대어휘 음성인식을 위한 의사형태소 분석 시스템의 구현†

(Implementation of A Morphological Analyzer Based on Pseudo-morpheme for Large Vocabulary Speech Recognizing)

양 승 원*
(Seung-Weon Yang)

요 약 교착어인 한국어를 대상으로 대용량의 대화체 어휘를 포함하는 연속 음성을 인식하는 데에는 인식단위를 결정하는 것이 매우 중요하다. 본 논문에서는 어절이나 형태소를 사용하는 기존의 음성인식 시스템에서의 난점을 해소하고 새로운 인식단위인 의사형태소를 제안하고, 입력되는 문장을 의사 형태소 단위로 분석하는 형태소 분석기와 태거를 구현하였다. 의사형태소를 이용한 음성인식/합성은 어절이나 형태소단위의 음성인식/합성에서 보다 개선된 결과를 얻을 수 있게 해주며, 인식의 출력을 인식의 다음 단계인 언어처리부의 처리단위와 일치시킬 수 있으므로 전체적인 음성언어 번역시스템의 성능도 높일 수 있다. 본 논문에서 구현한 시스템은 일반 형태소를 대상으로 하는 시스템과 동일한 수준의 성능을 보였다.

Abstract It is important to decide processing unit in the large vocabulary speech recognition system. we propose a Pseudo-Morpheme as the recognition unit to resolve the problems in the recognition systems using the phrase or the general morpheme. We implement a morphological analysis system and tagger for Pseudo-Morpheme. The speech processing system using this pseudo-morpheme can get better result than other systems using the phrase or the general morpheme. So, the quality of the whole spoken language translation system can be improved. The analysis-ratio of our implemented system is similar to the common morphological analysis systems.

1. 서 론

음성언어를 처리하는 분야에서는 사람이 직접 자연스럽게 발성한 음성을 입력으로 받기 때문에 입력되는 문장 중에서 인식단위의 경계가 모호하고 신호의 처리 결과에서 이미 오류를 포함하고 있기 때문에 정확한 입력을 기대하기는 어렵다. 이러한 특성을 갖는 대화체 문장을 대상으로 하는 음성 인식에서 보다 양질의 결과를 얻기 위해서는 인식의 단위를 결정하는 것이 중요하다. 인식의 단위를 결정하는 데에는 다음과 같이 대립되는 점이 고려되어야한다.

첫째, 언어학에서 말하는 형태소를 분리하지 않고, 어절을 그대로 사용하는 방법이다. 발성의 지속 시간이 제법 길기 때문에 어느 정도 이상의 인식률을 높일 수 있다.

그러나, 한국어는 교착어로서 하나의 문장소는 의미를 가지는 실질형태소에 조사나 어미처럼 어법적 관계를 나타내는 형식 형태소가 붙음으로서 문법적 기능을 수행한다. 게다가 각 낱말의 어미변화나 활용에 의하여 문장의 성분이 결정되고 불규칙 및 음운현상이 발달한 언어이다. 이러한 특성을 갖는 한국어를 인식할 때, 연속하여 발성된 대용량의 문장에서 나타나는 모든 어절을 사전에 직접 수록하면 인식 대상 단어 수가 늘어나고 많은 미등록어가 발생되어 활용을 크게 제약하기 때문에 인식률의 저하는 불가피하다. 둘째, 언어학에서 정의된 형태소를 분리

† 이 논문은 1999년도 우석대학교 교내 연구비 지원을 받아 연구되었음
* 우석대학교 정보통신컴퓨터공학부 조교수

하여 인식단위로 사용한다. 이렇게 함으로써, 음성 인식 모듈에서는 미등록어(out-of-vocabulary, unknown words)를 최소화하여 인식률을 높일 수 있고 인식의 다음 단계인 기계번역 모듈에서는 인식된 결과를 별도의 형태소 분리과정 없이 직접 사용할 수 있다. 그러나, 언어학에서 말하는 형태소의 많은 경우가 단음소로 이루어져 있다. 예를 들면, 'ㄴ', 'ㄹ', '이' 등은 단음소이나 하나의 형태소이다. 이와 같은 형태소는 음성인식에서 매우 짧은 시간 동안에 발성되기 때문에 이를 인식하기에는 많은 어려움이 있다.

결과적으로 보다 정확한 음성언어의 인식을 위해서는 적절한 길이의 발성 시간과 적절한 수의 인식 단어를 가질 수 있는 처리단위가 새로이 정의되어야 한다. 다시 말하면, 가급적으로 언어학적인 단위인 형태소를 유지하면서 음성인식에 그다지 무리가 가지 않는 범위 내에서 적절한 분리 기준을 따르는 새로운 인식 단위를 정의해야 한다. 본 논문에서는 이러한 요구 조건에 부합되는 인식단위를 정의하고 이것을 의사형태소(pseudo-morpheme)라고 하였다. 또, 정의한 의사형태소 단위의 형태소 분석기와 태거를 구현하였다. 의사형태소는 인식률의 제고는 물론, 통합된 음성언어번역 시스템의 관점에서 보면 음성처리모듈을 위해 일반 형태소보다는 발성의 지속시간이 길고 기계번역모듈에는 어절보다는 풍부한 정보를 갖고 정형화된 처리단위이다. 또한 의사형태소는 언어학에서 정의된 형태소에 그 기반을 두고 있으므로 인식 모듈의 결과를 직접 언어처리 모듈의 입력으로 사용할 수 있게 해주며 인식단계에서 형태소 및 형태소 접속정보 등의 다양한 언어적 지식을 이용할 수 있게 해준다.

본 논문의 남은 부분은 다음과 같이 구성된다. 제 2장에서는 의사형태소를 정의하고, 제 3장에서는 의사 형태소 분석기의 구현에 대해 기술한다. 제 4장에서는 태거에 관하여 설명하고 제 5장에서 결론을 맺는다.

2. 의사형태소

2.1 의사형태소의 정의

의사형태소는 주어진 어절의 소리값은 유지하는 범위 내에서 형태론적으로 최소한의 의미를 가지는 형태소를 말한다. 그러므로, 형태소를 분리해 낼 때 주어진 어절에 어떤 형태의 음소도 첨가되거나 삭제되지 않는다. 어절을 이룰 때 음소가 변하는 경우는 불규칙 현상이나 음운 현상에 의해서 주로 발생된다. 그런데, 불규칙 현상이나 음운 현상에 대한 형태소를 문리할 때 소리값을 유지하면서 형태소들의 원형을 밝히기는 어렵다. 또한 그들의 품사 결정도 매우 모호하다. 따라서 이들에 대한 적절한 분리 기준과 의사형태소에서 품사 결정 원칙이 필요하다. [정의1]은 의사형태소의 분리 규칙이다.

[정의1] 의사형태소의 분리 규칙

① 불규칙 현상이나 음운 현상(축약, 탈락)에 의해서 분리되는 용언일 때에는 <표 1>에서와 같은 기준을 따르며,불규칙이나 음운현상의 종류를 괄호안에 표시한다.

예1) "무엇을 도와 드릴까요." 도/V(b)+와/E

예2) "수건을 꺼내 닦았다." 꺼내/V(ae)

예3) "철수가 운다." 우/V(l)+ㄴ다/E

② 한 어절에 대해 내용어와 기능어로 분리함을 원칙으로 한다,

단, 원형을 분명하게 결정하기 힘든 경우에는 내용어의 품사를 따른다.

예4) "주울 게 많다." 게/Nb

③ ①과 ②를 제외한 것들은 일반 형태소 분석에서와 같다.

여기에서 예1)은 'ㅂ'불규칙, 예2)는 'ㅐ' 축약, 그리고 예3)은 'ㄹ'탈락 현상을 갖는 형태소를 분리한 예를 보여준다. 예를 들면 예1)의 도/V(b)에서 V는 이 형태소의 품사가 동사라는 의미이고 (b)는 이 형태소가 원형이 아니라 'ㅂ'불규칙을 포함하고 있다는 표시이다. 이와 더불어서 괄호 속에 표시하는 불규칙 및 음운현상의 표시는 의사 형태소 분리 결과를 일반적인 언어처리 시스템에서 사용하려고 할 때 의사 형태소를 일반 형태소로 직접 변환하는 데에도 사용한다. 예4)는 원형을 분리하기 어려운 '게'에 대한 품사결정 원칙을 설명하고 있다.

즉, 일반 형태소로 분리할 경우에는 '것/Nb+이/Jo'로 분리되지만 의사형태소에서는 '게' 그대로 분리되어야 하므로 이것의 품사를 Nb(의존명사)로 정할 것인가 Jo(조사)로 정할 것인가를 결정해야 하는데, 이 때, 내용어 '것'의 품사인 Nb를 '게'의 품사로 결정한다. <표 1>에는 불규칙과 음운현상에 의해 달라지는 의사형태소의 종류와 예를 보인다. 이 표에는 일반 형태소에 익숙한 독자들의 이해를 돕기 위해 의사형태소를 일반형태소와 비교하여 두었다.

2.2 토 의

음성의 인식모듈에서, 어절단위의 음성인식은 인식단위의 평균 발성 지속시간이 길어 높은 인식률을 기대할 수 있는 반면, 인식 대상 단어가 증가할수록 미등록어(out-of-vocabulary)가 증가하고 언어모델의 강건성이 떨어져 대어휘(Large Vocabulary) 인식기의 인식단위로는

부적합하다. 또, 형태소 단위의 인식에서는 미등록어를 줄일 수 있지만 발성의 길이가 짧아 인식률이 현저히 떨어지는 난점이 있다[2,6]. 의사형태소를 채택하면 어절단위의 음성인식에서보다 미등록어를 현저히 줄일 수 있고 일반 형태소보다는 발성시간이 좀더 길게 유지되므로 인식 모듈의 성능을 제고할 수 있다.

<표 >1. 불규칙과 음운현상에 의한 의사형태소

불규칙 및 음운현상	의사형태소의 분리예	일반 형태소의 분리예
'으'탈락 1	써서: 써/V(eu1)+서	써서(쓰+어서)
'으'탈락 2	모아: 모/V(eu2)+아	모아(모으+아)
'ㄹ'탈락	우니: 우/V(l)+니	우시고(울+시고)
'이'탈락	소다: 소/N+다	소다(소+이다)
'ㅅ'불규칙	이어서: 이/V(s)+어서	이어서(잇+어서)
'ㄷ'불규칙	물어라: 물/V(d)+어라	물어(묻+어)
'ㅂ'불규칙	도우니: 도/V(b)+우니	도와(돕+아)
'ㄹ'불규칙	흘러: 흘/V(leu)+러	흘러(흐르+어)
'러'불규칙	푸르러: 푸르/V(leo)+러	푸르러(푸르+어)
'우'불규칙	퍼서: 퍼/V(u)+서	퍼서(푸+어서)
'거라'불규칙	가거라: 가/V(geola)+거라	가거라(가+어라)
'너라'불규칙	오너라: 오/V(neola)+너라	오너라(오+너라)
'ㅎ'불규칙	파란: 파라/Pa(h)+ㄴ	파란(파랗+ㄴ)
'애'축약	맸다: 매/V(ae)+ㅆ+다	맸다(매+었+다)
'해'축약	했다: 해/V(hae)+ㅆ+다	했다(하+었+다)
'에'축약	배서: 배/V(e)+서	배서(배+어서)
'여'축약(체)	였다: 여/Jo(ye)+ㅆ+다	였다(이+었+다)
'여'축약(용)	옮겨서: 옮겨/V(ye)	옮겨(옮기+어)
'와'축약	봤다: 봐/V(wa)+ㅆ+다	봤다(보+았+다)
'워'축약	줬다: 줘/V(weo)+ㅆ+다	줬다(주+었+다)
'아'축약	갔다: 가/V(a)+ㅆ+다	갔다(가+았+다)
'어'축약	섰다: 서/V(eo)+ㅆ+다	섰다(서+었+다)
'왜'축약	괬다: 괘/V(way)+ㅆ+다	괬다(괴+었+다)

V:동사,E:어미,N:명사,Nb:의존명사,Jo:조사,Pa:형용사

합성을 하는 과정을 살펴보자. 일반 형태소를 이용하는 경우에 구문분석을 거친 각 문장들은 구문트리로 표현되고, 이 트리의 각 노드는 원형으로 변환된 형태소들이 모여진 문장소들이므로, 합성 시에는 음운 현상을 고려하여 원문을 복원한 다음 음성으로 합성해야만 했다. 이러한 작업은 매우 번거로울 뿐 아니라 완벽한 복원 또한 불가능하다. 그러나, 의사형태소를 이용하면 원문을 다시 복원할 필요가 없기 때문에 정확한 결과를 얻을 수 있다. 또, 복원 루틴을 거치지 않으므로 시스템이 단순해지고 실시간에 더욱 잘 적응할 수 있다.

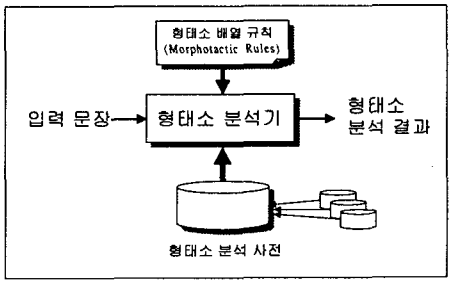
자연언어처리의 연구에 있어서 여러 가지 언어학적인 정보들을 얻어내기 위하여 태그된 말뭉치를 구축하는 것은 매우 중요한 일이다. 그런데 일반 형태소를 이용하여 태깅을 한 말뭉치는 원문과 태깅된 문장의 쌍을 보관한 파일(정렬된 말뭉치:aligned corpus) 이어야만 한다. 그 이유는 태깅 문자열들만 가지고는 원문을 다시 복원하는 것이 불가능하기 때문이다. 그러나, 의사형태소를 이용한 태깅을 한 말뭉치는 원문을 보존할 필요가 없어서 매우 경제적이다. 실제로 [5]를 정렬된 형태로 보관할 때 약 7.5 MB인 반면, 의사형태소 태그된 부분만을 보관하면 3.9MB이다. 이는 보다 정확하고 다양한 정보를 얻어내기 위하여 거대한 말뭉치를 운용하는 최근의 추세에 비추어 볼 때 기억공간의 경제성을 제고할 수 있음은 물론 시스템의 성능에도 지대한 영향을 미칠 수 있다.

의사형태소 분석에서는 불규칙용언의 활용형을 직접 사전에 등재하므로 사전의 크기가 커질 것이라는 우려가 있을 수도 있으나, 우리말의 용언 중 불규칙을 이루는 것은 약 900여 개[1]이고 이들의 활용형을 모두 등록한다해도 약 2800여 개의 사전 엔트리가 늘어나는 것이다.

3. 의사 형태소 분석기

본 논문에서는 챠트를 기반으로 한 형태소 분석기[3,5]를 토대로 의사형태소 분석기를 구현하였다. 의사형태소는 정의에서 본 바와 같이 일반 형태소와 거의 같다. 따라서, 의사 형태소 분석기는 [3,5]의 분석기와 엔진이 동일하므로 본 장에서는 구현된 의사 형태소 분석기를 일반 형태소 분석기와의 차이에 중점을 두고 설명한다. 구현된 시스템의 구조는 그림 1과 같다.

3.1 시스템 구조



<그림 1>. 의사형태소 분석 시스템

분석기의 입력은 음성인식기의 출력으로 대화체 한국어 문장이고, 출력은 어절 단위로 분석한 의사형태소 열이다. 본 논문의 의사형태소 분석기는 자료 구조로 챠트를 사용하였고 형태소 분석 사전과 형태소 배열규칙(Morphotactic)을 이용한다. 또, 내부에서 사용하는 코드

는 이성진[7]코드를 사용하였다. 형태소 분석 사전은 세 개의 소사전을 통합하여 만들어진다. 이 세 개의 소사전은 기본사전, 불규칙 용언 사전, 그리고 불규칙 패턴 사전이다. 의사형태소 분석기를 구현하기 위한 사전의 엔트리는 일반 형태소 분석에 사용되는 사전과는 그 내용이 제법 다르다. 다음에는 각 사전의 내용 중 일부를 보여주는데 이해를 돕기 위해 일반 형태소 사전의 내용을 같이 비교하여 두었다.

3.1.1 기본 사전

기본 사전은 태깅된 코퍼스[5]로부터 자동으로 추출하며, 이 사전의 엔트리의 자료구조는 다음과 같다.

표제어	품사수	품사1 빈도1	품사2 빈도2	.....

사전의 각 엔트리는 세 종류의 필드로 이루어져 있다. 품사수 필드는 하나의 표제어가 여러 가지 품사를 가질 때 그 수를 가지고 있고, 그 뒤에 각 품사와 그 품사로 쓰인 빈도수를 저장하는 필드가 이어진다. 예를 들어, "감기 2 V 3 N 2"가 사전에 등록되어 있다면 '감기'는 2가지 품사로 쓰일 수 있으며, 이들 중 동사로 쓰인 것이 3회, 명사로 쓰인 경우가 2회인 것이다. 또, 의사형태소 단위의 분리를 하다보면 일반 형태소 분석에서와 달리 변칙적인 품사를 갖는 단어들이 발생하는데 이들 역시 기본사전에 등록한다. 예를 들면, 어절 "도와"는 '도/V+와/E'로 분석해야 하므로 조사 '와/Jo' 이외에 새로운 어미 '와/E'도 기본사전의 엔트리에 포함시켜야 한다<표 2>).

<표 2>. 기본 사전

의사형태소사전	일반형태소사전
가깝 1 V 5 감기 2 V 1 N 1 강도 1 N 4 강력 1 N 22 .... 와 2 Jo 39 E 17	없음 감기 2 V 1 N 1 강도 1 N 4 강력 1 N 22 .... 와 1 Jo 39 ...

3.1.2 불규칙 사전

불규칙 용언 사전에는 불규칙 용언에 대한 원형 과 품사, 불규칙 유형이 담겨 있다. 불규칙 용언 사전에도 활용될 때 소리값이 변하지 않는 부분만을 수록해 두었다. 즉, <표 3>에서 보는 바와 같이 불규칙 사전의 엔트리는 원형을 가지고 있지 않으며 일반형태소 분석 시에 사용되던 엔트리는 더 이상 불규칙 정보를 가지고 있을 필요가 없으므로 일반사전에 등록한다. 예를 들면, 어절 "가까운"은 '가까/Pa(b)+운/E'으로 분석해야 하므로 불규칙 정보를 포함한 엔트리는 '가까'로 등록하고, "가깝게"는 '가깝+게'로 분석되어야 하므로 '가깝'은 기본 사전에 등록해야 한다. ('가깝'은 말뭉치로부터 일반 사전에 자동으로 등록될 것이다.) <표 3>에서 IRR_B는 'ㅂ'불규칙을 IRR_S는 'ㅅ'불규칙을 그리고 IRR_U는 'ㅡ'탈락을 사전에 내부적으로 표현하기 위한 기호이다.

<표 3>. 불규칙 사전

의사형태소	일반형태소
가까 Pa IRR_B	가깝 Pa IRR_B
가벼 Pa IRR_B	가볍 Pa IRR_B
짝지 V IRR_S	짝짓 V IRR_S
퍼 V IRR_U	푸 V IRR_U

이 불규칙 정보사전은 말뭉치로부터 자동으로 생성할 수 없으므로 수동으로 작업한다.

3.1.3 불규칙 패턴 사전

우리말의 용언은 어간과 어미가 결합할 때 활용한다. 그래서, 형태소 분석을 할 때 활용에 의해 변화된 어간과 어미를 원형 그대로 복원하려하는데 이를 복원하기 위해서는 변화된 문자열과 변화전의 문자열을 가지고 있어야 한다. 이 때, 사용하는 정보가 불규칙 패턴 정보이다. 불규칙 정보라 함은 어떠한 음운환경에서 탈락, 축약 현상이 발생하며 이 음운환경에서 어떻게 복원되어야 하며, 이 때 발생된 음운 변화가 어떠한 현상인지 등에 관련된 정보이다. 우리의 시스템에서는 이러한 불규칙 패턴 정보를 불규칙 패턴 사전에 모아 두었다. 의사형태소 분석을 할 때, 불규칙 현상이나 음운현상에 의해 활용된 부분을 복원하는 방법은 일반 형태소 분석과는 상당히 다르다. 일반 형태소 분석 시에는 불규칙 패턴사전에는 불규칙 동사들이 활용하는 환경도 가지고 있어야 하지만 의사형태소 분석 시에는 이미 불규칙 용언 사전에 복원할 필요가 없는 형태로 등재되어 있으므로 이곳에 중복하여 보관할 필요가 없다. <표 4>를 보면 일반 형태소 분석에 필요하던 불규칙 패턴 정보들이 의사형태소 분석에는 생략되어 있음을 알 수 있다.

또 의사형태소 분석에서는 축약이나 탈락된 음운현상을 처리 할 때 복원(확장)이 필요 없기 때문에 확장될 부분에 @로 표시를 해 두고 형태소 분석기가 이 표시를 만났을 때 더 이상의 확장은 하지 않는다. 예를 들어, <표 4>에서 "wa w.a o.@ IRR_wa"는 '와'를 보면 'ㅗ.ㅏ'가 축약이 되었다는 것을 알고 확장은 하지 않으며 이 축약의 종류는 'wa'라는 의미이다.

<표 4>. 불규칙 패턴 사전

의사형태소	일반형태소
	# 'ㅅ' 불규칙
	Ae AS.e IRR_S
	Aa AS.a IRR_S
	A_ AS. IRR_S
	# 'ㄷ' 불규칙
	ALe AD.e IRR_D
	ALa AD.a IRR_D
	AL_ AD._ IRR_D
	# 'ㅂ' 불규칙
	s_leN s_leB.N IRR_B
# 와 축약	Awe AB.e IRR_B
wa o.@ IRR_wa	Awa AB.a IRR_B
# '워' 축약	# '와' 축약
we u.@ IRR_weo	wa o.a IRR_wa
# '아' 축약	# '워' 축약
la la.@ IRR_a	we u.e IRR_we
# '어' 축약	# '아' 축약
le le.@ IRR_e	la la.a IRR_a
lye lye.@ IRR_e	# '어' 축약
# '왜' 축약	le le.e IRR_e
w8 wi.@ IRR_w8	lye lye.e IRR_e
	# '왜' 축약
	w8 wi.e IRR_w8

이것의 사용에 대해서는 2절에서 보다 자세히 설명한다.

3.1.4 형태소 배열 규칙

일반적으로 형태소 분석 단계에서는 사전에 등재된 모든 단어를 대상으로 형태소를 분리해 내기 때문에 과생성 문제가 심각하다. 실제로 어절 '소나무라고?'는 51개의 형태소 분석 결과를 갖는다[3]. 따라서, 본 논문에서는 연속되어 나타날 수 있는 형태소들의 bigram 정보를 이용하여 형태소 분석의 과생성을 제약한다. 예를 들면, 배열규칙 파일에는 명사 뒤에는 조사가 올 수 있다는 정보를 "N Jo"와 같이 등록되고 동사 뒤에 어미가 올 수 있다는 정보를 "V E"와 같이 등록해 둔다. 이러한 배열 규칙은 의사형태소 단위로 태깅된 말뭉치[5]로부터 추출하였다.

<표 5>에서는 우리의 분석기에서 사용한 의사형태소 배열규칙 중의 일부를 보여준다. 그런데, 이 의사형태소 배열 규칙에는 언어학적으로는 어색한 bigram 정보도 포함하고 있다. 일반적인 형태소 제약 관계와 의사 형태소에서 필요한 제약 관계를 살펴보자.

- 일반 형태소 분석기 : "나무다" => 나무/N+이/Jo+다/E
- 의사 형태소 분석기 : "나무다" => 나무/N+다/E

일반 형태소 분석 시에는 명사의 다음에 어미가 나타날 수 없지만 대화체 문장의 의사형태소 분석 시에는 "나무다"에서와 같이 명사 다음에 바로 어미가 오는 경우도 있다. 따라서 이렇게 새로운 형태소 제약 관계도 설정해 두어야만 한다.

<표 5>. 형태소배열 규칙

N E
E Jo
E E
Jo E
...
V space

<표 5>에서 배열규칙 'N E'는 명사 뒤에 어미가 바로 올 수 있다는 의미이다. 이 규칙에 의해서 어절 "나무다"는 '나무/N+다/E'로 분석에 성공한다.

3.2 의사형태소 분석기의 구현

앞 절에서 설명한 사전을 사용하여 의사형태소 단위로 분리해 내는 분석기의 알고리즘은 [알고리즘1]과 같다.

[알고리즘1]

1. Loading Dictionary;
2. Initialize edge;
3. Get_an_Edge(edge);
4. if (find_in_dictionary(edge, dict))
   write the information(tag, frequency, irregular) to edge;
5. if (the edge has irregular feature)
   expand_irr();
6. if (Get_an_Edge(edge) != NULL) goto 4;
7. Select the valid edges with morphotactic rule;

위의 알고리즘의 step4에서 에지에 포함된 모든 원소들에 대하여 사전에 있으면 해당 품사를 붙여 에지에 추가한다. 따라서, 기본 사전에 등재된 단어들은 물론 "ㅅ, ㅂ, ㅎ, ㄷ"등의 불규칙 용언들도 step4에서 모두 처리된다. 왜냐하면, 이러한 불규칙 용언의 활용된 형태를 사전의 표제어로 등록하였기 때문이다. 예를 들어, 불규칙 용언 사전에 "가까 Pa IRR_B"가 있으므로 "가까운"이라는 용언에 대하여 사전 탐색만으로도 "가까/Pa(b)+운/Em"을 분석해 낼 수 있다. step5의 expand_irr()는 일반 형태소 분석기에서는 불규칙 용언의 원형을 복원하기 위하여 확장을 하는 함수 인데, 본 분석기에서는 불규칙 패턴들을 이용해서 축약과 탈락현상의 종류를 판별해내 원래 모양으로 결과를 나타낼 수 있도록 하였다. 즉, 의사 형태소

분석의 경우, 음운 환경을 이용하기는 하지만 복원한 내용을 형태소 분석의 결과로 만들지 않는다. 다만 복원한 형태소를 이용하여 사전을 탐색하여 복원된 단어가 사전에 있는지를 검증할 뿐이다. 예를 들어, "옮겨"라는 어절에 대한 형태소 분석 과정을 불규칙 패턴 정보를 중심으로 살펴보자. '겨'의 음운 환경은 "기+어"가 축약 된 것이다. 따라서 입력 어절에 '겨'가 나타나면 아래와 같은 환경을 조사한다.

겨 : 기 + 어 (gye gy.e gi.@ IRR_ye_)

위에서 표현된 메타 표현으로부터 '겨'라는 음절이 발견되면, "기+어"의 축약임을 알 수 있으므로 '겨' 앞에 나타나는 음절과 '기'를 결합하여 사전에서 단어를 탐색한다. 좀 더 구체적으로 살펴보자. 음절 '겨(gye)'를 보면 '기(gy)'와 '어(e)'로 분석할 수 있음을 알고 '기(i)'라는 단어가 사전에 있는지 탐색한다. 또, 'ㅁ기', 'ㄹㅁ기', 'ㅭ기', '옮기' 순서로 계속 사전을 탐색하면서 사전에 있으면 모두 에지에 추가한다. 물론 에지에 추가되는 단어는 기본형이 아닌 의사형태소 단위이다. 예를 들어 사전에는 '기'와 '옮기'가 등록되어 있지만, "기+어"로 확장하지 않고 '겨', '옮겨'등으로 변환하여 에지에 추가한다. 이러한 현상을 반영하기 위해 '기'가 발견되어도 원형 복원하여 에지에 추가하지 않도록 적절한 마커(@)를 둔 것이다.

이와 같은 절차를 거쳐 모아진 에지에서 정의된 형태소 배열규칙에 위배되지 않는 것들만을 고르면 최종적인 분석결과이다. 이제 "도와"와 "세워"라는 어절에 대한 일반 형태소분석 과정과 의사형태소분석 과정은 <표 6>과 같다.

본 논문에서는 의사형태소 분석기의 구현과 비교를 위해서 일반 형태소 분석기로 98.7%의 성공률을 갖고있는 [3]을 선택하여 사용하였다. 이 형태소 분석기는 문어체 문장을 분석하기 위한 형태소 분석기인데 본 논문에서는 이 분석기를 모태로 의사형태소 분석기를 구현하였다. 본 논문의 분석기는 Sparc Station solaris 2.5하에서 C(gcc) 언어를 사용하여 구현되었다. 실험에 사용하는 사전과 배열규칙은 여행계획 영역에서 얻어진 대화체 12,934문장을 의사형태소 단위로 태깅한 말뭉치[5]로부터 생성하였다. 이 사전의 엔트리 수는 48,083개이다. 실험결과 의사형태소 분석기의 성능은 98.9%의 성공률을 보였다(분석에 실패한 경우는 주로 지명이나 인명 등이었다). 이는 동일한 실험 환경에서 일반 형태소 분석기의 성능과 차이가 없음을 시사한다.

<표 6>. 형태소분석과정

일반 형태소 분석 과정	의사 형태소 분석 과정
"도와" : step4; => "도"를 탐색(도/N) => "와"를 탐색(와/Jo) => 찾아진 어절을 에지에 추가 {(도/N+와/Jo),...,(도/Jo+와/Jo)} step5; => 음운 환경 조사 ㅗ+와 => ㅂ+아 (owa=>ob+a) /*IRR_B조건 복원*/ => "옵"을 탐색 => "돕"을 탐색(돕/V) => "아"를 탐색(아/E) => 찾아진 어절을 에지에 추가 {(돕/V+아/E)..} step7; => 형태소배열규칙으로 필터링 {(돕/V+아/E),(도/N+와/Jo)}	"도와" : step4; =>"도"를 탐색(도/N,도/V(b)) =>"와"를 탐색(와/Jo,와/E) =>찾아진 어절을 에지에 추가 {(도/V(b)+와/Jo),),..,(도/N+와/E)} step7; =>형태소배열규칙으로 필터링 {(도/V(b)+와/E),(도/N+와/Jo)}

4. 태 거

형태소 분석을 하는 과정에서 하나의 입력 문장에 대해서 여러 가지의 분석이 가능하다. 이 결과를 음성인식 시스템에서 직접 사용하려면 이 여러 가지 분석결과 중 발화된 환경에서 가장 적합한 하나의 결과를 찾아내어야만 한다. 본 논문에서는 제안한 의사형태소를 인식이나 언어처리부에 직접 사용하기 위하여 태거를 구현하였다. 의사형태소가 언어학적인 형태소의 정의를 최대한 수용하고 있으므로 태거의 알고리즘도 형태소 분석기와 마찬가지로 기존의 일반적인 태거와 같은 엔진을 사용할 수 있다. 단지 기존의 태거에서 사용하던 자료구조를 의사 형태소 분석기에서 사용하는 것과 일치하도록 수정하고 출력 루틴에서 의사형태소의 형태로 출력하도록 수정하면 된다. 본 논문에서는 태거의 엔진으로 HMM을 이용한 ETRI 태거를 이용하였다. HMM을 이용하는 태거에서는 여러 가지 확률값을 구해 태깅에 사용한다. 이 확률값들은 태깅된 말뭉치로부터 추출하는데 의사형태소 태거에서 사용하는 확률값은 의사형태소 단위로 태깅된 말뭉치로부터 얻어와야 하므로 확률값을 추출하는 데에 [5]의 말뭉치를 이용하였다. 구현한 태거를 여행계획 영역에서 얻은 문장들을 가지고 실험한 결과 96.8%의 성공률을 보였다. 이러한 성공률은 문어체 문장을 대상으로하여 발표된 기존의 태거들[3,8]에 비하여 1% 정도 낮은 것이지만 본 논문의 태거

가 지명, 인명, 숫자 등을 많이 포함하고 있는 구어체 문장을 대상으로 하는 것임을 감안하면 이러한 차이는 무시할 수 있는 정도이다.

5. 결 론

음성 언어 번역 시스템에서 각 모듈간의 통신 수단으로 잘 적응할 수 있는 의사형태소를 제안하고 이를 위한 형태소 분석기와 태거를 구현하였다. 의사형태소의 정의 시에 일반형태소분석의 결과를 이용하는 기존의 언어처리시스템과의 호환성을 고려하여 일반형태소로의 직접 변환을 위하여 형태소 태그의 괄호 안에 불규칙이나 음운현상의 종류를 부가하여 두었다. 이 변환은 별도의 도구없이 텍스트 변환과정으로 이루어진다. 시스템은 Sparc Station solaris 2.5하에서 C언어를 사용하여 구현하였는데, 여행계획 도메인에서 얻어진 문장을 토대로 실험을 한 결과, 의사형태소 분석기는 98.9% 태거는 96.8%의 성공률을 보여 일반 형태소단위의 분리 때와 동일한 성능을 보였다.

본 논문에서 정의한 의사형태소가 일반의 형태소보다는 발성의 길이가 길다고는 하나 언어학적인 형태소에 기초를 두고 있으므로 어절에 비하면 짧은 편이다. 더욱이 어절에서는 찾아 볼 수 없는 단음소도 하나의 형태소로 분리하기 때문에 인식률이 저하되는 요인으로 작용할 수 있다. 따라서, 실제 음성인식 모듈에서 본 연구의 결과를 이용하기 위해서는 의사형태소들을 적절히 결합하여 인식에 적용하는 노력이 필요하다.

참 고 문 헌

[1] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터 공학과 박사학위논문, 1993.

[2] 권오욱, 박준, 황규웅, "의사형태소 단위 대어휘 연속 음성 인식기 개발," 제15회 음성통신 및 신호처리 워크샵 논 문집, pp.320-323, 1998.

[3] 김재훈, 오류-보정 기법을 이용한 어휘 모호성의 해소, 한국 과학 기술원 박사학위논문, 1996.

[4] 남기심, 고영근, 표준 국어 문법론, 탑출판사, 1987.

[5] 양승원, 의사형태소 태깅과 의사형태소 해석기 및 태거의 개발, 한국전자통신연구원 용역결과 보고서, 1998.

[6] 이경님, 정민화, "의사형태소 단위의 음성언어 형태소 해석," 제 10회 한글 및 한국어 정보처리 학술대회, pp396-404, 1998.

[7] 이성진, Two-level 한국어 형태소 분석, 한국과학기술원, 전산학과, 석사학위논문, 1992.

[8] 임희석, 언어지식과 통계정보를 이용한 한국어 품사 태깅 모델, 1997.