 왔고, 만들어진 예문 데이타베이스로부터 [그림 7]과 같은 예문들이 선택되었다면 이들이 결합하여 [그림 8]과 같은 결과를 생성한다.

2.8 생성

영어 생성기는 기본적으로 Sheiber와 Pereira의 Semantic-Head Driven 생성기를 변형한 것이다[14]. 이 생성기는 엔진과

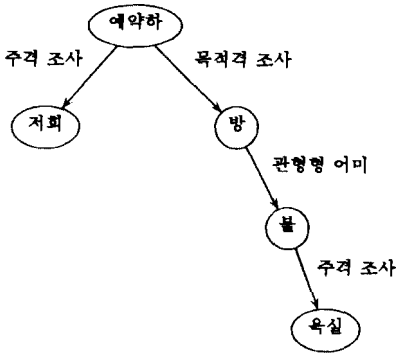


그림 6: 입력 문장:'저희는 욕실이 붙은 방을 예약했는데요.'

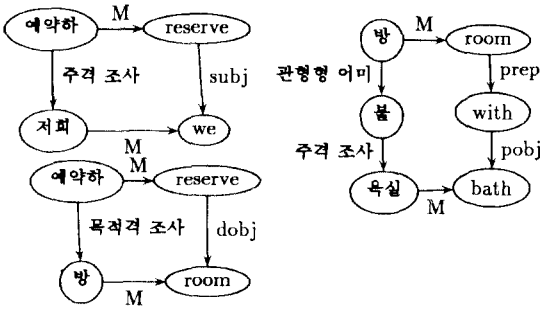


그림 7: 예문 집합

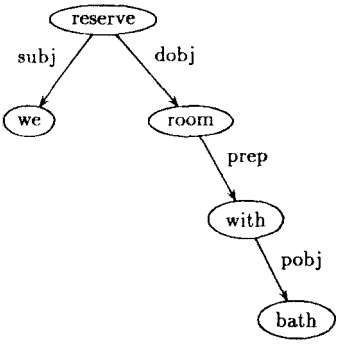
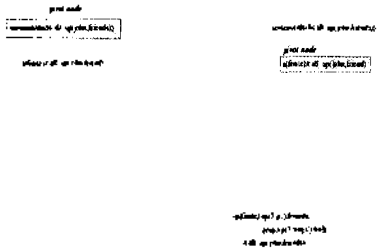


그림 8: 번역문:'We reserved a room with a bath.'

```
(1)   sentence/decl(S) --> s(finite)/S.
      sentence/imp(S) --> vp(nonfinite,[np(_)/you])/S.
(2)   s(Form)/S --> Subj,vp(Form,[Subj])/S.
(3)   vp(Form,Subcat)/S --> vp(Form,[Compl|Subcat])/S, Compl.
      vp(Form,[Subj])/S --> vp(Form,[Subj])/VP, adv(VP)/S.
      ....
      vp(finite,[np(_)/0,np(3-sing)/S])/love(S,0) --> [loves].
      ....
(4)   vp(finite,[np(_)/0,p/up,np(3-sing)/S])/call_up(S,0) --> [calls].
      ....
      vp(finite,[np(3-sing)/S])/leaves(S) --> [leaves].
      ....
(5)   np(3-sing)/john --> [john].
(6)   np(3-pl)/friends --> [friends].
      ....
      adv(VP)/often(VP) --> [often].
      ...
      np(3-sing,X)/friend(X) --> [friend].
```

그림 9: 예제에서 사용한 문법.

예 : John calls friends up.



(a)규칙(1)을 이용 (b)규칙(4)을 이용

그림 10: 간단한 예제에 대한 분석 트리.:(a)(b)

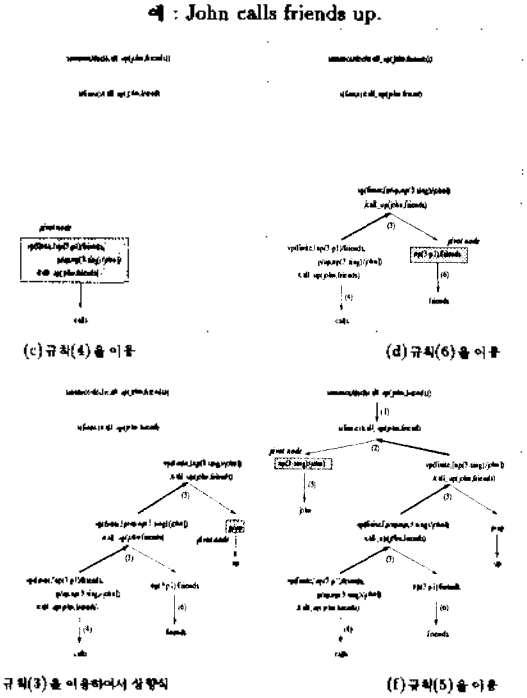
규칙을 분리하여 생성기의 확장성이 좋은 것으로써, 본 시스템에서는 규칙의 판독성을 향상시키기 위해서 보다 읽기 쉬운 규칙 표현법을 정의하고, 그것을 엔진에서 사용하는 형태로 변형시키는 컴파일러를 구현하였다. [문법 9]를 이용하여 영어 문장 "John calls friends up."을 생성하는 과정은 [그림 10, 11]이다.

3. 결론

본 논문은 한국 통신과 함께 수행하고 있는 한⇒영 대화체 기계번역 시스템 개발 현황 및 시스템 전반에 걸친 기술적인 측면을 설명한 것이다. 현재 1,000 단어 수준의 기계번역은 거의 100% 가까운 번역 성공율을 보이고 있으며, 3,000 단어 수준의 복잡성을 가지는 대화체 기계번역 시스템이 실현 단계에 있다.

지금 가장 어려운 문제점은 대화체 문장에 숨어 있는 화자의 의도를 파악하기 위한 대화모델을 견고하게 확장하는 문제이다. 화자의 의도를 파악하기 위해서는 문맥 정보와 각종 지식들이 기계적으로 처리가 되어야 하는데, 처리 단계에서 약간의 예측하지 못한 상황이 벌어지면 실패하게 되는 경우가 생긴다. 이 경우를 대비하여 변환 방식의 번역 방법을 사용하게 되는데, 변환 방식으로 번역된 결과의 정확도에 대한 문제와 변환 방식으로 처리된 문장을 기존의 문맥 정보에 어떻게 포함시킬 것인가 하는 문제는 좀더 연구해야 할 과제로 남아 있다.

예 : John calls friends up.



(c)규칙(4)을 이용 (d)규칙(6)을 이용

(e)규칙(3)을 이용하여서 성립식 (f)규칙(5)을 이용

그림 11: 간단한 예제에 대한 분석 트리.:(c)(d)(e)(f)

자동통역기의 경우에는 당사자들이 서로 마주 보며 언어 이외의 수단으로 정보를 전달할 수 있기 때문에 자동통역 전화기에서의 번역보다는 제약 조건이 약하다고 볼 수 있다. 예를 들면, 손으로 어떤 물건을 가리키면서 발화하는 경우 그 문장의 대명사나 생략문처리가 외의로 문맥 정보를 사용하지 않고 단순 변환 방법으로 번역하여도 의미가 전달될 수가 있을 것이다. 또한 상대방의 눈치를 보면서 자신의 의도가 정확히 번역된 것 같지 않으면 다른 쉽고 정확한 표현을 하여서 다시 번역시킬 수도 있는 상황이 되려라고 보기 때문에 통역전화에서와 같이 많은 대화 지식을 필요로 하지 않을 수도 있다.

대화모델에 기반한 대화체 기계번역 시스템의 강점은, 다음 대화로 어떤 유형에 속하는 대화가 나올 것인지에 대한 지식을 활용하여 음성 인식의 애매성 해소에도 도움을 줄 수 있다는 것이다. 더 중요한 강점은 성공적인 대화모델에 기반한 대화 이해 시스템의 개발로 보다 더 지능적인 인간 통역사와 같은 기능을 가진 번역기를 개발할 수 있다는 것이다. 즉, 컴퓨터가 인간과 같이 번역 도중에 애매한 것이 있거나 이해가 잘 되지 않는 발화가 나타나면 발화자와의 대화를 통해서 완전히 이해를 한 다음 번역을 할 수 있게 될

것이다. 이와 같은 기능을 "지능형 통역 인터페이스"라고 하는데, 이런 기능을 부가하기 위해서는 먼저 인간과 스스로 대화하면서 문제를 해결해가는 대화형 컴퓨터 시스템의 개발이 우선되어야 한다. 이러한 개발을 위해서 본 연구팀은 현재 같은 영역인 호텔 예약을 대화를 통해서 처리하는 대화형 컴퓨터 시스템 개발을 추진하고 있다.

그외 구문 분석의 모호성 및 의미 분석의 모호성 해소를 위한 통계적 처리 모델의 개발과, 최소한의 지식 처리만을 수행하면서도 문맥 정보를 효과적으로 처리할 수 있는 기계번역용 최소 대화모델 개발도 앞으로 수행해야 할 중요한 과제이다.

참고 문헌

[1] 강 승식, 통접 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터 공학과 박사 학위 논문, 1993.

[2] 김창현, 한국어 구문분석을 위한 오른쪽우선 차트파서, 한국과학기술원, 전산학과, 석사학위논문, 1993.

[3] 안동언, 기계번역을 위한 한국어 해석에서 형태소로부터 구문요소의 형성에 관한 연구, 한국과학기술원, 전산학과, 석사학위논문, 1987.

[4] 이상호, 김재훈, 조경미, 서정연, "부분 분석 결과를 공유하는 한국어 형태소 분석," 제 11회 음성통신 및 신호처리 워크샵 논문집, 한국음향학회, 1994.

[5] 이상호, 김재훈, 조경미, 서정연, "한국어 품사 모호성 해소를 위한 통계적 모델," 제 11회 음성통신 및 신호처리 워크샵 논문집, 한국음향학회, 1994.

[6] 이성진, Two-level 한국어 형태소 해석, 한국과학기술원, 전산학과, 석사학위논문, 1992.

[7] 이재원, 김재훈, 서정연, "대화모델을 이용한 대화체 기계번역," 제 10회 음성통신 및 신호처리 워크샵 논문집, 한국음향학회, pp. 104-107, 1993.

[8] S. Caberry, A Pragmatics-Based Approach To Ellipsis Resolution, *Computational Linguistics*, vol.15, no.2, pp. 75-96, 1989.

[9] E. Charniak , C. Hendrickson , N. Jacobson and M. Perkowitz, "Equations for Part-of-Speech Tagging," *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Washington,DC July, 1993.

[10] K. Goodman, S. Nirenburg, *The KBMT Project: A Case Study in Knowledge-Based Machine Translation*, Morgan Kaufmann Publishers, 1991.

[11] D. J. Litman, J. F. Allen, A Plan Recognition Model for Subdialogues in Conversations, *Cognitive Science*, pp. 163-200, 1987.

[12] S. Nirenburg, *Machine Translation - theoretical and methodological issues* -, Cambridge University Press, 1987.

[13] Sadao Kurohashi, Makoto Nagao, "A Method of Case Structure Analysis for Japanese Sentences Based on Examples in Case Frame Dictionary," *IEICE TRANS. INF & SYST.*, 1994.

[14] Stuart M.Shieber, Fernando C.N.Pereira "Semantic-Head-Driven Generation," *Computational Linguistics Vol 16.Number 1*, pp. 30-42, March.1990.

[15] E. Sumita, "Experiments and Prospects of Example-Based Machine Translation," *ACL* , 1991.

[16] H. Watanabe & H. Maruyama, "A Transfer System Using Example-Based Approach," *IEICE* , 1994.