의미패턴에 기반한 대화체 한영 기계 번역

정 천 영† · 서 영 훈††

요 약

본 논문에서는 한국어 대화체를 분석하고 의미패턴에 기반한 대화체 한영 기계번역 시스템에 대하여 기술한다. 한영 기계번역에서 구문정보를 이용한 한국어 분석의 모호성은 의미패턴을 이용하여 해결할 수 있다. 따라서 대화체 번역을 위하여 한국어 스케줄링 도메인으로부터 추출된 의미패턴에 기반한 시스템을 구성한다. 번역의 강건함을 위하여 한국어 문장 분석시 음절을 건너뛰어 분석할 수 있도록 하였으며, 패턴수를 줄이기 위하여 의미패턴에 옵션을 부가하였다. 실험을 위하여 사용된 데이터는 스케줄링 도메인으로 실험결과 88%의 번역율을 보인다.

Machine Translation of Korean-to-English Spoken Language Based on Semantic Patterns

Cheon-Young Jung† · Young-Hoon Seo††

ABSTRACT

This paper analyzes Korean spoken language and describes the machine translation of Korean-to-English spoken language based on semantic patterns. In Korean-to-English machine translation, ambiguity of Korean sentence analysis using syntactic information can be resolved by semantic patterns. Therefore, for machine translation of spoken language, we estabilish the system based on semantic patterns extracted from Korean scheduling domain. This system obtains the robustness by skip ability of syllables in analysis of Korean sentence and we add options to semantic patterns in order to reduce pattern numbers. The data used for the experiment are scheduling domain and performance of Korean-to-English translation is 88%.

1. 서 론

전통적인 기계번역에서는 특정언어의 문법적 구조와 언어간의 변환을 주로 규칙으로 기술함으로써 자연언어의 구문과 의미를 컴퓨터로 처리하고자 LFG[1], HPSG[2]과 Generative Lexicon[3] 등이 제안되었다. 그러나 이러한 문법은 기계번역 시스템을 만드는데는 효율적인 파싱 없이는 불가능하고, 파싱과정에서 많은 모호성이 발생하며, 또한 lexicon을 구축하는데 많은 어려움이 있다.

이러한 과제를 해결하기 Corpus-Based 나 Example-Based 기계번역[4,5], 통계적 기계번역[6], Pattern-Based 기계번역[7,8,9] 기법들이 제안되었으나 이러한 기법은 대부분 문어체 처리를 목적으로 진행되어 왔다.

대화체 문장은 대화체에서 갖는 단어의 축약이나 탈락, 조사의 생략, 수정 또는 반복 발화, 간투어 등의 특성으로 형태소 분석이나 구문분석 등이 현재의 자연

† 정 회 원 : 구미전문대학 전자계산과 교수
†† 정 회 원 : 충북대학교 컴퓨터공학과 교수
논문접수 : 1998년 5월 18일, 심사완료 : 1998년 7월 24일

언어 처리를 위한 문법이나 파싱기법으로 처리하기에는 상당한 문제점이 있다[10,11]. 대화체의 이러한 특성으로 이를 처리하기 위한 여러 가지 새로운 방법이 시도되고 있는데 단어간 확률정보를 이용하는 확장된 문맥 자유 문법을 이용하거나[12], 구문정보를 전혀 고려하지 않은 개념기반의 시스템을 구성하거나[13], 구문과 의미정보를 이용하되 기존의 자연언어 분석 기법을 자연발화 처리에 적합하도록 변형하여 강건한 특성을 포함시킨 기법[14] 등이 대표적인 예이다. 대화체 처리의 어려움으로 문어체에 비해 대화체에 대한 연구가 미흡한 편이며, 특히 한국어 대화체는 연구가 활성화되고 있지 않은 상태이다.

따라서 본 논문에서는 대화체 번역을 위하여 한국어 대화체에서 발생하는 특성을 분석하고, 한국어 스케쥴링 도메인으로부터 추출된 의미패턴에 기반한 시스템을 구성하며, 번역의 강건함을 위하여 한국어 문장 분석시 음절을 건너뛰어 분석할 수 있도록 하였으며 패턴수를 줄이기 위하여 패턴에 옵션을 부가하였다. 또한 패턴을 적용할 때 발생하는 모호성을 감소시키기 위하여 명사의 의미표지를 이용하여 패턴을 정의하는 방법을 제안한다.

2. 한국어 대화체 분석

기존의 자연언어 분석에 대한 연구는 입력 문장이 문법적으로 옳다는 가정하에 분석을 하지만 대화체는 문어체와는 달리 대부분의 발화문이 구문적으로 옳지 않을 뿐만 아니라 문장의 필수 성분 조차도 생략되어 발화된다. 대화체가 가지는 몇가지 특성을 요약하면 다음과 같다.

-. 짧은 말속에 의미를 전달하고 싶어하므로 단어의 축약이나 탈락 현상이 빈번히 일어난다. : 김진홉니다(김진호입니다), 재밌는(재미있는),...
-. 발음이 변하는 현상이 일어나는데 일반적으로 양성음이 음성음으로 발화된다. 보통 'ㅗ'가 'ㅜ'로 변하여 발화된다. : 하구요(하고요),...
-. 안높임체인 어미'요'가 체언, 부사, 조사, 어미 등의 뒤에 자주 붙는다. : 나는요, 빨리요, ...
-. 방언이나 은어 등의 비표준어를 사용한다.
-. 간투어 사용이 많다. : 음, 저, 어, ...
-. 수정 또는 반복 발화 현상이 일어난다. : 세시 아니 네시 네시가 좋겠네요, ...
-. 문장의 성분이 생략되는데 필수격까지도 생략된다. : 네시요
-. 문장의 경계를 찾기 어렵다.
-. 존칭어를 많이 사용한다.
-. 조사가 많이 생략된다. 이는 의미적으로 모호성이 발생되지 않을 때만 생략이 가능한 것으로 보인다.
  너는 나를 좋아하니 --> 너 나 좋아하니
-. 문장 부호가 나타나지 않는다.
-. 아라비아 숫자나 외국어 문자가 나타나지 않는다.
-. 격조사와 보조사의 결합이 빈번하다 : 께서가, 같이도, ...

기존의 대부분 형태소 분석기는 문어체를 처리하기 위하여 설계되었기 때문에 대화체의 특성으로 인하여 형태소를 분석할 때 많은 오류가 발생한다. 이러한 오류를 해결하기 위한 방안으로 형태소 분석기를 대화체 특성에 맞도록 확장할 필요가 있다. 특히 본 논문에서 실험의 대상으로 하고 있는 스케쥴링 도메인은 날짜 관련 표현이 가장 중요한 표현으로 정확한 분석이 요구되며 [11]에서 분석한 바에 의하면 스케쥴링 도메인의 형태소 분석 오류 중 날짜 관련 오류에서 가장 많은 오류가 발생하였으며 음운축약, 비표준어 순으로 나타났다.

따라서 본 논문에서는 스케쥴링 도메인 약 2,500개의 발화문을 대상으로 대화체 특성을 조사하여 기존의 형태소 분석기를 대화체 분석이 가능하도록 보강하였으며 복합명사 결합, 동사 결합 및 보조동사 처리, 명사(을/를) + 동사 (예:예약을 하다 --> 예약하다) 처리, 존칭어 처리 (전화드리다 --> 전화하다), 대역어 변환 등을 패턴을 처리하기 전에 형태소 결과를 입력받아 전처리 하였다. 또한 조사 생략이나 필수격 생략 등은 패턴에서 처리하였다.

3. 의미 패턴

3.1 패턴 모호성

한영 기계번역에서 조사와 어미의 번역 어휘를 선택하는 일은 그들의 의미를 결정하는 것과는 다른 문제이다. 주어진 문맥에 맞는 번역 어휘를 선택하여 목표 언어로 적절한 변환을 수행하는 것이 기계번역의

과제인 것이다. 원시문장에서 조사나 어미의 의미적 역할이 동일하다 하더라도 번역어가 갖는 용법과 어휘 특성에 따라 서로 다른 번역을 하여야 되는 경우가 있다[9].

(1) 나는 서울로 간다.
(1′) I go to Seoul.
(2) 나는 버스로 간다.
(2′) I go by bus.

문장 (1)과 (2)는 동사가 같고 조사 '로'가 같은 형태로 사용되었지만 영어로 번역될 때는 (1′)와 (2′)와 같이 각각 다른 전치사 to와 by를 취하는데 이것은 조사'로' 앞의 어휘의 의미로 발생된 것이다. 따라서 문장 (1)과 (2)를 같은 패턴에 적용할 수 없다.

(3) 철수는 영희와 사과를 먹었다.
(3′) Chulsu and Younghee ate the apple.
(4) 철수는 딸기와 사과를 먹었다.
(4′) Chulsu ate the berry and the apple.

문장 (3)과 (4)는 같은 구문 구조를 보이는 형태로 문장 (3)의 영희를 문장 (4)에서 딸기로만 대체하였다. 그러나 영어로 번역될 때 문장 (3)에서 영희는 주어로, 문장 (4)에서 딸기는 목적어로 번역되기 때문에 문장 (3)과 (4)를 같은 패턴에 적용할 수 없다.

이와같이 어휘 모호성 및 구문 모호성은 패턴 구성시 품사나 구문정보만으로 패턴을 구성 할 때 발생한다. 이러한 모호성을 해결하기 위하여 명사에 의미표지를 도입하여 패턴을 구성한다. 문장 (1)과 (2)는 '로' 앞의 명사 의미에 따라 전치사를 결정할 수 있고, 문장 (3)과 (4)는 패턴 구성시 의미 표지가 같을 때에만 결합되도록 함으로써 모호성 해결이 가능하다.

3.2 의미 패턴의 구성

문장 성분상 서술어는 주어, 목적어와 같은 필수 성분으로 관형어나 부사어에 비해서는 기능상의 상위를 차지한다. 한편 목적어는 서술어의 특성에 따라 필요 유무가 정해지므로 서술어와 주어가 문장에서 가장 중요한 성분이다. 그러나 주어는 생략이 일반화되어 있고, 서술어는 특별한 이유없이는 생략되지 않으므로 서술어가 문장내에서 기능이 더 크다. 국어는 개별언어로서는 물론 보편언어로서도 주어 중심언어라기 보다는 서술어 중심 언어라고 결론지을 수 있다[15]. 따라서 한국어의 문장을 서술어를 중심으로 패턴을 구성한다.

패턴은 의미 패턴의 집합으로 의미표지, 품사, 터미널 심볼, 옵션의 조합으로 번역 대상 문장인 한국어 부분과 한국어 부분에 대응되는 생성 문장인 영어 부분의 쌍으로 구성된다.

패턴에 이용되는 의미표지는 도메인 특성에 따라 다르게 분류될 수 있으나 본 시스템에서 사용되는 도메인은 스케쥴링 도메인으로 장소와 시간 등에 많은 의미가 있기 때문에 <표 1>과 같이 8가지로 분류하고 시스템 처리의 편의를 위하여 proper(대명사)와 conjunction(접속사)을 추가하여 구성하였다. 의미표지는 도메인이 확장됨에 따라 확장될 수 있으며, 다른 의미의 문장이 패턴에 일치되어 목적 문장이 잘못 생성될 경우 분류된 의미표지를 더 세분하여 모호성을 해결할 수 있다.

〈표 1〉 의미표지
〈Table 1〉 Semantic maker

코드 번호	의미표지	설 명
1	animal	동물
2	human	사람
3	location	장소
4	time	시간
5	vehicle	교통수단
6	number	수
7	eating	음식,식사
8	money	화폐
9	proper	대명사
10	conjunction	접속사

〈표 2〉 의미 패턴의 종류와 형식
〈Table 2〉 Sorts and format of semantic patterns

종 류	형 식
동사 패턴	동사 : 한글 동사 패턴 : 영문 동사 패턴
구패턴	한글 구 패턴 : 영문 구 패턴

의미 패턴은 <표 2>와 같이 동사 패턴과 구 패턴이 있는데 동사 패턴은 동사에 따라 패턴이 분류되며 동사 다음의 분리기호 ':' 다음이 한글 동사 패턴이고

한글 동사 패턴 다음의 분리기호 ':' 다음이 영문 동사 패턴으로 구성된다. 구 패턴은 분리기호 ':'를 중심으로 좌측은 한글 구 패턴이고, 우측은 영문 구 패턴으로 구성된다. 패턴 (5)와 (6)은 동사패턴의 예이고, (7)과 (8)은 구패턴의 예를 제시하였다.

(5) 여행하다 : @2*가*이 @4까지 @3*를*을 : @1 @0 @3 @2
(6) 여행하다 : @4간 @3로 : @0 @2 during @1
(7) @3의 @3과 @3를 : @2 and @3 of @1
(8) @4부터 @4까지 : from @1 to @2

한글 패턴에서 @다음의 숫자는 의미 표지 코드 번호이고, 영문 패턴에서 @다음의 숫자는 한글 패턴에서 나온 의미 표지된 비단말의 순서를 가리키며, 동사패턴에서 @0은 한글 동사에 대응되는 영문 동사를 의미하는데 대역어 변환사전에서의 의미와 다른 의미로 쓰일 경우 동사의 의미에 대응되는 대역어를 직접 기술하여 패턴을 구성한다. 또한 한글 패턴에서 의미 표지가 안된 단어는 영문 패턴에서 대역어로 직접 기술한다.

적은 패턴으로 많은 한글 문장을 처리하기 위하여 의미표지 외에 의미표지가 부여되지 않은 단어는 품사를 패턴에 추가하여 기술하였는데 한글 패턴에서 '$+숫자'로 표시하였다. 또한 품사정보 외에 입력문장 분석시 음절을 건너뛰어 분석하거나 반복처리가 가능하고 효율적으로 패턴을 관리하기 위하여 패턴에 옵션 기능을 부가하였다. 패턴에 추가된 옵션의 종류는 다음과 같다.

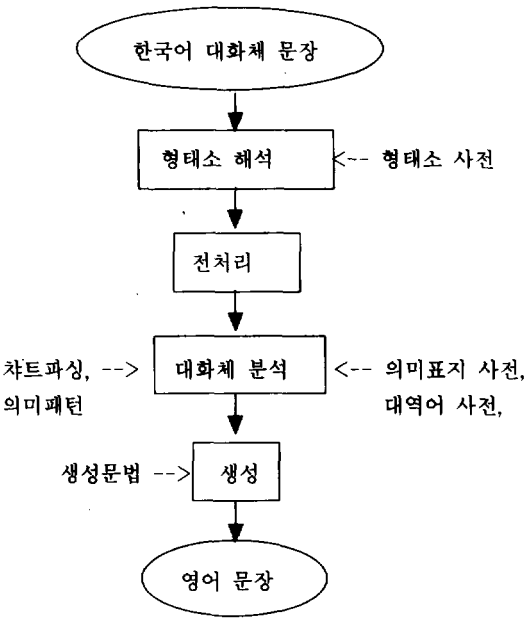
* : 적용될 수도 있고 안될 수도 있는 선택적 성분
+ : 한번 이상 적용이 되는 반복적 성분
*+ : 적용 안될 수도 있고 한번 이상 적용될 수 있는 선택, 반복적 성분
( ) : 상기 옵션을 적용할 때 적용 범위를 설정

한정된 도메인을 대상으로 패턴을 추출하였기 때문에 도메인에 포함되지 않은 단어 및 패턴은 올바른 번역을 위하여 추가되어야 하며, 대화체 특성으로 번역이 어려운 문장, 숙어, 속담 등은 영문 패턴에 번역된 결과를 직접 기술하여 번역의 정확성을 높이도록 패턴을 구성하였다.

영어 문장을 생성할 때 적용되는 의미 패턴은 구패턴, 동사패턴을 순서적으로 적용하여 생성된다. 입력문장에 구패턴을 적용하고 동사패턴을 적용하는데 일치하는 동사패턴이 존재하면 일치된 동사패턴에 의해 영어문장을 생성한다.

4. 생 성

4.1 시스템 개요



(그림 1) 시스템 개요
(Fig. 1) System overview

본 논문의 시스템은 한국어 자연발화문을 영문으로 번역하기 위한 시스템으로 의미패턴을 이용하여 분석과 생성을 수행한다. 본 시스템은 문장 단위로 처리하는데 형태소 해석 시스템이 대화체 기반의 시스템이 아닌 문어체 기반의 시스템이므로 대화체의 특성상 분석될 수 없는 어절을 처리하기 위하여 형태소 해석 결과를 전처리 과정을 통하여 재분석한다. 전처리과정에서는 숫자처리, 형태소 결합, 어미와 보조용언의 인식과 결합 등을 수행한다.

형태소 해석 결과를 전처리한 후 분석과정에서 사용되는 자료구조는 다음과 같다.

```
structure input_tag {
struct input_tag *next; /* 다음 input entry  */
```

```
int id;          /* entry id*/
uchar *str;      /* 원문          */
int sem;         /* 의미표지      */
int tag;         /* 품사          */
int josa;        /* 조사          */
int head;        /* 어근          */
uchar *eomi;     /* 어미          */
uchar *eng;      /* 대역어        */
```

entry id는 각 입력 어절마다 1번부터 일련번호가 부여되고, 의미표지는 <표1>의 의미표지에 따라 코드번호가 부여되는데 정의되지 않은 단어는 '0'번으로 부여하고 대역어는 대역어 설정단계에서 수행하게 된다.

```
typematchhdr*bind_of_sent(type_input *inlst)
{
   if ((verb_tmp=find_verb(fverb)) == NULL){  //한글동사패턴을 찾음
   return (NULL);
   } else {
        patt_tmp = verb_tmp->patt;
        while (patt_tmp) {     // 한글패턴이 남아 있으면 반복
           cstate=make_nethan(patt_tmp->han_patt,&rstr);
//network 만듬
           if (cstate) {
              cur_lst=inlst;
              while(cur_lst) {    //입력이 남이 있으면 반복
              localmhdr=state(cstate, cur_lst);   // 파싱
              if (localmhdr) {          // local 선택
                mhdr=selectbestmatch(localmhdr);
                bestmhdr=makebestmatch(mhdr,patt_tmp,
endlst);
              }
              cur_lst=cur_lst->next;
           }
           free_state(cstate);   //네트워크 반환
        }
        patt_tmp=patt_tmp->peer;  //다른 패턴에 대해 수행
     }
     mhdr=selectbestmatch(bmhdrbase);  //global 선택
     return(mhdr);
  }
}
```

(그림 2) 의미패턴을 선택하기 위한 알고리즘
(Fig. 2) Algorithm for choice of semantic pattern

입력문장을 분석하여 한글 의미패턴을 선택하는 수행 절차는 다음과 같다.

1) 의미패턴에서 입력문장의 동사에 해당하는 한글 동사패턴을 찾는다. 이때 한글 동사 패턴은 여러개 존재한다.
2) 한글 패턴을 유한오토마타(network)로 만든다. 하나의 한글 패턴은 옵션(+, *, *+, ())에 따라 어절단위의 네트워크 즉, 부분 네트워크를 만들고 이를 서로 연결하여 만들어진다.
3) 파싱을 수행하여 분석 결과를 얻는다. 분석은 챠트파서로서 하향식 파싱을 수행한다.
4) 분석결과에서 가장 많이 일치된 것을 저장한다.
5) 남아있는 한글 패턴이 있으면 2)를 반복하여 수행한다.
6) 저장된 분석 결과 중에서 가장 많이 일치된 패턴을 선택한다.

입력 음절과 비교할 때 패턴과 일치하지 않으면 다음 음절을 비교 대상으로 삼는다. 이때 최대로 건너뛸 수 있는 음절수(MAXSKIP)를 넘지않는 범위내에서 비교를 수행한다. (그림 3)은 패턴과 일치하지 않을 때 입력음절을 건너뛰는 알고리즘이다.

```
while (입력 entry가 남아있는 동안 && MAXSKIP) {
    if (비교해서 같으면) {
          path에 저장
          return proc_state (다음상태로, path)
    } else {
          비교할 입력 entry = 다음 입력 entry
    }
}
```

(그림 3) 음절을 건너뛰기 위한 알고리즘
(Fig. 3) Algorithm for entry skip

4.2 생성

목표언어의 생성은 분석결과에 따라 이에 상응하는 생성문법을 이용하여 수행된다. 생성문법은 동사패턴에 따라 구성되는데 선택된 생성문법에 따라 왼쪽에서 오른쪽으로 생성을 해 나가면서 비단말들에 대한 대역어를 그 위치에 삽입하여 생성을 한다. (그림 4)는 입력문장이 의미패턴에 완전 일치하는 경우, (그림 5)는 입력문장이 부분 일치하는 경우, (그림 6)은 중복발화인 경우에 대한 생성 예이다.

입력문장 : 내가 오일부터 팔일까지 영국의 런던과 에딘버러를 여행하려고 하는데요.
동사패턴 : 여행하 : @2*가*이 @4까지 @3*를*을 *+$3 : @1 @4 @-1 @0 @3 @2
구패턴 : @4 부터 @4 까지 : from @1 to @2

형태소가 결합되고 구패턴을 결합한 결과 :

원문	의미표지	조사	어근1	어근2	어미	대역어
제가	human	가	제	NULL	NULL	I
오일부터 팔일까지	time	까지	오일	NULL	NULL	from the 5th to the 8th
영국의 런던과 에딘버러를	locat	를	영국	NULL	NULL	London and Edinburgh in England
여행하려고 하는데요.	verb	NULL	여행하	NULL	want	travel

생성 결과 : I want to travel London and Edinburgh of England from the 5th to the 8th.

(그림 4) 입력문장이 의미패턴과 일치하는 경우
(Fig. 4) A case of the input sentence matched with the semantic pattern fully

입력문장 : 제가 이번에 어머니 환갑이어 가지구요 엘에이 주변을 관광할려고 하는데요
동사 패턴 : 관광하 : @2*가*이 @3 주변을 : @1 @-1 @0 the arround of @2
생성결과 : I want to tour the arround of LA

(그림 5) 입력문장이 의미패턴과 부분 일치하는 경우
(Fig. 5) A case of the input sentence matched with the semantic pattern partially

입력문장 : 십팔일에는 두 대의 항공기가 있는데 열두시와 여섯시 비해 열여섯시 비행기입니다.
동사패턴 : 이 : @4에는 @6 대의 @5*가*이 있는데 @4와 @4 @5 : there is @2 @3 at @1 and is @6 @4 and @5
생성결과 : there is two airplane at 18 th and is airplane 12 oclock and 16 oclock

(그림 6) 입력문장이 중복발화인 경우
(Fig. 6) A case of input sentence repeated

5. 실험 결과

본 연구에서 한국어 대화체 분석 및 생성을 위해 사용된 말뭉치는 한국전자통신연구원에서 소유하고 있는 스케줄링 도메인을 대상으로 실험하였다.

패턴 추출시 동사 패턴수를 줄이기 위하여 어미는 형태소 후처리 과정인 전처리에서 처리하고 본동사만을 대상으로 패턴을 추출하였다. 패턴 추출 및 실험을 위해 사용된 데이터는 93개의 대화로 1,000개의 발화문이다. 실험 방법은 2500개의 문장을 분석하여 패턴 및 사전을 작성하고 나머지 대화를 이용하여 패턴 및 사전을 확장하였으며 실험한 결과는 <표 3>와 같다.

〈표 3〉 실험 결과
〈Table 3〉 Test results

실험 문장수	번역 성공			번역 실패		
	옳은 번역	일부 번역	비율	번역 실패	패턴 없슴	비율
200개 문장	168	7	88	17	8	12
200개 문장	162	2	82	28	8	18
200개 문장	164	11	88	22	3	12
200개 문장	172	6	89	20	2	11
200개 문장	187	1	94	12	-	6
계	853	27	88	99	21	12

실험 결과의 평가는 객관적으로 제시된 평가 방법이 없기 때문에 번역이 올바로 되었는가의 정도에 따라 '옳은 번역', '일부 번역', '번역실패', '패턴없음'으로 구분하여 평가하였다. '옳은 번역'이란 번역이 정확하게 이루어져 확실한 의미전달이 가능한 것을 의미하고, '일부 번역'은 번역이 완벽하게 되지는 못하였지만 전달하고자 하는 기본적인 의미(예:날짜, 장소 등)가 전달되어 발화자의 의도가 전달 가능한 것을 의미하고, '번역 실패'는 패턴이 존재하지 않아 번역이 실패하였거나 다른 패턴이 적용되어 번역이 잘못되는 경우를 의미한다.

<표 3>에서 옳은 번역은 85.3%, 일부 번역은 2.7%로, 대화가 가능하도록 번역이 성공한 비율은 88.0%이고, 번역이 실패하는 경우는 12%로 다른 패턴에 적용되는 경우가 9.9%, 패턴이 없는 경우가 2.1%이었다. 의미패턴을 이용하여 실험한 결과 대화체 분석의 모호성을 줄일수 있었으나 문어체 번역에 비해 개선할 여지가 많이 남아 있다. 번역이 실패하는 경우 중 패턴 적용의 우선순위 알고리즘을 적용하여 최적패턴을 선택토록 하고, 패턴이 존재하지 않아 번역이 실패하는 경우 패턴을 추가하면 번역 성공률은 95% 이상 높일 수 있을 것이다.

<표 3>의 결과는 형태소 해석 결과에서 중의성이 없다는 가정하에서 수행하였다. 또한 입력문 자체의 의미가 애매하여 의미전달이 모호한 경우, 띄어쓰기나 인식이 잘못된 경우, 발화의 중복이 심하여 이해가 어려운 경우 등은 입력문을 수정하여 형태소 해석을 하였다.

6. 결 론

대화체는 문어체와는 달리 비문법적이거나 단어의 축약이나 탈락, 조사의 생략 등의 특징이 있어 분석하는데 상당한 어려움이 있다. 본 논문에서는 대화체 번역을 위하여 한국어 대화체에서 발생하는 특성을 분석하고, 번역의 강건함을 위하여 분석시 음절을 건너뛰어 분석할 수 있도록 하였으며, 패턴 적용시 발생하는 모호성을 감소시키기 위하여 명사의 의미표지를 이용하여 입력문장을 분석한다. 또한 패턴수를 감소하기 위하여 패턴에 옵션을 부가하여 시스템을 구현하였다.

실험을 위하여 사용한 데이터는 스케쥴링 도메인으로 93개의 대화 1,000개의 자연발화문을 대상으로 하였다. 실험결과 88%의 번역 성공률을 보이는데 의미패턴을 이용함으로써 대화체 분석의 모호성을 줄일수 있었으며, 예문을 기반으로 하는 방법보다 효율적으로 패턴을 관리할 수 있었다.

앞으로 의미표지를 더욱 세분화하고 의미패턴에 성(gender), 수(number), 일치(agreement) 등에 대한 정보를 추가하면 번역의 질을 높일 수 있을 것이며, 대동사나 대명사 처리를 하기 위하여 문맥정보를 고려하는 방안도 생각할 수 있다.

참 고 문 헌

[1] Kaplan, R. and Bresnen, J., "Lexicon-Functional Grammar : A Formal System for Generalized Grammatical Representation", MIT Press, pp.173-281, 1982

[2] Pollard, C. and Sag, I. A., "An Information-Based Syntax and Semantics, Vol.1 Fundamentals", CSLI Lecture Notes, Number 13, 1987

[3] Pustejovsky, J., "The Generative Lexicon", Computational Linguistics, 17(4), pp.409-441, 1991.12

[4] Sato, S. and Nagao, M., "Toward Menory-Based Translation", In Proc. of the 13th International Conferenceon Computational Linguistics, pp.247-252, Helsinki, 1990.8

[5] Sumita, E. and Iida, H., "Experiments and Prospects of Example-Based Machine Translation" In Proc. of the 29th Annual Meeting of the Association for Computational Ling-uistics, pp.185-192, Berkeley, 1991.6

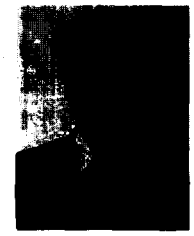
[6] Brown, P. F., et al., "The Mathematics of Statistical Machine Translation : Parametric Estimation", Computational Linguistics, 19(2), pp.263-311, 1993.6

[7] Maruyama,H., "Pattern-Based Translation:Context-Free Transducer and Its Applicat-ions to Practical NLP", In Proc. of Natural Language Pacific Rim Symposium, pp.232-237, 1993.12

[8] Koichi Takeda, "Pattern-Based Context- Free Grammars for Machine Translation", In Proc. of the 34th Annual Meeting of the ACL, June 1996

[9] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국

어 문장 분석과 한영 변환의 모호성 해결", 정보과학회 논문지 제23권 제7호, pp.766-775, 1996.7
[10] 서영훈, "음성언어 번역을 위한 개념기반의 한국어 분석 및 생성", 정보과학회 논문지 제23권 제11호, pp.1176-1184, 1996.11
[11] 최운천, 한남용, 김영성, "개념파서를 이용한 대화체 음성언어 번역", 정보처리학회 추계학술발표논문집, 제2권 제2호, 1995
[12] Seneff, S., "TINA : A Natural Language System for Spoken Language Applications", Computational Linguistics, Vol.18, No.1, pp.61-86, 1992
[13] Levin, E., and R.Pieraccini, "Concept-based Spontaneous Speech Understanding System", Eurospeech '95, pp.555-558, 1995
[14] Levie, A., "GLR*: A Robust Grammar Focused Parser for Spontaneously Spoken Language", Doctoral Thesis, Carnegie-Mellon University, 1995
[15] 이관규, "국어 대등구성 연구", 서광학술자료사, 1992



정 천 영

1986년 충남대학교 계산통계학과 (학사)
1992년 충남대학교 전산학과(석사)
1996년 충북대학교 컴퓨터공학과 (박사수료)
1986년~1997년 한국에너지연구소 연구원
1997년~현재 구미전문대학 전자계산과 전임강사
관심분야 : 기계번역, 자연언어처리, 정보검색

서 영 훈

1983년 서울대학교 컴퓨터공학과 (학사)
1985년 서울대학교 컴퓨터공학과 (석사)
1991년 서울대학교 컴퓨터공학과 (박사)
1988년~현재 충북대학교 컴퓨터공학과 부교수
1994년~1995년 미국 Carnegie-Mellon 대학 기계번역 센타 객원교수
관심분야 : 자연언어처리, 음성언어처리, 기계번역