韓國컴퓨터情報學會 論文誌
第14卷 第2號, 2009. 2.

2009-14-2-4-5

한영 모바일 번역기를 위한 강건하고 경량화된 한국어 형태소 분석기

여 상 화*

A Light Weighted Robust Korean Morphological Analyzer for Korean-to-English Mobile Translator

Sanghwa Yuh*

요 약

본 논문에서는 핸드폰, 스마트폰, PDA폰 등의 모바일폰에서 동작하는 강건하고 경량화된 한국어 형태소 분석기를 제안한다. 이들 모바일 장치들은 낮은 CPU 성능과 메모리 사용에서의 제약으로 인해 자연언어 인터페이스를 적용하기 곤란했다. 본 논문에서는 1) 키 이벤트 핸들러 (Key Event Handler)를 이용한 온라인 형태소 분석과 2) 붙여쓴 오류 문장에 대해 띄어쓰기 오류 교정의 전처리 과정 없이 강건한 형태소 분석 방법을 제안한다. 본 논문에서 제안된 경량화된 한국어 형태소 분석기는 모바일 한영 번역기 시제품에 적용하여 메모리 사용량은 5.8% 줄이고 평균 반환 시간은 19.0% 개선하였다.

Abstract

In this paper we present a light weighted robust Korean morphological analyzer for mobile devices such as mobile phones, smart phones, and PDA phones. Such mobile devices are not suitable for natural language interfaces for their low CPU performance and memory restriction. In order to overcome the difficulties we propose 1) an online analysis by using Key Event Handler mechanism, 2) and a robust analysis of the Korean sentences with spacing errors without its correction pre-processing. We adapt the proposed Korean analyzer to a Korean-English mobile translator, which shows 5.8% memory usage reduction and 19.0% enhancement of average response time.

▸ Keyword : 형태소 분석(Morphological Analysis), 온라인 분석(On-line Analysis), 모바일 번역기(Mobile Translator)

• 제1저자 : 여상화
• 투고일 : 2008. 12. 24, 심사일 : 2009. 1. 3, 게재확정일 : 2009. 2. 10.
* 경인여자대학 정보미디어학부 부교수
※ 본 연구는 2007년도 경인여자대학 교내연구지원 연구비에 의해 수행되었음.

Ⅰ. 서 론

외국을 방문하는 한국 여행객의 가장 큰 불편사항은 "외국어로의 의사 표시" 문제이다. 영어가 공용어로 일반화되어 있지만, 대부분의 한국인 여행객이 외국어로의 의사소통이 어려운 것이 현실이다. 최근 해외여행자의 급증과 WCDMA 등의 글로벌 로밍 서비스가 일반화되면서 해외여행자를 위한 언어 번역 기능을 내장한 핸드폰이 보급되고 있다. 이러한 언어 보조도구로는 전통적인 전자 사전, 여행용 외국어 회화집과 같은 단순한 형태에서 OCR 기능을 사용한 전자사전, SMS (Short Message Service)를 통한 자동번역 (예: 모바일 번역나라 (아이티플러스, 2005년), M-MTS (LNISOFT; 2005년)) 등 보다 진보된 형태의 제품이 출시되어 있다. 그러나 기존의 시스템들은 단발성 발화만 가능한 수준으로 사용자가 외국인과 자유로운 의사표현에 이용되기에는 불가능한 수준이다.

본 논문에서는 점차 고기능화되고 있는 모바일폰에서 단문 메시지(Short Message Service; SMS)와 같은 자유문장 입력 방식의 모바일 회화 번역기를 위한 경량화된 한국어 형태소 분석기를 제안한다. 모바일폰에서 키패드를 이용한 문자 입력은 가장 빠르고 보편적인 입력 수단이다. 문자메시지 사용은 청소년뿐만 아니라 중/장년층에도 일반화되는 추세이다. 본 논문에서는 핸드폰, 스마트폰, PDA폰 등의 모바일폰의 낮은 CPU 성능과 메모리 사용으로 인한 제약을 극복하기 위하여 키 이벤트 핸들러 (Key Event Handler)를 이용한 온라인 형태소 분석과 붙여 쓴 오류에 강건하면서 경량화된 한국어 형태소 분석기를 제안한다.

본 논문의 구성은 다음과 같다. 2장은 기존의 모바일폰용 회화번역기와 한국어 형태소 분석기를 살펴보고, 3장에서는 빠른 반응시간을 제공하기 위해 본 논문에서 제안하는 이벤트 기반의 온라인 형태소 분석을 설명한다. 4장에서는 실험 및 평가 결과를 제시한다. 마지막으로 5장에서는 결론을 맺는다.

Ⅱ. 관련 연구

1. 모바일 회화 번역기

컴퓨터를 이용하여 인간의 언어를 다루는 자연언어 처리 (Natural Language Processing) 시스템은 대량의 외부 지식과 O(n3)의 높은 시간 복잡도로 인하여 주로 고성능의 PC나 서버용으로 개발되었다. 따라서 200~600MHz의 낮은 CPU성능과 32MB~128MB의 주메모리(Main Memory)를 갖는 모바일폰 (PDA폰, 스마트폰 등)에는 적당하지 않은 것으로 인식되어 왔다. 모바일폰용 응용프로그램에서 사용할 수 있는 Heap Size의 경우, IM-8300 단말기(SK사)가 최대 12M를 사용할 수 있는 반면 열악한 폰은 1.5M인 폰도 있다.

최근에 PDA폰이나 스마트폰의 하드웨어 사양이 고성능화되면서 자동번역과 같은 언어 번역 기능을 탑재하려는 시도가 있어 왔다[1, 2, 3]. 그러나 대부분 자유문장에 대한 자동 번역보다는 자국어와 외국어의 대역 문장을 데이터베이스로 구축하고 이를 사용자가 상황에 맞는 대화 문장을 찾아가는 방식을 채택하고 있다.

그림 1은 W사의 전용 하드웨어 방식의 모바일 번역기인 OH-100의 동작 예이다. OH-100은 언어 쌍별로 6,100개의 한국어 문장의 영/일/중 대역 문장을 DB로 가지고 있다. 6,100개의 문장은 공항, 교통, 숙박, 식당, 은행, 쇼핑, 전화, 긴급상황, 회화, 기본용어, 스포츠, 시설 이용의 12개의 상황으로 분류되어 있다.

사용자는 원하는 문장을 찾기 위해 그림 1과 같이 최소 5단계의 탐색 과정을 거쳐야한다. 또한 각 상황별로 나열된 수십 개의 문장에서 원하는 문장을 찾기 위해 여러 번의 화면 스크롤 (Scroll)을 수행해야 한다.

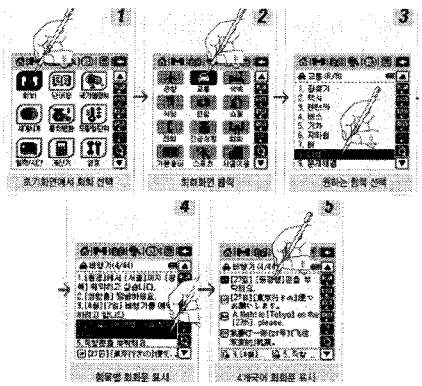

그림 1. 기존의 모바일 번역기의 번역 과정
Fig. 1. Translation Procedure of an Existing Mobile Translator

이와 같은 방식의 기존 시스템은 번역 문장의 정확도 측면에서는 완벽하지만 다음과 같은 단점이 있어 이를 이용하여 여행자가 실제로 외국인과 대화하기는 불가능하다.

- 자유 문장 번역이 불가능하므로 미리 수록된 문장 이외의 발화가 불가능하다.
- 한 번의 발화를 위해 대상 문장을 검색하기까지 많은 스크롤 과정과 탐색으로 시간이 오래 걸린다.
- 하나의 문장을 발화하고 나서, 다른 영역(Domain)의 문장을 발화하기 위해서는 좀 더 많은 시간과 검색 단계가 추가되어 대화를 할 수가 없다.

S전자의 SPH-C3250 PCS폰에 내장된 번역기를 사용하여 그림 2의 예에서와 같이 "호텔 예약"을 진행하던 대화(도메인: [호텔]-[체크인])에서 "화장실"의 위치를 묻고자 하는 경우(도메인: [기타표현]-[질문]), 초기메뉴에서 다시 원하는 문장을 검색해 들어가야 한다. 즉, 영역(Domain)간 변경이 필요한 대화의 경우 기존의 모바일 번역기에서는 실용적으로 사용하기가 불가능하다.

U1: "방을 하나 예약하고 싶습니다."
도메인: [호텔]-[체크인]
U2: "그런데 여기 화장실은 어디 있습니까?"
도메인: [기타표현]-[질문]

그림 2. 영역간 변경이 발생한 대화의 예
Fig. 2. An Example Dialogue with a Domain Shift

또한 기존의 모바일 번역기는 일부 제한된 어휘에 대해 대체(Replacement)가 가능하지만 자유입력을 허용하지 않고 선택 가능한 어휘를 나열해주고 사용자가 선택하는 방식을 취하고 있다. 따라서 시스템에서 제공하지 않는 경우, 자신의 의사를 표현하는 것이 불가능하다.

본 논문에서는 외국어에 익숙하지 않은 한국인이 외국 여행 중에 자유로운 의사 전달이 가능하도록 한⇨영 자유문장 번역이 가능한 모바일 자동번역기 개발을 위하여 경량화된 고속의 한국어 형태소 분석기를 제안한다. 본 논문에서 제안하는 한국어 형태소분석기는 주메모리의 사용을 줄이고 분석 시간 단축을 위하여 키 이벤트 핸들러를 이용한 온라인 분석을 제안한다. 제안된 방법의 주요 특징은 다음과 같다.

- 붙여 쓴 오류가 있어도 띄어쓰기 오류에 대한 전처리 교정 과정 없이도 형태소 분석이 가능하다.
- 키 이벤트 핸들러를 이용한 온라인 분석을 통해 입력 문장의 한글 자소 열(Sequence)을 획득하여 형태소 분석을 위한 별도의 코드 변환 과정 없이 즉시 분석이 가능하다.
- 이벤트 핸들러를 이용한 온라인 분석을 통해 한 어절의 경계가 확인되면 Trie 사전을 이용한 사전 검색 과정이 동시에 진행된다. 상대적으로 매우 느린 사용자의 문자 입력시간동안 어절별 사전 검색을 진행하여 형태소 분석 시 가장 많은 시간을 차지하는 사전 탐색 시간을 줄일 수 있다.
- 코드변환테이블 없이도 자모 열을 획득함으로써 주메모리 사용을 최소화한다. 또한 Trie 사전의 음절별 인덱스를 외부 메모리에서 오버레이(Overlay) 방식으로 접근하여 주메모리의 사용을 줄인다.

2. 모바일폰용 형태소 분석기

기존의 모바일폰용 형태소 분석기에 대한 연구는 주로 형태소분석 전처리과정으로 통신언어에서 발생하는 철자와 띄어쓰기 오류에 대한 교정에 관한 연구가 주종을 이루고 있다. 이를 위해 통계적 방법은 어절간 또는 음절간의 n-gram 정보를 이용하거나[4, 5] Noisy Channel Model을 이용한다[6, 7].

규칙을 이용하는 방법은 언어학적인 지식을 표현한 규칙과 사전 정보를 이용하는 것으로 오류를 포함한 단어로부터 올바른 후보 단어를 생성하기 위해 자소 또는 음절 단위의 변환 규칙을 작성하고, 이로부터 만들어지는 후보 단어와 올바른 단어 간의 유사도 측정을 위해 편집 거리 (Edit Distance) 등을 이용한다. 자주 사용되거나 규칙으로 해결하기 어려운 단어는 "통신용어 사전" 과 같은 별도의 시스템 사전에 수록하여 해결한다[8].

기존의 시스템은 띄어쓰기오류를 교정하는 전처리 과정을 형태소 분석기와 별도로 둔다. 이는 형태소 분석기 자체가 정확한 띄어쓰기 입력을 가정하고 제작되었기 때문에 띄어쓰기 오류를 포함하는 문장에 대해 취약하기 때문이다. 본 논문에서 제안하는 분석기는 기존의 연구에서와 다르게 띄어쓰기 오류를 포함하는 문장에 대해 띄어쓰기 전처리 교정과정 없이 형태소분석을 수행한다. 또한 모바일폰이라는 경량화된 장치의 하드웨어 제약을 고려하여 적은 메모리에서 빠르고 정확한 분석을 목표로 다음과 같은 입력을 가정한다.

- 문법적인 문어체/구어체 문장을 대상으로 한다.
- 모바일폰에서의 문장 입력 시간 단축을 위하여 붙여 쓴 오류와 일부 통신 용어 사용을 허용한다.
- 철자오류는 분석 후보를 지나치게 과다하게 생성하여 분석의 모호성을 증대시키고 이로 인해 최종적으로 번역의 정확률을 떨어뜨리므로 본 논문의 대상 범위에서 제외한다.

이는 모바일폰을 이용하여 외국인과의 자연스러운 대화 수

행을 위하여 빠른 문자 입력과 정확한 번역을 위해 적절한 제약으로 판단된다.

III. 강건한 온라인 한국어 형태소 분석

본 논문의 한국어 형태소 분석기의 가장 큰 특징은 온라인 분석이다. 즉, 번역 대상 문장이 온라인으로 입력되는 상황이다. 전통적인 형태소 분석의 경우, 웹 문서, PDF 나 텍스트 문서를 대상으로 하는 오프라인으로 입력 문장이 입력되는 것이 아니라 SMS와 동일하게 모바일폰의 키패드나 터치스크린에서 문자를 입력하는 과정과 동일하게 형태소 분석을 진행한다.

1. 키 이벤트 기반의 온라인 사전 검색

본 논문에서 제안한 한국어 형태소 분석기는 한국인 해외여행자가 문자 메시지(SMS)와 같이 발화하려는 한국어 문장을 입력하는 과정과 동시에 형태소 분석을 수행한다. 모바일폰의 문자 키패드를 누르는 것과 동시에 키 이벤트 핸들러와 한글 오토마타에 의해 한글 자모가 결정되고 어절단위의 인식과 동시에 Trie를 이용한 사전 검색을 수행한다. 이러한 과정을 통해 문장의 입력이 끝나는 것과 동시에 어절 단위로 모든 가능한 형태소의 사전 검색 결과를 이용하여 형태소 분석을 수행한다.

사용자가 입력 문장을 완성하는 것과 동시에 분석을 위한 가능한 모든 형태소의 사전 정보 획득이 동시에 이루어진다. 이를 통해 형태소 분석시 가장 많은 시간을 차지하는 사전 탐색 시간을 크게 줄일 수 있어 형태소 분석결과를 빠르게 반환할 수 있다.

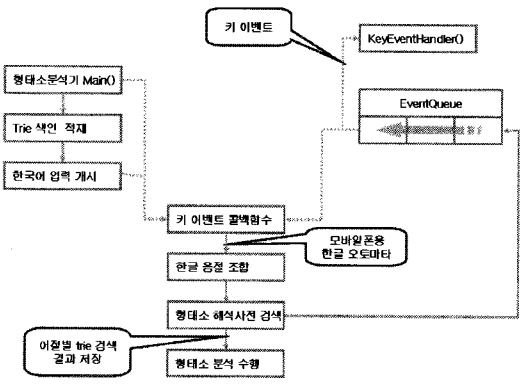

그림 3. 키 이벤트 기반 온라인 형태소 분석
Fig. 3. Key Event Driven On-Line Morphological Analysis

그림 3은 본 논문에서 제안하는 이벤트 기반의 온라인 형태소 분석의 처리 흐름을 보여준다. 첫음절의 Trie 인덱스가 주메모리에 적재된 후 사용자의 입력이 시작된다. 모바일폰의 키패드를 이용한 문자 입력 과정에서 핸드폰에 내장된 한글 오토마타에 의해 자모가 조합된다. 자모가 조합되는 과정에서 키 이벤트 핸들러를 이용하여 완성된 음절의 자모를 보관한다. 이벤트 핸들러에 의해 빈칸(Space) 키 또는 문장의 마지막에서 "번역" 수행을 위한 키 이벤트가 발생하면 한 어절의 완성을 인식하고 어절에 대한 Trie 사전 검색을 수행한다. 이러한 과정을 통해 어절별 자모 열과 Trie 사전 검색 결과를 얻게 된다.

한국어 형태소 분석을 위해서는 음절별 자소 정보가 필수적이다. 즉, "난"이라는 하나의 음절의 가능한 모든 형태소를 분석 하기위해서는 "ㄴ, ㅏ, ㄴ"과 같이 자소 분리가 필요하다. 표준 한글 코드인 KS C-5601-1987 완성형 한글에서 자소 분리를 위해서는 2Byte 조합형 또는 3Byte 조합형 코드 테이블을 이용한 코드 변환 과정이 필요하다. 그러나 본 논문에서 제안하는 온라인 한국어분석에서는 문장입력을 위한 키 이벤트(Key Event)에 따라 모바일폰의 한국어 오토마타를 통해 자소 분리가 자연스럽게 이루어질 수 있다. 따라서 2,350*2Byte (4.6KB) ~ 2,350*3Byte (6.9KB)의 불필요한 메모리를 줄일 수 있다.

2. 온라인 Trie 사전 검색

사용자가 문장을 입력하는 것과 동시에 음절별로 Trie 사전 검색을 수행한다. 그림 4는 본 논문에서 사용하는 Trie 사전의 구성도이다. 본 논문에서는 음절 Trie를 기반으로 압축률을 향상시킨 변형된 음절 Trie 구조를 사용한다[9].

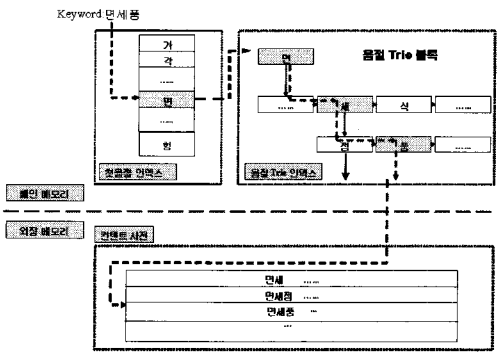

그림 4. Trie 사전의 구조
Fig. 4. Structure of the Trie Dictionary

번역할 문장을 입력하는 과정에서 모바일폰의 한글 오토마타(Automata)에 의해 한 어절의 첫 음절이 확정되면 그 음절의 '음절 Trie 블록'이 주메모리로 적재된다. 음절 Trie 블록은 동일한 음절로 시작하는 모든 표제어의 Trie 색인 블록이다. 음절 Trie 블록은 실행 메모리 (Runtime Memory)를 줄이기 위해 어절 단위로 주메모리에 오버레이된다. 하나의 어절이 입력되거나 한 문장의 입력이 종료되면 Trie 색인 블록을 사용하여 사전 검색이 이루어진다. 이는 형태소 분석시 가장 많은 시간을 차지하는 사전 탐색을 비교적 느린 사용자의 문자 입력 시간동안에 진행하여 분석 시간을 크게 줄인다. Trie 구조의 특징상 한 번의 사전 검색으로 동일한 음절로 시작하는 모든 부분 문자열에 대한 정보를 1회의 사전 검색과정에서 모두 검색된다.

국내 주요 핸드폰 제조사인 S사, L사, SK사의 문자입력 인터페이스는 각각 다음 그림 5와 같다.

그림 5. 주요 모바일폰의 한글 입력 키패드
Fig. 5. Hangeul Keypad Layouts of Major Mobile Phone Manufactures

그림 6은 "이것은 얼마 입니까"란 문장에 대해 S전자의 천지인 방식과 L전자의 나랏글/ez한글의 한글 입력 방식에 따른 사전 검색 과정을 보여준다.

한글 입력기	"이것은"의 키 열 (Key Sequence)
S사	ㅇ ㅣ ㄱ • ㅣ ㅅ ㅇ ㅡ ㄴ
L사	ㅇ ㅣ ㄱ ㅓ ㅅ ㅇ ㅡ ㄴ
완성된 자모	ㅇ ㅣ ㄱ ㅓ ㅅ ㅇ ㅡ ㄴ

한글 입력기	"얼마"의 키 열
S사	ㅇ • ㅣ ㄴ/ㄹ ㅇ/ㅁ ㅣ •
L사	ㅇ ㅓ ㄹ ㅁ ㅏ
완성된 자모	ㅇ ㅓ ㄹ ㅁ ㅏ

한글 입력기	"입니까"의 키 열
S사	ㅇ ㅣ ㅂ ㄴ ㅣ ㄱ/ㅋ/ㄲ ㅣ •
L사	ㅇ ㅣ ㅂ ㄴ ㅣ ㄱ/쌍자음 ㅏ
완성된 자모	ㅇ ㅣ ㅂ ㄴ ㅣ ㄲ ㅏ

그림 6. 모바일폰(S사, L사)에서의 한글 입력과정
Fig. 6. Hangeul Input Procedure on Samsung and LG Mobile Phones

3. 붙여쓰기오류 허용 강건한 형태소분석

회화번역기 사용시 빠른 입력을 위해 띄어쓰기를 하지 않은 경우를 허용한다. 즉 사용자는 "이것은 얼마 입니까?" 대신에 "이것은얼마입니까"와 같이 입력할 수 있다. 이와 같이 붙여쓰기 오류를 허용하는 입력 문장에서의 한국어 형태소 분석은 일본어와 중국어의 형태소 분석과 유사하다. 본 논문에서는 [9]에서의 분석 방법을 한국어 온라인 분석에 적용하여 붙여 쓴 문장에 대해 띄어쓰기 교정을 위한 전처리 과정 없이 강건한 형태소 분석을 수행한다.

그림 7은 붙여쓰기 오류를 포함하는 문장에 대한 한국어 형태소 분석기의 분석 과정을 보여주고 있다. 인접한 형태소가 사전에서 검색된 경우에는 접속 가능 여부를 조사하여 접속 가능한 경우에만 사전 검색 결과를 자소별 차트(Chart)에 저장한다. 그림 7에서 A, B, C, D, A1, B1, B2, C1, D1, D2, C3은 Trie 사전에서 검색된 형태소를 의미하고 화살표는 접속이 가능함을 의미한다.

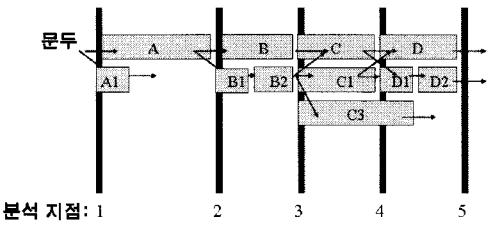

그림 7. Trie 사전 탐색 및 형태소 분석 과정
Fig. 7. Procedure of Trie Dictionary and Morphological Analysis

맨 처음 어절의 경우에는 '문두(Sentence Begin)'와 문장의 중간에 나타나는 어절의 경우에는 직전 어절의 마지막 형태소들과 접속 가능한 경우로 제한한다. 그림 7에서 온라인 사전 탐색 과정을 통해 문두와 접속 가능한 두개의 형태소 A와 A1 두 개의 후보가 Agenda list에 기록되어 있다. Agenda List는 Priority Queue로 관리되며 우선순위가 높은 후보가 선두에 위치한다. 우선순위는 다음 식 (1)과

같이 단어의 길이의 제곱과 어휘생성 확률(Lexical Generation Probability) 그리고 품사전이 확률(POS Transition Probability)을 이용하여 계산된다.

우선순위=단어길이2*어휘생성확률*품사전이확률 ····· (1)

만약 A가 우선순위가 높다면 A를 분석결과 스택에 저장하고 분석 지점 2에서 다시 두개의 분석 후보(B, B1)가 가능하다면 B와 B1은 다시 Agenda list에 기록된다. 후보 B가 우선순위가 높다면 B를 분석결과 스택에 저장하고 분석 지점 C에서 3개의 분석 후보(C, C1, C2)를 얻게 된다. 다시 C가 우선순위가 높다면 C를 분석결과 스택에 저장하고 분석 지점 4에서 동일한 방법으로 분석을 진행한다.

만약 분석 지점 4에서 얻은 단어 D와 D1이 C와 접속이 불가능하다면 분석결과 스택에 저장된 C를 Stop Word List에 보관하고 분석 지점 3으로 백트랙킹 (Backtracking)한다. Stop Word List는 다른 경로를 통해 동일한 위치에 도달한 경우 불필요한 분석을 반복 수행하는 것을 방지한다. 예를 들어 분석 지점 2까지 백트랙킹하여 다른 분석 경로인 A-B1-B2를 통해 분석 지점 3에 도달한 경우 C, C1, C3의 사전 검색 결과가 이미 차트에 존재하므로 분석지점 3에서의 Trie 사전 탐색을 하지 않는다. 또한 C가 Stop Word List에 있으므로 A-B1-B2-C로의 분석은 이미 실패한 시도임을 알 수 있으므로 C 다음으로 우선순위가 높은 C1로 분석을 진행한다. 이러한 분석 과정을 통해 가능한 n개의 후보 열이 최적우선 (Best-First) 탐색 기법을 사용하여 얻어지고 최종 분석 후보를 결정한다.

4. 통신용어 교정 전처리기

최근 통신용어의 발달로 문자메시지(SMS) 사용시 문자 입력 시간을 단축하기 위해 다양한 형태의 통신용어가 사용되고 있다[10]. 이러한 표현은 끊임없이 매우 빠른 속도로 새로운 형태가 출현하므로 단순하게 이들 형태를 사전에 모두 수록하는 것으로는 해결할 수 없다. 따라서 본 논문에서는 [11]에서와 같이 웹 게시판과 같이 통신용어가 사용된 대량의 말뭉치를 자동으로 수집하고, 미등록어를 추출하는 [11]에서와 달리 기존 사전을 이용하여 기계학습을 통해 자소 단위의 철자 교정 모델을 자동 구축하는 방법을 취한다. 통신 용어에서 나타나는 상당량의 철자 오류는 워드프로세서에 내장된 철자 교정기와 마찬가지로 구성 자소의 단순 변환만으로도 교정이 가능하며 약 35%의 철자 오류가 단순 자소 변환만으로 교정이 가능한 경우임을 밝히고 있다[7].

철자 교정 모델로는 자소단위의 교정모델[7, 12]과 음절 단위의 교정 모델[13]이 있다. 자소 단위의 철자 교정 모델은 어절이나 음절 단위의 학습모델에 비해 모델의 크기를 줄일 수 있고 자료 부족 현상 (Data Sparseness)에 효과적으로 대응할 수 있으며 모델이 학습한 통계 정보의 데이터 크기를 줄일 수 있어 모바일폰에의 적용이 용이하다.

본 논문에서는 다음과 같은 형태의 자소 및 음절의 통합 모델을 제안한다. 원형 복원 규칙을 사용하여 철자오류를 정정하기 위한 후보를 제시하고, 이는 Trie 사전 검색시 분석 후보 생성에 이용된다. 철자 교정 규칙의 형태는 다음 그림 8과 같다.

```
좌문맥{자소|음절}+/현재{자소|음절}+/우문맥{자소|음절}+  =〉 { 교정규칙+ }
```

그림 8. 철자 교정 규칙의 형태
Fig. 8. Template of Error Correction Rules

그림 8에서 '+'는 1개 이상을 의미하고 '{자소|음절}'은 자소 또는 음절을 의미한다. 예를 들어 "있자나요"==〉"있잖아요"의 경우 철자 교정 규칙은 다음과 같다.

```
있/자/나 =〉 {Replace(0, "잖"); }
```

교정 규칙은 Replace(), Add(), Delete() 등의 철자 교정 함수로 이루어지며 0은 음절전체, 1은 초성, 2는 중성, 3은 종성을 의미한다. 규칙 기술시 자소 또는 음절 단위로 기술이 가능하며 빠른 수행을 위해 flex 스캐너 (Scanner) 제작 도구를 사용하여 유한상태 전이 네트워크(Finite State Transition Network; FSTN) 형태로 이용된다.

IV. 구현 및 실험

본 논문에서 제안한 모바일폰용 경량화된 한국어 형태소 분석기는 MS Visual Studio 8을 사용하여 Windows Mobile 5.0 플랫폼과 Scanner 제작 도구인 flex[14]를 사용하여 개발되었다. flex는 강력한 정규 표현(Regular Expression)을 지원하므로 날짜, 영어 인명, 숫자표현 등을 효과적으로 대응할 수 있다. 철자 교정 규칙은 규칙 컴파일러에 의해 C Program으로 변환되고 형태소 분석기와 통합된다.

표 1은 여행용 영어 회화집[15] Part 16과 Part 17에서 추출한 506문장의 문서(18.96 Kbyte)에 대한 정보이다. 테스트 문서는 문장당 평균 5.3어절, 문장당 평균 11.8음절이고 문장 당 평균 형태소 수는 16.7개이다.

표 1. 테스트 셋
Table 1. Test Set

구분	Part16	Part17
Domain	해외여행	호텔
문장 수	269	237
어절 수	1,235	1,431
형태소 수	2,764	3,229
음절 수	3,963	4,607
파일 크기	8.86KB	10.1KB

표 2는 여행용 회화집 테스트 셋에 대한 실험 결과이다.

표 2. 형태소 분석 결과
Table 2. Evaluation Results of the Morphological Analysis

사전탐색 횟수	8,570
접속체크 횟수	35,649
백트래킹 횟수	1,117
형태소 수	9,248
문장당 사전탐색 횟수	16.9
문장당 접속체크 횟수	70.5
문장당 백트래킹 횟수	2.2
실패한 문장 수	0

실험 결과 문장당 평균 사전 탐색 횟수가 16.9회, 문장 당 평균 2회의 백트랙킹을 수행하여 최적의 분석 결과를 얻었다. 분석 실패하는 문장은 없었다. 그림 9는 본 형태소 분석기가 적용된 한영 모바일 번역기의 실제 동작 화면이다.

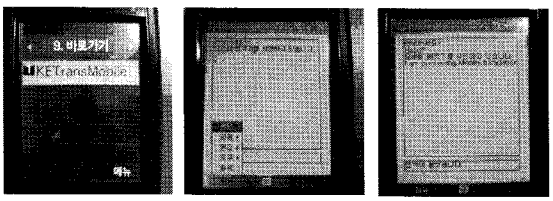

그림 9. 한영 모바일 번역기 실행 예
Fig. 9. A Demonstration of the Korean-English Mobile Translator

형태소 분석기를 포함한 전체 번역 엔진 크기는 6.31MB이고 사전의 크기는 47.2MB이다. 초당 번역 속도는 13.3 Byte/sec이다.

표 3은 S전자의 SCH-M450 모바일폰에서 본 논문에서 제안한 온라인 형태소 분석을 적용하기 전후의 한영 번역기의 메모리 사용량과 반환 시간 (Turn-around time)을 평가한 결과이다.

표 3. 온라인 분석 적용 결과
Table 3. Evaluation Results of the Online Morphological Analysis

구 분	적용 전	적용 후
실행 메모리(MB)	3.6	3.4
문장당 평균 반환 시간(sec)	1.69	1.42

메모리 사용량은 적용 전 3.6MB에서 적용 후 3.4MB로 5.8%를 줄일 수 있었다. 평균 반환시간은 적용 전 문장당 1.69초에서 적용 후 1.42초로 19.0%의 속도 개선을 이루었다.

그림 10은 본 형태소 분석기가 적용된 한영 모바일 번역기의 실제 동작 화면이다. 왼쪽부터 S전자의SPH-M4655 (L이동통신), SCH-M620 (S이동통신), 그리고 SCH-M450 (S이동통신) 모바일폰이다.

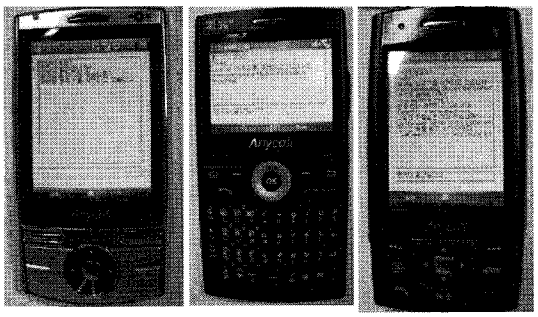

그림 10. 한영 모바일 번역기 시제품
Fig. 10. Prototypes of the Korean-English Mobile Translator

V. 결 론

본 논문에서는 모바일 한영 번역기를 위한 강건하고 경량화된 한국어 형태소 분석기를 설계 및 구현하였다. 빠른 반응 시간을 위하여 Trie를 이용한 온라인 분석을 수행하며 사용자

의 문자 입력 시간 단축을 위해 붙여 쓴 문장에 대해서도 별도의 띄어쓰기 오류 교정 과정 없이도 강건한 형태소 분석을 수행하는 방법을 제안하였다. 제안된 방법은 한국인 여행자용 한영 모바일 번역기에 적용하여 메모리 사용량은 5.8% 줄이고 평균 반환 시간은 19.0% 개선하였다.

본 논문에서 제안하는 강건하고 경량화된 온라인 한국어 형태소 분석기는 낮은 메모리 요구량과 빠르고 정확한 분석 성능으로 휴대폰/PDA 음성다이얼링, 로봇의 제어, 일정/날씨/교통 등 대화형 정보검색, 텔레매틱스 내비게이션, 홈네트워크 음성명령, 차세대 PC용 자연언어 인터페이스, 무인 콜센터, 음성타자기, 음성 SMS, 멀티미디어 음성검색, 휴대용 자동번역기 등 다양한 모바일 장치에 적용할 수 있을 것으로 기대된다.

참고문헌

[1] 박세영, 김병수, 이경일, "모바일 다국어 번역기술 동향과 그 구현 사례," 정보과학회지, 제 24권 제 1호, 37-47쪽, 2006년 2월.

[2] R. Isotan, K. Yamababa, et. al., "An Automatic Speech Translation System on PDAs for Travel Conversation," Proc. Fourth IEEE International Conference on Multimodal Interfaces, pp.211-216, October 2002.

[3] J. Zhang, X. Chen, J. Yang, and A. Waibel, "A PDA-based Sign Translator," Proc. Fourth IEEE International Conference on Multimodal Interfaces, pp. 217-222, Oct. 2002.

[4] D. Lee, H. Rim, and D. Yook, "Automatic Word Spacing using Probabilistic Models based on Character n-grams," IEEE Intelligent Systems, Vol. 22, No. 1, pp. 28-35, Jan.-Feb. 2007.

[5] S. Kang and C. Woo, "Automatic Segmentation of Words Using Syllable Bigram Statistics," Proc. Natural Language Processing Pacific Rim Symposium, pp. 729-732, Nov. 2001.

[6] J. Gao, M. Li, and C.-N. Huanh, "Improved Source-Channel Models for Chinese Word Segmentation," Proc. 41st Annual Meeting of the ACL, 2003.

[7] 노형종, 차정원, 이근배, "띄어쓰기 및 철자 오류 동시교정을 위한 통계적 모델," 제 18회 한글 및 한국어 정보처리 학술대회 (HLT06) 발표논문집, 포항공대, 25-31쪽, 2006년 10월.

[8] 강승식, "한글 문장의 자동 띄어쓰기를 위한 어절 블록 양방향 알고리즘," 정보과학회논문지, 제 27권, 제 4호, 441-447쪽, 2000년 4월.

[9] S.-H. Yuh, H.-M. Jung, et. al., "FromTo/JK: A Japanese-Korean Machine Translation System," Proc. Natural Language Processing Pacific Rim Symposium, pp.613-616, Dec. 1997.

[10] 권오경, 서은아, "*인터넷 통신어휘 사전*," 동인, 15-441쪽, 2002년.

[11] 박소영, "웹문서에서의 출현빈도를 이용한 한국어 미등록어 사전 자동 구축," 한국컴퓨터정보학회논문지, 제 13권, 제 3호, 27-33쪽, 2008년 5월.

[12] M. Li, Y. Zhang, M. Zhu, and M. Zhou, "Exploring Distributional Similarity based Models for Query Spelling Correction," Proc. 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp. 1,025-1,032, July 2006.

[13] J.-H. Byun, S.-Y. Park, K.-S. Han, and H.-C. Rim, "A Rule-based Spelling Correction Model Learned from Corpus for Refining Chatting Message," Proc. The First Europe-Korean Workshop on Spoken Dialog System Technology, Dec. 2008.

[14] D. Dougherty, "*Lex & Yacc*," O'Reilly & Associates Inc., pp.27-179, 1990.

[15] 이보영, "*이보영의 영어회화사전*," 두산동아, 716-786쪽, 2002년.

저 자 소 개

여 상 화

1990 인하대학교 전자계산학과 학사

1992 인하대학교 전자계산공학과 공학석사

2006. 서강대학교 컴퓨터학과 공학박사

1992.1~2000.2. 한국전자통신연구원 선임연구원

2000.3.~2001.4. L&H Korea 책임연구원

2001.5.~2002.8 유니소프트 책임연구원

2002.8.~현재 경인여자대학 정보미디어학부 부교수

관심분야: 자동번역, 한국어정보처리, HCI